An Introduction to Evolutionary Biology

Prof. Sutirth Dey

Biology Department, Population Biology Lab

Indian Institute of Science Education and Research (IISER) Pune

Week 10 Lecture 46

Evolutionary patterns in the genome

Hi, so in our last two discussions, we looked at the ways in which organisms generate variation. both in the long term in terms of creating new genes or in the short timescale in terms of phenotypic changes or cryptogenetic variation, and so on. And I told you that all these changes leave their mark on the genome of a species, and therefore, When you look at the genomes, they act like books that contain the tales of all the previous generations. So, what exactly are the stories that they tell us? That is what we are going to look at in today's discussion. What are the kinds of patterns that you can see in genomes that allow you to infer various things about how they have evolved? But before we get on with that, there is a certain concept that we need to clarify, and that is related to mutations. So, remember that when you have mutations that lead to the substitution of a base, If you know transition or transversion, then there are two kinds of mutations that are possible.

One type, known as non-synonymous mutations, occurs because of the mutation. The amino acid that is incorporated into the protein changes. So, these are denoted as N, and the others are the so-called synonymous substitutions. Where there is a mutation, you can see it in the gene, but it does not lead to a change in the protein structure.

Now, why do synonymous mutations occur? Because we have about 64 codons, 3 are stop codons, so about 61, you know, are amino acid coding codons. But we have only about 20 amino acids, which means that there is redundancy for many of the amino acids. which means that if you end up changing, you know, one of the base pairs, then it is entirely possible that you get to the same You know another version of the codon that codes for the same amino acid, which is why synonymous substitutions happen. Now, suppose you have two genes or the same gene from two different species. And now you are trying to figure out what the difference is between these two genes in terms of the sequences.

So, when you are trying to compare the sequence of a gene from two species, you end up calculating two quantities. You essentially need to see how many non-synonymous substitutions there are and how many synonymous substitutions there are. But the number of ways in which a non-synonymous substitution can happen is much greater than that. The number of ways in which a synonymous substitution can occur given an open reading frame. Therefore, what you need to do is scale the numbers.

The number of substitutions that you are seeing divided by the total number of ways in which those substitutions can happen. And if you do that, then you derive two quantities called dN and dS. So, when you take the number of non-synonymous mutations and divide it by the total number of sites in that gene Where a change would lead to a non-synonymous mutation, that ratio, that scale value, is what is called dN. And similarly, when you do the same thing for synonymous mutations, How many changes have actually happened in that gene, and how many changes would lead to synonymous mutations? So, that is the thing. So, that stuff is known as dS.

Now, we are going to look at the various patterns that one sees and the first. I am sorry, but prior to that, why are you calculating dN and dS? That is because the ratio of dN to dS actually gives you a very nice insight into the nature of the selection that is happening on a gene. So, suppose you have a situation where, between two species, you see your $dN/dS < 1$. I am sorry for a gene you see your $dN/dS < 1$. Now, what will that mean? That means that there is purifying selection or negative selection happening.

Why? Because $dN/dS < 1$, which means that non-synonymous mutations are happening much less frequently compared to synonymous mutations, Which means that anything that is leading to a change in the protein structure is typically not favored by selection. Typically, it implies that it is an important protein; it is a conserved protein, and that is why the selection on that is of the negative type. Any departure from the existing version is not easily tolerated. So, this is the kind of stuff that you see in, for example, housekeeping genes or ribosomal genes, right? I mean there are many housekeeping genes; ribosomal genes are one example of them. Now, suppose you see $dN/dS \approx 1$.

What does that mean? That means that synonymous and non-synonymous mutations are roughly occurring on the same date. What does that mean? That means that, as far as the gene sequence is concerned, it does not really matter whether you have Synonymous or non-synonymous mutations are essentially just happening randomly. And there is nothing that is altering the fixation rate of those mutations. And when that happens, then you typically infer that this is an unimportant gene. Or, more likely, it is a gene that does not even lead to a protein.

So, based on our previous examples and our discussion, you know that. You know dead genes, which are genes that are not producing proteins, are known as pseudogenes, right? So, pseudogenes typically end up showing this property. The third possibility, I mean <1, ≈ 1; then the only other possibility is $dN/dS > 1$. So, this is the situation in which non-synonymous mutations are accumulating at a faster rate than synonymous mutations. which means that this is a gene that is under very strong positive selection And that probably means that this is increasing the adaptation of the organism, and in that sense.

