

**Introduction to Complex Biological Systems**  
**Professor Dibyendu Samanta and Professor Soumya De**  
**Department of Bioscience and Biotechnology**  
**Indian Institute of Technology, Kharagpur**  
**Lecture 6**  
**Gene expression and the Central Dogma of Molecular Biology**

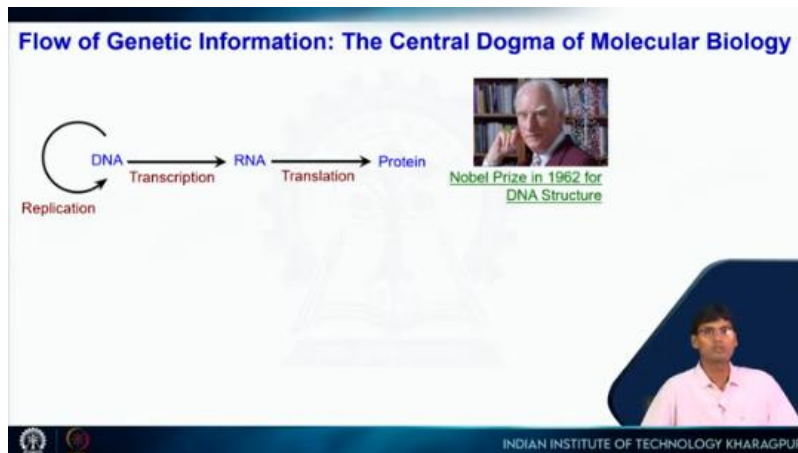
Hello everyone. So, we are entering the next module, module 2, where I will be discussing transcription and translation, particularly the making of products from the stored information in DNA. During the last module, module 1, I mostly concentrated on nucleic acids, particularly DNA, and we discussed how scientists proved that DNA is our genetic material, followed by how Watson and Crick's DNA structure helped us understand how information is present in DNA and how it can be replicated. This is a very important step in biology, as we have to keep our information, particularly genetic information, stored in DNA, and it needs to be copied precisely without errors for cell division, for the next generation, for offspring, for everything.

So, now let us start with module 2, and today I will be discussing gene expression and the central dogma of molecular biology. So, here I will particularly focus on understanding the central dogma of molecular biology, followed by a brief introduction to gene expression, and I will also discuss the correlation between DNA content and its downstream product. So, as I already mentioned, DNA is nothing but information. So, DNA alone cannot work. So, whatever information is present in DNA should be translated to make some product that will be functional in our body.

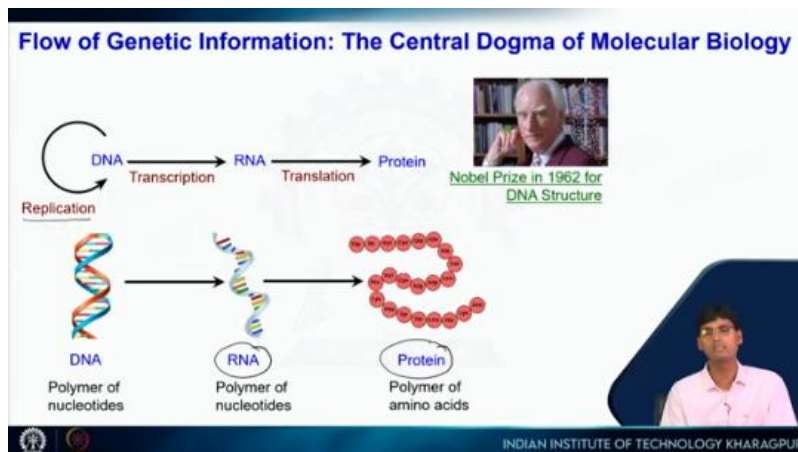


So, those products I am referring to here are downstream products, which could be proteins. So, now let us see the flow of genetic information, which is also known as the

central dogma of molecular biology. As I already mentioned before, DNA goes to DNA, which is called replication, and then here I will mostly discuss the next two steps, which are transcription and translation, in this module. And then, as you can see, after transcription, you will have RNA, which is also a polymer of nucleotides, but it is a polymer of ribonucleotides.



