An Introduction to Evolutionary Biology

Prof. Sutirth Dey

Biology Department, Population Biology Lab

Indian Institute of Science Education and Research (IISER) Pune

Week 4 Lecture 20

Dynamics of Selection 1: Numerical Simulations

So, in our last discussion, we defined selection. We looked at what the necessary conditions for selection to operate are, and then we requested that you go and look at two videos. I hope you have done that. And then we said that in the next discussion, which is this one, we would look at the dynamics of selection. We are going to look at how selection leads to changes in alleles, allele frequencies, and genotypic frequencies. Now, before we go there, we need to define one term which I have not really used until now, but going forward, we are going to require that term, and that is a term I am sure all of you have heard, which is fitness. Now, what exactly is fitness in an evolutionary sense? The problem with fitness is that there are many, many ways of Looking at fitness and, just like the concept of species, you know different people have defined it differently. They have used it differently, and philosophers have written books and papers on it. So, you know it is a bit of a mess.

Basically, different people have different views about what it is or what it should be. So, those who are interested in the philosophy of, you know, what fitness is. And what are the various definitions and how do they relate to each other, etcetera? Here is a very nice resource for looking at it. Of course, this is not a part of this course; it is only for those people who are interested in the philosophy of the subject. So, in the context of this course, we are going to define fitness as a measure of an individual's success in surviving or reproducing.

That is about it. So, if, for example, some individuals have 80% survivorship, And others have said 40% survivorship; then the ones who have 80% survivorship will be said to have greater fitness. Similarly, if some individuals produce 20 babies and some other individuals produce 40 babies, Then the one who is producing 40 babies will be said to have greater fitness. Now, this particular definition that I am talking about actually aligns with most textbooks. For example, Futuyma and Kirkpatrick, which we are, you know, following to some extent in this course.

However, I must caution you that we, I mean my colleagues and I, have actually argued. That there are some issues with the above definition, and we have suggested a very different definition. To be a better definition of fitness in the context of evolutionary biology. If you are interested in looking at our arguments, this is the specific reference. It is available on Evo Archive, so it is free for you to view.

Now you might well ask me here that if you are not very sure about this definition, then why are you giving it to us? The reason I am giving it to you in the context of this course is that you know in a course. We have to look at a lot of standard material, a lot of, you know, notions, and a lot of concepts to which the field generally adheres. And it will be easier for us to know about those concepts, etc., if we adhere to this particular definition of fitness. However, if you go deeper into the subject, for example, if you end up taking a more advanced course in evolution, Then you will figure out what the issues are with this particular definition, but as I said, that is an advanced thing.

So, for this course, this is what we are going to think about in the context of fitness. as a measure of an individual's success in surviving and reproducing. Okay, so with that definition out of our way, remember this figure, right? This is how we derive the Hardy-Weinberg principle, and I explicitly said that the good thing about thinking about, you know, Hardy-Weinberg in this context is that you explicitly figure out which stages the different assumptions are coming in. And therefore, when you study evolution, you study it in the framework of the violation of those assumptions, right? So, if you remember, in this entire framework, selection comes basically in three places. The first place is right over here at the top, where we assume that all individuals have equal fertility.

Now, if that condition is violated, you have a kind of selection known as fertility selection. But that is not what we are dealing with over here. We are dealing with this scenario here, where you are going from genotypic frequencies among the zygotes. that have been produced to the genotypic frequencies among the progenies who are going to produce the next generation. So, in this context, remember we said that you know as long as the viability of all the genotypes is equal, In other words, there is no selection in terms of viability at this step as well as at this step, then.

We are going to get what is known as "you are going to get no selection" if the viabilities are equal. So, obviously, what we are going to see is what happens if the viability or the survivorship of the various genotypes is not the same. So, how will we do that? In this particular discussion, we are going to do this using numerical simulations in Excel. And we are going to do a one-locus, two-allele model, which is the simplest kind of model in this context. In doing so, we are going to derive some interesting insights.

So, what I will do is show you step by step how to make this Excel file for yourself. So that you know, there is no magic anywhere, nor is there anything being shoved under the carpet. It will take a little bit of time, but that is okay. If you follow me through the entire thing and actually make the Excel file as you see me making it over here, You know, stopping and making it, and stopping and making it, and so on, then by the time we are done with this discussion, You will have the Excel file with you, using which you can verify every single thing that I am going to show you and a lot more, right? And once you do that, we are going to, you know, You are going to be able to explore this entire thing on your own, which, trust me, is a great thing. So, you know, it is going to be a little boring if you just see me doing Excel.

So, I strongly urge you to also do the Excel thing on your own at the same time. If you do not have access to Excel, OpenOffice, LibreOffice, or any other office suite, they will have some kind of spreadsheet program. I am using very simple functions, so all those functions will be available in those spreadsheet programs. So, you can use whatever you have. Let us go to Excel.

