Applied Environmental Microbiology

Dr. Gargi Singh
Department of Civil Engineering
Indian Institute of Technology, Roorkee

Lecture - 59
Bioinformatics IV

Dear students, welcome to our lecture on bioinformatics, where we continue our conversation on the importance of bioinformatics. And move on now to actually applying some easily accessible and readily available techniques from bioinformatics to make sense of our data. So, one of the most confusing data that we get in applied environmental microbiology is that of sequences whether we have genetic sequences, protein sequences or c DNA sequences. Now, if you remember c DNA is complementary DNA. So, basically you are you are getting the sequence for an RNA.

So, once you have got the sequence from your sequencer machine or from the sequence provider, the next step is to understand; what the sequences at the very minimum in an alphabetical way. So, in at the very least you should if it is a genetic sequence you should have some file, that you can read and it says atjcn so, on and so, forth. Now, if you have used applied bio systems platform, then it might have a particular different extension, and you might require certain software to open the file read it and then align your sequence annotate it.

You remember that for examples many years ago, applied bio systems sequencers would have would give us the output in form of dot ab 1, and there is a free sequence reader available finch tv and we used to use that finch tv in order to read the sequences. So, once you have read the sequences you can control c, that is you can copy the sequence and then now you can proceed online to do some very easy alignment of your sequences and get some basic idea what it might be actually; what actually might be your sample.

Now, what I am going to show you today in this particular lecture is, only relevant when we are dealing with small number of sequences. Let us say up to 50 or maybe up to 100, but definitely not 10000 or 20000 or 1 lakh, 1 million and so on and so, forth. And the reason for that is because these online platforms are not built to entertain such high amount of input, they are built for short in a small number of sequences and when we are talking about high number of sequences; typically we have meta genomic sequences, whether they are first, second, third or fourth generation sequencing, but whether they are from first, second, third or four generation sequencing techniques.

And in this case we have separate platforms, and I would; in this lecture in the next lecture I will be talking about one such platform, which is easily available and it is very accessible for people who have very little experience with programming and very little experience with Linux based software.

So, let us start with your let us start with Sanger sequencing. So, for example, if you get Sanger sequencing done for one of your clones, and your chief interest is to find out what is your microbe that from which you made this clone. So, in order to find the microbe that is answering the question, who is present in my sample? First of all you need to be able to open up the sequencing mush file and then if it is in faster form f a s d a, then simple notepad would open it if it is in fast q form again a notepad can open it, but if it is in other forms you might require specialized software as I mentioned earlier.

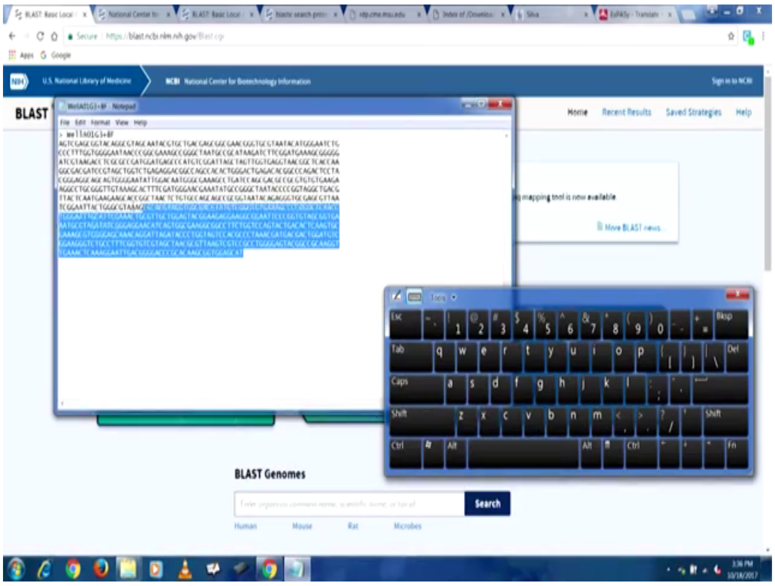
Now there is a chief difference between fast a file and fast q file fast a file basically is just your sequences. So, a, d, d, g, c, a so, on and so, forth would be FASDA file there is a particular format of FASDA file and I will be showing you very soon, what the format is fast q file for every nucleotide you will have a quality score.

So, let us say my sequencer said all right; I am reading a here, but then the sequencer is not very confident about, whether a is present in this position or not in that position in the sequence in the genetic element or not. So, in that case the quality score would be low, but if sequencing machine is very clear that yes, this is what is present, then the quality score will be high. This quality score which always is written below the corresponding nucleotide is very very useful; when especially in meta genomics when we tend to generate a lot of errors.

In Sanger sequencing is produces Sanger sequencing is one of the safe and reliable sequencing techniques, because the error rate is very low compared to other rapid and high throughput sequencing techniques. So, once you have opened a file and you know it is where the FASDA or fast q you need to convert fast q into FASDA in order to be able to use the online tools, because they do not account for the quality scores; that is for you to do some basic bioinformatics and eliminate the base pairs that you are not very sure about.

So, if there is a base pair that has very low score according to you can remove it and say gap I do not know what is here anyway. So, once you have removed the quality scores and you are left with FASDA file; your FASDA file would look something like this.

(Refer Slide Time: 05:28)



```
> wellA01G3+8F
AGTCGAGCGGTACAGGCGTAGCAATACGTGCTGACGAGCGGCGAACGGGTGCGTAATACATGGGAATCTG
CCCTTTGGTGGGGAATAACCCGGCGAAAGCCGGGCTAATGCCGCATAAGATCTTCGGATGAAAGCGGGG
ATCGTAAGACCTCGCGCCGATGGATGAGCCCATGTCGGATTAGCTAGTTGGTGAGGTAACGGCTCACCAA
GGCGACGATCCGTAGCTGGTCTGAGAGGACGGCCAGCCACACTGGGACTGAGACACGGCCCAGACTCCTA
CGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCGACGCCGCGTGTGTGAAGA
AGGCCTGCGGGTTGTAAAGCACTTTCGATGGGAACGAAATATGCCGGGCTAATACCCCGGTAGGCTGACG
TTACTCAATGAAGAAGCACCGGCTAACTCTGTGCCAGCAGCCGCGGTAATACAGAGGGTGCGAGCGTTAA
TCGGAATTACTGGGCGTAAAG[illegible]
TGGGAATTGCATTCGAAACTGCGTTGCTGGAGTACGGAAGAGGAAGACGGAATTCCCGGTGTAGCGGTGA
AATGCGTAGATATCGGGAGGAACATCAGTGGCGAAGGCGGCCTTCTGGTCCAGTACTGACACTCAAGTGC
GAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGACGACTGGATGTC
GGAAGGGTCTGCCTTTCGGTGTCGTAGCTAACGCGTTAAGTCGTCCGCCTGGGGAGTACGGCCGCAAGGT
TGAAACTCAAAGGAATTGACGGGGACCCGCACAAGCGGTGGAGCAT
```

. So, know this is a very good example to give you; how a FASDA file would look if you open it in your notepad. Now, here it starts with this particular sign the name of your sample follows this. So, you will put the sign and then you write the name of the sample. In this case the name of the sample is well A01G 3 plus 8F here 8F is describing the primer used. So, because this is my file data from my graduate work, I know that this is this sequence is of 16S rRNA gene.

So, if you remember what I told you about 16S rRNA gene; 16S rRNA gene (Refer Time: 06:11) is the measure of fine is a is a gold standard for finding out who is present in the sample if it is bacterial sample. And the reason for that is, because 16 SRRNA has conserved domains and as hyper variable domains. So, we aligned the conserved domains and all the 16S rRNA sequences that we have. And we notice how the variable domain is changing. And on basis of the variable domain differences we can decide, whether our sequence is similar to something, that is known and how similar it is and in this way we get an idea where there are sample is closer to bacteria a or bacteria b. 16 s 16S rRNA was the first that was proposed by doctor Sanger and it is very successfully used and even now, it continues to be a world standard despite some limitations which I have already taught in our lectures.

