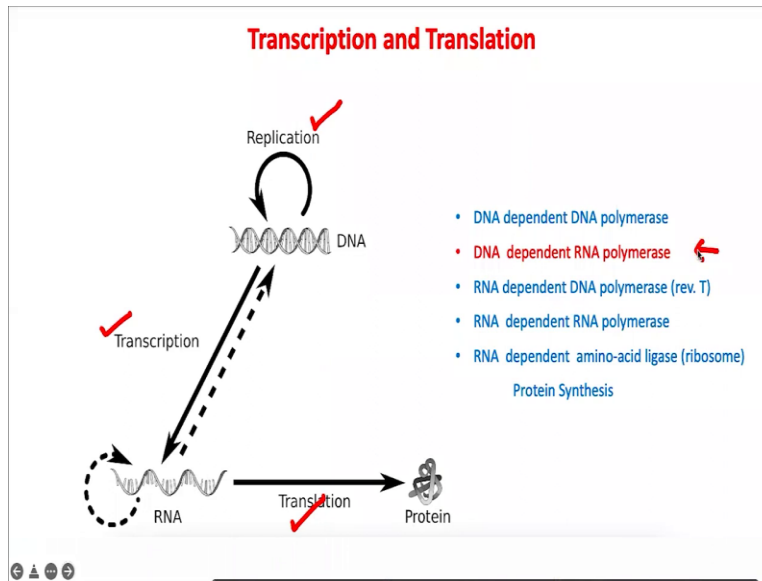


**Introduction to Cell Biology**  
**Professor Girish Ratnaparkhi and Nagaraj Balasubramanian**  
**Department of Biology**  
**Indian Institute of Science Education and Research Pune**  
**Central Dogma: Transcription - Part 1**

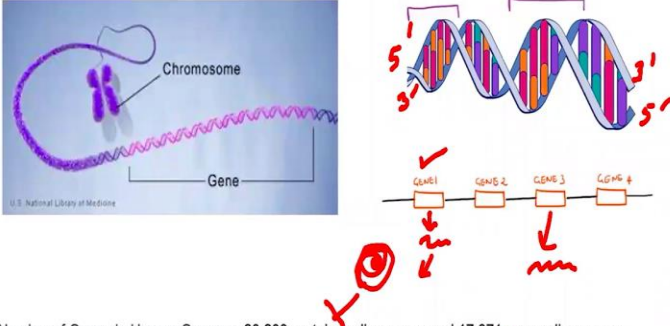
(Refer Slide Time: 00:15)



So, I will move on to the second step of the information transfer in the so called central dogma of molecular biology. We have already looked to a significant detail about DNA replication and we will now move on to looking at the processes of transcription and translation. And the molecule we will look at is DNA dependent RNA polymerase.

(Refer Slide Time: 00:47)

### Genes and DNA

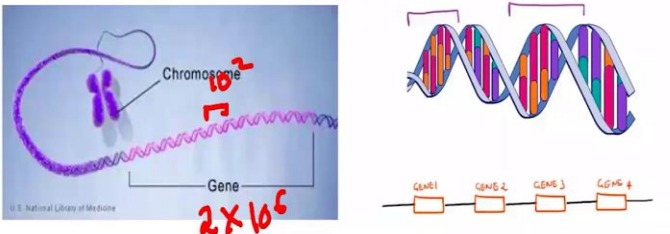


Number of Genes in Human Genome: 20,203 protein-coding genes and 17,871 non-coding genes

Number of bases in human genome =  $3 \times 10^9$  (3 billion). (on average  $10^5$  bases per gene)

Longest Gene: Dystropin, 2,220,390 base pairs ( $2.2 \times 10^6$ )

### Genes and DNA



Number of Genes in Human Genome: 20,203 protein-coding genes and 17,871 non-coding genes

Number of bases in human genome =  $3 \times 10^9$  (3 billion). (on average  $10^5$  bases per gene)

Longest Gene: Dystropin, 2,220,390 base pairs ( $2.2 \times 10^6$ )

So, let us make sure before we look at transcription about what exactly a gene is and what is the relationship between a gene, DNA and chromosomes. And obviously all of you have done biology in 11th and 12th would have a more clearer understanding of this as compared to the students who dropped biology in 11th and 12th. So, it is, think of it as a one slide refresher.

So, this is a nice pictorial view of the relationship between double stranded helical DNA which you have seen a lot about and a chromosome. So, a chromosome literally is nothing but a single long strand of double stranded DNA which is folded and compacted into a structure like this.

Now, it is important for you to understand that in a cell at any time in your body DNA is not really in this form or shape. Chromosomes are seen and this high level of packing of DNA into a chromosome is usually a process done just before cell division simply because a packed compact chromosome, DNA entity which is basically a chromosome is easier to move around especially when you have decided to break your house into two, give half of your material goods to one house and other half of your material goods to another house.

So, having a chromosome which is like a packed suitcase makes it easier to push it around and carry it around. It is as simple as that. So, under normal circumstances in a, let us say, a functional differentiated cell in your body which is not going to undergo mitosis, DNA is actually in a very unpacked and diffused state. And in this unpacked and diffused state, it is still very much in contact with the nucleosomes and it is in the form of a chromatin fiber and this of course is true only in eukaryotes and not in prokaryotes.

And when you look at naked DNA which is what we are doing now which is not the form in which it exists, they are always DNA binding proteins which are associated with DNA, then we can take a stretch of DNA as shown over here in pink and call it a gene. And in a simplistic and not too wrong visualization, as you look along the length of a double stranded piece of DNA, you will have genes along the length of the DNA and you can basically give them numbers or names called Gene 1, Gene 2, Gene 3, Gene 4.

So, here we are dealing with the double helix which has, which is an anti-parallel double helix. You have information coded in one strand and for redundancy complementary information coded in the other strand. Each strand is going in a different orientation. Let us just label it as 5 prime, 3 prime and 5 prime, 3 prime.

Now, I have already told you that there are approximately 3 billion bases in a human genome and for every animal, plant, bacteria, archaea, viruses the number of bases differ. Now, we also know that they are approximately 20,000 protein coding genes in the human genome and approximately again let us say 20,000 again, the number is very precise, which is shown over here, non-coding genes.

Now, what is the difference, the genes, and this is one gene, we will code for an m-RNA which in turn will code for a protein. So, this becomes a protein coding gene. Whereas, there are genes

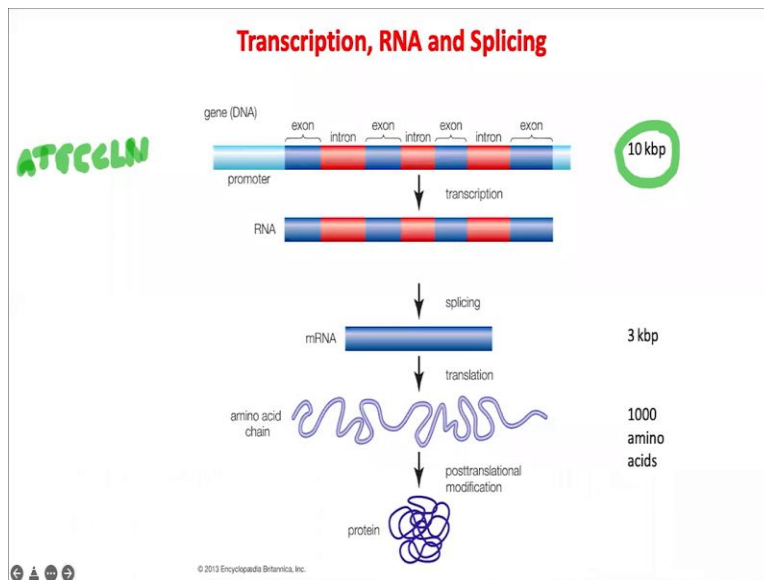
where, which will code for a piece of RNA, but this piece of RNA will not go forward to make a protein, it will remain as RNA and we call these non-coding genes.

