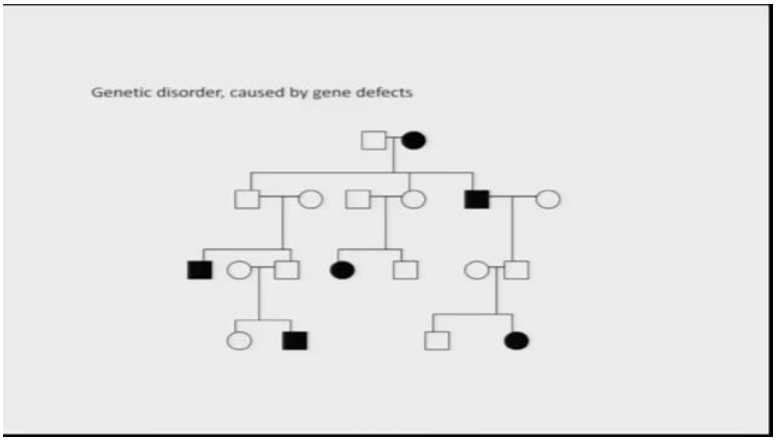
Human Molecular Genetics
Prof. S. Ganesh
Department of Biological Sciences and Bioengineering
Indian Institute of Technology, Kanpur

Module – 02
Lecture - 07
DNA Cloning and Hybridization Techniques - I

Welcome to the third lecture of second week. This particular lecture, we will be looking into techniques that we normally use for diagnosis and many other clinical genetic applications. Basically, we will be looking into the approaches we use for DNA cloning and hybridization. So, we will be covering some of the techniques like, southern hybridization, cloning, sequencing, PCR. So, these are the approaches generally being used to understand the genotype. So, when you talk about genotype, we are referring to changes in the DNA and how would you understand the sequence difference in the DNA. So, you need to clone or make copies of a particular segment of DNA, make a sequence of it and look into what sequence and how it may affect the individual or because it affects the gene or not. So, these are the approaches people use to understand the difference in the DNA. So, we will be looking into some of these issues and first let us look into some aspects.

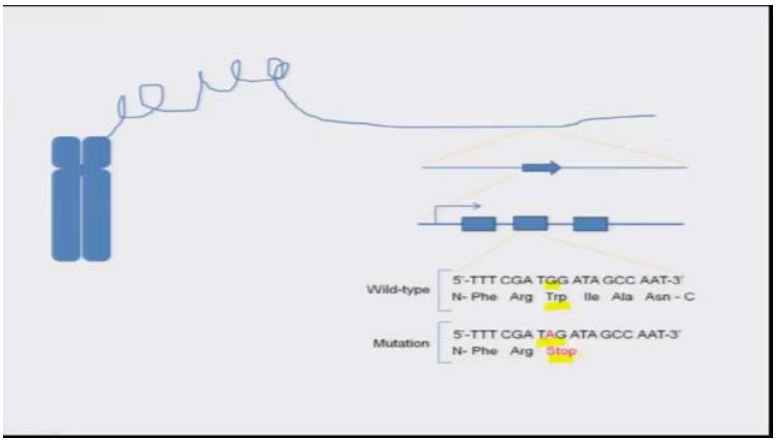
So, we in the last week, the last lecture, we looked into pedigree's, wherein we have looked at the phenotype in the individuals of various generations and tried to decipher whether the phenotype that you see is because of a defect in a gene and that has autosomal dominant phenotype, recessive phenotype or X - linked, Y- linked and so on.

(Refer Slide Time: 1: 47)



So, to understand how the disease is caused, we need to identify the gene and identity where is the variation. So, this is a very, very demanding task. So, we will be giving some of the approaches people use to identify the gene defect. So, when you talk about gene defect, the genes are nothing but a small segment of DNA that is present in our chromosomes.

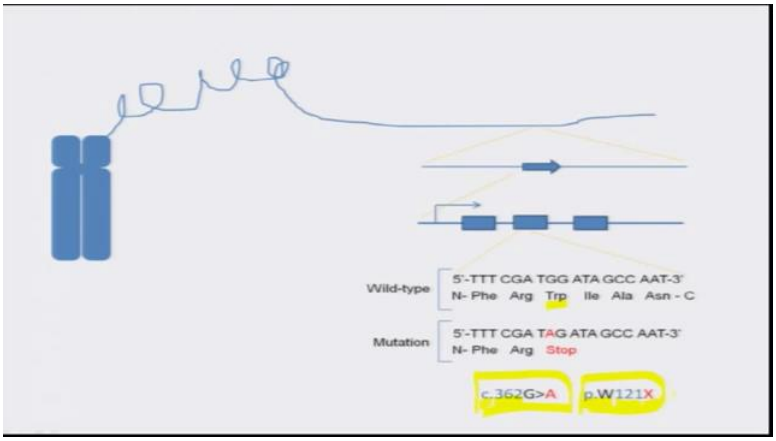
(Refer Slide Time: 02:18)



The cartoon that is shown here represent a chromosome and you have a long stretch of DNA that comes out and different segment may have different gene like for example, one that is shown here; the arrow indicates the orientation of the gene. If you look more carefully what is there in that region is that this is a segment of the gene which has got three exons and the three exons are regions that are present in the mature RNA. So, the RNA gets to the cytoplasm. You have the machinery that reads the code and then translates that code into amino acids. So, that is what is shown here. Some codes have been shown and which codes for, for example phenylalanine, arginine and so on.

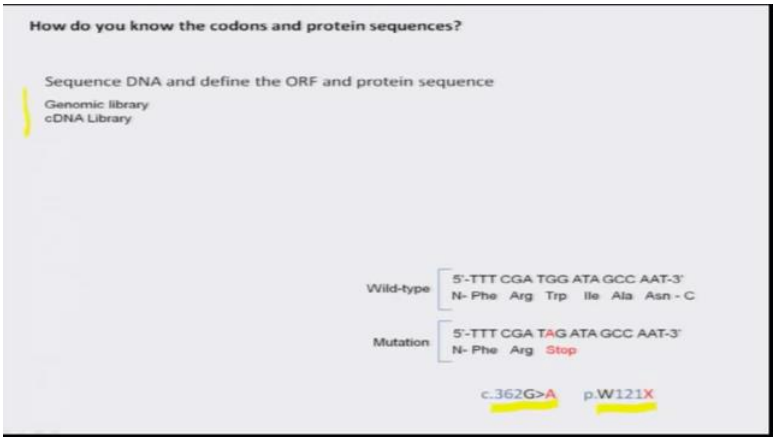
Now, what happens in case of disease? So, some change in the DNA sequence, which may affect the way the gene functions or it may affect the way the protein functions or the protein may not be made or made but not expressed or properly functioning and so on. So, what we have shown here is a point mutation, something that we discussed in our previous class. The point mutation is, you have particular base, and for example here the G got transferred to A. As a result, whatever amino acid that should be coded, now is not being coded there, but rather the synthesis of protein is stopped, because now the TAG represents a stop codon.

