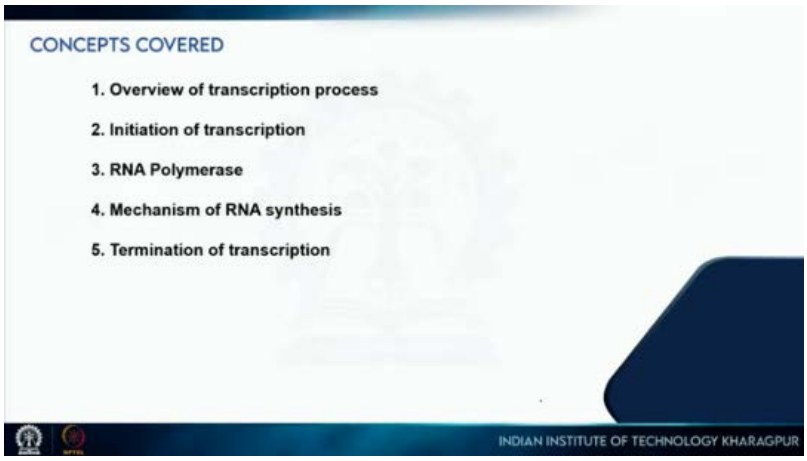
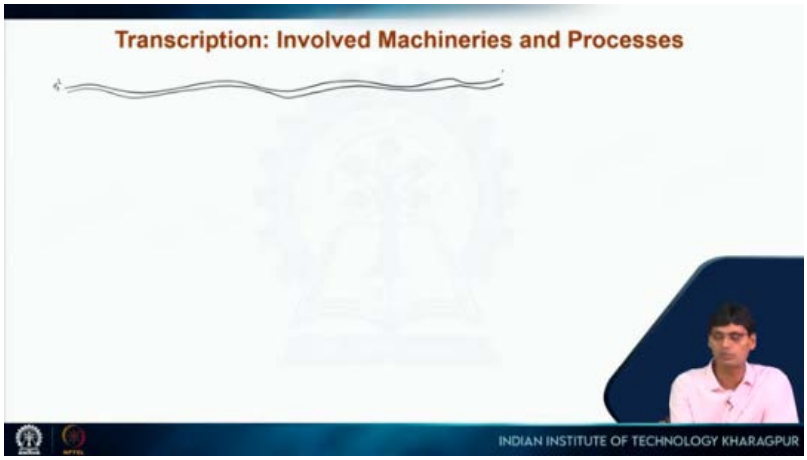
Introduction to Complex Biological Systems
Professor Dibyendu Samanta and Professor Soumya De
Department of Bioscience and Biotechnology
Indian Institute of Technology, Kharagpur
Lecture 7
Transcription: Involved Machineries and Processes

Hello everyone, I am Dibyendu Samanta from IIT Kharagpur, and today I will be discussing transcription, the involved machinery, and processes. This is the second lecture of module 2. Here, I will be covering these topics, an overview of the transcription process, initiation of transcription, followed by RNA polymerase, the mechanism of RNA synthesis and finally, how transcription stops. So, now, I would like to first give you an overview of the transcription process.



So, as I already mentioned in the previous lecture, we have around 20,000 protein-coding genes in our genome. So, now, in that respect, I can say that in one DNA molecule or in one chromosome, we have multiple genes.



So, for example, if I say that this is part of one double-stranded DNA, and since DNA is anti-parallel, those two strands, so I am just putting here something like this and DNA can be very, very long, and one DNA can have multiple genes, like 500, 1000, something like that, the number of genes present. So, now, I would say this is one gene that can code a protein, and this is another gene. So, something like that is present throughout our DNA molecule. But now, as we can see, for example, in our body, we have nerve cells; the function of a nerve cell is completely different from the function of a liver cell, and their shape, size, and morphology are also different.

But now, if we see their information, the DNA content present in a liver cell or a nerve cell, the information is the same. So, now the question is, how then are they doing different functions? Why are they completely different in terms of their shape and size, everything? So, this is because of how that information is being processed to make the product. So, now, here, in order to process the information present in this gene, for example, I am just naming this as gene A. How will we make the product from this gene?

So, the idea here is if I just show this portion, just I am targeting the rest of the DNA here now. So, I am just mentioning it here. So, this is my portion of this gene. So, here we have gene A.

So, in order to start transcription, that means in order to make or initiate the RNA, we have to understand where we have to start because most of the DNA or most of the nucleotides are not transcribed into RNA. So, as a result of that, something should be there just ahead of the gene which will initiate the process. So, here I would like to mention that we have

some specific sequence before the gene. So, this sequence is called the promoter sequence, some kind of specific sequence present here. So, this promoter sequence will invite RNA polymerase, right?

