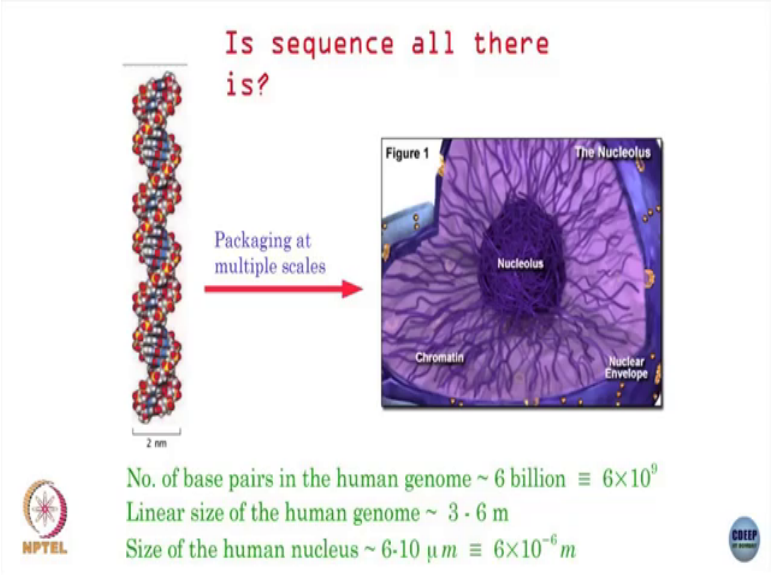
Physics of Biological Systems
Prof. Mithun Mitra
Department of Physics
Indian Institute of Technology, Bombay

Lecture – 02
DNA Packing and Structure

(Refer Slide Time: 00:19)

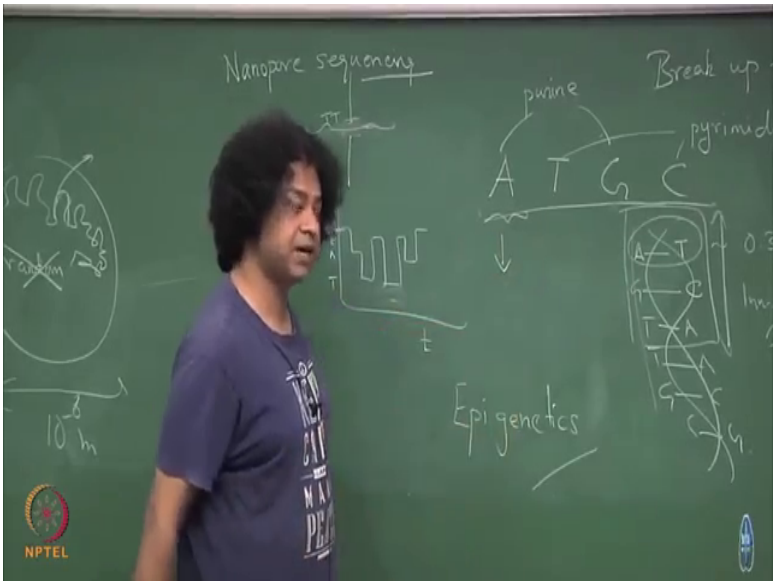


Moving on. Alright so, but is that all there is so, you know the whole idea of the human genome project was that once I know the sequence, I am sort of home. It is like knowing the fundamental laws of physics everything else is a matter of calculating things out. But all the information that is to be known is in principle encoded in the sequence.

But it turns out as it often happens whenever we think that we have cracked a problem, we have sort of we have all the ingredients that we need to understand something. It turns out that we were wrong and it is really underestimated the complexity of the problem. So, once