Now, very clearly I have been emphasizing the fact that a stretch of the double stranded helix is a gene. And since we know approximately that there are 3 billion base pairs in the human genome and we know how many genes are there and I am approximating, approximately 30,000 which is an approximation, on average each gene will occupy  $10^5$  base pairs. This should be pairs. This is a typing error. I am sorry about that.

Now, it turns out that genes are hugely a variable in lengths. For example, the longest gene in the human body is a gene for a protein called Dystropin and this region encompasses  $2 \times 10^6$  base pairs. So, that is a huge amount of a length of a gene, basically  $2 \times 10^6$ . At the same time, you will also have small genes, let us say,  $10^2$  which are 100 base pairs long, not too long at all.

So, you have a wide range of genes and a gene, the way we are defining it over here, is basically a stretch of a double stranded helix which codes for RNA and in some cases like this RNA continues to code for protein and in some cases like this the RNA does not continue to code for a protein. So, this is straight forward and simple enough.

(Refer Slide Time: 06:10)



Now, let us look at the structure of a eukaryotic gene and this is again simplistic but it is very accurate. So, now, this is a stretch of DNA, double stranded DNA, and we basically will look at one strand of the DNA and we will worry about which strand we are looking at, because each strand has complementary information and we usually read of one strand and the information makes sense on one strand.

Now, in this stretch we will basically say and start giving nomenclature in defining regions. So, each gene and now this is very approximate, this is a number I have created which is let us say 10 kilo base pairs, which means it is 10,000 base pairs in length. There will be a region which will be called as a promoter.

And this will be again remember this whole stretch is nothing but ATGCs. It is just ATGCs in different combinations, CCC, TTT. So, this is just DNA. And in this DNA we are basically trying to say that there will be defined regions which we can label with English, using the English language and we will call the region on the left hand side and I will define what the left hand side is in terms of orientation as a promoter region.

And after that we will divide the rest of the gene into two basic categories. We will call the ones we have labeled as blue as exons and ones we have labeled in red as introns. And there will be another light blue region in the end which is basically the terminator region. So, this is pretty much the structure of a gene. A eukaryotic gene will have a promoter region which will extend the range varies, let us say 500 base pairs to 1 kilo base, so 1,000 bases to 500 base pairs.

We will have pieces which we will call exons and we have pieces which we will call introns and then we will have a terminator region. And very obviously the machine DNA dependent RNA polymerase is going to read this gene and it is going to make a RNA, a piece of RNA.

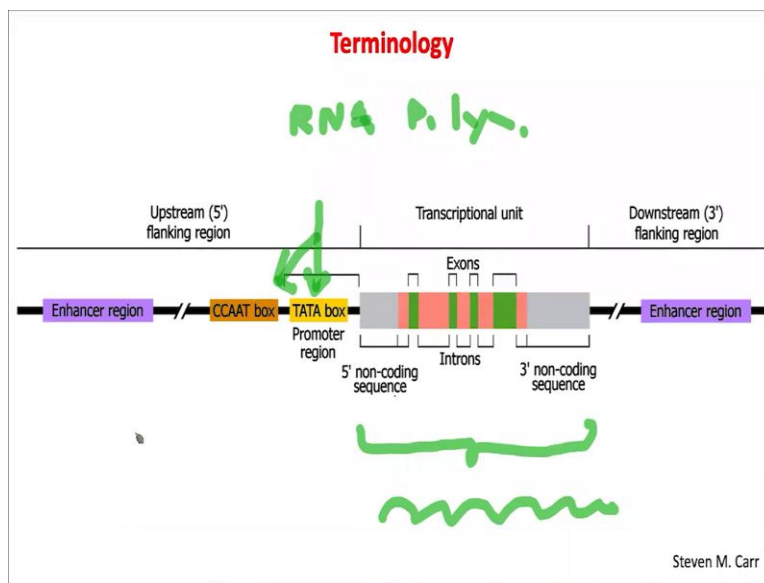
The piece of RNA will have both these exons and introns and again this is a simplistic picture. And the net result is the immature pre-RNA as we call it will become the mature RNA after it removes the red pieces from the total information it is copied from the gene. So, the mRNA which is now going to be translated into a protein basically does, is missing these red pieces and which why looking back knowing that these pieces are no longer represented in the code we call these as introns.

So, basically these are pieces of information which are available on the genome on the DNA but are not used to make a protein. So, effectively, there is an editing step which we term as splicing which happens and this editing step removes so called unwanted information from the pre-RNA to make the mature RNA which is going to be converted into a protein. And again these are all approximate sizes.

Let us say the gene encompasses about 10 kilo bases, the RNA will pretty much encompass something short of 10,000 bases. And after splicing a lot of RNA would have been removed before it is converted into protein and you end up with a much smaller piece of nucleic acid, single stranded mRNA.

So, you start with double stranded DNA. You use one of the strands to copy a complement which is RNA. The complement is spliced which continues to be the mature RNA. And the mature RNA is translated by the ribosome into a linear polypeptide chain which is amino acid chain which then folds into a special structure called as a folded state of a protein and this is all we are going to do in this class and in the next class. Is there a question?

(Refer Slide Time: 10:18)



So, now let us look at more detailed terminology. And this may look complicated, but it is not really all that complicated. And hopefully you will be able to, for those of you have not done biology before, you will be able to call back. Now, pretty much this is the region of interest which is going to make the RNA. The location where the protein binds is approximately in this

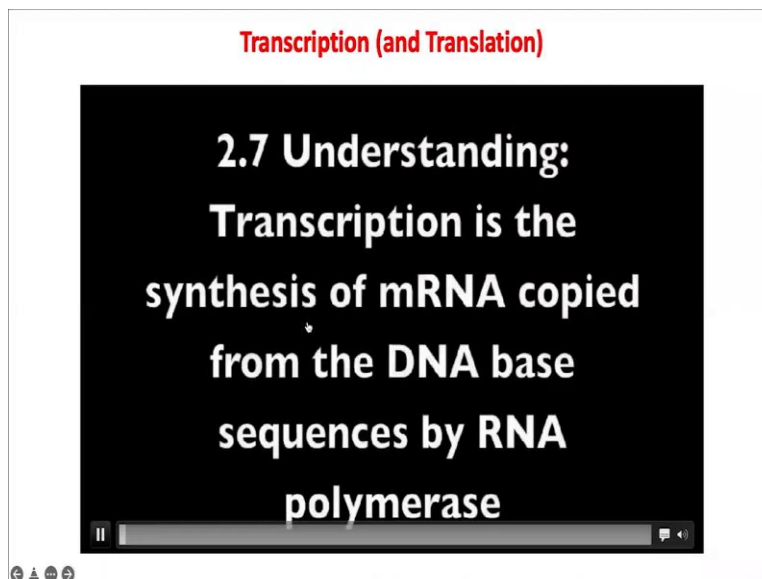
location. So, there is a stretch of DNA which works as a binding site for called dependent RNA polymerase.

Now, once RNA polymerase binds, it opens up the double stranded DNA and starts reading one of the strands and it makes a copy, an mRNA copy of that one strand of DNA. And the first part of is usually called the non-coding sequence, because this part never really is converted into protein.

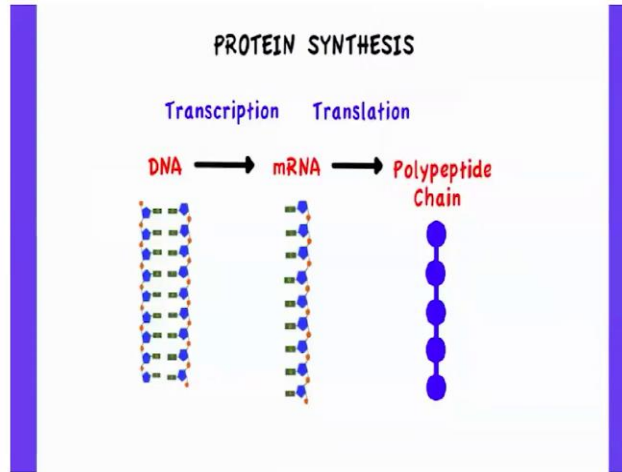
Then come the pink exons over here, sorry, the green exons over here and the pink introns. All of them are copied including the so called non-coding sequence. And at the end, the RNA will contain the 5 prime region, the 3 prime region, the main gene, but the introns are going to be spliced out. I will explain what enhancer regions are a little later.