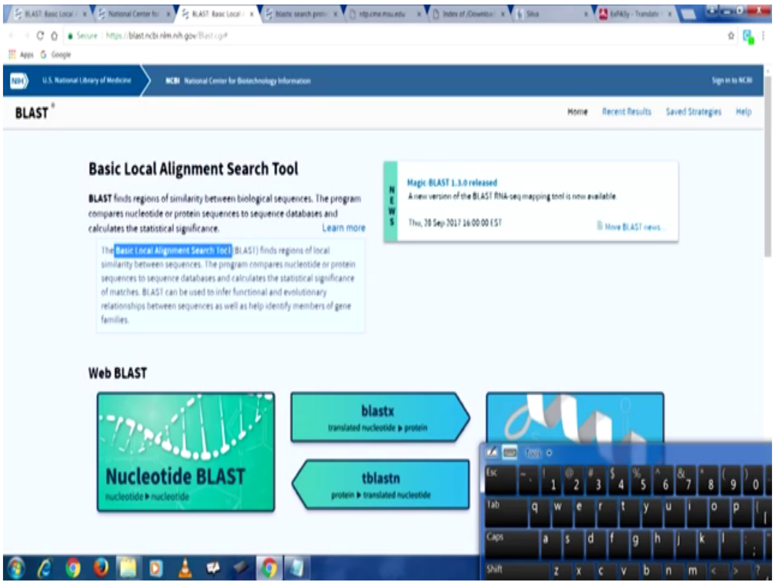
So, if you are not clear about, what 16S rRNA genus I highly recommend you go back to previous lectures skim through them and find out what this gene is why it is so important; and also with the general assumption that e bacterial cell will have a constant number of 16S rRNA gene beat 1 beat 2, and we know some back rubs that have very high amount of 16S rRNA genes by the way, but let us say if your gene 1.4 or 4.1 16S rRNA gene per microbial cell, then knowing the count of 16S rRNA gene by yq PCR which is quantitative polymerase chain reaction it will be easier for us to find out how many microbial cells are present.

So, if we do quantification of 16S rRNA gene, it tells me how many bacteria are present and; obviously, the next question is what if there is a bacteria it does not has 16 SRNA gene, there is no such bacteria that does not have 16S rRNA gene it is universal that is why it is used for quantifying bacterial populations all right.

So, this is how your FASDA file would look like, there will be the sign followed by the name of the sample and 8F here by the way, 8F here is the name of the primer 8 forward starts from the position 8 and 16S rRNA gene and I do remember that this particular sequencing we I supplied to my sequencing agency 8F primer to use that as the starting point of sequencing that is what they have labelled it 8F; F stands for forward and there's a corresponding reverse primer too, but I wanted it to be where forward.

Now, what follows this is a sequence of alphabets ATGC, and this is actually the sequence of your genetic material. So, once you have opened this file we can control see this, and then open up blast NCBI NLM dot nih dot gov slash blast dot CGI. An easy way to approach this would be just look up online NCBI blast and here you go. So, now, you can choose what kind of blast you want to do; but if you click the first or the main title; this is where you will end up.

(Refer Slide Time: 09:23)



BLAST®

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

NEWS

Magic BLAST 1.3.0 released

A new version of the BLAST RNA-seq mapping tool is now available.

Thu, 28 Sep 2017 16:00:00 EST

More BLAST news...

Web BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
Home
Recent Results
Saved Strategies
Help
Nucleotide BLAST
nucleotide ▸ nucleotide
blastx
translated nucleotide ▸ protein
tblastn
protein ▸ translated nucleotide

So, this is where we are now let us look at the different options we have here we have web blast which in this all this is web blast. So, all this would be done online, but would be nice idea first to find out what blast is? So, let us find out what blaster is? Blast stands for basic local alignment search tool. So, it is a very basic local alignment local alignment as in it does it locally, and it is a search tool because it searches what aligns the best. It finds the regions of local similarities between sequences.

So, overall similarity is not necessary, but if you look for local similarities here it is similar to this microbe, here it is similar to this microbe, almost everywhere it is similar to that microbe and that information it will report back to us. The program compares nucleotide or protein sequences now note here; blast can do both nucleotide comparison and protein comparison to sequence databases and who does it compare it to; with sequence databases.

So, there are two things happening here one we have the sequencer that we have generated from our environmental sample and second we have databases that already have annotated sequences annotated means the ones we know where they came from which microbe they are belonging to and we have all the information maybe they have been fully sequenced them also.

So, now, what the blast will do is; it will do a local alignment for all nucleotides in the same sequence that I have generated through my environmental samples and compare it to all the data as entries in the database. And then it will calculate similarity and then rank them in order of increasing similarity to decreasing similarity. So, in the decreasing order it will rank them.

Blast can be used to infer functional and evolutionary relationship between sequences as well as to help identify number of gene families. Now, if you remember we did talk about function similarities and phylogenetic similarities. So, if the genetic sequence is similar we assume that functional characteristics would be also similar. For example, if there are two microbes and both of them have MRSA gene which is which is responsible for methicillin resistant staph staphylococcus aureus, then we can assume that both the microbes that have MRSA gene are likely to be resistant to methicillin.

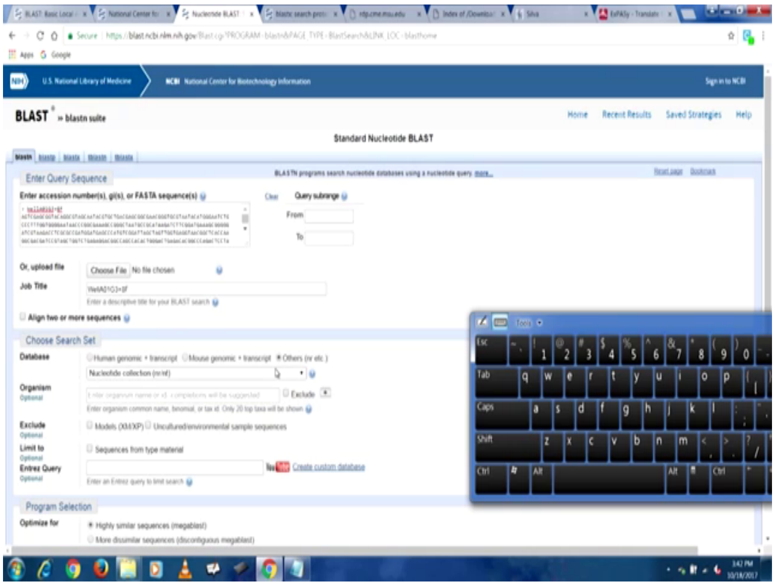
Now there is a word likely, because lot happens in on the regulation level, but likely similar resistant to methicillin exposure. So, the and if you do a blast, if it sequence MRSA genes and we do not know if your sequencing MRSA genes or we just sequence it and then we blast it, we align it, and with local we do local alignment and we find out that the genes that we have separated isolated from the two bacteria a very very similar to MRSA in fact, they are so similar we can say comfortably oh this is MRSA.

Then we can assume; we can infer that quite some probability quite some backing that the gene that we have isolated confers methicillin resistant to the microbe. So, in this way it gives us information or they related to functional of the gene function of the microbe the other is evolutionary. So, remember I told you in 16S rRNA what we do is we have conserved domains and then we have hyper variable domains. So, concern domains assumed to be almost same for all kind of bacteria, but hyper variable regions change.

So, depending on how far the hyper variable regions are from each other which means; how dissimilar they are, if you get an idea of how different they are from evolutionary perspective; how half distant they are, what it implies is that their common ancestor was very long back, because the ones that are close cousins will have more genetic similarity than the ones that are genetically that are evolutionary very far away. So, it gives us information on both functional and evolutionary perspective.