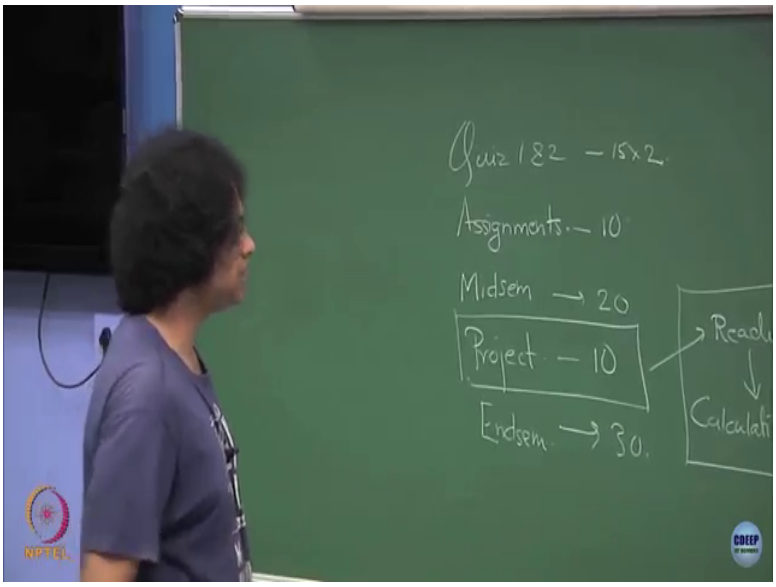
we did the sequencing, we sort of found out that genetics or the sequence information is not all there is, there is something above and beyond genetics and there is what is.

(Refer Slide Time: 01:10)



So, that is given a fancy name that is called epigenetics and explain a little bit a few as it comprises of a few different aspects.

(Refer Slide Time: 01:19)

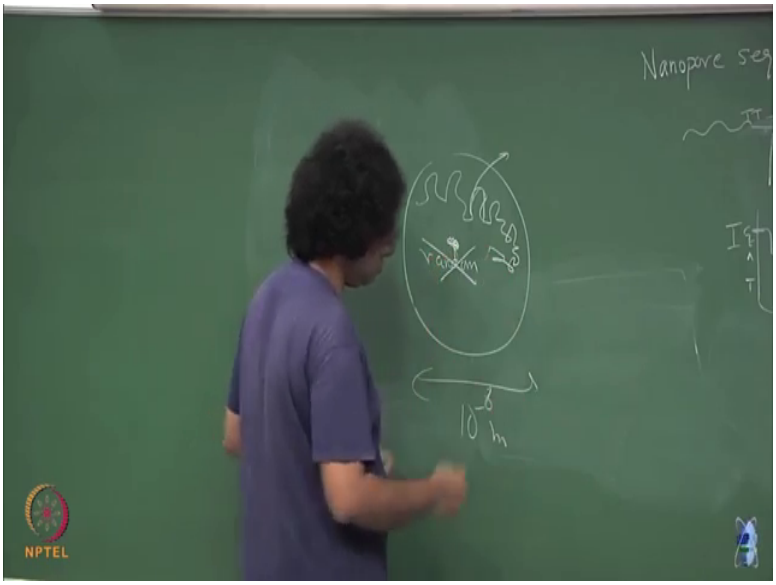


How do you control the genome basically so, it is at a level which is higher than the genome epigenome. So, I will talk a little bit about what comprises this epigenome ok. So, here is here is the thing that we were talking about right. So, if you think about the number of base pairs in the human genome, we have some 6 billion base pairs so, 6 into 10 to the power of 9 base pairs. So, if you stretched it out, it comes to order of meters which is around 3 to 6 meters and you need to package it into a nucleus which is of the order of microns; so, 6 to 10 microns so, 10 to the power of minus 6 meters. So, we take this meter long object and you are going to package it into a volume which is of the order of radius of 10 to the power of minus 6 meters.

And remember that you cannot do it sort of randomly see, I cannot take I cannot take a sphere. Let us say my nucleus is a sphere. So, here is my sphere and I need to put my DNA

into here. I can say that well I do not care, I will push it as much as I can I will push it until all of all of this 10 6 metre long string is inside this 10 to the power of minus 6.

