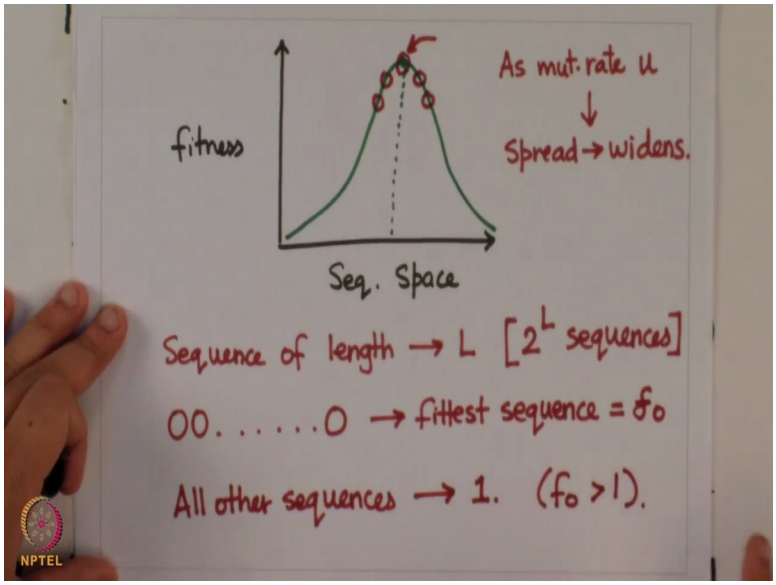
**Introduction to Evolutionary Dynamics**
**Prof. Supreet Saini**
**Department of Chemical Engineering**
**Indian Institute of Technology, Bombay**

**Lecture - 20**
**Modelling evolution on fitness landscapes – 2**

Hi everyone. Let us continue our discussion of localization of fitness, localization on a fitness peak of DNA sequence which was considered in terms of zeros and ones a binary DNA sequence and let us pick up from where we had left off last time.

(Refer Slide Time: 00:38)




So, I just want to start by reemphasizing the question that we are trying to answer if this is the sequence space and y axis is fitness. So, this is a highly simplified one dimensional sequence space which is an oversimplified picture and if we have a fitness landscape such as this a single peak fitness landscape and this represents the sequence which corresponds to the maximum fitness of an genotype which belongs to this the sequence space the question that we are trying to answer is that we understand that at steady state of this system.

And if we have finite mutation rates the steady state distribution of individuals is going to be centered around this peak sequence. And we are going to have majority of individuals belonging to this particular sequence, but we would also have some nonzero fractions of individuals which belong to the neighboring sequences. And we will have

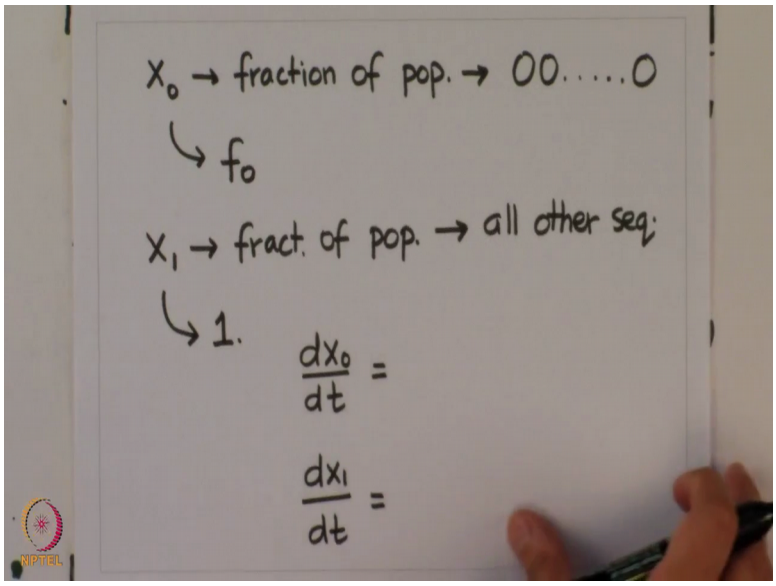
the steady state distribution around this sequence which corresponds to maximum fitness. And as mutation rate increases; as mutation rate u increases the spread around the peak widens.

