# **Phyloinformatics**

Assignment-2

Q1-Using BLAST to identify similar sequences

Answer: **FOR EXAMPLE:**

UniProt:

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data. The UniProt Knowledgebase consists of two sections: a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, and a section with computationally analyzed records that await full manual annotation. For the sake of continuity and name recognition, the two sections are referred to as "UniProtKB/Swiss-Prot" (reviewed, manually annotated) and "UniProtKB/TrEMBL" (unreviewed, automatically annotated), respectively.

Step 1: Go to the homepage of UniProtKB and in the search bar type 'hsp90' and click on search button.
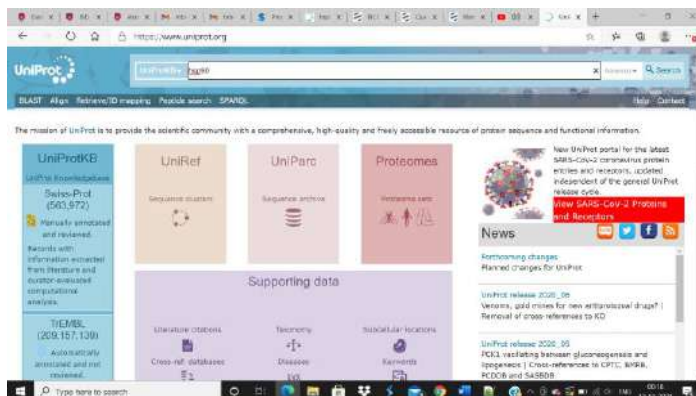

**Step 2:** You will see 1 to 25 of 133,802 number of hits.
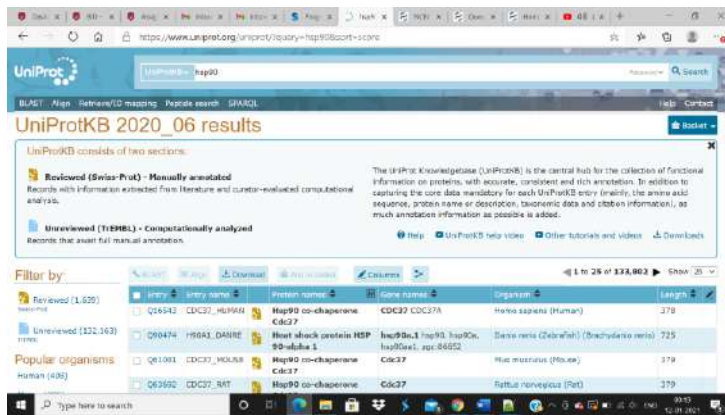
NCBI:

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper.The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for biomedical literature. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine. NCBI was directed by David Lipman,one of the original authors of the BLAST sequence alignment program and a widely respected figure in bioinformatics. He also led an intramural research program, including groups led by Stephen Altschul (another BLAST co-author), David Landsman, Eugene Koonin, John Wilbur, Teresa Przytycka, and Zhiyong Lu. David Lipman stood down from his post in May 2017.