(Refer Slide Time: 04:00)



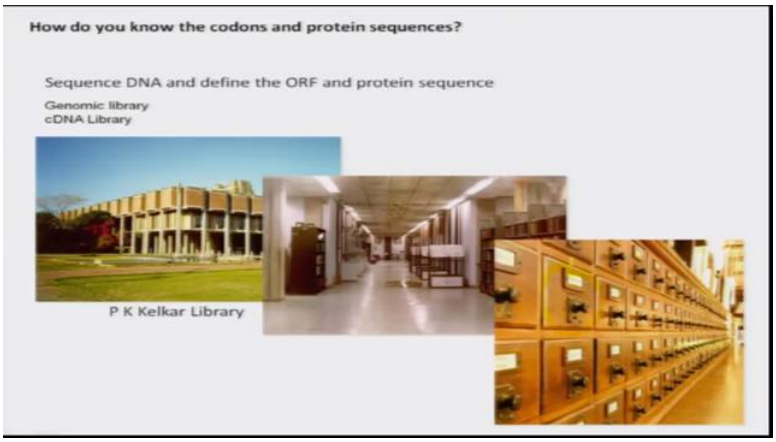
The question is how do you really identify? But before we get into that, let's see some of the nomenclatures or how do we name the mutations. This is one of the standard nomenclatures. For example, if we are talking about the DNA sequence you can see, a particular mutant allele is designated with this particular nomenclature here; it is hypothetical. So, the c stands for coding sequence, because it is present in a coding sequence, because it codes for a protein and the 362 is the number from the first nucleotide of the start codon, say the methionine that is coded by ATG. So, in that particular gene, A represent 1 and this particular gene which got converted to A represent 362. So, now this is the nomenclature is for the protein, because that is denoted by the letter P, which stands for protein. So here there is tryptophan, that is what is present here which is the residue 121 in the protein is replaced by a codon for terminating the synthesis that is the stop codon. So, this is how you represent as to what is the mutant allele, both at the level of DNA and at the level of protein.

(Refer Slide Time: 05:25)



Now let us ask the question, so how do we really know? So how do you know that this is the three hundred and sixty second residue of a gene and this is for example, the amino acid one hundred and twenty first? So, normally this information comes from sequencing different segments of the DNA that contain the gene sequences. So, we derive these sequences from these DNA fragments and these DNA fragments are derived from a kind of catalogue or library you call, one is called as genomic DNA library, other one is cDNA library. So, the genomic DNA library we will be discussing soon. So, this represents almost all the segments of your chromosomal DNA and we talk about cDNA library that represents your transcribed regions.

(Refer Slide Time: 06:21)

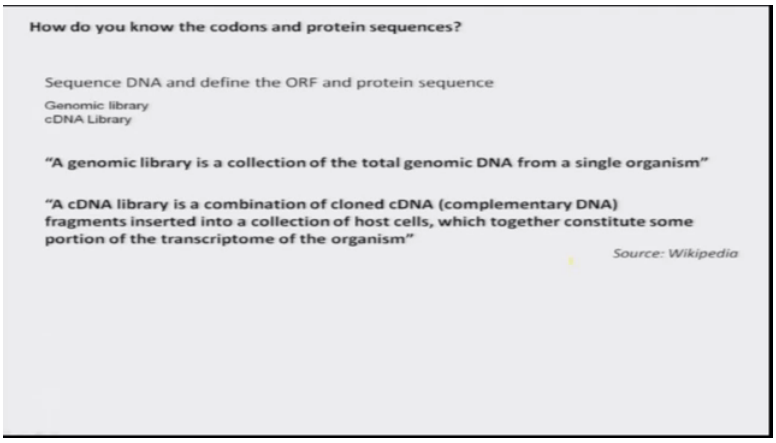


So, why do you call them as library? So, what I have shown here is a photograph of the library that is there in our institute. This is the famous Kelkar library of Indian Institute of Technology, Kanpur, which houses thousands and thousands of books, reference materials, journals and many other such literature that we normally consult for our research and

teaching and so on and what is interesting is that in any library if we get in, we will find racks of books like what is shown here and they all stacked there. Now, there are thousands and thousands of books. Now, how do you know where is the book that you are looking for? Say, I am looking for a book that discusses about human molecular genetics, and then I go and search for it. So, how would I really search? You know that in old, one of the traditional ways of doing it is that you have such drawers, which are attached with a particular alphabet. So, if the book starts with title 'human molecular genetics', I will go to a particular rack which, starts with H, pull out and then look at cards that are there and one particular card would have this particular book name and it would tell me in which rack, which floor this particular book is there. So, then with that information I go there and pick up.

So otherwise, a library is nothing but it is very similar; if you go to a book shop where they are selling the book, if you want to identify a given book it is extremely difficult for you to get without anybody's help. So, that is the difference. So, in a library you can go and get whatever you want, although it houses thousands of such literature, books and so on. So, that is something very similar what you see in genomic library or cDNA library. That is why they are called as library. They have everything that you look for, but there is also way, a way by which you will be able to go and get the particular fragment. Let's have a look at it.

(Refer Slide Time: 08:33)

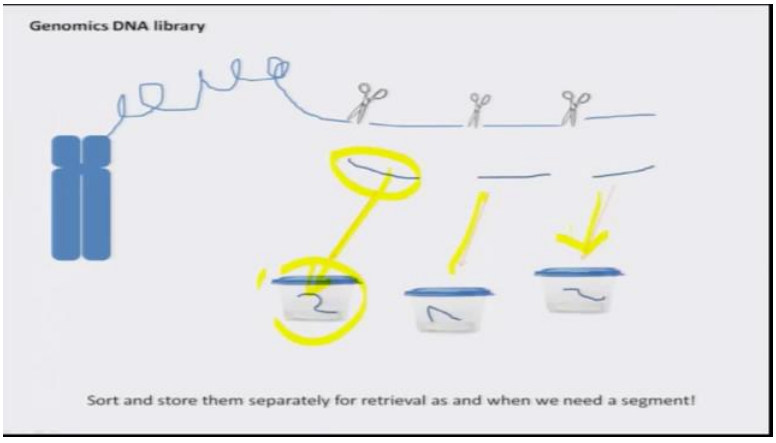


This is a definition I quote from Wikipedia, because it is easy for you to go and refer to the Wikipedia. So, if you give such a command called genomic library, it would define it as, "a genomic library is a collection of total genomic DNA from a single organism". Likewise, if you give such command for library it would also tell you that that also represents collection

of all the books that are available. So, this is one, it is collection and second, you will be able to retrieve whichever, you know, fragment that you want. That is, this is the way by which we will be able to get.

On the other hand, cDNA library is a combination of cloned cDNA. You call it as cDNA, because they are complimentary DNA and they are complimentary DNA, because these are DNA derived from messenger RNA. You copy the messenger RNA which is single stranded, therefore they are complimentary cDNA. The fragments of the cDNA are inserted into again some host cell. Likewise, for example in the DNA genomic library and this cDNA library represent your transcribed regions. For example, majority of the messenger RNA that is expressed in a given tissue if you have developed a cDNA library, for example brain and so on.