And the question that we want to understand is that; what happens if we keep increasing this mutation rate further and further do we keep getting into a situation where the spread around the fitness peak just widens or this is there a qualitative change in the behavior of this localization that we see as we keep increasing mutation rate. And to analyze the system the framework that we had started with the last time the last time around was that, we have a sequence of length L.

And because we are considering DNA to be binary this corresponds to a total number of 2 to the power L sequences. So, the stretch of DNA that I am interested in is of length L, but DNA being binary in this case we have 2 to the power L sequences and the way we had defined our fitness landscape was that the sequence of all zeros. So, this particular sequence is the fittest sequence among all the 2 to the power L possible sequences and the fitness given to this particular sequence was f naught and all other sequences have a fitness equal to 1.

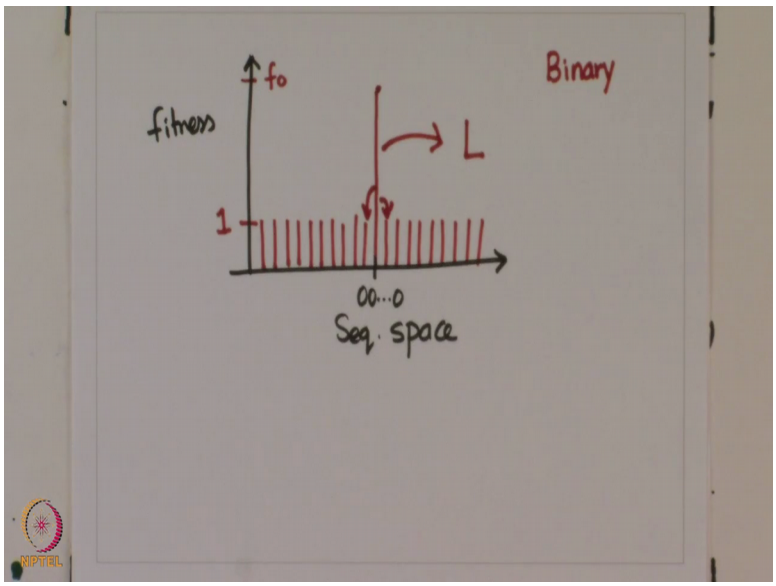
So, when we say fittest this is fittest this clearly implies f naught is bigger than one and what we had started was that we are interested in developing dynamical equations representing these 2 groups of sequences one which represents the fittest sequence with a fitness equal to f naught and the other group of sequences is all the other sequences clubbed together all of which have a fitness equal to 1.

(Refer Slide Time: 04:11)




So, then we say that let X naught be the fraction of population with the sequence all zeros. So, this is the fraction of population which is growing at growth rate f naught the fitness of the sequence and let X 1 be the fraction of population which has all other sequences and doing that is fine, because all other sequences have the same I. Same fitness which is equal to 1 and what we are interested in is coming up with dynamical equations coming up with expressions for d X naught by d t and d X 1 by d t. Before we get there let us try and visualize the fitness landscape that we are talking about here.

(Refer Slide Time: 05:16)

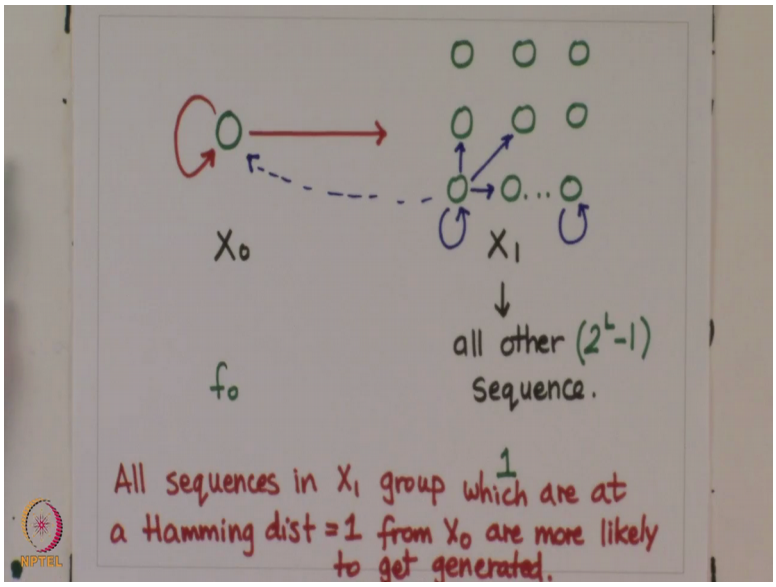


And what this corresponds to is we have sequence space here is my sequence of all zeros this is fitness and the and then my fitness landscape is such that this sequence corresponds to fit fitness equal to f naught and every other sequence in this plane has a fitness equal to 1.

