Evolutionary Dynamics
Supreet Saini
Chemical Engineering
Indian Institute of Technology Bombay
Week 01
Lecture 09

Hi everyone, we will continue our discussion on the central dogma and how information that is stored in DNA is passed on through mRNA molecules, finally resulting in these amino acid chains and protein synthesis. So we'll go through this example, a toy example that I have on the slide here. Given to us is a DNA sequence, ATG, CCC, AAA, AGC, and TGA. And obviously, I'm reading these in triplicates because, as we saw last time, when the ribosome machinery comes and sits on the mRNA molecule, it reads mRNA in these triplets and moves from one triplet to the next one. So let's imagine that this is part of a long DNA sequence.

We have the five prime end, a whole bunch of nucleotides, and this TGA is followed by the three prime end and a whole bunch of nucleotides here. And under certain environmental conditions, the machine RNA polymerase is recruited here. Let's call that RNA polymerase. And this leads to triggering the transcription of this fragment of DNA. And the resulting mRNA sequence will obviously look like this.

So we saw that during transcription, the two strands are separated from each other, and an identical copy to the five prime to three prime end is produced. The only difference being that instead of T, we have a U in the mRNA sequence. So after the process of transcription, What we have is an mRNA sequence, which is the following. So we have AUG, CCC, AAA, AGC, and UGA.

So transcription takes place, and we get the mRNA molecule. And now this mRNA molecule will obviously, let's again imagine. So I've purposefully written the DNA and, hence, the corresponding mRNA sequence starting with ATG because this is where protein synthesis will take place. ATG is the signal for amino acid assembly to start. But there'll be some nucleotides here also, and these provide the space for the ribosome, two subunits of the ribosomal machinery, to come and bind on the mRNA.

If you remember from our last lecture, the ribosome's two subunits come and bind at a sequence called the Shine-Dalgarno sequence, and then it starts scanning until it encounters

the first AUG. And that is a signal for it to read the mRNA sequence in triplets and start assembling amino acids. So through these three examples, what I hope to show you is how even a small change, a single nucleotide change, can have a large implication in what sort of protein a cell makes. So let's take this, the case that we have started with, as individual number one. And in this individual, when the ribosome begins the process of translation, the first triplet that it encounters is AUG.

So if we look at the genetic code, we have the first letter as A, the second letter as U, that is this row and this column, and the third letter is G. So AUG corresponds to methionine. So which means that when amino acids are assembled, the first amino acid is to be assembled as part of this peptide chain is going to be methionine. After that, the ribosome moves to the next triplet, which is CCC. That's the first C. This is the second C, and the third C. And that's this.

And that's an amino acid called proline. So this proline is linked with methionine, and we move to the next triplet, which is AAA. Which is A, A, and the third A, which is this. So the third amino acid is lysine. We keep moving.

Next one is AGC. So the first is A, the second is G, and the third one is C. So AGC corresponds to serine. And after that, we get UGA, which is U, G, and A. And that's a signal for the ribosome to stop the process of translation. And the two subunits disassemble, and we get this particular amino acid chain. Now, since this was only a toy example, what we have here is an amino acid chain which consists of only four amino acids.

If you look at real sequences of genes and proteins in E. coli, a typical protein will have It's a large range. We get very small proteins. We get very large proteins also. But a typical protein will have 200 to 300 amino acids in E. coli.

Obviously, for the purpose of this class, this example is very short and only has four amino acids. Now, let's imagine that in this individual two, This one nucleotide change happened, which means that this C was mutated into an A. And as you can see, when we are looking at these triplets, this has a change. Everything else remains the same. ATG, CSE, TG.

I think this T shouldn't be there. ATG, CAC, AAA. There's something wrong. Yeah, this should be an A. Let me rewrite this. So let's imagine that the second individual has a sequence of ATG, CAC,

A, A, A, A, G, C, and T, G, A. And the only difference that you can see is that instead of a C here, this individual has an A here. The rest of the sequence is the same. So let's try

and understand what this single nucleotide change will do to the amino acid sequence that we get. So as in the previous case, we first have the process of transcription take place, which results in the following mRNA. This sequence is essentially the same except for the fact that T changes to U, and we get this as the mRNA sequence.

And now when translation begins, we've already seen that AUG corresponds to methionine, and AAA corresponds to lysine. AGC corresponds to serine. So all of that remains the same. The only difference is that instead of a CCC, we have a CAC in the sequence. And if we see what CAC corresponds to, this is first letter C, second letter A, and the third letter C.

CCC corresponds to proline in the previous case, but now it has changed to CAC, which is first letter C, second letter A, and third letter C, which has become a histidine. Hence, because of this one nucleotide change between these two sequences, we get one amino acid change. What you should also realize here is that let's imagine that this individual, the change in this triplet, the change that happened here was CAC. But another change could also have happened where the CCC could have changed to CCT. In which case, after transcription, we would have gotten a CCU.