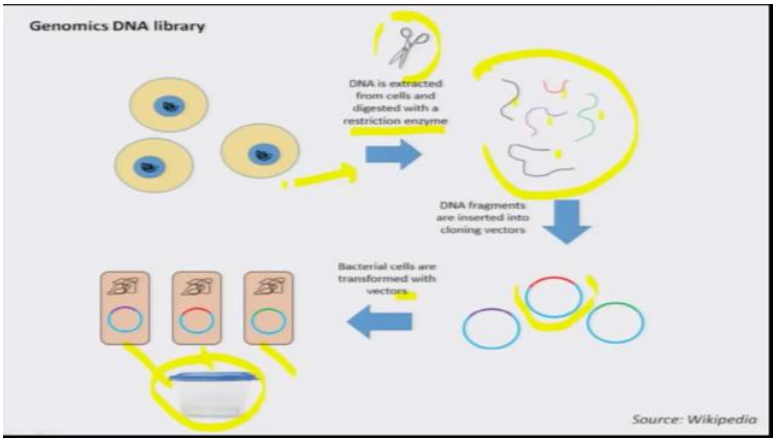
(Refer Slide Time: 09:50)



So, let us have a look at how genomic DNA library is made. This is what it is. You have the DNA that is running from top to bottom in a chromosome and what you need to do is normally, that you have to cut them into smaller pieces and not only that, you have to sort them and then store them and why do you store? For two reasons; one you don't mix one with the other, therefore the identity is lost. For example, if I pull out randomly different pages of different books and bind them together as a book, you will not be able to make any sense of it. So, you need to have them separately. So, that is exactly one of the goals when you make a library that have certain containers into which you are able to put one particular type of fragment. It could be multiple copies, but represent this particular region of this

chromosome and then sort them like this, and these are smaller DNA fragments and then you store them and whenever you want, you will be able to get them.

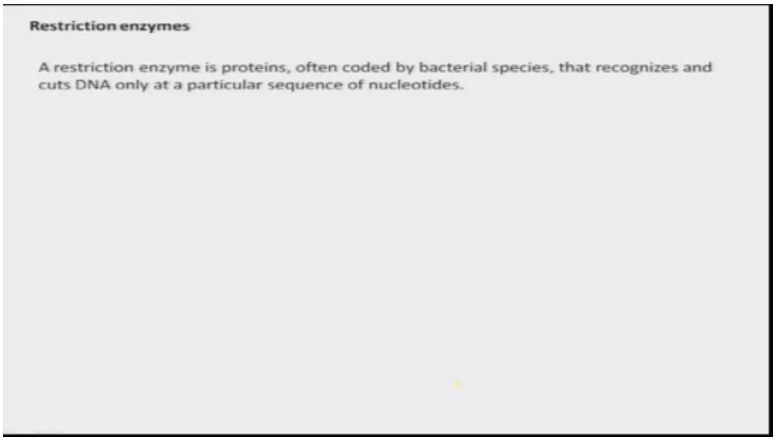
(Refer Slide Time: 10:50)



So, how do you really do that? So, this is the schematic. So, basically you extract the DNA. So, any of our tissue will give very similar DNA, because there is no big difference between the DNA that is present in my skin as compared to my DNA that is present in my white blood cells. So, I can use any tissue that I think, that is available; easiest thing is that, for example I can donate some amount of blood, isolate the WBCs, white blood cells and then extract the DNA and then I use the scissors to cut the DNA like what is shown here and what are these scissors? These are called as restriction enzymes, we will look into that. These are enzymes, which are proteins which cut the DNA at specified region and once you made these kinds of different pieces and you will be able to put them in specialized host called vectors that are shown here and then, these vectors get into a host organism; for example, here it is bacterial cells, which are like your containers.

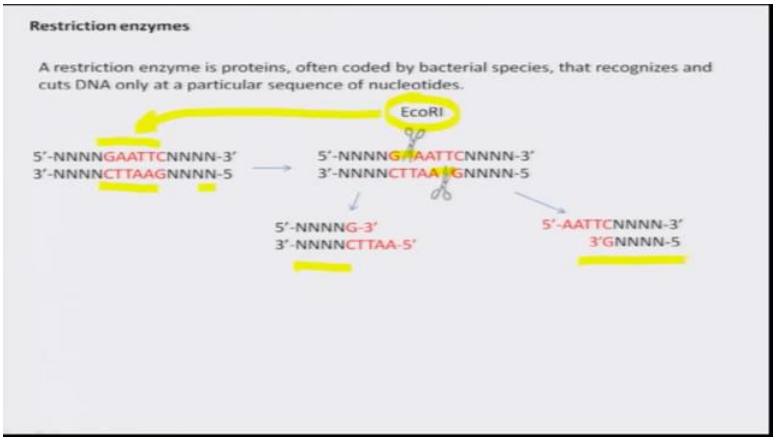
Now, each one can be put in different container and we can store them. Whenever you want, we can pull out one and make as much DNA that you require for any of your downstream applications. So, this process of cutting the DNA is done by restriction enzymes and then, putting each fragment, you use specialized host to carry the DNA. These are called as DNA vectors.

(Refer Slide Time: 12:30)



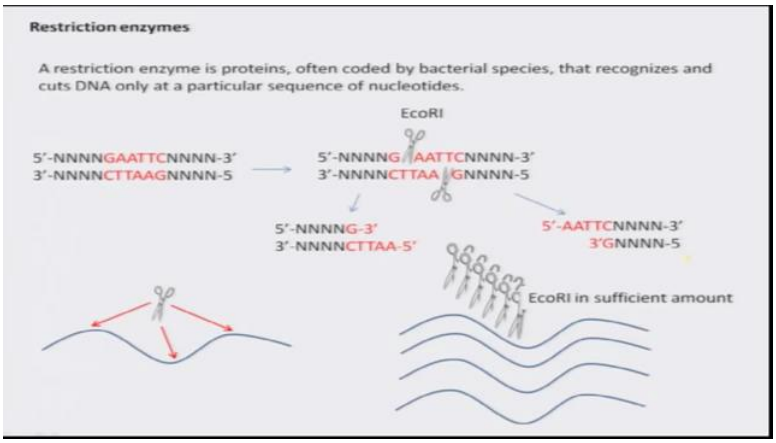
So, these are two important components of a DNA library that we are going to look into. So, what are restriction enzymes? As I said, these are like scissors, they cut. So, what is special about them is that, they recognize and cut DNA at a particular sequence, meaning if there are ten such sites in my DNA, if I add enzyme it will cut exactly at the ten such sites. For example, I have a DNA sequence that doesn't have any of that sequence that would be identified by the restriction enzyme it will not cut, no matter how much enzyme I add. So, they are so specific to the sequence and that specificity really helps us to make the libraries. So, these are proteins, as I referred to and normally they are expressed in bacterial cells and these are some sort of defense mechanisms for the bacteria to protect from foreign DNA getting into them. So, that is why they make these enzymes and we identified these genes and we exploited them for our own benefit that is to make the restriction in the nucleases, which we use it for such kind of applications.

