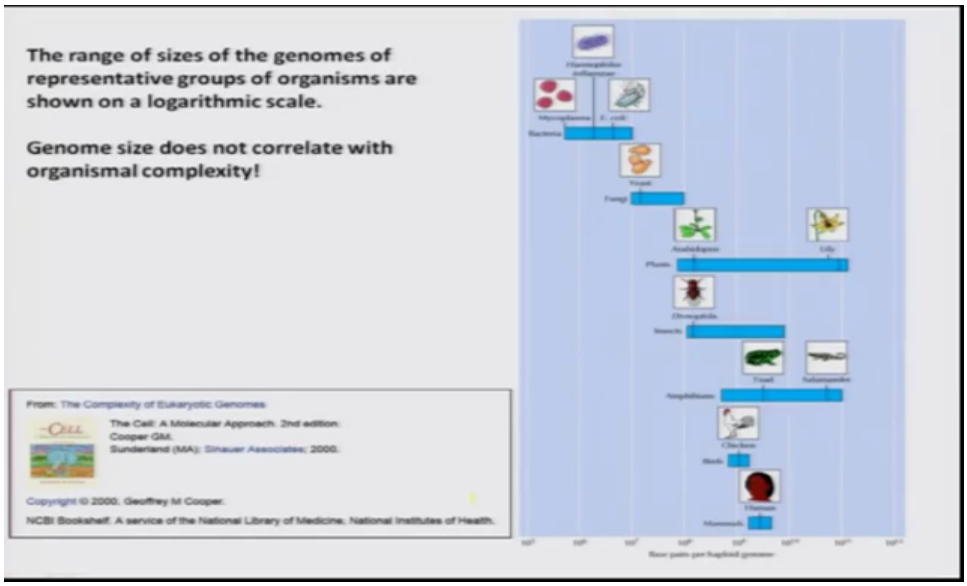
Functional Genomics
Professor S Ganesh
Department of Biological Sciences & Bioengineering
Indian Institute of Technology Kanpur
Lecture No 10
DNA Sequencing Methods Part 1

So welcome back to this course of Functional Genomics. In the previous lecture we looked into the various databases that are used to store the information that are derived from the genome including the functional aspect and how we can query them to understand the functional genes and build your hypothesis do more you know to you know understand the function of the genes. So in this class we are going to look into some of the methods that have been developed in the early years to understand how to sequence genome. So let us ask a question like how the genome is so special for human such that we became in species which study the other species.

(Refer Slide Time: 0:59)



So for example (())(0:59) go to the university or college and study science and use human as a model system but we do it other way. So what is so special about the human genome that made us so successful species in terms of your cognitive abilities, you are so successful you could domesticate other species and study them as a model system or for our own benefit. So what is shown here in this slide is the range of sizes of the genome of representative groups in log scale, so that is what shown here.

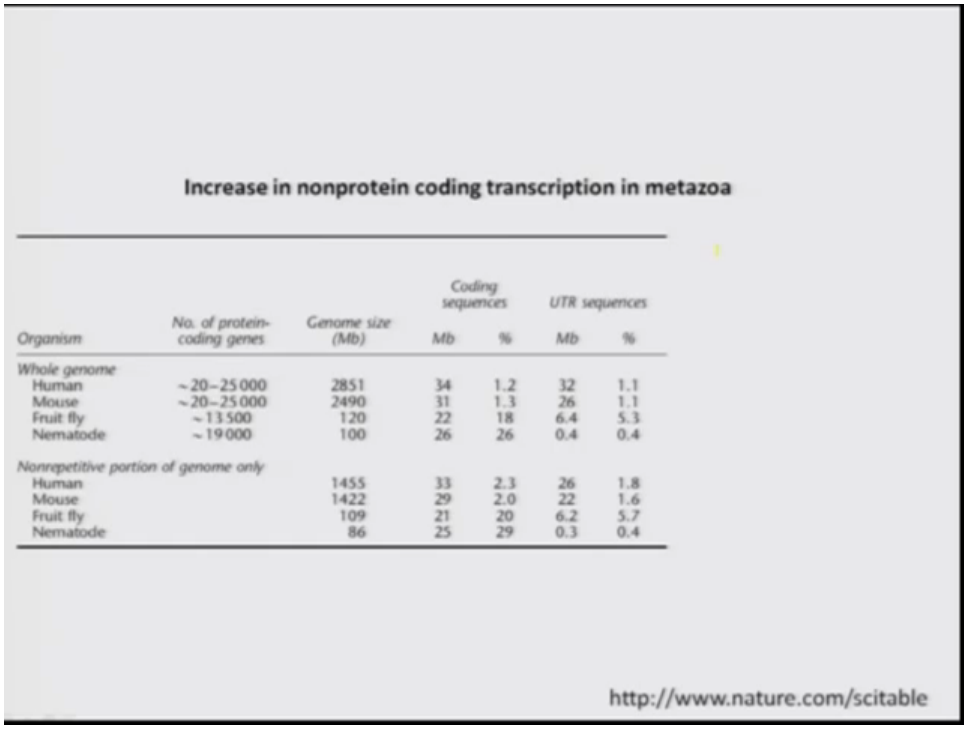
So we can see here that human is here, the genome size is comparable amongst all the mammals because we believe that the humans (())(1:44) from a common set of species. And if you compare the genome size of the humans with other popular model that we study is not that the genome size correlates with the complexity. So you can have for example a frog which we call as salamander which is having a genome which is in a (())(2:08) for larger than human likewise we have plants for exam the lily the genome size is much larger as compared to the humans.

Ours is not so different from the chicken or the tord or even for example the fly and so on. So what is that makes us so special right that become so successful species and we study the other species.

(Refer Slide Time: 2:41)

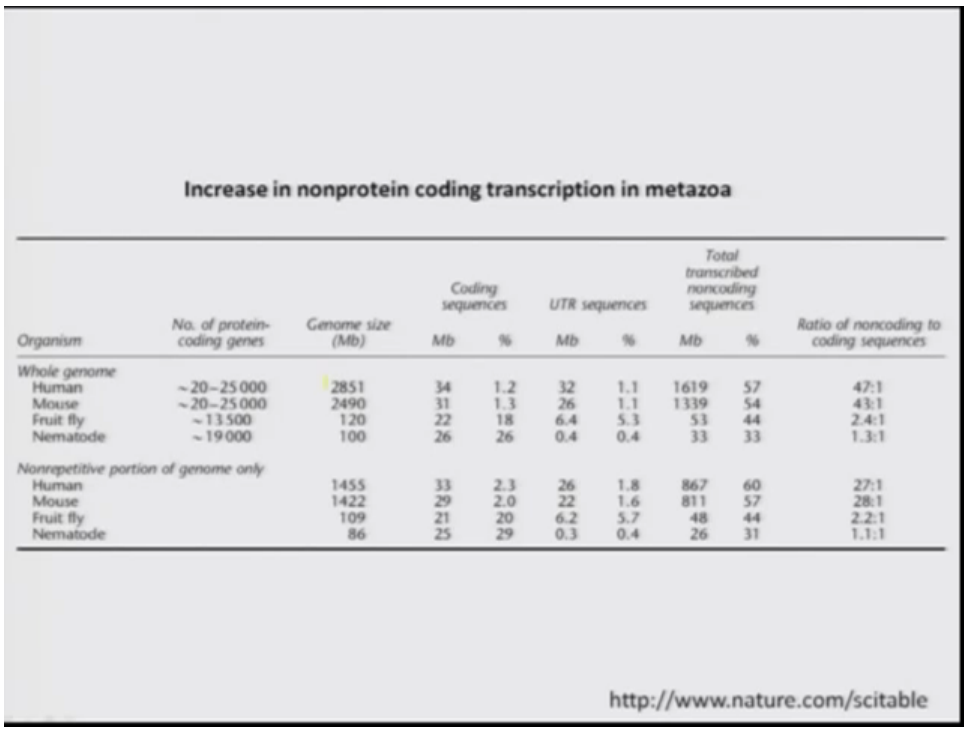
This has been the query for the scientist for a very long time, they try to investigate one of the earliest theory was that we have our genome size is larger that is why they looked at there is genome size. And this can be easily done by measuring the amount of DNA with our really knowing what combination bases and sequences so on. So that is what shown in the previous slide.

(Refer Slide Time: 3:05)



Then people thought it could be the number of genes you could have more genes as compared to other species that is what shown here so with the sequence of the genome we sought of predicted the number of genome, genes that we have, the human and for example the mouse we roughly calculate about 25000 genes that we have. It is not very different significantly as compared to for example Nematode has got of a 19000 genes and fruit fly about 13500 genes. So that is it looks like it is not really the gene number per say that makes the huge difference. And of course we have the genome size is much larger as compared to the other species.