So, this is the structure of a gene which basically consists of introns and exons, 5 prime and 3 prime leader sequences, a piece of DNA which is there to bind RNA polymerase and all this is nothing but a single stretch of double stranded DNA out of which only one strand is being read and copied into mRNA.

(Refer Slide Time: 12:12)



## Transcription (and Translation)

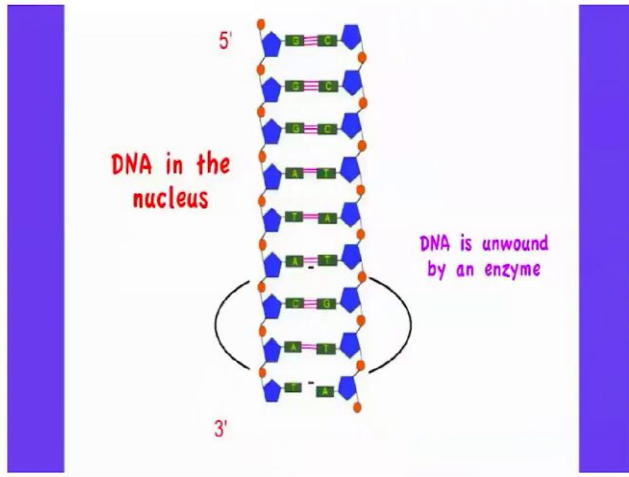


## Transcription (and Translation)

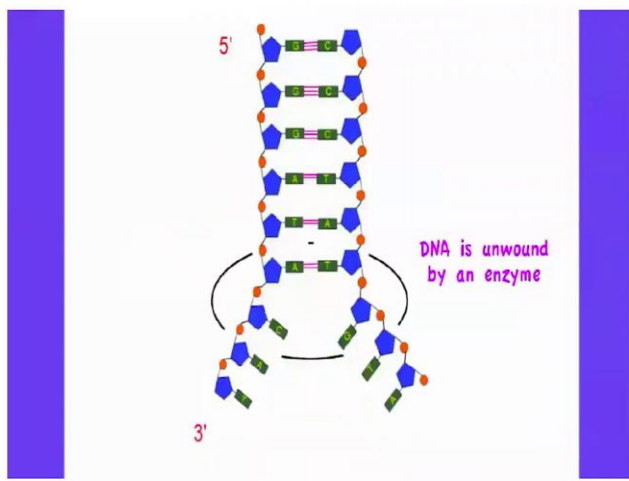
**Transcription**



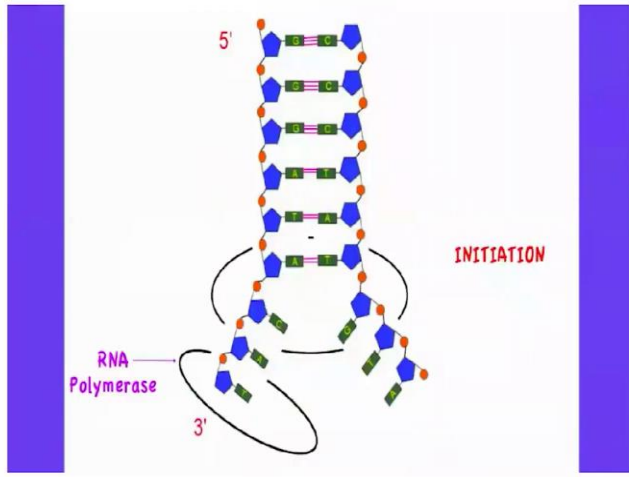
### Transcription (and Translation)



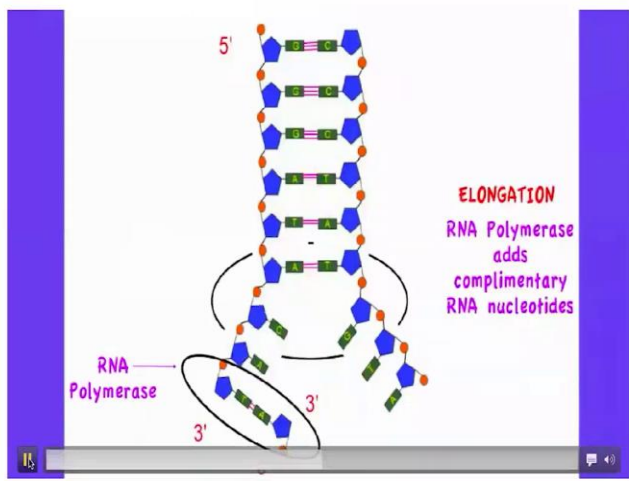
### Transcription (and Translation)



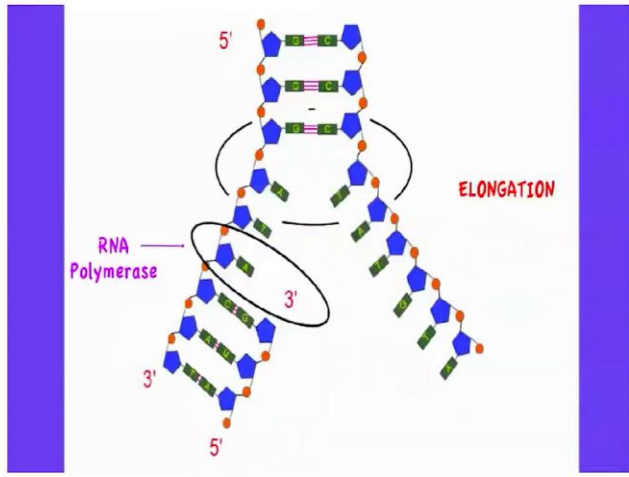
### Transcription (and Translation)



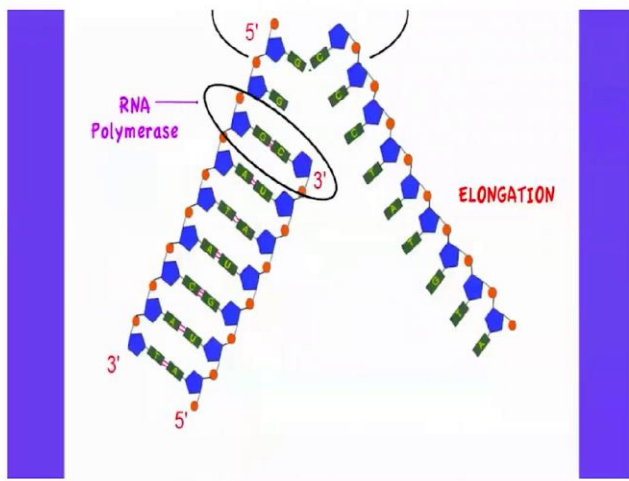
### Transcription (and Translation)



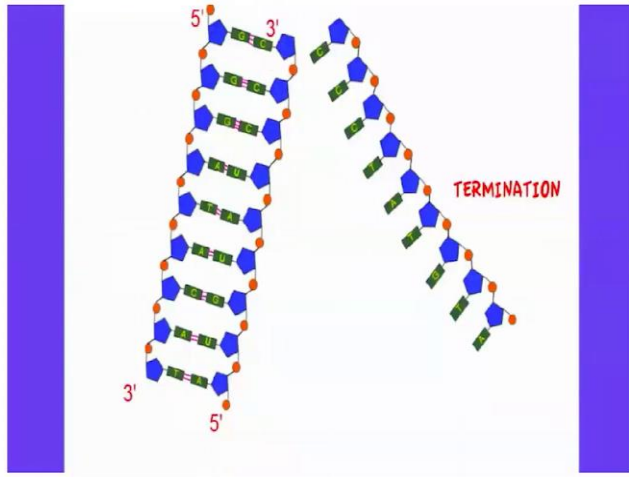
### Transcription (and Translation)