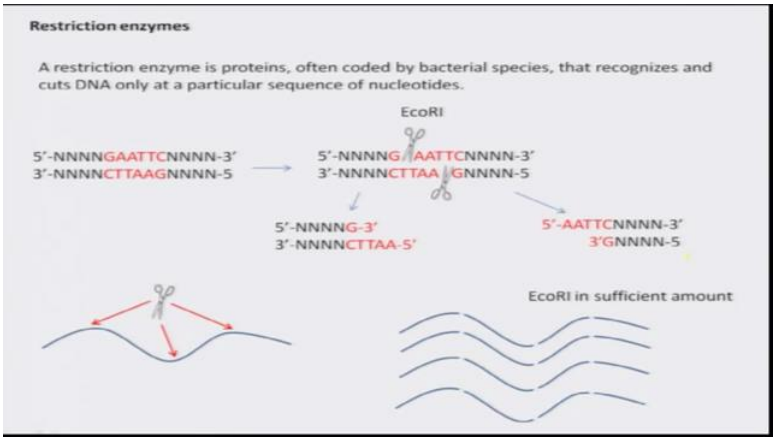
(Refer Slide Time: 13:41)



So, how does it work? So, let us see some cartoon here. So, what is shown here is DNA segment and you see a particular sequence, which is shown in the red color font, for example, GAATTC. So, this is a sequence that is recognized by an enzyme, restriction enzyme with a name called *EcoRI*. Let's not worry about the name, but this is a unique name and tagged with a unique restriction site. So, this enzyme cuts only this site. It doesn't cut anywhere else. So, how does it do? You can see that it pretty much breaks the backbone of the DNA exactly between G and A here, likewise at the other strand. So, as a result, one DNA strand, that double strand that we have seen here, is cut into two fragments. So, this is how the enzymes cut. So they cut only where such sequence GAATTC is there. Even if one nucleotide is not the same, for example it is GATTTC, the enzyme will not cut. So, it is so specific to the sequence.

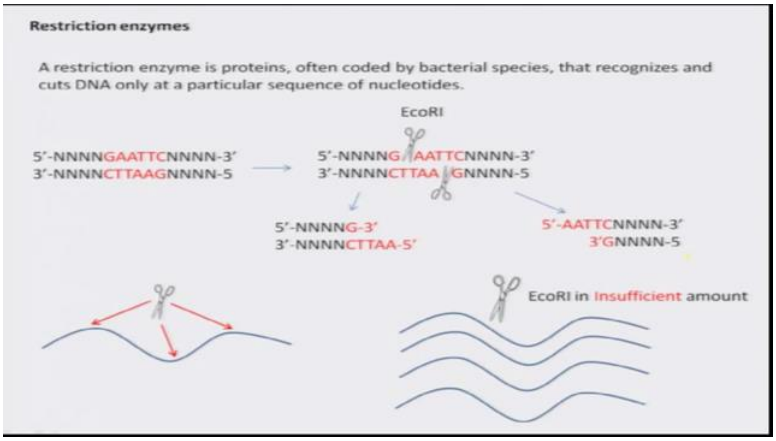
(Refer Slide Time: 14:54)





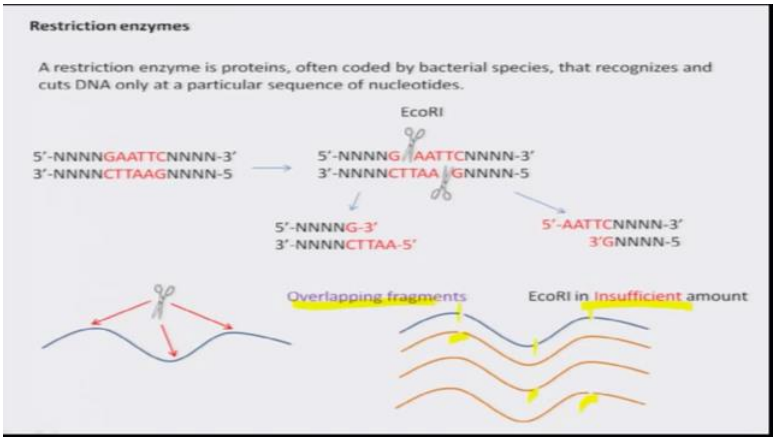
So, if for example I have a DNA sequence and this enzyme has got three sites and no matter how much, if I add multiple copies of the same DNA sequence, if I add enough amount of this particular enzyme, it is going to cut all the fragments exactly at the same site.

(Refer Slide Time: 15:17)



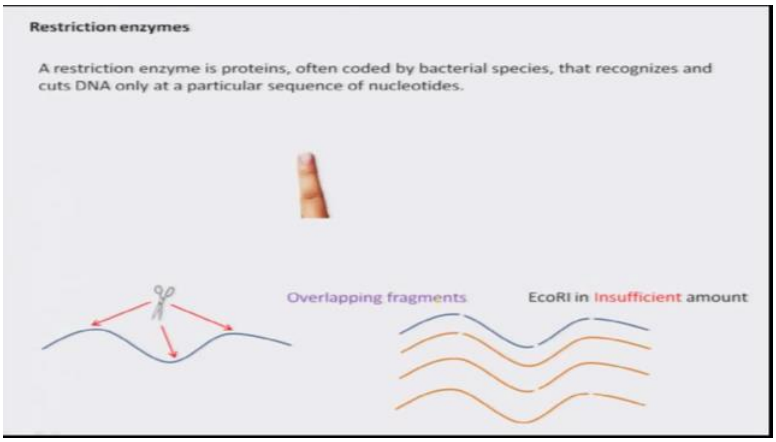
But, if I do the same the reaction with little amount of enzyme, where it is limiting, so what would happen?

(Refer Slide Time: 15:21)



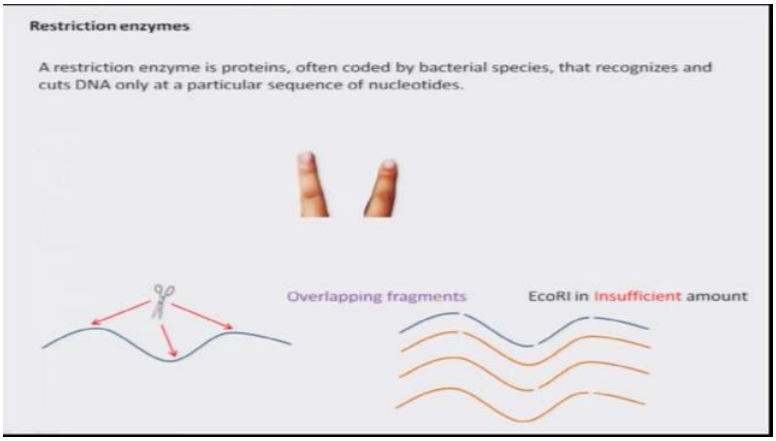
So here, it might randomly cut, because the amount of enzyme is insufficient, at times what happens, ideally you would like that it would cut here and here; but, if you give somewhat lesser amount of enzyme, then what happens, because of the kinetics, it may cut here in this fragment or here in this fragment and so on. So, not all the sites are cut if you, for example, reduce the amount of enzyme or if you reduce the duration of the reaction, if you give it for only short period. So, such reactions really help us to get what is called as overlapping fragments. What they are? What do you mean by overlapping fragments?

