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

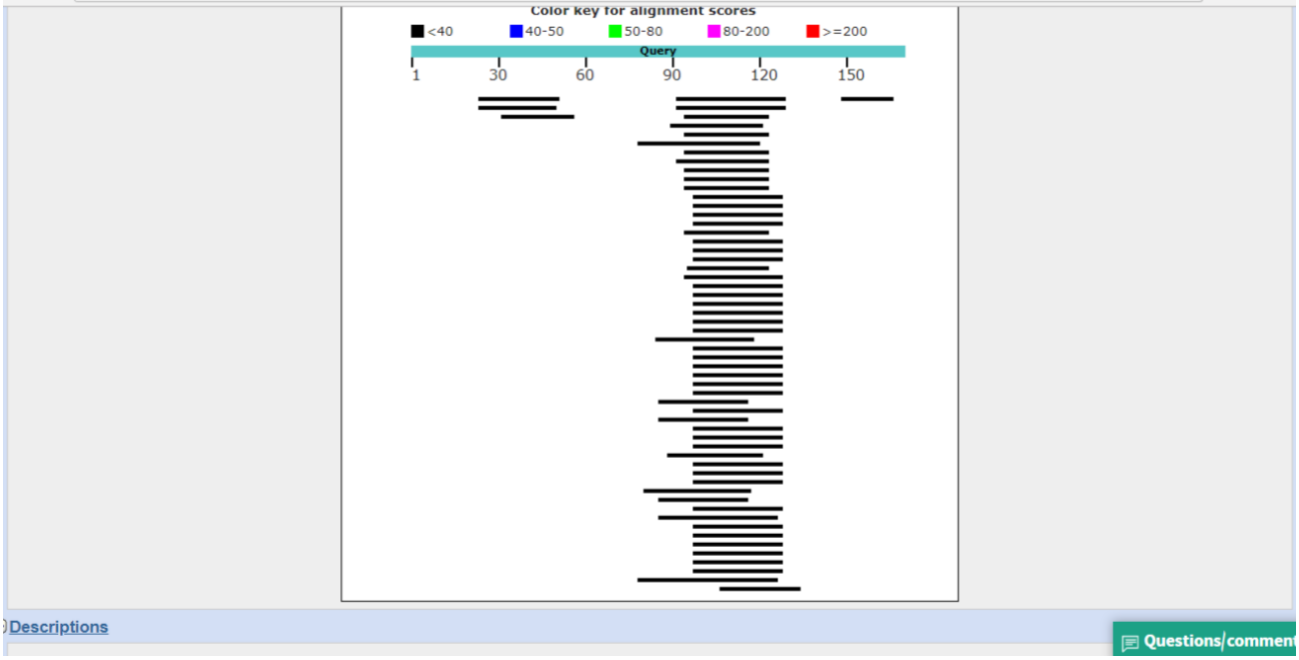
Expect threshold:1000

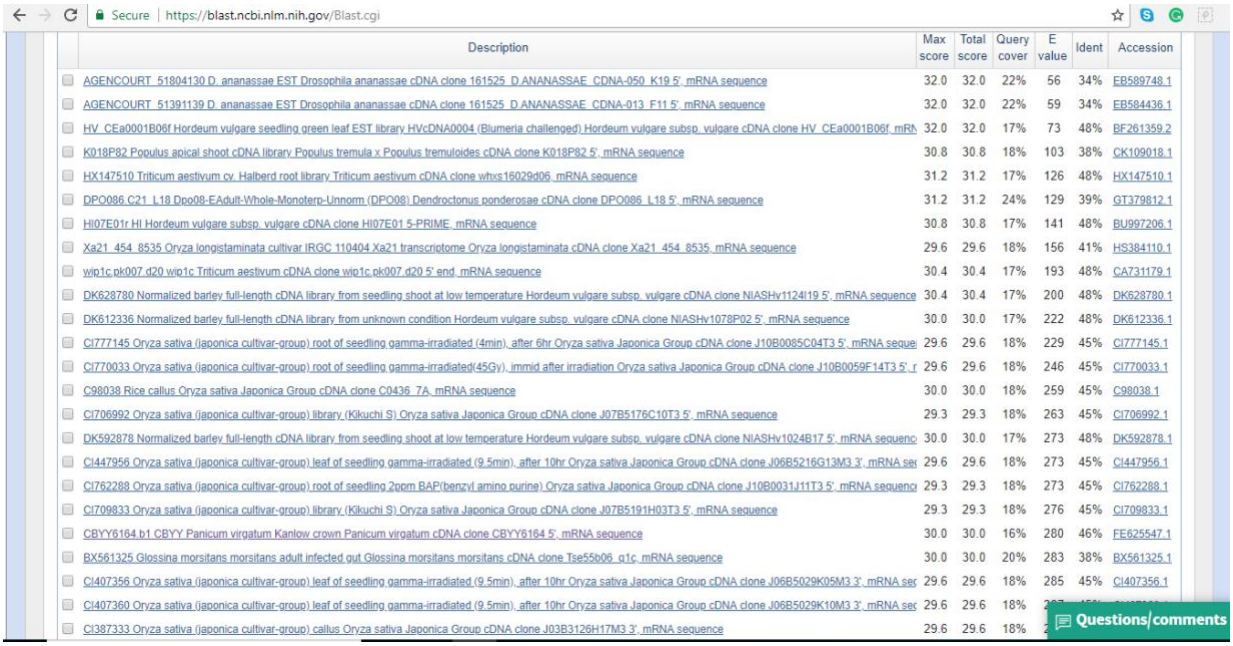
Word sixe:2

Matrix: BLOSUM62

Filter: Low complexity region

Blast output:





Description	Max score	Total score	Query cover	E value	Ident	Accession
AGENCOURT_51804130 D. ananassae EST Drosophila ananassae cDNA clone 161525_D.ANANASSAE_CDNA-050_K19 5', mRNA sequence	32.0	32.0	22%	56	34%	EB589748.1
AGENCOURT_51391139 D. ananassae EST Drosophila ananassae cDNA clone 161525_D.ANANASSAE_CDNA-013_F11 5', mRNA sequence	32.0	32.0	22%	59	34%	EB584436.1
HV_CEa0001B06f Hordeum vulgare seedling green leaf EST library HVcDNA0004 (Blumeria challenged) Hordeum vulgare subsp. vulgare cDNA clone HV_CEa0001B06f, mRN	32.0	32.0	17%	73	48%	BF261359.2
K018P82 Populus apical shoot cDNA library Populus tremula x Populus tremuloides cDNA clone K018P82 5', mRNA sequence	30.8	30.8	18%	103	38%	CK109018.1
HX147510 Triticum aestivum cv. Halberd root library Triticum aestivum cDNA clone whxs16029d06, mRNA sequence	31.2	31.2	17%	126	48%	HX147510.1
DPO086_C21_L18 Dpo08-EAdult-Whole-Monoterp-Unnorm (DPO08) Dendroctonus ponderosae cDNA clone DPO086_L18 5', mRNA sequence	31.2	31.2	24%	129	39%	GT379812.1
HI07E01r HI Hordeum vulgare subsp. vulgare cDNA clone HI07E01 5-PRIME, mRNA sequence	30.8	30.8	17%	141	48%	BU997206.1