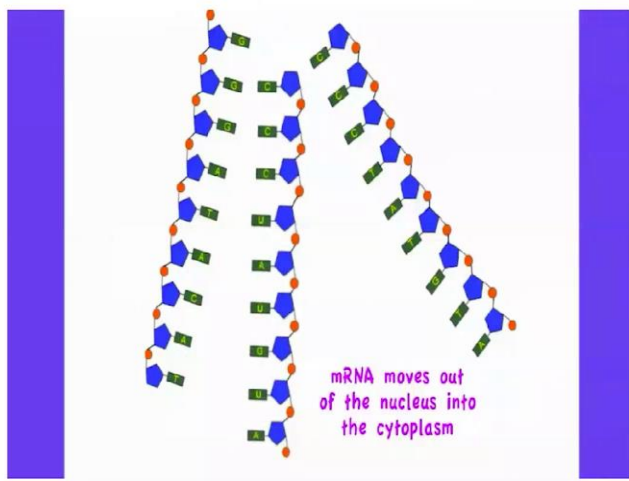
### Transcription (and Translation)

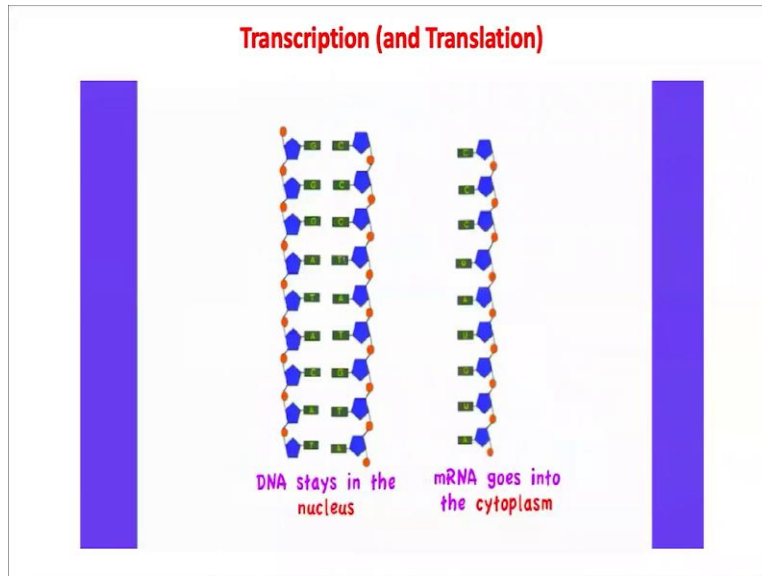


### Transcription (and Translation)



### Transcription (and Translation)





So, now let us look at a simple example of transcription. Before outlining DNA transcription, I feel it is important to give an overview of protein synthesis to put it in context. First of all, we have the DNA and where we find the gene on the DNA this would be transcribed to form a piece of mRNA this is known as transcription.

That mRNA would then move from the nucleus where transcription took place into the cytoplasm, where it is going to be used to form a polypeptide chain in a process known as translation. Ultimately that polypeptide chain would be folded to form a protein. In this video, we are just going to outline the first part the transcription of DNA to form a piece of mRNA.

Transcription begins when a gene is located on DNA and that particular part of the DNA is unwound or unzipped by an enzyme. Once that particular area has been unzipped you then find DNA as two single strands and one of those strands is going to be used as a template upon which it forms a piece of mRNA. So, please notice which strand and which direction is being copied.

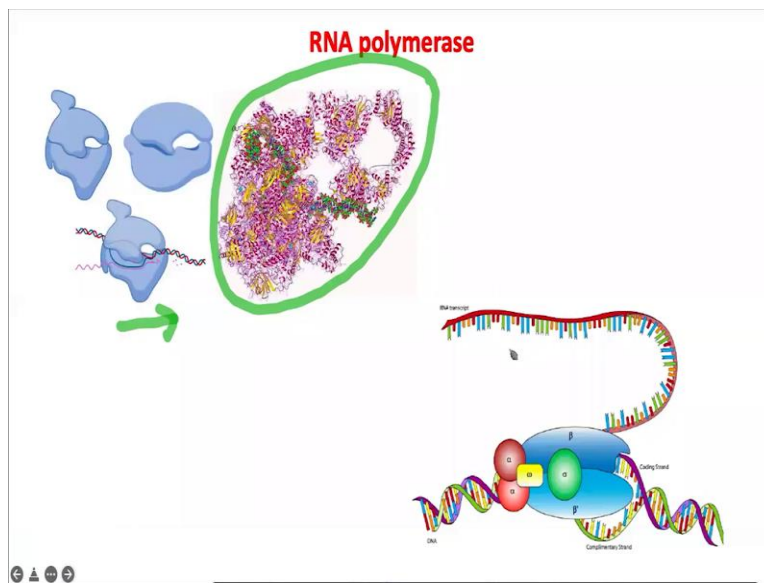
Piece of mRNA is formed by the enzyme RNA polymerase by adding complementary RNA nucleotides to the template strand of the DNA. Notice that only one strand of the DNA is used to do this and the other strand is not used at all. Furthermore, the piece of the mRNA formed is single stranded. In this animation I have shown you the section of DNA to be transcribed the gene. But remember that DNA contains many genes and therefore the piece of mRNA formed will ultimately be shorter than the complete length of the DNA.

Finally, notice that the DNA is using thymine as one of its nitrogenous bases and mRNA uses uracil instead. Once this mRNA has been formed it moves out of the nucleus where this transcription has taken place and into the cytoplasm where it later used in translation to form a polypeptide chain. It is been formed it moves out of the nucleus where this transcription, the later used in translation.

So, to highlight the changes again the sugar is, in RNA is, RNA is nucleic acid polymer. The sugar is going to be ribose instead of deoxyribose. That is one change. And instead of T we are using uracil, a nucleotide which contains uracil. But pretty much RNA is a cousin of DNA. It usually exists in single stranded strand but it always associates with proteins and folds pairly well. So, the word single stranded is basically textbook language.

And what you could clearly see is how a piece of RNA was made from DNA. And in your mind you should compare and contrast the two process, DNA replication done by DNA dependent DNA polymerase and RNA replication which is done by DNA dependent RNA polymerase.

(Refer Slide Time: 15:24)



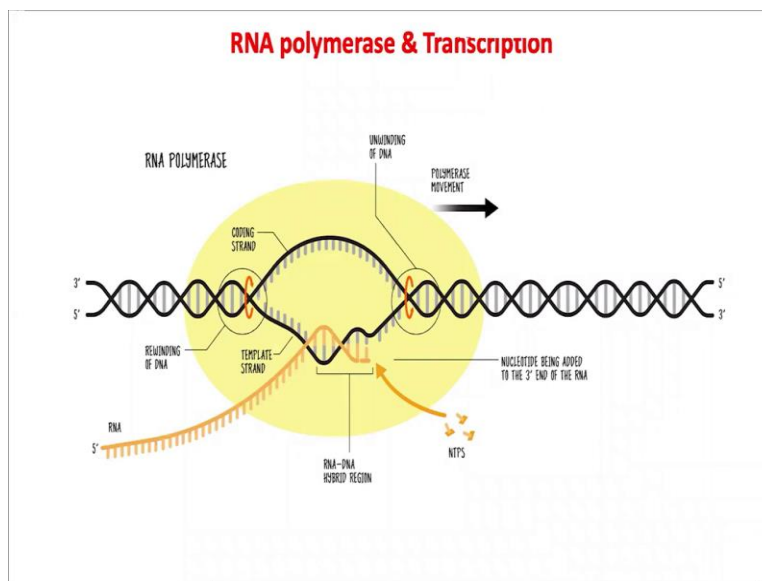
So, let us now just look at the machine is doing all this work and this machine is RNA polymerase. Now, RNA polymerase, this is a schematic on the left, pretty much like two hands club together. And you can see in this schematic there are gaps in the protein. This is a protein enzyme. And you can see DNA actually going into this enzyme being opened up in a small bubble, a bubble which is not as big as a transcription bubble, and you can see a piece of RNA

which is coming out, which is being copied from one of the strands which is complementary to one of the strands which is being read.