This is just a simple Excel file. So, what we will do is we will do it in a slightly complicated way. But the reason we will do it in that complicated way is that then you know looking at the simulations becomes very easy. So, the first sheet we are going to call ParamGraphs. It does not matter; you can call it anything.

And for the second sheet, I am going to call it simulations. So, now let us assume that we are doing a 1-locus, 2-allele model. Which means that let us assume that our allele A1; let me increase this a bit. So, let us assume that our allele A1 has a frequency of p. So, this is allele frequency, and therefore, allele A2 will have a frequency of (1 - p).

So, for this one, we put allele A1 as p. Let us go over here. So, for the frequency of A1, let me increase this size a bit over here and set it equal to what we take from here. So, let us assume—sorry, let us assume that we will start with p equal to 0.1, let us say. And we will mark this in yellow, so that we know that this is where we have to make our changes. So, here the frequency of A1 is equal to this value over here. Therefore, the frequency of A2 is equal to 1 - p. So, this is 1 - p. Great. So, now remember that we are saying that this is the frequency in the parent regeneration. So, now these parents are going to produce the offspring, and these offspring's genotypes are going to be in the frequencies p^2, 2pq, q^2. So, we will call P = p^2, Q = 2pq, and R = q^2, right? So, let us make this P = p*p, which is p^2, Q = 2*p*q. And let us say this Q = q^2, equal to q*q. So, until this point, the Hardy-Weinberg frequencies have been obeyed because we are going to have variability selection after this.

So, let us go back to this Param_Graphs sheet, and now let us have our genotypic frequencies. So, what are our genotypes? Our genotypes are A1A1, A1A2, and A2A2, right? So, these are our genotypes. Now, let us assume that all these genotypes have some kind of survivorship. So, let us assume this is what we are going to call fitness. The fitnesses are given as w11. So, the fitness of A1A1 is w11, the fitness of A1A2 is w12, and the fitness of A2A2 is w22. Let us give them some values. So, let us say, just to begin with, that this value is 0.9, this value is 0.7, and this value is 0.3. So, those are some arbitrary values. Again, what we will do is mark this as yellow. Meaning that these are the places where we can make changes. Okay, so now we are back to our simulation thing. So, these are the frequencies before selection: frequencies before selection, these three.

Now, the genotypes have been formed, selection has operated, and therefore they have survived differentially. So, in other words, now that the selection has happened, what will their frequencies be after selection? The frequencies after selection are p^2*w11 and 2pq*w12. Because that is the fraction that is surviving to the next generation; sorry, surviving to the next stage. So, this is going to be p^2*w11; frequencies after selection for A1, I will just call them P, Q, R. So, this is equal to p^2, which means this times this w11 over here, right? Now, here we have to do one extra thing.

What is it? So, here when we have this, you know, cell E6 selected, we need to put a $ sign. So, two $ signs, one before E and one before 6. So, E6. Why do we do this? So, when you do this, it means that when you are going to drag these, you are going to use this relationship for other cells. From this place, it will only pick this value, and it will not pick any other values in the relationship.

In other words, basically, we are using it as a constant; that is why. So, this is p^2w11, Q = 2pq*w12. What is my w12? My W12 is this, and again I need to put a $ here, and I need to put a $ here. And this is what? This Q is = q^2 * w22, which is this value. And I need to put a $ here and a dollar over there again.

Alright. So, these are the frequencies before selection and these are the frequencies after selection. Awesome. But now I have a problem. What is my problem? If I take the frequencies before selection, If I sum them, what am I going to get? I am going to get 1.

Great. That is how it should be. But now that these numbers, p^2, $2pq$, and q^2, have been multiplied by different fractions that do not sum to 1 or anything. Therefore, will this sum to 1? Let us quickly check. No, it will not sum to 1. It is not even expected to sum to 1, right? Now, obviously, if you are trying to figure out the actual proportions, then you need to make this sum equal to 1, right? You need to scale it in such a way that all these frequencies sum to 1. How will you do this? The simplest way to do this is to divide every single term by the sum of these.

So, we will look at the sum of this, and we will call it $\bar{w}$. Why we are calling it $\bar{w}$, you know, is something that we will postpone to the next discussion. But as of this moment, let us call this $\bar{w}$, and this $\bar{w}$ is simply the sum of these three things: these three frequencies. Great. So, now if our understanding is correct, then if we are going to divide these, Each one of these by this sum should sum to 1.

So, let us do that: =P/$\bar{w}$, great; =Q/$\bar{w}$, great; =R/$\bar{w}$, this. And if what we are thinking is correct, then this would sum to 1, and it does, which is not surprising. So, what have we done over here? We have simply scaled them by the, you know, average. By the sum of the frequencies, scaled frequencies, and these scaled frequencies are now summing to 1, great. So, these are the actual genotypic frequencies post-selection in the offspring generation.