Xa21_454_8535 Oryza longistaminata cultivar IRGC 110404 Xa21 transcriptome Oryza longistaminata cDNA clone Xa21_454_8535, mRNA sequence	29.6	29.6	18%	156	41%	HS384110.1
wip1c.pk007.d20 wip1c Triticum aestivum cDNA clone wip1c.pk007.d20 5' end, mRNA sequence	30.4	30.4	17%	193	48%	CA731179.1
DK628780 Normalized barley full-length cDNA library from seedling shoot at low temperature Hordeum vulgare subsp. vulgare cDNA clone NIASHv1124I19 5', mRNA sequence	30.4	30.4	17%	200	48%	DK628780.1
DK612336 Normalized barley full-length cDNA library from unknown condition Hordeum vulgare subsp. vulgare cDNA clone NIASHv1078P02 5', mRNA sequence	30.0	30.0	17%	222	48%	DK612336.1
CI777145 Oryza sativa (japonica cultivar-group) root of seedling gamma-irradiated (4min), after 6hr Oryza sativa Japonica Group cDNA clone J10B0085C04T3 5', mRNA seque	29.6	29.6	18%	229	45%	CI777145.1
CI770033 Oryza sativa (japonica cultivar-group) root of seedling gamma-irradiated(45Gy), immid after irradiation Oryza sativa Japonica Group cDNA clone J10B0059F14T3 5', r	29.6	29.6	18%	246	45%	CI770033.1
C98038 Rice callus Oryza sativa Japonica Group cDNA clone C0436_7A, mRNA sequence	30.0	30.0	18%	259	45%	C98038.1
CI706992 Oryza sativa (japonica cultivar-group) library (Kikuchi S) Oryza sativa Japonica Group cDNA clone J07B5176C10T3 5', mRNA sequence	29.3	29.3	18%	263	45%	CI706992.1
DK592878 Normalized barley full-length cDNA library from seedling shoot at low temperature Hordeum vulgare subsp. vulgare cDNA clone NIASHv1024B17 5', mRNA sequenc	30.0	30.0	17%	273	48%	DK592878.1