Every other sequence they are all equal to 1. So, that is the fitness landscape that i am talking about and again this is a idealized representation of this fitness landscape which is a higher dimensional quantity which is being represented in this 2 dimensional plane we had discussed last time that this being a binary sequence DNA being binary here the number of neighbors that this quantity that this particular sequence has the one mutant neighbors is equal to l, but in this highly simplified representation this sequence has only 2 neighbors which is a consequence of representing this higher dimensional structure on a plane.

But for our purposes we will consider that these 2 neighboring sequences represent all L neighbors associated with this particular sequence in this hyper dimensional space which represents the fitness landscape of the system that we are talking about. Now, again this is my simplified vision of the fitness landscape and we are interested in coming up with expressions for d X naught by d t and d X 1 by d t. So, how do we go about thinking about coming up with these expressions and the way we will start thinking about this is let us divide this entire.

(Refer Slide Time: 07:19)



Let us divide this entire 2 L number of sequences into 2 groups and in the first group is just X naught which is the fraction of population belonging to the fittest sequence and then the other group is X is represented by X 1 which is all other sequences fraction of the population which belong to all other sequences everything except than all 0 sequence. So, this one is we know is growing at fitness f naught this one every sequence is growing at with the fitness equal to 1.

But this X 1 is comprised of 2 to the power L minus 1 sequences and each one of them can be represented by a node there are 2 to the power L minus 1 such nodes the X naught sequence is just one sequence. So, that can be represented by 1 node and what we have to pay attention to here is that every time a division event happens what happens in terms of this node and edges diagram that we have over here and what happens is that whenever X naught divides if during that division event mutation does not happen.

So, the progeny is an exact replica of the parent genotype that it is coming from we get this particular arrow and every time there is a mutation the progeny cannot belong to group X naught the progeny has to belong to one of the sequences in X 1. So, every time a mutation happens frequency of X 1 increases if you think about it which ones which particular sequences in the X 1 are more likely to get generated via this via this mutational event not all nodes are equally likely to be generated via this mutation event and that is so because mutations are typically going to be small mutation rates are typically going to be small.

So, the probability that an error is going to happen during the replication event is also going to be small. So, all those sequences in the X 1 group so, all sequences in the X 1 group which are at hamming distance equal to 1 from the sequence X naught are more likely to get generated when this mutation event happens it is sort of easy to visualize that this group will have sequences which are hamming which are at hamming distance 1, 2, and so on and so forth.

And those which are at a lower hamming distance which means those sequences which differ from this particular sequence at only one nucleotide position would only require one mutation to happen to this mutant such that the resultant progeny belongs to one of these in X 1 those sequences which are at hamming distance to from this particular

sequence would require that 2 mutation events happen while the replication a parent from this group is dividing.

And 2 mutations should happen such that the resulting sequence belongs to group X 1 and is at a hamming distance equal to 2 from the genotype corresponding to the individuals in X naught and those are going to be harder, because mutation rates are going to be typically small; so anyway when X naught divides you could have generation of another X naught or you could have generation of one of the individuals which belongs to group X 1.

