

Bioengineering: An Interface with Biology and Medicine
Prof. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology - Bombay

Lecture – 35
Bioinformatics-I

So welcome to MOOC-NPTEL course on Bioengineering, an Interface with Biology and Medicine. In the last few lectures, I tried to give you some of the problem statements, some bigger problems towards which all of us have to work together whether we are biologist or engineers. Actually, you know, we have to think about the goal of how to solve those major problems.

I also try to provide you an interface with clinicians to give you some motivation that in which way engineering discipline can help medical scientist and clinicians for our day to day medical problems. To understand in which way the medical field can also benefit from the engineering tools and various insights you can bring in now from the engineering point of view. In this slide I thought it might be good idea to introduce you to some of the bioinformatics tools.

Some time even if you do not have an opportunity to work in a vet laboratory to do the experiments yourself, you can actually obtain lot of data which is publically available now a days in different databases and repositories. And then if you know some of the open access software and tools, you can then try to download the data available already and then make use of the data and then try to understand in which way people are studying genes and proteins.

So that is something which you can do from your own system, your own computers by sitting at your own room. You need not to work in a laboratory setup. So even before we dig deeper into the fundamental concepts, I thought it is good idea to begin with some of the bioinformatic assignment. To do this, I will take help of one of my teaching assistants. We will first try to understand and explain, you know, more information how you can retrieve for a given gene.

So in my previous lectures I had talked to you about, you know, how people target specifically a given gene, look at their sequences and how you can use technologies like polymerase chain

reaction to amplify the genes. So first of all you know how to get a gene sequence. Again you need to use the bioinformatics tools. You can go to NCBI portal where you enter the gene name or accession ID.

And then you will see how much information is already available for that given gene. Then in the tutorial sections you will see in a stepwise manner how this assignment could be performed. So I will hand it over now to my TA and after doing the assignment, we will also give you some problem statements and some assignments which you can do at your own place. (Video Starts 02:45 - Video Ends 10:54)

(Refer Slide Time: 02:45)



I am Apoorva Venkatesh. I am a final year PhD student under professor Sanjeeva Srivastava and I am your TA for this course.

(Refer Slide Time: 02:54)

Bioengineering: An Interface with Biology & Medicine

BIOINFORMATICS ASSIGNMENT 1

How to use NCBI to study genes and evolutionary relationships



This bioinformatics assignment will teach you how to study genes using NCBI and their evolutionary relationships.

(Refer Slide Time: 03:02)

A screenshot of the National Center for Biotechnology Information (NCBI) website homepage. The page features a search bar at the top with the text 'Search All Databases' and buttons for 'Search' and 'Clear'. Below the search bar, there is a 'Welcome to NCBI' section with a brief description of the center's mission. A prominent 'Genome' section highlights that '1000 prokaryotic genomes are now completed and available in the Genome database'. The page also includes a 'Resources' sidebar with various categories like 'All Resources (A-Z)', 'Data & Software', and 'Training & Tutorials'. On the right side, there is a 'Popular Resources' section listing tools like BLAST, BioCatal, and Gene, and a 'NCBI News' section with recent updates.

NCBI provides a variety of databases and computational tools which are freely accessible for use by the research community. NCBI is also home to PubMed which many of you may have used for literature searches. In today's assignment, we will learn how to retrieve the sequence of a gene? How to identify similar sequences and how to compare multiple sequences to study their evolutionary relationships.

(Refer Slide Time: 03:29)

EXAMPLE Assignment 1A. Enter the following details to complete the table

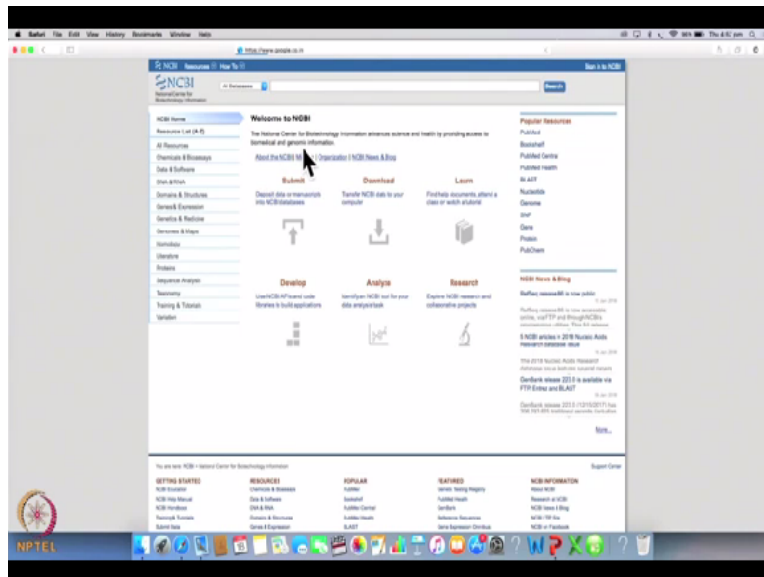
Gene ID	3040
Official full name	
Date of last update	
Gene symbol	
Gene type	
Location	
Size of gene (bp)	
Organism	
Superkingdom	
Size of Chromosome (bps)	
# Genes on chromosome	

Hint: Data can be obtained from <https://www.ncbi.nlm.nih.gov/>



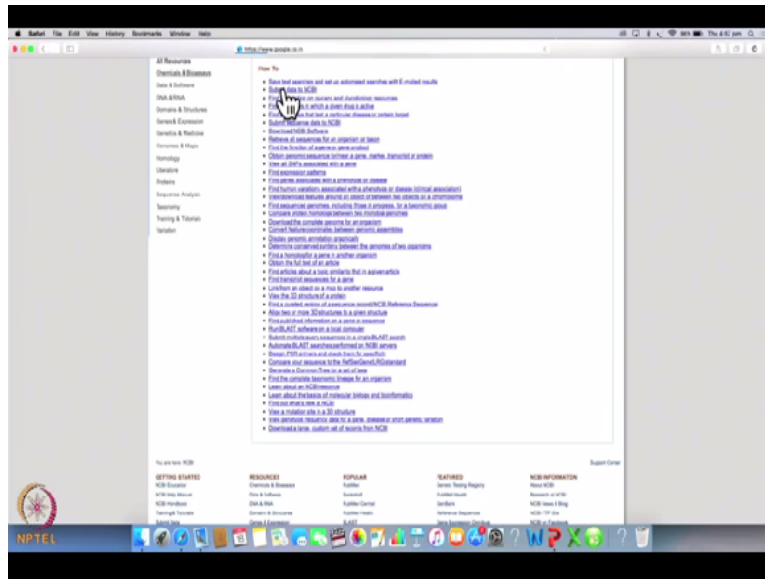
Let us begin with the first question. You have been given an accession number 3040 which is a unique identifier given to a DNA or protein sequence. We need to fill this table by getting all the information about this gene. So let us get started. It would be great if you could do this along with me on your computers. Type NCBI into your Google browser.

(Refer Slide Time: 03:57)



So this brings us to the NCBI website. We will not be able to explore all the features of NCBI today. However, I would highly recommend that you go through some of these resources after this tutorial. You can begin with the how to link, that is up here. This will teach you how to accomplish specific tasks at NCBI. Now I am going to click this link and this will give us a list of how tools.

