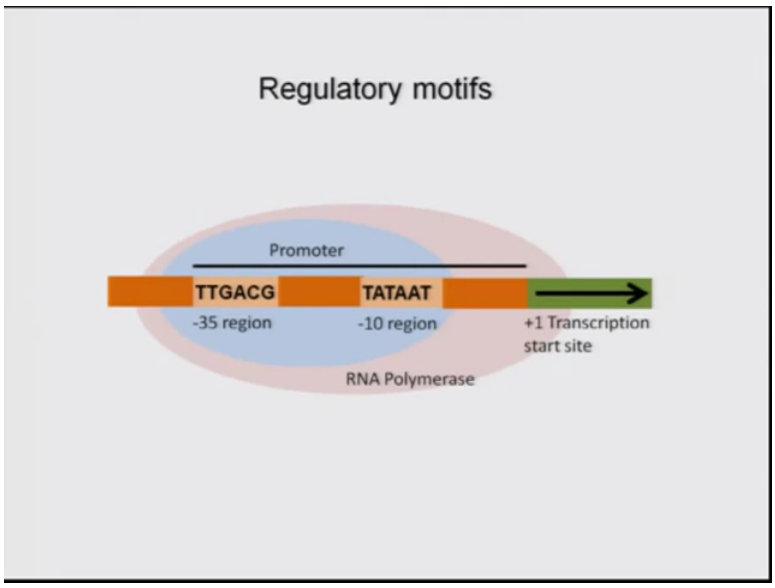
Functional Genomics
Professor S Ganesh
Department of Biological Sciences & Bioengineering
Indian Institute of Technology Kanpur
Lecture No 14
Tutorial Part 2

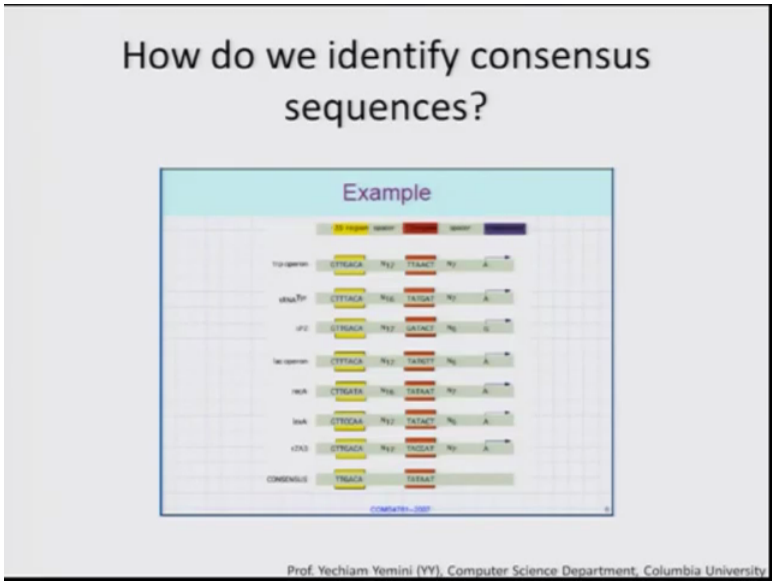
Hello everyone, I am Anshika your teaching assistance for the functional genomics course and today I will be taking a tutorial lecture in which I will be telling you about some of the interesting tools that we use to study genomics.

(Refer Slide Time: 0:34)



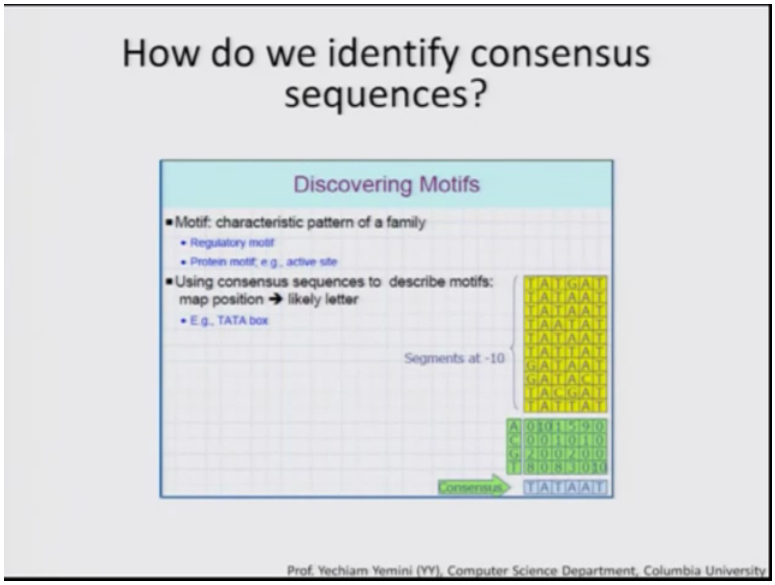
I will start with this slide in which I am show you schematics of a transcription cassette. So we all know that RNA polymerase binds to the promoter region and transcribe genes and to bind to the promoters they scan through the, they look for the consensus regions which are present upstream of the promoter so as to locate the promoter. We have been reading all these facts in our textbooks but ever wondered how these consensus regions were identified in the first place. So it was a bit experimental techniques and mutation studies at these consensus regions were identified but are these why are they called consensus because consensus means these regions should be conserved throughout species and how was that done, how was how are these consensus regions mapped across the species whether they are conserved or not.

(Refer Slide Time: 1:31)



That was done through DNA sequences, oborin sequences were compared massively compared across the across different species. Here I am showing an example of prokaryotic operons such as the tryptophan operons or the lactose operons or the lambda the bacteria fudge some of the bacteria fudge genes which are involved in those operon systems so comparing the DNA sequences upstream of the transcription start site revealed that they were several conserved regions in that region of the gene and these were it was then believed that these regions which are conserved are indeed the consensus region which helps the polymerase to locate the promoters.

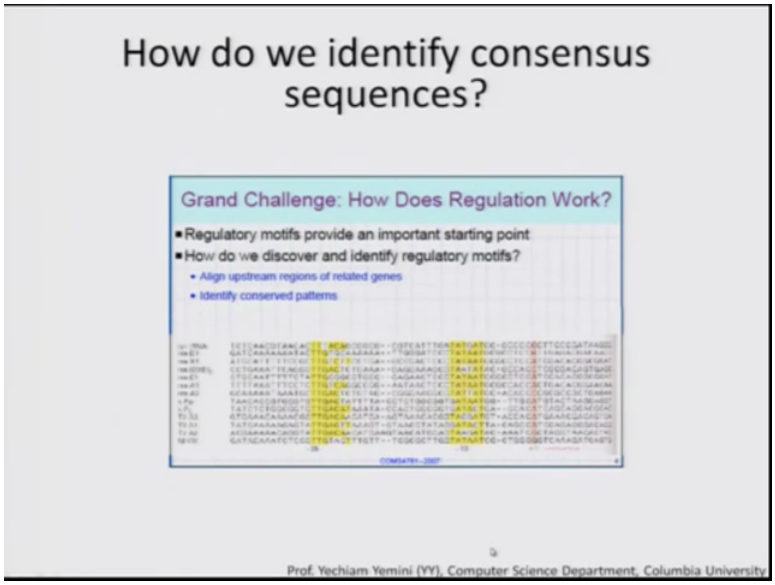
(Refer Slide Time: 2:24)



So these conserved regions could also be called as motives which are which form the characteristics of the protein family and I would call them these are called as regulatory motives or in a DNA sequence or protein motives which even form the active sites where for docking of ligands. How are these conserved sequences across the species are mapped like what was done after it was located that there is some conservation indeed present.

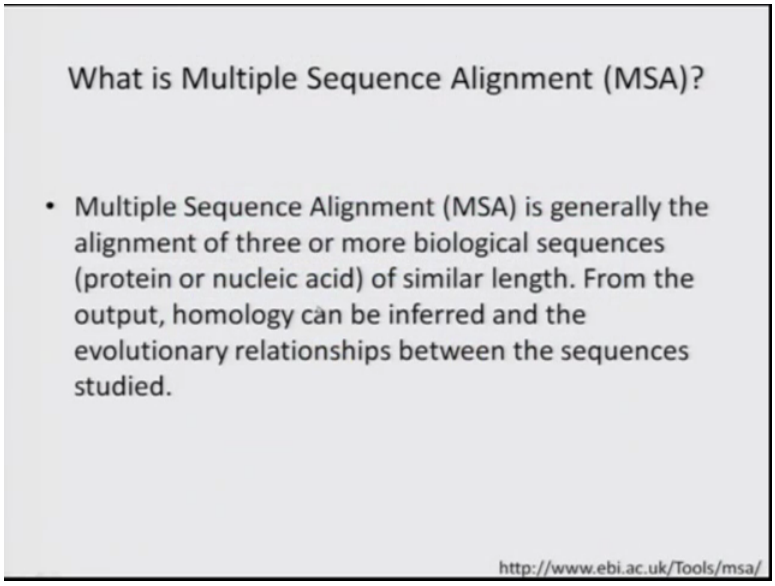
In case of DNA sequences the letters of the nucleotides ATGC were counted across the sequence and the one which is present as maximum was thought to be the most important residue at that region and was thought to be conserved throughout evolution and hence matrices were built by scoring by these scoring and these kind of concepts algorithms were built for the to develop tools which would be used for massively for comparing thousands and thousands of sequences across different species.

(Refer Slide Time: 3:21)



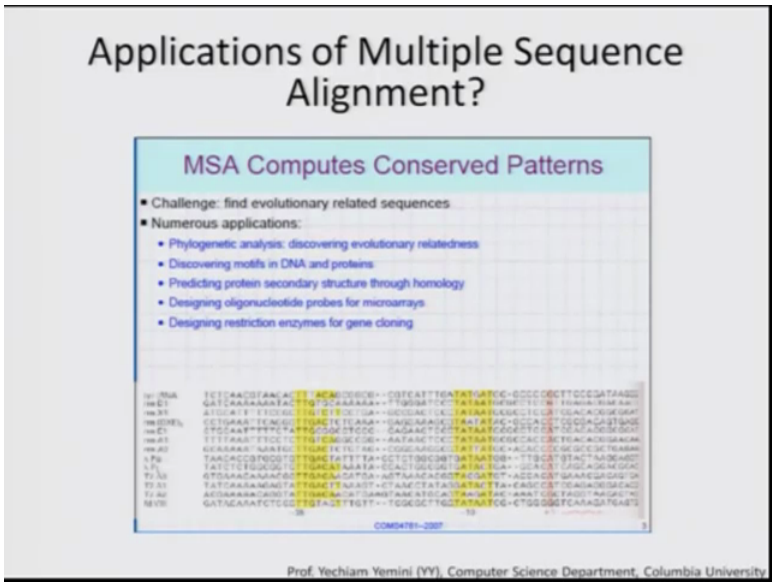
So this is a an example a diagram which I am showing you which shows different DNA sequences which are matched and the yellow colour here close you the conserved regions across these different sequences, so like I told you regulatory motives or these conserve sequence they provide important starting points. These were discovered by aligning the upstream regions of related genes before the transcription upside and hence and this is how the conserved regions were identified.

(Refer Slide Time: 3:52)



Tools such as multiple sequence alignment were thus developed on these lines to find such conserved regions. So multiple sequence alignment is generally align. This is the classical definition as stated by the European Bio-informatics institution EBI you can go to this website and learn more about MSA. So it is generally the alignment of three or more biological sequences proteins or nucleic acid DNA or RNA of ideally the same length and from the output the homology can be inferred and thus the evolutionary relationships between the sequences can be studied such as building of a phylogenetic tree which I will tell you later in my talk.

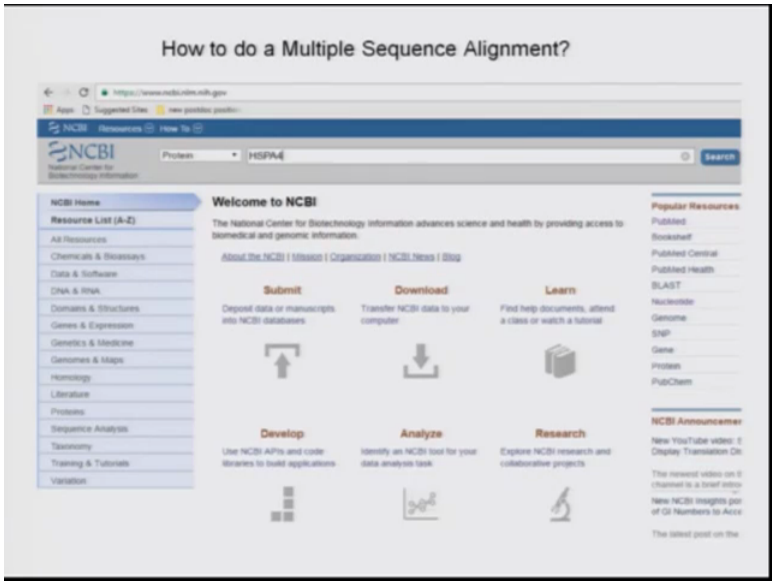
(Refer Slide Time: 4:34)



Multiple sequence apart from phylogene multiple sequence alignment has various applications. Phylogenetic analysis can be done for discovering evolutionary relatedness that which species are evolutionary more closer than discovering motives in DNA and proteins that which could be the motives and proteins which could be the docking site for ligands predicting a secondary structure of protein through homology modelling which could help again help in recognition of different sites. Which could be very import for that protein for the function of that protein?