So, as a result of that, this promoter sequence will invite RNA polymerase, which is the enzyme RNA polymerase. From the name, you can understand, just like DNA polymerase. So, this is also making a polymer, but a polymer of ribonucleotides and that is why this is called RNA polymerase. So, now compared to replication, this is much more simple, not that many topological and logistic problems are present here. So, we do not need helicase, topoisomerase, or a lot of enzymes because RNA polymerase alone can

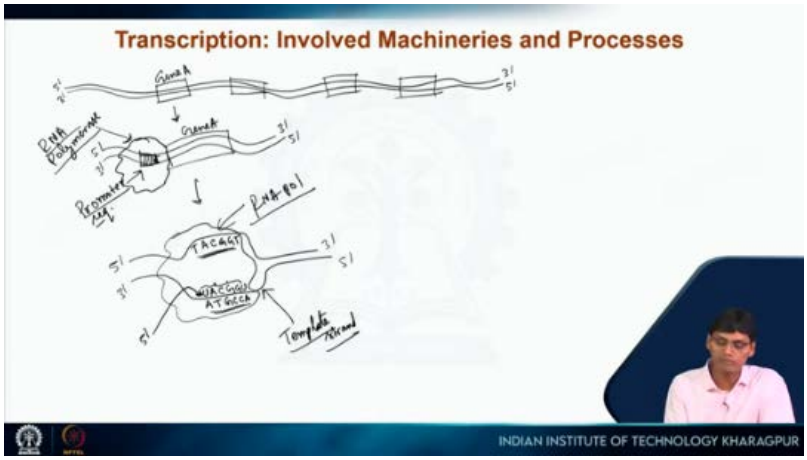
break double-stranded DNA just in a small part so that we can have access to a single strand and we can make the RNA. So, here if I go to the next step, it looks something like this. So, I would say this is 5', this is 3', and 3', 5'. So, RNA polymerase is present somewhere here. So, this is our RNA polymerase.

The synthesis of RNA is just similar to DNA synthesis, as it will always grow from the 5' to 3' direction. So, as a result of that, I would say that whatever sequence is present here in the template, the RNA will be synthesized in this way. For example, in this DNA, if I say the sequence here is A T G C C A in this strand, the complementary strand sequence should be T A C. G G T, and we can continue like this. So, that means when the RNA is being synthesized, this is template-directed synthesis because whatever is present in the template, it will make a complementary sequence.

So, as a result of that, in this RNA, what will be the sequence? So, here instead of A, T will be coming here, but T is not present in RNA. RNA polymerase will not recruit T, instead it will have U in this RNA, and then A C G G U; in this way, this RNA will grow. So, now if we see the sequence of this RNA that is being synthesized, this sequence of this RNA is similar to this sequence, exactly similar to this sequence and complementary to this sequence because, in this case, this strand of DNA is acting as a template

strand. So, that is why we are getting the complementary sequence here, which is similar to this strand, and that is why we say this is the template strand and this one is the coding strand or sense strand. Because whatever code or sense is present in this strand of DNA, it

is coming into this RNA, and that is why this is called the sense strand or coding strand, and this is the template strand. So, I will discuss this in more detail in the next slide.



So, here, this is just the RNA polymerase synthesizing this RNA from the 5' to 3' direction. And then, we should have some mechanism so that the RNA polymerase will understand that the gene is over; otherwise, it will transcribe the whole DNA, but that is not our goal. So, as a result of that, some stop signal should be there, just like the promoter sequence where this process started. Similarly, there should be some stop signal, and therefore, the RNA polymerase will stop there, RNA synthesis will be done, and RNA will be coming out from this complex. This is a little bit of a summary of transcription. I want to discuss in more detail about this promoter first, which recruits RNA polymerase, followed by RNA polymerase's action that how it actually helps in RNA synthesis, followed by the mechanism of RNA synthesis, and finally, how the chain terminates, that means, how transcription stops at the right position.

So, as you can see here the promoter, promoter is just like a pointer that points to the location of the information, in this case, particularly the location of the gene to be copied into mRNA. So, this is, if I give you an analogy, similar to, for example, in a DVD player, like in a DVD. We have, I would say, 100 different songs which I really like, and for example, this morning I want to play a particular song from this list. So, whatever is on that list, maybe the 50th song, I am really liking to hear today.

So, I do not have to listen to number 1, 2, 3, 4, all these songs. I can directly select the 50th song by some pointer device's pointer mechanism. This is similar, based on the requirement

inside the cell, we have a very complex regulation mechanism. The cell will guide properly so that the particular gene will be transcribed. So, that is why before the gene, we have the promoter sequence, and it will help to start the process.

So, as we can see here in this sequence, whatever we are writing here, like double-stranded DNA and 5' to 3', the sequence we are mentioning here. So, here the starting point of the gene, as you can see, this is the starting point of the gene, this A. So, the starting point of the gene, the first nucleotide, we denote as plus 1 site. So, I would say plus 1, then the next nucleotide T here, this is plus 2, plus 3, it will go like this.