It is, you know, the continuous evolution of this gene is very important for the fitness of the organism. So, for example, this is the kind of stuff that happens in all those genes. which allows the organisms to evolve and adapt to new challenges. So, for example, let us say I have bacteria and you know you are hitting it with different kinds of antibiotics in a hospital situation. So, there are all kinds of genes that are responsible for antibiotic resistance.

And those genes are continuously being, you know, strongly selected for. So, any positive non-synonymous mutation is any non-synonymous mutation which is Leading to a positive fitness effect is going to be beneficial under that situation, and therefore, The total number of non-synonymous mutations is going to be greater than the number of synonymous mutations. So, these values quickly tell us what kind of selection is happening on the organism or rather on that gene. Now, based on this, we can look at our first property, which is that mutations accumulate at a roughly constant rate. What do I mean by that? So here I am showing you a graph in which the x-axis shows the time in millions of years.

Since the most recent common ancestor of humans and nine other species you know, numbered 1 to 9 on this blue line. You know the time since they diverged. So, what they did was let us say, "Let me show you what the species are." So, let us say there is a chimpanzee over here. So, what this is telling you is that humans and chimpanzees are based on their fossil data.

They diverged somewhere around, you know, a few less than 10 million years ago. Similarly, number seven is chicken. So, the ancestor of humans and chickens diverged somewhere around 320-330 million years ago, and so on. Now that is on the x-axis. On the y-axis, you have the number of various kinds of substitutions per site between humans and those species.

And this is not based on one or two genes; this is based on about 4,200 loci. Now there are three kinds of substitutions that have been discussed here. The blue one in this line is the synonymous substitutions. The red one, this one is amino acid substitutions, and then this one is non-synonymous substitutions. And of course, the corresponding lines—the blue, the red, and the green—these are the regression lines.

Now, what do you observe? The first and most interesting thing that you observe is that there is a strong correlation. between when the species diverged and how much sequence difference there is between them. In other words, you can see that this is almost linearly increasing. In the context of synonymous, it is slightly, ever so slightly, noisy. But in the context of the amino acid substitutions and the non-synonymous substitutions, it is very strongly linear.

In other words, the rate at which the mutation is accumulating is almost constant. And the second thing that you see is that if you look at these three classes, Synonymous, non-synonymous, and amino acid; the rates are different, right? I mean you can see straight away that synonymous substitutions are accumulating at a much faster rate. And in this particular case, this is about 5 times faster compared to the non-synonymous ones, the green ones here. But within a particular class, you know that things are still changing linearly, which means that the rate is constant. Now, obviously, if you have an observation like this, there is a very important practical consequence of it.

What is that? The practical consequence is that you are seeing that at least some genes. They accumulate mutations at extremely constant rates, and all those genes that do that, They can be used as molecular clocks to estimate the time of divergence of species for which we do not have any fossils. Now, how do you do that? So, just to show you what I mean. So suppose you have two species and you want to see what their time to divergence is. So what you do is compute the number of differences in amino acid sequences.

For proteins or nucleic acid sequences for genes for the given pairs of species, such as humans and some snakes. Now, why did I say snakes? Because if I look at this particular diagram, I do not see snakes in this data. But now what we can do is suppose you have a calibrated diagram like this, where you already know the relationship. Between divergence time as determined from fossil data and the substitutions per site, you know the lines look like this. And what you can do is take this calibrated relationship where both fossil data for divergence and substitution rates exist are available, and then use that for your unknown species for which you do not have fossil data. So, for example, let us say that we observe our synonym for the snake that we have in mind. The substitution rate for that is, let us say, somewhere over here. So, from this data, what we can do is say that the ancestor of that snake and humans is now clear. They, you know, diverged from each other, say, somewhere about 250 million years ago.