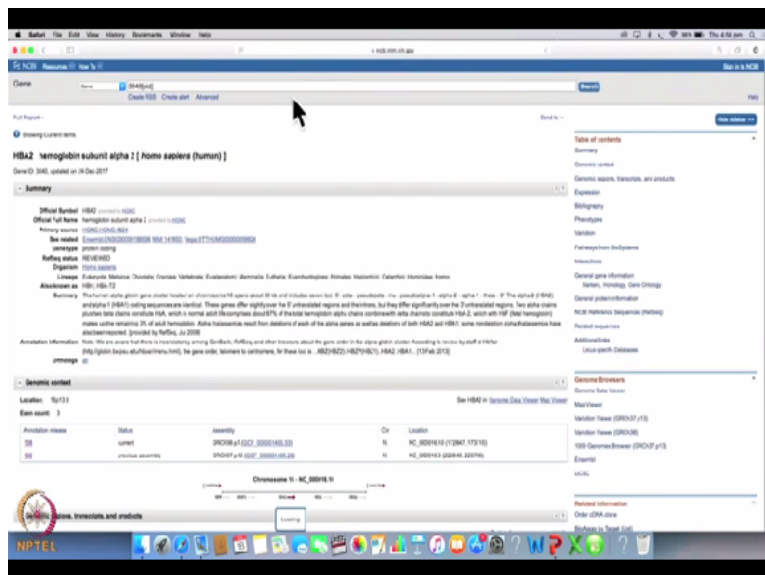
(Refer Slide Time: 04:24)



For example, you may want to know how to retrieve all sequences for an organism or taxon. You may want to know how to find the function of a gene or gene product. View all SNPs associated with a gene. Or you may want to know how to obtain the full text of an article. So on the left is a list of resources the NCBI provides. You may want to use some of them depending on your objectives. So now let us get back to our assignment.

The first thing we will have to do is to type the accession number in the box here and click gene from the dropdown menu. Now we say search.

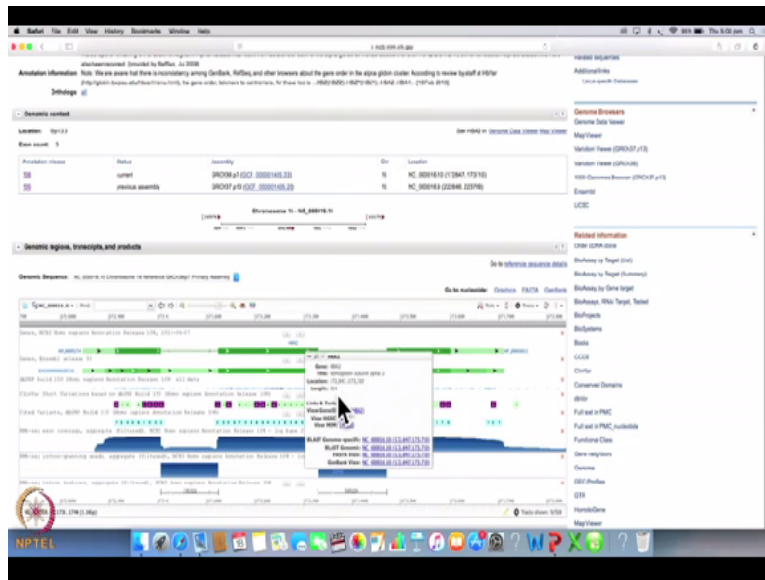
(Refer Slide Time: 05:21)



So this page will give us all the details we are looking for. Please fill the table as you move along. Our first task is to find out the gene name. So this gene encodes a subunit of the protein hemoglobin. I am sure you have heard of this protein earlier. Hemoglobin is a protein molecule in red blood cells that carries oxygen from the lungs to body tissues. Here you can see that this page was last updated on 24 December, 2017.

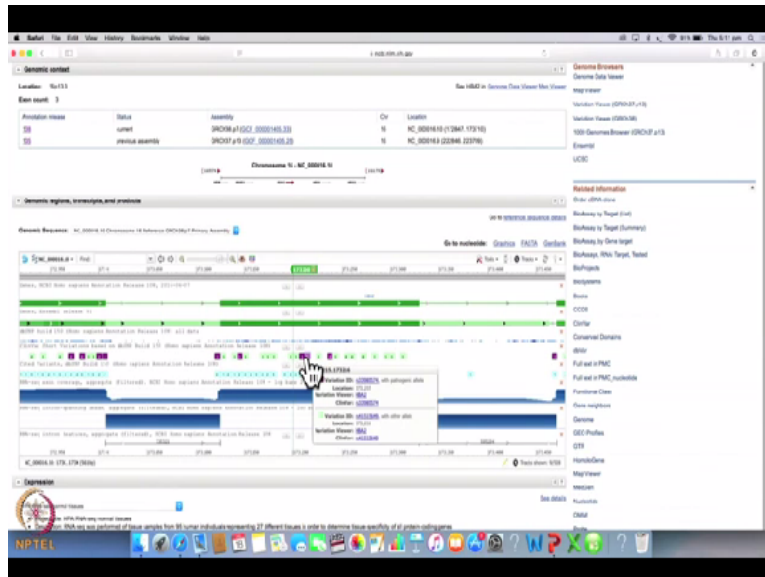
The official gene symbol is HBA2. But however, you will also see from this list that not everyone uses this official symbol in their publications. This can also be called HBH, HBA-T2, etc. Gene type is a protein coding gene and this summary here gives us information about the gene and its functions. Now if you scroll down further, you see genomic context. So genomic context describes where the gene resides. So this gene resides on chromosome 16 in the p arm at 13.3 region.

(Refer Slide Time: 06:49)



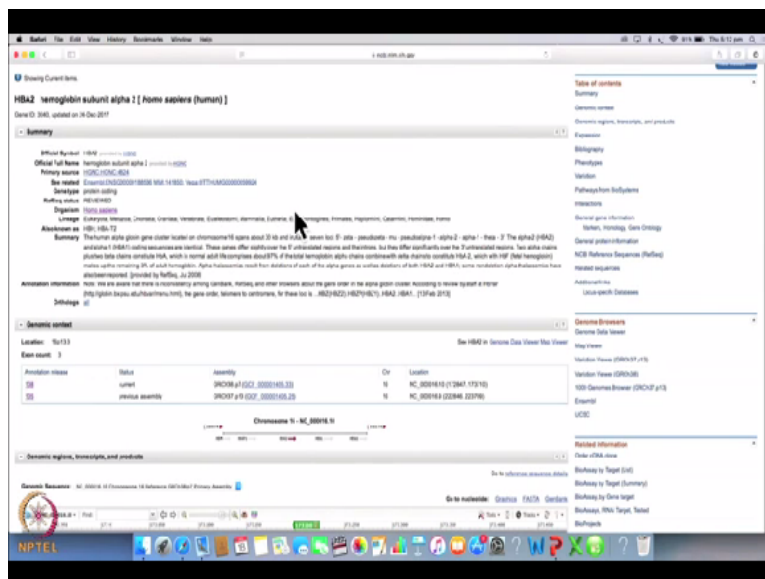
Now if you scroll down further, this section here shows us the direction in which the gene is read. So this particular gene HBA2 is read towards the right. Now we scroll down further and we go and click on HBA2 to get us more details about this gene. So this will give us the length which is 864 base pairs. So if you click on this gene, you will get all the information on the gene. And if you further click on this green section, you will get information on the messenger RNA as well as in the translated protein.

