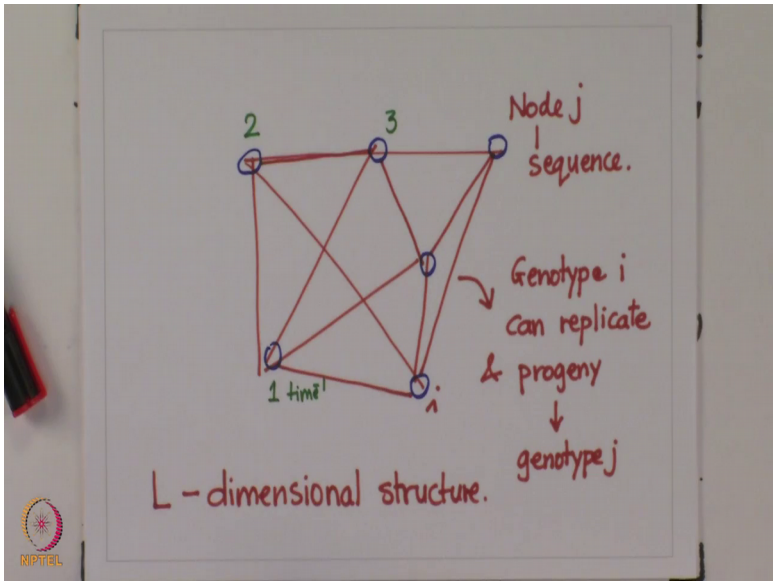
Introduction to Evolutionary Dynamics
Prof. Supreet Saini
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture – 15
Properties of Fitness Landscapes and Quasi–species

Hi everyone. Let us continue our discussion of fitness landscapes. So, we have been discussing in the last lecture of movement of bacterial populations on a fitness landscape. And we talked about how populations might move in might move on this landscape where they might get to the global optimum which represents which corresponds to the sequence which is growing at the maximum possible growth rate or stochastic events. Such as, we do not know a priori which mutations are going to happen in the culture in a we do not know whether mutations are going to happen which mutation is going to arise at what point in the sequence and that might lead to populations getting stuck in a local optimum which represents a node which is fitter than all of it is neighbors, but it does not represent the highest growth rate among all the nodes that are available in this graph.

(Refer Slide Time: 01:12)



So, this graph that we have been talking about can be represented such as this where each node represents a sequence which is what we have been talking about. And now connections between these nodes represent the fact that node I can muted can a bacterium at note I can divide and acquire a point mutation and give rise to the sequence

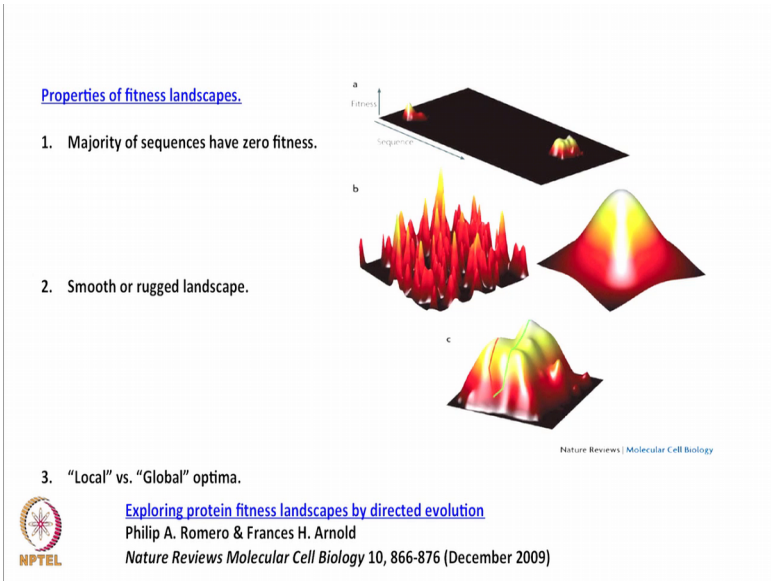
which is corresponding to node j, and you have a graph like structure suggest this. So, this is a node which corresponds to sequence and an h represents the fact that node I and j genotype I can replicate and mutation event can happen a single point mutation which can lead to the progeny having genotype j.

In that sense this represents a 2 dimensional structure, but what is going to happen for real fitness landscape is that there is going to be a lot of criss cross and it is not going to be as easy to visualize as that we have been doing in this course. So, far the real fitness landscapes is going to be a very hyper dimensional structure. And in fact, the dimension dimensionality of the fitness landscape is going to be L. This here what we have described on this slide is in L dimensional structure. And what we are going to impose on this is another dimension which is imagine that to be perpendicular to this plane and the height which corresponds to each of these nodes represents the fitness value associated with that particular node.