Designing oligo nucleotides probes for microarrays now this is important because microarrays they can be used across different species which are closer to each other so designing such chips which has a region, which has a probe which can wind to more than one species is of important also for gene cloning finding out restrictions enzymes, restriction sites which are conserved across species is can be of importance.

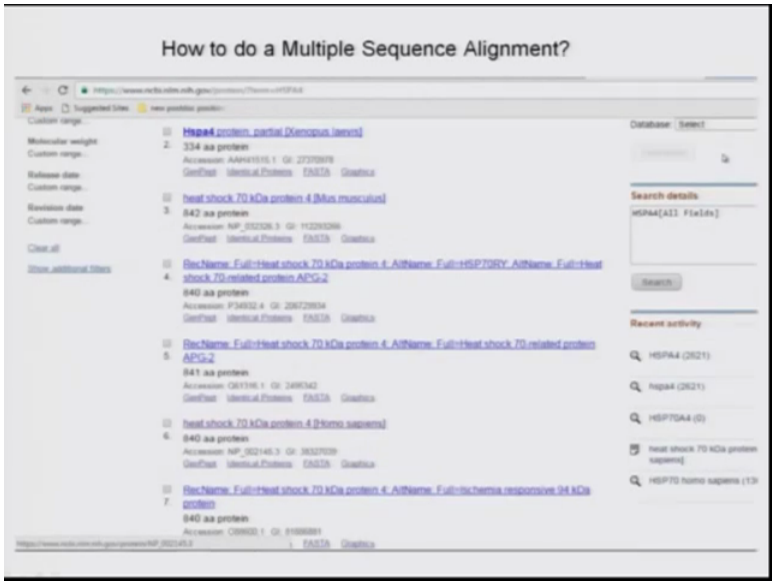
(Refer Slide Time: 5:44)

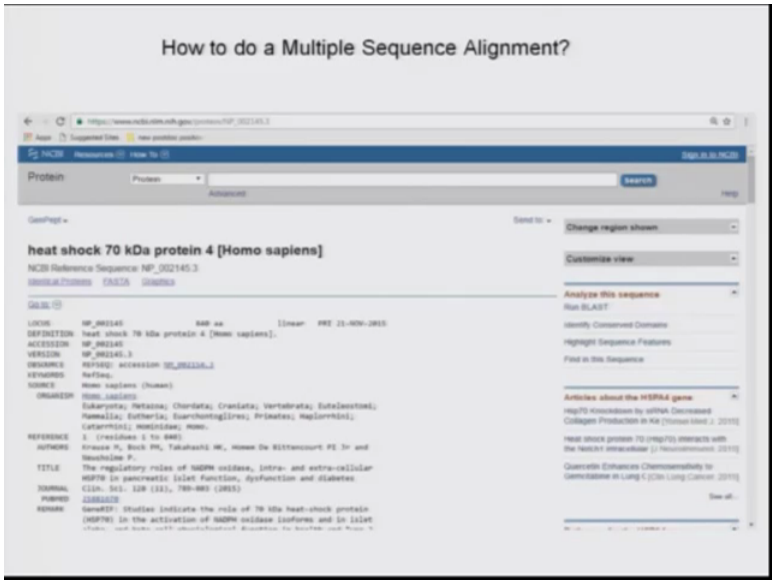


And let me tell like how to do a multiple sequence alignment so we are all familiar with this website NCBI.NLM.NIH.GOV it has a data base of all the sequences the DNA –RNA protein sequences available and I am going to look for the evolutionary relationship or conserved regions across the genes which I have been working on. It is called the heat shock protein and it is expressed it is elevated in the cells in case of a stress condition such as a bacterial or viral infection to help the cell combat that condition.

So heat shock proteins are know now they are known to be highly conserved through a different species and this is one of the reasons I took it in my study over here so HSPA4 is the heat shock protein that will be aligning so across different species, so to start with you can type the name of your gene, whichever gene you are interested in aligning in this column and then from the drop down you can select protein and click on search.

(Refer Slide Time: 6:54)



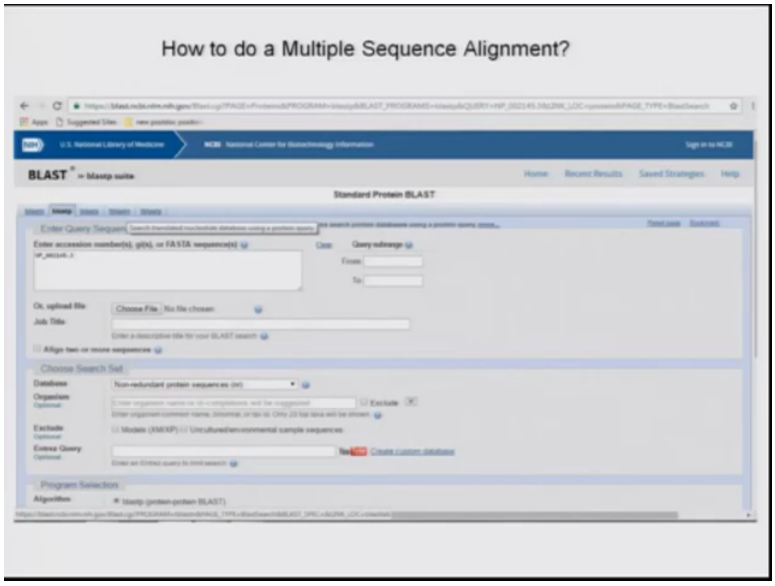


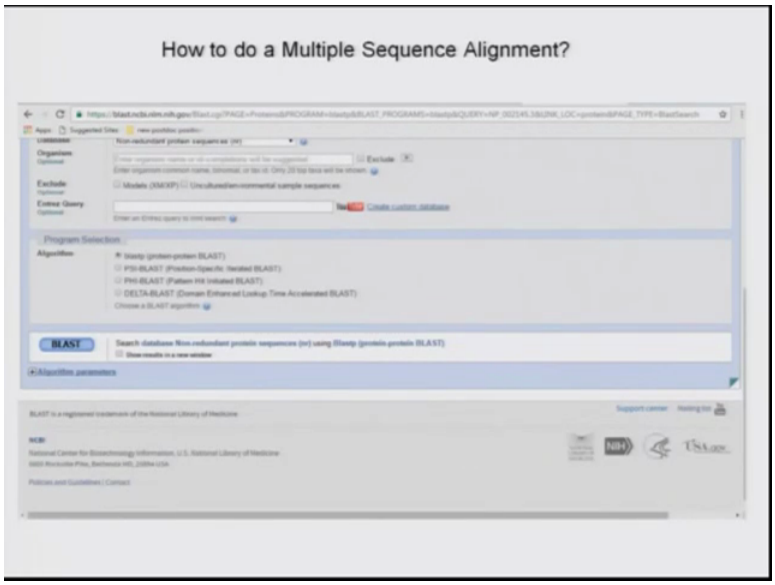
So it will give you different HSPs that are across that are across different species so I thought this analysis i am starting with the homosapiens HSP70A4 gene, so you just click on HSP70 homo sapiens A4 it will give you all the details of HSPA4 and as you browse down it will tell you more about the protein the what kind of protein it is and what kind of domains it has, NCBI will tell you all about it. You can browse down in read about it.

And here on the right side you can see this option as run blast now our first aim is to retrieve the sequences of HSPA4 from different species and then only we can subject it to a tool which can

look for conservation across the species. So when we click when we do a blast then it will retrieve the HSP70 sequences across different species.

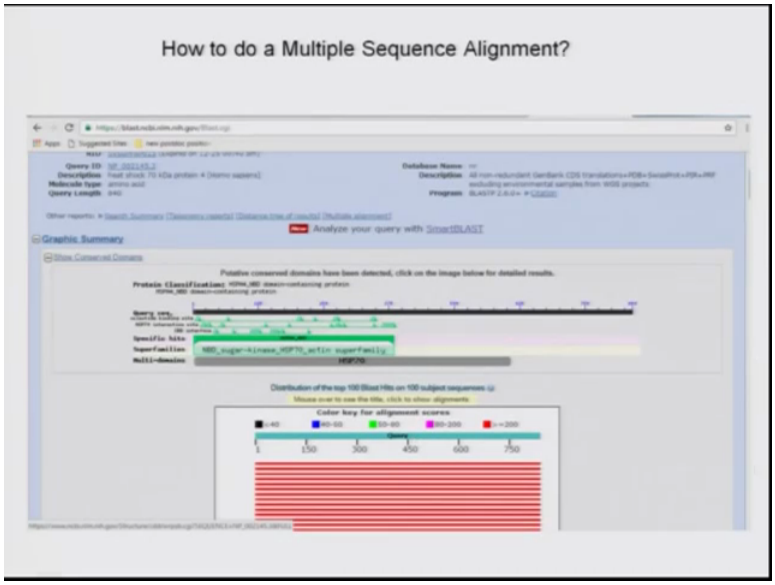
(Refer Slide Time: 8:05)





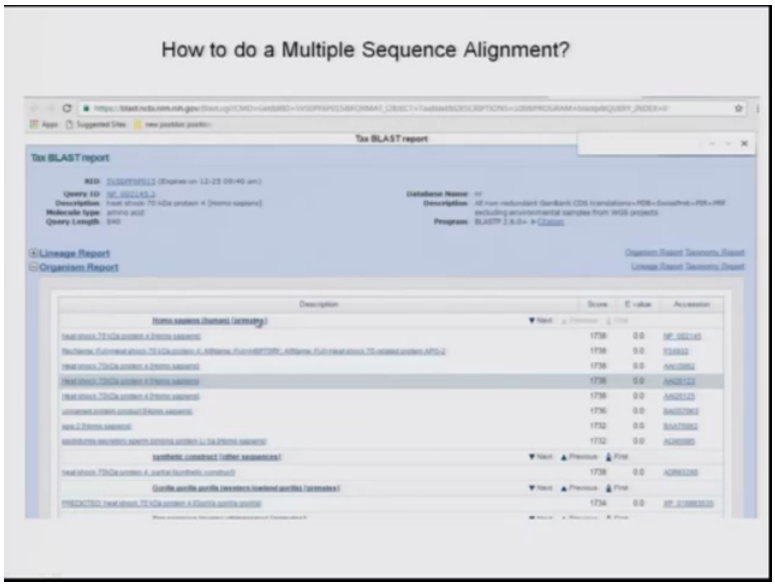
Click on run blast and what happens is that there is this NCBI tool called blast which you will read in another lecture taken by Nagma in so blast it will you just accession number of the protein will automatically appear over here and then you can simply go down without changing any options just click the blast button and this window shows you a blast running.

(Refer Slide Time: 8:27)



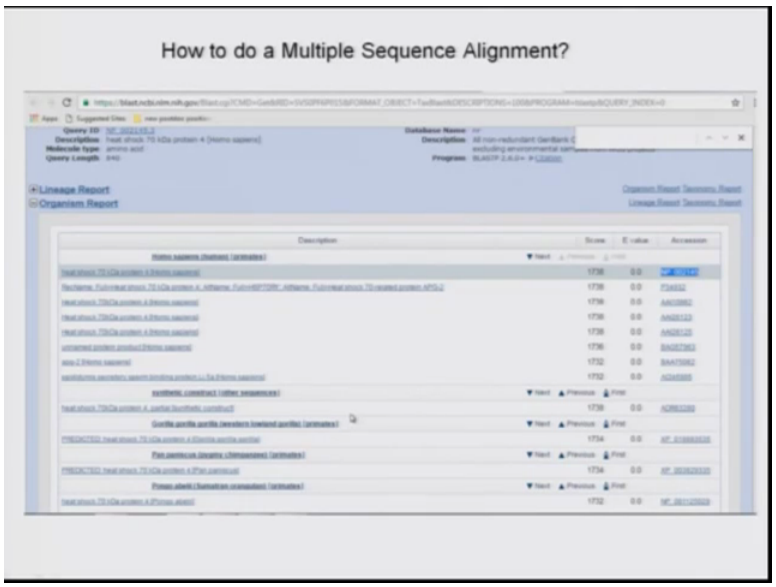
So this is the output of the blast for HSPA4 gene. So to to go directly to the different organism sequences of this HSP we click on the taxonomy reports.