(Refer Slide Time: 3:56)



Increase in nonprotein coding transcription in metazoa

Organism	No. of protein-coding genes	Genome size (Mb)	Coding sequences		UTR sequences		Total transcribed noncoding sequences		Ratio of noncoding to coding sequences
			Mb	%	Mb	%	Mb	%	
Whole genome									
Human	~20–25 000	2851	34	1.2	32	1.1	1619	57	47:1
Mouse	~20–25 000	2490	31	1.3	26	1.1	1339	54	43:1
Fruit fly	~13 500	120	22	18	6.4	5.3	53	44	2.4:1
Nematode	~19 000	100	26	26	0.4	0.4	33	33	1.3:1
Nonrepetitive portion of genome only									
Human		1455	33	2.3	26	1.8	867	60	27:1
Mouse		1422	29	2.0	22	1.6	811	57	28:1
Fruit fly		109	21	20	6.2	5.7	48	44	2.2:1
Nematode		86	25	29	0.3	0.4	26	31	1.1:1

http://www.nature.com/scitable

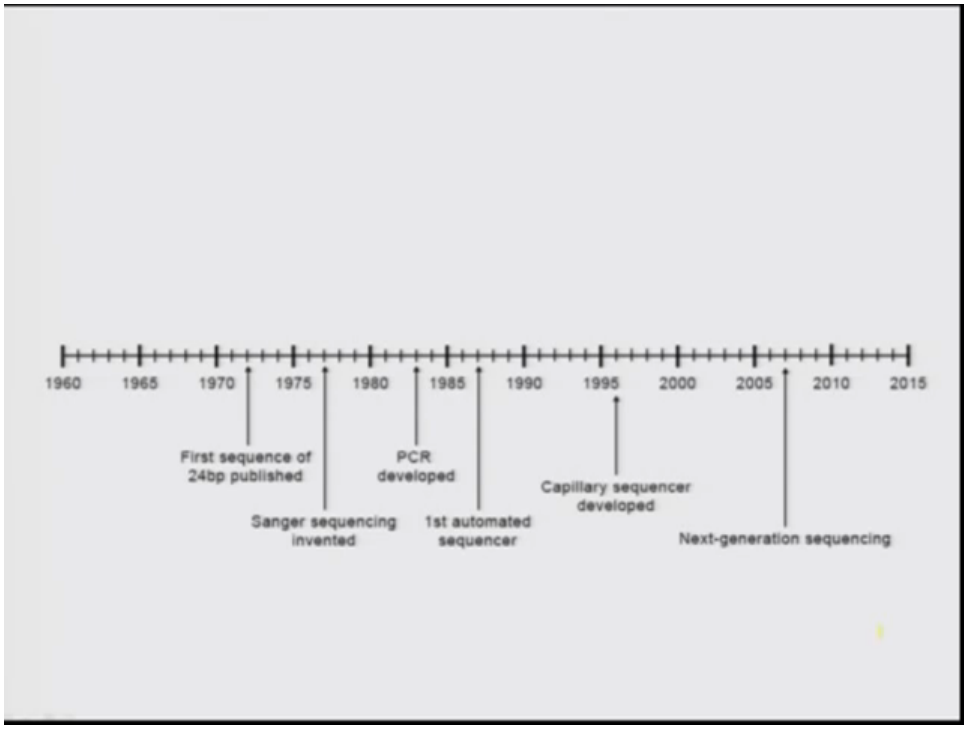
And then with other approaches that we will be talking about little later is by sequencing the transcripts now we understand that possibly the difference is the transcript. So you can see here the total the if you look into the relative proportion of the genome that is transcribed into RNA that do not have any coding potential meaning they may not really code for genes. So they are not classified as genes that really is gone up. For example it is 33 percent of the genome, that as you come up to the human is much higher and if you look into the ratio of noncoding sequence that is showed here is gone up here, right.

So it really shows that you know from Nematode of fly which is and if look into that coding versus noncoding, which is not very different for these two groups but as you come to highly about species like mouse and human, you find that complexity is gone up. So probably the difference is the genome size of course plus the proportional genomes that code for transcripts that do not code for any protein. So these are what is called as noncoding transcripts which have very critical function and some of which we had already discussed in the previous lectures. But for majority we really do not know what is that function that is something being investigated.

(Refer Slide Time: 5:22)

So what is your genome, so you know if you look into it is very different from for example microbe you have here the bacteria you have circular genome but humans you have distinct chromosomes like what is shown here 23 pairs and if you look into the chromosomes then the genes are spread out all over the chromosomes and that is something that is shown here. So let see how the genome sequence began and what are the techniques people have used and how it has change our time.

(Refer Slide Time: 5:56)



So 1972, the first sequence that is 24 base pair was published, small effort and they came up with an approach to sequence part of you know sequence and then the Sanger sequencing was invented, that has really changed the way that you know sequence you know DNA sequence was done. And you know the one of the (())(6:21) for the Sanger sequencing is you have to start with DNA that are identical, multiple copies of a given segment of the DNA then only your sequencing would be good so we will see why it is so.

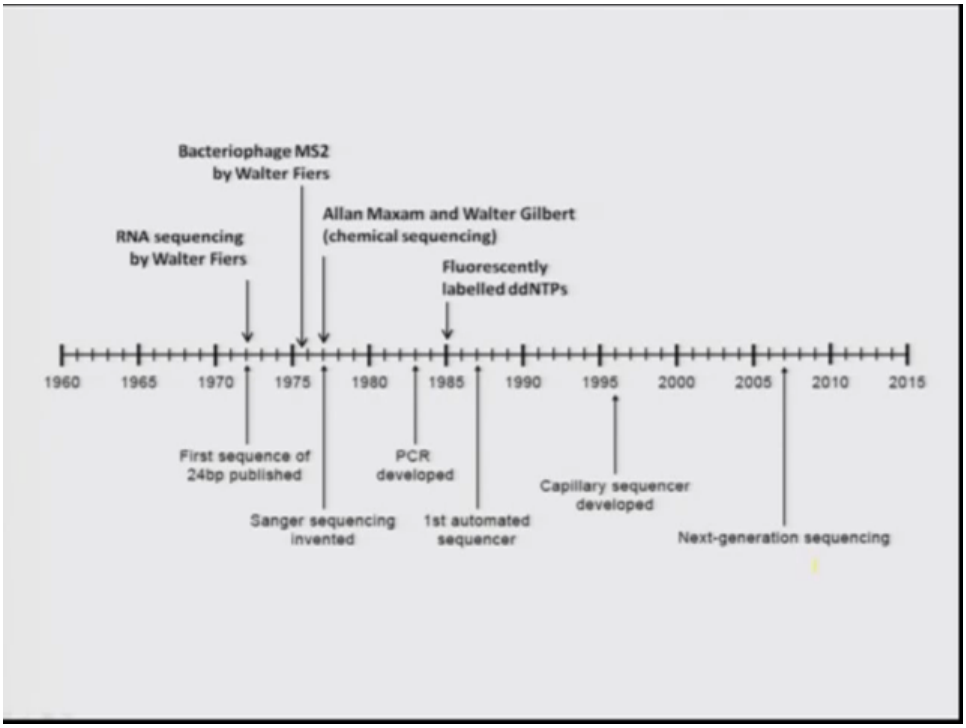
And that was a limitation because then you have to prone every segment and make multiple copies of with and then sequence. But with the invention of PCR approach it has changed the way the Sanger sequencing is done can because now we can amplify, meaning make millions of copies of any small region of the genome and this PCR product can straight away be sequence.

So that came what is called as the cycle sequencing like using PCR approach to sequence the genome. So that has led to what is called as automated DNA sequencer where you set up the reaction and leave it and the sequencer machine would run and decide further sequence and give you as a data. Now this again was something undergone one more round of you know evaluation

and revolution by 1995 instead of classic gel based sequencers we will discussed, we have what you call as a capillary so when do 96 (react) sequence runs at the same time.

But everything is now you know is changed with the introduction of what is called as Next-generation sequencing or sometime this people call as second generation because the Sanger based Sanger method is called first generation and we have now second generation which you call as next generation but now it is also a new approach coming up is called as a third generation. But as of now this is called as (())(8:01) next generation sequencing.

(Refer Slide Time: 8:19)



But it is not that the DNA sequence approach that came first because people started using RNA much before the DNA could be sequenced most often the choice was reversible RNA are virus RNA because then you have the RNA in multiple copies. These are identical sequences you have multiple copies you could sequence but that was a challenge with the DNA. So how would you isolate is mass segment and make multiple copies that sought of that main approach was possible because the cloning techniques that evolved that we discussed is called as recombinant DNA technology.