And just upstream, in this case, on the left side, we will go minus 1, minus 2, there should not be 0. So, whenever the gene starts, the first nucleotide is plus 1, and towards the left side on the upstream, it is minus 1, minus 2 position. So, now what happens in case of this gene, when it gets transcribed, as I just mentioned in the previous slide, for example, after transcription, we are getting this product, this A U U G C is written here.

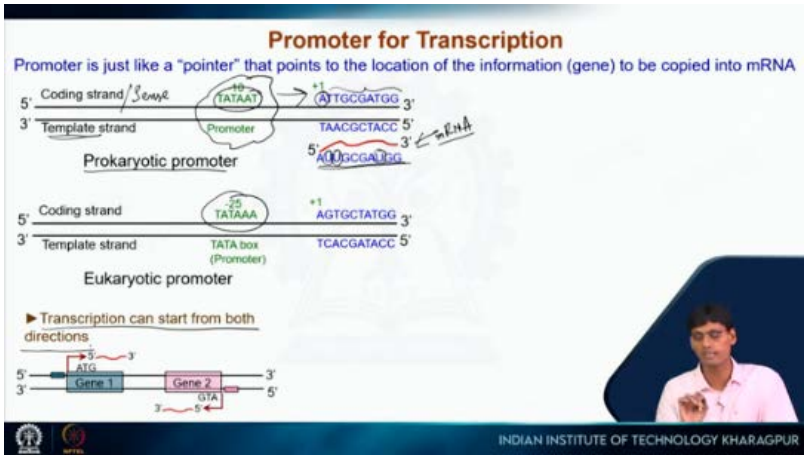
Now, if you see this sequence, this sequence is exactly similar to this sequence. Only instead of T, we have U here; we have U here and that is all, otherwise exactly the same. So, as a result of that, this strand, the top strand here, we would say this is the coding strand or sense strand; whatever the information present, we are directly writing this information in RNA. So, that is why this is the coding strand, and in order to do that, we use this bottom strand as the template strand; sometimes we also mention it as the non-coding strand.

So, now then, where is our promoter present in the case of bacteria, particularly the prokaryotic promoter? Approximately at the minus 10 position, that is why I mentioned which one is the plus 1 and minus 1 side. So, at the minus 10 position, we have some kind of conserved sequence, some kind of consensus sequence. This is called TATAAT, this kind of sequence is present in the prokaryotic promoter and it recruits the RNA polymerase. Therefore, RNA polymerase will come here and then it will go in this direction, and finally, we will have our RNA. So, if it is protein-coding RNA, then we can say this is mRNA.

So, now, the fundamental mechanism is the same in prokaryotes, like bacteria, as well as in eukaryotes, like in humans, plants, everywhere but the position of the promoter, the sequence of the promoter, that could all be a little bit different but the basic understanding

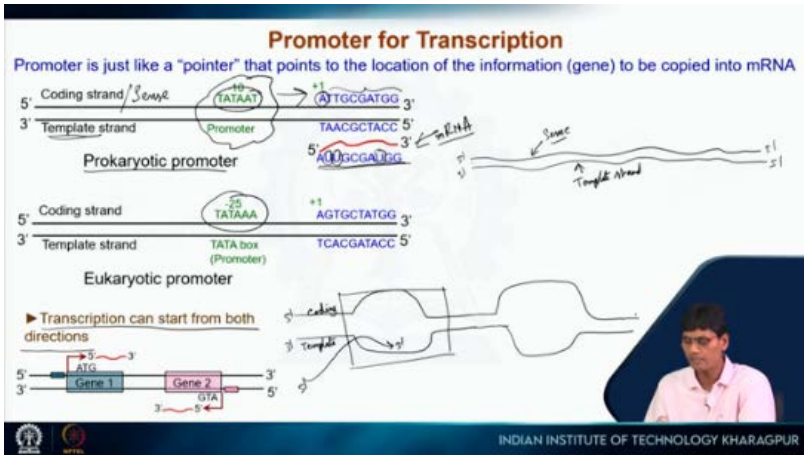
is the same everywhere. So, now, as you can see here, the promoter sequence of the eukaryotic gene is a little bit away from the starting point of the gene. So, this is around at the minus 25 position. This is not just one promoter, and some other regulatory elements present in DNA are also contributing to gene expression or in transcription itself. We are not going into that much detail.

Now, one important aspect I want to mention here is that transcription can start from both directions. What is that? So, RNA and DNA always get synthesized from the 5' to 3' direction, but since I discussed the coding strand and template strand here. So, as a result of that, I want to clarify here that if this is one long DNA. So, maybe here we have a few million bases present like 5' to 3' and 3' to 5' and few 100 genes are also present.