Now, let us look at what options we have here in blast. So, we have web blast online blast we have nucleotide blast x, t blast n and protein blast; and it is pretty self evident actually. Nucleotide blast is when I take a nucleotide sequence this is by the when you get that sequence and you can look at it very carefully and you can look at it very carefully that the only alphabet you will find here would be ATGC. So, you know definitely that this is a nucleotide sequence. So, if you want to blast this, if you want to compare this and get data related nucleotide.

(Refer Slide Time: 13:43)



BLAST » blastn suite

Standard Nucleotide BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Or, upload file: Choose File No file chosen

Job Title: [illegible]

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.)

Nucleotide collection (nr/nt)

Organism Optional: □ Exclude

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional

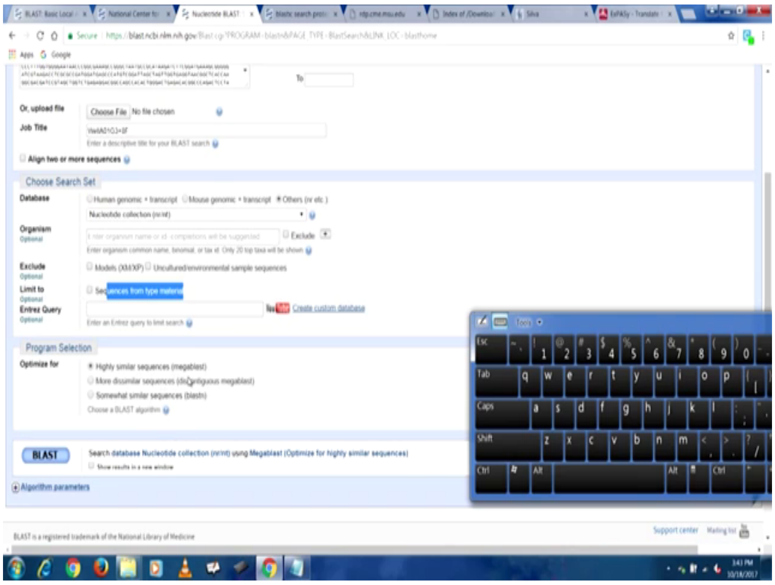
Program Selection

Optimize for: ● Highly similar sequences (megablast) ○ More dissimilar sequences (discontiguous megablast)

So, basically what will happen is that; the NCBI blast will take your sequence will take this nucleotide base sequence; and then compare it to all the nucleotides in the databases all the sequences and the nucleotides in the databases and then tell you what it is most similar to. So, let us try this. Blast n which is the first one. So, this is what it will look like. So, we just need to do control c and my suggestion is let us select all of it ok.

So, now, that we have input our data and look I have still kept the name of the file here, and the name of the file is automatically come being input here. I am interested in nucleotide collection, others, nr I am not interested in mouse genomes and it is transcript, I am not interested in human genome, I am not interested in anything, but let us write uncultured environmental samples. And I can choose that I want to limit let us remove this, because something fun we will see here. I want to limit my sequences from one type of material I only want matches the from hydrocarbon contaminated soils.

(Refer Slide Time: 15:04)



Or, upload file: Choose File No file chosen

Job Title: WellA01G3+8F

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.)

Nucleotide collection (nr/nt)

Organism (Optional): Enter organism name or id - completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional): Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for: ● Highly similar sequences (megablast) ○ More dissimilar sequences (discontiguous megablast) ○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST — Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

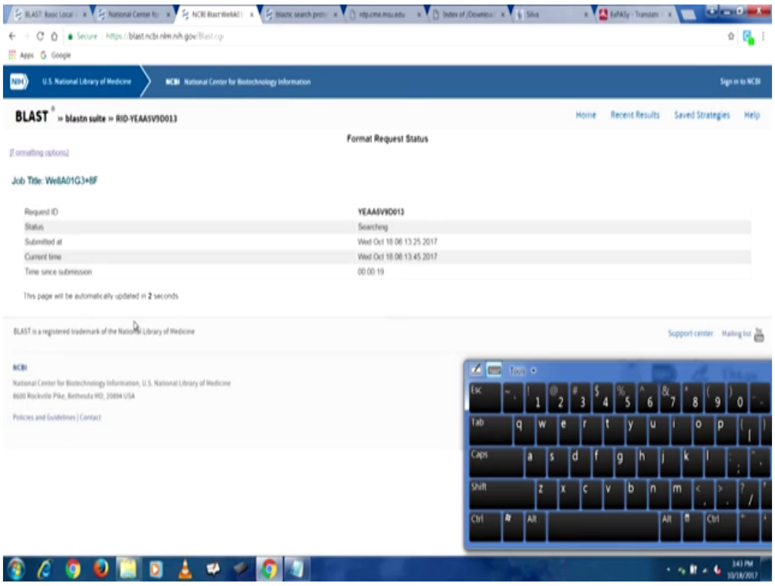
□ Show results in a new window

Algorithm parameters

BLAST is a registered trademark of the National Library of Medicine

And then I can choose highly dissimilar more dissimilar or somewhat dissimilar; somewhat similar sequences and then I can just press blast here; they are more algorithmic well parameters that I can change.

(Refer Slide Time: 15:17)



U.S. National Library of Medicine — NCBI National Center for Biotechnology Information

BLAST » blastn suite » RID-YEAASV9D013

Format Request Status

Job Title: WellA01G3+8F

Request ID	YEAASV9D013
Status	Searching
Submitted at	Wed Oct 18 06:13:25 2017
Current time	Wed Oct 18 06:13:45 2017
Time since submission	00:00:19

This page will be automatically updated in 2 seconds

BLAST is a registered trademark of the National Library of Medicine

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine

8600 Rockville Pike, Bethesda MD, 20894 USA

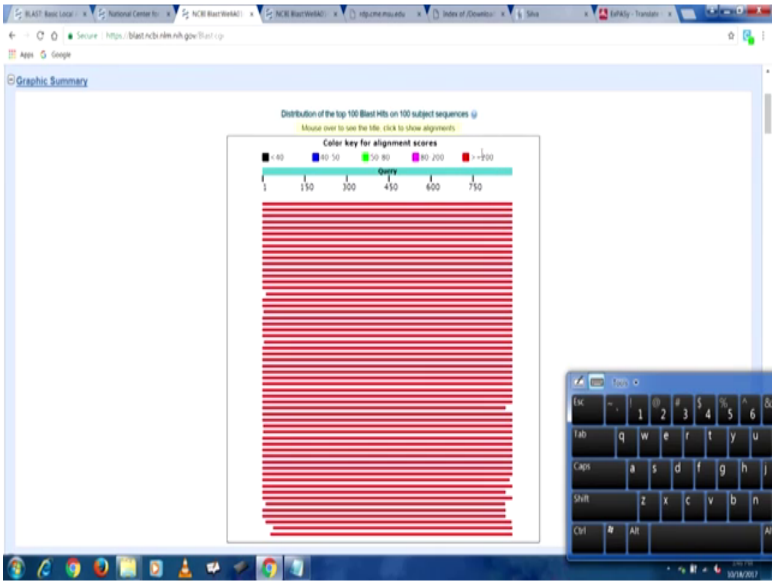
Now, this is the page you will get. So, what will happen is that you will be aligned a job title; and then you will be given a request id and then it will let you know in how much time has elapsed; since you made the request and how much more time you need to know for the next trial. So, it will continue doing this meanwhile let us move on to other this is blast x.

So, if you look at blast x, what it does is; it translates a nucleotide sequence. So, what alpha had ATGC, but what blast x will do for us; if you post the same thing here so we have pasted the same thing here same sequence exactly same name. And, now if I try to do blast x for it not blast n, what it will do is; it will take my nucleotide sequence look at all the reading frames in the 6 reading frames possible converted into 6 different proteins and then align the proteins and then the one that matches the best it will say all right this might be this all right.