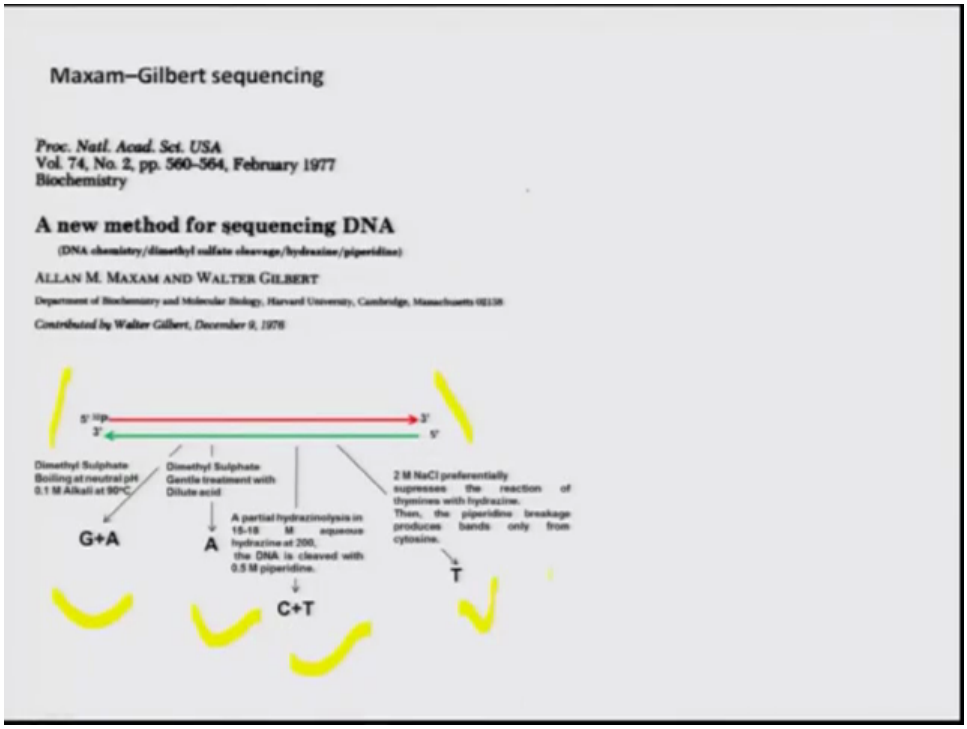
And because of the RNA sequence, chemical based sequencing was possible that was developed by Walter in the early 1973 or so this group is able to you know publish one of the Bacteriophage viral Bacteriophage which has got RNA genome, they are able to sequence and publish it in 1976. So that was the first major sequence initiative and by then you know the other people including Sanger it is the inventor this approach that is what you are going to discuss.

So per after the viral RNA sequence was completed and then Maxam and Gilbert they together used very similar approach to sequence the DNA this is called as chemical sequencing. And they use a very similar approach but by then you know you have the recombinant DNA methods have come in and therefore you are able to have a copy of the sparse segment of DNA, multiple copies you could have so you can sequence it.

But as I told you that automated DNA sequencer are possible because now we are able to fluorescently label the four bases of the DNA, all these we are talking about from here to here these approaches use radioactive mightiest attach to one of the four bases or all the four bases, then that is where you are able to see the sequence of the DNA RNA, because you are labeling the DNA or the bases.

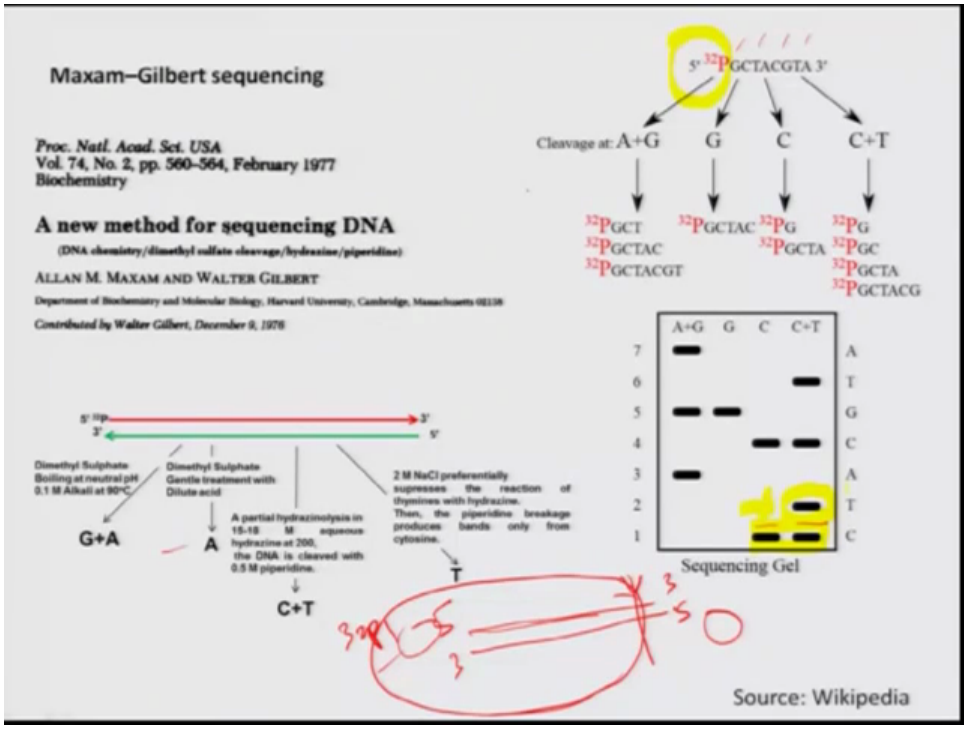
And then by 1985 know we have had you know inventions in which they developed what is called as fluorescent mightiest that are attached to the bases and that was pretty successful with the Sanger method we will see little later. And let us change the way that sequence that that is what let to what is called as the automated DNA sequencing. So this is the first paper that we talked about, the Maxam Gilbert sequencing paper published in 1977 a new method for sequencing DNA.

(Refer Slide Time: 11:19)



So they use the chemical methods that is shown here. So you basically have a short stretch of DNA and the DNA is put into for example four tubes and each in tube we are adding different chemicals with these are methods that would cut the DNA at different bases for example this method what is shown here would cleave the DNA and that has in the point where you have a GRA. Whereas this chemical method would cleave the DNA at A and this one wherever C and T and this one wherever you have T as a result now you could have sequences.

(Refer Slide Time: 11:44)



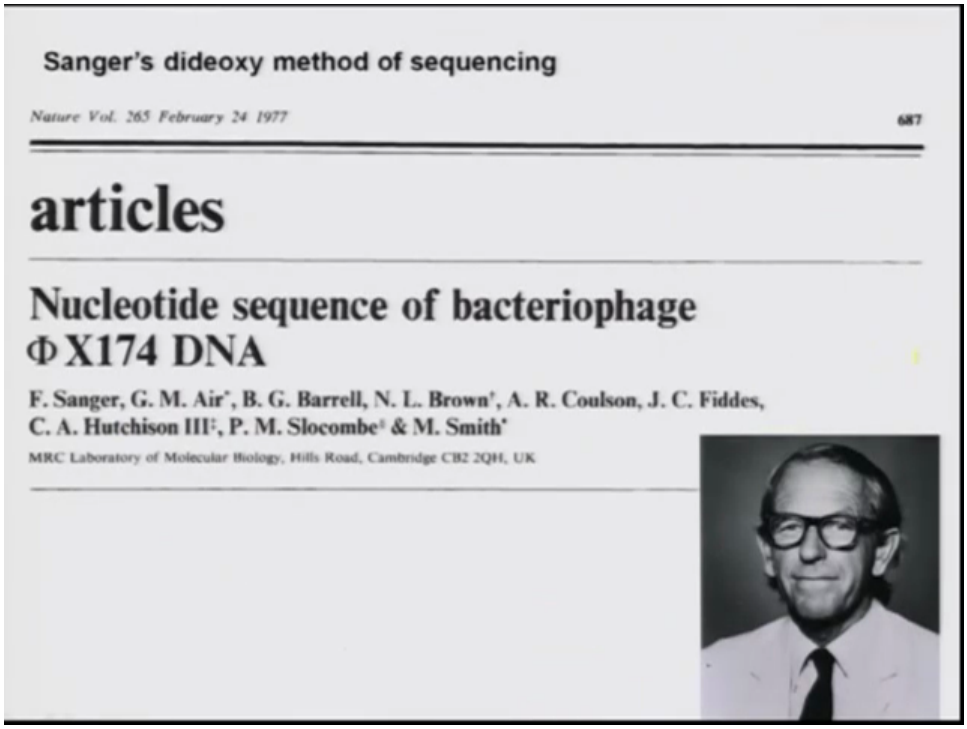
So basically what is happening is that you have the DNA that are you know labeled at one end with radioactive mightiest, for example you have a DNA and this DNA can be radioactively labeled by a method for example, this is your DNA and you have 5 prime, you have 3 prime. So you can add a phosphate (())(12:05) group here which is attached to a for example radioactive mighty by using a (())(12:11) and this would do it on the (())(12:13) but after that you can cut with the restriction enzyme and take only this part.