(Refer Slide Time: 8:47)



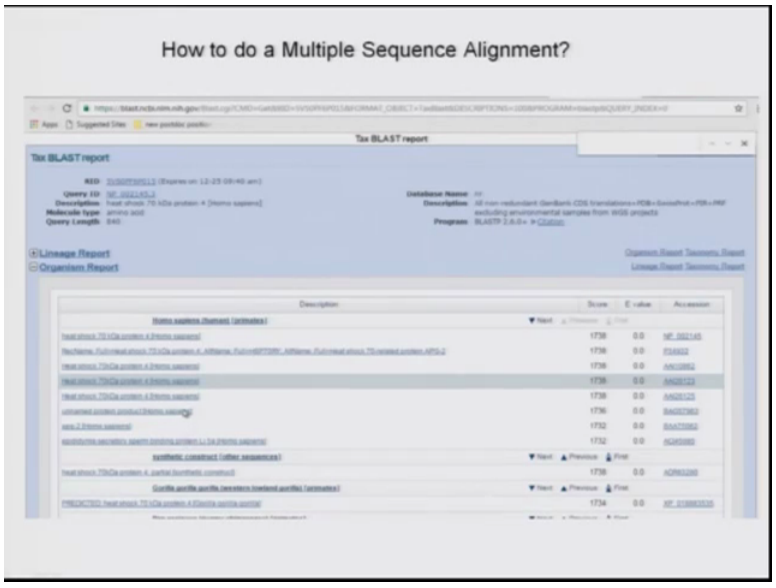
And here below the organism report window option we get this window which shows as you browse down you will see that this for homosapiens and then this is for gorilla so as you browse down you will see the HSP protein sequences listed for different organisms.

(Refer Slide Time: 9:07)



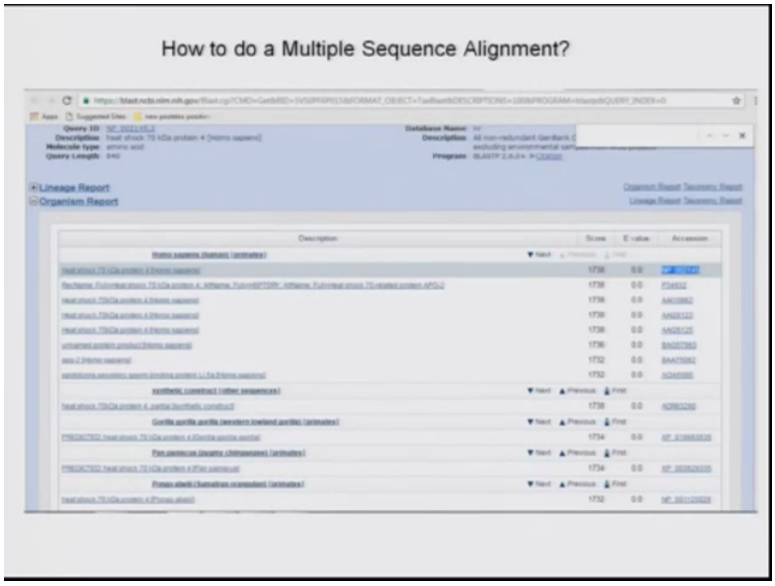
So as we browse down we can see that the HSP sequences for different sequence for different organisms like gorilla and chimpanzee or orangutan, so you can simply select the organisms for which you can you want to do homology search and you can simply select those organisms how its a good idea to copy the accession numbers of those protein sequences from here.

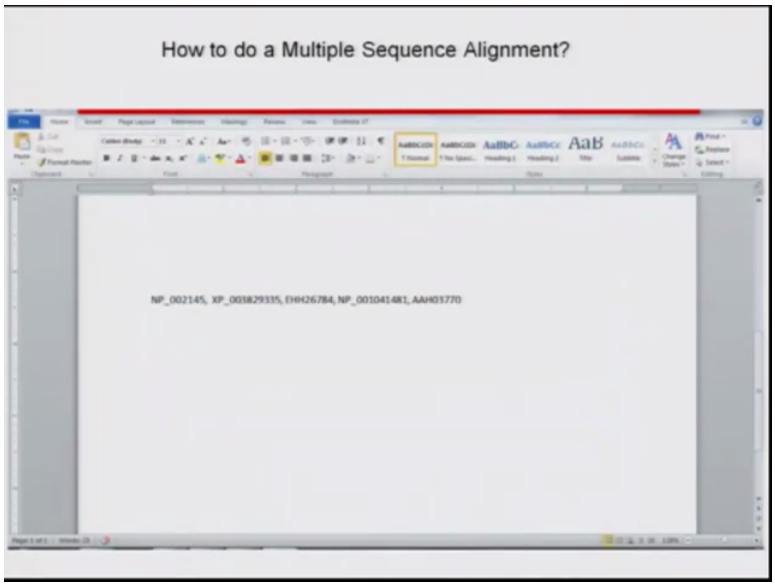
(Refer Slide Time: 9:37)



So like you see in the previous slide there are below homo sapiens there are lots of hits generally the first hit is the one with maximum similarity so in case you are interested in a organism we go for the first hit and copy the accession number.

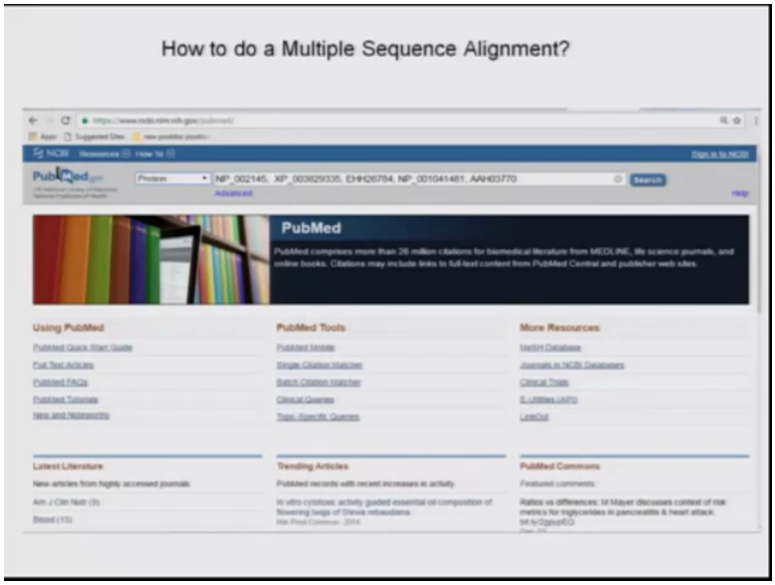
(Refer Slide Time: 9:51)





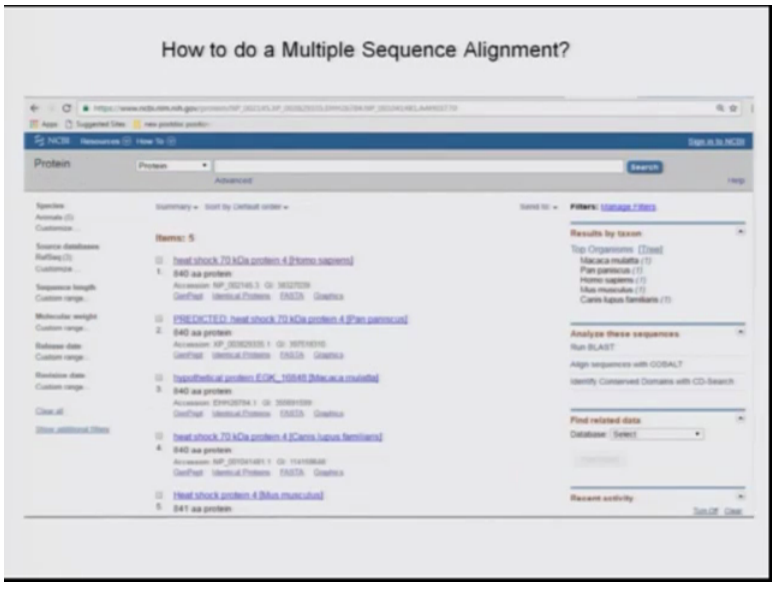
So likewise for all the organisms we are interested in we copy all the accession numbers and paste it on a word file just like I did it over here.

(Refer Slide Time: 10:05)



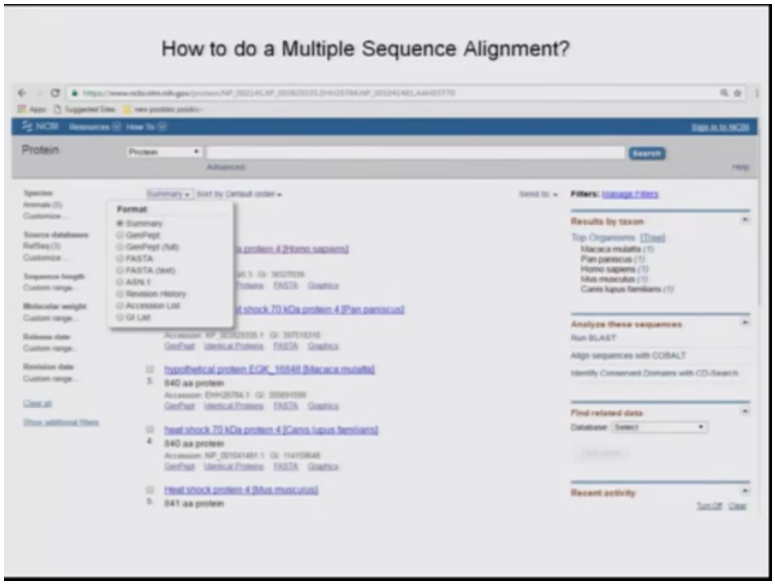
We copy the accession numbers and then copy back the accession number from here and put it on the home page of the first page that I showed you of NCBI website and from the drop down again we select protein again search now what this will do?

(Refer Slide Time: 10:14)



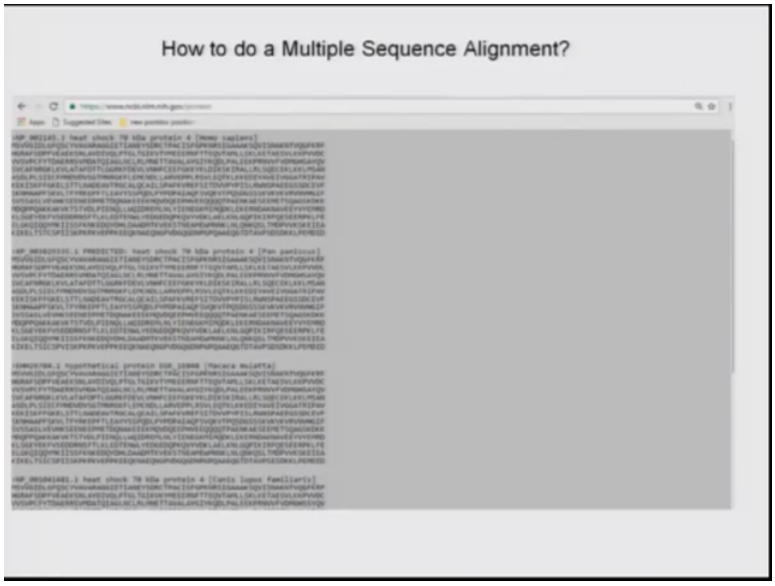
It will give you the sequences of just those proteins which you selected no other organisms.

(Refer Slide Time: 10:41)



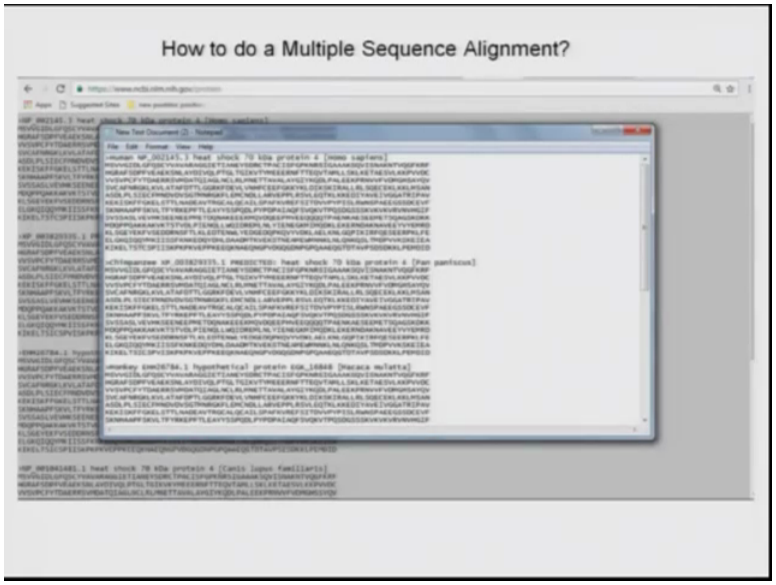
Now what we want is we want the protein sequence of these organisms which we have sorted out in a faster format so for that we click on summary and you see this window pops up where there are two options fasta and faster text. So fasta text gives you specifically the sequence of the protein with no other information which will be easier for you to copy paste and retrieve the sequence, so we do that we click on fasta text and this is the kind of windows that appears.