In other words, even if you do not have the fossil for that snake, given this relationship. And given that you can compute this substitution rate, you can figure out how long back the divergence was. And this concept is actually used extensively in molecular phylogeny to estimate the divergence times between species. Because, obviously, you know fossil records are very, very sparse. You do not always get a nice fossil to say, Okay, this is the time at which snakes diverged from the ancestors of humans. So, this ends up being very useful, but this leads to a massive question. Why should sequences evolve at such constant rates? I mean, we know that mutations themselves are, you know, not always, you know, we saw that they end up. Having very different rates, you know transition rates are different from transmission rates, and so on and so forth. But if all that is the case, then why should it be at the level of the genome? Why should it look as if they are evolving at a constant rate? So the answer to this question came from the work of the famous Japanese molecular evolutionary biologist, Motoo Kimura. And he proposed what is known as the neutral theory of molecular evolution.

So, what exactly did Kimura say? So he said that mutations that lead to beneficial or harmful fitness effects have either already been fixed or will be eliminated. Now then, he says that such mutations are beneficial or harmful. They actually have not contributed significantly to the variation of the existing sequences. Now, prima facie, this will look like a very bold assumption, but if you think about it closely, it is not. So if you remember when we were discussing the, you know, distribution of fitness effects (DFEs), We said that most of the mutations are actually not going to have any effect on fitness, but, Among the mutations that will have an effect on fitness, a large fraction of them are actually going to be deleterious.

So, all these deleterious mutations, since they are deleterious, what will happen? You know over time long time there is going to be selection against those mutations and they are just going to disappear from the population. At the same time, those mutations which are beneficial, if the benefits are reasonable, should be considered. They are going to slowly accumulate, and at some point, they are going to get fixed. So, that stuff is not going to lead to any variation at the level of the population because all individuals will have it. All the harmful mutations are, anyway, getting selected out.

So, when you think about the variation that is present within the population, it is actually going to be contributed. Primarily by those variations that do not have any effect on fitness, essentially those mutations that are neutral. Now we already saw that if there is something that is neutral, there is no selection acting on it. The primary thing that will lead to the changes in their frequency in the population is going to be random genetic drift. So, Motokimura essentially ended up saying to look at the level of the molecule at the level of the sequence.

The primary force leading to the variation we see around us is actually random genetic drift and not selection. Now this, of course, led to a massive debate: the so-called selectionist versus, you know, neutral mutation debate. But over time, people essentially figured out that, you know, Kimura is not really trying to say that selection does not work. He is essentially saying that, at the level of the sequence, it is the neutral mutations that play a greater role. And as we went forward, we realized that, you know, sometimes there are situations under which that works; there are other situations under which they do not work, and all those things we are going to be part of, we are going to see in the context of today's discussion. But the main thing I want you to take from this is the observation that Random genetic drift of selectively neutral mutant alleles has led to the majority of the intra- and interspecific variation. This is what the neutral theory of evolution tells us. Now, if this is the case, we already know that, or rather, we know that the magnitude of drift depends on the effective population size. Therefore, the amount of genetic variation within a species is going to be proportional to its effective population size.

Why is that so? Think about it. We have seen that when the effective population size is very low, the effect of drift is going to be very high. And when the effect of drift is very high, we know that there are only two possible things. Either a particular mutation or a particular allele is going to get lost, or it is going to get fixed, right? Frequency = 0 or frequency = 1, and once these happen, barring any back mutation, that is going to be the stable state. So, obviously the smaller the effective size, the lower the genetic variation is going to be. That is present due to neutral alleles in the population which is the same way of saying that you know, or another way of saying that if the population size or the effective population size is high, That is when the amount of genetic variation in the population will be higher. The next major thing that neutral theory claims, and that Kimura actually shows using mathematical derivation, We are not going there, but he claims that the rate of neutral mutations is going to be constant over time; therefore, The rate at which these neutral mutations become fixed in a population also becomes constant. So, that is the derivation that I am talking about. Now, think about it. If the rate of neutral mutations in a population is constant, and the rate at which these neutral mutations are being fixed is also a constant, Then, taken together, what is going to happen? That means that for these, you know any two species.

The rate at which these mutations will continue to accumulate will become a constant over time. Which will show up in the kind of figure that we saw in the previous slide, where, you know, as time progresses. As the divergence time progresses, the amount of mutation accumulating increases linearly at a constant rate. So, this is the explanation for why you have molecular clocks.