So in this fragment only one of the two strains is labeled with the radioactive mighty. And then you use this chemical method therefore this would cleave either here, here, here, here based on where it happens it is the random event. And therefore you know when you run a gel you are going to separate the DNA fragment based on the size and what you are looking at is only the radioactively labeled DNA using extra film, so that is what shown here.

So you by looking at the sequence you are able to you know read the looking at this bands you will able to read the sequence. For example I have a band here and here, that shows that what I am seeing here is you know C, because it is present in the C lane as well. But the one that is here present only in this lane though it is cannot be C, therefore you call it as T, again C and then

likewise you go on reading the sequence. So this is the first method for developing or understanding the sequence of the DNA.

(Refer Slide Time: 13:14)

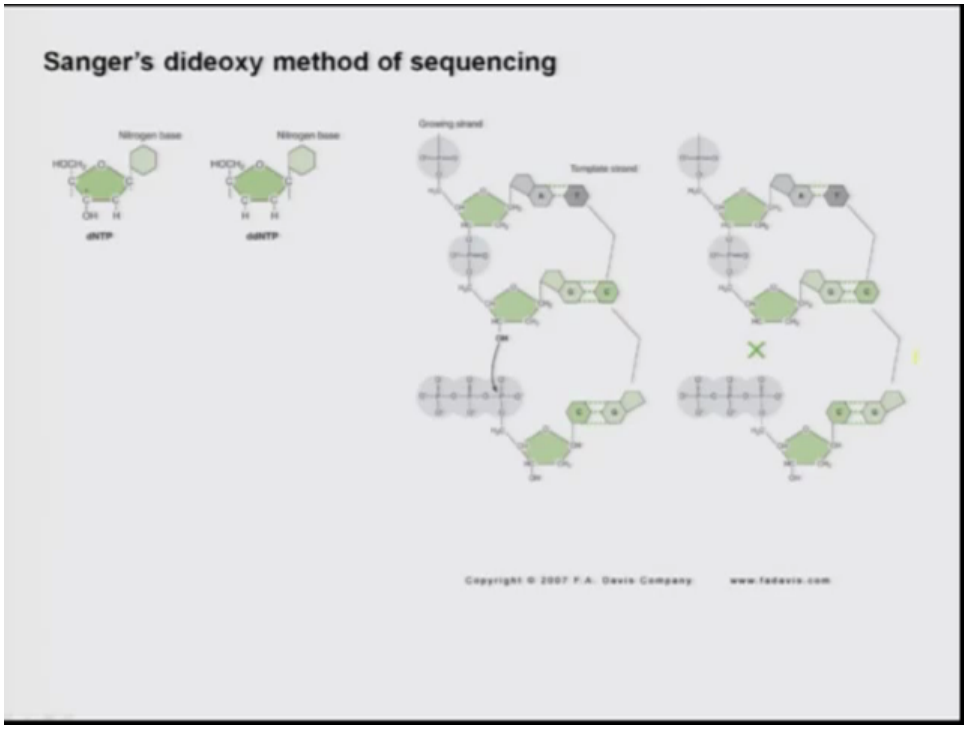


But still it was very cumbersome, so this gentlemen what you are shown on the right side is Sanger who developed a method for sequencing DNA which is based on very similar principle that ourselves use to copy the DNA, so as you must have studied for a DNA replicate you have enzyme called DNA polymerase which reads one of the single strains and keep adding the complimentary bases to make a new DNA and that is the method that you pretty much used to sequence.

So what is unique in that so (())(13:55) sought of chemically modified that base such that you know that base if it gets added the polymerase will not be able to continue the polymerization, in other words you terminate the polymerization. So these bases are added in a smaller amount therefore you have a normal base with small amount of this modified base which terminate and as a result you not terminate that polymerization at every point but in certain points. That is what the method was and in fact he was awarded noble prize for this particular invention because it

has really changed the way then has been sequence. So first paper that came in (1990) 1977 this is Nucleotide sequence of bacteriophage DNA, right.

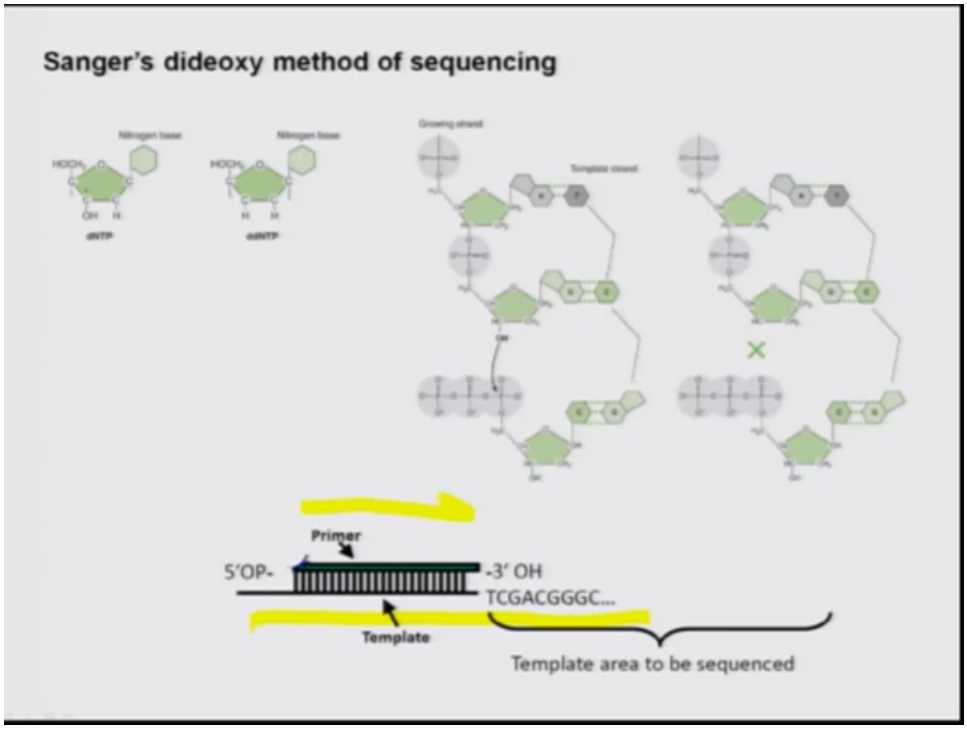
(Refer Slide Time: 14:50)



So this is called as Sangers dideoxy method or Sanger method or chain termination chemistry there are several names, normally it is called as Sanger method. So this is a normal base so you have the sugar and then the nitrogen base, so what it develop is an analog of this base which is called as dideoxy, meaning you can see here you have for the DNA this sugar you have OH as well as here but here it has been removed, right so that is what it is. So if you look into the way the DNA being made you can see here that every base is connected the way it is showed here, right you have the bonding.

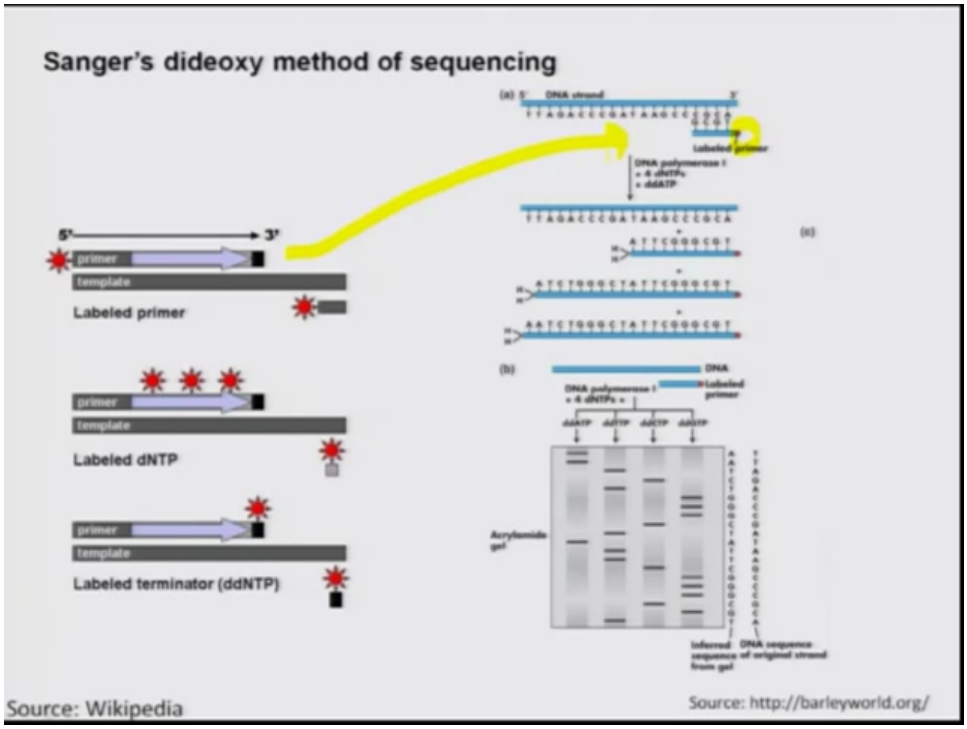
Whereas if you have a base which has got dideoxy sugar then this if you there is a polymerization, there is a DNA polymerase that help in copying the strain that is shown on the right side. Then this base would terminate the polymerization because it cannot really help in continuing the polymer because it does not have an oxygen over there. So that is the invention and that really changed the way it is done.