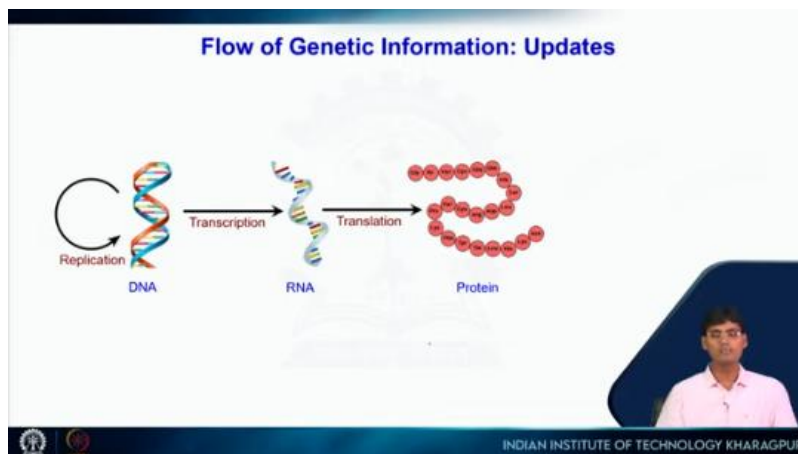
And finally, we will get protein after translation. So, this is a polymer of amino acids and proteins that actually do the work for us. Now, in terms of discovery, I would like to mention here that DNA and protein were understood by scientists much before the discovery of RNA. How DNA works, how it replicates, and also how scientists purified proteins, and people knew that proteins, particularly enzymes, have specific functions. But RNA came a little later.



So, later scientists understood that RNA actually connects the DNA and protein together. So, RNA is the intermediate step; without RNA, we will not have the protein. So, from DNA, we have to make RNA, and finally, from RNA, we will get protein molecules

which are functional in our body. Now, as I already told you, in this module, we will mostly focus on this part, which is, you know, I just put some boxes here, the transcription and translation and different aspects of these two steps. Those are very fundamental steps in molecular biology, I would say, in biology itself.

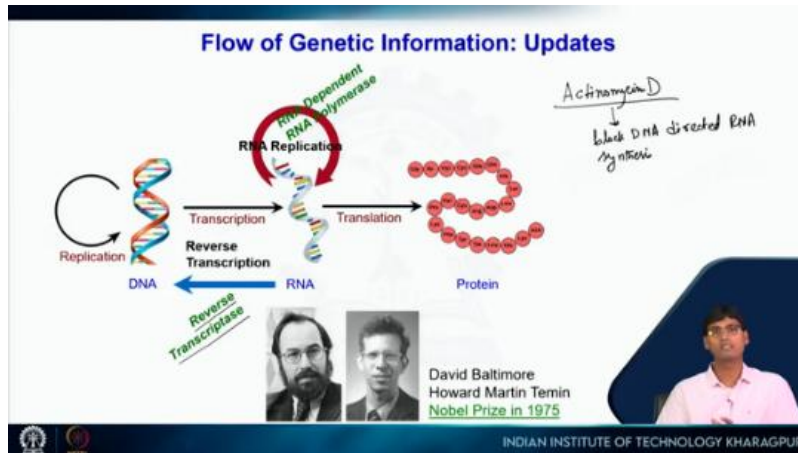
The term 'central dogma' is, you know, proposed by Francis Crick. So, here, like I already mentioned, Francis Crick and James Watson received the Nobel Prize for the DNA structure, but also they proposed the idea of the central dogma. So, central dogma means, here, you know, generally we say this: DNA goes to DNA, that is replication, then transcription, and then translation, that is all. But now, if we go into a little more detail, we will get some updates in this flow of genetic information. So, as you can see here, in the 1970s, scientists were working with different RNA viruses.