Now, that was the first observation. The second observation that is very interesting is that the different parts of the genome actually evolve at different rates. Now, what do I mean by that? So here is, you know, some figure about the evolutionary rate for different components of the genome. So this is estimated from the differences between humans and chimpanzees. And these are measured as the number of differences per nucleotide site multiplied by 10^3. So that is what is on the x-axis; that is what we are calling the evolutionary rate.

Now we have four different kinds of things here. These are the pseudogenes, and you can see that the rate at which they are accumulating differences is the highest. These are coding genes; you can see that the rate at which they are accumulating mutations is the lowest. Now, in some sense, both of these are obvious. Why? Because no selection is happening in pseudogenes.

So, you know, all mutations are selectively neutral. So it will keep on accumulating mutations without any, you know, cost to the organism. So obviously, the evolutionary rate of divergence over there is the highest. Similarly, coding stuff, non-synonymous mutations, many of them will end up affecting fitness. And therefore, the rate at which it will accumulate non-synonymous mutations in the coding gene is going to be much less. So these two are taken care of; these two are understandable.

Why are we getting this scenario where intergenic regions, that is, you know, if you have two genes, there are, you know, The nucleotide sequence between them is what is known as the intergenic regions, and then you have introns. What are introns? So you know that you have eukaryotes; you have these. In the genes, you have parts of the gene that are known as exons and parts of the gene that are known as introns. So the exon-intron-exon-intron is the structure of the gene. And when the mRNA is formed, the entire thing is transcribed, but after that happens, the intron portions are cut out.

And all the exons join together, and it is this, you know, edited version of the mRNA that goes for translation. Now, obviously, if there are changes that are happening at the intron level, Or if there are changes happening in the intragenic regions, then they should be selectively neutral. In other words, in them, you should have the evolutionary rate. The rate at which mutations are accumulating should be equal to that of the pseudogenes, but you are not really finding that. You find that these two things are actually intermediate between the pseudogenes and the coding genes. Why? That is because we now know that both the intergenic regions and the introns can actually affect the fitness of an organism. Why? Because the introns play a role in gene regulation, why is that? Because we know that there is something known as alternative splicing. So, you know, depending on which introns are present. The same version of the mRNA can lead to multiple proteins through different combinations of exons and which exons are being combined with each other, that the introns play a role in determining. And secondly, it has also been shown that the introns play a role in determining RNA stability. That is why, if there are mutations happening in the introns, they are not completely selectively neutral. Secondly, if you look at the intergenic regions, they do not have the coding region for sure, but they have regulatory sequences. They have promoters, they have enhancers, and what we now know is that they have many non-coding RNAs, you know we talked about these piRNAs, or you have what are known as siRNAs, and so on. And all these things play a role in determining the fitness of the organism. You know they affect the organism's functions. And that is why, if you have mutations in these regions, then Sometimes, they can affect the fitness of the organism depending on where they happen.

But obviously they are not as critical as crucial, you know. I am sorry, I will take that back. If there is a mutation that happens in a regulatory sequence or a promoter sequence, it is extremely crucial. But what I am trying to say is that not the entire stretch of, you know, the intergenic region, or not the entire stretch of the intron. is as crucial as, say, the entire stretch of the coding region, which is why the mutations in these regions, Intergenic regions and introns accumulate at a rate that is intermediate between pseudogenes and coding regions.

Different sites in the genome will also evolve at different rates. Now, how is this different from what I just told you? So what I just told you is in terms of the different kinds of regions, pseudogenes, coding regions, and so on. What I am now telling you is in terms of the positions in the open reading frame. So you know that we have codons that are made of three base pairs, right? But the three base pairs do not accumulate mutations at the same rate. So if you look at it, what we are showing you here is the proportions of base pairs. that differ in the DNA sequences of the mitochondrial gene called COI between pairs of vertebrate species.

So multiple pairs have been taken, and this is plotted against the time. Since their most recent common ancestor diverged, that is why this is in terms of millions of years. So they have taken pairs of species for which the fossil data are available. So, based on the availability of that fossil data, they have figured out what the time duration is between their divergence. That is what is there between, you know, that is what is there on the x-axis, and then on the y-axis, they have plotted. What is the sequence divergence in terms of the third base pair in the codon? The second base pair in the codon and the first base pair in the codon.