So, now, let us look at what we are interested in and let us blast it. So, now, we have two blast requests going on; when is this one and when is this one it might take time around 2, 3, 4, 5 minutes depending on the traffic. So, if it is very active in the part of the globe, if this is the timing people are actively blasting their sequences it is likely to take longer time. Let us look at what other options we have. We have t blast n. So, what t blast n will do? If you have a protein sequence so, protein sequence will not look like ATGC alphabets are again used to write amino acids name like MRTP, but they are not limited to ATGC they are much more variable; obviously,. So, when you have a protein sequence you want to do t blast n.

Ah n n here is the thing you want to do t blast n; when you have a protein sequence and you want to compare it to the nucleotides not to other proteins, if you want to compare your protein sequences to other proteins, they need to protein blast or here you will read it as blast b all right. So, we have some data here, ok.

(Refer Slide Time: 17:13)



Graphic Summary

Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

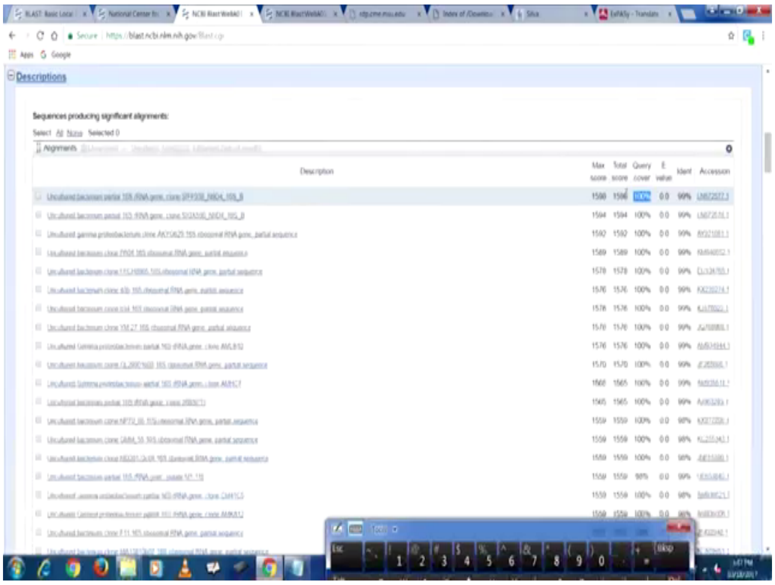
■ <40 🟦 40-50 🟩 50-80 ■ 80-200 🟥 >=200

This was blast n when very very compared our nucleotide to other nucleotide. So, you can read it here this is a blast n sweet and this is our RID number, this is our query id, this is the description name, this is the nucleic acid our query was 886 base pair long, the database you wanted to use was nr which is nucleotide collection and we did blast n. So, let us take a look at it.

So, look here colour key for alignment score; whenever the alignment is really good, the score is really good, the colour would be red and then pink and then green and blue and then black. So, notice here everything is mostly red there are some gaps. So, basically it is saying there are some gaps in the sequences and remember NCBI blast one of the key feature is that it allows for gaps.

If it would not allow for let us say this initial small gap, then all of these would have suffered; all of these would have been green, green blue black and the simple reason for is that when I am not allowing a gap then; I am introducing a frame shift mutation all right. Let us see what it is similar to.

(Refer Slide Time: 18:18)



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Descriptions

Sequences producing significant alignments:

Select All None Selected 0

Alignments

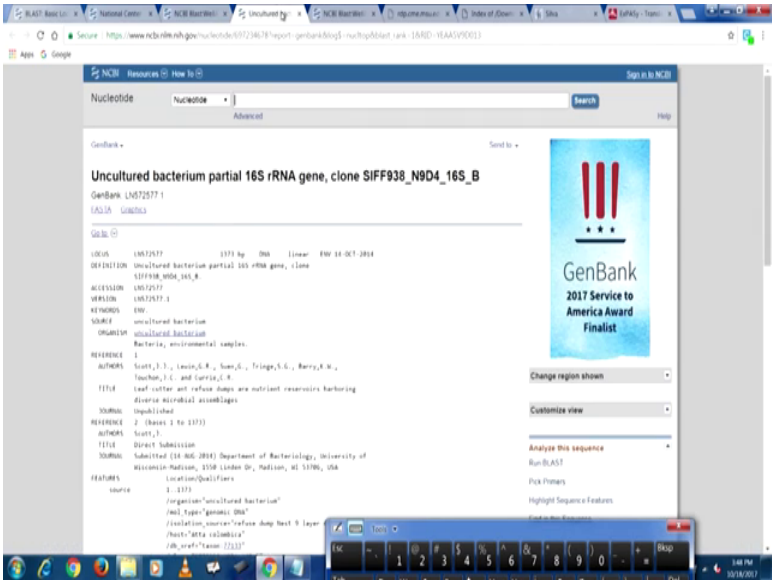
Description	Max score	Total score	Query cover	E value	Ident	Accession
Uncultured bacterium partial 16S rRNA gene, clone SPH938_N9D4_16S_B	1598	1598	100%	0.0	99%	[illegible]
Uncultured bacterium partial 16S rRNA gene, clone [illegible]_N9D4_16S_B	1594	1594	100%	0.0	99%	[illegible]
Uncultured gamma proteobacterium clone AKYG625 16S ribosomal RNA gene, partial sequence	1592	1592	100%	0.0	99%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1589	1589	100%	0.0	99%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1578	1578	100%	0.0	99%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1576	1576	100%	0.0	99%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1576	1576	100%	0.0	99%	[illegible]
Uncultured bacterium clone YM 27 16S ribosomal RNA gene, partial sequence	1570	1570	100%	0.0	99%	[illegible]
Uncultured Gamma proteobacterium partial 16S rRNA gene, clone [illegible]	1576	1576	100%	0.0	99%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1570	1570	100%	0.0	99%	[illegible]
Uncultured Gamma proteobacterium partial 16S rRNA gene, clone [illegible]	1568	1565	100%	0.0	99%	[illegible]
Uncultured bacterium partial 16S rRNA gene, clone [illegible]	1565	1565	100%	0.0	99%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1559	1559	100%	0.0	98%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1559	1559	100%	0.0	98%	[illegible]
Uncultured bacterium clone [illegible] 16S ribosomal RNA gene, partial sequence	1559	1559	100%	0.0	98%	[illegible]
Uncultured bacterium partial 16S rRNA gene, isolate [illegible]	1559	1559	98%	0.0	99%	[illegible]
Uncultured gamma proteobacterium partial 16S rRNA gene, clone [illegible]	1559	1559	100%	0.0	98%	[illegible]
Uncultured Gamma proteobacterium partial 16S rRNA gene, clone [illegible]	1559	1559	100%	0.0	98%	[illegible]
Uncultured bacterium clone F11 16S ribosomal RNA gene, partial sequence	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

So, if you scroll down you will see what it is similar to. So, we notice that there are lot of uncultured bacteria that is what it is similar to. So, this uncultured bacterium partial 16S rRNA gene so, we are right here is it, because I know that the sequences from 16S rRNA uncultured gamma proteobacteria so and so forth; what open interest people is find out the first cultured; that shows up in the sequencing list oh all uncultured here all right. So, learn that we will go through and we will find out one that is cultured ok.

So, this is a maximum score this is your total score this is query coverage which means at how many percentage of base pairs were covered in this alignment process and here it 100 percent e value, it is not 0 exactly it is nearly 0. We want e value to be as low as possible it is the possibility of having false error false positive and this is identity. So, my semi my sequence is 99 percent similar to this particular sequence and this accession idea of the sequence is this, if I click on this it will open another page and it will give me more information about the particular bacteria.

Now, remember I did not submit this bacteria; I did not submit this uncultured bacterium partial 16S rRNA gene cloned (Refer Time: 19:45) 938 9 N 9 D 416 SB somebody else did but.