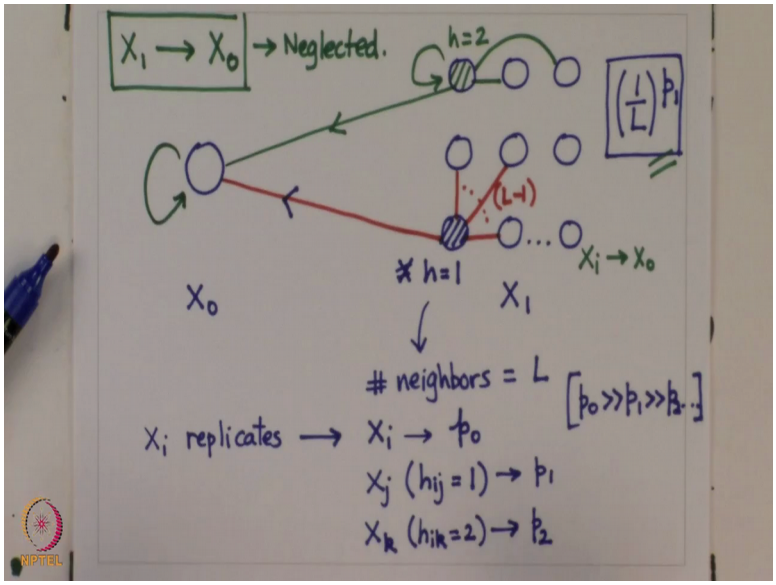
What happens in the other case when individuals in the group X 1 divide? Again you are going to have the similar process that sort of more often than, naught when an individual divides mutation is not going to happen. And you are going to get a progeny which also has a genotype equal to x 1, but every once in a while there will be error and now these errors are going to mean that the progeny belongs to the progeny can have one of two fits when error happens during replication of one of the individuals in group X 1 when error happens when suppose an individual with this particular sequence is dividing and an error is happening during the process of replication the progeny which is a mutant could either be a sequence which also belongs to the group X 1 or it could be a sequence which corresponds to the sequence x naught.

So, we do not know when that when a mutation event happens for an individual in group X 1 where does the mutant go in this case because this is the only sequence which belongs to group X naught it is obvious that every time a mutation happens the progeny belongs to group X 1. But when a mutation happens during replication of one of the sequences which belongs to the group X 1, we do not know where the progeny lies and that is the challenge that we have to resolve before we can come up with the dynamical equations which represent this system.

Let us think about this a little bit more. So, what is happening that every time what is happening is that every time and individual of sequence individual with sequence which belongs to X 1 divides it is going to give me a mutant sequence which could belong to either one of the sequences in group X 1 or it could give me the sequence which is the sequence corresponding to the X naught group which is the fittest sequence how do I

resolve this conflict and how do I get a quantitative estimate of which of the ones is more likely.

(Refer Slide Time: 14:06)




So, if I look at that again these are my nodes which belong to the group X 1 and this is the node which belongs to group X naught let us look at one of the individuals which belongs to group X 1 such that x such that the hamming distance between this individual and the X naught sequence is equal to 1. Now how we understood that how many neighbors does this individual have how many neighbors does this individual have and a neighbor being defined as all those sequences which are at a hamming distance equal to 1 and that number of neighbors the number of neighbors is equal to Laughter, because one distant mutants.

Means, that those sequences differ from the sequence of this individual at one nucleotide position only and that one difference could be at one of at any one of the L positions along the length of the DNA hence the number of neighbors is equal to L. And if the hamming distance between this individual and the X naught sequence is one this is one of the is one of the edges that is present in the graph and the other L minus 1 neighbors L minus L minus 1 neighbors of this individual are just present in the X 1 group only. So, these number L minus 1.

So, now if we if we sort of ignore the probability of generation of a mutant is going to be smaller if the number of mutants if the number of mutations required to go from one

sequence to another sequence increases for instance it is most likely that when in sequence x i replicates you get the progeny has sequence x i let us say that the probability of this happening is p 0. P 0 means, probability that 0 errors happened during replication the probability that when x i replicates you get a sequence x j such that hamming distance between h i j is equal to 1.

That means, during the replication process you get a sequence where one nucleotide error happened and the other L minus 1 nucleotides were replicated correctly let us say the probability corresponding to this is p 1 pro which stands for probability that during the replication process one error took place. Similarly let us say x k represents that when x i replicates the sequence that got generated was such that the hamming distance between h i k between sequence i and k was 2 and let us say this is p 2 and so on and so forth.

Now we know how mutation grades of organisms are present in nature and we know that these probabilities are such that p naught is much bigger than p 1 is much bigger than p 2 and so on and so forth; DNA replication machineries are for the most part very very accurate and they have error repair systems which makes sure that not too many errors happen when DNA is replicating itself. So, what; that means is that every time an error happens during replication of this individual most likely that the resultant progeny is going to belong to one of the L neighbors associated with this particular node? So, the chance that the mutant of this parent turns out to be an X naught is equal to 1 by L which is one by L times p 1.