So now let us zoom into 1 particular section of this gene and click zoom.
(Refer Slide Time: 07:35)



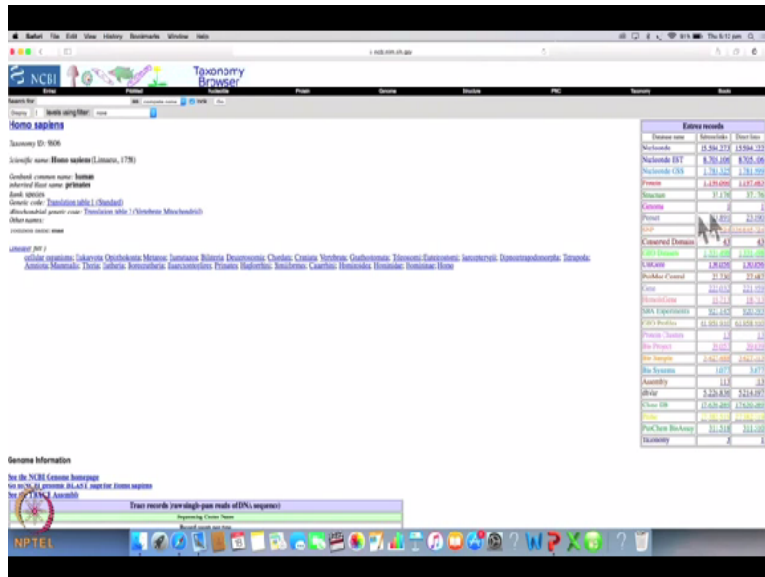
Now if you zoom into 1 of the locations, you will be able to see the variations associated with this gene, darker the colour, more deadly the variation. So notice this particular variation. As you will notice, this is in a very dark colour. This indicates a very deadly variation and if you zoom into this particular section, you will notice that this also has a PubMed citation. So now let us scroll upwards.

(Refer Slide Time: 08:10)

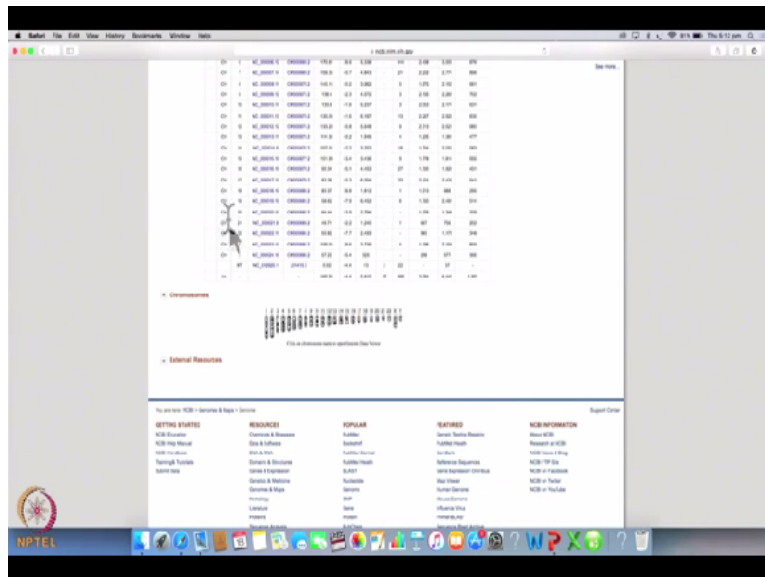


You will see that this gene belongs to homo sapiens, obviously because this is human hemoglobin subunit alpha 2 gene. This particular organism belongs to the superkingdom

Eukaryota. To know more about this organism, we can go and click on to this link.
(Refer Slide Time: 08:31)



And we can click on to genome to get more information about the genome of this organism.
(Refer Slide Time: 08:00)



Now as you know that this particular gene belongs to chromosome 16. So we go to chromosome 16 and you will see that the size of this chromosome is 90.34 megabases and it in codes 1920 genes. So with this information, you will be able to complete this table.
(Refer Slide Time: 09:10)

EXAMPLE Assignment 1A. Enter the following details to complete the table

Gene ID	3040
Official full name	hemoglobin subunit alpha 2
Date of last update	24-Dec-2017
Gene symbol	HBA2
Gene type	protein coding
Location	16p13.3
Size of gene (bp)	864
Organism	Homo sapiens
Superkingdom	Eukaryota
Size of Chromosome (bps)	90.34Mb
# Genes on chromosome	1920 genes

Hint: Data can be obtained from <https://www.ncbi.nlm.nih.gov/>



Now let us go back to our assignment. What we have is a Gene ID 3040. NCBI had thus fill in all these details. The official full name, hemoglobin subunit alpha 2. Date of last update 24 December, 2017. The gene symbol is HBA2. It is a protein coding gene. The location is on chromosome 16 in the p arm at region 13.3. The size of the gene is 864 bases. The organism is Homo sapiens. In the superkingdom, Eukaryota.

The size of the chromosome 16 is 90.34 megabases which encodes 1920 genes. So this brings us to the end of the first part of this assignment.

(Refer Slide Time: 10:03)

Assignment 1B. Enter the FASTA sequence of 3040 in the box below



The next assignment is a very small one. You have been asked to enter the FASTA sequence of

this gene in the box below. So let us go to NCBI website again.

(Refer Slide Time: 10:14)

The screenshot shows the NCBI Genome browser interface. At the top, the search bar contains the text '3040'. Below the search bar, there are tabs for 'Genome', 'Gene', 'Protein', 'Map', 'Variation', and 'Tools'. The main content area is titled 'Homo sapiens (human)' and provides information about the reference genome (GRCh38.p11). It includes links to download genome annotations in GFF, GenBank, or RefSeq formats, and a list of publications. On the right side, there are sections for 'NCBI Resources', 'Related information', and 'Search details'. The 'NCBI Resources' section includes links to 'Genome Data Viewer', 'FTP Human annotation (GFF)', 'FTP Human chromosomes', and 'Map Viewer'. The 'Related information' section includes links to 'Assembly', 'BioProject', 'Gene', 'Components', 'Protein', 'PubMed', and 'Taxonomy'. The 'Search details' section shows the search criteria '11116161[Organism:exp]'.

Let us type 3040 in the search box. We type gene and press search.

(Refer Slide Time: 10:28)

The screenshot shows the NCBI Gene browser interface. The search bar contains the text '3040'. Below the search bar, there are tabs for 'Gene', 'Protein', 'Map', 'Variation', and 'Tools'. The main content area is titled 'Genomic context' and shows the location of the gene on Chromosome 16. It includes a table of annotations with columns for 'Annotation release', 'Status', 'Assembly', 'Chr', and 'Location'. The table shows three annotations: 'current' (GRCh38.p11), 'previous assembly' (GRCh37.p13), and 'previous assembly' (GRCh37.p13). Below the table, there is a genomic map showing the gene structure and its location on Chromosome 16. The 'Genomic Sequence' section shows the sequence of the gene, and the 'Genome regions, transcripts, and products' section shows the gene structure and its location on Chromosome 16. On the right side, there are sections for 'Related information', 'BioAssay by Gene target', 'BioAssay by Gene target', 'BioAssay, RNAi Target, tested', 'BioProjects', 'BioSystems', 'Brands', 'CCDS', 'UniProt', 'Conserved Domains', 'Full text in PubMed', 'Full text in PubMed', 'Functional Class', 'Gene neighbors', 'Genome', 'GEO Profiles', and 'GTR'.

