Shanzida Jahan Siddique

SRavichandran

BIFX-550

23 April 2018

Find -a -gene project (1-4)

[1] Tell me the name of a protein you are interested in. Include the species and the accession number. If you do not have a favorite protein, select a protein that is associated with a disease.

Name of the protein: Alpha-synuclein isoform X1

Species: Homo sapiens

Accession number: XP\_011530509.1

[2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC). It is not necessary to print out all of the blast results if there are many pages.

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score.

In general, step [2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of step [4]), and a non-homologous result.

**First step:**

Program: TBLASTN

Database: est

Search against organism: Include: plants, insects, arrow worms, Nicotiana tabacus; Exclude: All homologous organism such as human,X.tropicalis,G.gallus,R.norvegicus,M.musculus,B.taurus,C.lupus,P.troglodytes.

Algorithm parameter:

Max target sequence:100

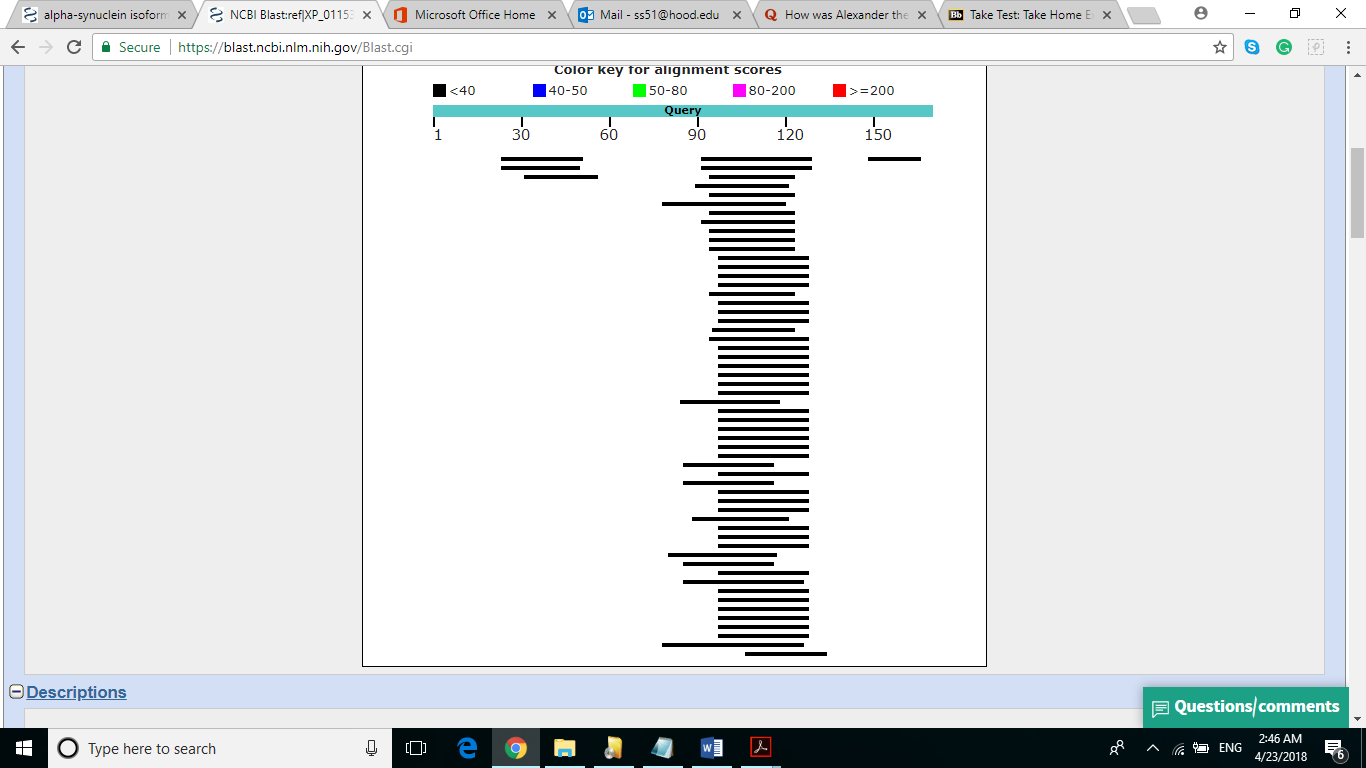
Expect threshold:1000

Word sixe:2

Matrix: BLOSUM62

Filter: Low complexity region

Blast output:





CBYY6164.b1 CBYY Panicum virgatum Kanlow crown Panicum virgatum cDNA clone CBYY6164 5', mRNA sequence. GenBank: FE625547.1