(Refer Slide Time: 19:49)



Uncultured bacterium partial 16S rRNA gene, clone SIFF938_N9D4_16S_B

GenBank: LN572577.1

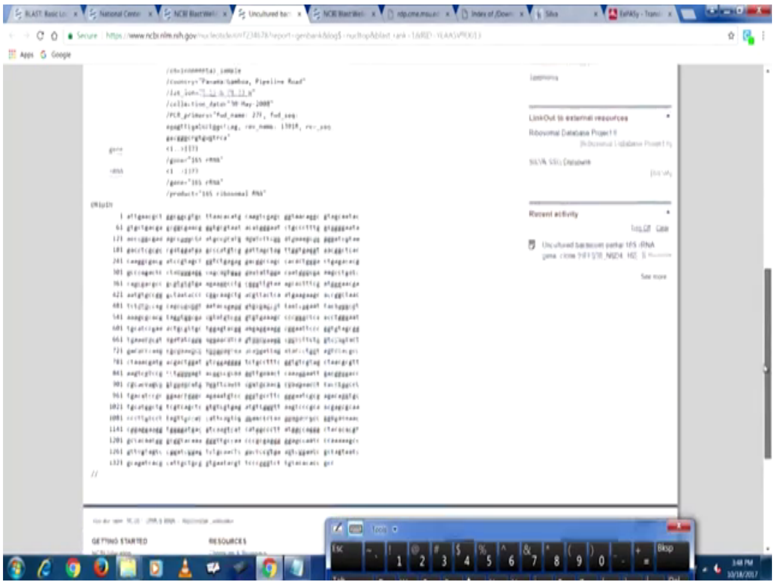
FASTA Graphics

```
LOCUS       LN572577               1373 bp    DNA     linear   ENV 14-OCT-2014
DEFINITION  Uncultured bacterium partial 16S rRNA gene, clone
            SIFF938_N9D4_16S_B.
ACCESSION   LN572577
VERSION     LN572577.1
KEYWORDS    ENV.
SOURCE      uncultured bacterium
  ORGANISM  uncultured bacterium
            Bacteria; environmental samples.
REFERENCE   1
  AUTHORS   Scott,J.J., Lewin,G.R., Suen,G., Tringe,S.G., Barry,K.W.,
            Touchon,J.C. and Currie,C.R.
  TITLE     Leaf-cutter ant refuse dumps are nutrient reservoirs harboring
            diverse microbial assemblages
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1373)
  AUTHORS   Scott,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-AUG-2014) Department of Bacteriology, University of
            Wisconsin-Madison, 1550 Linden Dr, Madison, WI 53706, USA
FEATURES             Location/Qualifiers
     source          1..1373
                     /organism="uncultured bacterium"
                     /mol_type="genomic DNA"
                     /isolation_source="refuse dump Nest 9 layer
                     /host="Atta colombica"
                     /db_xref="taxon:77133"
```

Now, I have information and I NCBI will allow you to actually upload your document here. So, this is Scott Suen Tringe Barry Tasha and Curie.

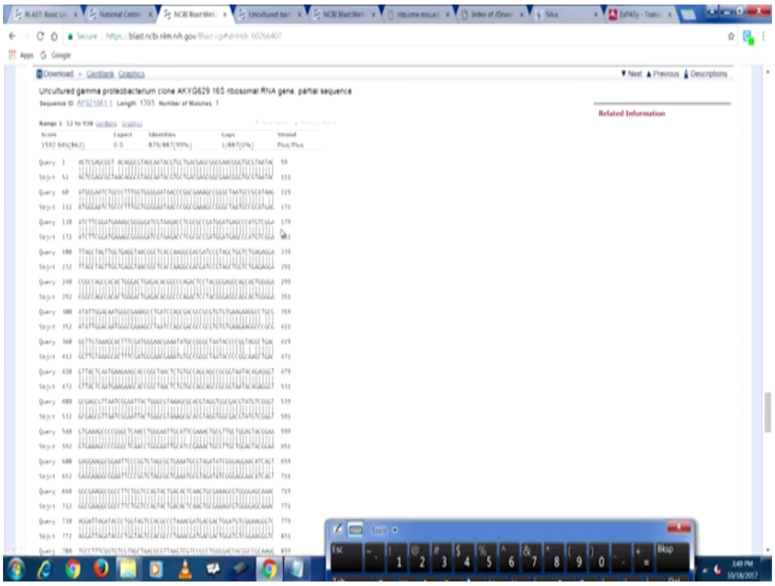
And the title of the paper their work is leafcutter ant refuse dups on nutrient reservoirs harbouring diverse microbial as in assemblages. So, we are looking at some particular ant leaf cutter ant refuse dups. So, we are looking in the excretory material of an ant and at the time of submission the paper was not published and it hit right had been submitted, but it was not published all right. So, I give some information that this Sequence that I submitted matches sequence found in ant tub all right.

(Refer Slide Time: 20:25)



Now let us look for one that is all of them are uncultured all right. So, I can go and take a look at the other one.

(Refer Slide Time: 20:44)



Download · GenBank Graphics

Uncultured gamma proteobacterium clone AKYG629 16S ribosomal RNA gene, partial sequence

Related Information

Score	Expect	Identities	Gaps	Strand
1592 bits(862)	0.0	875/887(99%)	1/887(0%)	Plus/Plus

And look if you go scroll below, let us look at the one that matched the best; if you scroll below NCBI has left a gap here. So, wherever NCBI has gaped left a gap you can take a look and notice here that this is a really good match all right.

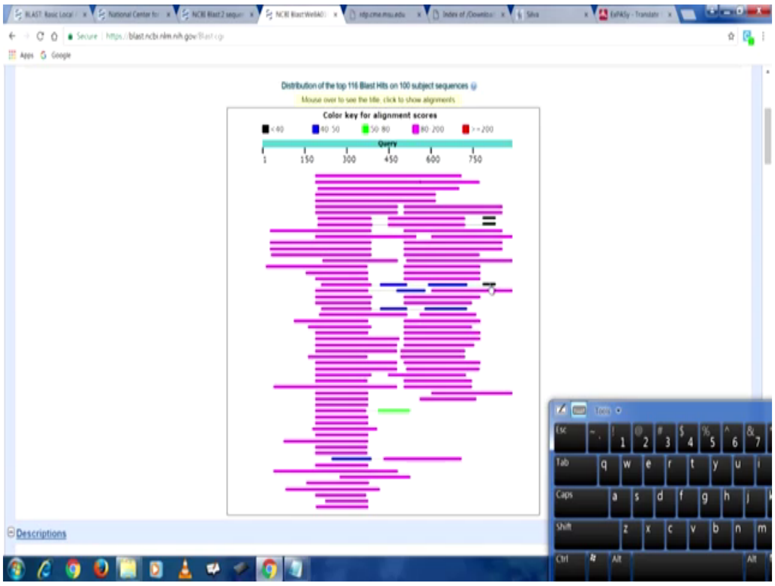
But here is the problem with this I still do not know what bacteria it is most similar to. So, in that case what I can do is; I can download the entire analysis and I can find out where is the first cultured bacteria and what it is. So, uncultured bacteria I do not know it could be an actino bacteria it could be acineto bacteria it could be (Refer Time: 21:24) I have no idea it could be a proteobacteria.

So, in order to know that; we need to download the whole file, and then you find out, what it is; because it will have all the information all right. And then what we can do is if you submit multiple files you can create distance string of results . So, let us try that. So, let us edit and resubmit ok. So, this is one file that I have and let us see if we have other files too. This is another file we have. Let us copy paste this here in this file.

There now we have two samples. So, now, we can ok. So, now, we have two samples here yes. So, let us try blasting this. So, here I do not want to have uncultured environmental samples last time remember the entire list was full of uncultured. So, let us say I do not want uncultured, I want to see the cultured ones. So, there I have some information that I can share and then let us blast it ok.

Let us see let us move on and see what is this will take time. So, let us move on and see what happened to our protein.