And as RNA polymerase keeps on moving forward, it will keep on moving forward on the DNA, you will get a mRNA strand which is a faithful copy of DNA and you get back the DNA helix. And this is pretty much what the crystal structure, atomic structure of DNA polymerase looks like and shown over here are the different domains, the different pieces of this enzyme. So it is basically multiple pieces of protein put together and this is something we will discuss later.

You can see it sitting on DNA. You can see this very small bubble that it opens up and you can also see the copied RNA which is called as a RNA transcript and this whole process is called as transcription.

(Refer Slide Time: 16:52)



Now, this is a sort of a picture of this entire process. In yellow is schematic of RNA polymerase. This is DNA which is opened up. This is the DNA, RNA hybrid region where you have extension of RNA bases in the 3 prime direction which is copied from this strand. Now, terminology is calling this the template strand. And the other strand over here as you can see is called as the coding strand.

And it will be of no surprise to you that this sequence over here of RNA is exactly equivalent to the sequence over here, because the sequence in brown, light brown over here is basically a

complement of the template strand and obviously the coding strand is, has the same sequence as the RNA strand except for of course the fact that you have uracils instead of thymine.

And you can see these three NTPS which are being used which are present in a pool, DNA is unbound, polymerase moves towards the right hand side. And as the polymerase makes RNA and moves ahead, the DNA is rebound back to its double helical strand. So, are there any questions at this point?

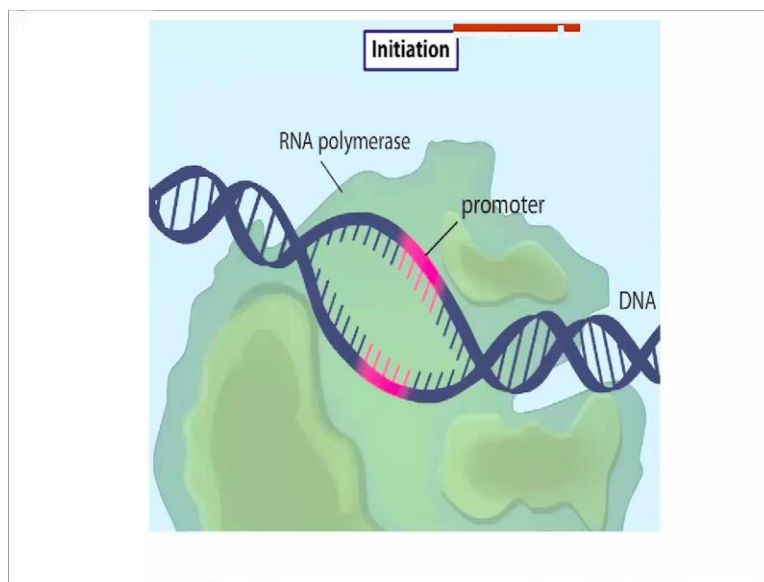
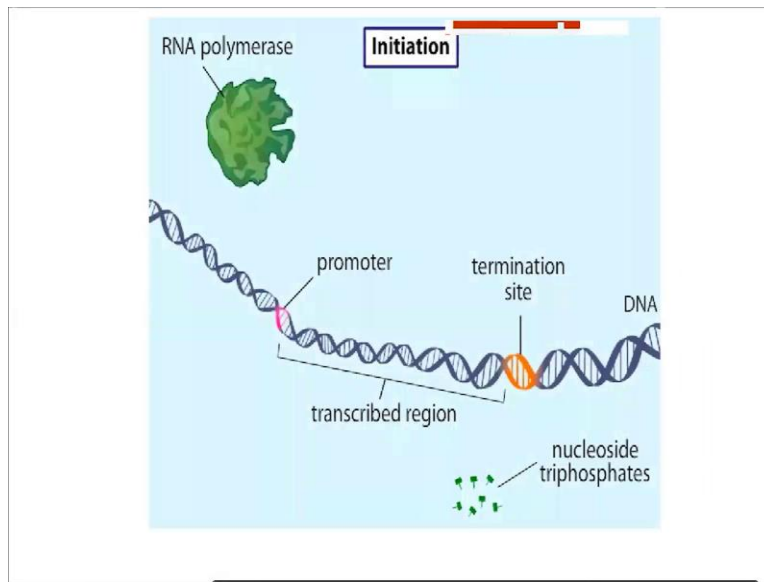
Student: Sir, like in the making of the RNA, the RNA joins to this template strand. So the RNA polymerase also acts as a helicase to remove the hydrogen bonds between them.

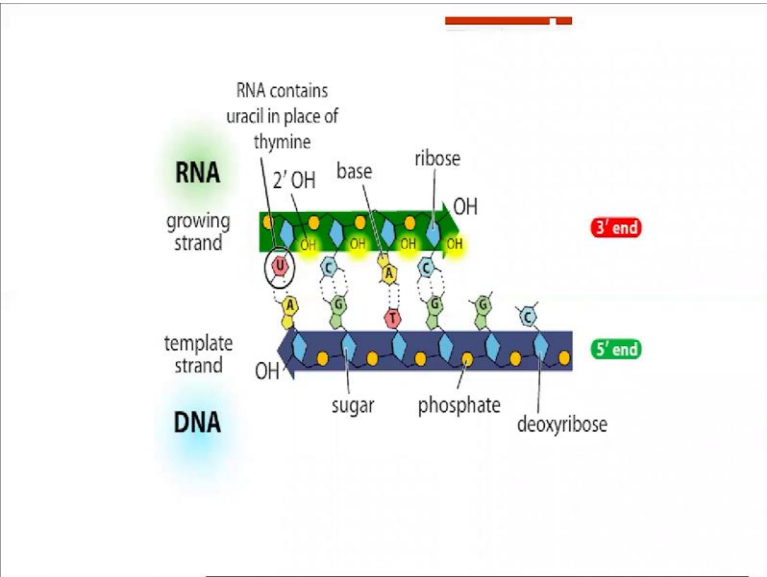
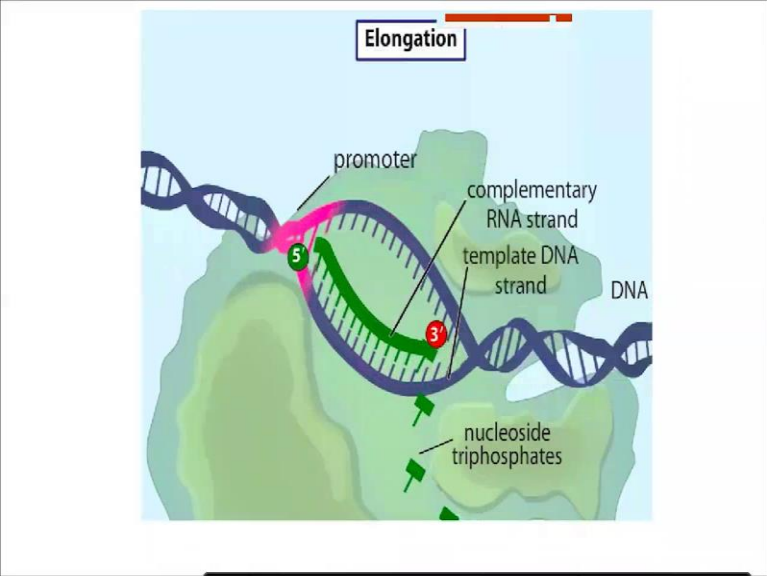
Professor: Yes. So, there are actually, I am giving you a simplified view. There are many proteins which are functional over here. There are actually about somewhere between 20 to 100 proteins which are working over here. RNA polymerase is the machine doing this particular job, but it is assisted by a large number of proteins. So, there is what is called as a transcriptional initiation complex over here. So, different proteins are different doing jobs including opening and closing of DNA. But yes RNA polymerase is also involved in this process.