(Refer Slide Time: 02:34)



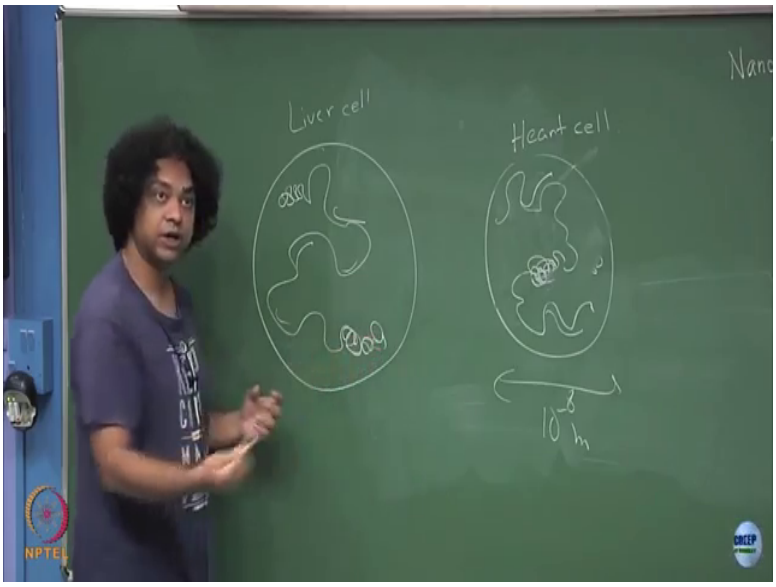
This is 10 to the power of minus 6 meter long its nucleus and I do not really care, how I am pushing it in. All I need to ensure is that this goes in and it stays in, but that does not really work for the cell because the cell needs to read information from this DNA right. Inside the DNA, there are genes and these genes code for proteins right.

So, when a cell needs to manufacture a certain protein, it needs to know that where the corresponding gene for that protein is. So, that it can go there and it can read that information and produce the appropriate protein. So, if you are packaging your DNA into million different cells in a million different ways, the cell has no way of knowing where exactly it should go in order to read a particular information particular piece of information right.

So, this arrangement it cannot be random, there needs to be some sort of there needs to be some sort of structure as to how you take this object and how you package it inside the nucleus ok. So, here is where the physics well the physical nature of this string becomes important. It is not just the information content which is the sequence, but it is also the fact that this information content is embedded in a string which is a physical object and that physical object is packaged inside a physical volume right.

So, how you do that packaging that also has implications for how for the regulation of the properties of the cell itself right. So, if the gene that you want to read is hidden somewhere here very tightly packed, then the cell might have difficulty accessing it. So, it may not be. So, for example, let me say this a little better. So, we do not know the answer to this ok. So, we do not know how the cell package is it, we know that there is some algorithm; we do not know what it is.

(Refer Slide Time: 04:19)



But let us say somehow the cell has done it and here is my here is my whole genome ok. This is how it looks; some parts are a little loose, some parts are a little tight ok. So, here it is coiled up a lot here, it is coiled up loosely. It turns out that if you take so, let us say this is some cell type. Let us say this is a heart cell all right, this is the heart cell. It turns out that you take it; if you take a different cell if you take a different cell, let us say a liver cell.

This packaging is different from a heart cell ok. So, for example, this part may be packaged very tightly and then you may have very floppy bits and then you may have a very tight bit over here again. What is the difference in these two? The difference in these two is that the proteins that are required for the heart cell to function, they would be sort of loosely packed so, that the cell can access it more easily.

The proteins that are not required for the heart cell to remember the DNA as a whole codes for all the possible proteins that your bodies can produce right, but not all proteins are going to be required by all different cell types. Some cell types will require some protein, some other cell types will require some different proteins right. So, a heart cell might require a protein which is encoded here, a protein which is encoded here. But none of the proteins which are encoded here and it turns out that similarly a liver cell might require proteins which are encoded here, here, here, but not over here.

And it turns out that depending on this on the particular type of cell that you are talking about. In fact, even depending on the time that you are talking about for the same type of cell, you can have different sorts of packaging. So, its a very dynamic sort of algorithm, it is not a fixed algorithm; it is a dynamical algorithm that dependants changes depending on the type of cell the changes depending on the state of the cell cycle, if the cell is dividing, if the cell is resting and so on and so forth and we do not know how the cell does it. So, it is a very important open question.