(Refer Slide Time: 22:56)

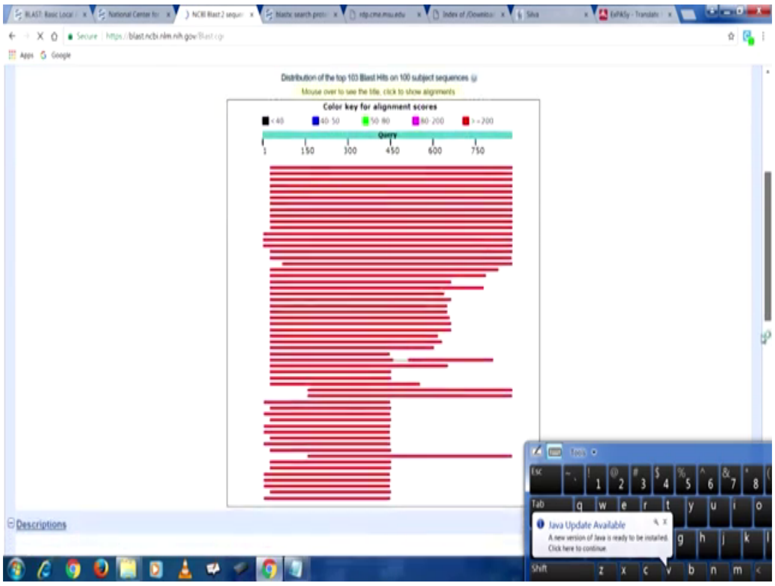

Distribution of the top 116 Blast Hits on 100 subject sequences
Mouse over to see the title, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
150
300
450
600
750
Descriptions

So, we took the same sequence we converted it into protein, we translated it into protein in silico and then let us see what our alignment result our alignment is really poor. Remember in the previous one there were very few gaps all of it was red which means very good alignment, but here alignment score is low sometimes it is very low and they are multiple gap. So, this is not reliable.

So, often where what we do is we take our nucleotide sequence we write to translate it and then blast it with protein database, but at times that might lose the information. So, if you look here, we have where a lot of hypothetical proteins, but look at the then there is also say homo sapiens analogy. So, there's a lot of confusion here just does not make sense I do not see a 16S rRNA in this. So, notice this is one limitation of your blast p. So, you have to be careful what do you want to do blast x or blast n or blast p all right this is still working I guess yes ok.

So, let us select this control a backspace control v there ok. So, last let us try to do blast x this is blast x now with two samples.

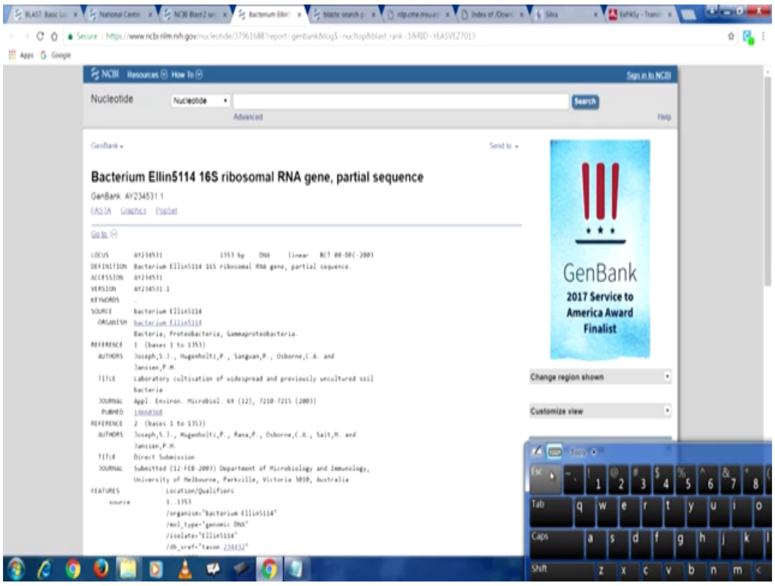
(Refer Slide Time: 24:23)


Distribution of the top 103 Blast Hits on 100 subject sequences
Mouse over to see the title, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
150
300
450
600
750
Descriptions
Java Update Available
A new version of Java is ready to be installed.
Click here to continue.

So, you can I have put with 30 samples at a time. So, I just got copy paste the whole notepad file. So, what if I want to look at the first sample, which have already seen then I just need to click select first sample from here, I am interested in a second I just need to select second sample from here, and it will and it will give me the similarity list. So, here is a similarity notice it has more gaps. Now the another reason why it might have more gaps is, because we said we do not want environmental samples or uncultured samples. We want samples whose information we have very clearly, because unlike examples are not very informative.

So, let us see when we blasted this with known samples we saw more gaps now. So, let us see what is the first one that a bacterium Ellin 5114, 16S rRNA gene all right. So, this is the best match to and, what percentage it is; let us look at the percentage. It matches nearly it covers 100 percent of the query of the entire sequence and 89 percent is your similarity identity. So, this is very helpful information let us look at it is accession number what kind of bacteria this is bacterium ellin all right.

(Refer Slide Time: 25:37)



Bacterium Ellin5114 16S ribosomal RNA gene, partial sequence

GenBank: AY234531.1

FASTA Graphics PopSet

```
LOCUS       AY234531                1351 bp    DNA     linear   BCT 08-DEC-2003
DEFINITION  Bacterium Ellin5114 16S ribosomal RNA gene, partial sequence.
ACCESSION   AY234531
VERSION     AY234531.1
KEYWORDS    .
SOURCE      bacterium Ellin5114
  ORGANISM  bacterium Ellin5114
            Bacteria; Proteobacteria; Gammaproteobacteria.
REFERENCE   1  (bases 1 to 1351)
  AUTHORS   Joseph,S.J., Hugenholtz,P., Sangwan,P., Osborne,C.A. and
            Janssen,P.H.
  TITLE     Laboratory cultivation of widespread and previously uncultured soil
            bacteria
  JOURNAL   Appl. Environ. Microbiol. 69 (12), 7210-7215 (2003)
   PUBMED   14660368
REFERENCE   2  (bases 1 to 1351)
  AUTHORS   Joseph,S.J., Hugenholtz,P., Rana,P., Osborne,C.A., Sait,M. and
            Janssen,P.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-FEB-2003) Department of Microbiology and Immunology,
            University of Melbourne, Parkville, Victoria 3010, Australia
FEATURES             Location/Qualifiers
     source          1..1351
                     /organism="bacterium Ellin5114"
                     /mol_type="genomic DNA"
                     /isolate="Ellin5114"
                     /db_xref="taxon:238832"
```

So, bacterium ellin 5114, 16S this is very good news; it matches 16S ribosomal RNA gene, because that is what I actually amplified. It is a bacteria among which it is proteobacteria gamma proteobacteria. These are the authors this is the title and they have published the paper already. So, I can actually go to the paper and brief what the research was all about wonderful ok.

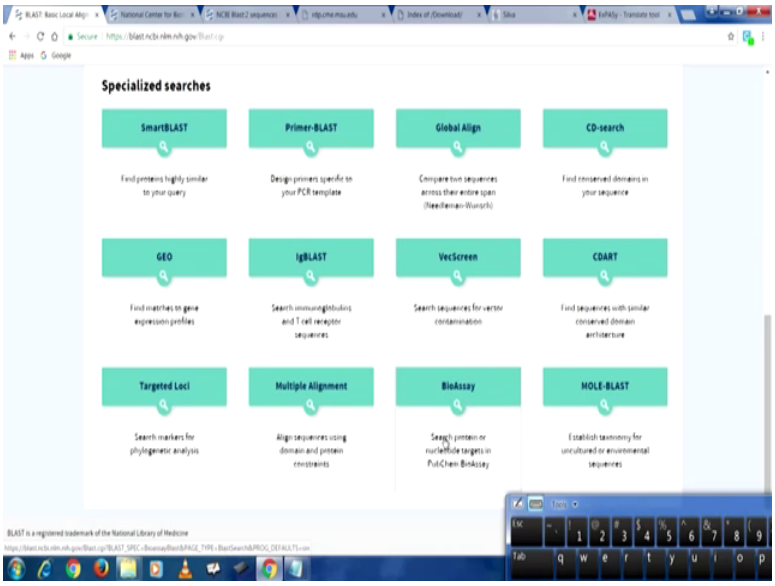
So, now let us go to the second one. So, in order to go to a second one, I just need to choose the second one here and hold a breath all right. So, this one has 882 base pairs I think the other one had 886 something and lot of gaps. So, let us see what the first cultured bacteria that matches it, wow; bacterium ellin 5290, 16S rRNA gene partial sequence. Wow! Same study similar study ok; this is what you have. So, you can look at where the gaps are present, where the gaps are not there lot of gaps here and then other thing we can do is the similarity is 89 percent comparable, but the coverage is 96 percent the 4 percent of the of this sequence was not even covered and we can take a look at.