Now, from this, we need to go to the allele frequency in the offspring generation, right? So, remember this is the frequency of A1 and A2 in the parental generation. Now we have to go into the offspring generation, and for that, what will we do? We will simply do P+Q/2. So, = (P+0.5*Q), right and this should be equal to actually, you know, you can just drag this, right. So, what we have done is start from the parental generation allele frequency.

And we have gone all the way to the next generation's allele frequency, right? And the only thing that is missing here is generation. So, let us assume that this is my generation 0, the parental generation. This is generation 1, and let us say, okay, we will talk about that in a moment. So, now all we have to do is drag this for the next generation, and we have the next set of values, right? So, this is the same thing for the next set, and now all we have to do is take this thing, select this entire stuff, and just drag this.

Let us see how long we want to go. We have gone up to 160 generations. Let us say we would like to go another 40. So, we make it 200 generations. There is nothing magical about 200; it is just a round number; that is all. Okay, that is it; our simulation is done, right? But these are just the numbers; you know it is good to get some graphs out of these things.

So, those graphs we are now going to draw at this end will show how p changes over time, the allele frequency of allele A1. So, what we will do is go over here, insert, and go over here, you know. We are going to press this; this and this is the kind of graph, you know, a dot with lines is what we are going to select. So, it gives you a blank chart, and what we will do is go to, we are already on chart design. So, we will go to select data, and now it allows us to select the chart data range.

So, we will go to the simulations page, the simulations worksheet, and we will select the entire column called frequency of A1. And that is it. We have our graph. So, this is what it looks like, and we will put the legends. Although there is only one legend, we do not actually need it.

There is just one point, you know, one variable over here. So, we need the axis titles, right? So, what is this? On the y-axis, we have the frequency of the A1 allele. So, we will select this and increase the font size a bit. Similarly, on the x-axis, we have generations. So, we will select this entire thing, and we will go to Home and We will increase the font size.

Let's see, this is good enough, or maybe a little less. Yeah, this is good enough. Okay. And yeah, this frequency on this side should only go from 0 to 1.

So we will cut this at 1.0. So that is done, and yeah, we can, you know, maybe make it look slightly better. But overall, what we want to do that is coming out of this graph is clear. We can see, you know, we can see the patterns. So that is our first graph, and for the second graph that we would like to draw, where do we draw this? Maybe next to it.

The second graph will show the frequency of the genotypes. So we do the same thing again. What we do is go to insert, and we go over here again. We select the one that says "Insert Line or Area Chart," and then we select this particular one again. It comes out here, and what we do is go here and say, "Select data." And in the data thing, we go back to our simulations page, and we are interested in knowing the scaled frequencies.

So we just select the entire thing for 200 generations; we select the entire stuff. P, Q, R is fine. The horizontal category is 1, 2, 3, 4 (Generations), so this is fine. There you have the graph. Okay, so this is, you know, the genotypic frequencies over time, and yeah, we need the legends over here.

So we go to this plus sign, and we need the axis titles, and here we definitely need the legends. Okay. Again, the y-axis cannot go beyond 1, so we'll make it 1.0. What else can we do? Yeah, the axis titles need to be given. So, this is generations, and this is genotypic frequency. Genotypic frequency again, we press Ctrl+A, go to home, and slightly increase the size to 14. Okay, so what we do now is we slightly reduce the size of the graph. Okay, so I think this is good enough for us to visualize or maybe what we can do is bring one of them here. And we can bring the other one over here so that we can visualize both of them in the same view. Okay. Excellent. So now what will happen is that we can play with the initial allele frequencies. Or we can play with the genotypic fitnesses right from here without even looking at what is happening to the numbers over here.

I mean, of course, we can go back and look if we want to, but we do not really need to do that. So these two graphs are going to tell us a lot. Let us start playing with them. Now, even before we start doing so, note that in this particular case, A1A1 had the highest fitness, and A1A2 had the second highest fitness. A2A2 had the lowest fitness, and therefore, even though A1A1 started with a low value of 0.1, Very, very fast, it actually climbs to a value of about, you know, 1, which basically means the other one goes close to 0. Now remember in the previous context when we were looking at mutation going from 0.001 to 0.1, it took, you know, Depending on what kind of mutation rates we were using, it took hundreds to thousands of generations.

So let us see what happens if we want to go from 0.001 to 0.1. So here we have to look at the numbers. So we have started with 0.001, and from 0.001 to 0.1, if you can see, it takes just 6 generations. Now, of course, the precise number of generations that it will take will depend on the values of those things, but You can see that for some reasonably realistic values, it is actually very, very fast. And this is what I mean by saying that selection is a strong force; it causes changes very fast. So now you can see that there is a lot of difference between the three genotypic fitnesses and the survivorship values.