(Refer Slide Time: 16:13)



So it is using the same machinery that ourselves uses because you have the template DNA, so what you give is small primer when (())(16:13) that has complimentary to the known sequence, it goes and binds and then you have this DNA polymerase comes and since here when you have all the four bases available it is going to just keep on copying it until it incorporates base that is shown here, right. So if that comes here then it is gone to terminates let us see how does it happen.

(Refer Slide Time: 16:41)



So there are several methods by which you are able to approaches by which you are able to read the new DNA stand that is being made, one of them is that you label the primer itself, so what is shown here. So you have the five prime end of the primers labeled with for example radioactive mighty, therefore every new fragment that is being made you are able to detect in a gel, right.

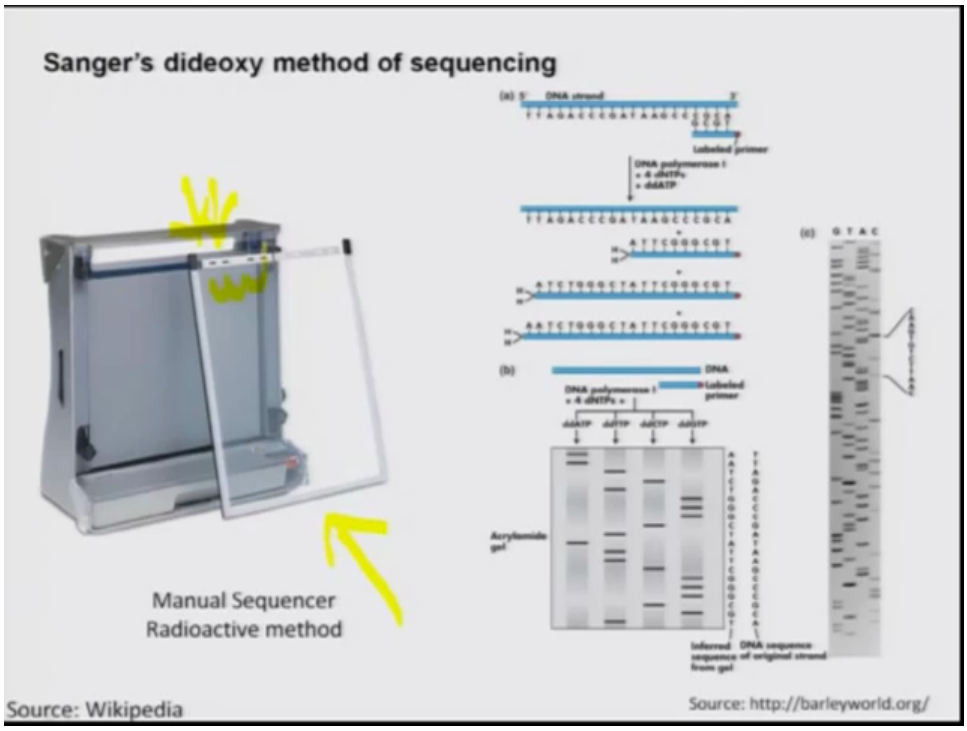
Or you could use in the reaction bases that are labeled. For example the four bases that you are using you can use them in a labeled way, so that every base that gets incorporated has got a radioactive mighty. So therefore it give you the signal when you are visualizing. The other method that has developed is that whatever the modified base that are using for termination the dideoxy one that alone is labeled.

So as in when it you know the DNA gets the growing strain gets terminated that would have the dideoxy one as well as it will have the radioactive mighty therefore you are able to visualize them on a in the gel. So these are the base that you do it and this is what the reaction is, so you have the primer which is labeled let us say that is what we are talking about this particular approach.

And then the primer goes and binds to and then you have the (RNA pol) DNA polymerase copies it and depending on when do you get this dideoxy base it would be arrested. So when you separate the fragments and a gel separate it then you could really read the sequence. So this is the smallest fragment and this is the largest fragment and you can see that this lane represent your reaction in which you have used dideoxy T or C or G or A.

So by looking at the fragment you will be able to read it, we read it as for example this is T, G, C, G, G, G, C and so on. So that is what it is shown here and this is the sequence of the DNA that is being made and this is the sequence the reverse compliment is the sequence of the original DNA that was used as a template to sequence, right. So in this way we are able to sequence the DNA and that is remarkable invention and the way we have done is (())(19:02) time.

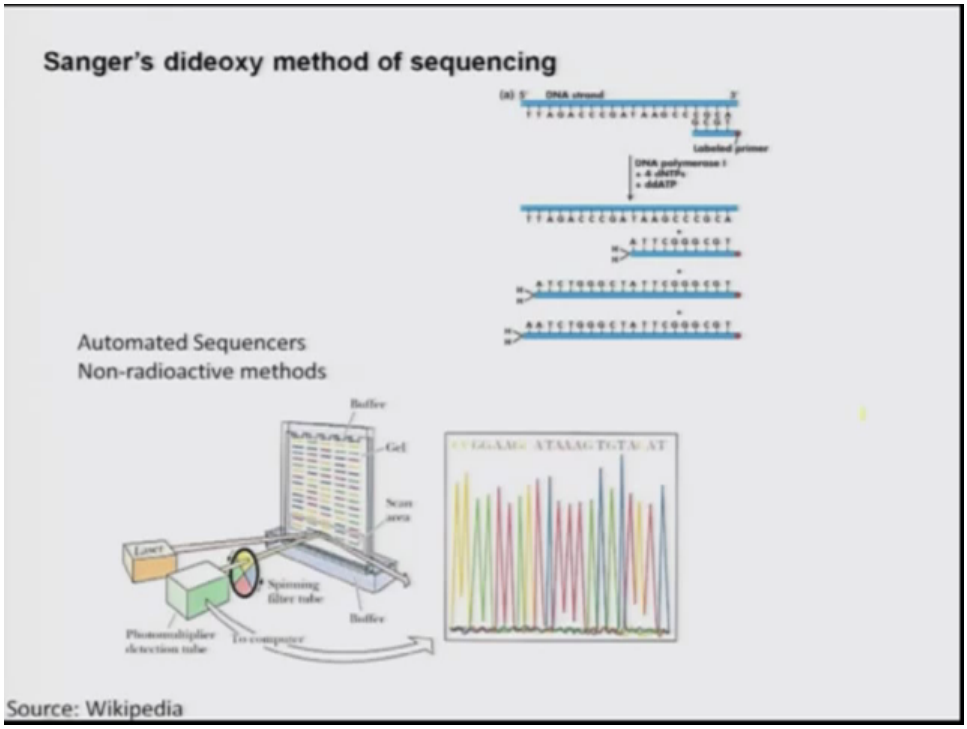
(Refer Slide Time: 19:18)



This is some of the classic you know manual sequences for radioactive gels. So you have you have to apply the samples here and as I told you there are four lanes for each DNA A, G, T, C and separate them like what is shown here and this is one such example of the gel sequence after autoradiogram after separating the DNA put extra film wherever you have the radioactive mighty

you would have this kind of black color now you are able to read it based on this, right so this is how it is done.

(Refer Slide Time: 19:42)



But things have changed as I told you in 1995, 94, different chemistry have come in one of them is to use what is called as nonradioactive method instead of using radioactive mightiest. Now the people started using fluorochromes different you know chemicals that can and when you excite with a laser then they florescent different wavelength therefore you are able distinguish one base from the other.

So in other words we can do all four reactions A, T, G, C termination anyone of the bases in single (())(20:17) and depending on what kind of florescence they emit upon excitation we will be able to call it as which base it is. So in other words what you have done is instead of loading four independent lanes for a given DNA. So now you have a gel in which each lane represent one sample and you are separating it as the DNA pass through at the end of the gel.