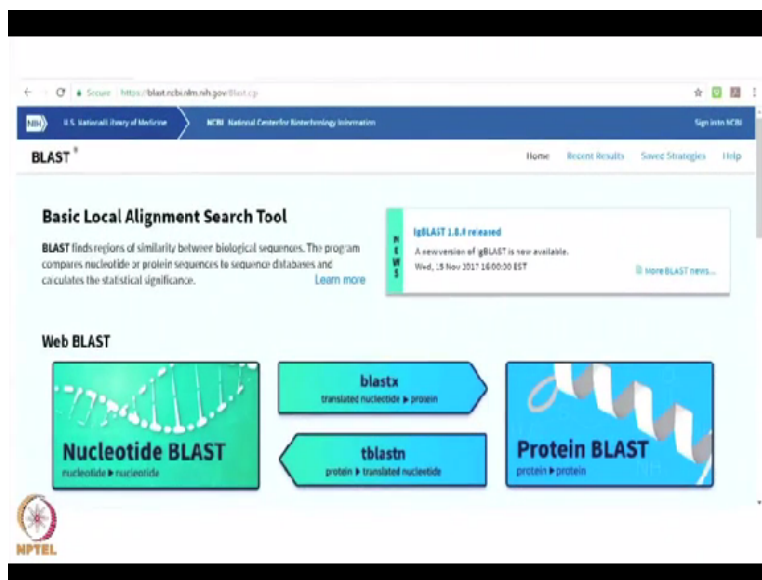
So this brings us to the same page. If you scroll down, you will be able to see FASTA here. If you click on FASTA, you will get this particular nucleotide sequence for this gene.

(Refer Slide Time: 10:39)

of information which is biologically relevant and it may sometime gives us the clues for even process like evolution how it has happened. Such sequence similarity searches can actually also help us to do to study evolutionary relationship.

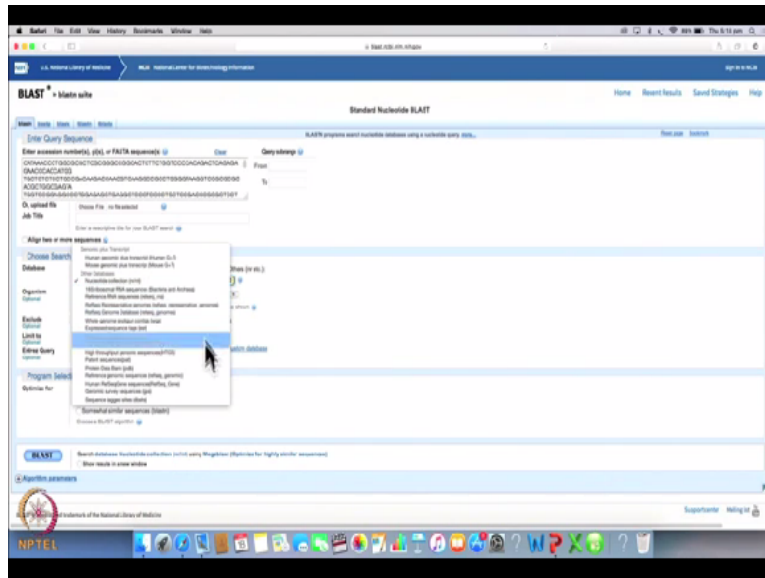
Or it can allow us to identify the proteins with similar functions. So this type of analysis which you can do using bioinformatics tools, is known as multiple sequence alignment and it can be best performed using various open access tools including the Clustal W or Clustal omega which we are going to demonstrate you in the tutorial section. (Video Starts 12:03 - Video Ends 22:10)

(Refer Slide Time: 12:03)



Before we move on to the next assignment, I would like to talk to you about a very popular feature of NCBI known as BLAST. BLAST stands for basic local alignment search tool. It allows you to take an input sequence and compare it to a database to see if anything similar has already been found. It can be very useful for studying unknown sequences or if you want to see how similar your sequence is to other sequences. The most commonly used feature is nucleotide BLAST. So let us click on NBLAST which is Nucleotide BLAST.

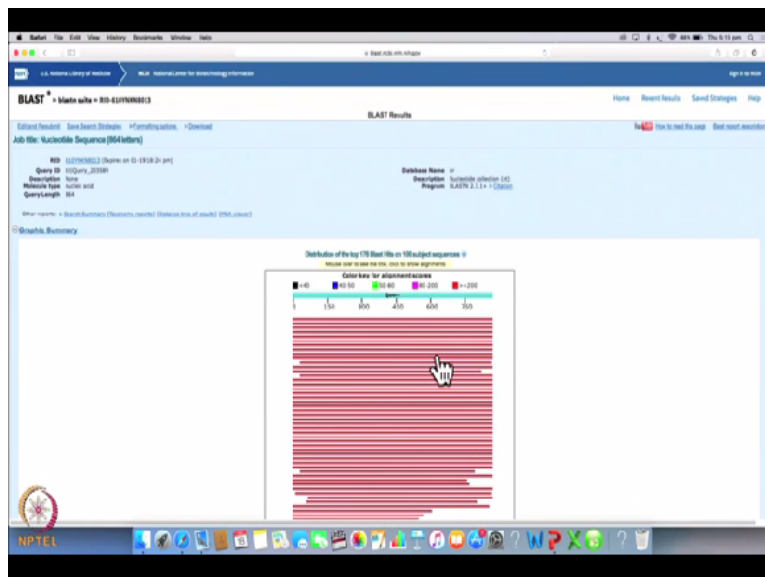
(Refer Slide Time: 12:40)



This will take us to the BLAST page where you will see that we need to enter a query sequence which is nothing but our input sequence which we need to search against a database. Let us take the FASTA sequence of the hemoglobin alpha chain that we just pasted. You can choose a human database if you know your protein is a human protein or a mouse database or if you want to see results from other species, you can choose other.

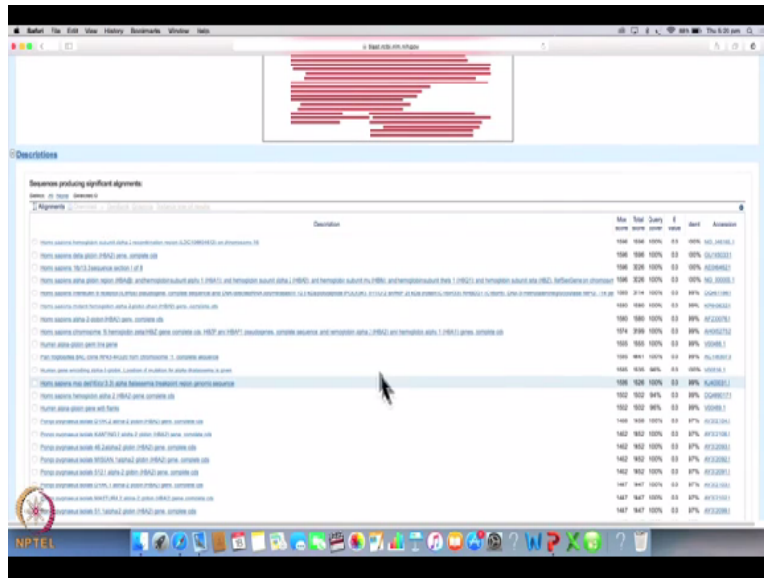
We will click on other. This dropdown menu here gives us a list of many databases. We will select nucleotide collection which is a combination of all nucleotide databases. Now let us click BLAST.