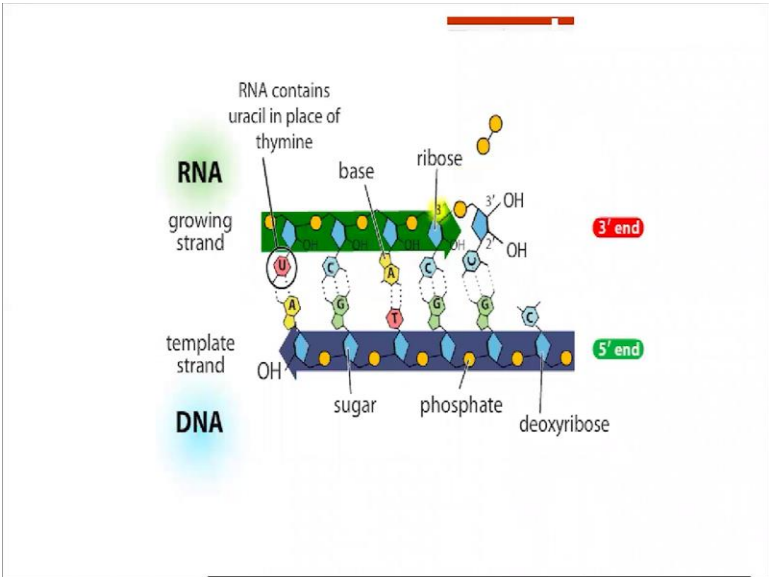
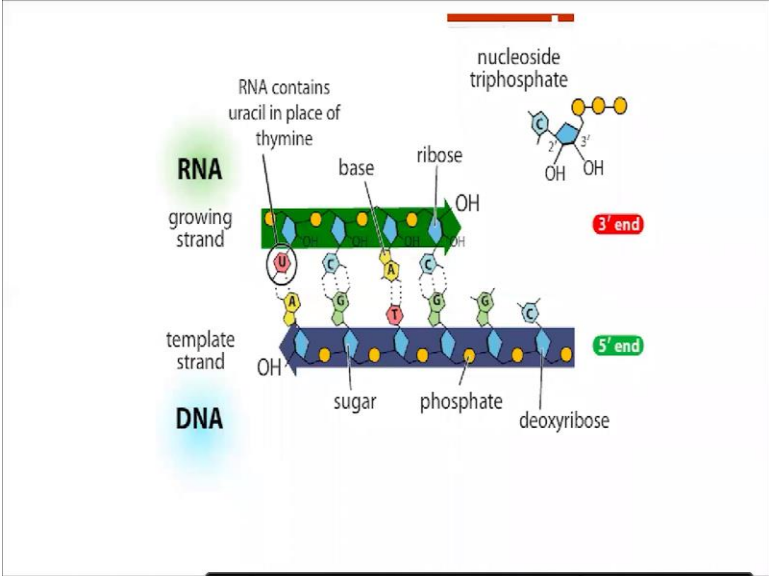
So, even though I am telling you this as a one protein doing one job, the primary job of of this protein, but do realize that there are a series of associated proteins which are also helping this whole process above. So, transcription is not about only about RNA polymerase. Transcription is about a group of proteins with RNA polymerase doing the key job in this process. So, let us look at this second movie.

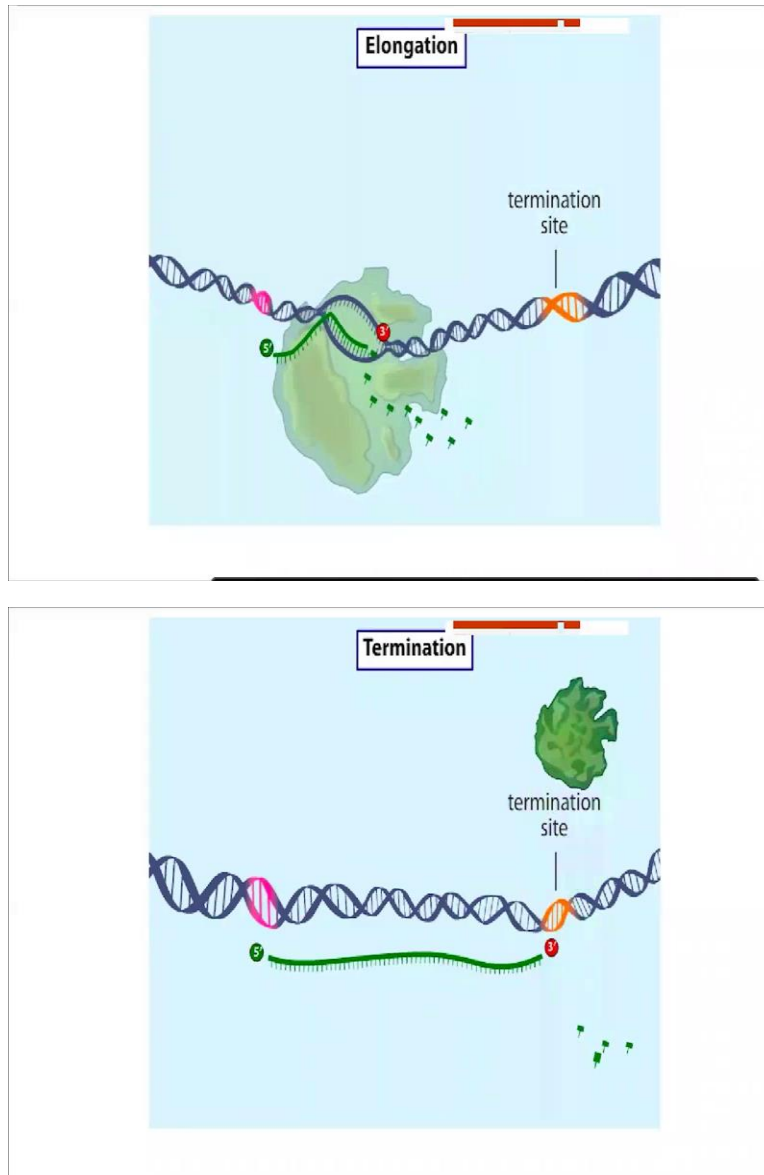


(Refer Slide Time: 19:30)









Transcription is the first step in the expression of a gene. It is the formation of a specific RNA sequence from a specific DNA sequence. So, I want you to now focus on the fact that this is a stretch of DNA. There is an area defined as a promoter. There is an area defined as a termination site. So, pretty much your gene is now between these two areas. And let us see how polymerase goes and makes a copy of RNA, of DNA.

Transcription initiation occurs at a promoter, a region on the DNA, where RNA polymerase binds. RNA polymerase attaches to the promoter and begins to unwind the DNA. At the initiation site, the polymerase begins reading the DNA template strand and building a complementary RNA strand from three nucleoside triphosphates. The RNA strand grows by the

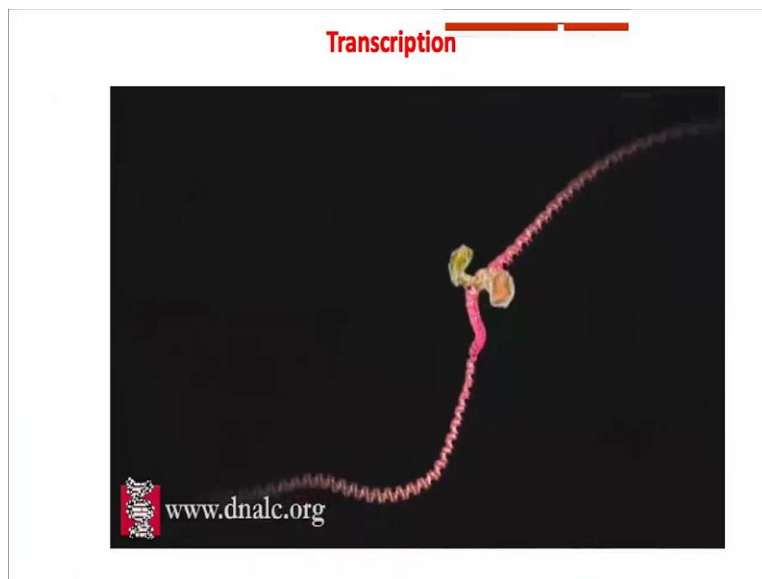
addition of these substrate molecules to its 3 prime end. This is the elongation phase of transcription.

