Evolutionary Dynamics

Supreet Saini

Chemical Engineering

Indian Institute of Technology Bombay

Week 9

Lecture 45

Thank you. Hi, welcome everyone. So this is the last discussion before we delve into the final part of the course, which is how we do experimental evaluation in the lab and what sort of insights that body of work has given us in terms of understanding, in terms of enhancing our understanding of evolutionary processes. So in this last bit, we are going to differentiate between two different aspects of evolution.

One is that if we have a starting population and let us say this is their genome, let us make them straight lines. They're circular and microbial organisms for the most part, but that's okay. Then what's going to happen? Let's imagine that this population gets split into two different groups.

Maybe they were separated by a mountain range or one of the populations got stuck on an island, and so on and so forth. So these two populations are separated by some individuals on this side of the mountain, some individuals on that side of the mountain. And now these will evolve in their respective environments, and they have no contact with each other. And DNA changes will take place in their respective environments.

Let us say Mutations will keep on happening because mutations happen at a fixed rate, errors keep on happening and so on and so forth. Let us say that the errors that are happening are purely neutral in nature. So you have an error that took place here and this is a completely neutral mutation that happened. We saw a few videos ago that neutral mutations occur and they have a finite probability of reaching fixation.

In fact, if the population size was N, then this neutral mutant has a probability 1 upon N to reach fixation. So, eventually some chance event will happen and these neutral mutations will reach fixation such that going forward in time and it may take a long time, the population structure might look like this. That is because this neutral mutation that

took place actually reached fixation. On the other hand, in the other population, a similar process might be going on and you could have other neutral mutations reaching fixation. It's practically impossible that you will have the same mutation take place, but let's say this is the neutral, again a neutral mutation that took place.

And we know again we know that neutral mutations fixation probability is 1 by N . And sooner or later some of these neutral mutations will reach fixation. Eventually the population structure would look something like this. Remember this entire picture that we have painted, selection has had no role to play in driving this change. This has been purely drift, which has driven evolution of these two species, leading to changes in their genome. But now, if I were to compare these two sequences, when I do that comparison, I will note the following.

This is sequence of, let's call this group A. and let us call this group B. So, for the A group, when I compare the sequences of the two groups, let us say I am a biogeographer, and I go to one island and collect specimens of this species, I go to the next island and get this species, get their DNA sequenced, and when I compare the two sequences, I will note the following. I will see that species A has a mutation here and I will see that species B has a change here. What I will note is that at all other places, the DNA of these two species is identical, except for these two locations on the genome where there is something here and something else here. And except at this position where, again, the nucleotides that are present in the two species are distinct from each other.

Of course, I do not know whether these changes were driven by selection or drift. Alternatively, these changes could have been driven by non-neutral mutations. They could have been driven by selection where beneficial mutations happened and then they were driven to fixation. So, this green mutation in this alternate prism could have been a beneficial mutation whose fixation probability was not 1 by n, but was in fact S naught, where S naught is the selection coefficient associated with the green mutation.

And as a result, and of course, not all beneficial mutations will reach fixation, but some beneficial mutations will reach fixation. As a result of that, what I see today on the island is is that every individual in the population is carrying this mutation. And same deal with the other group, that this mutation could have been non-neutral with a fixation probability of S0 dash, and then that beneficial mutation going to fixation. By non-neutral, we are essentially saying beneficial.

So, the question is, if I look at the sequence of these two species, I will see that they are different from each other at these two respective positions. If that is the case, what I want to find out is whether these changes... were driven by, as we discussed in the first case, drift only or were they driven by selection. That is the question we have, and at our disposal, all we have is the sequence of these two species. And this is the question we want to answer through an analysis of these two species.

So, I hope that question makes sense. As you will see, this is, of course, a picture of the genetic code. And as you will see, some mutations in the genetic code lead to a synonymous change, and some lead to a non-synonymous change. So, a synonymous change is a mutation in DNA that leads to the same amino acid. So, imagine a DNA sequence where you have, let's say, anything—you have CCC.