(Refer Slide Time: 10:53)



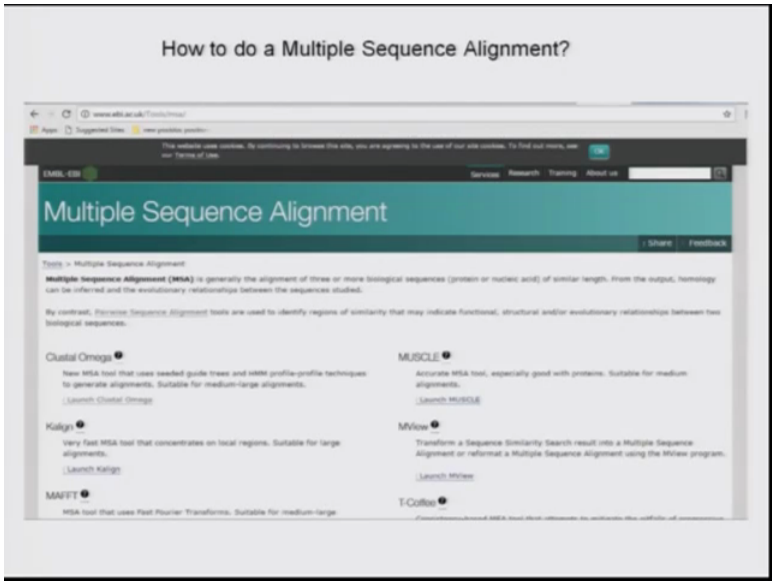
You see there are protein sequences in the order mentioned in the for the different organisms the protein sequences appear the same order. Now we can do a little modulation here is that once we feed this these sequences to a alignment tool multiple sequence alignment tool it will take the first letter that comes that appears after the arrow and the that will denote the sequence so we just modified this a bit and manually add the name of the organism the common name of the organism, so that we can for the ease of understanding. So that is what we do and I have copied this I first copied this entire thing to a notepad.

(Refer Slide Time: 11:37)



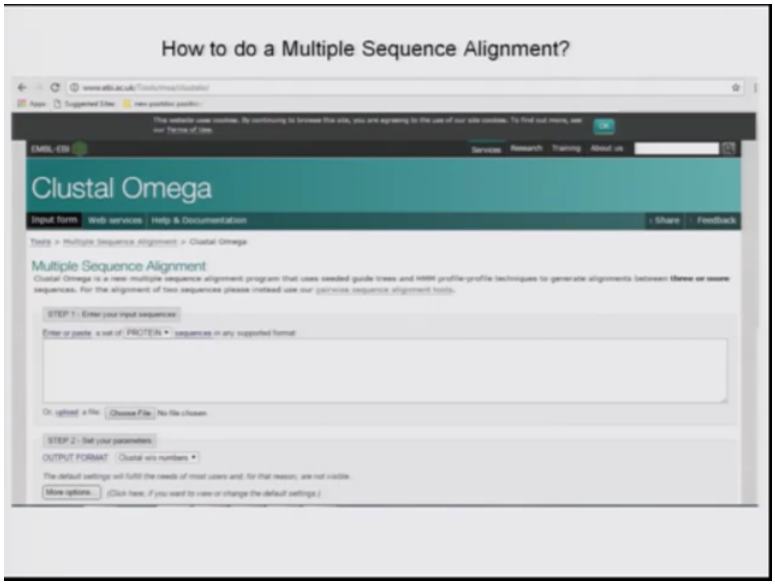
Any text editor you can use and you see before the accession number I have added the common name of the organism for the ease of my understanding like for humans for homosapiens, chimpanzee for (())(11:49) and monkey for (())(11:51) and so on.

(Refer Slide Time: 12:04)



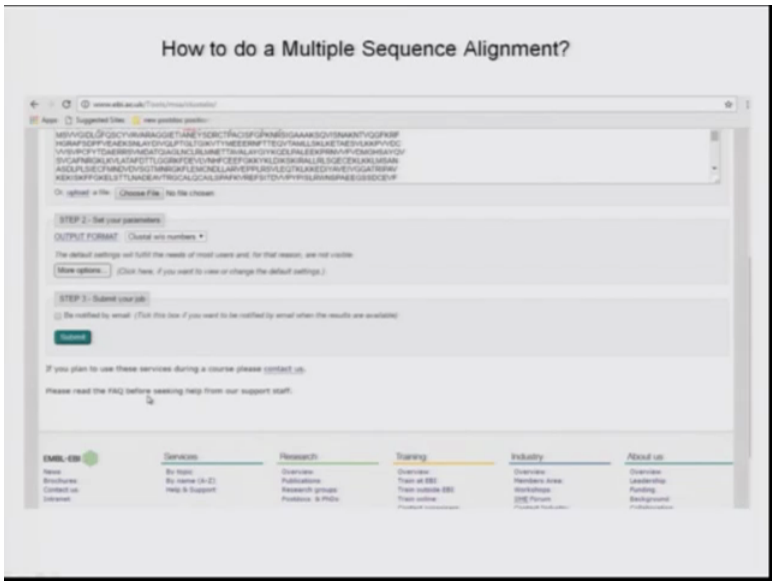
Copy this information again and now we go to one of the multiple sequence alignment tools on the EBI website the European Bio-informatics institute and here like I have showed the definition of MSA to begin with. They have defined MSA and the various tools for doing multiple sequence alignment mentioned at this website all these tools employ used different algorithms have been built to have been used to build these softwares and you can explore each one of them individually and learn about their pros and cons.

(Refer Slide Time: 12:45)



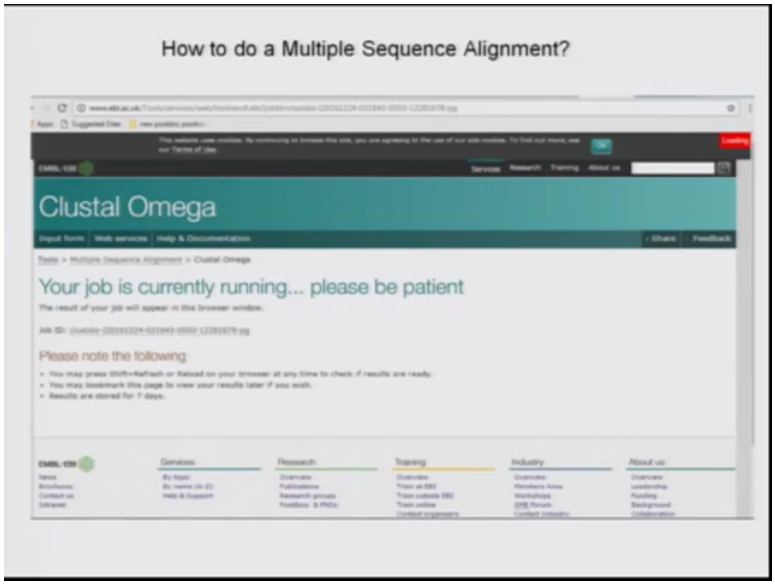
Here I will use the most updated the recent one which is called Clustal Omega and you just click on launch Clustal Omega here and you get this window where you can paste all the sequences which you want to align and if you want to do a protein alignment like I told you DNA or RNA can also be aligned by using these tools. So you select protein from the drop down menu and simply paste your sequence from the notepad in this window.

(Refer Slide Time: 13:13)



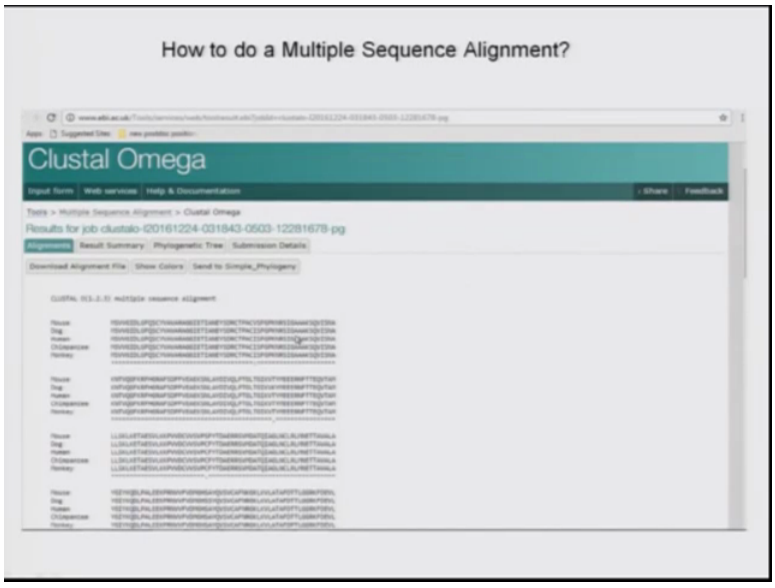
We do not want to make any major alteration in the options let them read the default options, you can always explore the other options available and click on submit.

(Refer Slide Time: 13:20)



The kind of window you get when the job is running please be patient.

(Refer Slide Time: 13:22)

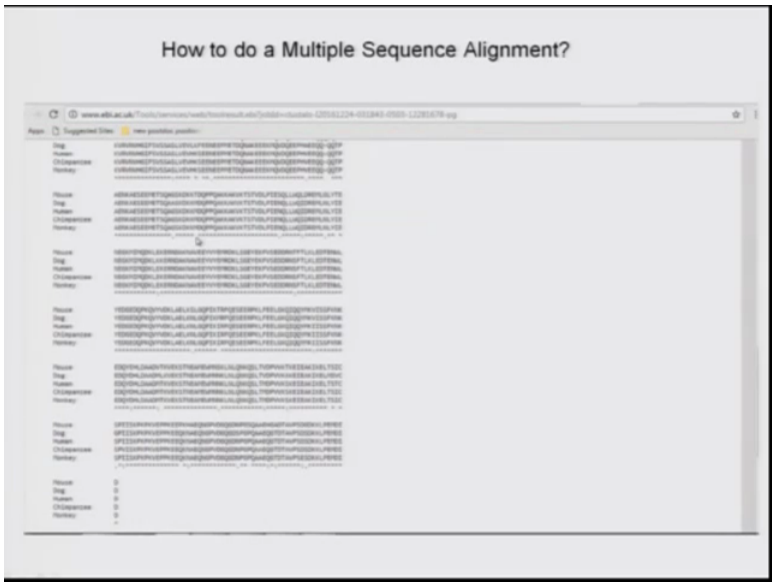


This is the output that we get in case of Clustal Omega like you saw I changed the names of the organism till their common name now it is very easy to compare the different species. You see the sequences are aligned to each other and below the last sequence there are certain mark. These symbols signify, a star or an asterisk what you call signifies a complete match as you can see for yourself wherever this HSP gene like I told you is very is a highly conserved protein known and

you can see almost all below every nucleotide column you will see a star which means its highly conserved and protein is conserved throughout.

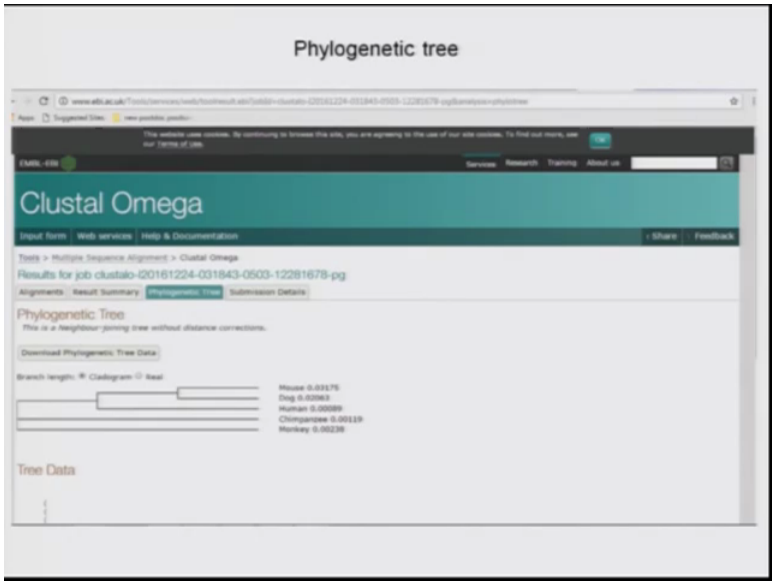
So an asterisk like I told you it means complete it means identical amino acids in case of a single dot that you see here there is change of one single amino acid which means all other, single dot means there is just one amino acid different and all others are the same. Whereas two dots as you see here means there is more than one variation which is not evident in this over here but you can explore more and find out that is how the case is, two dots means more than one variation. So this is the kind of output that we get upon doing a multiple sequence alignment and this can be used for various purposes which I have already talked about in the beginning here you see the options on the top.

(Refer Slide Time: 15:08)



The favourite application of multiple sequence alignment is building of phylogenetic tree that tells you relatedness of the species and you can draw it for yourself simply click on multiple sequence alignment. This is you can browse below and see the output that you have got, so you can see for the HSP there is very high degree of conservation and the asterisks are present all over the sequence. Like you see here two dots so these two dots signify more than one amino acid is different whereas a single dot defines they just one variation and one of the sequence all other are the same.