Range 1: 693 to 773[GenBank](https://www.ncbi.nlm.nih.gov/nucleotide/FE625547?report=genbank&log$=nuclalign&blast_rank=20&RID=DUZYEG0Y015&from=693&to=773)[Graphics](https://www.ncbi.nlm.nih.gov/nuccore/FE625547?report=graph&rid=DUZYEG0Y015%5bFE625547%5d&tracks=%5bkey:sequence_track,name:Sequence,display_name:Sequence,id:STD1,category:Sequence,annots:Sequence,ShowLabel:true%5d%5bkey:gene_model_track,CDSProductFeats:false%5d%5bkey:alignment_track,name:other%20alignments,annots:NG%20Alignments|Refseq%20Alignments|Gnomon%20Alignments|Unnamed,shown:false%5d&v=689:777&appname=ncbiblast&link_loc=fromHSP)Next MatchPrevious Match

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Alignment statistics for match #1 | | | | | | |
| **Score** | **Expect** | **Method** | **Identities** | **Positives** | **Gaps** | **Frame** |
| 30.0 bits(66) | 280 | Compositional matrix adjust. | 13/28(46%) | 18/28(64%) | 1/28(3%) | +3 |

Query 96 KKDQLGKKHPKYKPSKRQENVVMFLVQV 123

KD L HP +PS+ +EN ++ LVQV

Sbjct 693 PKDDL-LPHPLPRPSREEENKLILLVQV 773

**Second step:**

Collection of mRNA sequence:

>FE625547.1 CBYY6164.b1 CBYY Panicum virgatum Kanlow crown Panicum virgatum cDNA clone CBYY6164 5', mRNA sequence

GTGCGCCACCTTGCCCCTCCCCTCCCTCCCGGAGAAGCGCCGCGCCGCCTGCCTGCTTGCCCATCCACTC

CCCGACGAACTCACCCATCTATAAAACCCTCCACCTCCGCCGCTGCACATAACTCCAAGCAACCACCACC

ACCACTCCCAATTTCACTTCACCTCGACAGCGCACTCCCCACCCACCGGTCTCCCTCTCCCTCCCCTCCA

CCACCACCGAATCCCAGGCAGGGGCAGGCAGAGAGCAGAGCGGCGGCAGGGGGACCTCTCCCTCCTCTCC

TCTCCTTCCCTTCCTCACCCCCTTCGCCCGGCGAGCCAGCCAGCCATGGAGCCGCCGGCGAAGACGGTGG

AGCGGCTGGCGCAGCGCCTGGTGCCGCCGGCGGAGCCCACGCCCACCGGCCCGCACCGCCTGTCCTGGCT

CGACCGCTACCCGACCCAGATGGCGCTCATCGAGTCGCTGCACGTCTTCAAGCCCGACCCGGCGCGGGAC

GGGGTCAGCCCCGCGGAGACCATCCAGCGCGCGCTGGCGCGGGCGCTCGTCGACTACTACCCGCTCGCGG

GGCGCCTCGCCGTGTCCGACGGCGCCGGCGGGCTCCACGTCGACTGCAACGGCGAGGGCGTCTGGTTCGT

CGAGGCCGCCGTGCGGTGCCGGCTCGAGGACGTCGAGTACCTCGAGTACCCGCTGCAGATCCCCAAAGGA

CGACCTGCTCCCGCACCCGCTGCCGCGCCCCAGCCGCGAGGAGGAGAACAAGCTCATCCTGCTCGTCCAG

GTGACCACGTTC

Program: blastx

Database: Non-redundant protein sequence

Algorithm parameter:

Max target sequence:100

Expect threshold:100

Word size:6

Matrix: BLOSUM62

Filter: Low complexity region

Blast output: According to blastx result there are no protein which is 100% identical.



[3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from step [2]. In some cases, you will be able to do further BLAST searches to obtain even more sequence of your novel gene.

 Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or primates or protozoa.

Name of the novel protein: Unnamed protein

Family: Transferase superfamily

Sequence of the novel protein:Met S L F S S S R L G R G S G C G S R S S F G D L Q R V L E V L D V L E P A P H G G L D E P D A L A V A V D V E P A G A V G H G E A P R E R V V V D E R P R Q R A L D G L R G A D P V P R R V G L E D V Q R L D E R H L G R V A V E P G Q A V R A G G R G L R R R H Q A L R Q P L H R L R R R L H G W L A R R A K G V R K G R R G E E G E V P L P P L C S L P A P A W D S V V V E G R E R E T G G W G V R C R G E V K L G V V V V V A W S Y V Q R R R W R V L

Species from which it derives: Panicum Virgatum (Switchgrass)

[4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (step [3]), and use it as a query in a blastp search of the nr database at NCBI.

--If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as “unknown”). Someone has already found and annotated this sequence, and assigned it an accession number.

--If there is a match with less than 100% identity, then it is likely that your protein is novel, and you have succeeded.

--If there is a match with 100% identity, but to a different species than the one you started with, then you have succeeded in finding a novel gene.

--If there are no database matches to the original query from step [1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

Third step:

Novel gene’s mRNA sequence has translated by ExPASy and taken the longest open reading frame sequence.

Program: blastp

Database: Non-redundant protein sequence

Algorithm parameter:

Max target sequence:100

Expect threshold:200000

Word size:2

Matrix: PAM 30

Compositional adjustment: no adjustment

Blast output: According to blastp result there are no protein which is 100% identical.