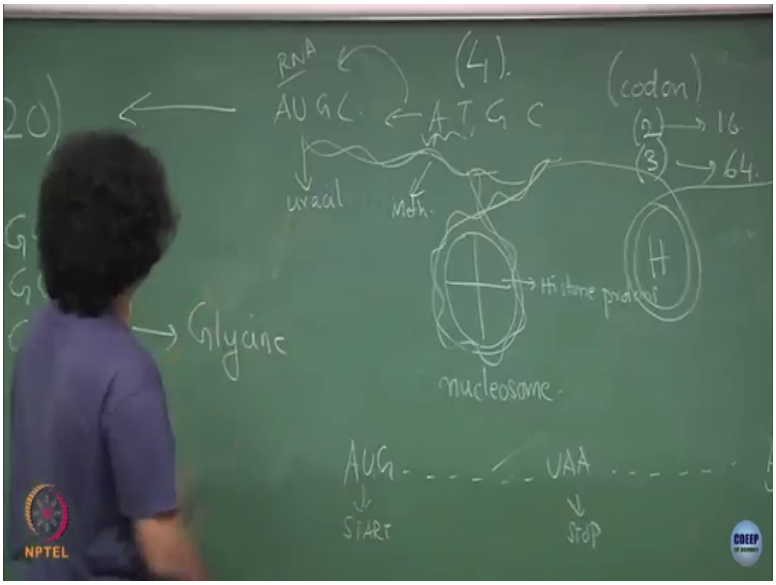
So, this is one level of epigenetic regulation which is packaging of the chromosome itself. There are other levels such as histone modifications or acetylations which I will talk about as we go on ok. So, the sequence is not all there is, there is other information. So, there is an information theoretic content which is the sequence, but there are other physical contents which is how this physical string stringy objects is itself packaged which is that I have my DNA double helix ok, that is a very bad double helix.

But whatever so, here is my DNA double helix. This double helix wraps itself around a protein complex which is called a histone complex or histone octamer. So, this contains 8 histone proteins histone proteins histone proteins and the DNA comes and sort of wraps itself around this histone protein.

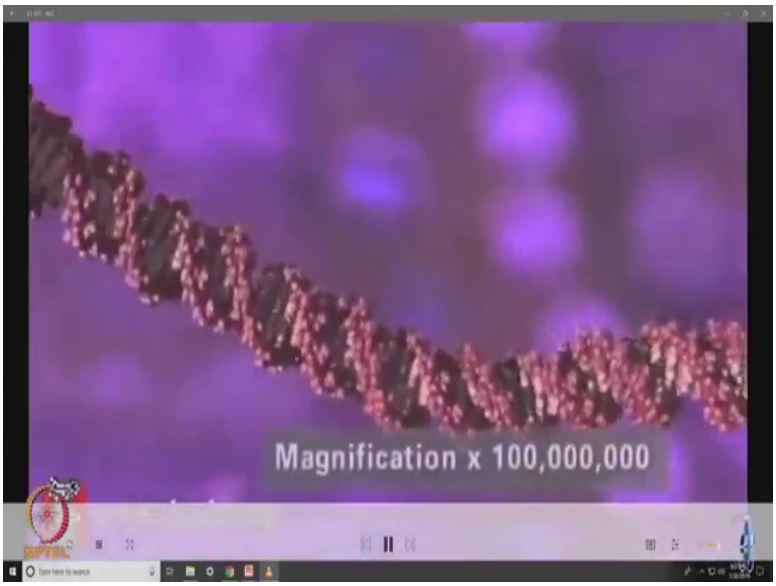
So, let me see if I can draw ok. So, it wraps itself around this histone protein to form the sort of beads on a string structure. So, if you if you zoom out and look at the DNA, you will see

that they are wrapped around these histone octamer proteins histone histone; this complex as a whole the DNA plus the histone is what is called a nucleosome is called a nucleosome.