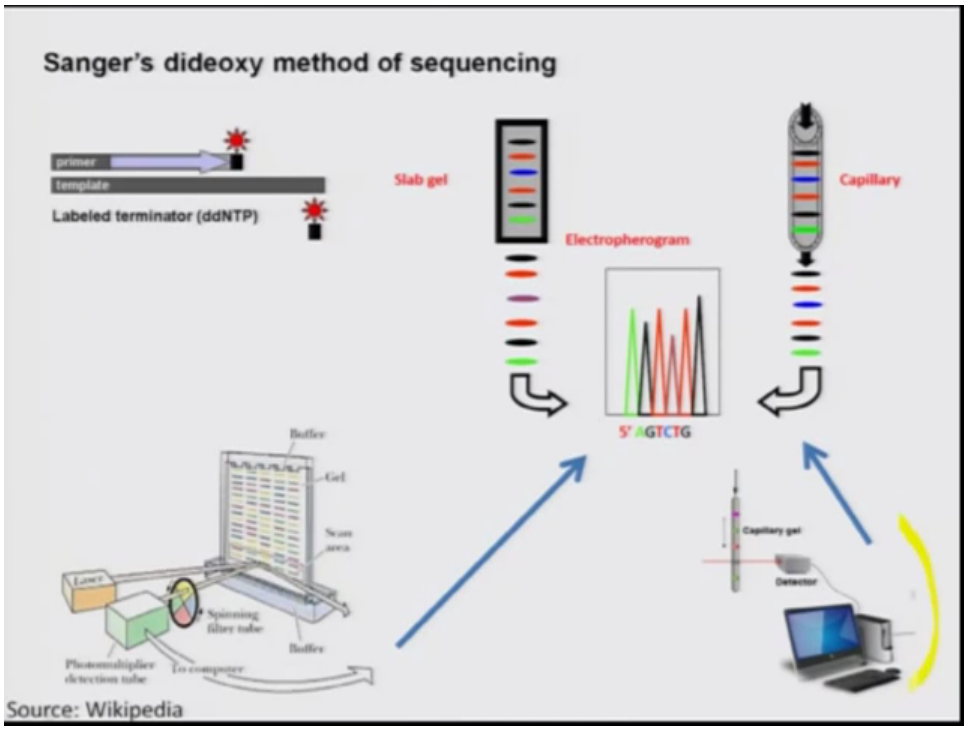
So you have a laser which excites and then you have a camera which basically measures the florescent intensity at different wavelengths. For example you have a filter which basically looks at the emission at different wavelength and depending on in which you know filter you are able

to get the signal you will be able to call it as A, T, G, R, C. So in this way you know it is taken care because you load the gel and forget about it rest the machine will take care.

So you have a camera that reads here in one lane we can do all the four reactions and at the end of the run we have this kind of electropherogram which talks about the intensity the peak intensity for given fluorochrome in a given segment and this is one base and this is another base, if both are giving maximum florescence that denotes C then the machine calls it as C and if likewise you have this sequences, right.

So this is how you are able to read the sequence it has chased. So this is the first generation automated DNA sequencers, right. So still you use gel cast gel and you have to load sample do reaction and load sample but separation and reading the sequence was automated, so it really changed.

(Refer Slide Time: 22:05)

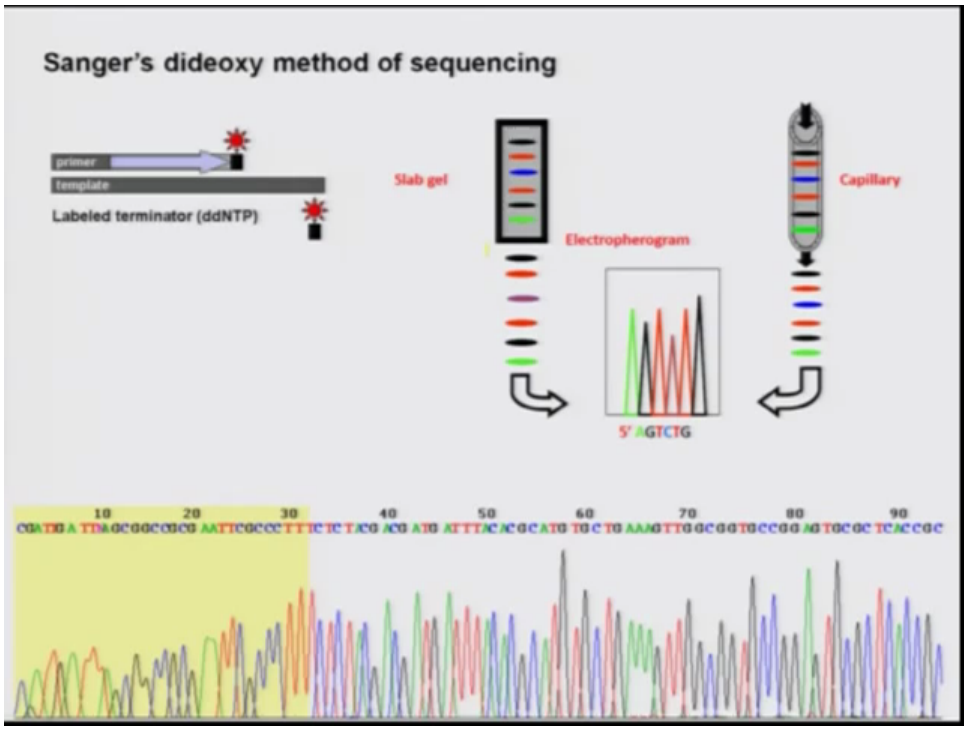


But things have gone much better, so this is a method we just now discussed that you have a slab gel which you have to cast and run the sample. But then people have gone to use different method called as capillary instead of gels you have thin wires now which have whole inside in which you can fill the gel and allow the system to pass through do the electrophoresis of the

DNA. And you know when they are coming out, you read what is the florescent that comes and you are able to read it.

So this is called as capillary DNA sequencers, something like that what is shown here so you have done away with the gels, so no longer you have to make the gel and you have thin capillaries each one now can go and take the samples, separate them and read out. So you can increase the number of capillaries, you can increase the number of sample that you would analyze not only that once the sample is you know read through then you can fresh remove the gel and fill in new set of gel and you can start going doing another reactions. So in this way we can automate it better as compared to this you know gel which normal traditional (())(23:13) of slab gels.

(Refer Slide Time: 23:27)

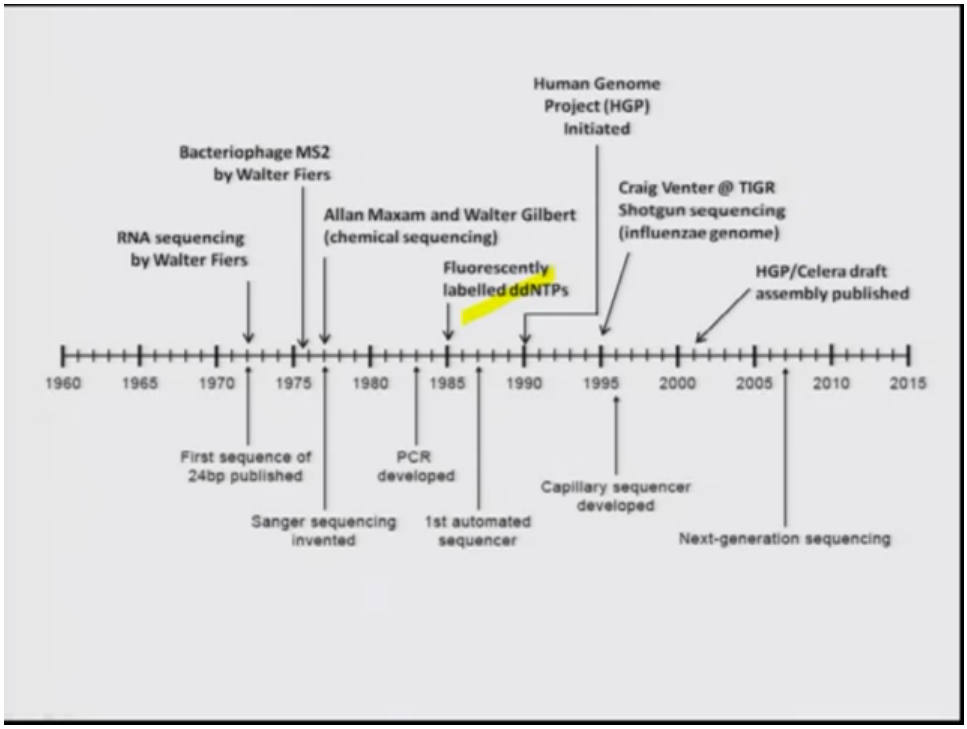


So that has really increased the throughput like number of samples that one can analyze and so on. So that has changed and in fact that has led to faster ways of sequencing the human genomics, so the capillary sequence is really helped in hastening the speed at which the sequence of the human genome was done.

(Refer Slide Time: 23:38)

This is one such example so that is the DNA capillary electrophoreses which again you know you have all the sample that been kept here and then you have capillary that is running here for the sequence to do.