So, although I mentioned that DNA is our genetic material, sometimes RNA can be the genetic material of some viruses. And scientists who, you know, were working with RNA viruses found that they have some specific enzyme called reverse transcriptase. So, from the name itself, it is understandable that this enzyme will make DNA from RNA; this is just the reverse reaction of transcription. So, RNA to DNA synthesis, and because of this enzyme, those viruses can reverse transcribe their RNA into DNA, and because of this work David Baltimore and Howard Martin Temin

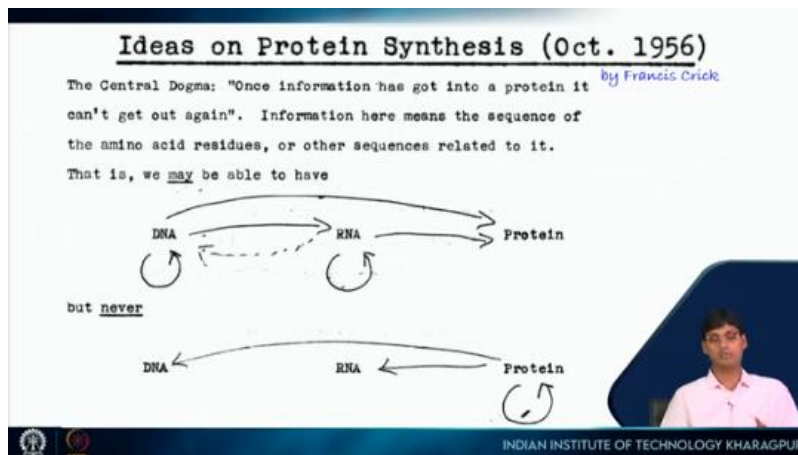
received the Nobel Prize in 1975. This is very interesting work and as many of you know, even the HIV virus, for example. So, the HIV virus has RNA as its genetic material, and it also has reverse transcription, which helps in this reverse transcription. If you see another group of scientists, who were working with some RNA viruses, particularly polio virus and some other viruses. They found that those viruses are not sensitive to one drug.

So, the name of this drug is Actinomycin D. So, this drug blocks transcription, particularly DNA directed RNA synthesis. So, if DNA is the template and RNA is being synthesized, this drug will block DNA directed RNA synthesis. So, now when they found that these viruses are not sensitive to this drug that means some other mechanism exists that might be DNA to RNA synthesis is not happening. And they found that



they have another enzyme called RNA-dependent RNA polymerase. So, it can synthesize RNA out of RNA, RNA-dependent RNA polymerase. So, as a result, RNA replication can also happen. So, these are some additions to whatever the normal thing I am saying again and again that DNA goes to DNA that is replication, transcription, and translation. Now because of this, many people, even many scientists, were a little bit skeptical about the idea of the central dogma. Central dogma means dogmatic, which means it's a kind of belief that should be true, so that is the central belief of molecular biology.

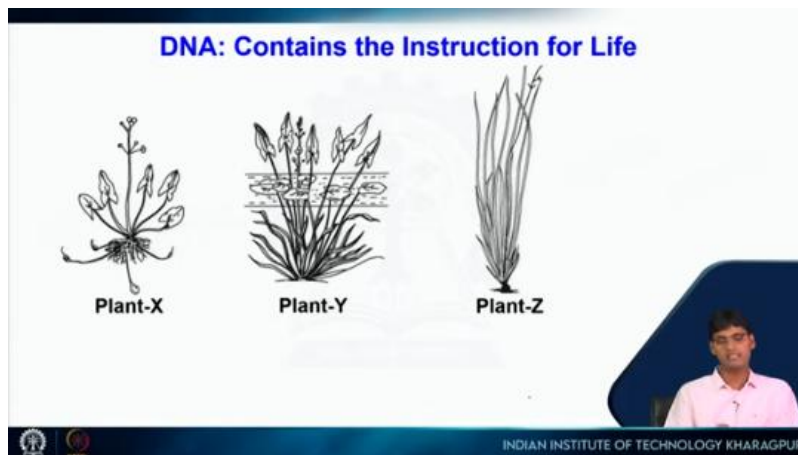
So, that is why people sometimes say that no, the central dogma is wrong, some updates are there, something is happening here, but if you see what Crick actually proposed, that is not exactly what we are telling here. So, this is what Crick actually proposed. He proposed ideas on protein synthesis in a light-hearted manner. He termed this the central dogma. You can see whatever he mentioned here, this is very important like once information has got into a protein, it cannot get out again. This is the major thing.