Step 3: Go to homepage of NCBI in the search bar type 'hsp90' and search. Select the database as 'nucleotide' in the drop box.
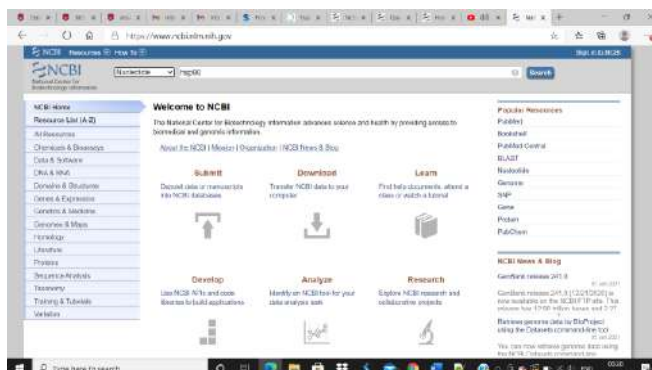
Step 4: In the extreme left corner select the species according to your choice. I have selected 'animal' and after clicking you will see 1 to 20 of 8075 number of hits.
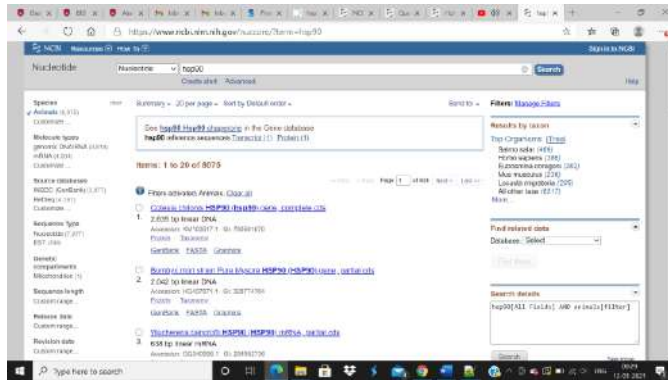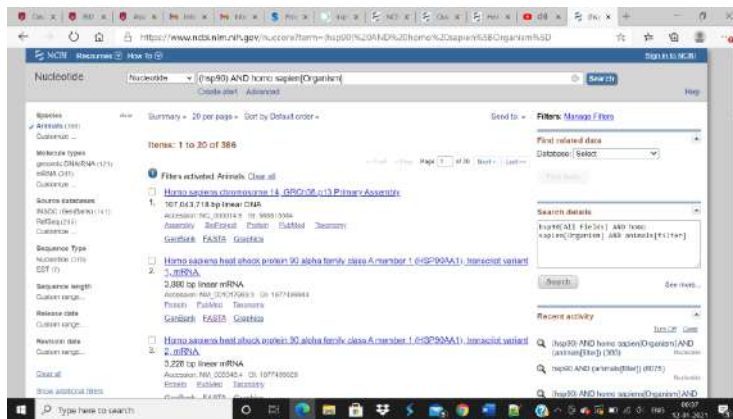
Step 5: Click on 'Advanced' below the search bar on the ncbi page.
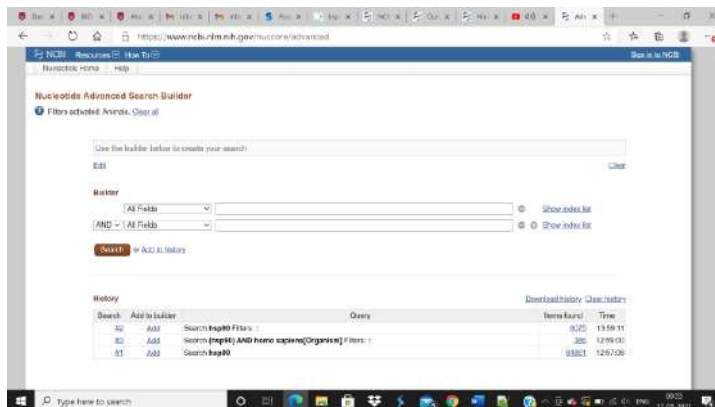

Step 6: Under the builder heading type 'hsp90' and in the second drop box select 'organism' as the category and 'homo sapien' as the name and click on search.
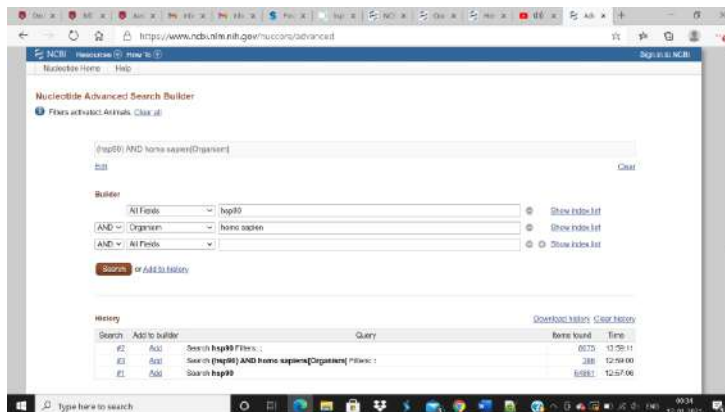
Step 7: You will get 1 to 20 of 386 as maximum number of hits.