(Refer Slide Time: 15:58)



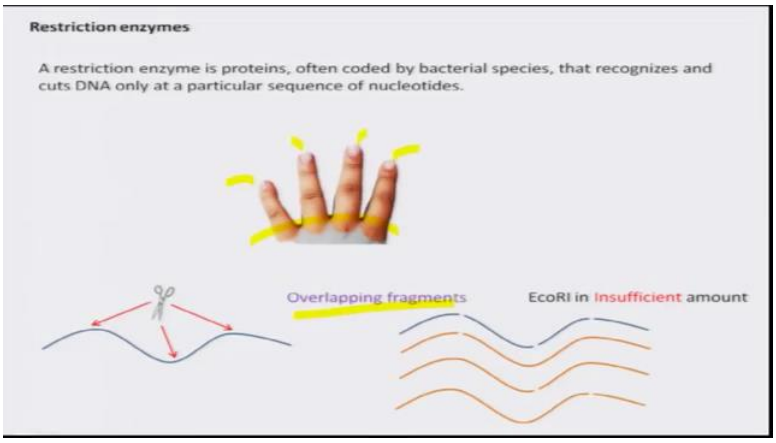
These are very, very important. I will explain to you. Let's see this particular figure. What is that? It's a finger, right? It is very obvious to you that it's a finger, but what is not obvious is, is it an index finger, is it the middle finger, is it the ring finger, is it the small finger? Except probably thumb, you can think of any finger that it represents.

(Refer Slide Time: 16:25)



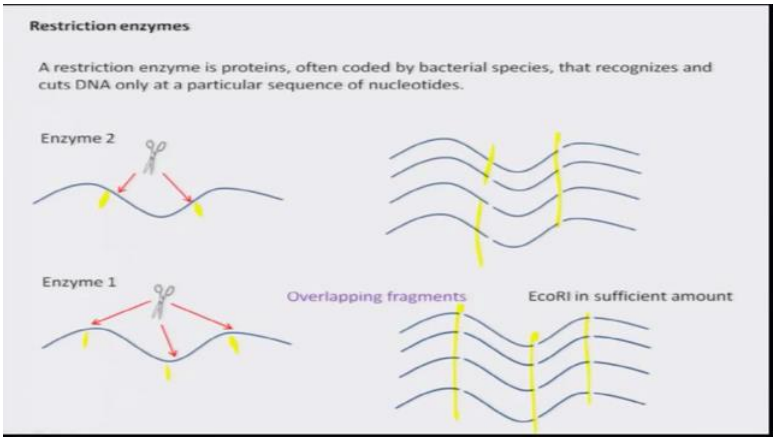
Even if I add one more finger it would be very difficult for you to tell whether these two fingers have come from the same hand or different hands. So, which one is index, which one is middle, or it is very, very difficult. So because each one, is independent of each other. It looks like from this figure.

(Refer Slide Time: 16:43)



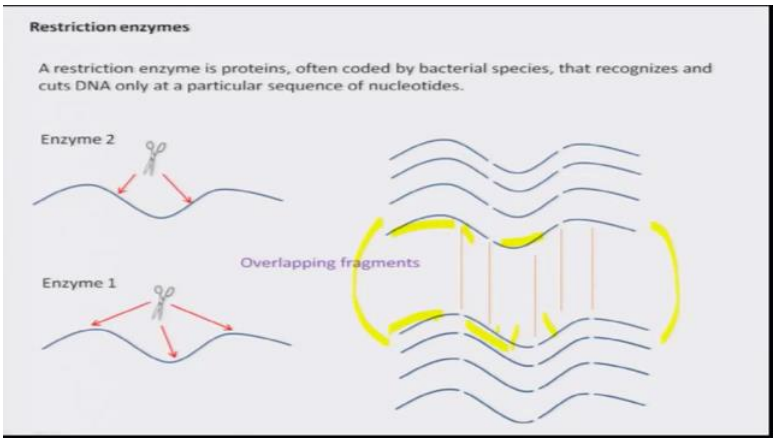
However, if I show you everything, then you will be able to tell that this is the small finger, this is index, middle, ring finger and so on, because you have a connection that, that each one connects to each other. That really helps you to relate and position them in certain order and that's where the overlapping fragments, even the DNA helps us to connect two different fragments and tell the relative order. Let's see how it is.

(Refer Slide Time: 17:10)



For example, here I am taking enzyme 1 and enzyme 2, but the same DNA fragment. Enzyme 1 has got three sites as shown here and enzyme 2 has got two sites as shown here. So, if I do or perform two different digestions, for the same DNA, you will find that that for this, it is cut here, for this it is cut at three different sites.

(Refer Slide Time: 17:38)

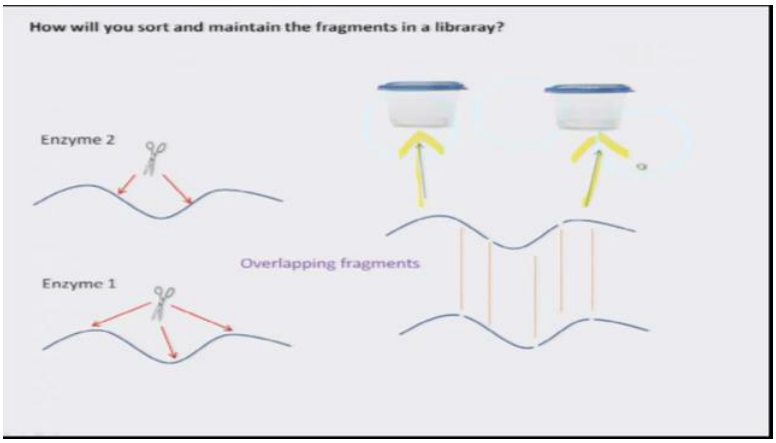


What is interesting here is that if you look into these two fragments which represent two different digest, there are overlaps. For example, this fragment overlaps with this fragment, whereas this fragment overlaps with the same fragment. So, this fragment is able to connect these two fragments. Likewise, this fragment is able to connect this and this and so on. So, by using these DNA fragments, although they are independent fragments, by looking into the overlapping segments, I am able to stitch back probably the entire chromosome, if I have a way to do such kind of overlapping. So, all the genomic libraries that are made for human or any other species, have collection of such fragments. This is not only fragments that represent different parts of the genome, but they also have overlap. That's why I am able to connect

one fragment with the other and I am able to recreate the entire, know, chromosomes. So, this is how you know restriction enzymes help us to create fragments of our genome, yet you are able to retain some information with regard to the overlap and so on. So, that's very, very important in generating DNA libraries.

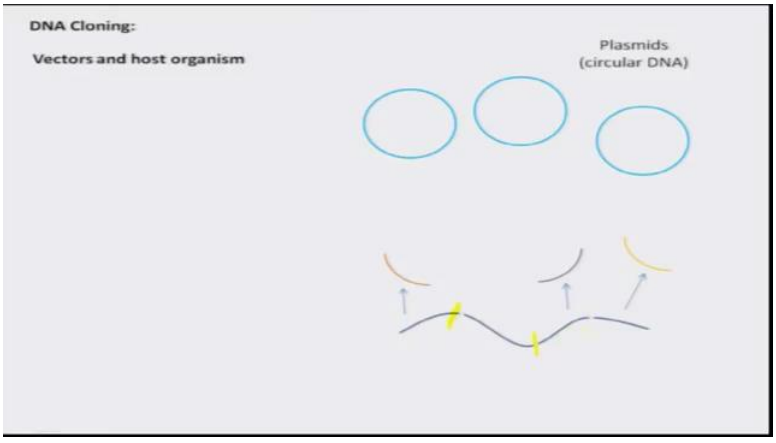
So, by doing this what do we do?