(Refer Slide Time: 13:46)



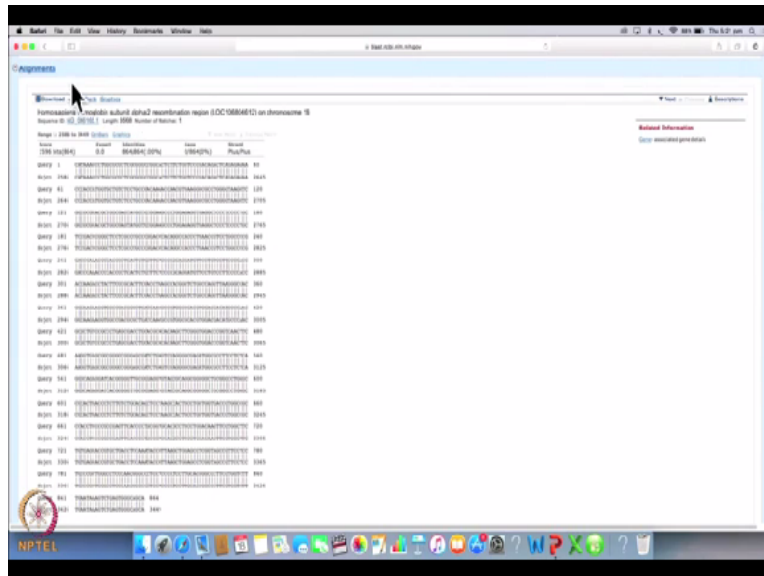
Here you will see a whole bunch of red lines which BLAST has found. These represent sequences similar to our input sequence. Top one here, is our query sequence which goes from 1 to 864 base pair. The colours here tell us how closely the results match our query or input sequence and the length of these bars tells us how much they line up with our query. So you will see there are some which line up perfectly, where some do not.

(Refer Slide Time: 14:35)



When you scroll down, you will get a list of sequences producing significant alignments. You have a maximum score, a total score and a query cover which is the percentage match with the query and an E value which is the measure of how likely something can occur by chance. Lower the E value, better the result. You will notice that the first few matches have 100% query cover which means that they are 100% identical. While you will also see that as you go lower, down this table, you will find matches with different species.

(Refer Slide Time: 15:22)



The alignment section which is further below here, gives you the actual nucleotide sequences and matches for each of the results. For example, in the first case, there is a 100% match; therefore, no gaps. So now let us get back to our assignment.

(Refer Slide Time: 15:41)

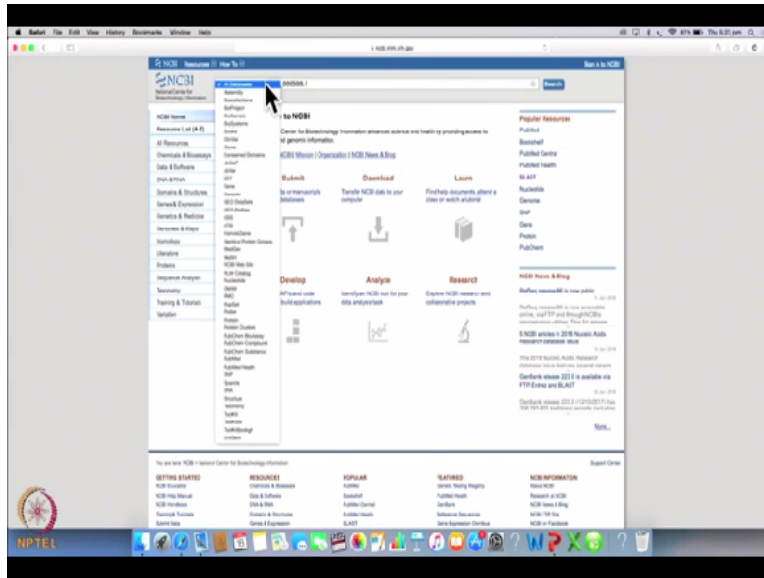
Assignment 1C. Enter the following details to complete the table

Basic Local Alignment Search Tool (BLAST)	
Accession number	NP_000508.1
Protein Name	
Organism Name	
Sequence Length	
Locus	
No. of BLAST hits	
PDB ID	



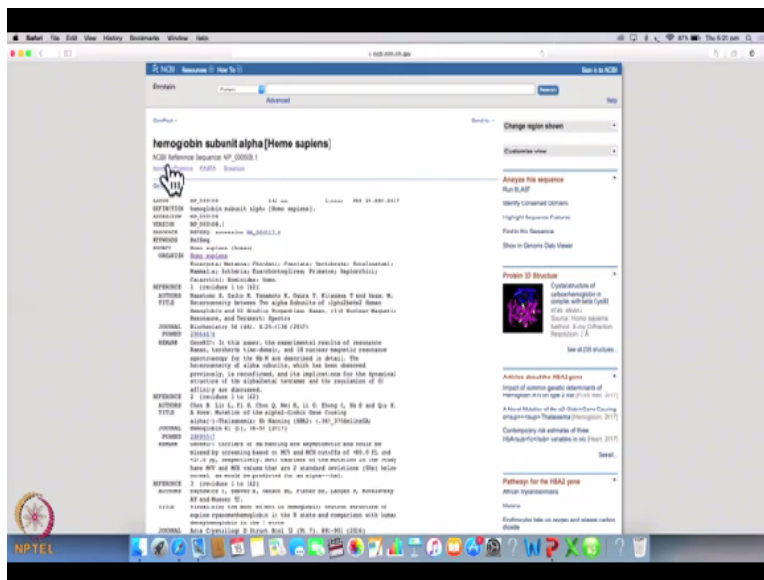
You have been given a protein accession ID which is NP 000508.1. You have to get details for this protein using NCBI. We have done this before for a gene. We will now do the same for this particular protein. Let us type the accession ID in the search box.

(Refer Slide Time: 16:03)



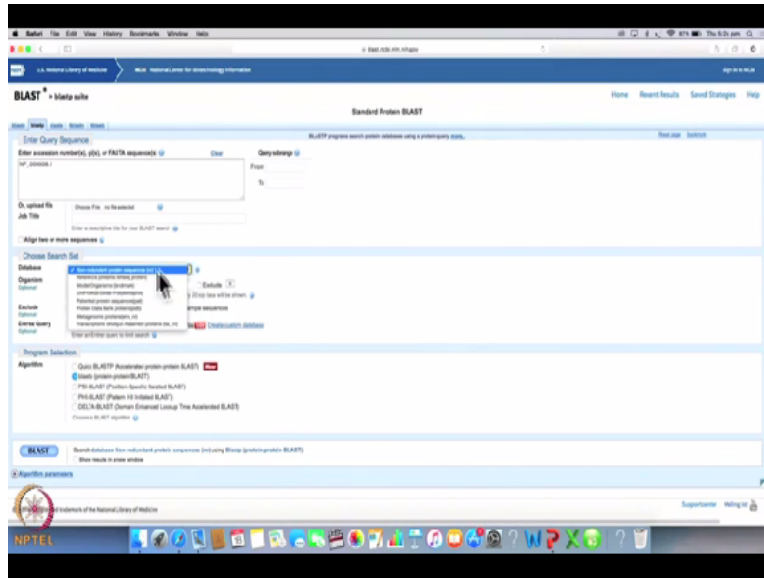
So let us type the accession ID in the search box and now we will select protein from this dropdown menu and click search.