This CCC encodes for proline. Should a mutation happen such that, should a mutation occur at the third position, and you get CCG. Now, the amino acid corresponding to CCG is also proline. So, while a change has happened at the DNA level, there is no change at the amino acid level that will be incorporated when the mRNA associated with the gene is translated.

So, this mutation is referred to as a synonymous mutation. Synonymous because, as far as amino acid identity is concerned—which is going to get incorporated into the growing peptide chain—whether you have CCC or CCG, it means the same thing. And hence, it means the same thing, which is proline, and hence the word synonymous. So, as you can see, there are lots of synonymous mutations possible in the genetic code. You could have GGC changing to GGA, but you would still be encoding the same amino acid.

I mean, they are everywhere, as you will see. On the other hand, you have something called non-synonymous changes. These are mutations as a result of which a different amino acid is incorporated. To see an example of a non-synonymous mutation, let us say you have TTG—that means you are here, and the amino acid you are talking about is leucine. But you have a mutation taking place at the second position, and you see TCG; then you basically jump from here to here, and now the amino acid that is being incorporated—

so it was leucine here, but now it is serine. This is an example of a non-synonymous mutation where the identity of the amino acid that is going to be incorporated has changed. So, as you will see, the way you read this genetic code—we discussed this in an earlier video—is that a mutation in the second position makes you change columns. A

mutation in the first letter makes you change rows, and a mutation in the third letter of a codon just switches you within each of these boxes. So, as you can probably imagine, most of the synonymous mutations are at the third position because, for instance, here—all four—it doesn't matter; the third nucleotide here is actually irrelevant because all four, when it starts with T and C, encode for the amino acid serine.

And as you will see, same deal here. If the codon starts with GC, all four possibilities code for alanine. Same deal here, same here, same here. And you will see this spread around the genetic code differently. You should realize in case you are wondering why this is a U because mRNA instead of a thymine has a uracil.

So we don't have ATGC, but we have AUGC in mRNA. This particular picture is written from the context of mRNAs and not DNA. So third positions have a lot of synonymous mutations. Sometimes first position also has a synonymous mutation. For instance, if you have CTG on the DNA and it becomes TTG, the identity of the amino acid does not change.

It remains from leucine to leucine. So, in some cases, first mutation at the first position can also lead to synonymous mutation. Mutation at the second position can never lead to the same amino acid because the way the genetic code is structured, if you change the column, amino acid identity always changes. The only sort of an exceptional exception to that is if you have TAA, which is a stop codon and it becomes a TGA, it is also a stop codon. But that's not amino acid, that's just stop codons.

So, at different positions of the codon, there is a different propensity to acquire a mutation and encode for either a synonymous or a non-synonymous change. A second mutation in the second position of a codon, 100% of the time, results in a change in amino acid identity. A mutation in the third position of a codon very often leads to only a synonymous change. And sometimes, a mutation in the first position can lead to a synonymous change. Primarily, it leads to a non-synonymous change.

What does this have to do with the problem we discussed of identifying whether a mutation is driven by drift or selection in the genomes of two groups of these populations? So, this is sort of a first approximation to study this idea. But what we are going to say is that synonymous mutations, by and large, are neutral in nature. So, synonymous mutations are neutral in nature. OK.

So. And the idea is that a synonymous mutation just encodes for a different codon but the same amino acid. And hence, the amino acid chain is not going to be altered in the sequence of the protein. As a result of that, the protein identity doesn't change. And ultimately, it's the protein, which is the workhorse of the cell, that is going to do all the work.

And as a result, if that doesn't change, then this mutation cannot have any effect on the fitness and hence synonymous mutations must be neutral. On the other hand, in non-synonymous mutations, we are changing the identity of the amino acid and hence the protein is going to change and its performance is going to change. And if the performance of the protein changes, that means fitness of the individual carrying that variant protein will also be different from the ancestor. And hence, this will be a non-neutral mutation. Now, it could be beneficial.