(Refer Slide Time: 18:56)



We have digested the DNA and we have fragments and we are able to put them in different containers. So, what are these containers?

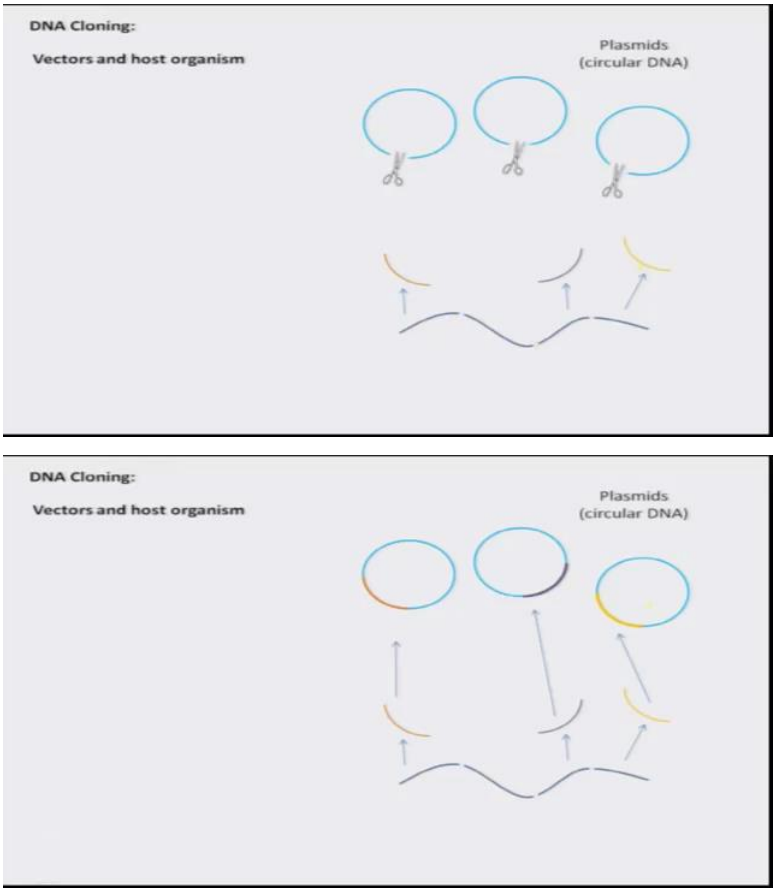
(Refer Slide Time: 19:04)



I just told you that these are vectors. So, what are vectors? There are various types of vectors available. What is being discussed is the most simpler form of vector called as plasmids. These plasmids are circular DNA that are present naturally in many of the microbes, but the

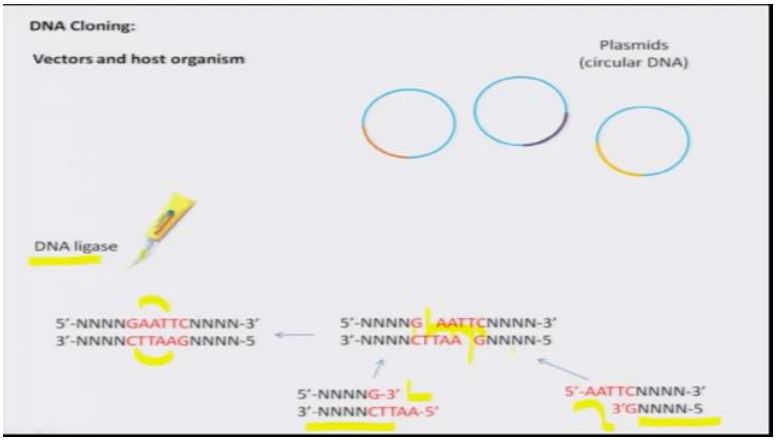
one that we use for a purpose, what is called as DNA cloning, are the vectors that are man-made. So, we have modified a lot; we will come to that little later. Therefore, they are very, very useful for cloning DNA, meaning to sort the DNA into different containers. So, how do you do? This is what it is. So, you have a DNA fragment, you have digested them with enzyme which cuts here and here. So, I am showing three different pieces. What do you do?

(Refer Slide Time: 19:55)



Each piece you separate them and then you use the same enzyme or a combination of enzymes to cut or to make a cut in the circular DNA and you are able to stick this DNA piece into each one of the plasmids. So, in this way, again you have created a circular DNA, but now having a part of the foreign DNA into them.

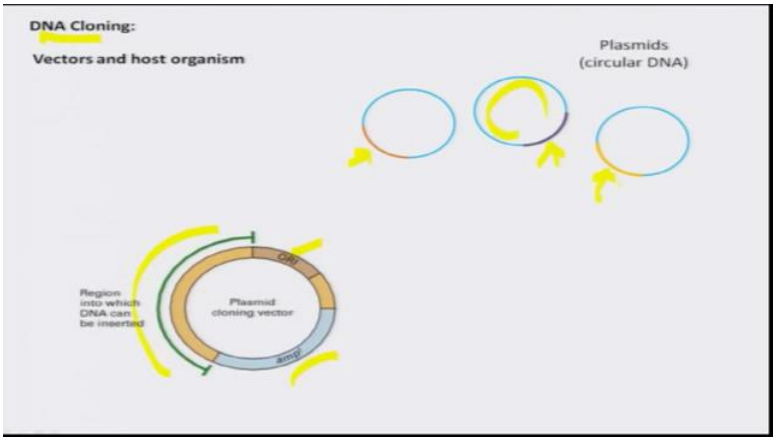
(Refer Slide Time: 20:17)



How is it possible? It is something like, you are able to cut a rope and then join with a small piece using glue. For example, well known glue that we use at home is the fevikwik, very fascinating advertisement; you must have seen that within a second that it is able to stick together. So, very similar kind of glue we will use to join DNA and this glue is called as ligase and this glue is there in your cell. So, whenever there is damage in your DNA and this enzyme is expressed and is able to join. So, that is how we are able to maintain your DNA and such enzymes, which are again proteins, are extracted from many *E. coli* and other such sources and we use this protein to stick the two different DNA.