The bacterium ellin that matches perfectly with it same paper and this is not a gamma proteobacteria this is a gemmat monand monadetes perfect. So, this is how you use blast x and this is also how you use blast n. Now let us look at this is blast x by the way.

So, we have same and led to blast x and the reason is remember when we did blast x for the first sequence they were not very good matches. So, we do not expect very good matches here. Especially, the 16S rRNA gene it is not wise to expect a lot of good matches, but I want to show before we turn off I want to show you one particular example where with nucleotide data, that we have it was better to used better to translated into protein and then dust it then not and we will see why already; now let us come back here to blast. So, we have blast x t blast n and here blast p.

If your protein sequence you want to blast it to other proteins you use blast p. If you have protein sequence and you want to blast it to other nucleotides you use this. Let us see what other facilities NCBI has given us. This is beautiful; NCBI has given a standalone and API blast which basically means you do not have to rely to in on internet you can download the database you can download their commands and then do it in house in your computer.

(Refer Slide Time: 28:14)



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

GEO
Find matches to gene expression profiles

IgBLAST
Search immunoglobulins and T cell receptor sequences

VecScreen
Search sequences for vector contamination

CDART
Find sequences with similar conserved domain architecture

Targeted Loci
Search markers for phylogenetic analysis

Multiple Alignment
Align sequences using domain and protein constraints

BioAssay
Search protein or nucleotide targets in PubChem BioAssay

MOLE-BLAST
Establish taxonomy for uncultured or environmental sequences

BLAST is a registered trademark of the National Library of Medicine

So, you do not need any connection to internet and you do not have to share the information, this is really nice what else you can do if you are interested in programming you can actually contribute to making blast and you can you also use blast in Google cloud all right. So, now, let us look at some specialized search as a blast oops; there is a particular option called smart blast that will allow you to find proteins very similar to your query there is a primer-blast is very important have uses before what it does is if I give it my amplicon sequences it will help me design the primer. Now primary design is not very easy we just do not take those ends of an amplicon.

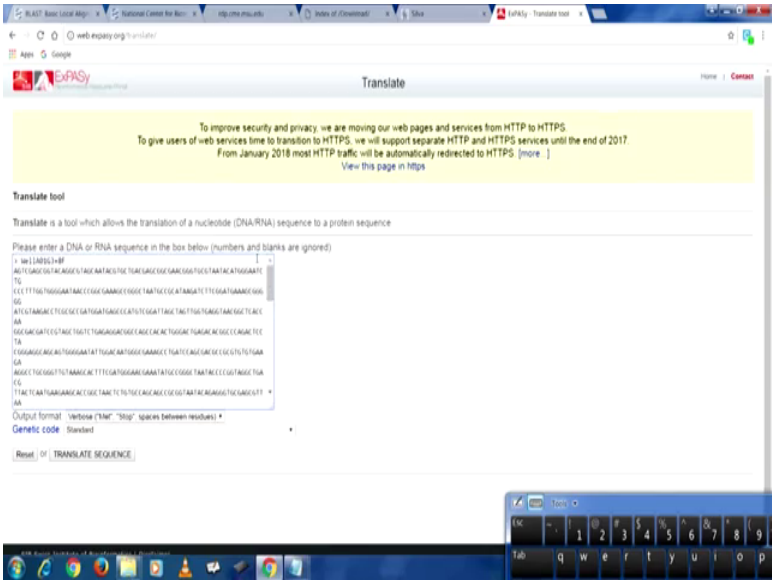
And then say the complementary of this would be a primary design. In fact, they have to be the right the GC content has to be the right and we have to avoid that avoid any possibility that they will stick with each other. So, they will make dimers or they will collapse into themselves and then disrupt all the PCR and downstream products you were interested in. So, primer blasts will help you with all that, then we have global aligned which will compare two sequences across their entire span.

So, not just part of 16S rRNA with part of 16S rRNA, but the entire span then we have CD search which will actually find conserved domains in your searches this is very very important, because once you know conserved domains once you are not worried about specificity, then you can design your own primers and you can design your own sensors next in geo technology. We have geo which is which finds matches to gene expression profiles and, then ig blast it searches immunoglobulin t cell receptor sequences we have vecscreen that looks for vector and termination we have CDART which find sequences with similar conserved domain architecture.

We have mole blast which establishes taxonomy for uncultured environment for sequences which is a very important for us environmental engineers, then it has bio essay which searches protein or nucleotide targets in pubchem bioassay, it is copyright bioassay; then we have multiple alignment what it can do is it can align sequences using domain and putting constraints and then we have targeted loci I am only interested in a particular kind of analysis.

For example, if I have a 16S rRNA gene sequence I know it is 16S rRNA nothing else I can ask NCBI only to 16S rRNA blasting which save my time ok. Here we are and this is your blast x and look the similarity index is very low and the hypothetical protein is lactobacillus Jensen knee percent a similarity 61 percent which is very low. So, we do not want to be very sure of this blast x does not seem to be working for this what you can also do is .

(Refer Slide Time: 31:03).



web.expasy.org/translate/

ExPASy

Translate

Home | Contact

To improve security and privacy, we are moving our web pages and services from HTTP to HTTPS.
To give users of web services time to transition to HTTPS, we will support separate HTTP and HTTPS services until the end of 2017.
From January 2018 most HTTP traffic will be automatically redirected to HTTPS. [more...]
View this page in https

Translate tool

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence

Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored)

```
[illegible]
```

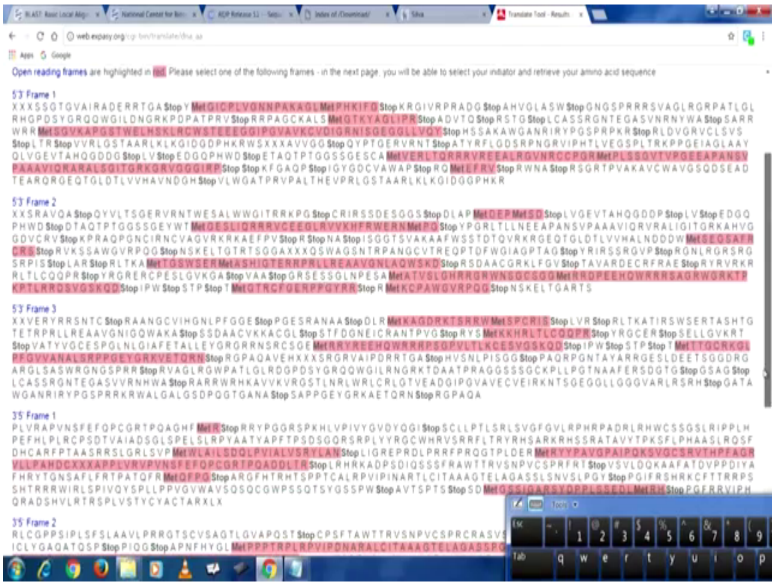
Output format: Verbose ("Met", "Stop", spaces between residues)

Genetic code: Standard

Reset or TRANSLATE SEQUENCE

We can go and use to such as (Refer Time: 31:05) that will allow you. So, this is my faster sequence. Now (Refer Time: 31:15) should ideally allow me to convert my faster sequences into protein sequence.