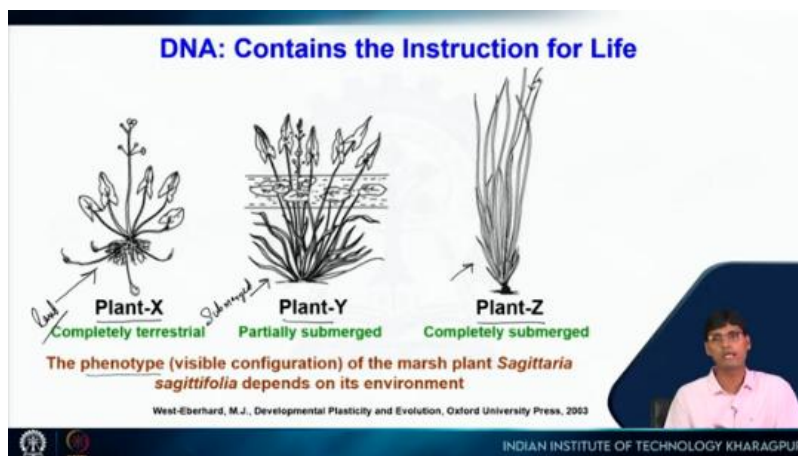
This is the central dogma. So, information here means amino acid residue. And generally, in textbooks, we do not see this figure, but I would like to mention whatever this update I mentioned that RNA to RNA synthesis or RNA to DNA synthesis, which were discovered much later compared to this proposition made by Francis Crick. If you see, Francis Crick also mentioned all these possibilities. Here, as you can see, he mentioned that RNA to RNA synthesis, like here, RNA replication can happen.

Similarly, he also mentioned here that RNA can be converted into DNA but never protein to DNA synthesis or protein to RNA synthesis and protein to protein synthesis, which is not possible. So, this is the central dogma, and in my understanding, whatever he mentioned about protein to protein, protein to RNA, protein to DNA, it should not happen. This is true also today, like we cannot have protein being formed from protein. So, in all these aspects, the central dogma, whatever its purpose, is very much what we are saying now. Now, here, as I just discussed the central dogma, particularly how the information flows from gene to functional molecule, but here, I want to discuss a slightly different aspect.

Here you can see three different plants; I named them plant X, plant Y, and plant Z. As you can see, all these three plants look different, but I want to mention here that all these three plants have the same DNA because they are the same plant. They look different because of their habitat; this plant grows on land. So, this is a terrestrial plant, and this is submerged; it is slightly inside water and outside water. So, this is submerged, and this is completely inside the water. So, that makes this huge difference as you can see that it is completely terrestrial, partially submerged and completely submerged.