(Refer Slide Time: 07:48)

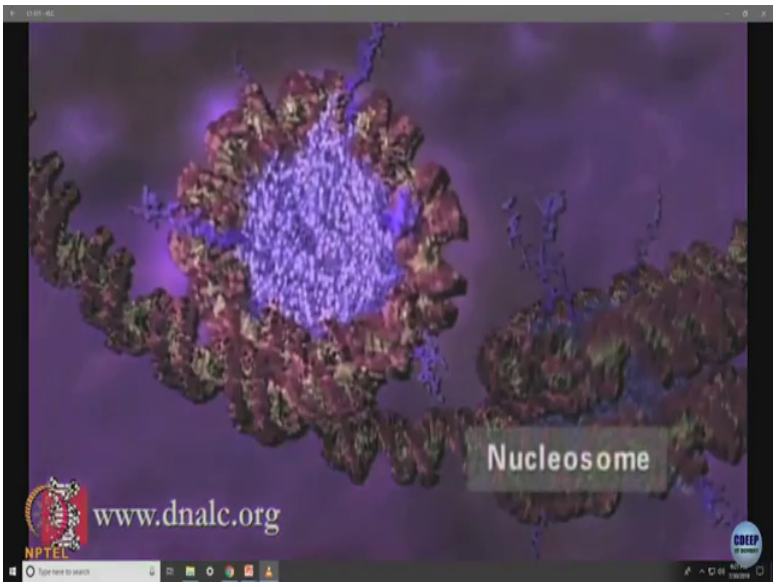


(Refer Slide Time: 08:04)



In this animation, we will see the remarkable way our DNA is tightly packed out to fit into the nucleus ok. So, here is my DNA double helix it wraps around this protein this protein complex. So, this is the histone protein complex to form these beads on a string structure ok.

(Refer Slide Time 08:24)



And this whole structure as a whole is called this nucleosome.

(Refer Slide Time 08:36)

So, this up till here is sort of what we know that after here is all conjecture there.

(Refer Slide Time 08:43)

So, the idea is that these nucleosomes sort of coil on top of each other to form a very thick fiber which is called this 30 nanometer chromatin fiber which is actually now we think is not correct, but anyway. So, from here on everything is conjected.

(Refer Slide Time 88:55)

And then this fiber somehow so, till there we at least have a conjecture beyond that we do not even have any conjecture how it exactly does.

(Refer Slide Time 09:07)

But somehow that 30 nanometer object at least that was the idea that the 30 nanometer object coils on itself to form this final folded structure of the nucleosome.

(Refer Slide Time 09:19)

And then during cell division, it organizes into this familiar sort of egg shapes that you might remember from your biology books ok.

(Refer Slide Time 09:22)

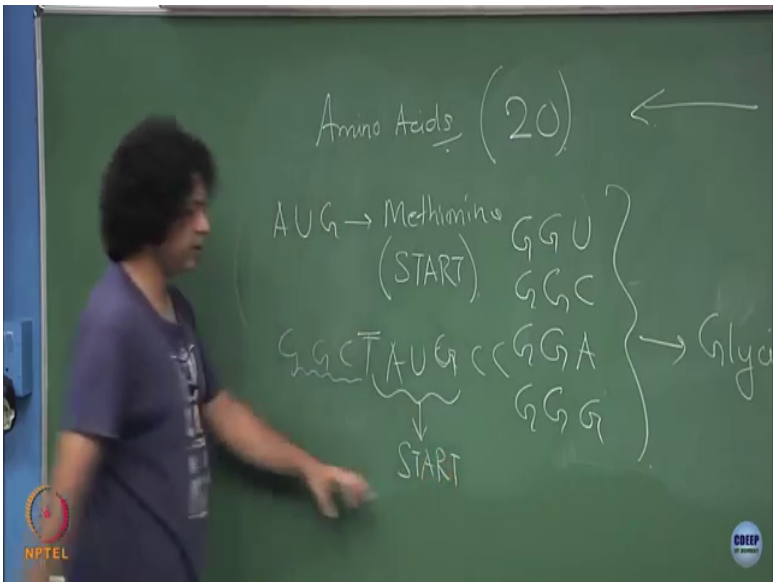
When the cell is actually not dividing, this is not how the chromosomes look. The chromosomes and then look like this sort of random soups noodle soup inside the nuclesome ok. So, this is an open question as to how this how this chromosomes actually packaged inside the nucleosome, we understand till the level of nucleosomes, but not beyond that.