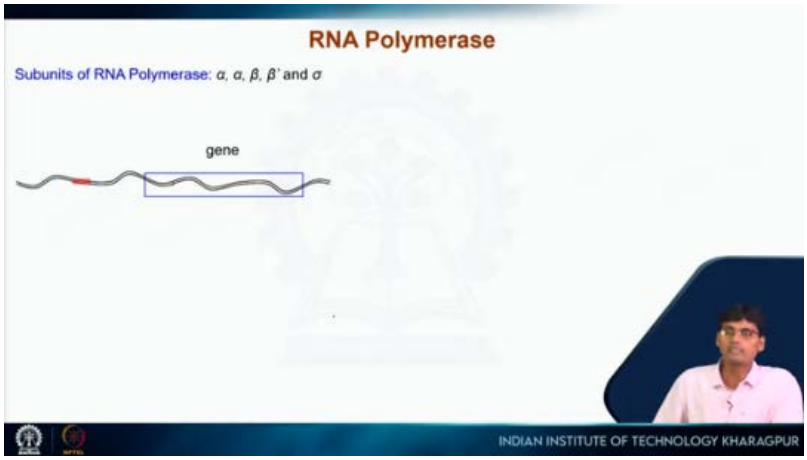
So, as a result of that, like whatever the way I just explained, I would always say that this top strand is the sense strand and the bottom strand is the template strand; not like that; generally, what we see in textbooks, we always see this kind of image where you will see that 5' to 3', 3' to 5', and the RNA is being synthesized in this direction. So, 5' to 3' synthesis of RNA, and as a result of that, if this is the case, then here, this is the template strand, and this is the coding strand. But maybe this is true for this gene, this portion of the DNA, but maybe for the next gene, this will be completely different. Maybe the next gene is present here, and I would say transcription can happen in this way also. So, this is 3' and this is 5'. I would say here transcription can happen like this, so again, 5' to 3' direction of synthesis. As a result of that, what we would see is whatever sequence is present here, like A, T, G, C, and the template strand of DNA, it should be here T, A C, G, and in this RNA, it will be

U, A, C, G. So, as a result of that, now this RNA sequence is completely similar to this, just the difference is U and T. So, as a result of that, now this is the coding strand, the coding strand.



So, for this gene, the coding strand is this one here. So, that means the promoter sequence present here, whatever the promoter sequence we are telling that T, A, T, A, A, T, is present here. Similarly, in this gene, the promoter sequence, this T, A, T, A, T is present here, and that is why it is happening. So, what I want to mention here is that in our DNA, it is a long sequence, but here information can be present in both strands. Both strands can be coding strands or template strands, depending on the gene, its orientation, and everything about that.

So, the system is very complex, but we have a very good mechanism, we have very good machinery to read it properly and make the correct product out of it. That is all. The next thing I want to discuss is the RNA polymerase. I just mentioned that the promoter sequence invites RNA polymerase to the site of the gene or just at the starting point of the gene. So, as you can see, this is the promoter sequence, and here the RNA polymerase Now, I would like to discuss a little bit more about RNA polymerase.



RNA polymerase is an enzyme, but now this enzyme can be a little bit more complex, that means it has multiple subunits. So, subunits mean that multiple polypeptides can be together and they can function together. So, that kind of structure is called the quaternary structure of a protein, which will be discussed by Professor Soumya De in the next module. So, that is called the quaternary structure of a protein. RNA polymerase is very complex, and as you can see, there are 5 polypeptide chains present in the polymerase to the alpha subunit, then beta, beta prime, and sigma. Those are just, you know, names given by scientists. That is all. So, all these five polypeptides together make the RNA polymerase. Now I mentioned that RNA polymerase is getting recruited to the promoter site, but how? So, if we see the sigma subunit, this is the sigma subunit of RNA polymerase. So, it first recognizes the promoter sequence. So, it has some affinity. Whatever the protein, like the polypeptide sigma, it can somehow interact with the promoter sequence, and therefore it first gets recruited to the promoter site.

This is followed by the recruitment of other subunits like alpha, beta, and beta prime. Now, this is the total RNA polymerase, right? This is RNA polymerase, RNA polymerase. Now, this total enzyme together is called the holoenzyme. But now, in order to start transcription, the sigma factor should be released from the complex. So that the RNA polymerase can move in this direction and it can make RNA out of this gene. Why? Most likely, this is because sigma has a very high affinity towards the promoter site.

So, as a result of that, the function of the sigma factor or sigma subunit is to find out the location, like finding the promoter. Then, other polymerase subunits will identify sigma