And if we look at the amino acid corresponding to CCU, we get C, C, and U, which also corresponds to proline. So in this case, CCC corresponded to proline, and CCU also corresponds to proline. So even though a nucleotide change happened, no change in the amino acid sequence occurred. The amino acid sequence would still be methionine, proline, alanine, and serine. These types of changes have a name.

This type of change where a nucleotide leads to a change in an amino acid is called a non-synonymous change. This is a non-synonymous mutation. This, on the other hand, this change of C to a T, because it did not result in a change of amino acid, both the sequences have the same amino acid sequence corresponding to it, is called a synonymous change. Meaning that even though the DNA sequence changed, the resulting amino acid sequence would be exactly the same in either one of these two cases, whether this was CCC or CCT. That's an example of a synonymous change.

In the last example, the mutation that we have is that this A has changed to a T. Let's quickly work through this example and see what this does to the amino acid sequence. So, as before, after transcription, the mRNA sequence is going to look like this. And we will look at them in triplets. Now, AUG corresponds to methionine, and that's fine. CCC corresponds to proline.

That's also as before. But now, instead of a AAA, we have a UAA. And if you look at what UAA corresponds to, that's the U, that's the second A, and that's the third A. UAA corresponds to a stop. So this is actually a stop. Nothing happens, and we get an amino acid chain, which is only two amino acids.

And the process of translation stops here. So what comes after it is irrelevant because the ribosomal subunits have already disassembled after they read this triplet UAA, which corresponds to a stop codon. So what you can see through these examples is how information contained in DNA is first transcribed into an mRNA molecule, and then ribosomes act on it to make the necessary proteins. A small nucleotide change, even the minimum change that we could have, which is just one nucleotide, could result in changes that lead to the alteration of the amino acid sequence in the protein. This one nucleotide change could also result in cases where there is no change in the amino acid sequence.

Or this one nucleotide change could result in cases where the protein synthesis process is truncated even before the entire transcript has been read. So, which of these three scenarios plays out obviously depends on where the mutation occurred, in which codon it happened, which triplet was affected, at what position the mutation occurred, and so on. But this sort of tells us the range of possibilities that exist for mutations in the genetic code. So before we go any further, let's look at the fact that most of our discussion in this course is going to be about E. coli as a model organism because we know the most about this organism. We'll also look at examples from yeast.

Saccharomyces cerevisiae, again a model system about which we know a lot. So before we go any further, let's just take a look at some numbers about the biology of the organism that we are going to be dealing with the most. So E. coli is the organism about which we know the most, and hence most of our examples are going to be picked from studies that have used E. coli as the model organism. It's a gram-negative bacterium, which means it has a cell wall and a cell membrane. It's a rod-shaped bacterium.

It looks something like this. This dimension would be of the order of 1 micron. This dimension would be of the order of 1 to 2 microns. The exact size of a bacterium is contingent on the environment in which it is grown. If we grow bacteria in rich conditions where there are lots of nutrients present, the temperature is optimal, bacterial cells divide faster, and not only do they divide faster, they tend to be larger also.

Cells grown in poorer conditions tend to be smaller, and they divide slowly also. The genome of E. coli is circular in nature, so we have this E. coli cell, and in it is its DNA. It's

circular, and the total number of nucleotides here is of the order of four to five into ten power six nucleotides. Depending on different strains of E. coli that we are working with, the number could be variable, but that's the order of magnitude of the genome size of E. coli. Again, as compared to us, human beings, our genome size is 3 into 10 power 9, something like SARS-CoV-2.

Its genome size is 3 into 10 power 4. So that gives you the relative scale of the diversity of the genome sizes that exists across the tree of life, although viruses are not really classified as living organisms. Okay, so last time we saw that when it came to lactose utilization, there is a transporter protein which is called lacY. This is the transporter which brings lactose from outside the cell to inside the cell.

So that's lacY. And corresponding to this, this protein is obviously a chain of amino acids, and its information is again obviously encoded in the DNA. So The chain of amino acids that encodes for this particular protein starts with ATG, as it does for every protein or nearly every protein, and it ends with TAA, which is one of the three stop codons. This sequence that starts with ATG and ends with one of the three stop codons corresponds to an amino acid chain, which would be formed when transcription and translation take place.

As a result, this particular sequence on the genome of the organism is called one gene. And every gene has a unique name, and the name for this particular protein, the name for this particular gene whose product, this protein, is responsible for the transport of lactose, is called lacY. Similarly, we saw that when lactose is brought in, it's broken down by another enzyme. We called it enzyme 2, which breaks it down into its constituent glucose and galactose. At some part of the E. coli genome, there is another sequence that starts with ATG and ends with TAA or one of the other two stop codons.

This particular sequence, when transcribed and translated, leads to this particular amino acid chain, which is responsible for the hydrolysis of lactose. And this gene has a name, LacZ. Similarly, we saw that there was a protein called lacI, which came and bound to the region upstream of lacY and lacZ and stopped transcription from taking place when lactose was not around. This, obviously, is again encoded at another location on the genome, ATG TAA, and this gene is called lacI. And you can think that these are just zoomed-in versions of some part of the E. coli genome, where if I zoom in and see, I see this particular gene sequence.