What if these are close to each other? So let's say 0.9, 0.85, and let's say 0.8. So now, although the order has been maintained, but now I have brought them closer to each other. Look what happens. The allele A1 still reaches fixation; it still reaches a value of 1. But it is just that it takes a long time to go there; it goes there nonetheless. Now, of course, this is happening because we are starting with a very, very small value of 0.001. If we start with, let us say, a slightly higher value, say 0.1, again the overall relationship remains the same. But the whole thing happens much, much faster. So this tells you that it is not merely the values that matter. Values of the fitnesses for the various genotypes also matter, as does how separated they are. If all the values are close to each other, then it will take longer to reach, you know, a fixation or a high value.

If they are separated from each other like it was in the first case 0.9, 0.7, 0.3, then things will happen much faster. So basically, it is the relative values related to each other of the fitness that matter. So this was a situation where we saw that if you changed the initial allelic frequency The outcome is not really changing; it is just becoming faster or slower.

So let us go back to our initial case: 0.9, 0.7, 0.3, sorry, and ask ourselves what happens if we keep changing this value of p. Am I going to get a qualitatively different thing? Is there any point at which, instead of allele A1, it is allele A2? which gets you know selected, which allele A2 which gets fixed. Will that happen? And it turns out that you can satisfy yourself that whatever values you put in these conditions, that is not going to happen. So now let us slightly change the things we are going to play with.

Now let us say, let us make this w11 as 0.7 and let us make this w12 as 0.9. Let us interchange these two values and see what happens. The moment we do that, we find that now allele A1 is not really going to fixation; neither is allele A2 going to fixation. Instead, they are being maintained somewhere in the middle. So, if you look at the actual values, it will become clear. You can see they are going to some kind of equilibrium where both alleles are being maintained in the population, right? Both of them have frequencies of 0.75 and 0.25. I mean, sorry, one has a frequency of 0.75; the other has a frequency of 0.25. So will this change if I change my initial allele frequency from 0.3? Let us make it 0.5; it does not change; let us make it 0.2 again; it does not change, and so on. So you can satisfy yourself by playing with it that whatever values you put in this situation, this is what is known as polymorphism; both alleles being present simultaneously is always going to be maintained.

Starting population frequencies will not matter. Now, let us interchange these two. So let us make this w12 0.3 and let us make this w22 0.9. Now what happens? Now you see that, again, one of them, R, which basically means A2A2, goes to a frequency of 1. And the other one, both genotypes go to a frequency of 0. And if you look at the simulations, you can straight away find that that is what is happening. Very soon, within 12 to 13 generations, this has a frequency of 1. Allele A2 has a frequency of 1, while allele A1 has a frequency very close to 0. As you can see, these are extremely small numbers; they are kind of going towards 0. For all practical purposes, they are 0. So what is it telling you? It is telling you that when you have a situation in which the heterozygous is less fit than the two homozygous conditions, Then again, you have a situation where one allele goes to fixation while the other alleles do not. Will the starting frequency matter in this case?

Let's see. So this is 0.2; let us make the starting frequency of p 0.8, and suddenly you have a completely different situation. Now it is P that has gone to fixation, which basically means that It is allele A1 that has gone to fixation instead of allele A2. Think about what has happened. All we have done is gone; we have kept the genotypic frequencies the same. In one case, we started with an allele frequency of A1 equal to 0.8; allele A1 gets fixed, and if you start with 0.2, allele A2 gets fixed. In other words, for the same set of fitness values, which genotype gets fixed? Which allele gets fixed depends on where you are starting your allele frequencies. So there are many, many other things that one can get out of this, but, let us get back to our PowerPoint and see what the major observations are that we just made. So our first observation was that when the surviving sheep are close to each other, selection is slow in changing the frequencies. When they are very different from each other, the change in frequencies is fast. Due to selection, sometimes one allele gets fixed, sometimes the other allele gets fixed, but there are situations.

When both alleles are present in the population, this typically happens when the heterozygote is the fittest. And in terms of the long-term outcomes, sometimes the starting allele frequencies matter, and sometimes they do not. Now, of course, you can keep on playing with all kinds of combinations of the genotypic frequencies. The starting allele frequencies can keep on getting these kinds of observations. But why exactly are you getting these observations? What is happening underneath? That is something that is not very clear from these simulations.

And therefore, the only way to gain an understanding, a deeper understanding of what exactly is happening, Under what conditions do I expect starting values to matter? Under what conditions do the starting allele frequencies not really matter, etc.? One has to get into the mathematics of this. One has to do an analytical treatment of what we just saw. and that analytical treatment is what we are going to do in the next discussion. See you.