(Refer Slide Time: 09:47)

So, moving on from moving on from DNA with the sequence where the DNA sequence does as we know is that it codes for proteins right. You have this whole process of transcription and translation, what is called the central dogma of biology that from DNA you get RNA; from RNA, you get proteins and then these proteins are again made up of another sort of language which is the language of amino acids.

So, just like the DNA language, the genetic code language is made up of 4 alphabets A T S G and C the protein language the protein alphabet which is the minor acid alphabet is made up of how many amino acids how many? 20. So, this is a slightly more complicated language still not as complicated as English, it only has 20 alphabets ok.

(Refer Slide Time: 10:42)



And so, this is very nice database called the protein database or the protein databank where you can actually go. So, whenever people manage to sequence a protein and by sequencing a protein, it means that finding out exactly the exactly like sequencing the chromosome finding out the sequence of amino acids that comprise a particular protein; so, tryptophan, this methionine, this that what are the exact amino acids that make up a protein.

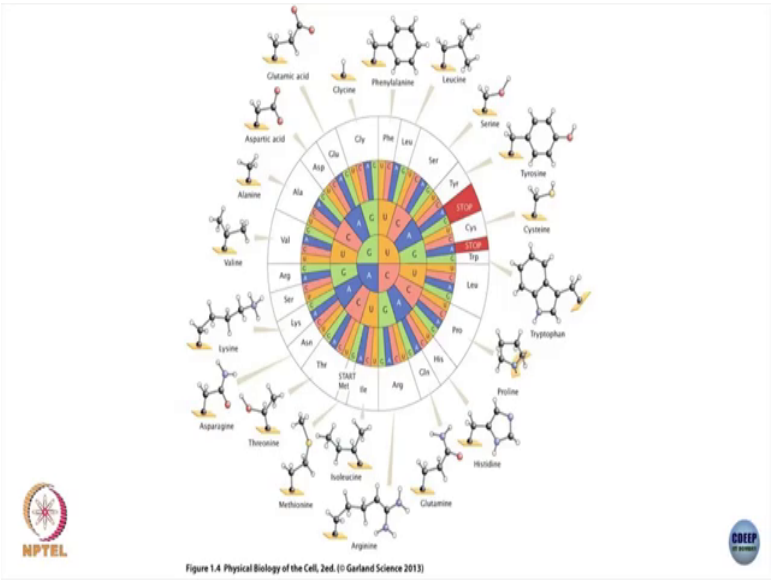
So, whenever people manage to do that and they managed to find the structure of a protein, they upload it in that protein database which is called the PDB database [FL] change. So, here for co connects this is actually an open source repository which means that anybody can go and play around with it. Whenever you find a new structure you go and deposit that structure over there and then you can download that and see sort of what is the structure of that folded protein.

So, for example, here is the human deoxyhemoglobin which is in the blood. Here is the human insulin in the second panel which regulates your blood sugar and then the final one is the human kinesin which is a motor protein which transports stuff inside the cells. We will talk a little bit about all of these proteins and how to model these proteins which is why I tried to show them as we go and go along in the course using different sort of techniques.

So, till now or well at least still last year which is when I downloaded this slide people had sort of sequenced 41407 distinct proteins. I am sure that number has increased even more if you were to go and look at it now. But so, this is what I mean when I say that there is a huge amount of data that you can be playing around with all the information. So, this when you have a protein off in the structure of the protein determines the function and the sequence of the protein determines the structure.

So, if you have a sequence, you might be able to take a guess as to what the three dimensional folded structure of the protein will look like and then from there on you can sort of see how the protein performs whatever function, it is supposed to do.

(Refer Slide Time: 13:01)



So, here are the 20 amino acids and so, when we say that the DNA codes for RNA and the RNA codes for proteins, you can sort of say that how many how many bases do I need to code for each individual protein.

So, code for each individual amino acid. So, I have four I have four nucleotides ok. Somehow I must combine these four nucleotides in order to build twenty amino acids right and then from these 20 amino acids, I will build whatever protein that I want to build. So, what should the unit of unit be so, that unit is called a codon? How many nucleotides to the codon contain such that an alphabet containing 4 letters can encode for this alphabet containing 20 letters? Is the question clear? So, let me say.