(Refer Slide Time: 16:18)

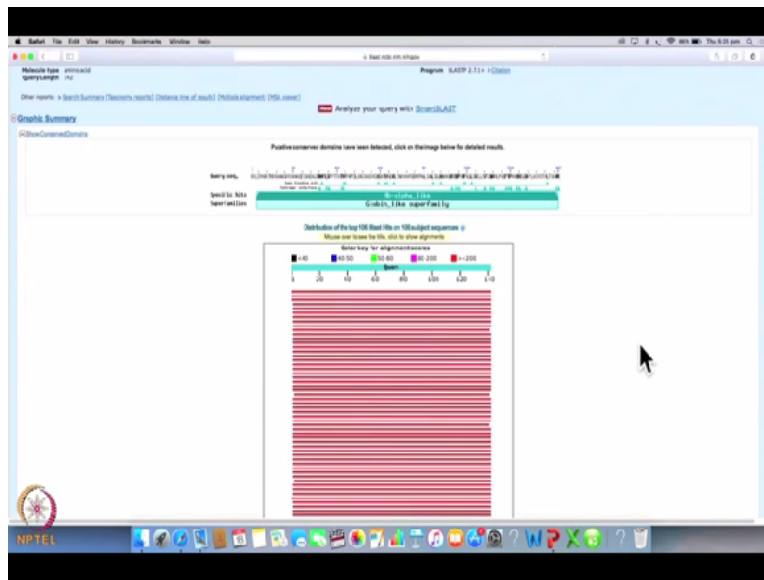


So here you will see a very similar to what we saw for the gene. You can actually get all the details for this particular protein. So the protein name is hemoglobin subunit alpha. The organism name is Homo sapiens. The sequence length is 142 amino acids. The locus is NP 000508.

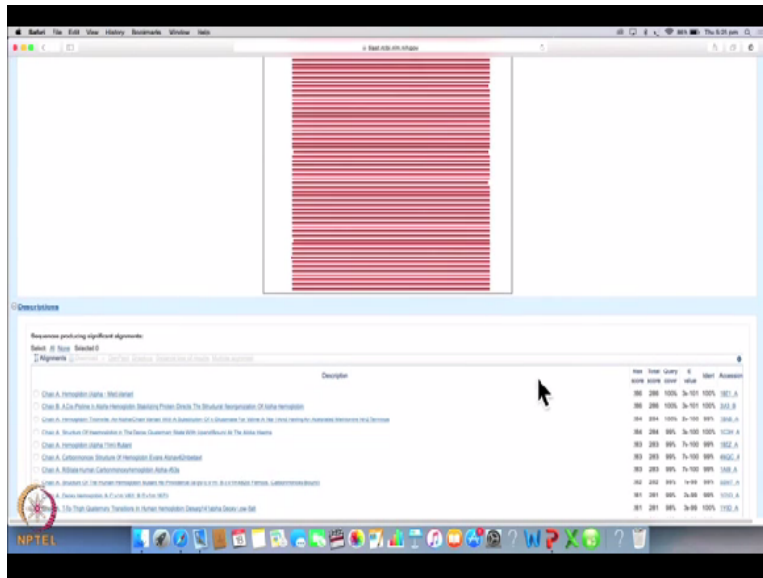
(Refer Slide Time: 17:56)



And now when we run a BLAST search, so now let us run a BLAST search and select PDB here. The accession number is given in this particular query box and we click on BLAST. (Refer Slide Time: 17:22)



So this particular page will give you details about the BLAST search as you see here, very similar to what we saw earlier. Just as you saw earlier, there are 106 BLAST hits on 100 subject sequences and you will find similar bars here which represent similar sequences. (Refer Slide Time: 17:59)



And if you scroll down further, you will see that the first hit is identical to your query sequence and you will take the PubMed ID from here which is your PubMed accession ID. So now let us go back to our assignment and fill the table.

(Refer Slide Time: 18:18)

Assignment 1C. Enter the following details to complete the table

Basic Local Alignment Search Tool (BLAST)	
Accession number	NP_000508.1
Protein Name	hemoglobin subunit alpha
Organism Name	Homo sapiens
Sequence Length	142 aa
Locus	NP_000508
No. of BLAST hits	106 Blast Hits
PDB ID	1E2I A

6

The details for this protein are the protein name is hemoglobin subunit alpha. The organism name is Homo sapiens. Sequence length is 142 amino acids. The locus is NP 000508. The number of BLAST hits 106 and the PDB ID.

(Refer Slide Time: 18:37)

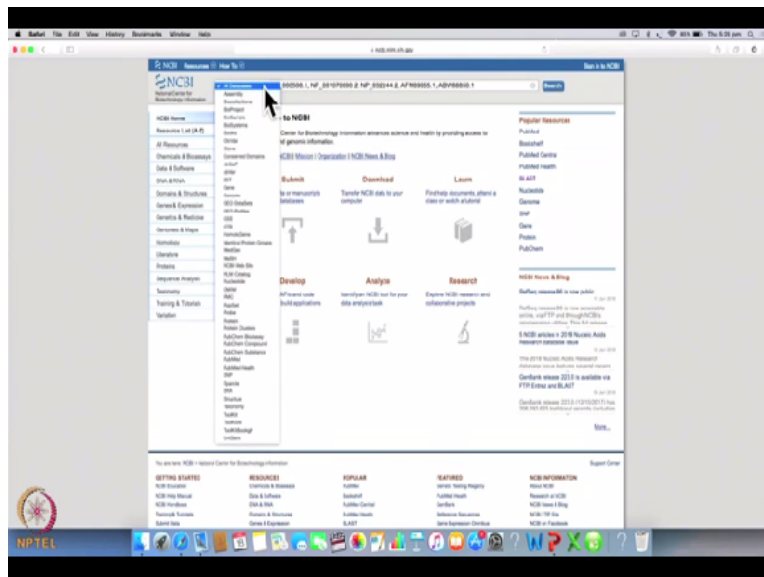
Assignment 1D. How are these sequences related?

Multiple Sequence Alignment (MSA)	
Accession No.s	a) NP_000508.1, b) NP_001070896.2, c) NP_032244.2, d) AF188055.1, e) ABW88850.1
Evolutionary related	



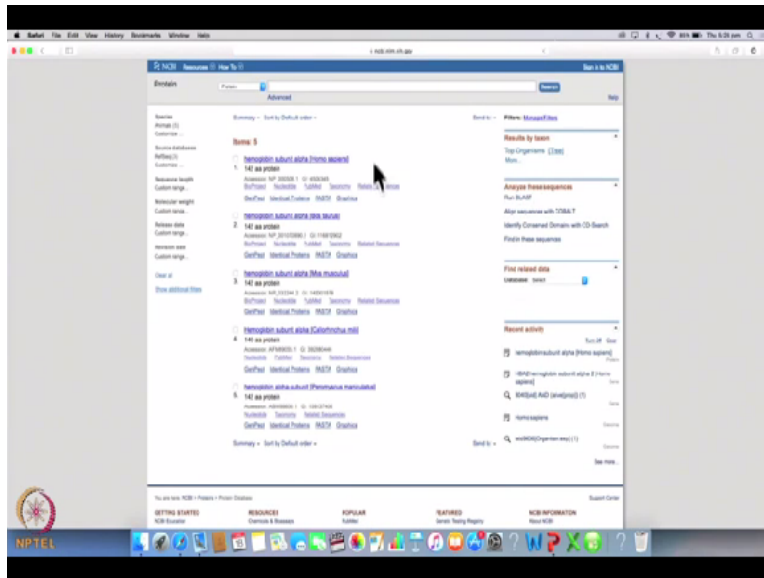
In the fourth part of the tutorial, you have been given a list of accession IDs. You have to find out how the sequences are evolutionarily related. For this, we will type all the accession IDs in the search box on NCBI home page.