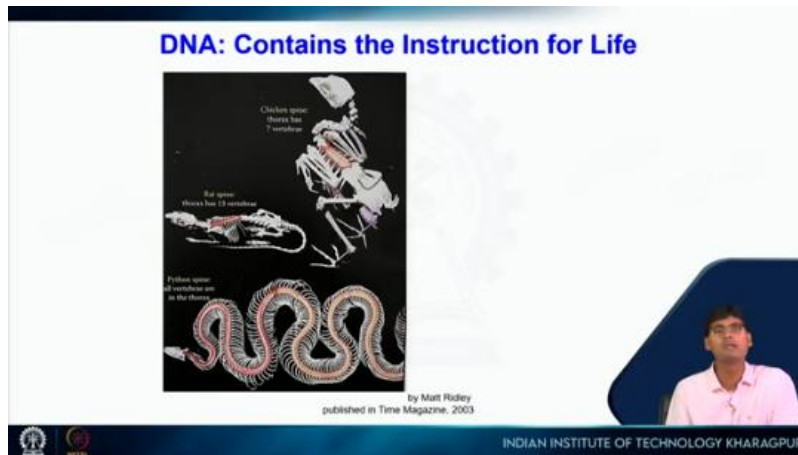


So, what I want to say here is, the phenotype, the visible configuration of an organism, the visible configuration of this plant, a marsh plant scientifically named *Sagittaria sagittifolia*, depends on its environment. So, it depends on its environment. So, although this is the same plant, they have the same information in their DNA, but they look completely different. So, this is controlled even by the environment. So, this is very interesting that means the environment can also somehow regulate the expression of genes.

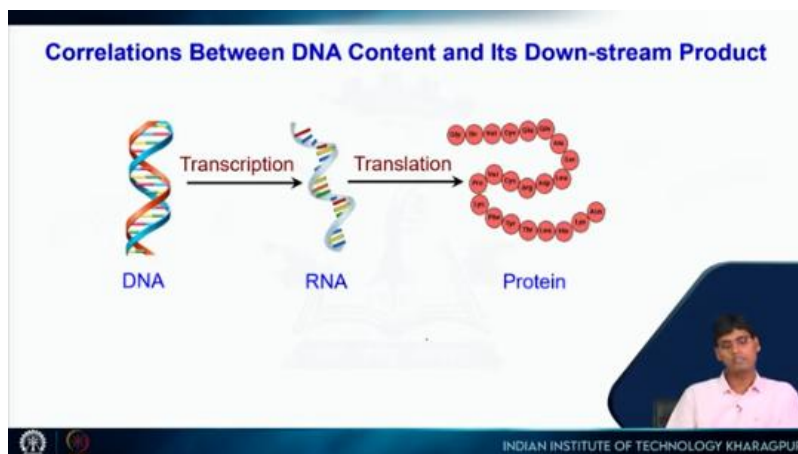


So, that these three plants look completely different. Now, in the same line, I would like to mention this, here you can see the skeletons of three different animals. So, here the first one is a chicken, here is a rat, and here is a snake or a python particularly and here, one gene *HoxC8*.



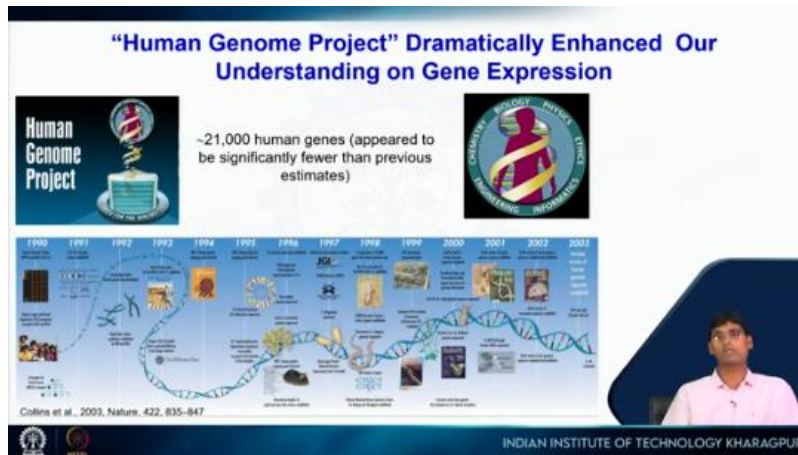


So, this gene is present in all these three organisms, and this Hox C8 is responsible for the formation of vertebrae in the thorax region. So, as you can see, this gene is present in all these three organisms, and they are very closely related. If we see the sequence identity, which is also very close because these three animals are not very far apart in terms of their evolution, right? So, python, rat, chicken, but still, their skeletal structure, particularly the number of vertebrae, differs a lot, as we can see in the case of the python, where everything is about vertebrae and the thorax. Because this gene, how it is getting expressed during the early developmental stage, is too active in the case of the python, and it expresses a lot, determining the number of vertebrae in the thorax region.