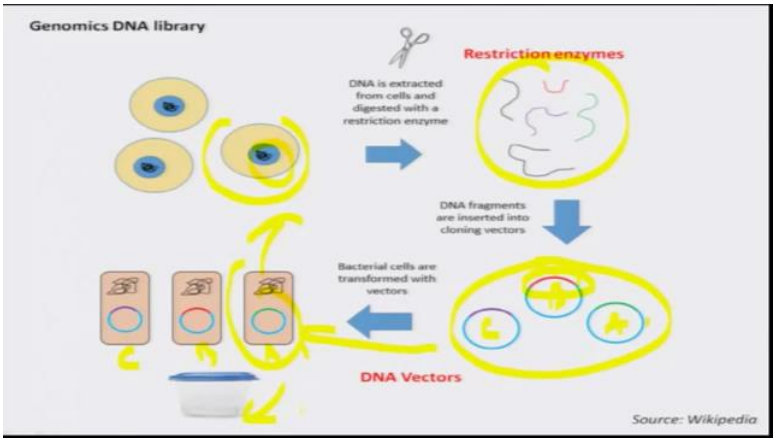
Like for example, if you have a DNA fragment digested by an enzyme *EcoRI* that we discussed earlier, so it would have this kind of what is called as sticky and which are complimentary to each other now, which can now go and form hydrogen bonds. When they do form hydrogen bonds, this ligase can come and seal it and make it again as a single DNA fragment. So, this is an approach people use to combine two different DNA fragments that were digested with an identical enzyme, which has an overlapping sequence; you are able to recreate the DNA. Again, this ligase is very, very specific. If the sequences that are, joining together are not identical, it will not join. If the overhangs are not complimentary to each other, it will not join. Therefore you will not be able to join two different DNA fragments which have sequence difference in the overlapping regions.

(Refer Slide Time: 22:17)



So, this way, we are able to put the foreign DNA inside this circular vector, which you call as plasmid. So, this is again a schematic of a plasmid which is used for cloning DNA and what it has? For example, it has sequence called ORI which is the region that is required for the replication. So, because the moment this DNA gets into *E. coli*, it has to make copies. Otherwise, whenever the *E. coli* cell divides, this DNA will be lost during cell division. So, it has to make multiple copies. So, this ORI, origin of replication really helps in making copies of the DNA. Then, you also have a region called marker region, called here for example, it is written AMPR, which represents ampicillin resistance, meaning this DNA segment produces a protein which is able to degrade antibiotic, for example ampicillin. So, this protein would inactivate this antibiotic, therefore the *E. coli* can survive, if it has got the plasmid and then you have a long region in which you will be able to clone different fragments like what is shown here. So, that is the region where you can put the foreign DNA.

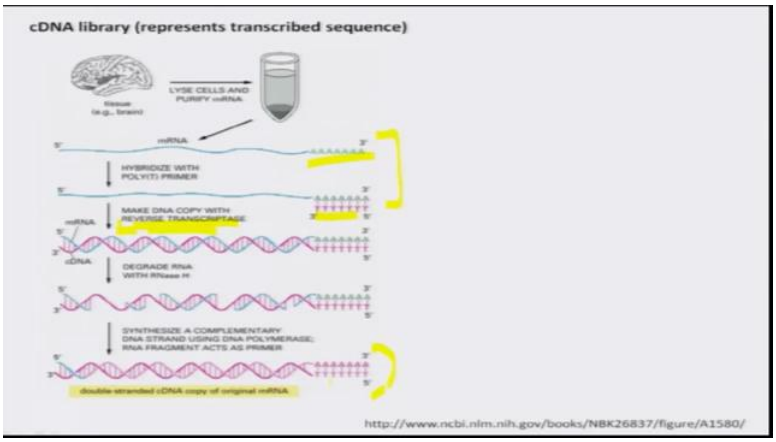
(Refer Slide Time: 23:39)



So, this way we are able to pack the vector and once this vector is ready like what you have shown here, now we can put them inside a host. The host here, what is shown here, is bacterial *E. coli.* So, basically you force the DNA to get into the *E. coli* and normally the *E. coli* cell would accept only one such DNA in a cell. So, for example, if you say A, B, C, you see that A plus vector has come here, B has come here, C has come here. So, this is the way they are able to sort them from a complex mixture that you have seen here. Each piece of the, we are able to put them inside a plasmid like here and each plasmid we are able to put inside *E. coli*.

Now, if you culture only this particular *E. coli*, you are going to make millions of copies of this particular DNA fragment and that is why it is called as cloning, because we are making multiple copies of the same DNA sequence, , in a very, very, cheaper way, because *E. coli* really doesn't require any to, special reagent, to make copies of the DNA. So, you give them all the food that it require, they have the machinery to make it. So, this is the way we are able to get a complex DNA that is there, present in your cell, extract them and make smaller pieces and put them inside the vectors and then sort them and then you have cells that are put into containers, and then you can go back and get whichever region that you want; that is something that we will discuss little later. So, this is the process for genomic library.

(Refer Slide Time: 25:21)

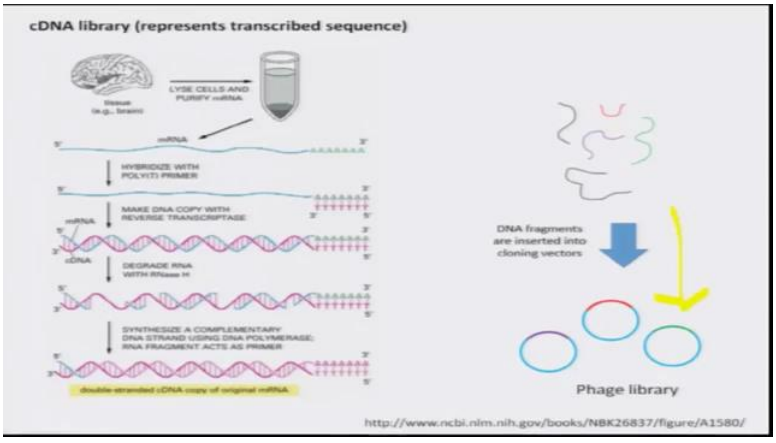


Very similar approach is also used for making a library representing your mRNA, which you call as cDNA library, which represent the transcribed sequences. So, here what we do is, starting material is RNA obviously, whichever tissue that you are interested in understanding expressions of genes, you take the tissue, extract the RNA and then, the messenger RNAs

have unique properties that is they have, most of them have what is called poly A tail, which are polymers of A base and then we can use this sequence to make what is called as a complimentary sequence.

So, if you have a primer, oligonucleotides that are artificially made which are complimentary to this poly A sequence and then, if you add an enzyme which, which we described in the central dogma of biology, which is a reverse transcriptase, which uses RNA as the template and makes a DNA out of it, that is what shown here, it is able to make copy of the mRNA, it copies mRNA to make a DNA. So, this DNA, single stranded, can be converted into double stranded by a process which we need not get into, but basically what we are trying to say is that we can use the reverse transcriptase to make the complimentary DNA, which is cDNA and then make into double stranded DNA.

(Refer Slide Time: 26:54)

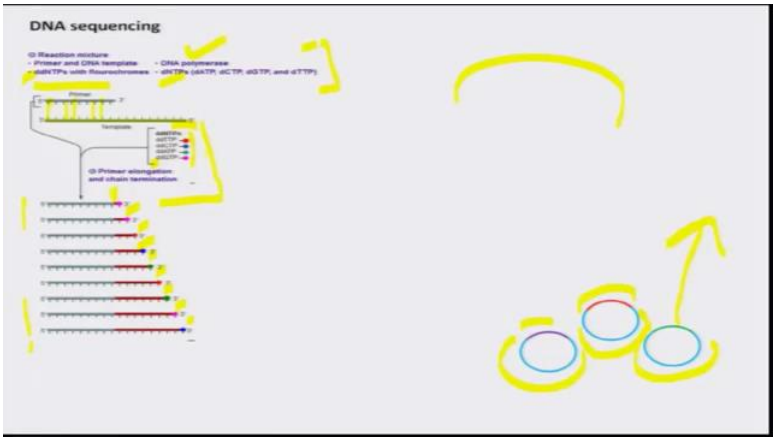


So, this double stranded DNA can now be used to clone into vectors and that can be made again into library. So, that is how we make cDNA libraries. So, you have a genomic library that represents different segments of the genome. Likewise, you have cDNA libraries that represent different transcripts that are expressed in a given tissue. We can make one library for brain, one library for liver, one library for kidney and so on and so forth; that would tell you what are the genes that are expressed and once you have made the library, the next big task is to sequence them. So, how do we know what are the proteins that are made in our brain? We don't sequence proteins that are extremely complex, difficult to do.