So, if this node if the sequence at this node is growing at value one per time. And if this node is the growth rate associated with this particular node is 2 and 3 and so on and so forth. Imagine there to be this axis which is perpendicular to this plane and the value of this node is the height is represented by the growth rate associated with that node. So, when you have a population at this particular genotype, it is actually existing at value of 1 and when there is a mutation event that happens such that a progeny corresponding to this particular sequence is produced that particular progeny lies at this point which is at a height of 2 units. So, there is been an increase in fitness when one of the individuals of the population has moved from this note to this particular node.

So, that increase represents that increase in height represents increase in fitness and as we saw in the last lecture populations will keep try to find keep try to find ways along this fitness landscape map such that there is an increase in fitness and they keep on going to higher and higher levels eventually getting trapped in either local optima for being able to reach the global optima associated with the map that we have talking about. So, what I am going to do is now move to a couple of slides through which I want to emphasize some of the salient features which are associated with fitness landscapes.

(Refer Slide Time: 05:04)



So, let us move to the slides and these slides are taken from a review paper which was published in 2009 and they talk about fitness landscapes the emphasis of this paper when there talking of landscapes is from the perspective of proteins.

But the results hold as far as DNA protein la and DNA fitness landscapes are concerned as well. So, this is one of the figures that have taken from this from this paper the reference is at the bottom of the page. And we want to emphasize 3 particular characteristics which are associated with landscapes. The first one is that majority of sequences have 0 fitness. So, on the figure that you see on the right here again the 2 dimensional plane represents all particular sequences you can think of these as protein sequences or DNA sequences whose length is equal to L. So, these are connected why an edges and nodes which we cannot see because there are so many of them and this axis here represents the fitness associated with that particular protein or the DNA sequence.

So, height of a particular node represents how fit that individual is which is carrying this particular sequence. So, as you can see from this example here that most of the sequences do not correspond to any fitness. The fitness level of most of the sequences is just equal to 0. That makes a lot of intuitive sense because if you think of DNA sequences if you think of the example that we have been talking about if you think of sequences as sequences belonging to bacteria and we are measuring their growth rates. Now when we are talking of all possible sequences of length L that includes many

polymers one of the sequence is going to be all is that represents one of the sequences in the sequence space that we have been talking about now clearly all a's is just a DNA polymer of length L it could not per hap it could not possibly correspond to living bacterium which is actively able to grow and divide.

One of it is neighbors one of the neighbors of the node which corresponds to sequence all a's will be all a's, but with the t at add again this sequence could not possibly correspond to a bacterium which is actually which is actively growing and dividing hence the fitness associated with all these nodes will be equal to 0. Which is what this graph is trying to show on the on the slide that we see here that most sequences do not correspond to actively growing bacteria. In the example that we have been talking about this figure is from the perspective of proteins. You only have a little island you only have a few islands here and there in the sequence space where there is actual which corresponds to actual living bacteria in the case that we have been talking about or actual functional proteins from the perspective of this review paper.

So, that is the first property of fitness landscapes that you want to talk about which is majority of sequences have 0 fitness. The second property is that the fitness landscape if we were to zoom in on this on these patches where there are non 0 values of fitness, the local structure of these fitness values could be very different. You could have a structure such as this which is lot of different peaks and each one of these peaks represents what we discussed in the last words end of last class a local optimal peak. This peak here represents the global peak in the sense that the sequence corresponding to this particular peak has a fitness which is higher than any of the sequences on the on the sequence space that we see here.

But every single peal that we see here corresponds to a local peak. What is going to happen in this cases that if I were to start a population somewhere over here, then that population is eventually going to depending on which is the first mutation that arises in the medium that I am doing the experiment in depending on that first particular mutation the population is going to move in a particular direction, and get trapped in one of the local optima. Once it is trapped in a local optimum in order to escape this local optimum it has to undergo a decrease and fitness which is not going to be permitted by selection. Hence this is an example of what is called a very rugged landscape a lot of local optima being present in the in the fitness landscape.