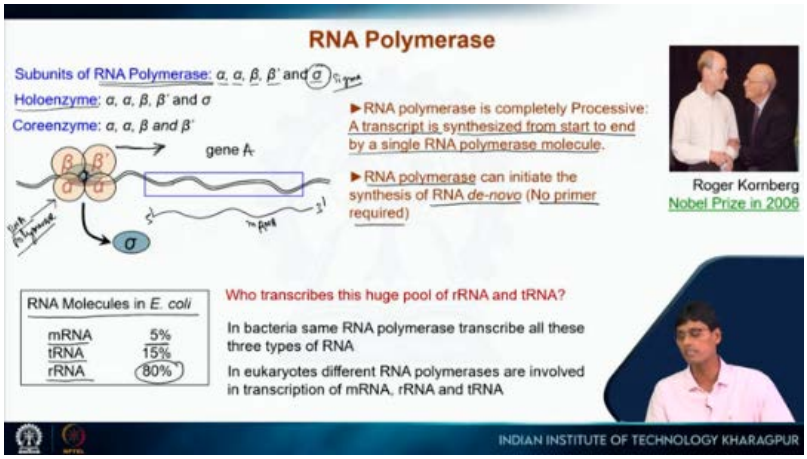
and come and make this complex. But now, if sigma gets dissociated, then RNA polymerase can move. Because of the high affinity between the sigma factor and the promoter sequence, it might be difficult to move the RNA polymerase in this direction. So, as a result of that, whenever sigma is released, RNA polymerase can efficiently move towards the downstream of this gene and it can make the product. It can make the RNA out of it. Now, a few important things I should discuss about RNA polymerase. The RNA polymerase is completely processive.

So, what is the meaning of processiveness? That means our transcript is synthesized. Transcript here means the RNA, the result of transcription, that is called a transcript here. A transcript is synthesized from start to end by a single RNA polymerase molecule. So, that means if this RNA polymerase starts to transcribe this gene, for example, I am saying this gene A, then we will get the RNA here. This is after transcription. This is, I would say, mRNA. mRNA is in the 5' to 3' direction, but this is being synthesized by a single RNA polymerase. Here is the difference between replication and transcription.

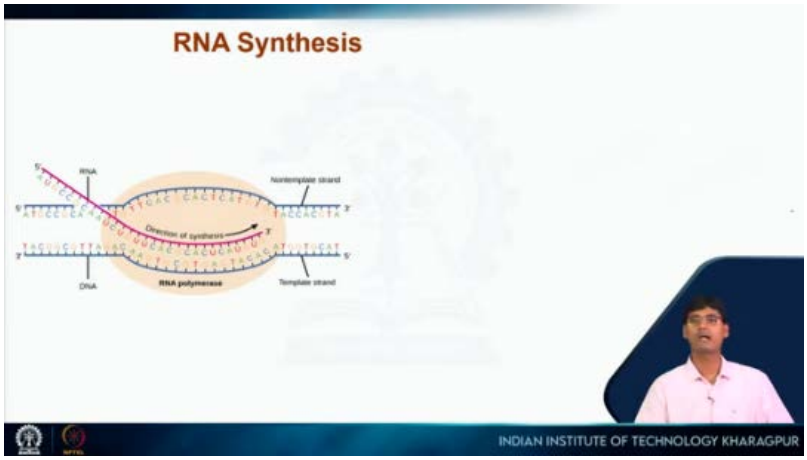
In replication, as you know by now, multiple DNA polymerases take part to replicate one single DNA molecule. Particularly, as you can remember, in the case of lagging strand synthesis, many DNA polymerases are recruited to synthesize DNA in a discontinuous manner. In addition to that, another important thing about RNA polymerase is that it can initiate the synthesis of RNA de-novo. That means no primer is required.

Like during replication, we discussed a lot about the primer, where an enzyme specifically called primase adds some RNA primer, followed by DNA polymerase synthesizing the DNA from that point, starting from the RNA primer itself. But in this case, no primer is required, meaning RNA polymerase has the ability to directly start the process of transcription. That is why we say it has de-novo activity, the capability to synthesize or start RNA synthesis. Here, if you see some statistics, RNA molecules are present in E. coli. So, if we see the total RNA molecules present in bacteria, particularly in E. coli, we will find that only 5 percent of RNA are mRNA or messenger RNA. So, that means they directly code for the protein.

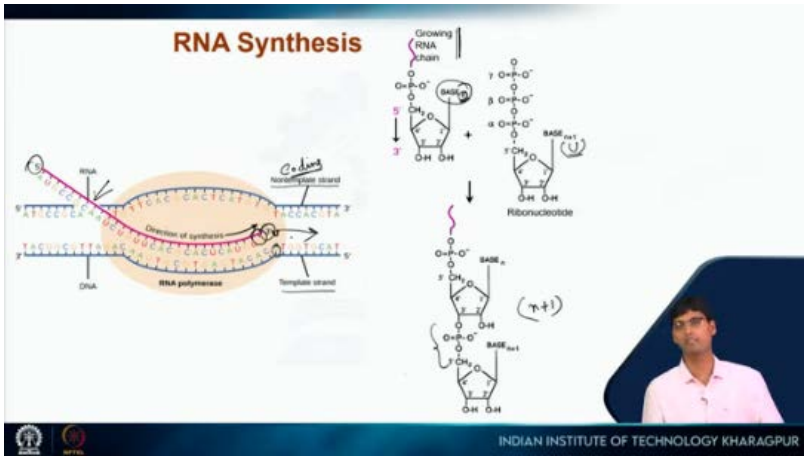
Then, the rest of the RNA are tRNA and rRNA. A huge portion of the RNA is actually ribosomal RNA. Now, as you can see here, who transcribes this huge pool of rRNA and tRNA? In bacteria, the same RNA polymerase transcribes all these three types of RNA, but in the case of eukaryotes, different RNA polymerases are involved in the transcription of mRNA, rRNA, and tRNA. So, in eukaryotes, we have three major types of RNA polymerases, they are RNA polymerase 1, 2, and 3, and they have specific functions. Particularly, RNA polymerase 2 transcribes mRNA in the case of eukaryotes. Here is an interesting fact. During the replication class, I mentioned that Arthur Kornberg received the Nobel Prize because of his contribution to understanding the process of replication. So, he is Arthur Kornberg here, and here you can see Roger Kornberg.