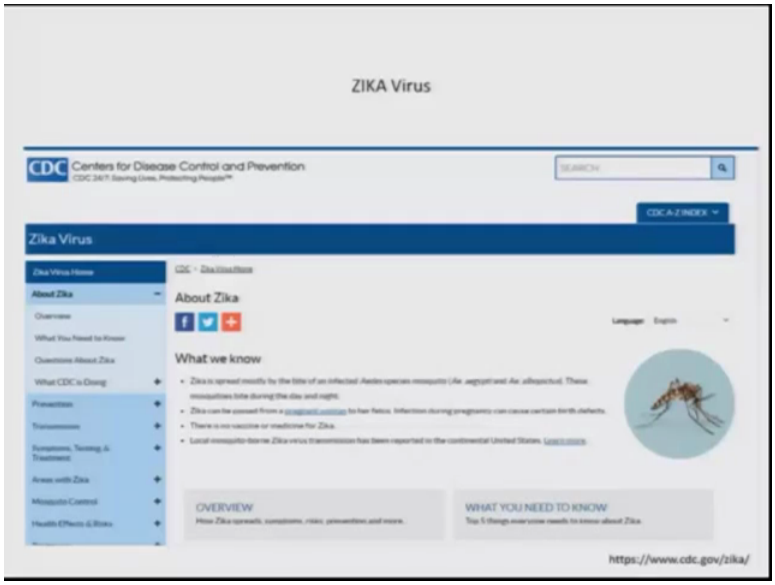
(Refer Slide Time: 15:45)



So like I told you just click on the phylogenetic tree and here is the phylogenetic tree available for the different organisms of your protein that you are analysing and here we can see that the distances the genetic distances calculated for each of the origins that will help us tell about the relatedness or the farness across the species like the monkey, human and chimpanzee seems to be close as evident from the genetics distance calculated here.

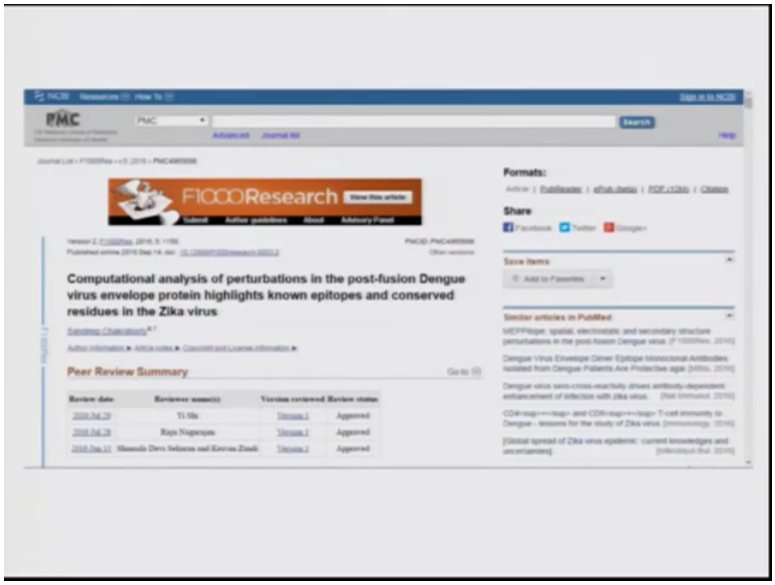
Let me tell you some interesting we have learned so far that what multiple sequence alignment is and how to do a multiple sequence alignment but let me tell you about some interesting cases where people have used multiple sequence alignment and in clinical studies or to solve real life problems.

(Refer Slide Time: 16:29)



One of this is an attempt to build a vaccine against the Zika virus. Now we have all heard about the deadly Zika virus it originated from it was identified in South African country I think Uganda and later it spread to several different countries. So Zika virus like its commonly known now is infected is transmitted by the bite of infected aides species of mosquitos and it is the passed through from a pregnant woman to her fetus and can cause birth defects. So like I told you there is no vaccine for medicine vaccine or medicine for Zika.

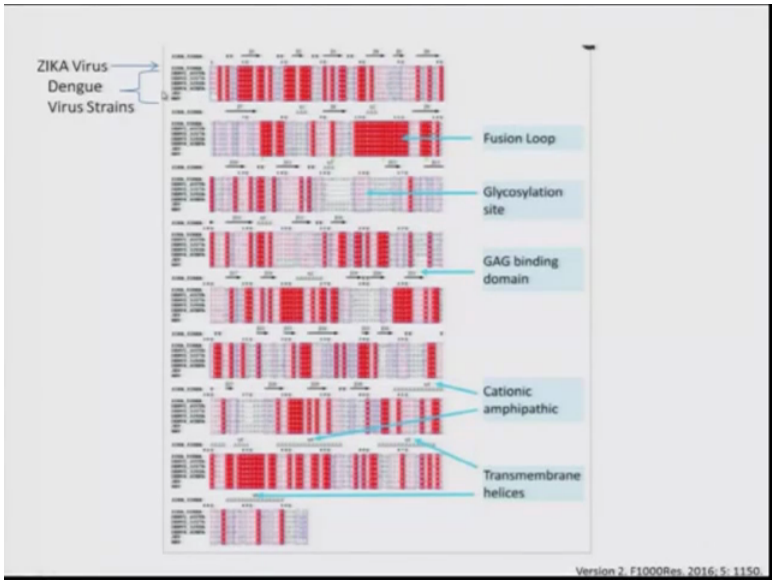
(Refer Slide Time: 17:21)



And so this interesting study there I am going to show you here it resotic to other viruses which other viruses which are similar to Zika virus such as the Dengue virus and this like a common approach for building vaccines so what people do is that they look for viruses which are common to the one which is the common target and for the conserved domains. Generally vaccines against viruses are built looking for epitopes which are perturbed when the virus fuses with the host and those epitopes are considered as the favourite epitopes for building a vaccine.

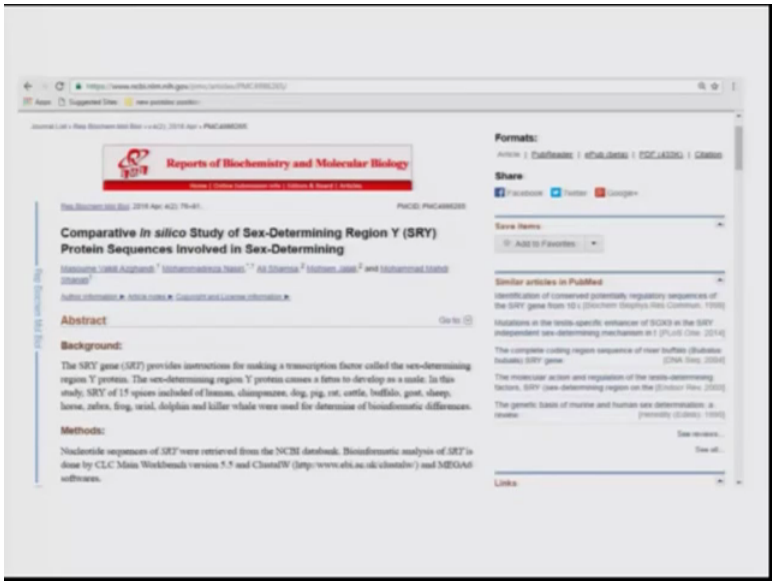
It was this study what the author is tried to do was look for epitopes such epitopes which for which the vaccines are already being made. For the Dengue virus and if those epitopes are conserved in the Zika virus and then if hence if that is the case then it will be a beginning to build and vaccines against those epitopes, that is what they did and this is a multiple sequence alignment output that I am showing from their study the first call out the naming listed over here.

(Refer Slide Time: 18:39)



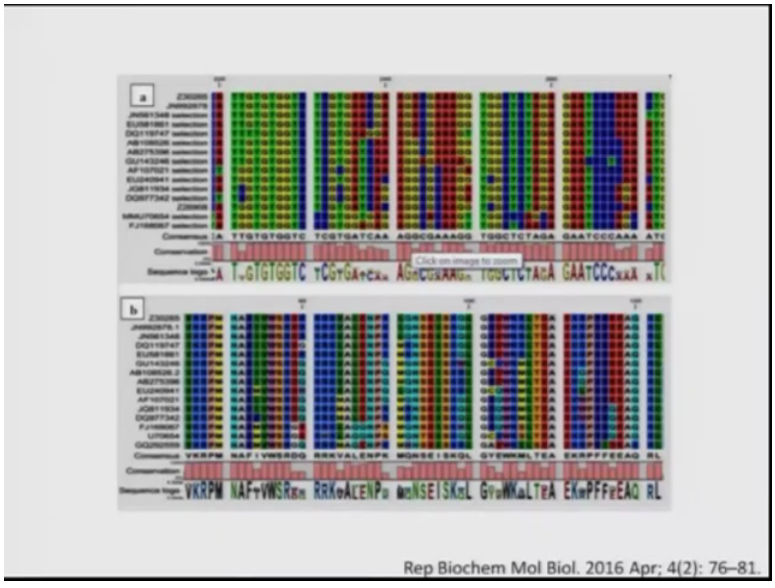
The first sequence is of the Zika virus and all others below it are from the different strains of Dengue virus, so her we can see the colour pattern denotes the different degree of conservation. The red colour show a high degree of or almost complete conservation and in the study they interestingly found that this this the fusion loop which which I just told you is important for finding the epitopes for preparing vaccines was form to be highly conserve between Zika virus and other strains of Dengue. Which was good for the study such, many more such studies were done to in an attempt to do the same.

(Refer Slide Time: 19:27)



Also we have heard about you must have heard about the SRY gene or the sex determining region Y which is responsible for male determining the sex of the fetus to maleness. So what the study did was that they wanted to check across several species the homology of this SR they conserved pattern of this SRY gene if any. So since this is such an important gene it determines the sex of the fetus, the authors tried to compare it across different species such as human chimpanzee, dogs, pig, rat, cattle, buffalo, goat, sheep and so on many more different species.

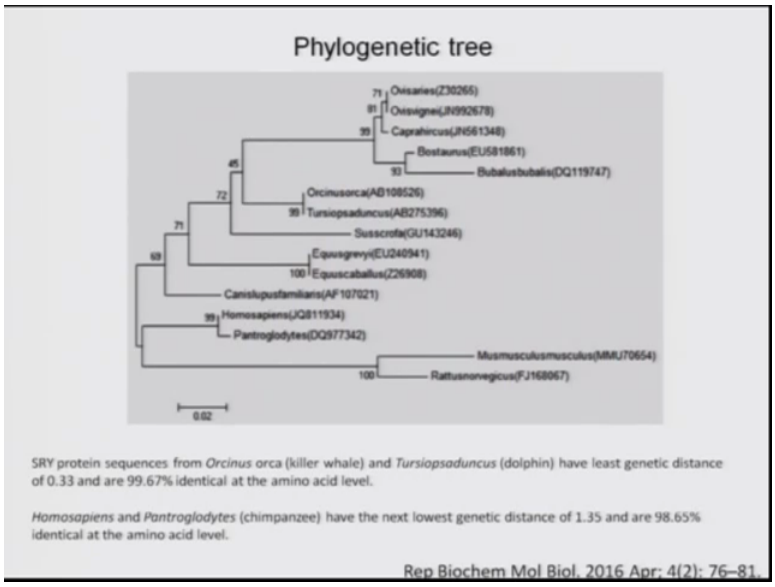
(Refer Slide Time: 20:21)



And this is the output from their studies that I am showing you the one on the top shows a multiple sequence alignment of the at the nucleotide level and the one below shows multiple sequence alignment at the protein level and the different colours here show different degrees of conservation and at the bottom you can see the sequence logo so the height denotes the degree of the conservation.

More the length of the nucleotide more it is conserved at that particular position and that is how consensus regions or conserved regions were figured out across the different species. This is a very interesting exercise to do you can always go back to this research paper and do it for yourself, you can retrieve the sequences of all these species for SRY gene and see if you can get similar output or not.

(Refer Slide Time: 21:16)

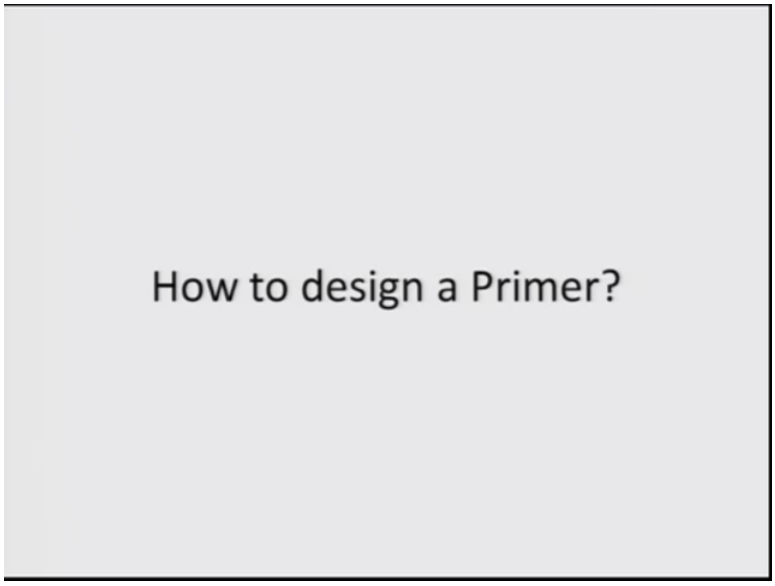


This is about aligning the sequences again you can built a phylogenetic tree using the same tool which I told you in this paper the same they built that phylogenetic tree from their output from this tree they inferred lot of interesting facts that the SRY protein sequences from the killer whale Orcinus Orca which we see here and the dolphin has the least genetic distance of 0.33 and our 99.67 percent identical at the amino acid level.