It could be deleterious. That is something that we do not know. We cannot tell. But the assumption is that non-synonymous mutation is non-neutral and synonymous mutation is neutral mutation. and non-synonymous mutations are non-neutral.

So, if the synonymous mutations are neutral that means their fate is driven by drift only because by definition if they are neutral then selection has nothing to do with synonymous mutations. So, this is drift only. And if non-neutral, if non-synonymous mutations are non-neutral, then their fate is obviously being determined by selection. And hence, there is a qualitative difference in how these two kinds of mutations are going to be treated when looked at through the prism of selection and drift. I mean, obviously here drift will matter, every mutation is susceptible to drift.

But the main idea is that selection has nothing to do with synonymous mutation. So now when I go back to groups A and B, when I go back to groups A and B, and they may carry any number of mutations between them. We just showed two, but maybe there are more. All these mutations that occur, what we are going to do is, is what if we measure that between the changes that these two groups are carrying, how many are synonymous changes and how many are non-synonymous changes.

And the idea that I'm going to pose here through this analysis, this was work that was first proposed in 1960s. The basic idea behind this concept is that suppose these four changes are all synonymous. Mutations have taken place at the DNA level. But at the amino acid chain level, no mutation has, no change has taken place as far as the identity

of the amino acid that is going to get incorporated is concerned. So, all four are synonymous mutations, which means all four are neutral mutations.

which means these two population groups' fate is decided by exclusively drift because there is no non-synonymous mutation that has taken place. So, there is nothing that is being selected for. It is just random fluctuations and drift which is deciding the fate of these populations. And so on and so forth. So the broad idea is by making a relative comparison between the frequency of these two mutations, I can tell that whether the evolution of these two groups is driven by selection plus drift or just drift.

One more thing before we move forward is that different amino acids have different propensities to undergo synonymous or non-synonymous changes. For example, let's take a look at methionine. This is also the start codon. Remember, translation starts with ATG primarily. So when I have ATG, it's either a signal for the start of translation or the incorporation of methionine if it occurs within a gene.

So in the DNA sequence, if I have the codon ATG, Now, it turns out that for methionine, this is the only possible codon. There is no other codon that encodes methionine, which means that for this particular codon, there is no synonymous mutation possible because, other than ATG, there is no codon that encodes methionine. This is different from, let's say, serine, where TCA and TCG both encode serine. So, an A-to-G mutation at the third position will give you a synonymous change.

Such a change is not possible for methionine. The only other amino acid that has this property is tryptophan, which is TGG. This is the only codon that encodes tryptophan, and hence tryptophan also doesn't have any possible synonymous mutations in its genome. Now, and so on and so forth. But some others, like serine, for instance—any mutation at the third position leads to a change in the DNA sequence but no change in the amino acid identity; it still remains serine.

And hence, different amino acids have different propensities to encode for a synonymous mutation or a non-synonymous mutation. Okay. So now, with these ideas, let's imagine that we have two sequences. We have a sequence from species A and a sequence from species B. Let's say I compare the sequences, and the idea is that I want to compare these two populations and understand whether their differences have been driven by drift or selection. And in order to do that, the first thing I do is count the number of differences that exist.

Count the number of differences between these two populations. So, there might be one difference here, another one here, and so on and so forth. So, let us say the number of differences is N naught. When we say number of differences, this refers to differences in nucleotides. So, there are four positions in this DNA sequence where the identity of the nucleotides differs between these two sequences.

Now, the next step is that among these N naught, I count how many lead to how many lead to synonymous and how many lead to non-synonymous changes. So now I am asking: since the identity of the codon here is different because there is a mutation, do these two codons encode for the same amino acid—in which case it will be a synonymous mutation—or do the codons present at these two positions in these two groups encode for different amino acids—in which case it will be a non-synonymous mutation? Since the total number of differences is N naught, maybe S of them are synonymous mutations and N naught minus S are non-synonymous mutations. So now I have a relative estimate of how many synonymous and non-synonymous changes have taken place in these two groups that I am comparing.