Roger Kornberg is actually the son of Arthur Kornberg, and he also contributed a lot to understanding transcription. He also received the Nobel Prize in 2006 for contributing a lot to understanding transcription. So, from the same family, one contributed to understanding DNA replication, and the next generation contributed to transcription, and both of them received the Nobel Prize. This is really interesting. Now, the RNA synthesis mechanism. The overall mechanism is exactly the same as what I discussed during DNA synthesis because of the 3' OH present in the sugar. In this case, it is ribose sugar, where both 2' and 3' are OH groups, and from there, the reaction starts, and phosphodiester bond formation takes place.

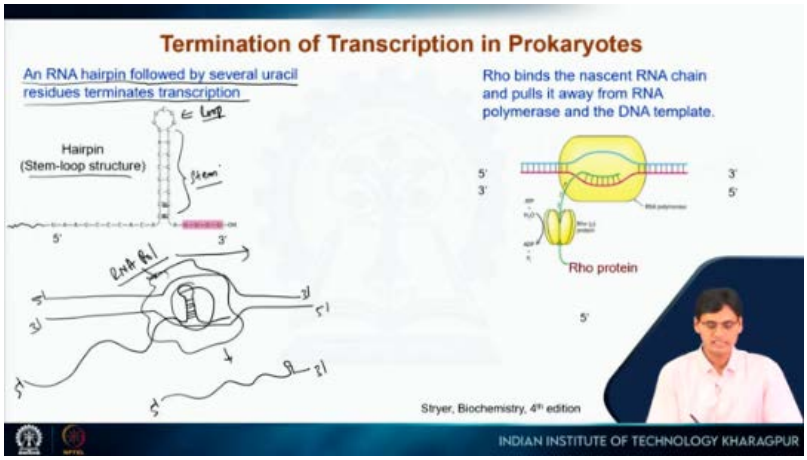


As you can see in this slide that is the RNA, the RNA which is being synthesized from 5' to 3', and we mentioned here the template strand and this is the non-template strand, whatever I mentioned, the coding strand, the same thing. Now if we see the chemistry, I would say this part, this is the growing RNA chain, this is the part. So, as you can see here, the growing chain, If I say the total number of nucleotides present is n, n number. Now, in the next step, we have here an A residue, an A nucleotide in the template strand. So, as a result of that one residue, that will be U should be added here. So, if I say this is a U nucleotide, U uracil. So, it will be added here through a phosphodiester bond here. So, then the total number of nucleotides present in this RNA is n plus 1. So, this way, this RNA will be growing in this direction, from 5' to 3' direction, that is all about RNA synthesis. Now, the synthesis started, then at some point, it should be stopped, otherwise, it will be going and going, it is unnecessary making RNA out of DNA. So, as a result of that, we have some mechanism, particularly in the case of bacteria, an RNA hairpin followed by several uracil residues terminates transcription.



As you can see, this is the growing RNA chain. So, we have many residue nucleotides in this side, the 5' so in the 3' side, because of the specific sequence present at the 3' side of this RNA. So, they make some intrastrand base pairing, as you can see here, that G C. So, they will make some hydrogen bonding. So, this is intrastrand base pairing, and this is the hairpin structure or stem-loop structure, this is the stem, and this is the loop.

We say this is a stem-loop structure or hairpin-like structure, but this is happening because of a specific sequence present at the end of the gene. However, this will destabilize the complex mechanically and will inhibit the progression of RNA polymerase through the gene. If I try to show here something like this, I would say the transcription is happening. So, if I say this is 5' to 3' and 3' to 5' of DNA, and RNA is being synthesized. If this stem-loop structure is like here, Watson-Crick base pairing, intrastrand base pairing between G-C and A-T, not actually it should be A and U. So, now this structure actually destabilizes this complex and it will inhibit the progression of RNA polymerase present here. This is our RNA polymerase. So, it will inhibit the progression in this direction. So, as a result of that, finally, this RNA will be dissociated from this DNA and we will get the RNA. So, as a result of that, transcription will stop in this way. So, now if you see carefully, following this stem-loop structure, we have multiple U residues here in mRNA.