And what you see straight away is that when it comes to the third base pair, Then mutations over there accumulate extremely fast, but after some point, they more or less stabilize. Whereas, in terms of the first and second base pairs, The mutation accumulates much more slowly, particularly for the second base pair; it accumulates really slowly. Now, why is that so? Remember the Wobble Hypothesis. So we know that all three base pairs in a codon are not equally important. In most cases, the redundancies that you have in terms of the genetic code are because of the vowel in the third base pair.

So if you end up changing the third base pair, then in many cases the changes are going to be synonymous which is why it is very easy for a mutation over there to be selectively neutral. And anything that is going to be selectively neutral, as we know, is going to get you know accumulated much faster. Now, of course, there is a limit to this; there is only so much wobble that exists at the third base pair. So that is why, after some time, when you know in terms of divergence, After some time, there is no further increase in terms of divergence at the third base pair.

However, when it comes to the first and the second, then the redundancy over there is much less; in the context of the second, there is hardly any redundancy. So, because of this, if there are mutations in the first or second base pair, then, Most of the time, they are non-synonymous, and that is why they are not selectively neutral; therefore, they do not end up accumulating as much as you know what happens at the third base pair. So, in that sense, this is again congruent with the fact that anything that is selectively neutral will typically accumulate much faster. Now, this brings us to the very important relationship between genome size and the complexity of the organism. Now, let me start by saying that there is no accepted definition of biological complexity.

In fact, there is no accepted definition of complexity to begin with. But for practical purposes, you can roughly think of those organisms. which have let us say more types of cells or which have more kinds of tissue or, you know, let us say organisms that have a greater division of labor for various functions. Those are the ones that are more complex compared to those organisms that, let us say, are lesser along these axes.

So, for example, think of an E. coli. The E. coli does everything on its own in the single cell; it has to find its own food and reproduce. It has to, you know, if there is a danger, it has to go away from it; everything has to be done by that one cell. However, think of an organism like a human; I use a different set of cells for seeing. A different set of cells for hearing, a different set of cells for reproduction, and so on and so forth.

So, it is in this context that we have a larger number of cells. We have a greater number of tissues, and there is a division of labor across those cell and tissue types. And therefore, we can reasonably say that a human body is more complex than an equine. Of course, there are philosophical arguments that one can make. But let us, for the moment, assume that that is how we think about complexity. So, fossil and DNA evidence tell us that, in general, this is very important, not always.

But in general, evolution has led from simpler forms to more complex forms. So, that is why you know the earlier forms; the ancestral forms were typically unicellular, and then at some point multicellularity evolved. The ancient multicellular organisms typically had very few kinds of cells. But then the number of types of cells started growing, and so on and so forth. So, we have, of course, situations like parasitism, where complexity has gone towards simplicity.

An organism has ended up losing its function, various tissue types, and so on. But in general, it goes from simpler to more complex forms. But, mechanistically speaking, where is this increased complexity going to be coded? How are you going to get the features that are needed for this increased complexity? And as we have been talking about, you know whenever you need novel traits, Then the place from which they will come at some level will have to be some genes. You can do certain things by changing gene expression levels, but beyond a point, you need two genes. Therefore, one can reasonably argue that as you go from simpler to more complex forms, the number of genes that have to increase. In other words, we expect to see a positive relationship between genome size and the complexity of a species.

So, this is the expectation. Does it really happen that way? So, what I am showing you here are the ranges of the genome size for a bunch of organisms. So, we have the genome size on the x-axis and we have different kinds of organisms on the y-axis. And as you can see, these are the ranges, right? Now, of course, you can see that for any class of organisms, there is a huge variation that is not entirely obvious. So, for example, you know all prokaryotes are supposed to be unicellular.

So, why do I expect this much variation? Okay, maybe. But what ends up becoming more interesting is how, if you look at the overlaps. So, for example, right at the bottom, we have the birds, the mammals, and the reptiles, and you can see that. The range in their genome size is somewhere about one order of magnitude between 10^3 and 10^4. But now look at, for example, algae or look at protists; these are the best.

Look at the amount of variation that you have in protists. This entire zone is actually greater than this stuff, you know, than all birds, mammals, and reptiles. that there are quite a few protists who have genomes that are much larger by orders of magnitude. Remember, the x-axis is in the log scale. So, they have orders of magnitude larger genomes compared to birds, mammals, and reptiles.