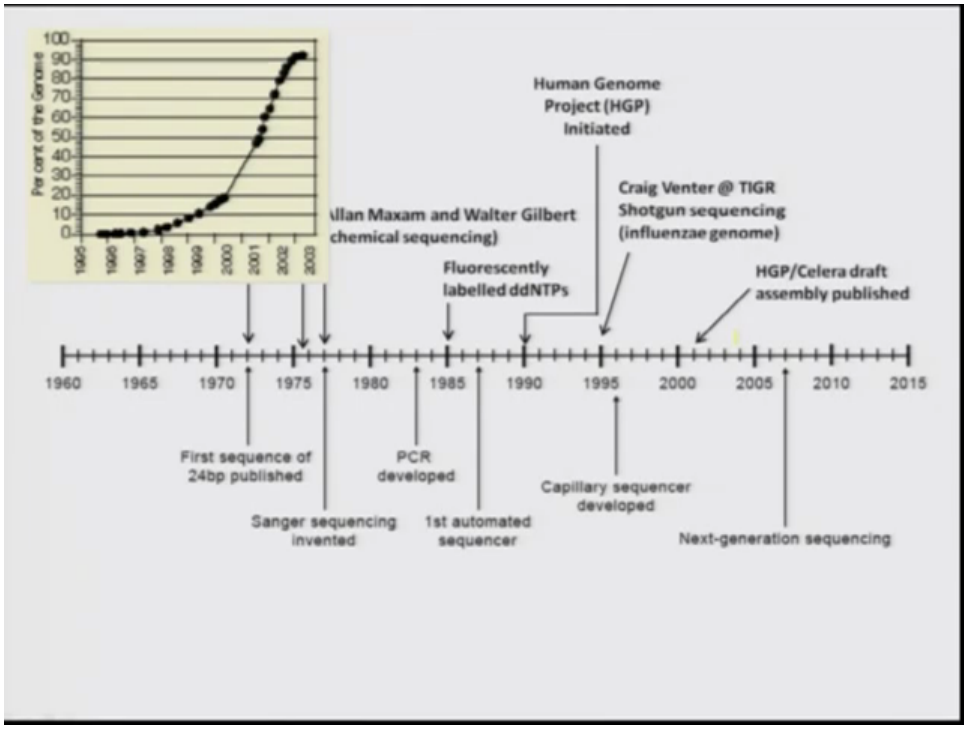
(Refer Slide Time: 23:57)



So that is what we discussed, so you know you have the fluorescently labeled dideoxy NTPs change the way for automated sequencer and then you have the capillary sequences have come and you know so it has changed. But it is not only this chemistry of sequencing that has changed even how do you prepare the DNA and how do you analyze the sequence over time to get the sequence assembly that has changed.

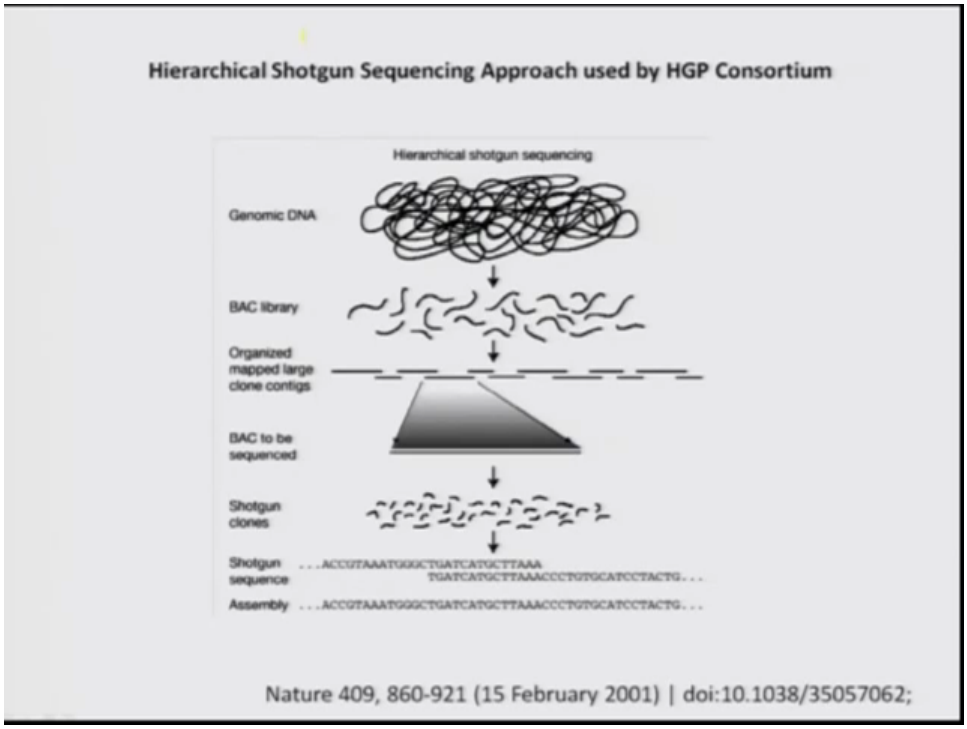
So this is one is human genome project that has used a different way of getting small segments the DNA for sequencing and how to analyze it. And then in the same time we talked about the other person call Craig Venter who developed a different method of sequencing it is called the Shotgun sequencing. He what he has done is he has used the influence genome to sequence using a Shotgun sequencing method which we will discuss about and he has then used you know same method to you know look at human genome sequencing. So the way it has been done has dramatically changed work course of time.

(Refer Slide Time: 25:06)



So as a result you can see that if you look into the percent of the human genome sequence you can see when the automated sequencers came in and when that capillaries sequencers came in 1995 somewhere around that the rate at with the genome sequence completed as dramatically changed. So that really helps to us to understand how the technology has transformed the way we look at the sequence.

(Refer Slide Time: 25:38)



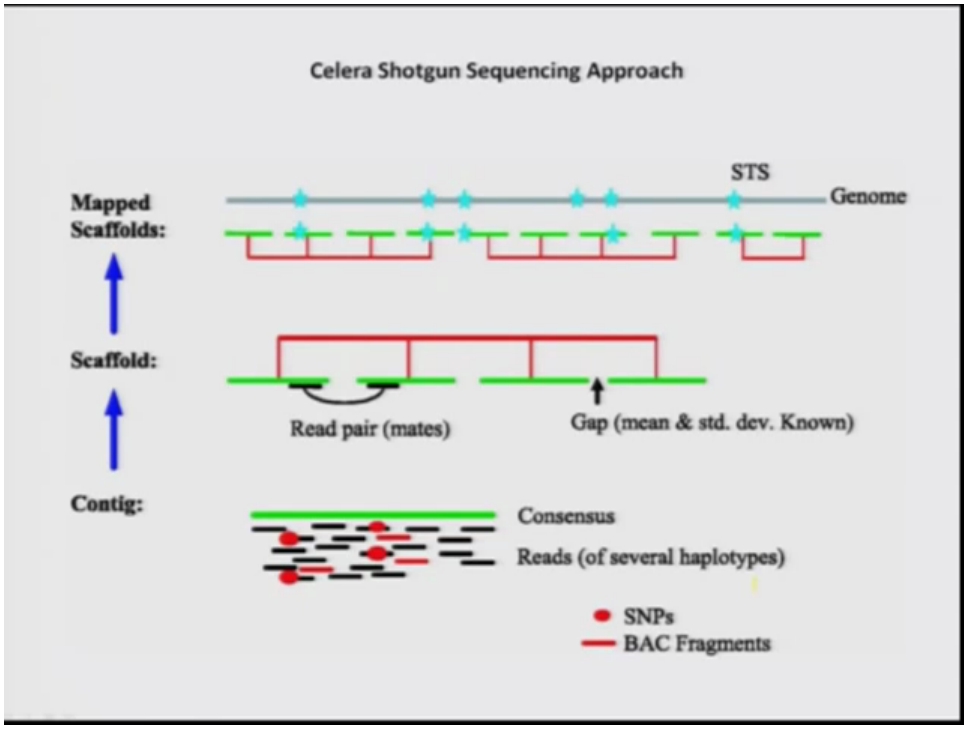
Let us look into the two different approach people have used. The human genome consortium used in approach called hierarchical shotgun sequencing. So basically what they do is that or what they have done is that they have isolated the genomic DNA and then the DNA is digested with restriction enzyme and then the fragments are cloned into vectors called as BAC which stands for Bacterial Artificial chromosome.

So they created library meaning you know the bacterial has got different segments of the human genome and different you know cell, right. And after that you use, you take one of the BACs and then stream for the library and you are able to identify the BAC clones that have overlapping in a DNA segment. So we are able to get different pieces like puzzle piece that overlap with each other, we are able to first make what is called as contig, right. This has got you know overlapping regions.

And then each clone was digested cut into smaller pieces and then they were sequenced and what we have done is we have each segment that was sequenced now use try to align them get a larger sequence and then go and align with the other and they are able to assemble this, ok. So this is a

method that the human genome project consortium used and it was time consuming because you have to create a library.