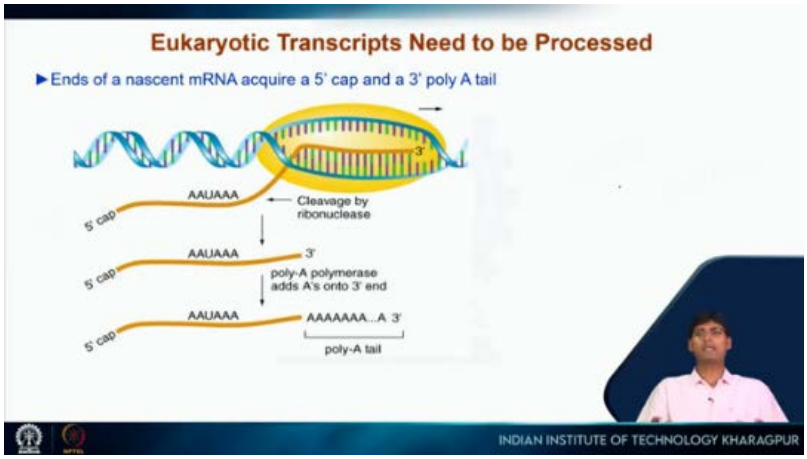
Now, if we have multiple U residues in mRNA, that means in the template, we have multiple A residues present. So, as a result of that, stem-loop structure followed by multiple U residues means that the template and the nascent RNA have Watson-Crick base pairing. If you have now U on mRNA and A in the template strand, that means you need less energy to dislodge this nascent mRNA. So, as a result of that, I would say that the stem-loop structure followed by multiple U residues even helps in removing nascent RNA from this transcription complex.

But scientists observed that when they are trying to do transcription, they observed that in-vitro transcription that means we can carry out transcription inside a tube also. So, we just need some template DNA, some RNA polymerase and the raw material ribonucleotides and we can start transcription. RNA polymerase should be there. So, now while they were trying to synthesize RNA in-vitro, they found that sometimes the product, the transcript particularly, is much longer than the gene itself. Then why is it happening?

If all genes have some specific sequence at the 3' side which can make this RNA hairpin structure, then the length should correspond to the length of the gene, but they found some product, some transcripts are much longer. They found that this is not the only mechanism, another mechanism exists inside the cell, inside bacteria that is called Rho-dependent transcription termination. Rho is a protein, which requires some energy. It recognizes some sequence and this Rho protein is a complex protein, so it will migrate and finally it will dislodge this nascent RNA or the mRNA from the transcription complex. But during in-vitro transcription, it is happening inside the tube, and we are not adding any other protein,

no Rho protein there, that is why the length of those RNA are much bigger than the gene itself.

So, as a result of that, those are the two major things in case of prokaryotic transcription where hairpin structure or stem-loop structure as well as Rho protein stop transcription. Now, in eukaryotes the overall mechanism, like I always say, the fundamental mechanism is the same, but there is a little bit of variation in eukaryotes. We have some more complexities, that is all.

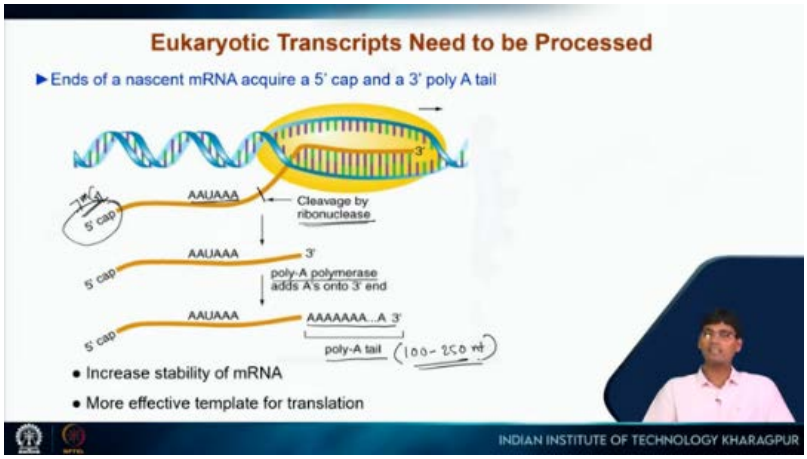


So, as you can see here, eukaryotic transcripts need to be processed. Transcripts mean the RNA, particularly here we are considering mRNA that need to be processed. As you can see that the RNA, while it is being synthesized, the 5' end of this nascent RNA, messenger RNA is attached to some kind of modified nucleotide, here modified guanine residue that is called 7 Mg. Some methyl group is attached to this G residue, 7 Mg cap should be added at the 5' end and then when the transcription is going on, some particular sequence for example, here A A U A A A, this sequence will recruit some enzyme called ribonuclease. From the name again you can understand that it will cleave RNA, the ribonuclease. So, it will cleave RNA. So, ribonuclease will cut this RNA here.