Also like we all know that homosapiens and chimpanzee are very close are known to be very close also evident from the study that they have the next closer genetic distance of 1.35 and 0.65

percent identical at the amino acid level. So it is indeed an interesting thing to study and you can always explore more from whatever I have told you in this talk.

(Refer Slide Time: 22:12)

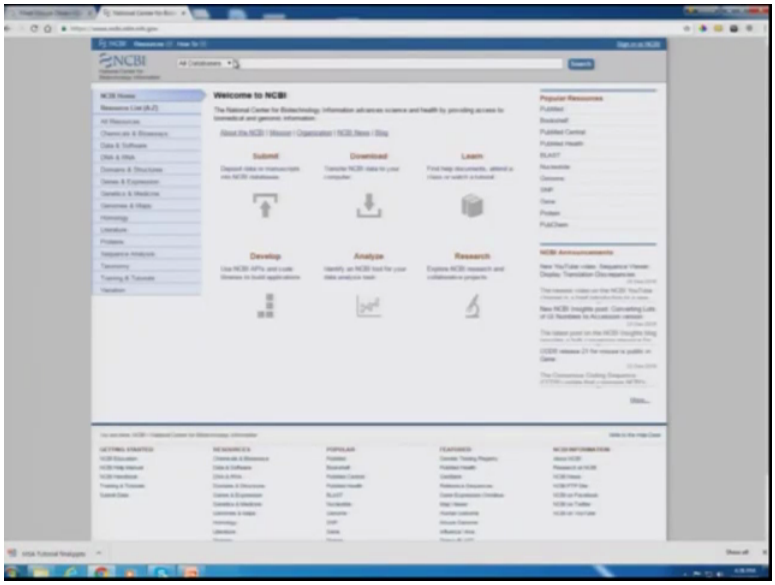


So another important tool that we use in studies functional genomics studies is to run a PCR which you have which Dr. Ganesha has taught you in of the lectures. But to do a PCR we need a primer so how do we design a primer, so let me tell you primers can be designed for DNA or the RNA which is coded from the DNA from the at the genomic level or at the transcript level. So today I will tell you how to design a primer to amplify a transcript present in the cell.

(Refer Slide Time: 22:53)

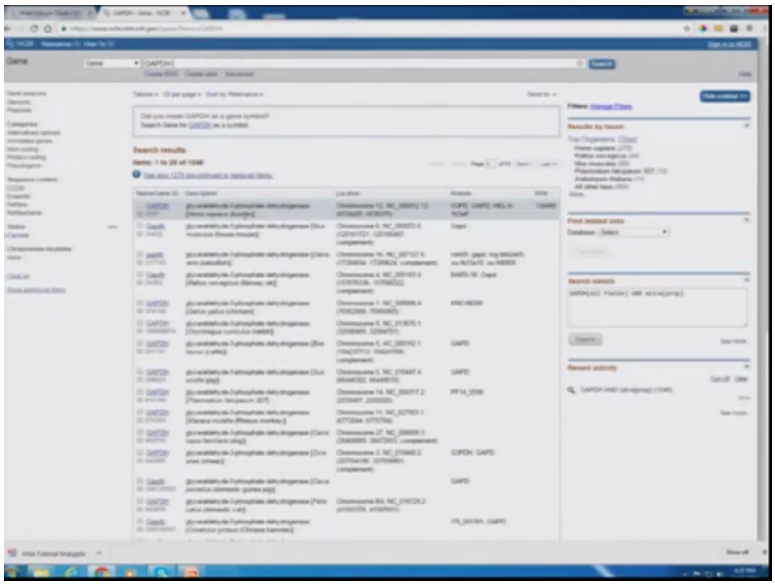
So here you see a PCR gel this is a normal PCR gel where we have DNA bands evident over here which are stained by ethidium bromide tin which is eliminated in the U.V, light and on the sides you see the ladder to know the size of the DNA band. So this is a simple polymerase chain reaction which is done using primers and the DNA and then loaded on the gel to show you the product of the PCR.

(Refer Slide Time: 23:19)



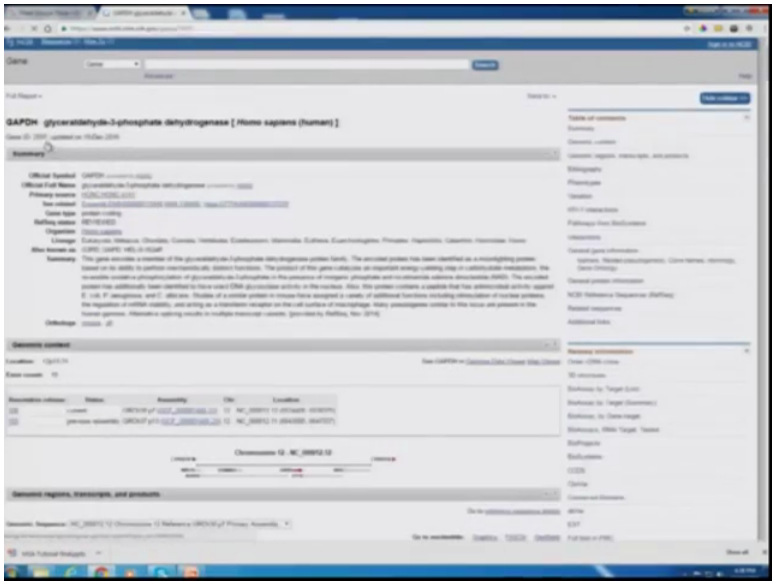
So to design the primer for the quantify or identifying the presence of a transcript in a cell. We need to design a primer for that and I will show you here, I will show you here about designing a primer for a gene called a GAPDH which is considered to be a constitutive gene expressing under all conditions and its level is not does not generally change across different conditions in the cell.

(Refer Slide Time: 24:02)



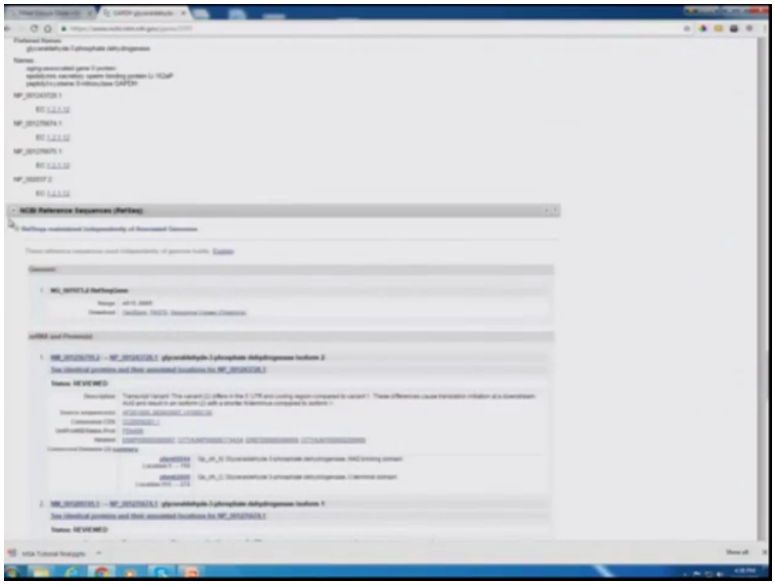
So I will simply type GAPDH this is the symbol for the protein GAPDH, it is one of the enzymes in the glycolysis cycles. Select gene over here so these are the different GAPDH across different species that they are listed over here and for which ever organism you want to design a primer you can select that organism.

(Refer Slide Time: 24:19)



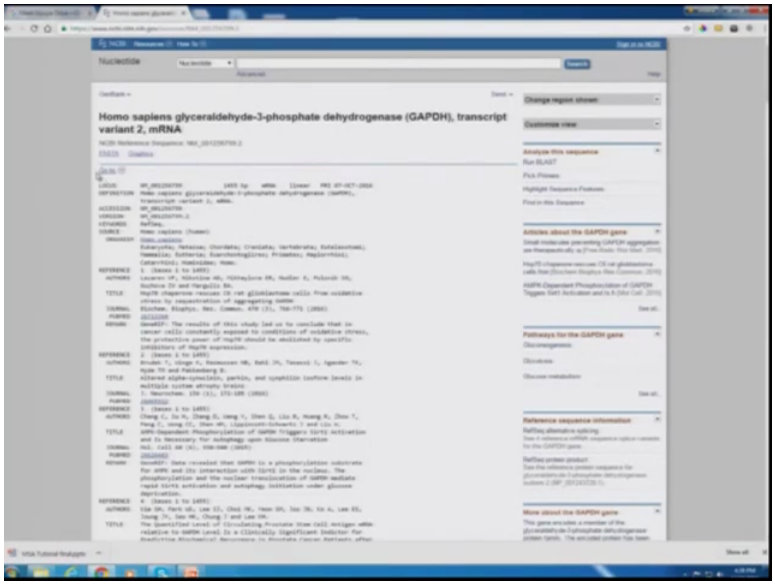
I will select human over here, so you see all the information about this gene is listed over here. The complete name glyseraldehyde 3 phosphate dehydrogenase and you see a little summary over here which tells you about the functions of this genes the known functions of the gene. You can browse down and explore more about the gene but here we want to design the primes so first we want to retrieve the sequence the transcript sequence of this the RNA sequence of this gene.

(Refer Slide Time: 24:51)



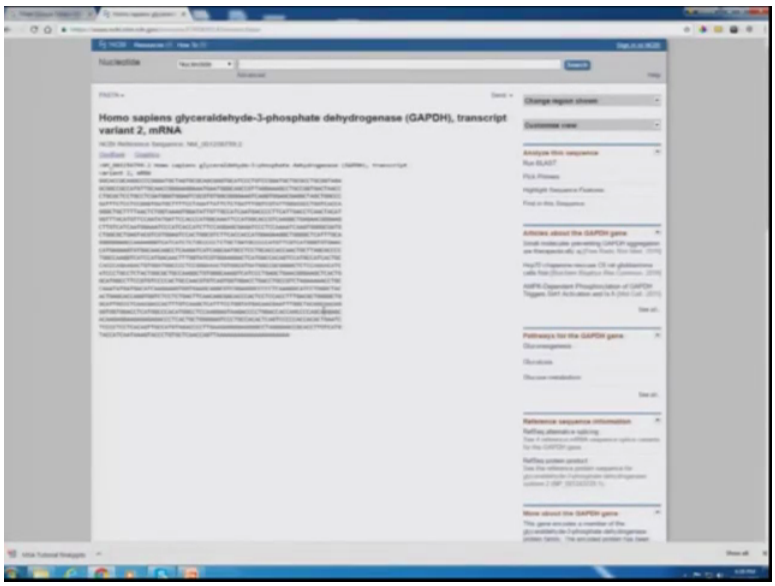
So we go down here you see the genomics sequence the RNA sequence the protein sequences all are listed for this gene. The NM represents the transcript sequence click on that so this is all about the mRNA of this gene.

(Refer Slide Time: 25:19)



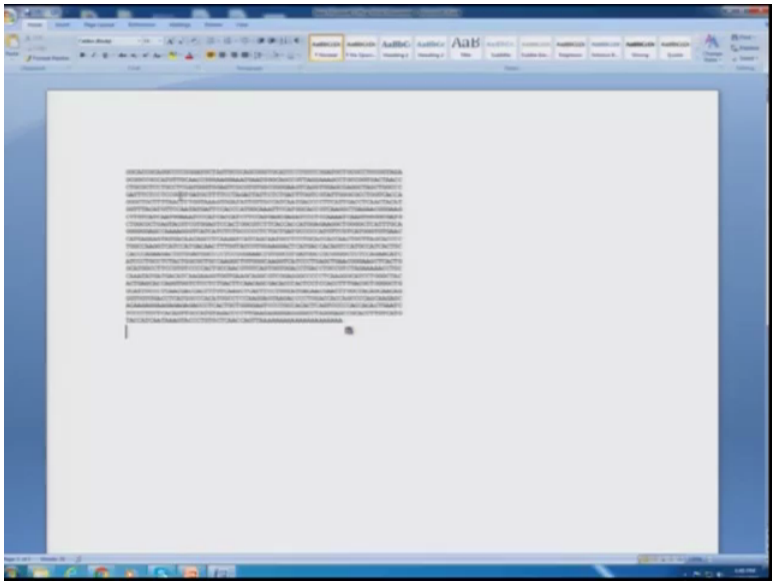
It is a 1455 base pair long mRNA and all about it is listed below you can go through it to learn more feature about and we are interested in retrieving the FASTA sequence of this mRNA.

(Refer Slide Time: 25:28)



So just click on FASTA so this is the sequence mRNA sequence that we have obtained here.