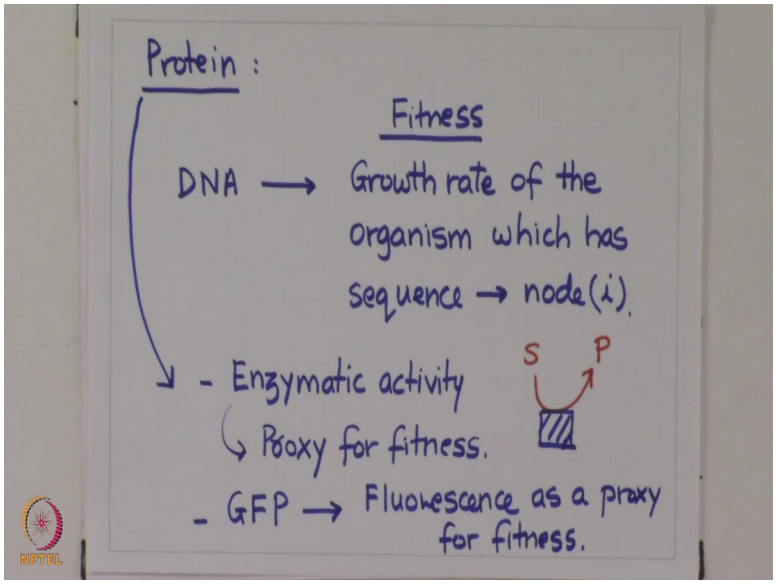
As compared to this you might also have regions in the sequence space where fitness is a very smooth function of the sequence space associated. The consequences there are big implications when it comes to consequences of differences between these 2 local structures of fitness landscapes. Here depending on where I start what is my initial genotype I am going to get trapped in a local optima and stay there, and which local optima I get trapped in is a very strong function of what is my initial genotype that I am starting with. Whereas, on this case no matter where I start on this sequence space if I were to start from here I am going to move up this hell and eventually reach the top. Similarly, if I want if I want to start from here I would I would still just have acquired mutations and move higher in the fitness scale and eventually reach the same top.

So, irrespective of what is your starting position in this landscape you eventually reach the same genotype which corresponds to the maximum fitness in this local structure. So, the second property of fitness landscaped is that fitness landscapes could be very rugged in nature rugged landscape corresponds to the figure on the left or it could be very smooth landscape which is the figure on the right. And this will as we have seen will dictate how populations move and where do the eventually end up as there moving with time. And the third property associated which is highlighted in this review paper is the effect of stochasticity. Now when you have something that we have already touched on in this part of the figure, but they have highlight it again that now imagine this to be the starting point of your of your simulation. So, you are doing of your experiment. So, imagine you are doing an experiment where all in all cells at time t equal to 0 belong to this sequence are genotypically identical.

So, this is your starting genotype now because mutations are random you could have first mutation take place in this direction or you could have the first mutation take place in this direction. If the first mutation was in this then you eventually reach this peak why have consequent additional acquisition of beneficial mutations leading up to this peak and once you reach here you stay in this peak because this is now a local optima all it is neighbors have a fitness which is lower than the fitness associated with this peak, and hence selection would permit movement of population out of this peak and hence you get stuck in a local optima. Whereas, of the first mutation was in this peak you keep moving up the fitness scale and eventually end up on the local fitness.

So, even from the same starting point you could get trapped in the local optima or be able to reach the global optima depending on precisely what happened in that first mutational event in which direction did you take the first step when you are an evolving population. So, again this is a very nice paper which reviews protein fitness landscapes, but I think the results are easily generalizable to fitness landscapes in general all right. So, let us let us move back to the slides and talk about some of the results that we have been talking about in the context of protein landscapes. So, what happens how these DNA results that we have talked about? So, far carry forward to protein landscapes.

(Refer Slide Time: 13:19)



First of all, question you should ask is that when we were talking of DNA, the fitness landscape corresponded to growth rate of the organism, we have been talking bacteria growth rate of the organism, which has sequence corresponding to any node I in the in the map that I have.