Step 8: Select any one type , for example I am selecting the second one : Homo sapiens heat shock protein 90 alpha family class A member 1 (HSP90AA1) and click on its FASTA formate.
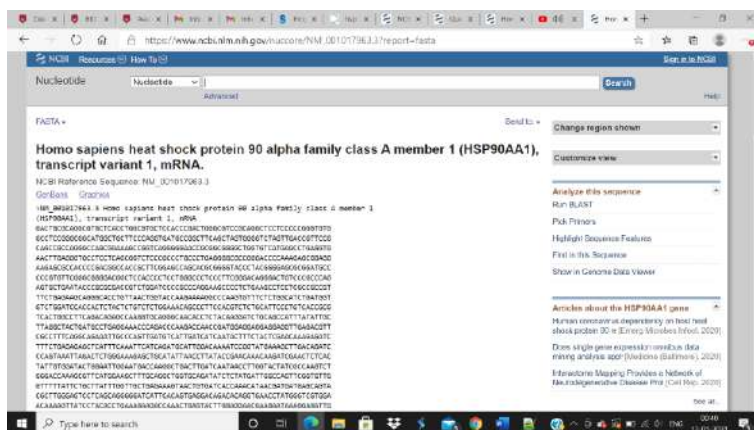
>Sample_1

CGAGGGACCTTTACGGCGTAATCCTGGAAACCATGACAAATCCAGAACCCCAAGGCTCCCCTCTT
CAGCTGATGTAGAATTTTGCCTGAGTTTGACCCATGGAAGGATTTGCTAGTCCACTTACTGGGAT
AGCGGATGCCTCTCAAAGAGCATGCACAATGCCTTGCACATCTATATGAATGGAACAATGTCCCA
GGCAGGGATCTGCCAACGATCCTATCTTCCTTCTTCACCATGCATTTGTTGACAGTATTTTTGAG
CAGTGGCTCCGAAGGCACCGTCCTCTTCAAGAAGTTTATCCAGAAGCCAATGCACCCATTGGACA
TAACCGGGAATCCTACATGGTTCTTATACCACTGTACAGAAATGGTGATTTCTTTATTTCATCCA
AAGATCTGGGCTATGACTATAGCTATCTACAAGATTCAGACCCAGACTCTTTTCAAGACTACATT
AAGTCCTATTTGGAACAAGCGAGTCGGATCTGGTCATGGCTCCTTGGGGCGGCGATGGTAGGGGC
CGTCCTCACTGCCCTGCTGGCGGGCTTGTGAGCTTGCTGTGTCGTCACAAGAGAAAGCAGCTTCC
TGAAGAAAAGCAGCCACTCCTCATGGAGAAAGAGGATTACCACAGCTTGTATCAGAGCCATTTAT
AAAAGGCTTAGGCAATAGAGTAGGGCCAAAAAGCCTGACCTCACTCTAACTCAAAGTAATGTCCA
GGTTCCCAGAGAATATCTGCTGGTATTTTTCTGTAAAGACCATTTGCAAAATTGTAACCTAATAC
AAAGTGTAGCCTTCTTCCAACTCAGGTAGAACACACCTGTCTTTGTCTTGCTGTTTTCACTCAGC
CCTTTTAACATTTTCCCCTAAGCCCATATGTCTAAGGAAAGGATGCTATTTGGTAATGAGGAACT
GTTATTTGTATGTGAATTAAAGTGCTCTTATTTTAAAAAATTGAAATAATTTTGATTTTTGCCTT
CTGATTATTTAAAGATCTATATATGTTTTATTGGCCCCTTCTTTATTTTAATAAAACAGTGAGAA
ATCT

- **Select the BLAST program that is required protein / DNA**

- **Copy and paste your DNA** into the search box.

- Select the required **Database**

- Select the Megablast for DNA

- For the first time, **leave the Options section settings** on the **Default** values and keep **homo sapiens** as organism

- **Click** on the **Blast button** at the bottom or top of the screen.

- **A new window will appear** gives an estimate of how long the search will take and which lists conserved domains in your query sequence.

- **Check your result** - check **graphic** - red, green, blue, black lines and their meaning

- **Identify the top 10 similar sequence** and analyse **Accession No**. **E value**, **coverage**, **identity** with other sequences (use a table)

- What is the source of the given query sequence?

- Show the Alignment, Graphic summary and taxonomy of top ten similar sequences

- List out the set of sequences showed good similarity with query sequence

BLAST:

In bioinformatics, BLAST is an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. A BLAST search enables a researcher to compare a subject protein or nucleotide sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. **PHI-BLAST** performs the search but limits alignments to those that match a pattern in the query. DELTA-**BLAST** constructs a PSSM using the results of a Conserved Domain Database search and searches a sequence database. PHI-BLAST stands for Pattern-hit initiated BLAST. The program uses an input sequence and a defined pattern to query a protein database. **The BLAST Homepage** : The BLAST homepage consists of several sections, each provides a specific set of functions:

1. The common header , present in most BLAST-related pages, provides easy access to other content or functions not directly accessible from the homepage.