(Refer Slide Time: 25:45)



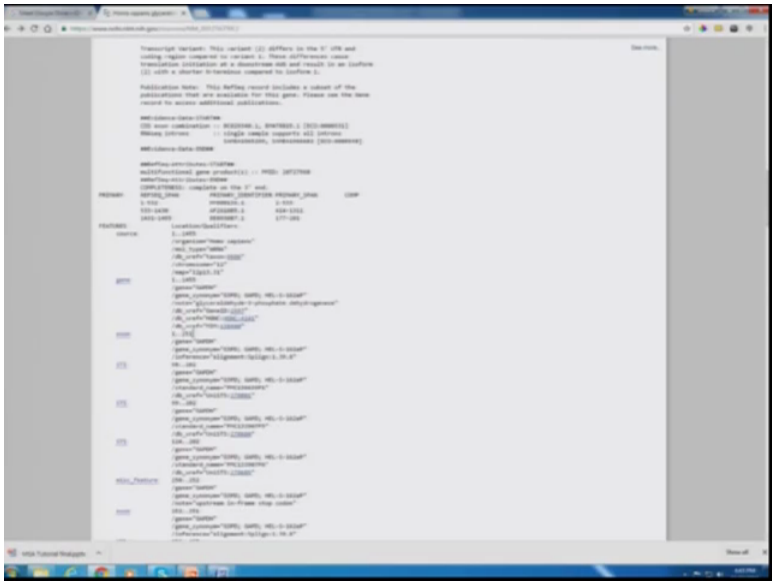
Copy this sequence from here and open a word file paste a sequence over here. Now an important thing to keep in mind while designing primers for detecting mRNA levels in the cell is that we do not want any non-specific bands present in the amplified in the PCR. The primers that we design so it is important to keep in mind to design primers from different exons which bind to different exons panning in intron in between.

So that is important because during the process of splicing the intron is cleaved out the exons are ligated together, so in the mature mRNA only the exons are stitched together. So if we design primers from different exons we will make sure that the final amplicon that we will get on our PCR is indeed from the mRNA and not from a genomic sequence where the primer might have a nil.

Since in the RNA sample there can be some contamination some genomic DNA contamination and if there is an intron present in between then it will be a huge size and the primers will never be able to amplify that larger sequence. Apart from this another thing which is important is that while we are for designing a primer to amplifying mRNA sequence, we experimentally what we do is that we isolate RNA and then prepare cDNA.

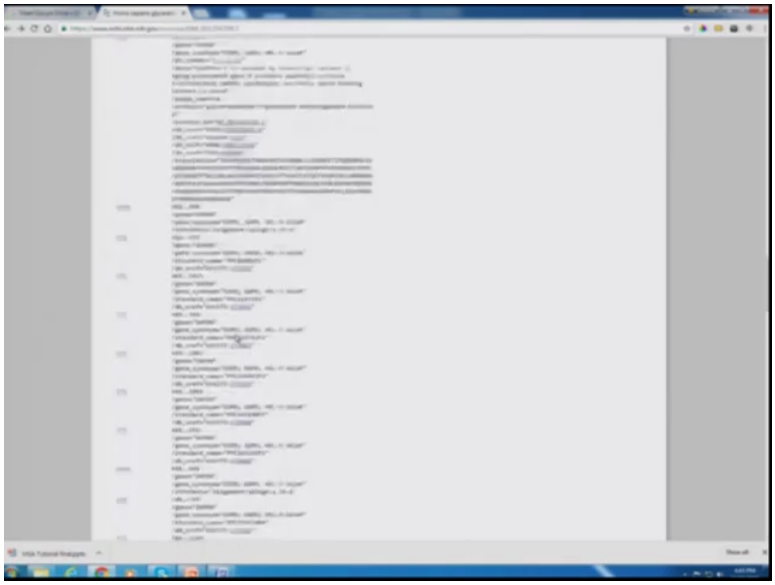
So and from that that cDNA is what Is the template for this primers so it is important that in the synthesis of cDNA from the mRNA the polymerase might not have a very good fidelity all the time and complete the transcription and complete the synthesis of the cDNA to till the end of the mRNA transcript and might just fall down, so it is a always a good idea to design the primus towards the end of the transcript. Since the end is 3 prime end since the revserse transcript is starts synthesising the cDNA so we look for two exons which are last which are present towards the end of the transcript.

(Refer Slide Time: 28:14)



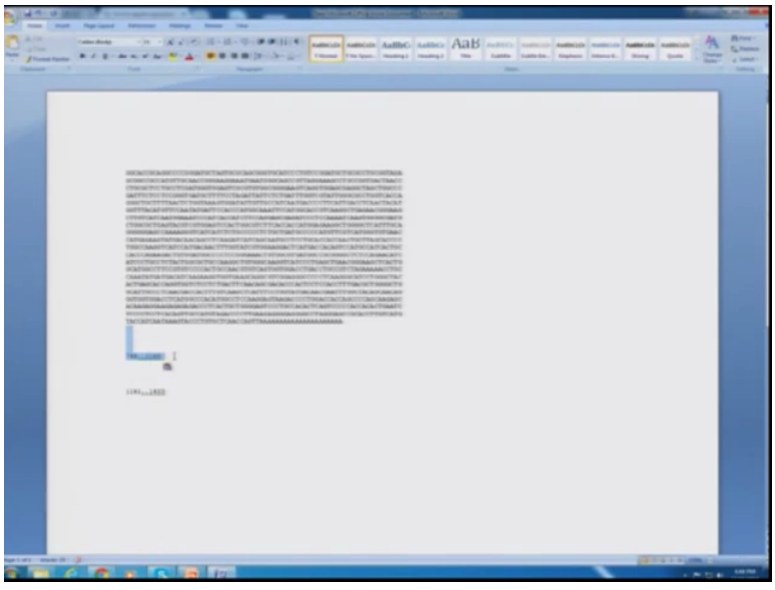
So we go back to the NCBI website where the information about the gene was given and here are the different exons that are listed here so the first exon is present from first base pair to 251 and so on.

(Refer Slide Time: 28:27)



We can go down and for the exons, the second one is from 352 to 458 and lets go towards the end so the last exon is from 1161 to 1435 we copy this information to our word file and the one just before the second last exon is from 748 to 1160 just paste it before the last exon.

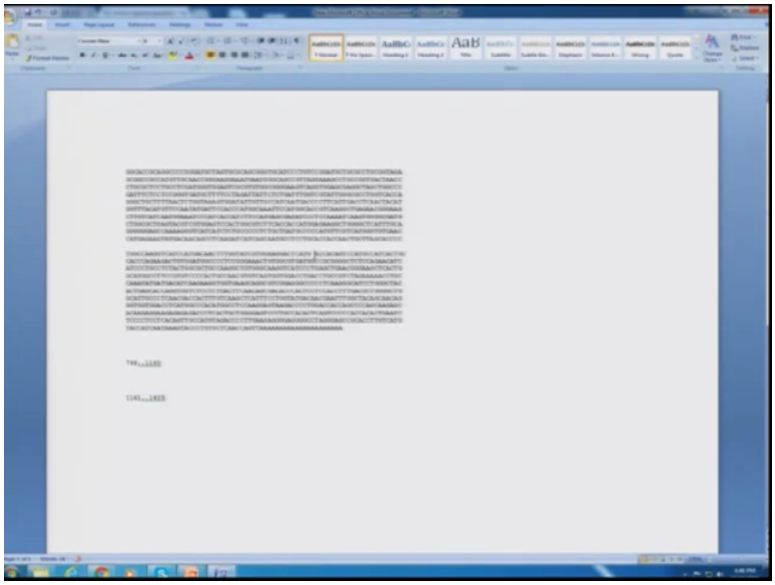
(Refer Slide Time: 29:06)



We now have the positions of the two exons and we will try to design the primes, the forward from the last second exon and the reverse primes from the last exon. So word will help us do that and we need to an estimate of the number of nucleotides and in a row and we just select that and

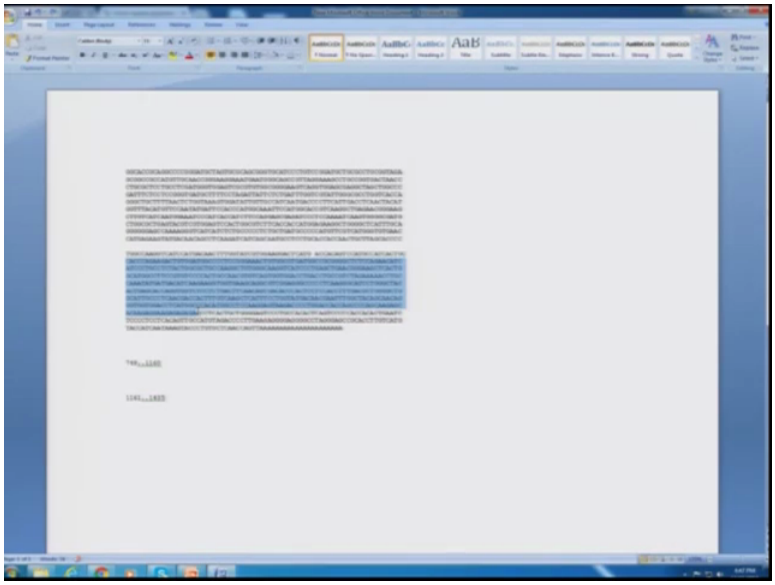
this tells us there are 17 nucleotides in the first row and we need to go until 748, so we count 10 rows that will be 700 and the 11 row will be containing the 748 nucleotide this is hopefully the 11 row and we give a space here. Just to confirm this should make 700 thats what it is and now we need to go till 48.

(Refer Slide Time: 30:34)



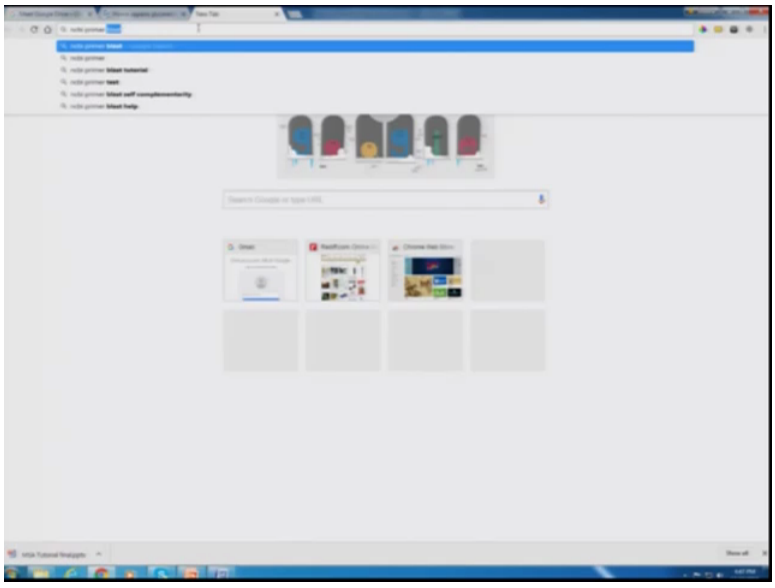
So we will just randomly select some region and this is 28 so we need to go for the, this is 39 give a space so this is 48 so in fact the space should be over here because 748 the exons starts with delineating the exons boundaries and then this goes until so this where the second last exon starts and.so this is the region we are interested in for designing the primers. So for amplifying targets you need to determine what kind of product size that you want so you it is always good to keep the product size low, so that the conditions can be for the PCR can be set such that the gene can be easily amplified.

(Refer Slide Time: 31:11)



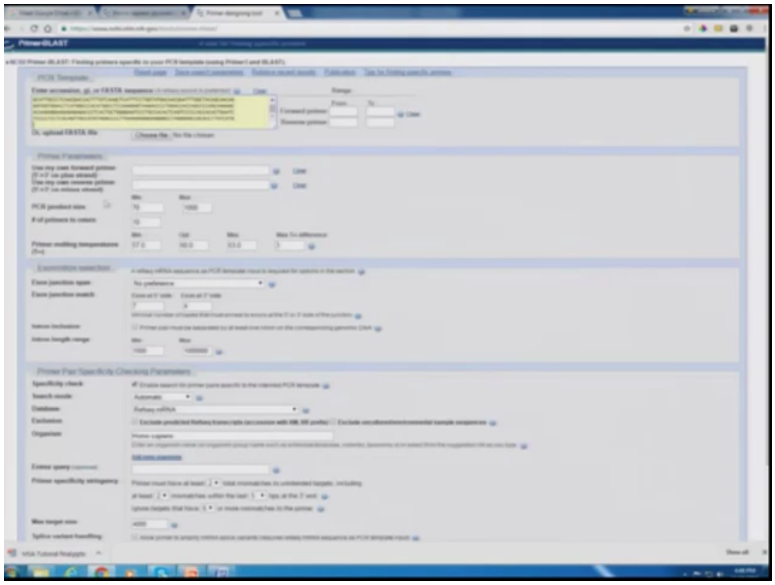
So we will select a region such that we can get a 100 to 200 base pair amplicon and I select this region for designing the primer.