So now I define a quantity which is number of synonymous divided by number of non-synonymous changes. And that is simply equal to S divided by N0 minus S. But this does not tell us what we want to know. Because as we saw in this picture, suppose I have a DNA sequence which is simply ATG, ATG and nothing else. This sequence cannot have a synonymous mutation because the amino acid it encodes for Because the amino acid in it encodes for doesn't have another codon.

Hence, every mutation that happens in this sequence is going to be a non-synonymous mutation. Same with tryptophan. However, if I have a mutation, if I have a genome sequence, which is TCG, TCG repeated over and over again, then this mutation is only serines. And hence, as a result, many of the mutations that can happen in this sequence are going to lead to a synonymous change because serine has six codons allotted to it. So the propensity to acquire synonymous and non-synonymous changes varies from one DNA sequence to another.

And hence, we cannot directly compare the synonymous and non-synonymous mutations in absolute terms. We have to study the DNA sequences and understand the propensity of synonymous and non-synonymous mutations that the underlying DNA sequences had. To do that, we look at the DNA sequence. Let us say the number of codons is in this DNA sequence is equal to L. Note that for every codon there are 9 mutations possible.

Suppose I have a codon CCC, and I want to introduce one mutation. Then, I could introduce one mutation at position 1. I could change this to T, A, or G, or I could make that one mutation here, or I could make that one mutation here. So, there are nine ways in which I can introduce a single mutation into a codon. So, if this sequence—let us say I took A—if this sequence has L codons, then for each codon, there are nine mutations possible.

Out of these nine possible mutations, some are going to be synonymous, and some are going to be non-synonymous. For instance, CCC is proline. So, some are going to be synonymous, and some are going to be non-synonymous. So, we count that, and let us say this we call synonymous, this we call non-synonymous.

This is the number. And we do this for every codon. So, for codon 1, I introduce 9 mutations and count them. For codon 2, I introduce the 9 mutations and count them, and so on and so forth. So, for every one, I have how many are synonymous possible and how many non-synonymous are possible.

For the second codon, how many synonymous are possible and how many non-synonymous are possible? The idea of what I am doing through this exercise is getting the propensity of the sequence that I am dealing with to have synonymous and non-synonymous mutations. Because, as we have seen, there could be DNA sequences where introducing a synonymous change is just impossible. Hence, we simply cannot compare these two ratios. This has to be weighed in by what is the propensity of synonymous and non-synonymous mutations in the underlying sequences from which we have gotten this data.

So, we calculate the number of synonymous and non-synonymous for every one of these codons, and we add them up. So, we add all the synonymous, we add all the possible synonymous mutations in every codon, and let us call—so we add them all up, and this, let us call this K and And we add all the non-synonymous ones for all codons. And let us say we call this K1, let us call this K2. So, suppose this was a codon where 4 were synonymous, 5 were non-synonymous, and the second codon 3 were synonymous and 6 were non-synonymous.

The total has to add up to 9 for every codon. In that case, the synonymous ones here are 7, that is K1, and the non-synonymous ones is 11, that is K2. But we do this for all codons in the sequence that I am dealing with. This tells me that for the DNA sequence that I am dealing with, the relative propensity to have a synonymous to a non-

synonymous mutation is K1 is to K2. Whereas, in reality, what I have is S to N naught minus S. So, I will normalize this, and I will call this ratio as dN by dS.

$$\frac{dN}{dS} = \frac{(N_0 - S)/K_2}{(S)/K_1}$$

And this dN by dS is simply the non-synonymous changes that actually happened, N naught minus S, divided by the non-synonymous propensity of the sequence, which is K2. Divided by the synonymous changes that happened, which is S, and finally divided by the propensity of synonymous mutations in that sequence. This ratio gives us an estimate of what the relative strength of the two forces—selection and drift—is in driving the evolution of these two groups of populations. We will begin our next lecture with a brief discussion of dN by dS, what it tells us, what actual values of dN by dS indicate, and then we will start with a discussion of experimental evolution in the lab. Thank you.