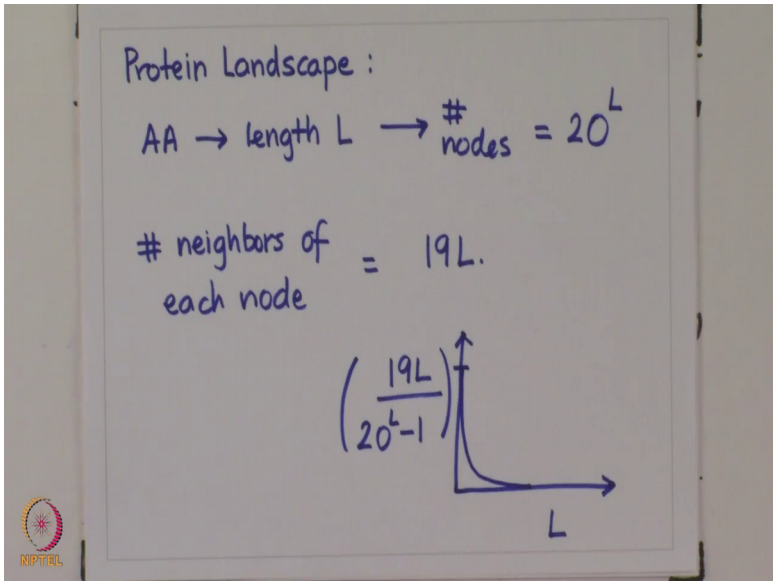
But what does clearly proteins own grow by themselves. So, what does the fitness? So, for DNA fitness was defined as the follows as follows, but what is the corresponding fitness when we are talking proteins. Which is what this what this review paper that we just saw also talks about. So, for proteins you could have this defined a couple of phase one could be that you could have the enzymatic activity. So, if you if you have an enzyme e and this catalyzes are reaction going from substrate to product a proxy for a

proxy for fitness is just the enzymatic a proxy for the protein fitness is just the enzyme enzymatic activity of a particular sequence of a protein.

Another way to characterize this would be there are these fluorescent proteins which flow. So, you have green fluorescent protein for instance. Now you could have fluorescence has a proxy for fitness. The question here is not so much, determining whether these whether these quantities actually represent fitness of a protein or not, but the goal is to understand the structure and the properties of these landscapes in general. So, defining them as proxy does not really mean that we are strictly saying that this is the fitness of a protein, but this helps us relationship, but this helps us relation understand the relationship that exists between sequence and function of a protein.

So, that sequence is; obviously, (Refer Time: 16:24) what is acid sequence and the function associated with protein as we just saw in the 2 examples that we wrote could be enzymatic activity or the fluorescence address associated with any florescent protein all right. So, just as an very quickly just as an we have been talking about the DNA sequence for a protein landscape.

(Refer Slide Time: 16:47)

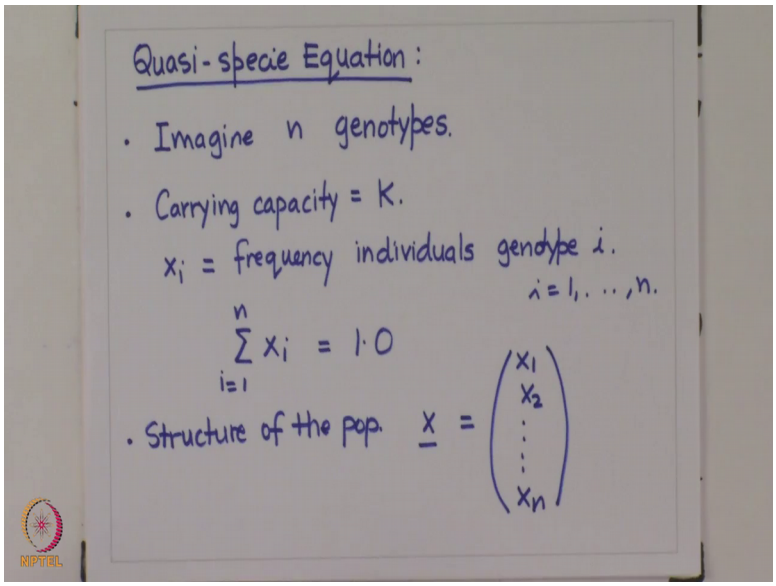


If we are talking of an amino acid sequence of length L the total number of nodes is equal to 20 to the power L. Number of neighbors of each node is equal to 19 L. And again if we plot L versus fraction of the total number of nodes that one particular node is connected to this decays even more rapidly as compared to what we saw in the DNA

results. And we will come back to that when we are talking of movements on populations and we are trying to define mathematical frameworks as to how movements how population will move on these hyper dimensional spaces.

For now, let us start with what is called quasi specie equation.