So this region of E. coli corresponds to these three genes. So if I were to ask you the question of how many genes E. coli's genome has, it's really difficult to answer if you don't

already know the answer. But E. coli's genome has somewhere around 4,500 genes. Of which we've just listed three, which help the organism utilize lactose for growth and energy purposes. But as we saw last time, genes like lacZ and lacY are only going to be transcribed and translated when lactose is the carbon source in the environment.

If there is no lactose, then I would ensure that transcription and translation of these genes are in the off state. So as a result, what we should learn from that is that in any given environment, even though E. coli has the capacity to make proteins corresponding to each one of these 4,500 genes, not all these genes are going to be transcribed and translated in a given environment. For many environments, these genes would be in the off state, such as when cells are growing in glucose, then E. coli has no need to utilize lactose and hence lacZ and lacY will not be transcribed or translated. So while all these genes are there, some of them are only transcribed and translated when needed. On the other hand, there are some genes which are always transcribed and translated because their products are essential for survival.

And you can think of an example of that as the proteins which make constituents of the machine called RNA polymerase. RNA polymerase is necessary for transcription of every gene and hence its presence is required for the cell to begin the process of protein synthesis, and hence proteins that constitute RNA polymerase are always transcribed and translated. So these can be branched into these two groups. One are sort of essential genes. These are

Always in the on state, which means their transcription is always happening. On the other hand, we have these non-essential genes. And their expression is based on need only. So, if the environment of the organism dictates that we need the product associated with this gene, then E. coli would have mechanisms to upregulate its expression. Otherwise, they would be kept in the off state. Okay, so we know that E. coli has 4,500 genes or so.

The typical length of a gene we saw was 200 to 300 amino acids. Let us imagine that the mean length is about 200. This is just for example. So, which means that the genes corresponding to it were 600 nucleotides. And this is 4,500 such genes, each of 600 nucleotides.

So, this is the number of genes. This is the length of each gene. And what that means is that if I were to just multiply these two numbers, that would tell me how much of the genome of E. coli is just taken up by genes only. And this is four zeros here, and 6 into 5 is 33, and 24, 27. This is something like 3 into 10 to the power of 6.

We've taken the average gene length as 200. It's perhaps a little more. It's about 250, which tells me that the E. coli genome, which is about 4.5 into 10 power 6, A large part of this genome is only comprised of regions which actually get transcribed and translated into proteins. So, bacterial genomes in general have this property that they have a lot of regions of the DNA which are actually transcribed and then translated.

And this region is approximately 80% or higher for bacteria. Contrast this with human beings where this region is actually only about 2% or so. 98% of the regions in human beings and 15-20% in bacteria is this region which is called the intergenic region. This is the region of DNA which does not itself get transcribed or translated into protein molecules. However, what regions such as this do is we just saw in the example of LacI that this might be the region where the LacI molecule comes and binds and stops transcription of LacY from taking place.

So these regions are often involved in controlling the dynamics, the amount of the protein molecules that they are setting upstream of. And these regions for this reason are called promoter regions. So we know the number of genes, phenotypic length, what fraction of the genome is comprised of the genic region. The last bit that we want to discuss here is how many protein molecules an E. coli has. What we mean by that is let's imagine this bacteria growing in an environment and in this environment, it has 4,500 genes.

Let's imagine that what are needed are 2,500 genes. These are needed and are being actively transcribed and translated, while not needed and hence not being expressed in this environment are another 2,000 genes. For each one of these, for each one of these, the process is going to look like first transcription takes place, then translation takes place, and we get the amino acid chain, which folds into protein, and the protein does its job. However, what we should realize here is that in this process of transcription and translation, it's not as if one gene sequence will only give rise to one protein molecule.

Very often what you will have is that several mRNA molecules will be produced from the same gene, and then ribosomes will come and sit on each one of them. So let's say there are these five transcripts that are made via transcription of this particular gene. And what you would have is that ribosomes would come here also. Ribosomes would come here and start scanning. And obviously, each results in its own amino acid sequence.

What also happens is that while this is scanning, by the time a ribosome moves ahead on this mRNA transcript, another ribosome will come and bind and start the process of translation. So if you were to look at one mRNA sequence, one mRNA sequence, we would

find several ribosomes sitting on it and translating simultaneously. Obviously, this one has just started, so it doesn't have many amino acids.

This one is a bit further along, so it has several amino acids. This one is way further along, so it has lots of amino acids. And so on and so forth. So, what we are trying to say is that for every gene that is being transcribed and translated, you could have many, many protein molecules corresponding to that same gene. Hence, 2,500 genes being transcribed and translated does not mean that E. coli only has 2,500 protein molecules at any given point in time.

In fact, the number of protein molecules in E. coli at a given point in time is around 10 to the power of 6, or 1 million. So, that is the diversity associated with E. coli. With the proteome of E. coli at any given point in time. So, with this basic understanding of a cell—what it looks like, what some of the numbers associated with it are—we'll start looking at the processes which generate molecular diversity in E. coli and how selection then acts on the diversity that is so generated, so that evolutionary change can take place. .