(Refer Slide Time: 19:06)

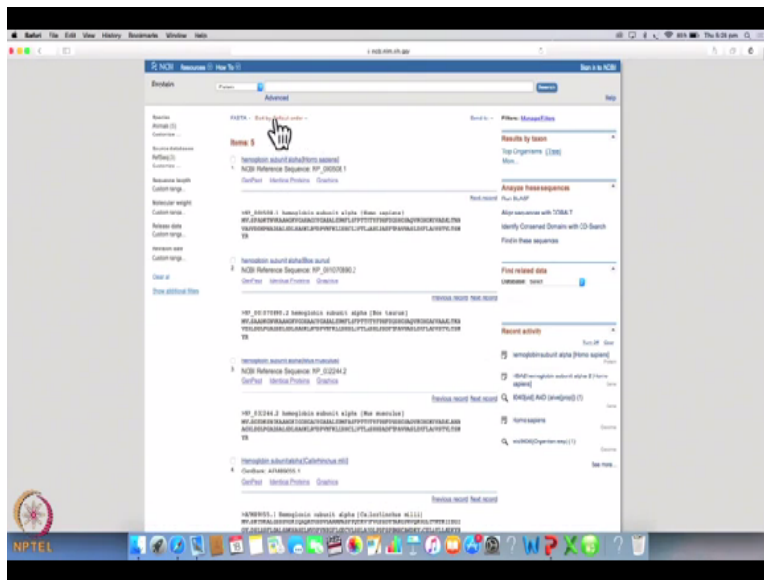


Now we will choose protein from the dropdown menu and we will click search.

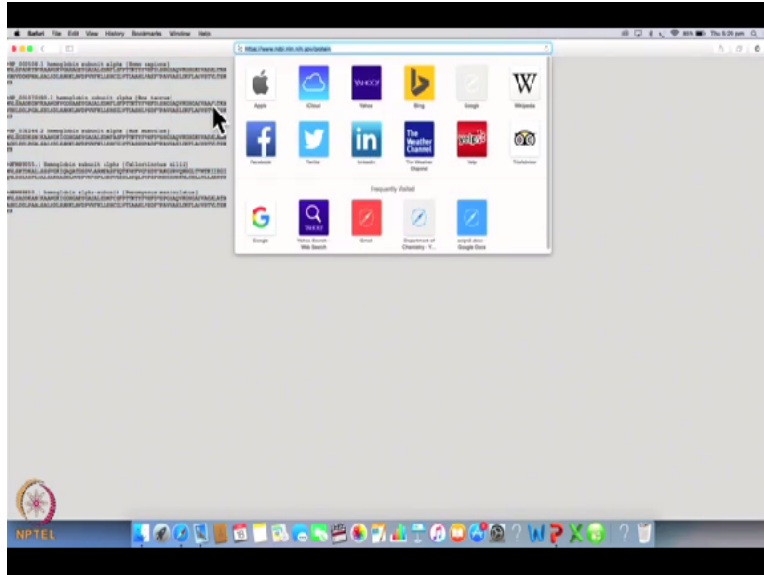
(Refer Slide Time: 19:18)



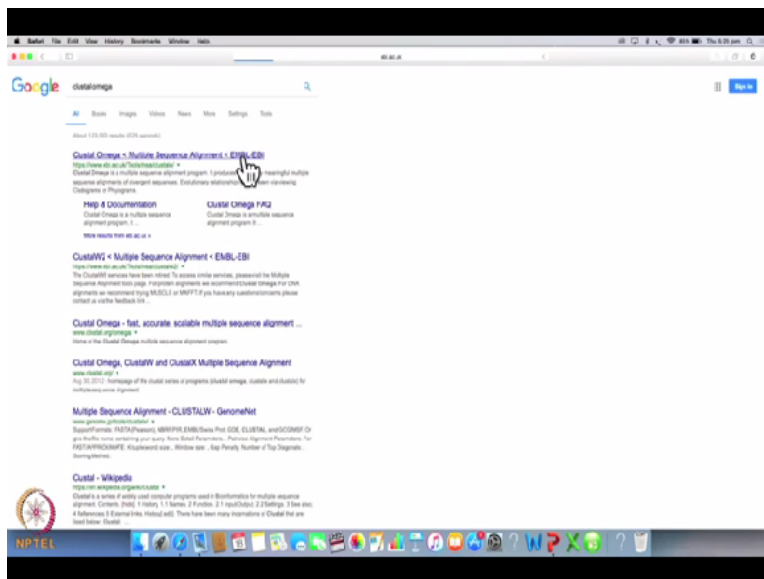
We will now be able to see all 5 proteins on a single page. So we do not have to go to each protein individually to take the FASTA sequence. Simply click on summary and click on FASTA. (Refer Slide Time: 19:40)



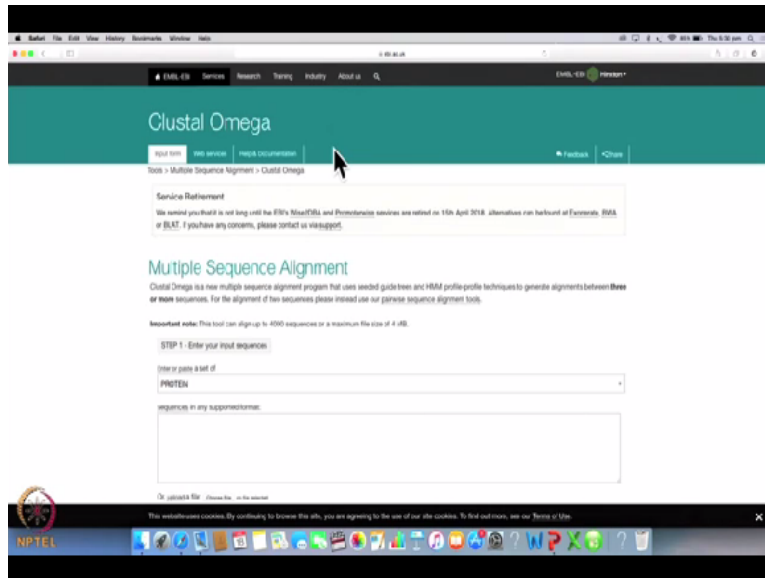
Or we can also go to FASTA text so that we get all sequences on 1 page. (Refer Slide Time: 19:51)



So let us copy them, go to Google browser and type Clustal Omega.
(Refer Slide Time: 20:14)

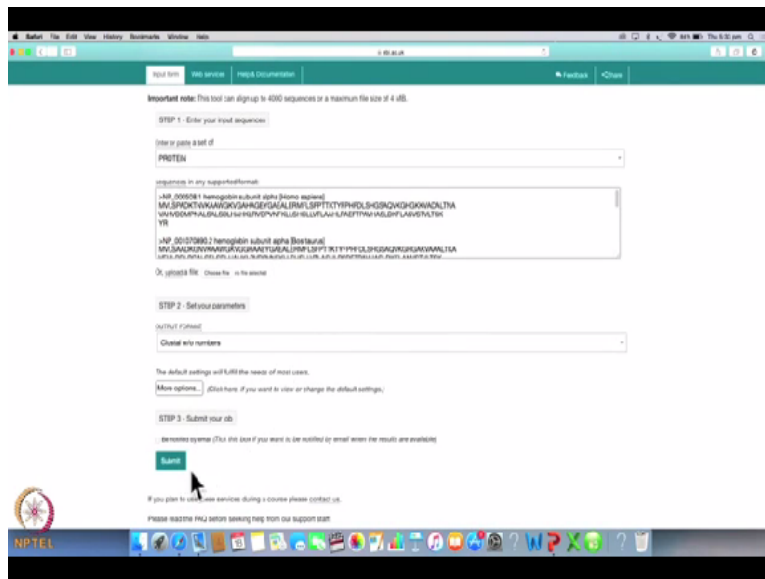


Clustal Omega is a multiple sequence alignment program. It allows us to perform multiple sequence alignments of divergence sequences. Evolutionary relationships can be seen via cladograms or phylograms.
(Refer Slide Time: 20:39)