Note that the bases in the nucleotide to the RNA strand form pairs with the bases in the DNA strand. Also, note that the RNA nucleotides contain the sugar ribose which has a 2 prime hydroxyl group, whereas the DNA nucleotides contain the sugar deoxyribose which lacks this hydroxyl group. RNA polymerase functions similarly to DNA polymerase. It adds nucleotides to the 3 prime hydroxyl group of the last nucleotide.

The DNA double helix rewinds as the RNA polymerase moves through. Like the unwinding, the rewinding is an energy requiring process that is accomplished by RNA polymerase. When RNA polymerase reaches the termination side, the RNA transcript is released from the template. The DNA rewinds completely and the RNA polymerase dissociates from it. The DNA and RNA polymerase can then participate in other rounds of transcription.

So, hopefully this was fairly straightforward and clear. It introduced you to the concept of promoter sites, termination sites. Let us now look at a more complicated movie.

(Refer Slide Time: 21:44)



## Transcription



## Transcription



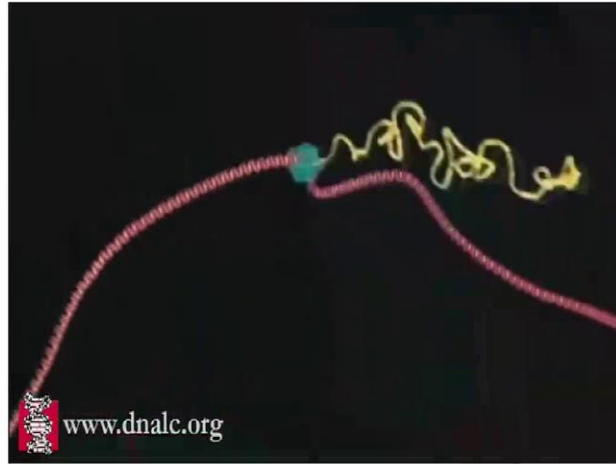
## Transcription



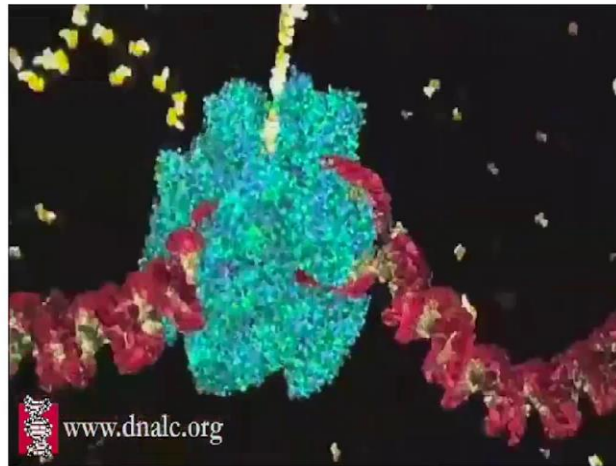
## Transcription



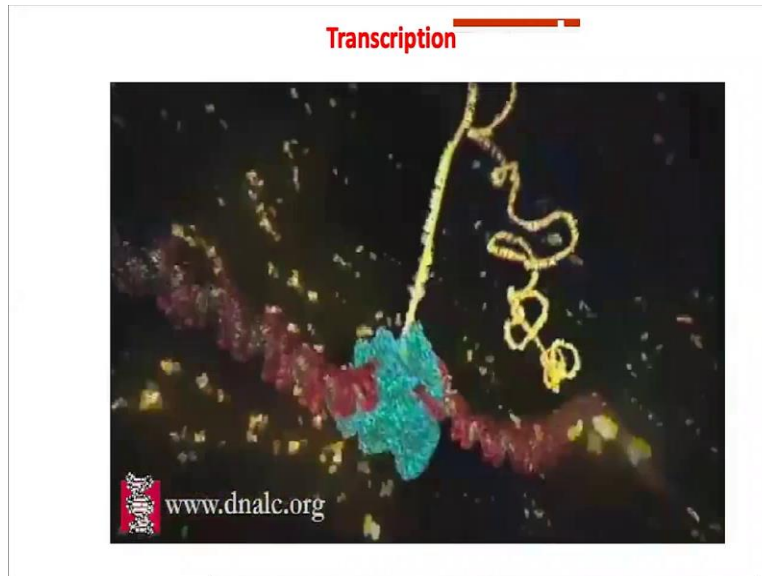
## Transcription



## Transcription





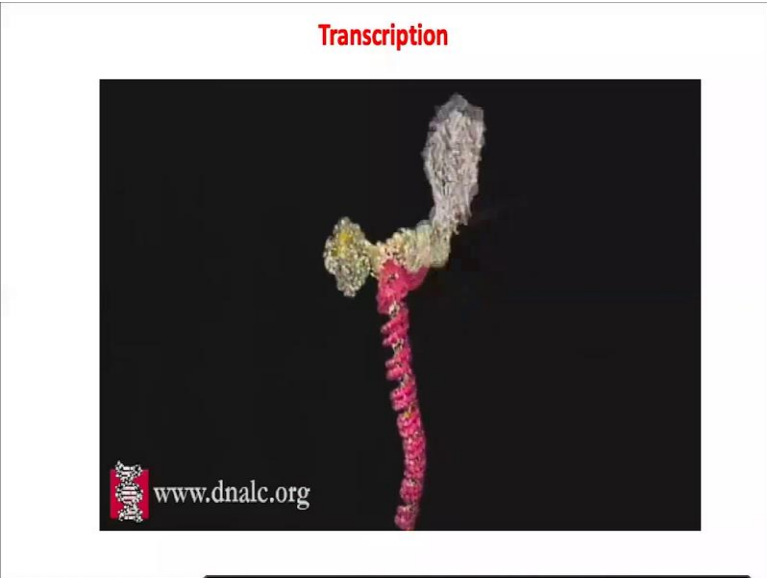
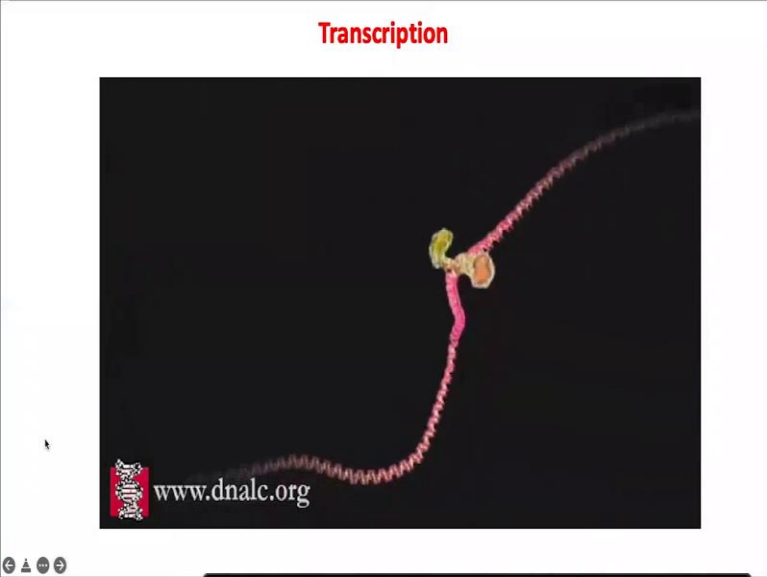


What you are about to see is DNA's most extraordinary secret how a simple code is turned into flesh and blood. It begins with a bundle of factors assembling at the start of a gene. A gene is simply a length of DNA instruction stretching away to the left. The assembled factors trigger the first phase of the process, reading of the information that will be needed to make the protein.