2. The "Recent Results" lists a set of links to recently completed search results.

3. Pages with web forms for submitting searches are listed as links in the body of the BLAST homepage. These links are organized into three categories, "Web BLAST" , used to be called "Basic BLAST", "BLAST Genomes" , and "Specialized searches" .

4. The search box in the "BLAST Genomes" section takes the name of an organism as input and suggests a list of candidates. Selecting from the suggested list and clicking the "Search" button locates the best genomic sequence dataset for BLAST alignment purposes.

Step 1: Copy the give 'Sample1' nucleotide sequence. Open the blast homepage and click on Nucleotide blast..
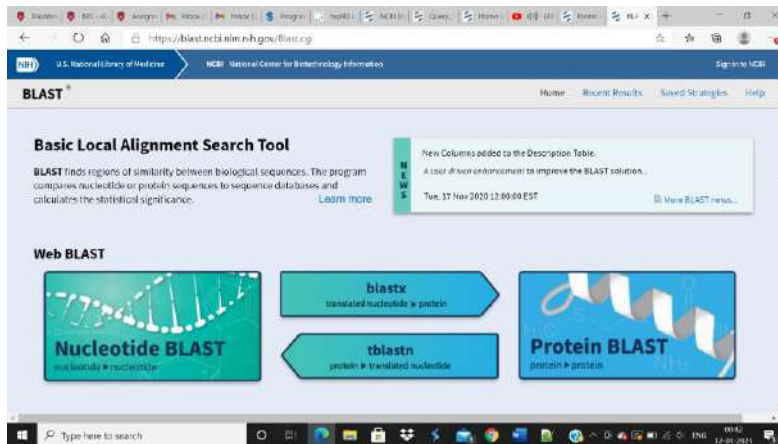
Nucleotide blast:

- ➤ Enter query sequence(s) in the text area. It automatically determines the format of the input. To allow this feature, certain conventions are required with regard to the input of identifiers.
- ➤ Enter coordinates for a subrange of the query sequence. The BLAST search will apply only to the residues in the range. Sequence coordinates are from 1 to the sequence length.The range includes the residue at the To coordinate.
- ➤ Use the browse button to upload a file from your local disk. The file may contain a single sequence or a list of sequences. The data may be either a list of database accession numbers, NCBI gi numbers, or sequences in FASTA format.
- ➤ Title:Nucleotide collection (nt)
- ➤ Description:The nucleotide collection consists of GenBank+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA, patent sequences as well as phase 0, 1, and 2 HTGS sequences and sequences longer than 100Mb. The database is non-redundant. Identical sequences have been merged into one entry, while preserving the accession, GI, title and taxonomy information for each entry.
  - o Molecule Type:mixed DNA
  - o Update date:2021/01/10
  - o Number of sequences:65815883
- ➤ Start typing in the text box, then select your taxid. Use the 'plus button to add another organism or group, and the 'exclude' checkbox to narrow the subset. The search will be restricted to the sequences in the database that correspond to your subset.
- ➤ ou can use Entrez query syntax to search a subset of the selected BLAST database. This can be helpful to limit searches to molecule types, sequence lengths or to exclude organisms.
- ➤ Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast. Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is

intended for cross-species comparisons. BlastN is slow, but allows a word-size down to seven bases.
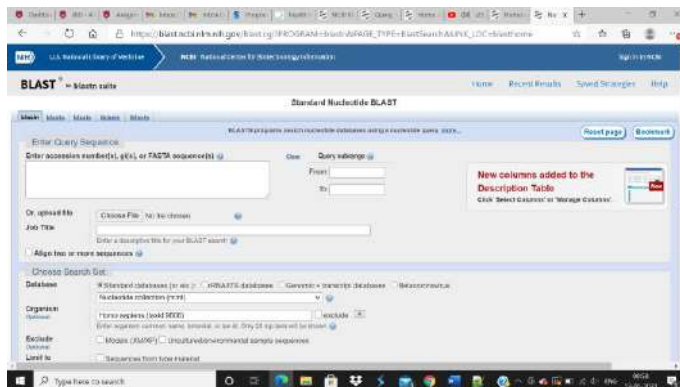
Step 2: Paste the nucleotide sequence and click on blast. You will see 100 sequences selected.