Similarly, you now look at, for example, this fungi, alright. Again, look at the amount of variation that they have over here, but look at the protists. These are protists, which are single-celled organisms, but they are far outscoring the fungi. Similarly, many vascular plants, even the simple Arabidopsis and all Many of these guys have genomes far larger than those of mammals, birds, fishes, and chordates. Whereas, as far as we can see, they are not as complex and are much simpler.

So, what exactly is going on? Now, here I have to introduce a term, and that term is C-value. What is it? So, the amount of DNA that is present in the haploid phase of the life of that organism is what is known as the C-value. Now, why do we need to think about the C-value? Because some organisms are haploid, others are diploid. So, you cannot really do an apple to oranges comparison. So, in order to make a proper comparison, you bring everything down to the haploid complement.

So, when I was talking about it over here, this is all about haploid genome content. So, this observation suggests that there does not seem to be a relationship between the C-value of a species and its complexity. This is what is known as the C-value paradox. Genome size typically does not correlate well with complexity, and this is our fourth pattern. Now, the question is, why is that so? So, one potential solution to this emerges if you look at the composition of the genome.

So, we are showing you the composition of the human genome here. So, you see this little tiny sector over here; these are the coding sequences, the exons. Less than 2% of the human genome is actually devoted to protein-coding sequences. So, what are the major things? So, it turns out that the major things are actually different kinds of transposable elements: the jumping genes. About 50% of the total genome is composed of these transposons, and about 25% are introns which you know is actually a little more than 25%, to be frank. So, now how does this solve the C-value paradox? Because we said that an organism's complexity should be related to its genome size, But implicit in that is the assumption that the whole genome is contributing to the organism's complexity. But if the whole genome is not contributing to the organism's complexity, as we can see here, Only 2%, less than 2%, is doing so; then that probably means that you know that will explain why we are seeing this discrepancy. Maybe the organisms that actually have larger genomes, they are the ones who end up having a lot of quote unquote junk in their genome, which is not contributing to the complexity, which is why even if the genome size is large, the complexity is still small. But in order to understand whether that is indeed the case or not, We have to ask the question: is this thing that we saw in the context of the human true for all organisms? So, is it true that in all organisms the amount of non-coding DNA we find is similar? So, that is the question that was asked in a paper by, you know, Michael Lynch himself.

So, this is a very interesting graph. So, this shows us the relationship between a genome's total DNA content and the genome size. And the amount of its protein-coding sequence. What has been done is that there are three lines here. So, this line is for 100%, which means that if a point is on this line, then for that organism, 100% of its genome is a coding sequence.

This line over here is 10% coding sequence, and this line over here is 1% coding sequence. So, note what is happening here. So, if you look at this legend over here, you realize that these guys are the ones who have 100% coding sequence. Who are they? These are the bacteriophages; these are the eukaryotic DNA viruses. So you know viruses that affect eukaryotes, and these are the prokaryotes; these are the guys.

So, these are the ones with, you know, one little exception here and there. These are the ones in which the coding sequences actually occupy 100% of the genome, and they hardly have any non-coding DNA. But then the moment you come to, let us say, here, okay. So, here are the animals and here are the land plants; you see that they are diverging away, which means that These are the organisms that have a lot of non-coding DNA in their genome: unicellular eukaryotes. Sorry, these are the ones that sit somewhere intermediate between these guys, the plants, and the animals.

So, this is telling us that there is a lot of variation in the genome size, as we can see over here. But it is only in the animals and the plants that you have a lot of non-coding DNA. All these other organisms, such as bacteriophages, DNA viruses, and prokaryotes, etc. All those which have to typically replicate extremely fast have an extremely small generation time. Those are the ones that do not have all this non-coding DNA business, whereas all those organisms that have relatively longer life cycles, Those are the ones that have all the, you know, non-coding DNAs and unicellular eukaryotes that are somewhere in between in terms of life cycle, they are also somewhere in between in terms of coding and non-coding DNA fraction. So, this tells us that the coding to non-coding ratio of genomes varies across taxa. With simpler organisms generally having more streamlined genomes. More streamlined means a smaller amount of non-coding DNA. Great. But if we have found this, then that simply means that, look, it is non-coding DNA that is screwing up the calculation.