(Refer Slide Time: 18:08)



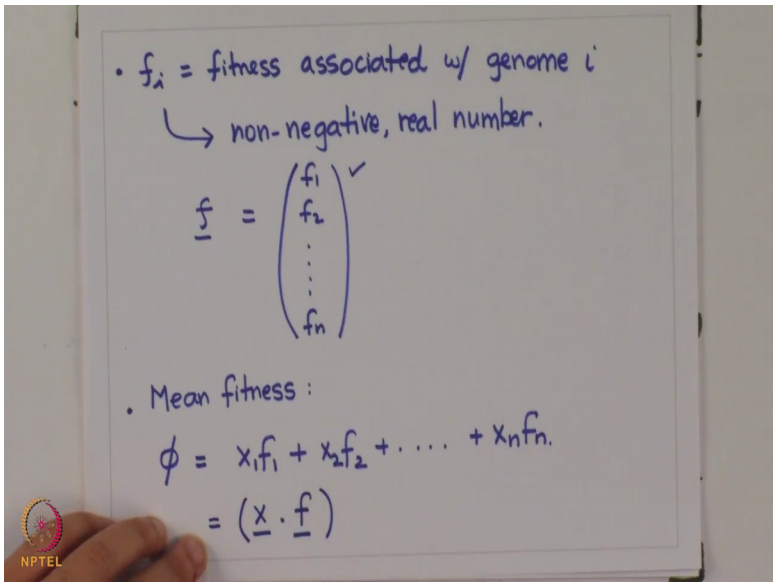
And we will define the framework for this imagine that there are n genotypes. So, you have a specie again will think in terms of bacteria, you have a bacterial species and you have n distinct genotypes which exist this experiment is being done in an environment where the carrying capacity is held constant and is equal to K. And x I represent the frequency associated with individuals belonging to genotype i right and i equals from 1 to n. We have done this before they should immediately tell you that this automatically implies that sigma x i, i going I goes from one to n is just equal to 1. Because x i is just the fraction of individuals which belong to a particular genotype i.

So, at any given point in time the structure of the population is given by a vector x, who is consequence are x 1 x 2 and so on and so forth going all the way up to x n. So, this is the vector which gives me the makeup of the population at any given point in time t, x 1 again corresponds to fraction of individuals which belong to genotype one and so on and so forth. And the sum of elements of this vector of course, has to be equal to 1. So, structure of population is going to be defined by this spectral x we have already done this before when we were talking in terms of mutations, but what we are going to introduce

now is another change that in addition to mutations will also incorporate the effect of selection in this.

So, all and genotypes are going are growing at growth rates which are distinct from each other's and each has it is own growth rate associated with it is particular genotype.

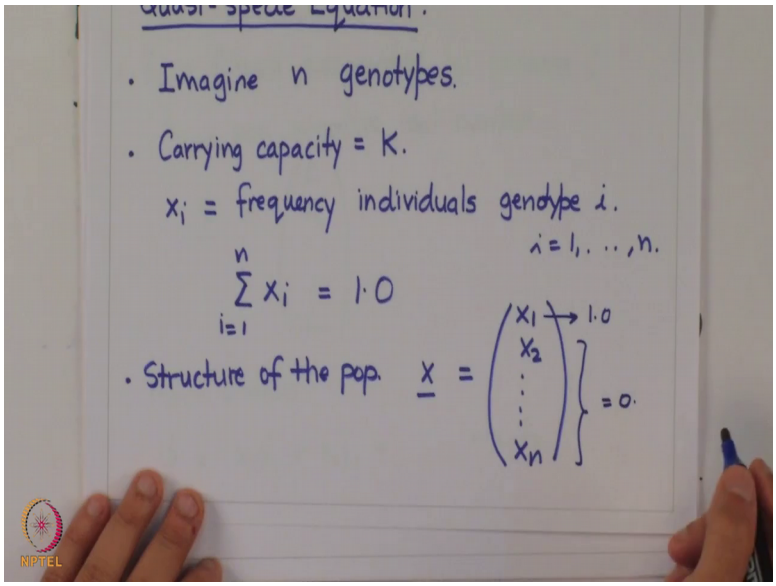
(Refer Slide Time: 20:49)



So, will define that as f i let f i be the fitness associated with genome i. Now f i because it is fitness growth rate this is a nonnegative real number. And we have going to define vector f which gives the fitness of all genotypes in the environment that are growing at any particular time. So, this is f 1 f 2 going all the way up to f n. So, we have structure of the population by given by a vector x, which defines the const fraction individuals belonging to each particular genotype and we have a vector f which defines the fitness corresponding to each one of those genotypes that exist in the environment.

So, first thing that we node is what is the mean fitness of the environment, if the structure of the population is defined by this vector x and the fitness corresponding to each genotype is defined by this vector f. Mean fitness which is what we have been calling five. So, far we will then be equal to x 1 times f 1 plus x 2 times f 2 going all the way up to x n times f n. So, it just the weighted mean of all the finesses and the weights being equal to frequency of that particular genotype that exists in the population. You can do a sort of a superficial check of this.