(Refer Slide Time: 31:30)



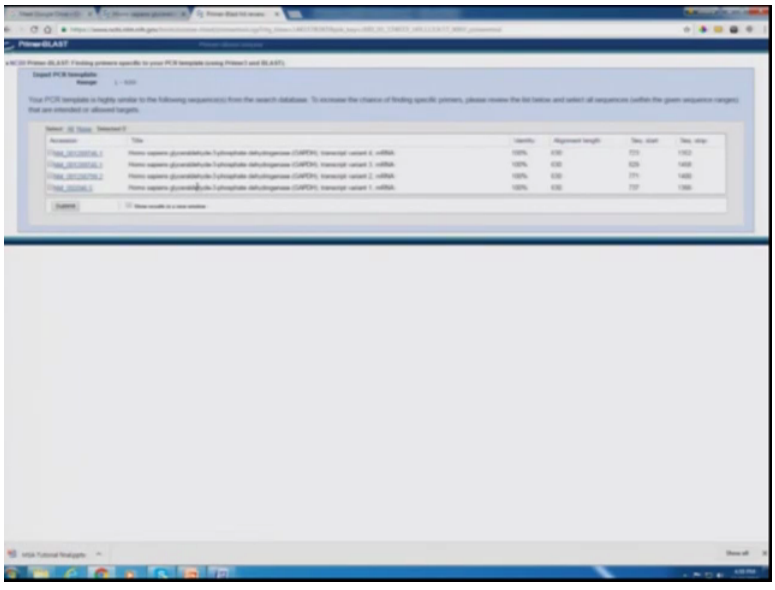
And now we go back to the NCBI blast tool, NCBI primer blast tool. We can simply type that on Google and you will get that tool so this tool will help you design primers with the template sequence that you have to provide it.

(Refer Slide Time: 31:41)



This is the template sequence we can change the amplicon size that we desire. Like i told you 100 to 2000 base pair and you can keep all other options the same and you can click on get primers.

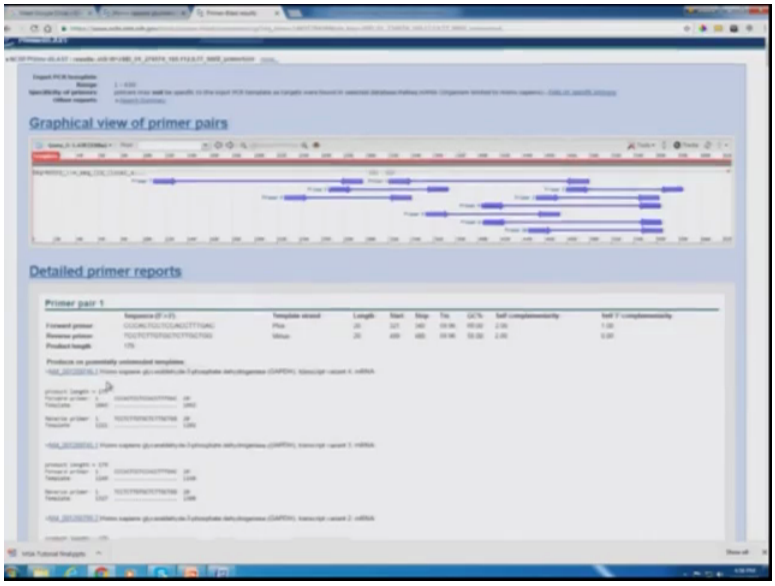
(Refer Slide Time: 32:02)



Primer blast tool is designing primers for the sequence that we provided. So we see here the tool has given us the sequence the different iso forms of GAPDH that a primer could be annealing to

so we can either select one particular iso form that we want the primers to target or if you want to generally target the gene as a whole we cannot there is no need to select.

(Refer Slide Time: 32:37)



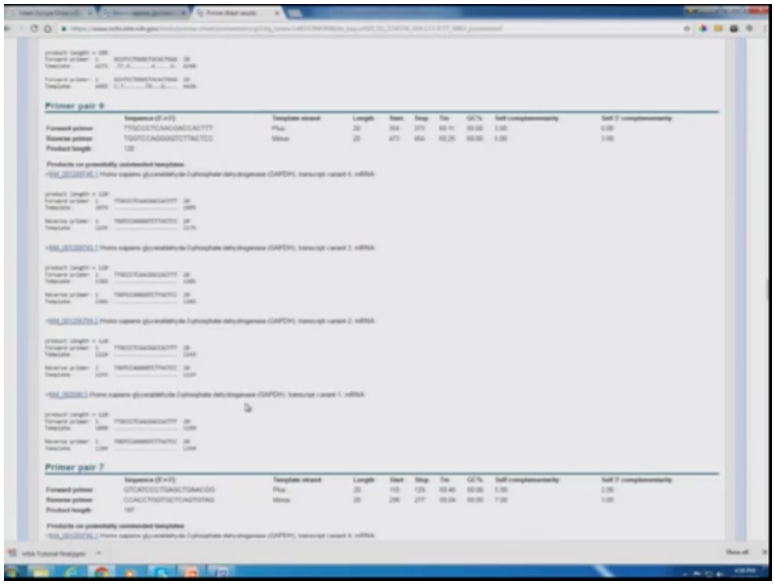
Simply click on submit so the tool has design primers for us the different primers pairs that the tool has designed and our job is now to select a primer pair which is a which satisfies the requirement of a good primer that it should not bind non-specifically to other sequences which we can find out from the results available here also it should not have repetition of nucleotide repeats which would result in self-annealing of the primer.

And there should be a uniform distribution of the nucleotide if they are if the primers are G series then that would pose difficulty in there binding and denaturation to the template so these are the conditions that things that we need to keep in mind while preparing while picking up a primer set. So the first one is of 179 base pair apart from GAPDH it also binds well to the other transcript as well and the dots you see are due matches and the nucleotides specified are the mismatches so we will not go for this primer pair. Let s look at the second one.

It is a 104 nucleotide primer pair and it matches as well with other transcripts primer pair 3 is 110 base pair and it gives, it does match with other transcript but thats a very large amplicon of 3 more than 3 kb which is difficult to be amplified with the conditions which we will be setting for

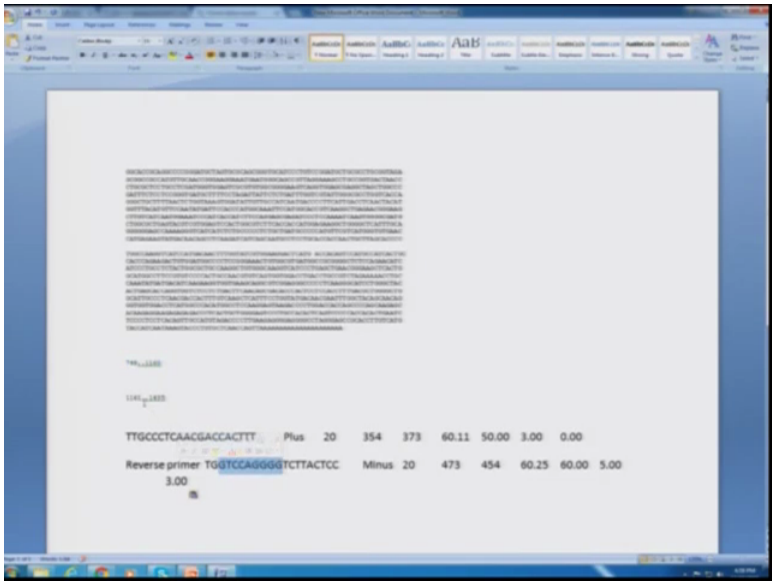
this primer which will be like very stringent conditions so this is a good primer pair the primer pair 3. Again primer pair 4 is a good option, so this is how you can always explore.

(Refer Slide Time: 34:54)



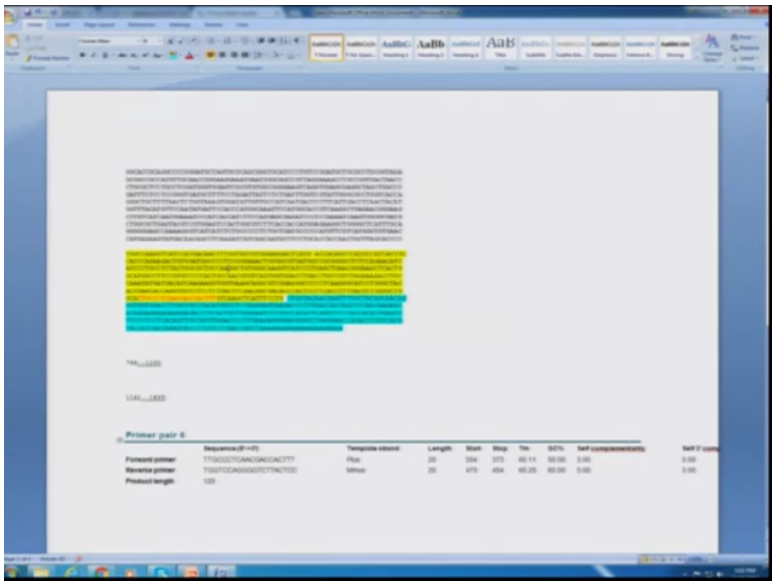
So the primer pair 6 does not seem to give us any non-specific matches so this seems to be a good option and all we now have to do is that pick up these primer pair from here.

(Refer Slide Time: 35:29)



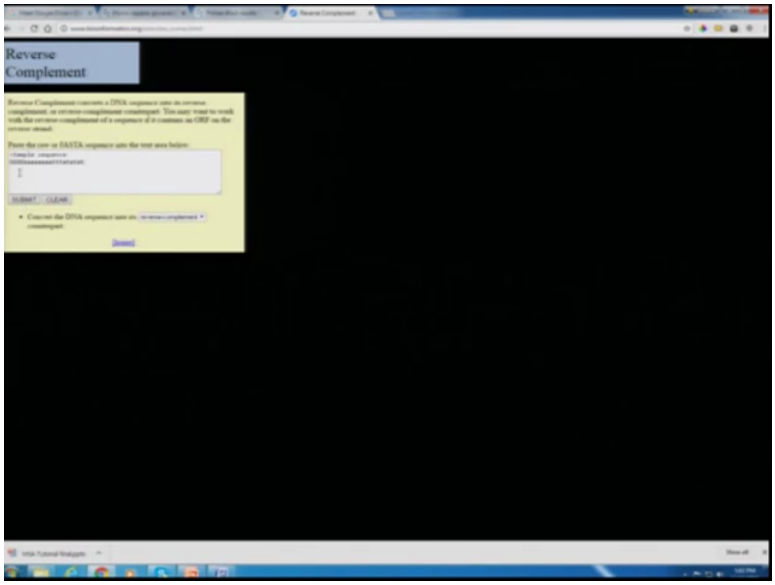
And go back and match it to our sequence in the word file and make sure that these primer pairs they (())(35:20) to the different exons that the the forward (())(35:24) to the second last exon and the reverse (())(35:27) to the last exon, so to do that we will have to first determine the boundaries of the 2 exons. So the second last is until 1160 base pair so we can also colour these exons so there is no confusion.

(Refer Slide Time: 36:23)



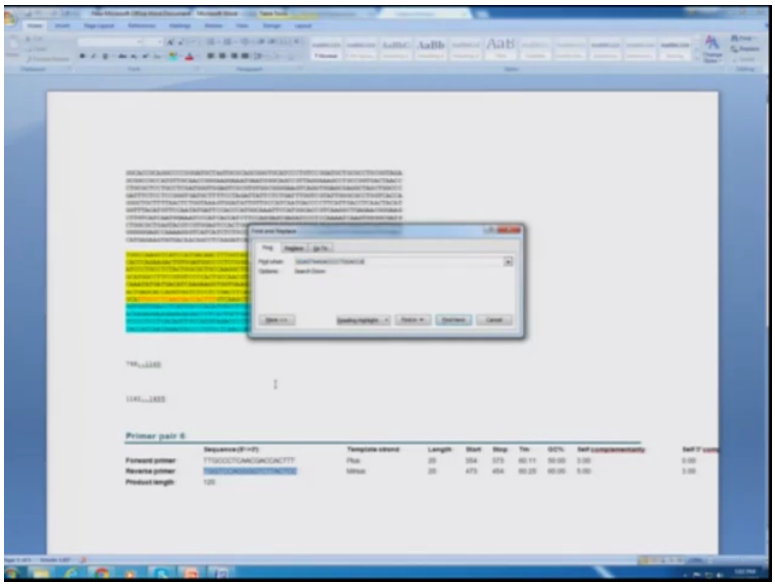
This is yellow and this is to a different colour and now we can see we can match the primer pairs so the forward primer pair. Lets so this is the primer pair and the forward will match exactly to the sequence, that is over here whereas the reverse will match the sequence in a reverse complimentary manner.

(Refer Slide Time: 36:34)



So we can find the reverse compliment orientation of the reverse primer.

(Refer Slide Time: 36:48)



And copy it from here and then try to match it with the last exon. So this is where it matches. So we are successful in designing primers which bind to different exons and do not give non-specific matches in the NCBI prime last 2. So this is how we can design primers and I hope you enjoyed learning these two functional genomics tools which in today's class and do go and

explore more such tools and if you have any queries you can write to us at the course portal. Thank you.