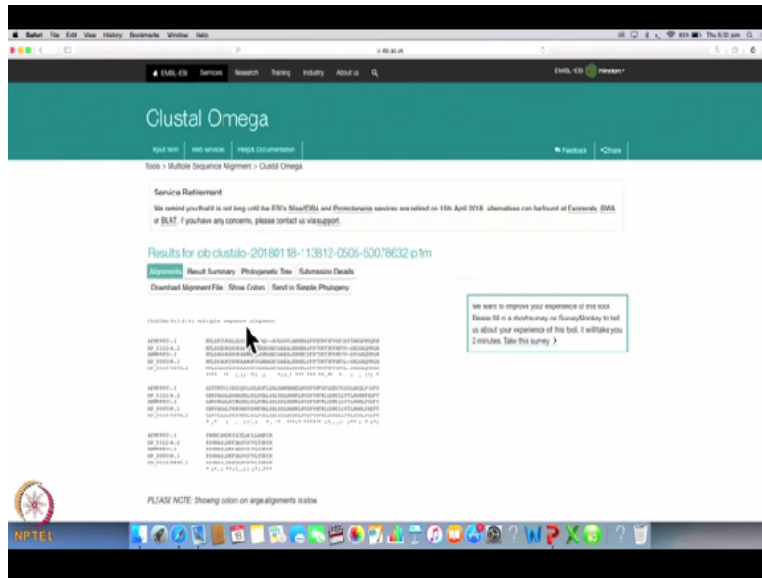
Copy the FASTA files into the box here.

(Refer Slide Time: 20:50)

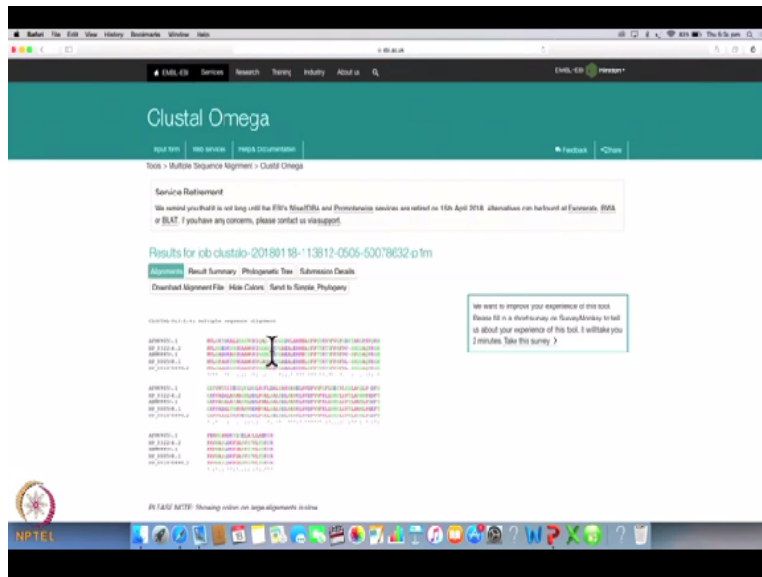


Keeping all other default parameters constant, click submit.

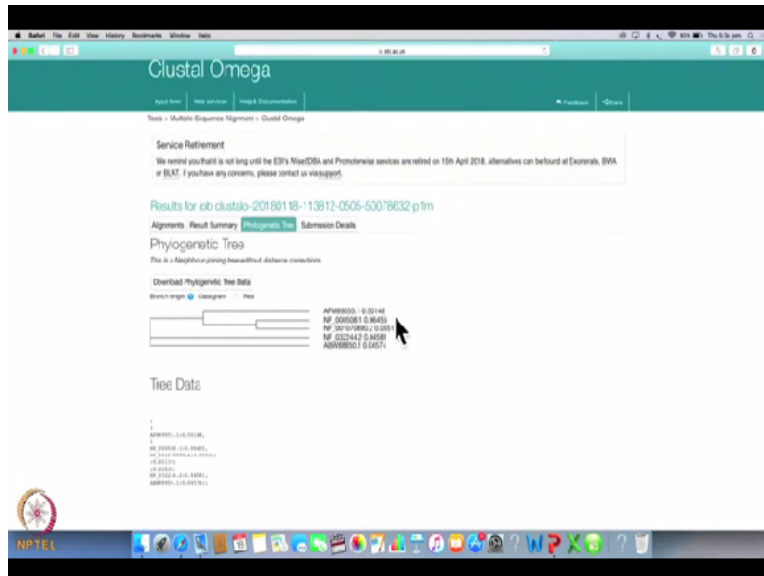
(Refer Slide Time: 20:59)



So as you see the sequences of all the 5 proteins are lined up against each other.
(Refer Slide Time: 21:13)



If you select show colours, you can visualize variations and gaps in the sequences.
(Refer Slide Time: 21:240)




Now click on the phylogenetic tree to get an idea about the evolutionary relationships. What you see here is the cladogram. This is a neighbour joining tree without distance corrections. Now let us copy this into our document.

(Refer Slide Time: 21:41)

Assignment 1D. How are these sequences related?

Multiple Sequence Alignment (MSA)	
Accession No.s	a) NP_000508.1, b) NP_001070890.2, c) NP_032244.2, d) AFM89055.1, e) ABW88850.1
Evolutionary related	



AFM89055.1 0.50148 d- *Callorhinchus milii*

NP_000508.1 0.06455 a- *Homo sapiens*

NP_001070890.2 0.05616 b- *Bos taurus*

NP_032244.2 0.04581 c- *Mus musculus*

ABW88850.1 0.04574 e- *Peromyscus maniculatus*

A and B are closely related

While D is somewhat related to A and B

C and E are distantly related from all

7

So let us paste the cladogram in this document. You will see here that A and B are closely related while B is somewhat related to A and B, while C and E are distantly related from all. So this is how you will get an idea about evolutionary relationships. This brings us to the end of the bioinformatics assignment 1. Hope this was very useful. Thank you.

(Refer Slide Time: 22:02)

Bioengineering: An Interface with Biology & Medicine

BIOINFORMATICS ASSIGNMENT 1

How to use NCBI to study genes and evolutionary relationships

THANK YOU



Alright so in conclusions I am sure you are now convinced that just by doing bioinformatic analysis using some open access tools, you can get lot of understanding of biologically relevant information. In this assignment, we slowly built your, you know, confidence, you try to do some information for looking at the gene, its sequences, its sequence homology and its different similarity with that sequences.

We then try to explain you how computational tools could be used to study the evolutionary relationship just based on the protein sequences. In another lecture in the coming week, we will also try to study how to visualize protein structures and protein-protein interactions using some of the bioinformatic tools. We will also study the pathway analysis and we will learn how to perform molecular assimilation and docking experiments to study the protein drug interactions. Thank you and see you next week.