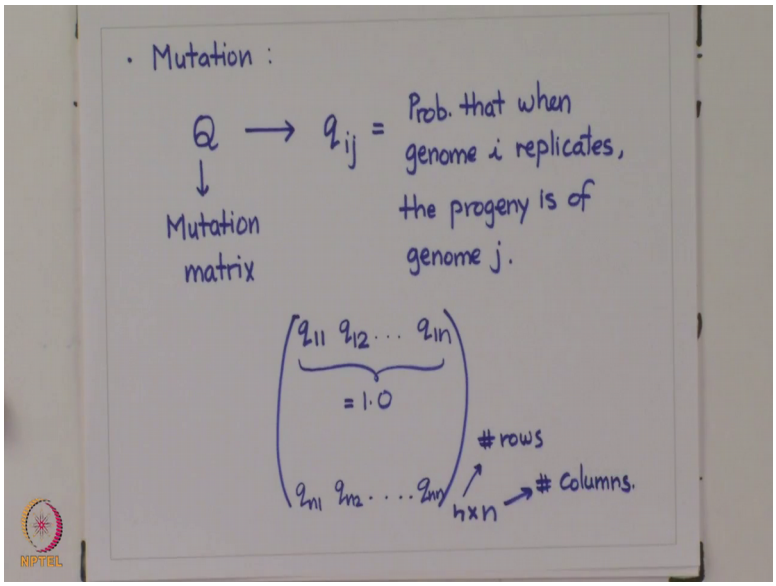
(Refer Slide Time: 22:46)



Imagine that the structure of the population is such that x 1 is equal to 1 and all the other exercise at equal to 0. If that is the case what you are saying is that every single individual in the population belongs to genotype 1; that means, every single individual in the population is at fitness f 1 and there are no individuals which belong to fitness f 2 to f n.

If every individual in the population is at fitness f 1 then; obviously, the main fitness should also be equal to f 1 and this formula gives us exactly that when we plug in x 1 equals to 1 and f 1 equal to f 1 all the other excise are equal to 0. So, these terms drop out and we will get 5 equal to f 1 that is it is a very crude check of the relationship, that we have talked we have been talking about this is the main fitness this can be represented as dot product of the 2 vectors x and f. So, this is fitness we have n distinct genotypes.

(Refer Slide Time: 23:53)



We are also going to incorporate mutations taking place in the environment. So, this is the L now we are going to sort of bring together for the first time all 3 tenets of evolution that we introduced in the first few lectures mutation selection and reproduction and bring them together in this framework.

So, we have mutation taking place and mutations rates are defined by the mutation matrix that we have been defining in the last few lectures. So, this is my matrix Q which is called the mutation matrix and it is elements Q i j is equal to probability that when genome i replicates the progeny is of genome check, that represents my mutation matrix and it is structure is going to be an n cross n matrix. The I will start the first (Refer Time: 25:16) Q 1 1 which is what is the probability that genotype one replicates and genotype one is that of the progeny Q 1 2 represents the genotype one replicates and the progenies of genotype 2 going all the way up to Q one n and similarly Q n 1 n replicates one is the progeny Q n 2 going all the way up to Q n n.

So, this is my mutation matrix and remember the property associated with that that the rho sum of this matrix is equal to 1. The sum of all these elements is equal to 1 because every time you have individual of genotype 1 replicating it has to come it to 1 of the n genotypes the progeny that is going to be produced has to be 1 of the n genotypes in the environment. And hence some of these probabilities has to be equal to 1. So, the rho

some of this of this matrix mutation matrix is equal to 1 for every each one of the n rows the n cross and here again represents the number of rows and number of columns.

So, we have sort of lay down the groundwork for incorporating all the 3 fundamental process is associated with an evolutionary process. We have a framework for growth rates associated with each of the genotypes that is reproduction. Then these are there is going to be selection acting because each of these genotypes are growing at a this at a growth rate which is distinct from each other. And finally, we have defined this mutation matrix which allows for one genotype to be converted to another genotype and hence we have incorporated sort of the all 3 processes that we have talk we have been talks saying are fundamental to an evolutionary.

So, in the next class we will we will continue development of this framework further and derive some dynamical equations which represent how do these frequencies associated with each particular genome change with time, and how can we compute the steady states associated between these n genotypes when there is reproduction selection and mutation all 3 taking place. So, we will continue that in the next class.

Thank you.