So, if a codon was one unit in that each nucleotide coded for a specific amino acid right. So, if I read an A in the sequence that meant I take some amino acid, let say this makes

methionine**.** If that was my unit if my codon was one unit, then how many amino acids could I make given that I have 4 nucleotides, 4 right. So, that does not work right. If my codon was 2 units so, if A T together coded for an amino acid, then how many how many amino acids could I make?

Student: 16.

16 so, that also does not work. So, 2 means I could make 16 amino acids that does not work. If I had 3 in a unit, then I could code for 64 amino acids right in principle. So, that is at least larger than 20 so, that is the minimum number that is the minimum number that you can should have in a codon in order to build an alphabet of 20 amino acids and indeed that is what biology does.

The codons that code for these amino acids are three bases comprised of 3 bases comprised of 3 bases. So, for example, here is the way you should read it. If I look at let us say glycine which is on the top over there, I will ensure I have a pointer next time. You start off from the inner circle which is A G and then the out next middle circle which is again A G and then the outer circle which has four different options U C A and G which means that all of these code for glycine over there.

So, let me see. So, the inner circle contains A G and then again A G and then U G G C G G A and G G G ok. So, because you have 64 possible combinations and you only have 20 amino acids that you are making, you have redundancies built into the system. It is not an uniform redundancy. So, you will see that some amino acids have multiple sequences some are fewer sequences. So, for glycine you have these four sequences or these four codons all of which whenever ribosomes seize any of these sequence it is going to produce a glycine.

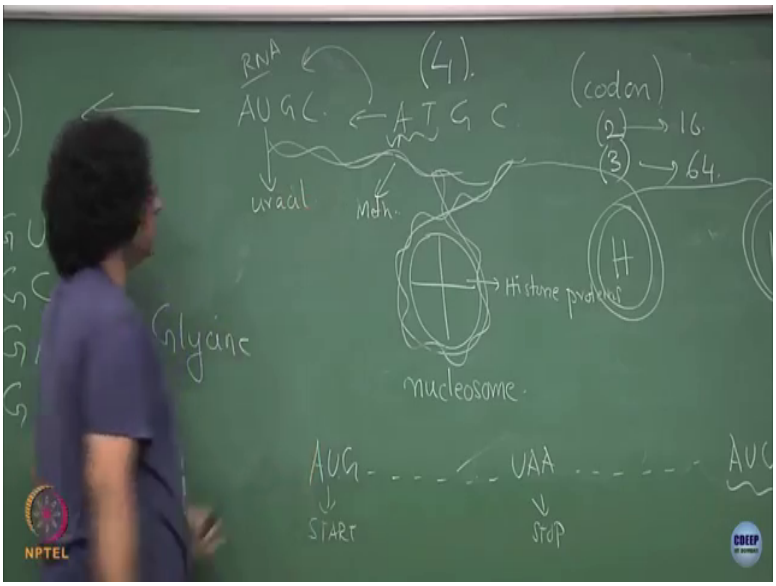
Similarly, so you will see that for example, the next one which is on the top over there glutamic acid that has only two redundancies. There are some which has a single redundancy; for example, this one over there in the bottom methionine has a single one which is A U G. So, A U G which is methionine does not have any redundancies. There is a unique codon

which codes for methionine and methionine is often called the start codon in that proteins sort of translation starts whenever you see this sequence ok.

So, you can have a random array of sequences G G A T A U G C C whatever. So, whenever transcription happens, it sort of keeps on reading until it finds this A U G sequence which tells it that this is the start position. It should actually record information from and then it starts reading the sequence in producing the appropriate protein ok. I think there is one more which has a single redundancy somewhere tryptophan over there, everything else has at least two codons that code for it.

And similarly just like there is a start codon, there is also a stop codon over there; there is two stop codons. In fact, so U G A is a stop over there and then there is another stop U A A A which means that whenever you see these codons, you stop reading and you say that is the end of my protein; I am not going to read anymore.