CI447956 Oryza sativa (japonica cultivar-group) leaf of seedling gamma-irradiated (9.5min), after 10hr Oryza sativa Japonica Group cDNA clone J06B5216G13M3 3', mRNA se	29.6	29.6	18%	273	45%	CI447956.1
CI762288 Oryza sativa (japonica cultivar-group) root of seedling 2ppm BAP(benzyl amino purine) Oryza sativa Japonica Group cDNA clone J10B0031J11T3 5', mRNA sequenc	29.3	29.3	18%	273	45%	CI762288.1
CI709833 Oryza sativa (japonica cultivar-group) library (Kikuchi S) Oryza sativa Japonica Group cDNA clone J07B5191H03T3 5', mRNA sequence	29.3	29.3	18%	276	45%	CI709833.1
CBYY6164.b1 CBYY Panicum virgatum Kanlow crown Panicum virgatum cDNA clone CBYY6164 5', mRNA sequence	30.0	30.0	16%	280	46%	FE625547.1
BX561325 Glossina morsitans morsitans adult infected gut Glossina morsitans morsitans cDNA clone Tse55b06_g1c, mRNA sequence	30.0	30.0	20%	283	38%	BX561325.1
CI407356 Oryza sativa (japonica cultivar-group) leaf of seedling gamma-irradiated (9.5min), after 10hr Oryza sativa Japonica Group cDNA clone J06B5029K05M3 3', mRNA se	29.6	29.6	18%	285	45%	CI407356.1
CI407360 Oryza sativa (japonica cultivar-group) leaf of seedling gamma-irradiated (9.5min), after 10hr Oryza sativa Japonica Group cDNA clone J06B5029K10M3 3', mRNA se	29.6	29.6	18%			
CI387333 Oryza sativa (japonica cultivar-group) callus Oryza sativa Japonica Group cDNA clone J03B3126H17M3 3', mRNA sequence	29.6	29.6	18%			