(Refer Slide Time: 31:21)



web.expasy.org/cgi-bin/translate/dna_aa

Open reading frames are highlighted in red. Please select one of the following frames - in the next page, you will be able to select your initiator and retrieve your amino acid sequence

```
5'3' Frame 1
XXXSSGTGVAIRADERRTGA Stop Y Met GICPLVGNNPAKAGL Met PHKIFG Stop KRGIVRPRADG Stop AHVGLASW Stop GNGSPRRRSVAGLRGRPATLGL
RHGPDSYGRQQWGILDNGRKPDPATPRV Stop RRPAGCKALS Met GTKYAGLIPR Stop ADVTQ Stop RSTG Stop LCASSRGNTEGASVNRNYWA Stop SARR
WRR Met SGVKAPGSTWELHSKLRCWSTEEEGGIPGVAVKCVDIGRNISGEGGLLVQY Stop HSSAKAWGANRIRYPGSPRPKR Stop RLDVGRVCLSVS
Stop LTR Stop VVRLGSTAARLKLKGIDGDPHKRWSXXXAVVGG Stop QYPTGERVRNT Stop ATYRFLGDSRPNGRVIPHTLVEGSPLTRKPPGEIAGLAAY
QLVGEVTAHQGDDG Stop LV Stop EDGQPHWD Stop ETAQTPTGGSSGESCA Met VERLTQRRRVREEALRGVNRCCPGR Met PLSSGVTVPGEEAPANSV
PAAAVIQRARALSGITGRKGRVGGGIRP Stop Stop KFGAQP Stop IGYGDCVAWAP Stop RQ Met EFRV Stop RWNA Stop RSGRTPVAKAVCWAVGSQDSEAD
TEARQRGEQTGLDTLVVHAVNDGH Stop VLWGATPRVPALTHEVPRLGSTAARLKLKGIDGGPHKR

5'3' Frame 2
XXSRAVQA Stop QYVLTSGERVRNTWESALWWGITRRKPG Stop CRIRSSDESGGS Stop DLAP Met DEP Met SD Stop LVGEVTAHQGDDP Stop LV Stop EDGQ
PHWD Stop DTAQTPTGGSSGEYWT Met GESLIQRRRVCEEGLRVVKHFRWERN Met PG Stop YPGRLTLLNEEAPANSVPAAAVIQRVRALIGITGRKAHVG
GDVCRV Stop KPRAQPGNCIRNCVAGVRKRKAEFPV Stop R Stop NA Stop ISGGTSVAKAAFWSSTDTQVRKRGEQTGLDTLVVHALNDDDW Met SEQSAFR
CRS Stop RVKSSAWGVRPQG Stop NSKELTGTRTSGGAXXXQSWAGSNTRPANGCVTREQPTDFWGIAGPTAG Stop YRIRSSRGVP Stop RGNLRGRSRG
SRPIS Stop LAR Stop RLTKA Met TGSWSER Met ASHIGTERRPRLLREAAVGNLAQWSKD Stop RSDAACGRKLFGV Stop TAVARDECRFRAE Stop RYRVRKR
RLTLCQQPR Stop YRGRERCPESLGVKGA Stop VAA Stop GRSESSGLNPESA Met ATVSLGHRRGRWNSGCSGG Met RRDPEEHQWRRRSAGRWGRKTP
KPTLRRDSVGSKQD Stop IPW Stop STP Stop T Met GTRCFGERPPGYRR Stop R Met KCPAWGVRPQG Stop NSKELTGARTS

5'3' Frame 3
XXVERYRRSNTC Stop RAANGCVIHGNLPFGGE Stop PGESRANAA Stop DLR Met KAGDRKTSRRW Met SPCRIS Stop LVR Stop RLTKATIRSWSERTASHTG
TETRPRLLREAAVGNIGQWAKA Stop SSDAACVKKACGL Stop STFDGNEICRANTPVG Stop RYS Met KKHRLTLCQQPR Stop YRGCER Stop SELLGVKRT
Stop VATYVGCESPGLNLGIAFETALLEYGRGRRNSRCSGE Met RRYREEHQWRRRPSGPVLTLKCESVGSKQD Stop IPW Stop STP Stop T Met TTGCRKGL
PFGVVANALSRPPGEYGRKVETQRN Stop RGPAQAVEHXXXSRGRVAIPDRRTGA Stop HVSNLPISGG Stop PAQRPGNTAYARRGESLDEETSGGDRG
ARGLSASWRGNGSPRR Stop RVAGLRGWPATLGLRDGPDSYGRQQWGILRNGRKTDAATPRAGGSSSGCKPLLPGTNAAFERSDGTG Stop GSAG Stop
LCASSRGNTEGASVVRNHWA Stop RARRWRHKAVVKVRGSTLNRLWRLCRLGTVEADGIPGVAVECVEIRKNTSGEGGLLGGGVARLRSRH Stop GATA
WGANRIRYPGSPRRKRWALGALGSDPQGTGANA Stop SAPPGEYGRKAETQRN Stop RGPAQA

3'5' Frame 1
PLVRAPVNSFEFQPCGRTPQAGHF Met R Stop RRYPGGRSPKHLVPIVYGVDYQGI Stop SCLLPTLSRLSVGFGVLRPHRPADRLRHWCSSGSLRIPPLH
PEFHLPLRCPSDTVAIADSGLSPELSLRPYAATYAPFTPSDSGQRSRPLYYRGCWHRVSRRFLTRYRHSARKRHSSRATAVYTPKSFLPHAASLRQSF
DHCARFPTAASRRSLGRLSVP Met WLAILSDQLPVIALVSRYLAN Stop LIGREPRDLPRRFPRQGTPLDER Met RYYPAVGPAIPQKSVGCSRVTHPFAGR
VLLPAHDCXXXAPPLVRVPVNSFEFQPCGRTPQADDLTR Stop LRHRKADPSDIQSSSFRAWTTRVSNPVCSPRFRT Stop VSVLDQKAAFATDVPPDIYA
FHRYTGNSAFLFRTPATQFR Met QFPG Stop ARGFHTRHTSPPTCALRPVIPINARTLCITAAAGTELAGASSLSNVSLPGY Stop PGIFRSHRKCFTTRRPS
SHTRRRWIRLSPIVQYSPLLPPVGVWAVSQSQCGWPSSQTSYGSSPW Stop AVTSPTS Stop SD Met GSSIGARSYDPPLSSEDL Met RH Stop PGFRRVIPH
QRADSHVLRTRSPLVSTYCYACTARXLX

3'5' Frame 2
RLCGPPSIPLSFSLAAVLPRRGTSCVSAGTLGVAPQST Stop CPSFTAWTTRVSNPVCSPRCRASVS
ICLYGAQATQSP Stop PIQG Stop APNFHYGL Met PPPTRPLRPVIPDNARALCITAAAGTELAGASSPG
```

Now, this is my protein sequence in 6 different frames, 5 prime to 3 prime frame 1 frame 2 frame 3 and then 3 prime to 5 5 5 frame one frame to frame 3 perfect and then you see it has noted all the stop codons I only see stop (Refer Time: 31:38) not start codons which is very interesting and ok. So, now, I can what I can do is I can blast all these proteins. So, I can copy all this and then I can put it in my here in protein blast, and then I can get my data I can get, but because we already have a nucleotide we can just use blast x it will automatically translate it into protein for us and we do not have to worry.

All right dear students this is all for today we learnt about NCBI blast which is a very very essential and very very important tool for anybody doing environmental microbiology even remotely. Thank you very much in the next class we will look at what tools are available to us if we have meta genomic data which is high throughput sequencing data.

Thank you.