Step 3: By default the 100 sequences are selected already. We have to select on 10 sequences as shown below:

Step 4: To get the result of graphics , we have to click on the 'Graphics' besides GenBank you will get - red, green, blue, black lines and their meaning as shown below:
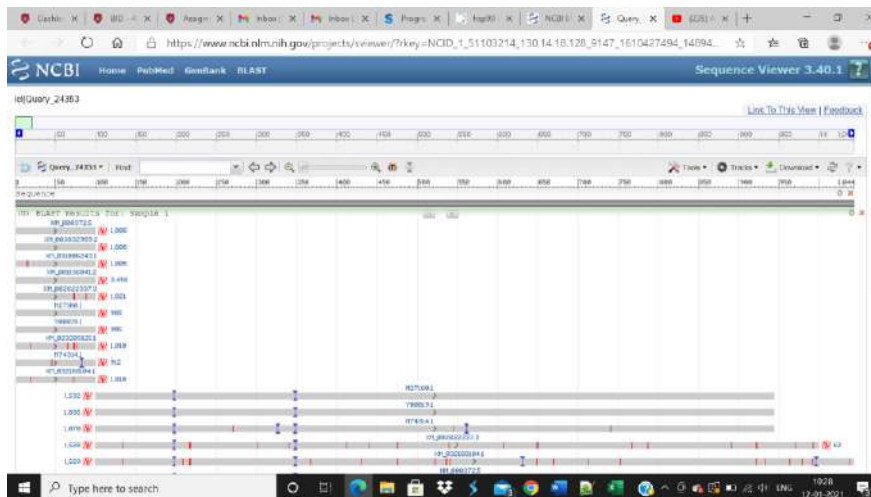
(U) BLAST Results for: Sample 1 (Non-NCBI data)

Alignment:	**Query_24353 x NM_000372.5**
Anchor:	**Query_24353 (1..99)**
Query:	**NM_000372.5 (958..1,056) Homo sapiens tyrosinase (TYR), mRNA**
Relative orientation:	**forward**
Span on Query_24353:	**99**
Segments:	**1**
Coverage:	**4.8%**
Identity:	**100.0%**
Mismatches:	**0**
Gaps:	**0**

Download FASTA: Query_24353

Links & Tools
BLAST nr:	**NM_000372.5**
BLAST to Genome:	**NM_000372.5**
FASTA record:	**NM_000372.5**
GenBank record:	**NM_000372.5**

Alignment:	**Query_24353 x M27160.1**
Anchor:	**Query_24353 (100..944)**
Query:	**M27160.1 (1,533..2,384)**
Human tyrosinase (TYR) mRNA, complete cds	
Relative orientation:	**forward**
Span on Query_24353:	**852**
Segments:	**1**
Coverage:	**35.4%**
Identity:	**100.0%**
Mismatches:	**0**
Gaps:	**4**

Unaligned regions: 5
Query_24353 position: 196
M27160.1 position: 1,630

Download FASTA: Query_24353

Links & Tools
BLAST nr: M27160.1
BLAST to Genome: M27160.1
FASTA record: M27160.1
GenBank record: M27160.1

Step 5: What is the source of the given query sequence?
Select seq NM_000372.5   Homo sapiens tyrosinase (TYR), mRNA

Step 6: To identify the top 10 similar sequence and analyse Accession No. E value, coverage, identity with other sequences (use a table) we need to click on 'Download' and then click on 'Descriptions Table (CSV)'.

Step 7: To identify the top 10 similar sequence and analyse Accession No. E value, coverage, identity with other sequences (use a table)

Accession No:

Accession
NM_000372.5
XM_003832989.2
XM_031006243.1
XM_001136041.2
XM_032168104.1
XM_002822337.3
M27160.1
Y00819.1
XM_023209125.1
M74314.1

E value:

The Expect value (E) is a parameter that describes the number of hits one can 'expect' to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background

noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance. The lower the E-value, or the closer it is to zero, the more 'significant' the match is." E value and the matrix score/identity score go inversely., exponential graph.

E value
0
0
0
0
0
0
0
0
0
0

Coverage:

Query Coverage
100%
100%
100%
100%
99%
95%
90%
90%
100%
90%

Identity:

Per. Identity
99.26
98.84
98.74
98.53
96.66
97.35
99.18
99.18
94.99
97.91

Step 8: To show the Alignment, Graphic summary and taxonomy of top ten similar sequences we need to click on 'Alignment'

'Graphic summary'

Taxonomy:

Step 9: List out the set of sequences showed good similarity with query sequence