Questions/comments

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

Range 1: 693 to 773GenBankGraphicsNext MatchPrevious Match

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
30.0 bits(66)	280	Compositional matrix adjust.	13/28(46%)	18/28(64%)	1/28

```
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

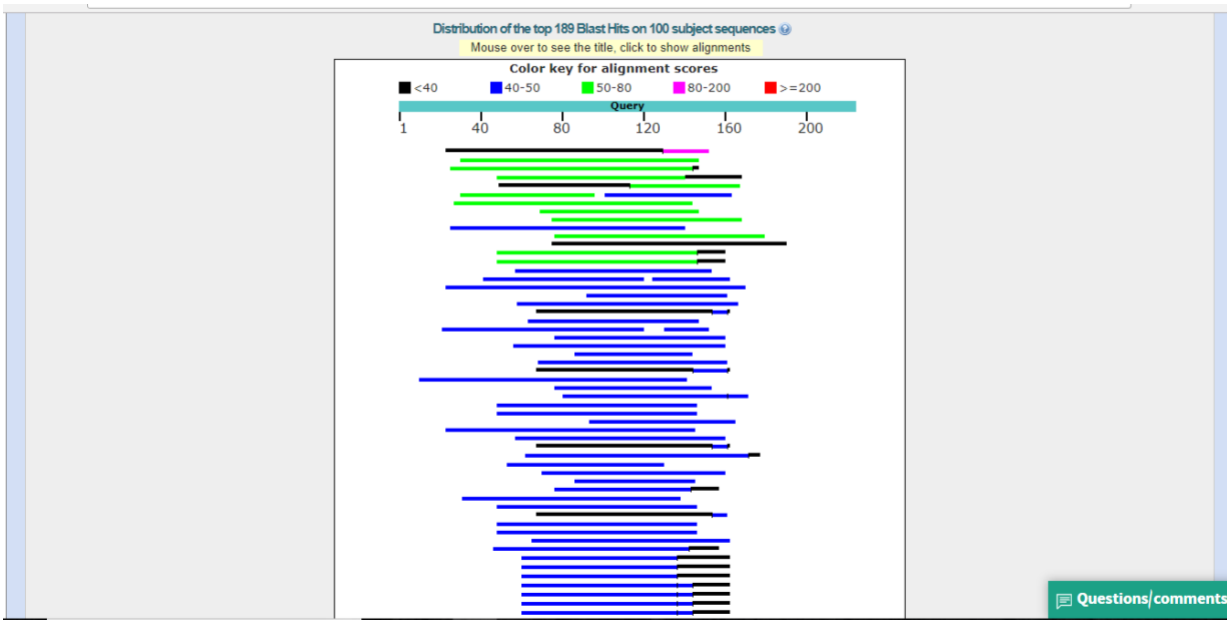
Expect threshold:200000

Word size:2

Matrix: PAM 30

Compositional adjustment: no adjustment

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



Description	score	score	cover	value	Ident	Accession
hypothetical protein BRADI_2g33977v3 [Brachypodium distachyon]	160	194	57%	1e-36	59%	PNT71699.1
predicted protein [Hordeum vulgare subsp. vulgare]	68.5	68.5	52%	1e-08	43%	BAJ96099.1
Os10g0503250 [Oryza sativa Japonica Group]	68.1	95.4	54%	2e-08	40%	BAT11547.1
hypothetical protein PAHAL_F02379 [Panicum hallii]	67.7	106	53%	3e-08	44%	PAN25424.1
hypothetical protein SD81_43115 [Tolypothrix campylonemoides VB511288]	56.2	82.3	52%	1e-04	38%	KIJ71079.1
hypothetical protein ACMD2_15061 [Ananas comosus]	54.5	54.5	29%	3e-04	49%	OAY83759.1
Os07g0209201 [Oryza sativa Japonica Group]	53.7	53.7	52%	6e-04	36%	BAT00581.1
Os11g0530650 [Oryza sativa Japonica Group]	52.0	52.0	34%	0.002	43%	BAT14280.1
amino acid ABC transporter permease [Actinokineospora bangkokensis]	52.0	52.0	41%	0.002	39%	WP_084794244.1
unknown [Zea mays]	51.5	100	51%	0.003	44%	ACN25787.1
expressed protein [Aureococcus anophagefferens]	51.1	82.7	45%	0.004	35%	XP_009036307.1
hypothetical protein [Jatrophihabitans endophyticus]	50.3	89.9	51%	0.007	41%	WP_084180928.1
hypothetical protein DM75_4932 [Burkholderia mallei]	50.3	175	50%	0.007	37%	KGC80022.1
hypothetical protein DM77_4801 [Burkholderia mallei]	50.3	174	50%	0.007	37%	KOT16474.1
hypothetical protein HMPREF0731_4776 [Roseomonas cervicalis ATCC 49957]	49.8	49.8	42%	0.008	39%	EFH09005.1
hypothetical protein ACMD2_15060 [Ananas comosus]	49.8	49.8	35%	0.008	43%	OAY83754.1
HAD-IIB family hydrolase [Microtetraspora niveoalba]	49.8	49.8	27%	0.009	47%	WP_084517790.1
Os05g0278550 [Oryza sativa Japonica Group]	49.4	49.4	65%	0.010	35%	BAS93128.1
hypothetical protein PBRA_000732 [Plasmodiophora brassicae]	49.4	49.4	30%	0.012	42%	CEO97387.1
hypothetical protein Cus16_3018 [Curtobacterium sp. ER1/6]	49.4	49.4	48%	0.012	39%	OEI67418.1
hypothetical protein DP57_4663 [Burkholderia pseudomallei]	49.4	143	42%	0.013	35%	KGC69272.1
signal peptidase I [Quadrisphaera sp. DSM 44207]	49.0	49.0	37%	0.016	40%	WP_092866369.1
LOW QUALITY PROTEIN: carrier protein membrane protein [Streptomyces viridochromogenes DSM 40736]	49.0	49.0	16%	0.016	54%	EFL31896.1
ABC transporter ATP-binding protein [Nocardia caishijiensis]	49.0	49.0	44%	0.016		