So, let us forget about the non-coding DNA. Let us go back to our C-value paradox and instead of saying that, There should be a positive association between genome size and complexity; let us simply say that. There will be a positive association between the number of protein-coding genes and the complexity of the organisms. Does this prediction hold? I will just show you a few examples. So, here we have some species, and I have organized this in ascending order of the number of genes you know, based on whatever we know from their whole genome sequences. So, this is Drosophila melanogaster right at the bottom; this is C. elegans; this is a nematode. These are us humans; this is a zebrafish; this is Arabidopsis; this is Daphnia pulex. Now, look at their genomes.

Drosophila has, you know, the smallest number of genes among all these examples that I have taken. But this is a nematode, which is, for all we can see, a far simpler organism than Drosophila melanogaster. Yet, it has about 6,000 extra genes. Homo sapiens, as humans, have about 19,000 to 20,000 genes which is roughly in the same zone as C. elegans, which is a small, tiny nematode. Zebrafish have way more genes than we do. Arabidopsis, you know, this small plant, the thale cress, is very widely used by plant molecular biologists. That damn thing has 27,000 genes, and Daphnia, that microscopic, you know, 4-5 millimeter-long organism. That we talked about in one of the previous lectures, that class, that thing has 31,000 genes, roughly speaking, it is a little less than double compared to the humans. So, taken together, it is pretty obvious. That there is no correlation between the number of protein-coding genes and the complexity of the organism either. And this is what is known as the G-value paradox in literature. No association exists between the number of protein-coding genes and complexity. So, now obviously comes the question: why? Why do we expect the G-value paradox, or rather, what explains the G-value paradox? So, it turns out that quite a few mechanisms have been proposed to explain the G-value paradox and I do not think there is one mechanism, all of these probably act simultaneously. So, for example, in many cases, you know organisms that have a lower genome and a lower number of genes. For example, as humans, we have more sophisticated controls over gene expression. So, for example, you know about epigenetic mechanisms. So, that even if you have a relatively smaller amount of genes, fewer number of genes, If you have good epigenetic ways of turning them on and off, thereby controlling their spatial and temporal aspects of expression that might lead to more diverse characters.

Similarly, in many of these organisms, the proteins are multifunctional. Instead of one protein doing just one job, maybe one protein does multiple jobs. Or maybe the proteins are such that the number of interactions happening in the metabolic pathways is much greater. Or maybe, I mean, this we already know; you know, there are, you know, alternative splicing and post-translational modifications. All these things are leading to the fact that you have just one gene, but the number of functional products. The complexity that is coming out of that gene is much greater, which obviously can lead to greater complexity with a lower number of genes.

And finally, this is again something that has been extremely well shown in many of these organisms that have Lower genome size; they have lots of non-coding RNAs, miRNAs, siRNAs, and all those other RNAs that I talked about. And all these things not only provide greater control of gene expression, but in some cases, they might end up doing. You know certain functions on their own, none of which will obviously show up in the gene count. So, the basic idea is that the complexity of the organism can be modulated.

Even with a lower number of protein-coding genes, you can have all these other features.

Great. So, until this point, and when I say until this point, I mean right from the first discussion we had in week one, Everything that we have discussed till now, you know about the various forces of evolution. Selection, wave, mutation, migration, etc., and how these forces act with each other. Everything that we have discussed has been within the so-called modern synthesis or the extended evolutionary synthesis frameworks.

And this framework supposes, presupposes that life has arisen. And then it starts talking about how things might have changed after that. But the crucial question of how exactly life originated, that question which is perhaps it is supposed to be the biggest question in the whole of biology, That question actually falls outside the purview of both MS and EES. Now, it is a very, very hard question. Why? Because there are no fossilized remains of those absolutely early ancestral life forms. Therefore, even if we end up having some nice theories about how those things might have worked, It is very difficult, close to impossible, to have proper evidence that This is the way it happened and not the 15 other alternative hypotheses that have been proposed.

So, it is a tough one, but it is something that is extremely interesting, nonetheless. And once you understand that, the other interesting thing that we have not worked on until now is about us, us humans. How did we originate, exactly? From where, what are the various forces that have worked? Evolutionary forces have worked on us to bring us evolutionarily to where we are right now. So, these are the kinds of questions that we are going to discuss in our next discussion, which is going to happen next week. Okay. See you then. Bye.