So, as a result of that, what I want to say is that although the information is present, how this information is getting expressed, how we can read this information, because of that, the trait, the phenotype can be very different between organisms. Now, here, I want to mention the correlation between DNA content and its downstream product. As I told in that slide where I was showing that three plants look different, but they have the same DNA, which means there should be some kind of correlation between DNA and its

downstream product, and that is why the phenotype is different. So, here, the Human Genome Project, which is one of the biggest projects in biology, and it dramatically enhanced our understanding of gene expression.

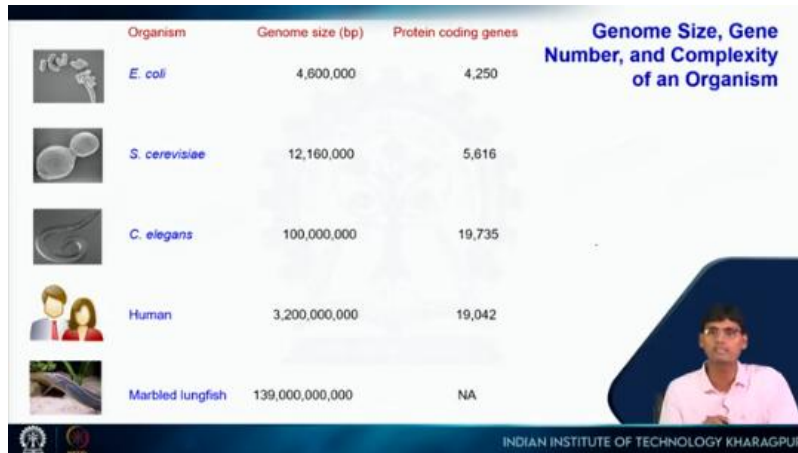


So, this project, the Human Genome Project, was conceived and started in 1990, particularly in the United States of America, who took this initiative, but many other countries like Germany, France, and UK also joined in this collaborative project. And this is a government funded project, and in 2003, they actually completed this project, and we came to know the total information present in the DNA of humans. This is very important, because of that, it revolutionized the biomedical field, so for medicine, treatment, genetic diseases, and many things, it completely revolutionized the field of biomedical science. But what I want to mention here is that because of the Human Genome Project, we came to know that we have approximately 21,000 genes, which are responsible for making different types of products and it appears to be significantly fewer than previous estimates, before that we used to think we have around 80000 ah 1 lakh genes and but after this human genome project the number is very much low compared to the previous estimate. So, it actually helped us to understand that how this information is processed, how many genes are present, their specific location in particular chromosomes and all sorts of information. And during this project, scientists decided to sequence the whole genome of a set of organisms, not only humans, because the size of the human genome, the amount of nucleotides present in the whole human genome, is really high. So, it is better that they should try with different organisms also and we have to progress slowly, as you can see that it was started in 1990 and was completed in 2003.

The full sequence was confirmed and mapped in individual chromosomes or DNA in the human genome. Now, here, because of all this information coming out from the Human



Genome Project, we can see, particularly, I am showing here a few organisms for example, *E. coli*, which is a bacteria and a prokaryotic cell. And their genome size, so genome size means the total amount of DNA in terms of nucleotides, in terms of base pairs, how much DNA is present. So, this is the total number of base pairs of nucleotides present in *E. coli*. And here, this number is the number of protein-coding genes.



Organism	Genome size (bp)	Protein coding genes
<i>E. coli</i>	4,600,000	4,250
<i>S. cerevisiae</i>	12,160,000	5,616
<i>C. elegans</i>	100,000,000	19,735
Human	3,200,000,000	19,042
Marbled lungfish	139,000,000,000	NA

So, what I want to mention here is that initially, we used to think that complex organisms like humans, we have many cells approximately  $10^{13}$  cells in our body in an adult human, approximately. And we have a very complex system, complex organs like we have eyes, ears, kidneys, and so many complexities, the brain. So, that means we must have many genes and much information, and that is why we are very complex. So, this is kind of a general idea that the complexity of an organism can be related to its genomic size also, but later on, it has been established, as I am showing here, you can see here.