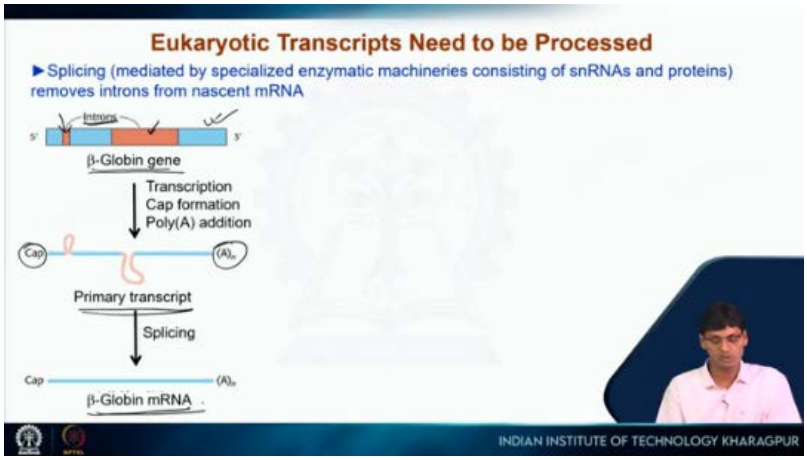
So, as you can see here, some special things happen in case of eukaryotic mRNA. That one enzyme is called poly A polymerase and from the name itself, we can say what is the function of this enzyme polymerase. But it is poly A polymerase. So, as a result of that, it will add only A residue at the 3' end of this mRNA. So, that is why this is called poly A polymerase, and this is template independent. So, that means, in the template we do not

have many corresponding residues, complementary residues present, not like that. This is template independent synthesis.

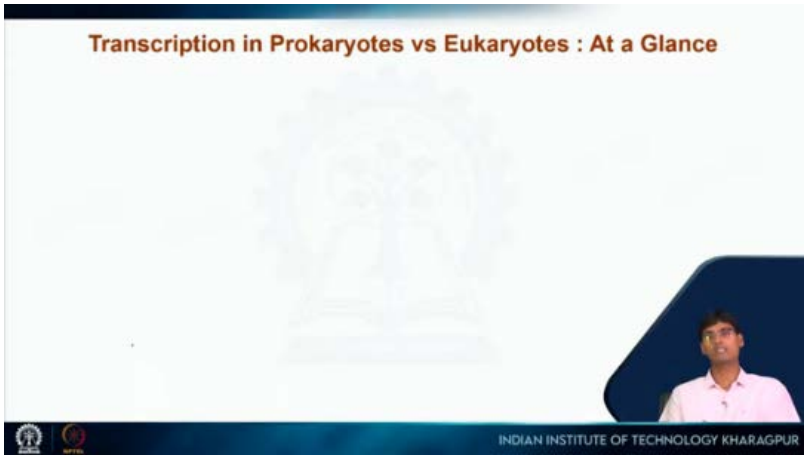
So, it adds a Poly-A tail at the 3' end. So, generally, the length of this poly-A tail is 100 to 250 nucleotides long. So, all these things, like the 5' cap as well as the poly-A tail, increase the stability of mRNA, and because of this, it is a more effective template for translation, which I will discuss during the translation. Now, here, if we see, eukaryotic transcripts also need to be processed. So when we are getting the mRNA from this gene.



So, this is the DNA; this is the gene, the beta-globin gene, for example, here we are mentioning, and this is RNA here, as you can see we have the 5' cap and 3' poly-A tail. This is called the primary transcript as it is not yet processed. So the meaning is that some portions of this gene called introns should be removed. So, finally, we will get something like this. This is the mature RNA. So, this is the processed RNA where the introns are removed.



This process is called splicing. Splicing is also required for eukaryotic transcripts to make the eukaryotic mRNA, and that will be used for translation in the next step. So, as a result of that, a few things are different in prokaryotes and eukaryotes. Particularly if I say about the RNA polymerase in case of eukaryotes, we have different types of RNA polymerase and they have different roles. Like RNA polymerase 1, 2, 3, but in the case of bacteria or prokaryotes, we have the same type of RNA polymerase, which transcribes all different types of RNA, including mRNA, rRNA, and tRNA.



If we see here, the processing part is really different between prokaryotes and eukaryotes. In case of prokaryotes, no additional steps are there but in eukaryotes, we need to have these processes a step that is the addition of the 5' cap, the 3' poly-A tail, as well as removing the introns from the mature mRNA. I should also mention the location of this transcription. So, in the case of prokaryotes, as we know, we do not have the nucleus so there we have just one compartment. As a result of that, transcription is happening there.

But in the case of eukaryotes, in the case of human cells, we have the nucleus, and all of us know that DNA is present inside the nucleus.

So, as a result of that, transcription, as well as replication, these two steps are happening inside the nucleus but translation is happening outside the nucleus. So, whatever I told you, this mature mRNA, after processing, should come out from the nucleus into the cytoplasm in case of eukaryotes and then it will be translated into protein, which I will be discussing in the next class. So that is all about this class. You can refer to these books for additional reading. That is all.