(Refer Slide Time 18:25)



Whenever the next thing was going to code for a new protein right; so, you have an A U G and then you have a lot of sequence in between and then you have something which is let us say U A A U A A. So, this is my start, this is my start, this is my stop. So, it is going to continue reading until it reads a stop and then again, there will be many things in between and then again the next coding region will start with another start codon A U G and end at another stop codon.

Student: Sir

Yes what is?

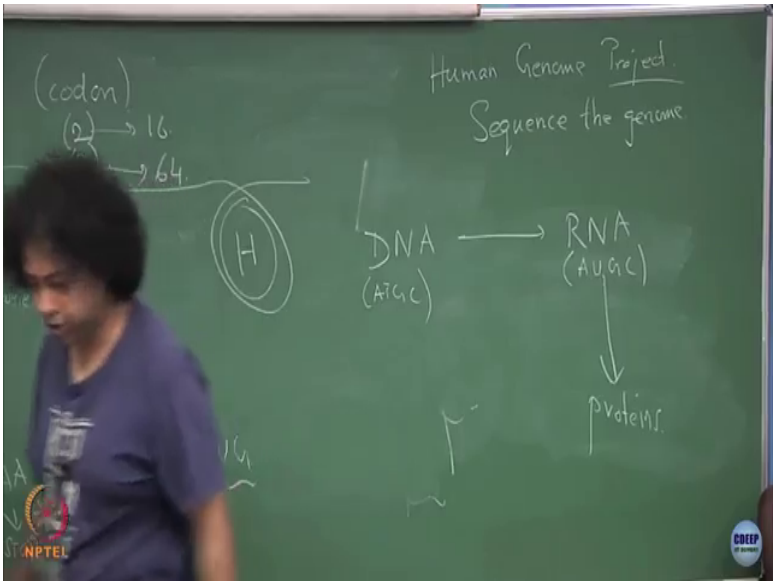
Student: (Refer Time: 18:58).

U? U is R ok, thank you. So, you will see that instead of A T C and G in this, I have A U C and G right. So, instead of this A T G and C, I have A U G and C and that is because in the

DNA, it is A T G and C. When the DNA gets read into RNA, the thymine gets converted into a close analog which is called uracil which is called uracil. It is like thymine except in RNA ok. So, it is a very close analogue of T.

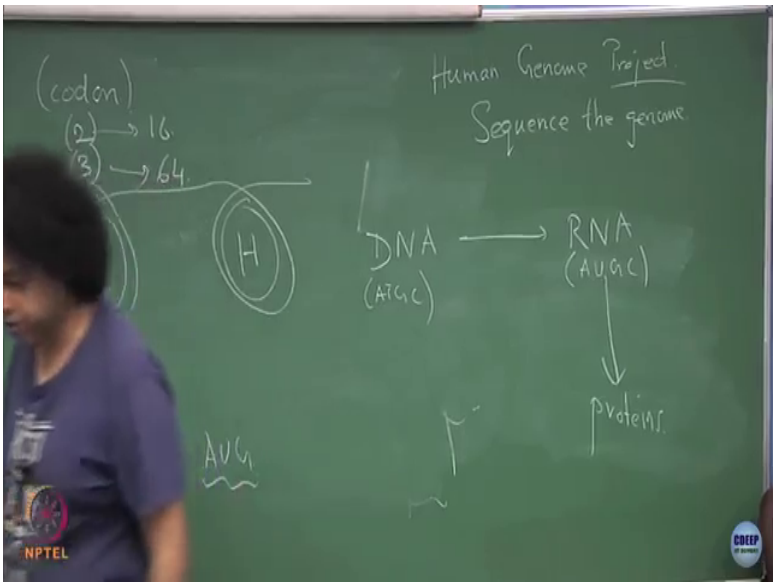
So, whenever this transcription machinery sees at T on the RNA, it produces are U. So, when this so, what is the central?

(Refer Slide Time 19:54)



So, you have DNA so, you have DNA; going to RNA going to proteins right.

(Refer Slide Time: 19:54)



So, DNA has this A T G C, the RNA has an a U G C and then that gets read in these three three codon combinations to produce the appropriate amino acid and there on the protein.