For example, this is *E. coli*, whatever I mentioned, this is yeast, some kind of fungus, and this is the genome size and the protein-coding genes here and interestingly here, this is *C. elegans*. So, this is some kind of worm, a very small worm, and this is the genome size and the number of protein-coding genes present in *C. elegans*. So, here I want to mention that *C. elegans*, it is a very small worm, it has approximately around 2000 to 3000 total cells in its body, only 2000 something like that in that range. But humans, I told you that approximately  $10^{13}$  cells are present here.

And also, as you can see, *C. elegans* is very simple compared to humans in terms of their architecture and organization, body plan, etc. But as you can see here, the number of protein-coding genes is almost equivalent in *C. elegans* and humans. Literally, I would say, in the case of *C. elegans*, the number is a little higher than in humans. But then why

is it happening? So, that means it suggests that the number of genes is not directly proportional to the complexity of an organism.

That means in humans, maybe the number of genes, the protein-coding genes, is less compared to *C. elegans*. But the regulation, the crosstalk between the products of multiple genes, is more complicated, and that is actually deciding different types of functions, different types of morphology, organ development and all sorts of things. So, the regulation is very important, the expression of genes and the regulation thereof. So, that is why from here we would like to mention that the number of genes or just the content, how many nucleotides are present, is not directly dictating the configuration of a particular organism. So, as you can see here, the marbled lungfish is one primitive fish.

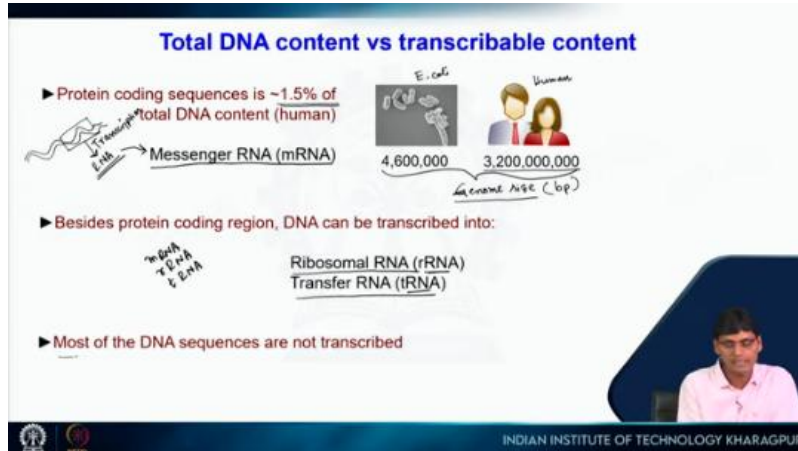
So, if you see, the genome size is much bigger than that of humans or *C. elegans*, whatever you compare. But as you can see, this is not that evolved, right? So, as a result of that, how genes are getting expressed is very important. That is actually dictating many things in our life. So, here I would like to mention now about the total DNA content versus transcribable content.

Why? Because, as I just mentioned, in the marbled lungfish, although the DNA content is very high, maybe the products are not that many or the regulation is not that complex. So, that is why here we are trying to address this issue. So, as you can see here, I just mentioned that this is the genome size, genome size in terms of base pairs of nucleotides, and as you can see, this is *E. coli* and here human.

Now, the protein-coding sequences in humans is approximately 1.5 percent, that means although we have a huge amount of nucleotides, here as you can see, 3.2 billion nucleotides present. but out of those, only 1.5 percent of DNA can be transcribed into RNA and finally, which will be forming some protein or which will help to make the protein, that is all. So, now, the protein-coding sequence of DNA. So, if I say this is one DNA, double-stranded DNA, and here if this portion is the protein-coding portion.

So, as a result of that, after transcription, it will make some RNA. Right? That RNA is called messenger RNA because the message is present in that RNA to make the protein so that is why it is called messenger RNA. So, messenger RNA is the direct template for protein synthesis. So, apart from messenger RNA, we have a few more RNAs which are also very important in making protein.