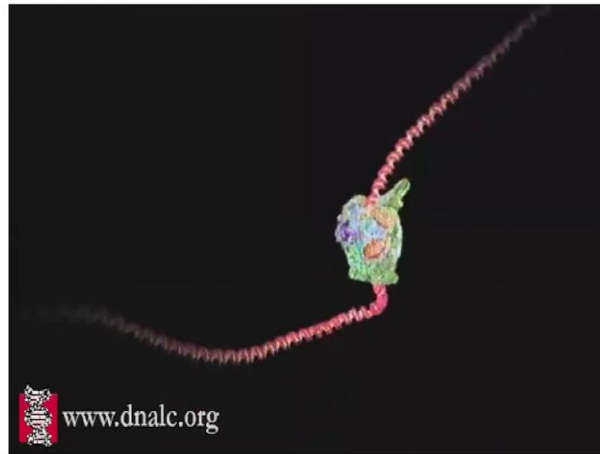
Everything is ready to roll, three, two, one, go. The blue molecule racing along the DNA is reading the gene. It is unzipping the double helix and copying one of the two strands. The yellow chain snacking out of the top is a copy of the genetic message and it is made of a close chemical cousin of DNA called RNA.

The building blocks to make the RNA entered through an intake hole. They are matched to the DNA letter by letter to copy the As, Cs, Ts and Gs of the gene. The only difference is that in the RNA copy, the letter T is replaced with a closely related building block known as U. You are watching this process called transcription in real time. It is happening right now in almost every cell in your body.

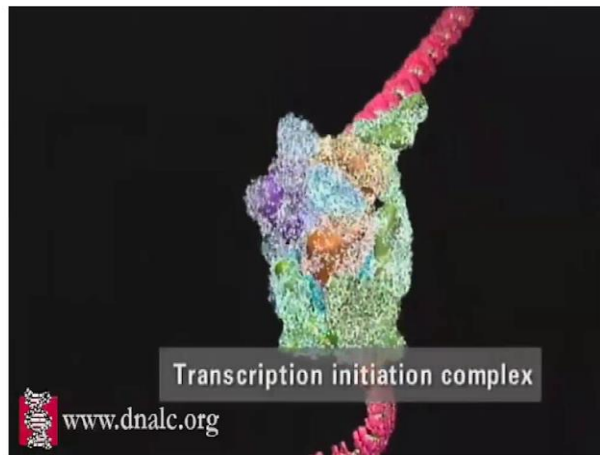
(Refer Slide Time: 23:37)



## Transcription



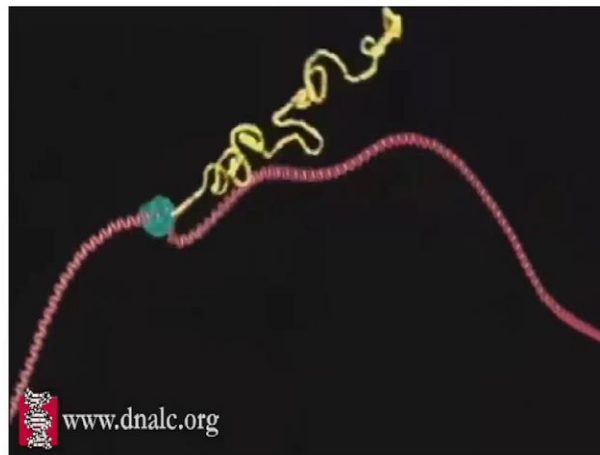
## Transcription



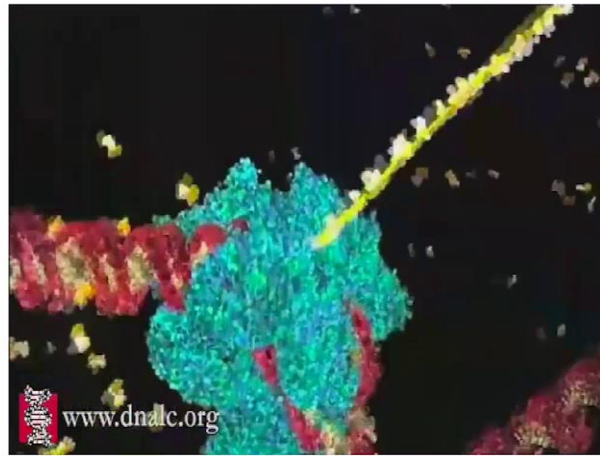
## Transcription



## Transcription

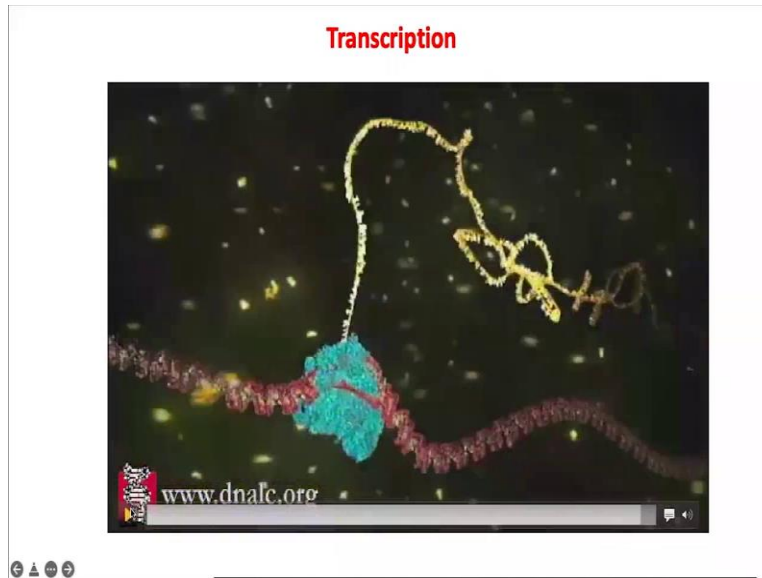


## Transcription



## Transcription





Another movie slightly different from the first one. The central dogma of molecular biology, the central dogma of molecular biology, DNA makes RNA makes protein. Here the process begins. Transcription factors assemble at a specific promoter region along the DNA. The length of DNA following the promoter is the gene and it contains the recipe for a protein.

A mediator protein complex arise carrying the enzyme RNA polymerase. It maneuvers the RNA polymerase into place, inserting it with the help of other factors between the strands of the DNA double helix. The assembled collection of all these factors is referred to as the transcription initiation complex and now it is ready to be activated.

So, in this part of the movie the sort of, if this is a race, the person who is firing the gun appears to be coming from a slightly different site. You have already seen a large number of proteins including transcription factors, the mediator complex and of course RNA polymerase all assembling over here. But the way this is described is everything is ready, but the race cannot start unless it is triggered.

And the trigger, the way they have shown it in this movie is coming from a protein which is a little far away which is actually bound to a stretch of DNA called as the enhancer DNA. Now, you will see this enhancer protein sitting on DNA which is called as enhancer DNA coming and triggering this whole process.

The initiation complex requires contact with activator proteins which bind to specific sequences of DNA known as enhancer regions. These regions may be thousands of base pairs distance from the start of the gene. Contact between the activator proteins and the initiation complex releases the copying mechanism.

The RNA polymerase unzips a small portion of the DNA helix exposing the bases on each strand. Only one of the strands is copied. It acts as a template for the synthesis of an RNA molecule which is assembled one subunit at a time by matching the DNA letter code on the template strand. The subunits can be seen here entering the enzyme through its intake hole and they are joined together to form the long messenger RNA chain snaking out of the top.

Now, remember that what the polymerase is doing is everything starts with the polymerase binding to the promoter there is a initiation site, then copying starts, and there is a termination site where copying stops. Now, the size of genes vary dramatically as I said from a few hundred base pairs to tens and thousands to 100,000 base pairs. And the polymerase whether it is a short gene or a long gene its job is to stay on the DNA and keep on making RNA till it hits a sequence which tells it that it is a termination step in making RNA. So, it hits the termination piece of DNA at this point.