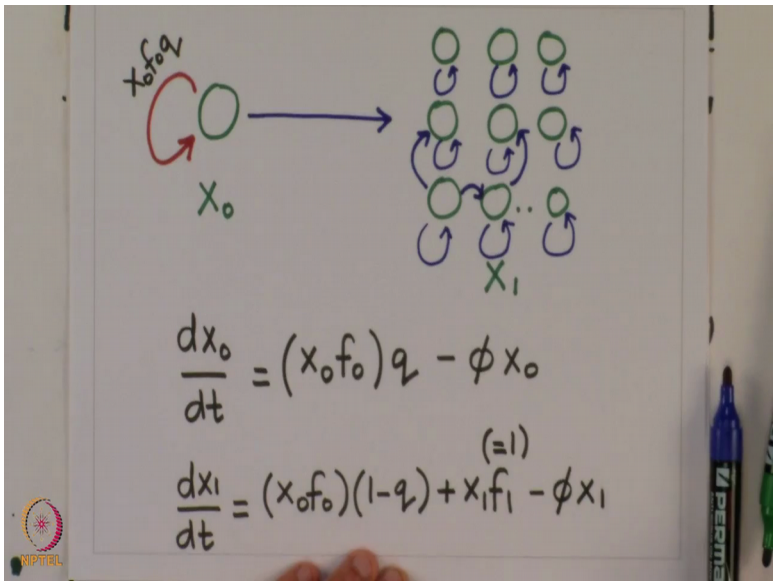
And p 1 remember is always going to be much less than p naught. So, that is a very crude estimate of the probability that when a when parent of this sequence divides the progeny acquires one mutation and that one mutation is such that it leads me to genotype corresponding to the fittest individual which is X naught this event happens that is a very crude estimate of that event happening here. Let us look at one more node associated with this which is at a hamming distance 2 from the sequence X naught. Now whenever an individual belonging to this sequence divides more often the naught it is going to lead to a progeny which has an identical sequence of a identical sequence as compared to the parent should a mutation happen it is more likely to lead to an individual which is at hamming distance equal to 1, because chances are if there were no if replication process was not entirely faithful one error would have happened.

So, there is a chance p 1 that the resultant progeny belongs to the sequences which are at hamming distance one from this particular sequence this sequence is at a hamming distance 2 and hence only if there are 2 mutants; 2 mutations happen which during the during the replication process for this individual lead may to would this individual lead to production of a progeny which belongs to this particular genotype. And that probability is very small compared to this probability, because p 1 is much bigger than p 2 the chances that during the replication process more than one error happened in genomes in an organism's genome of length L are smaller than one error happening during that process and so on and so forth. So, we can go to each particular node at a at a given hamming distance on this on this group of sequences which is being referred to as X 1.

And see that as hamming distance from the sequence X naught increases the probability that one of the progenies a acquires a mutation and a sequence x i converts to X naught decreases very rapidly as the hamming distance between the sequence i and X naught increases. So, increasing hamming distance makes the likelihood of generation of sequence X naught from replication of the sequence x i very very unlikely which is the trick that we are going to use in our analysis here that because of this analysis that we had talked about we are going to ignore that the X 1 group could lead to a progeny belonging to sequence X naught this event can be safely neglected because this event is going to be.

So, rare because of this reason that we had just talk they have that we just talked about we are going to ignore this very very small contribution of X naught sequences being generated via mutation of sequences which belong in the X 1 group that is the contribution that is very very small as compared to say this particular contribution which is number of X naught individuals being generated by division of X naught individuals themselves and no occurs and no errors occurring during that replication process. So, this contribution will be very small as compared to this contribution, and hence we ignore that. So, where does that lead us what that leads us to is let us go back to the node and edges diagram and see where how can we develop our equation further.

(Refer Slide Time: 23:06)




So, we have individuals in the X naught group and we have individuals in the X 1 group now what we are saying is that when it comes to we are interested in when is X naught generated and X naught is generated whenever X naught a sequence individual with sequence X naught divides and there are no errors happening.