What we do is, we sequence the RNAs and predict what is called as open reading frame, something that we discussed in the first week lectures and this open reading frame help us to

predict what the protein sequence that you could make is and you use the protein sequence to predict what could be the function of the protein. So, this is how most of the protein sequence that we know is derived at. So, how do you really sequence?

(Refer Slide Time: 28:07)



So, this is a quick introduction to DNA sequencing, which also uses same concept that our body uses for making copies of DNA during replication process. So, what we have is, is the DNA segment that is cloned in the vector. So, this is our starting material, say for example, I am interested in understating what are the genes that are expressed in brain. So, I pull out a clone from, say a cDNA library derived from the brain tissue and then I try to sequence it. So, what I know in this clone is, the vector backbone is common for all the DNA present in the library. What is new is the fragments that are present here. So, I don't know the sequence of the fragments, but I do know the sequence of, let's say, the vector that I am starting with.

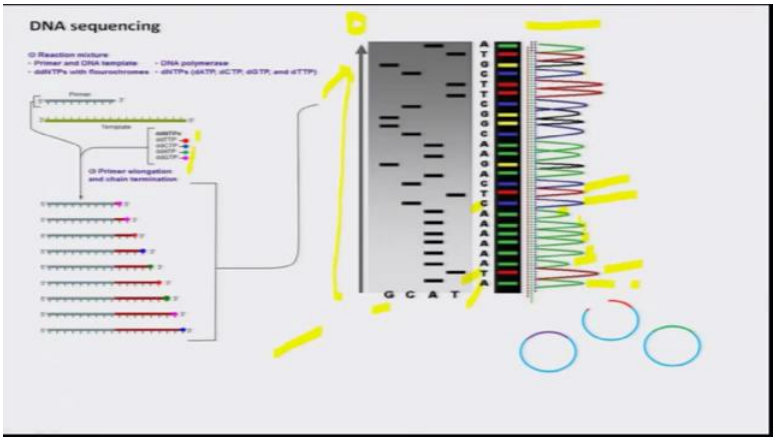
So, what I do is that if I have a vector and what I do is that I make a small oligonucleotide, meaning these are synthetic DNA which are complimentary to the particular DNA region and as shown here for example, then I, denature the DNA, meaning I boil the DNA, for example. As a result, all the hydrogen bonds are broken, so the DNA is in a single stranded form and you give very high concentration of this oligonucleotides, they go and bind to the region that are complimentary to the single stranded DNA that is present there and then, you add to that tube all these regents.

What are these? These are bases, which are specialized bases. For example, these are not the normal base, these are dideoxy, what you call as, like what is shown here. So, you add the

primer, you add the DNA template I already explained and then of course, you have to add the DNA polymerase, which is an enzyme which copies the DNA and then, we add the normal bases that are required for copying the DNA, the deoxy NTPs and also you add some specialized bases, which is called dideoxy. These bases are called as dideoxy, because once that base gets added, the DNA cannot be polymerized any further, meaning the synthesis of a new DNA strand would stop as soon as one of the dideoxy base is added. So, that's why we are able to stop.

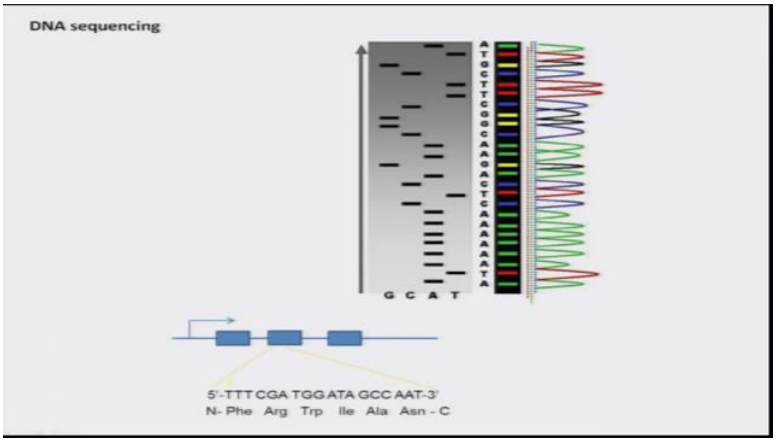
But, the trick is that, you keep the concentration of dideoxy and deoxy to some ratio, such that more often than not, your normal base is added. Therefore the DNA continues to be made, but once in a while your dideoxy base will be added, but if that happens the synthesis would stop there. So, when you do that kind of a reaction, what you are going to see is something like this. For the same region of the DNA, the primer is bound. For some, here the termination, that is the DNA synthesis, stops in the first base, some it stops in second base, some it stops in third, fourth, fifth and so on. So, you are going to have different lengths of new DNA that is being made and what you have done is, for each one of the dideoxy bases you have attached some unique chemicals, which when you excite with certain wavelength of light, will give some fluorescence, will emit some light, which could be of different wavelength. So, by looking at the light, I will be able to call it as base A, or base T or base C or base G. So, by looking at the fluorescence, I am able to look into and call them as which base it is.

(Refer Slide Time: 32:03)



So, what I do is, I separate the DNA fragments based on their size. This is something that we will come back to little later. So, when I separate them depending on their size, the one that are shorter will be here, one that are very long are going to be here and when I do that, I will be able to call what fluorescence they have, which kind of the four different fluorescence, that I discussed which are attached to each one of the four bases. I look into what is the fluorescence and call them as a base, something like what is shown here. So, what is shown here is the fluorescence intensity? So you see, you get a green fluorescence that means I call it as A, if it is red I call it as T and then, I have a stretch of AAAA and then I get a green, blue that is C, again I get a red, which is T and so on. So, I am able to read the sequence, by calling what is the fluorescence that comes out. So, this is the way, I am able to sequence the DNA. So, this is called as one of the robust approach these days' people use. This is called as automated DNA sequencing.

(Refer Slide Time: 33:17)



So, this way I am able to sequence a fragment of the cDNA and I am able to get what is a nucleotide sequence and based on the sequence, I will be able to tell what the protein is. So,

this is how these are derived at. So, that is how we are able to predict what kind of protein in a given gene can code for. We look into the sequence and then we predict the protein sequence. This tells you the protein that a gene can code for. So, what we need to do now next is to understand how a change in the DNA sequence can cause a disease, because our course deals with human molecular genetics, so it is something I am going to discuss in the next class. For this class, we have discussed the basic approaches that people use to generate libraries and then to sequence them. So, this is what it is and we will see in the next class.