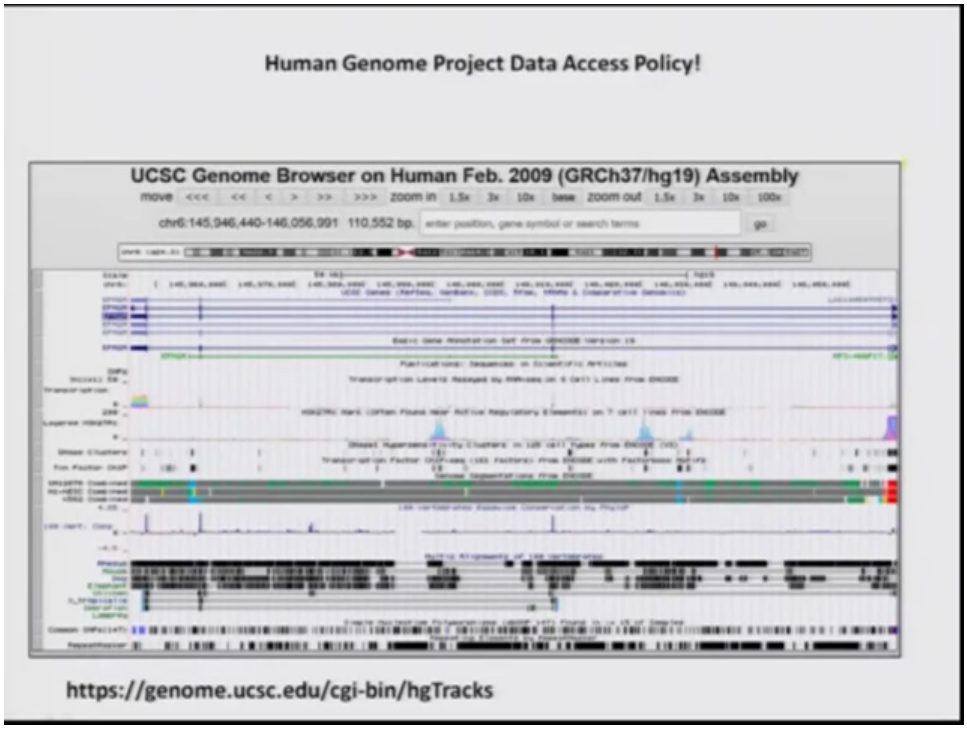
(Refer Slide Time: 27:08)



But the Celera the private organization that sequence the human genome use a different model this called the shotgun sequencing approach wherein they basically fragmented the DNA into smaller pieces, they went on reading each base the sequence so and then based on the sequence they use a powerful computing tool to identify the sequence that overlap and they get the consensus and then you try to you know use different methods to sought of identify sequence that could match each other and make a much larger sequence.

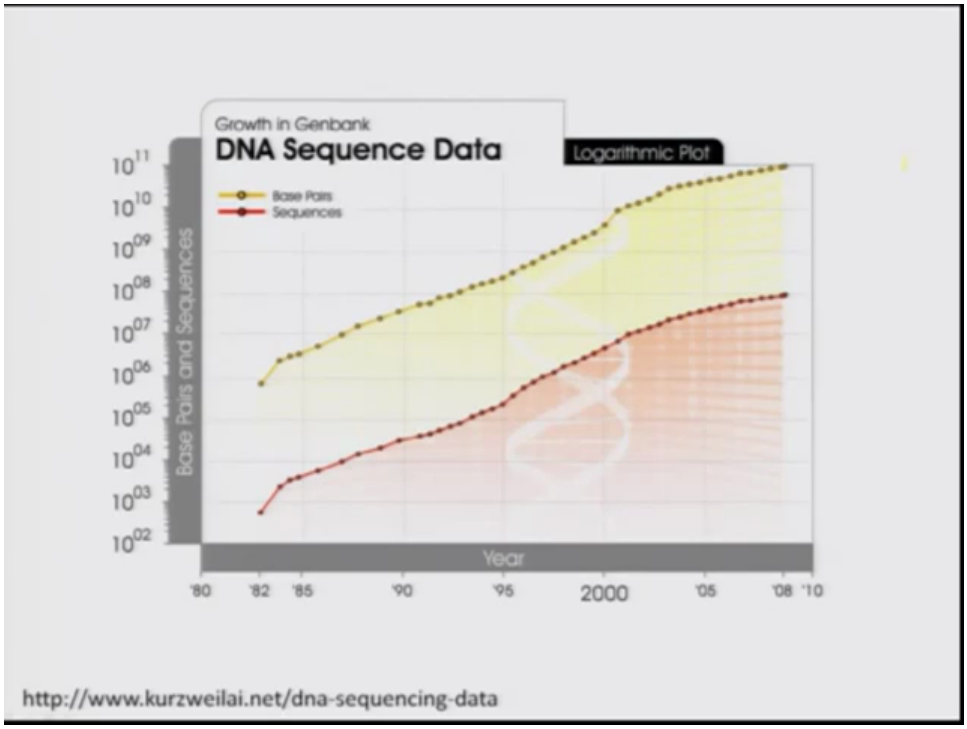
What they have done is there are several gaps in their approach but they have used the reference sequence that were given by the human genome consortium, they used it and quickly they filled in the gaps and they are able to complete the sequence based on the publicly available data so that is the other approach that they have used.

(Refer Slide Time: 28:02)



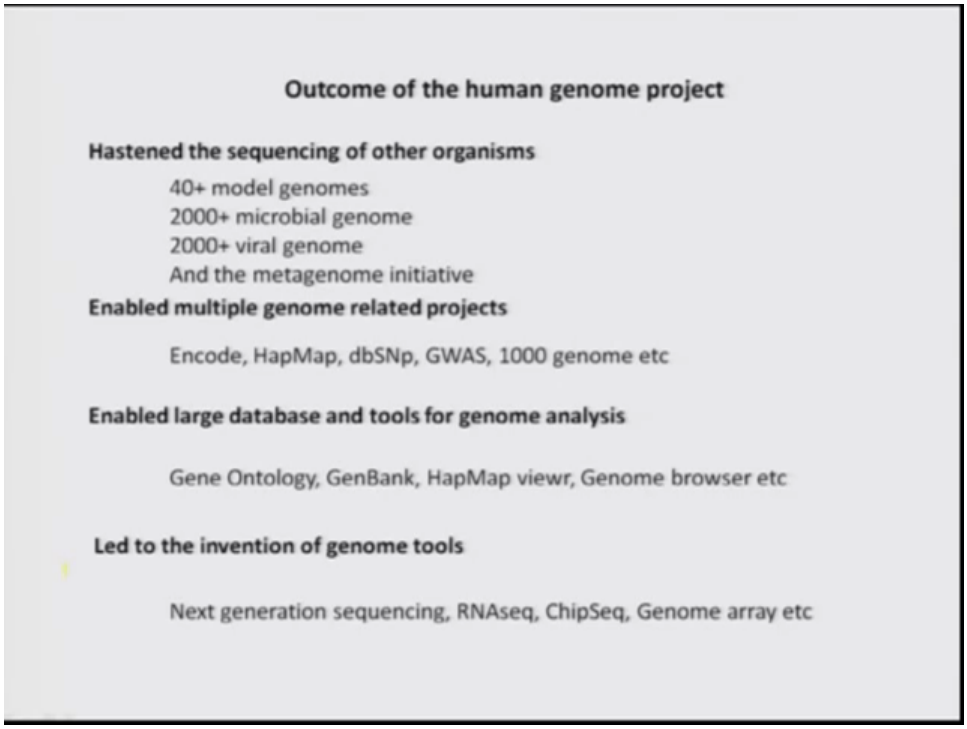
So based on that you know you have all these sequence now so now we can look into where is the transcribed region, where is the region that course for the protein, where is the region that are regulatory. The different element that we have seen in the previous lecture all are derived from this kind of approach, right. So that is how it led to you know this kind of genome browsers right which gives you all the information.

(Refer Slide Time: 28:25)



Expectedly with you know we can see that from 80s to 95, you know there is a dramatic increase in the sequence and there are deposited in the gene bank and ofcourse with the in the last one decade it is really changed, what is show is the log scale you understand is going to exponentially is increased in the what being deposited, right so that really helps us.

(Refer Slide Time: 28:54)



So what is the outcome of the human genome project, use it simply kind of a mission that completed showing that we can do something or it really made any difference to us. So it hastened the sequencing of other organisms, now we are able to we have you know his is somewhat out dated we have 40 plus model genome sequenced, more than 2000 microbial genome sequence, 2000 viral genome sequence and you know large initiative we discussed in the previous lecture what is called as metagenomics initiative where you try to understand that genome of all the microbe that are around us, including our digestive system.

So this helps us to understand the relations, the conservation and the gene functions and so on. And enabled multiple genome related projects, for example the Encode, HapMap. For example human genome variation projects, the genome wide association studies to identify genetic risk factors and understand the genome diversity across human population all these things have happened because of the technologies developed, because of the understanding that we have all the genome.

And then obviously one of the important outcome is that the databases we have Gene Ontology, if you have identified the gene then you are quickly able to understand the function that we

derive from other model systems we have GenBank, Genome Browsers and so on. And of course you know the new generation of techniques, right that is what one of the greatest outcome, development of what is called as next generation sequencing, RNA sequencing, ChipSeq and many other that we have already discussed now.