What we are ignoring here is the when an individual belonging to sequence which belongs to the group X 1 divides and mutations happen and which leads to an individual belonging to group X naught that is the contribution that is being ignored here and when are individuals belonging to group X 1 being generated. There are 2 ways to do that one is that when replication is error free each sequence is giving rise to progenies which are identical with its own sequence.

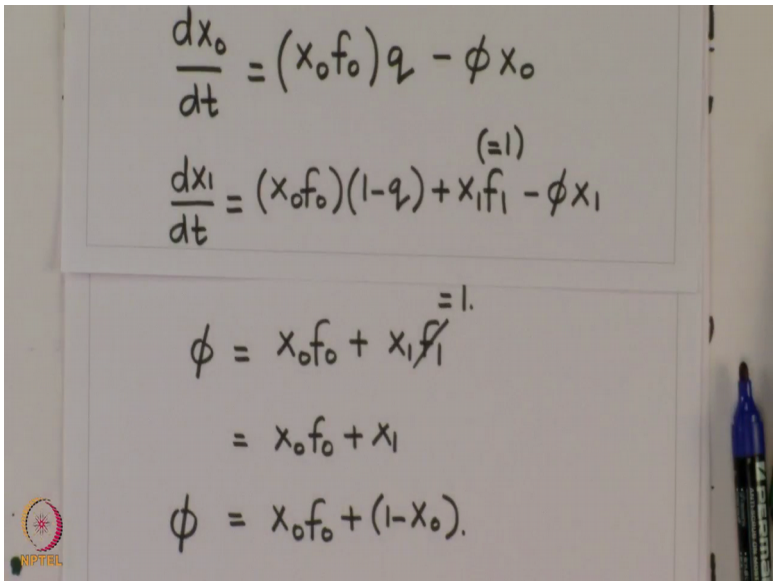
So, you have this particular way of generating individuals with whose sequence belongs to group X 1 and in addition you have X 1 individuals being generated whenever there are there is error happening. So, every time error happens during replication of X naught you get an individual which belongs to group X 1 every time an error happens during replication of X 1 you get an individual which belongs to another belongs to another sequence, but that sequence also belongs to group X 1 and this is true, because we have ignored the generation of sequences to group X naught when errors happen during happen during replication of individuals and group x 1.

So, if that is the case then d X naught by d t can be written as X naught times f naught which gives me the rate at which the individuals belonging to genotype X naught are growing and this should be multiplied by q which is the probability that the replication process was error free. So, this quantity represents this arrow X naught f naught q minus 5 which is the mean fitness times x naught.

Because this is the only generation term and we have deaths taking place in the system to account for a constant population size similarly d X 1 by d t can be estimated as first let us account for this arrow which is X naught f naught which is the total rate of reproduction of individuals belonging to genotype corresponding to this group times one minus q whatever fraction is not replicating faithfully this resulting sequences are not identical with sequence corresponding to this is leading me to an individual which belongs to the group X 1.

In addition I also have X 1 times f 1 which is the frequency of this entire group times its own fitness which we have said is just equal to 1 we do not have a fraction here because we are saying every time a mutation happens in group one the resulting progeny also remains in group one and that is assumption is being enabled. Because we have ignored the generation of sequence X naught every time a mutation happens when an individual belonging to X 1 is dividing minus phi times X 1. So, these are my dynamical equations and what is phi? Phi is the variable again that we have been talking about which represents the mean fitness.

(Refer Slide Time: 27:19)




So, phi is equal to mean fitness which is X naught f naught plus X 1 f 1 f 1 is just equal to 1. So, this is X naught f naught plus X 1 and I know X 1 is just 1 minus X naught that is equal to phi. So now, I have my dynamical equations here and I have a description of phi only in terms of X naught. The next steps you should be able to guess. The next steps in our analysis here, what we can do is substitute and express this expression for X naught in this equation that gives me a differential equation d X naught by d t which will solely be in terms of X naught. Now we have this phi variable, but we can write phi in terms of X naught. Then X naught will be the only variable which is present in this equation and that is 1.

We will analyze the steady states and their stability is associated with this particular equation and that is something we will start off our next lecture with.

Thank you.