So, as you can see here, there are ribosomal RNA and transfer RNA. So, ribosomal RNA constitutes the ribosome, and transfer RNA also participates in the translation process, which means it helps in protein synthesis. I will discuss that in a different class. So, these are the three major groups of RNAs, the messenger RNA, ribosomal RNA, and transfer RNA. Apart from that, some other RNAs also exist, which we are not going to discuss in detail, for example, micro RNA, siRNA, all those things that help in gene expression and gene regulation, but for our discussion, this should be fine, like mRNA, tRNA, and rRNA. Now, only 1.5 percent of the human DNA is being transcribed. So, in that case, most of the DNA sequences present in our genome are not transcribed. So they are just present but are not getting transcribed. So, here particularly, I want to mention something because, during every round of replication when a cell divides, for example, we have to copy all this information.

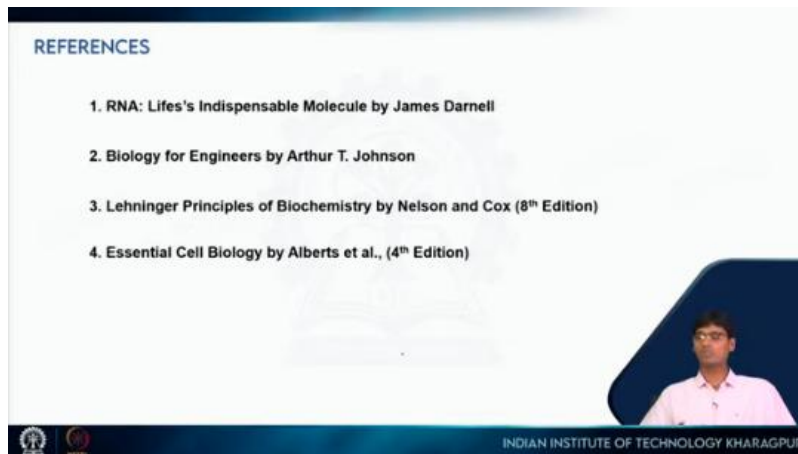


Now, during copying information, that means replication, cells need to spend a lot of energy and time, energy in terms of ATP or some kind of molecule that is required for these processes. Then why are we actually maintaining that big chunk of DNA, although those DNAs are not directly making and they are not getting transcribed and not making a product? So, actually, I would say evolutionarily it is very much helpful. So, there are many reasons, many hypotheses, logic behind this, but very simply, if I want to mention something, for example, throughout our life, for example, if I say the human lifespan is around 80 years, 90 years, something like that.

So, during that time, we are exposed to different types of chemicals, sometimes radiation for different reasons. But now, if we have only those precise things in our genome, even a little bit of a problem, it will create and show some defects in our body. But I want to mention here, just in this huge chunk of DNA, those important DNA segments are also

getting hidden. As a result, a little bit of change, like one or two mutations, might not affect this crucial information present in the DNA. And, apart from that, we have different pathways in our body, in our cells, where mutations or some little bit of damage in DNA can be rectified. But this is another reason I would say that will be helpful.

For example, in the case of bacteria, most of their DNA is transcribable, not like humans, where just around 1.5 percent is being transcribed. In bacteria, it would be around 40 to 50 percent of DNA that can be directly transcribed. But those are unicellular organisms. If some big defect happens, the cell will die, but it will not be a big problem since that cell has already divided into many, many cells. So, as a result, it is very important for eukaryotic, complex organisms like us, with very long lifespans, to maintain this information properly for many, many years. So, this is all about this lecture, and that is all.



**REFERENCES**

1. RNA: Life's Indispensable Molecule by James Darnell
2. Biology for Engineers by Arthur T. Johnson
3. Lehninger Principles of Biochemistry by Nelson and Cox (8<sup>th</sup> Edition)
4. Essential Cell Biology by Alberts et al., (4<sup>th</sup> Edition)

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR