BGGN-213: FOUNDATIONS OF BIOINFORMATICS

The find-a-gene project assignment
http://thegrantlab.org/bggn213
Dr. Barry Grant

Juan Gonzalez; jhg001@ucsd.edu; PID: A69036681

Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name: Longifolia 1 (LNG1)
Accession: NP_197062.1
Species: *Arabidopsis thaliana*
Function: Regulates leaf morphology by promoting cell expansion in the leaf-length direction

[Q2] 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 details of the BLAST method used, database searched and any limits applied (e.g. Organism).

LNG1 *Arabidopsis thaliana* DNA sequence:

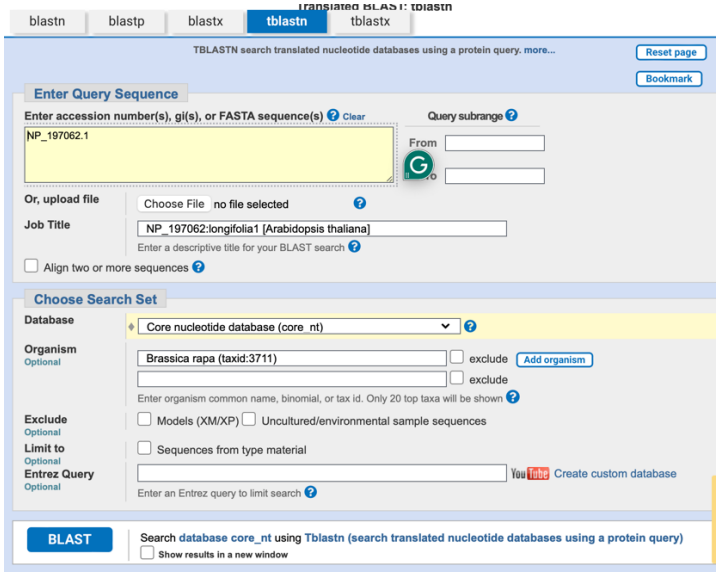
```
ATGTCGGCGAAGCTTTTGTATAACTTGTCGGATGAGAATCCAAATCTGAATAAACAGATTGGATGTATGAATGGGATCTTTCAGGTGTTTTACCGGCAACATTATCCACCGAGACGTGTCACC
GGAGATGAGCTCAAGTCTCTTCCTTCAGGCAAAGCAAGTGACAATGTCGGTGATACCAACATTTCAGCGGACAAGAAGGAAACGGAGAAGAGTAAGAAGAAGAAGACTGCAAAGGAGA
AACAGAGGGGAGTATCCTCTGAATCGTCCTCGAGGTTGTCGTTTTCTTCATCACCATGCTCCTCGAGCTTCTCATCTGCAGATATTAGCACCACGGCTTCTCAGTTTGAACAGCCCGGTTTG
AGTAATGGTGAGAATCCTGTAAGAGAACCGACCAATGGGTCGCCAAGGTGGGGCGGTTTAATGATGCCAAGTGATATAAGGGAGCTTGTGAGAAGCTCTATTCATAAGGAGACAAGAACC
AGAGATGAAGAAGCCTTGTCTCAGCAGCCTAAATCAGCCAGAGCTAATGTGTCTCTTCTCAAAGAATCATCACCATCTCGGAATTCTAATGAATGGAGTGAGGGACGGAGAGTAGTGAAG
CTGAAAGACAGTCCTCGGTTCTCTTACGATGAGAGGGAGACGAGAAAGACAGGGGCCAAGTTGAAAGAGACCCCGAGGTTGTCATTAGACAGTAGATCGAATTCCTTTAGGAGCGCAAG
ATCTAGTTGTTCACCAGAGCCACAAGAGCTTGTAACAGGTCACAGAAGAACAACATCAAGCGTCGTTGCAAAACTGATGGGTCTTGAAGTAATTCCAGATGAGCCTGTGACCATTCAGAAT
AGAGAAAATCGCTTCTGCGACTCCCCGAGGCCAACTTCCCGAGTGGAAGTAGATCTACAAAGATCAAGAGGTTTTGATTCAATCAAAAAGATGATGCCTGCTAAGTTTCCTATGAAAGCAT
CTCCATGGGCGCAAGTTGATGGTGCCAAGAACCAAGTCAAAATACCAGACGCTACTACGCTGACAGTTTATGGTGAGATACAGAAGCGGCTTTCACAGCTTGAGTTCAAAAAGTCCGAGA
AAGACCTCAGGGCTCTTAAGCAAATACTCGAAGCAATGGAGAAGACGCAGCAGTTGATAAGCAAAGATGATGATGACAACAAAACTCTATGTTCAAGCAATTTTATGCAGAGAAATAATCA
GCCAATTCCATCTGCAATAAACACCTCGTCCATGAATTTTAAATCATCCTCTATCGTGGTTATGAAAGCAGCTACCGCTCCAGTCTTCAAAGACACAGGCATTGCAGGTTCTGCGTCTTTCTC
CCCGCGGAATGTTGCTTTACCAAATGTCAAGGTTGGAAACCTGAGGCAAGCCCAGAAAGTCATTCCGAGGAAGCAGAGTGCTATGGATGTGACCCCGAGGCCAGGATATTACAAGGGCC
AGACAGAATCCACAATGAAAAACACTAGTACCAGACCATTACAATCGAAAAGCGACATGGCCAAGTCAGGGAAGATCCAGAAGCCTAGTGTCAGCCTAAGAACACCACCAAAGAAGCTT
GGGTTCGAGAAGCAGTCTCGGCCAACATCCCCAAAACCAGAACTGAACAAGAACCAGAGACAACAACTCAGTAGGCAACAGACAGAGTCAGCCTCCCCGAGAAGAAAGCCAGGGATA
AAATCTCGTGGCCTACAGCAATCTGAAGACCGTTTAAGTGATGAAAGCAGTGACTTGAGAAGTCTAAGATCTGACAGCAACGTAAGCTTGGCCTCTAACCTTGATACTGAGGTTACAAGCA
GGTATAACTATGAGAGGAACAGCGACATTACGGAGCAGCACACCCCGAAACAAAGGAGTCCAGACTTGGGAATGAGGTCGCTGTCAAAACCTCTGAAAGTTACAGTGGAGCAGCCCAG
CCCGGTGTCAGTTCTTGATGTAGCTTTCGACGAAGATGATTCACCATCCCCTGTGCGGAAGATATCCATTGTCTTTAAAGAGGACGACAATCTAAGTTCTGAAGAGTCCCATTGGATGAACA
AGAACAACAACTTATGTAGATCGATTGTGTGGCCTGAGAGTAACACGAGTCTAAAGCAACCTGATGCTGAACTTACGGAGGGTTTCATGGAAGACGATGCCGAATTCAAAAATGGTGACC
ACAAGTACATCTCAGAGATAATGTTGGCATCAGGGCTTCTACGAGATATCGACTACAGCATGATAAGCATCCAGCTGCACCAAGCACACCTACCGATCAACCCGAGCCTTTTCTTTGTACTG
GAACAGAACAAGACAAGCAATGTGAGTCTACAGGACAACAAGCACAAAGGCAGAGGATTTGGACAACAACAAACGGTGAACCTGGTTGAGAGAAGTAAGAGGAAGCTCATATTTGACA
CCATCAACGAGATCTTAGCTCACAGATTCGCTGCAGAAGGGTGTACAAAGCAACCATCCATAACATTATCAATCAGCACGCAAAGGACACACGAAAAAAGTTCAAGAGGGGAAGAGCTTC
TGCAAACTCTGTGTTCAGAGATTGATCGATTACAAGATAACTCAAAGTGTATCTTGGATGAGGACGATGAAGACCTCATTTGGGAGGATCTGCAAAGCCATGGCATGAACTGGAAGGAGAT
TGAAGGAGAGACACCAGGGTTAGTGTTAGACATTGAGAGGCTAATCTTCAAAGACTTGATTGGTGAAGTTGTGACAAGCGAGTTTGCAGCTTTTCCAAGGATGCTCAGTGGGCAACCAAG
GCAGCTTTTTCATTGCTAA
```

LNG1 *Arabidopsis thaliana* protein sequence:

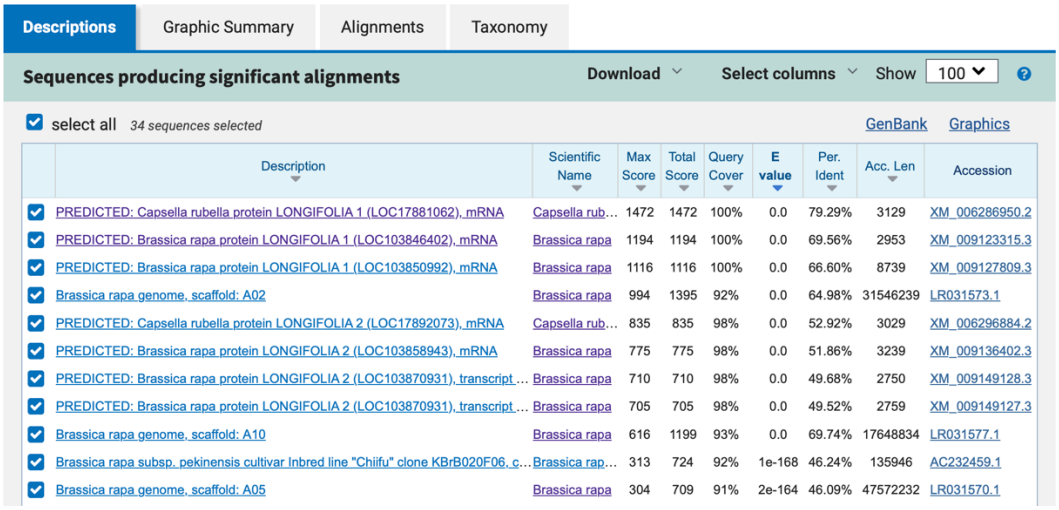
```
MSAKLLYNLSDENPNLNKQIGCMNGIFQVFYRQHYPPRRVTGDELKSLPSGKASDNVGDTNISADKKETEKSKKKKTAKEKQRGVSSESSSRLSFSSSPCSSSFSSADISTTASQFEQPGLSNGE
NPVREPTNGSPRWGGLMMPSDIRELVRSSIHKETRTRDEEALSQQPKSARANVSLLKESSPSRNSNEWSEGRRVVKLKDSPRFSYDERETRKTGAKLKETPRLSLDSRSNSFRSARSSCSPEPQ
ELVTGHRRTTSSVVAKLMGLEVIPDEPVTIQNRENRFCDSPRPTSRVEVDLQRSRGFDSIKKMMPAKFPMKASPWAQVDGAKNQVKIPDATTLTVYGEIQKRLSQLEFKKSEKDLRALKQILEAME
KTQQLISKDDDDNKTLCSSNFMQRNNQPIPSAINTSSMNFKSSSIVVMKAATAPVFKDTGIAGSASFSPRNVALPNVKVGNLRQAQKVIPRKQSAMDVTPRPGYYKGQTESTMKNTSTRPLQSKS
DMAKSGKIQKPSVSLRTPPKKLGFEKQSRPTSPKPELNKNQRQQLSRQQTESASPRRKPGIKSRGLQQSEDRLSDESSDLRSLRSDSNVSLASNLDTEVTSRYNYERNSDITEQHTPKQRSPDL
GMRSLSKPLKVTVEQPSPVSVLDVAFDEDDSPSPVRKISIVFKEDDNLSSEESHWMNKNNNLCRSIVWPESNTSLKQPDAELTEGFMEDDAEFKNGDHKYISEIMLASGLLRDIDYSMISIQLHQAH
LPINPSLFFVLEQNKTSNVSLQDNKHKGRGFGQQQTVNLVERSKRKLIFDTINEILAHRFAAEGCTKQPSITLSISTQRTHEKSSRGEELLQTLCSEIDRLQDNSKCILDEDDEDLIWEDLQSHGMNW
KEIEGETPGLVLDIERLIFKDLIGEVVTSEFAAFPRMLSGQPRQLFHC
```

Method: blastn (to use protein sequence to search nucleotide databases)
Database: Core nucleotide database (core_nt)
Organism: Brassica rapa (taxid:3711); Capsella rubella (taxid:81985)

Search Screenshot:



Blast Result Screenshot:



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *34 sequences selected* GenBank Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Capsella rubella protein LONGIFOLIA 1 (LOC17881062), mRNA	Capsella rub…	1472	1472	100%	0.0	79.29%	3129	XM_006286950.2
PREDICTED: Brassica rapa protein LONGIFOLIA 1 (LOC103846402), mRNA	Brassica rapa	1194	1194	100%	0.0	69.56%	2953	XM_009123315.3
PREDICTED: Brassica rapa protein LONGIFOLIA 1 (LOC103850992), mRNA	Brassica rapa	1116	1116	100%	0.0	66.60%	8739	XM_009127809.3
Brassica rapa genome, scaffold: A02	Brassica rapa	994	1395	92%	0.0	64.98%	31546239	LR031573.1
PREDICTED: Capsella rubella protein LONGIFOLIA 2 (LOC17892073), mRNA	Capsella rub…	835	835	98%	0.0	52.92%	3029	XM_006296884.2
PREDICTED: Brassica rapa protein LONGIFOLIA 2 (LOC103858943), mRNA	Brassica rapa	775	775	98%	0.0	51.86%	3239	XM_009136402.3
PREDICTED: Brassica rapa protein LONGIFOLIA 2 (LOC103870931), transcript …	Brassica rapa	710	710	98%	0.0	49.68%	2750	XM_009149128.3
PREDICTED: Brassica rapa protein LONGIFOLIA 2 (LOC103870931), transcript …	Brassica rapa	705	705	98%	0.0	49.52%	2759	XM_009149127.3
Brassica rapa genome, scaffold: A10	Brassica rapa	616	1199	93%	0.0	69.74%	17648834	LR031577.1
Brassica rapa subsp. pekinensis cultivar Inbred line "Chiifu" clone KBrB020F06, c…	Brassica rap…	313	724	92%	1e-168	46.24%	135946	AC232459.1
Brassica rapa genome, scaffold: A05	Brassica rapa	304	709	91%	2e-164	46.09%	47572232	LR031570.1

[Q3] 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 [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don’t have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

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 plants or protozoa.

Chosen sequence of “novel” protein:

```
>B. rapa protein LONGIFOLIA 1 (sequence taken from BLAST result)
MSAKLLYNLSDENPNLNKQFGCMNGIFQVFYRQHYPARRVSVAGDELKSLPSGKTSDNVGVTNGSTDKKETEKSKKK
KAAKEKQKVVSSESSSRLSFSSSPCSSSFSSADISTTTSQFEQPMSNGETPAREPTYGSPRWGGLVMSSDLRELVRS
SIHKETRTRVEEEALSQQPKSARANVSLLKELSPSRSSNEWSEGRRVVKLKDSPRFSYDEREARKTGAKFKETPRLS
LDSRSNSFRSAKSSCSPEPQELVTGHRRTTSSVIAKLMGLDVVSDEPVTDQSRENHFCDSPRPAPRVEADLPRSRGS
DSFKKMMPAAKFPAKTAPWTQADGARNQVKAADAAATLTVYGEIQKRLSQLEFKKSEKDLRALQQILEAMEKTQQLM
SKDDDNSSLSSTNFMQPSPSSKSIRSSSIVVMKAASAPVFKETGSSSSTSSSPRSVALPNVKVSNQKGTTRKQSAMD
VTPRPATKNTSTRPLQSKIEMAKSGKPSVSPRTQPKKLGFEKQSRPTSPKPEPNKNQRQQLSRQQTESPSPRRKPGM
KSRGLQQSEDRSSDESSDLRSLRSDSNVSSASNFDIEVTSRHKCDLTEQHTPKQRSPELGMRSLPKPLKITVEQPSP
VSILDVAFDDDESPSPVRKISIVFKDDDHIRSEESLWMKKHNNLCRSIVWPESNTSLNQPDAVLTESFMEEGADLRN
GDRKYISEILSASGLLKDIDYSMLSIQLHQAHLPINPSLFFVLEQNKTSNVTHRGRGFGQQTANLIGRSRRKLVFDT
VNEILARKFAAEGCTKQPYITSSISPLMKTDKSSRGKELLEALCSEIDRLQDNSNCILDEDDEDLIWEDLQSQGMNW
KEIEGETPGLVLDIERLIFKDLISEVVTSEVAAFPGNKLSGQPRQLFHC
```

Name: protein LONGIFOLIA 1
Species: Brassica rapa
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Brassiceae; Brassica

[Q4] 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 (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.

Chosen match: Accession ID: XM_009123315.3 (see below for alignment details)
Query Coverage: 100%, E-value: 0.0: Percent Identity: 69.56%

PREDICTED: Brassica rapa protein LONGIFOLIA 1 (LOC103846402), mRNA

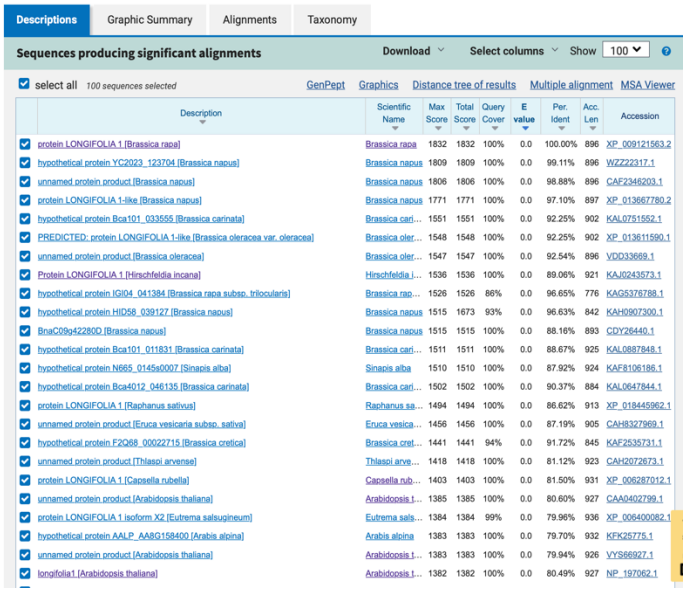
Sequence ID: XM_009123315.3 Length: 2953 Number of Matches: 1

Range 1: 86 to 2773 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
1194 bits(3088)	0.0	Compositional matrix adjust.	746/933(80%)	800/933(85%)	43/933(4%)	+2

```
Query  1     MSAKLLYNLSDENPNLNKQIGCMNGIFQVFYRQHYPPRRVT--GDELKSLPSGKASDNVG
Sbjct  86    MSAKLLYNLSDENPNLNKQ GCMNGIFQVFYRQHYP RRV+  GDELKSLPSGK SDNVG

Query  59    DTNIsadkketekskkkktakekQRGVssessrlsfssspcsssfssadisttasQFEQ
              TN S DKKETEKSKKKK AKEKQ+ VSSESSSRLSFSSSPCSSSFSSADISTT SQFEQ
Sbjct  266   VTNGSTDKKETEKSKKKKAAKEKQKVVSSESSSRLSFSSSPCSSSFSSADISTTTSQFEQ

Query  119   PGLSNGENPVREPTNGSPRWGGLMMPSDIRELVRSSIHKETRTR-DEEALSQQPKSARAN
P +SNGE P REPT GSPRWGGL+M SD+RELVRSSIHKETRTR +EEALSQQPKSARAN
Sbjct  446   P-MSNGETPAREPTYGSPRWGGLVMSSDLRELVRSSIHKETRTRVEEEALSQQPKSARAN

Query  178   VSLLKESSPSRNSNEWSEGRRVVKLKDSPRFSYDERETRKTGAKLKETPRLSLDsrsnsf
VSLLKE SPSR+SNEWSEGRRVVKLKDSPRFSYDERE RKTGAK KETPRLSLDSRSNSF
Sbjct  623   VSLLKELSPSRSSNEWSEGRRVVKLKDSPRFSYDEREARKTGAKFKETPRLSLDSRSNSF

Query  238   rsarsscsPEPQELVTGHRRTTSSVVAKLMGLEVIPDEPVTIQNRENRFCDSPRPTSRVE
RSA+SSCSPEPQELVTGHRRTTSSV+AKLMGL+V+ DEPVT Q+REN FCDSPRP  RVE
Sbjct  803   RSAKSSCSPEPQELVTGHRRTTSSVIAKLMGLDVVSDEPVTDQSRENHFCDSPRPAPRVE

Query  298   VDLQRSRGFDSIkkmmpakfp-mkaSPWAQVDGAKNQVKIPDAT-TLTVYGEIQKRLSQL
 DL RSRG DS KKMMPA     K +PW Q DGA+NQVK  DA  TLTVYGEIQKRLSQL
Sbjct  983   ADLPRSRGSDSFKKMMPAAKFPAKTAPWTQADGARNQVKAADAAATLTVYGEIQKRLSQL

Query  356   EFKKSEKDLRALKQILEAMEKTQQLISKDDDDNKTLCSSNFMQRNNQPIPSAINTSSMNF
EFKKSEKDLRAL+QILEAMEKTQQL+SKDDD N +L S+NFMQ +          SS +
Sbjct  1163  EFKKSEKDLRALQQILEAMEKTQQLMSKDDD-NSSLSSTNFMQPS---------PSSKSI

Query  416   KSSSIVVMKAATAPVFKDTGIAGSASFSPRNVALPNVKVGNLRQAQKVIPRKQSAMDVTP
+SSSIVVMKAA+APVFK+TG + S S SPR+VALPNVKV N   QK  RKQSAMDVTP
Sbjct  1313  RSSSIVVMKAASAPVFKETGSSSSTSSSPRSVALPNVKVSN-----QKGTTRKQSAMDVTP

Query  476   RPGYYKGQTESTMKNTSTRPLQSKSDMAKSGKIQKPSVSLRTPPKKLGFEKQSRPTSPKP
RP          KNTSTRPLQSK +MAKSGK   PSVS RT PKKLGFEKQSRPTSPKP
Sbjct  1481  RPA---------TKNTSTRPLQSKIEMAKSGK----PSVSPRTQPKKLGFEKQSRPTSPKP

Query  536   ELNKNQRQQLSRQQTESASPRRKPGIKSRGLQQsedrlsdessdlrslrsdsNVSLASNL
E NKNQRQQLSRQQTES SPRRKPG+KSRGLQQSEDR SDESSDLRSLRSDSNVS ASN
Sbjct  1625  EPNKNQRQQLSRQQTESPSPRRKPGMKSRGLQQSEDRSDESSDLRSLRSDSNVSSASNF

Query  596   DTEVTSRYNYERNSDITEQHTPKQRSPDLGMRSLSKPLKVTVEQPSPVSVLDVAFDEDDS
D EVTSR+      D+TEQHTPKQRSP+LGMRSL KPLK+TVEQPSPVS+LDVAFD+D+S
Sbjct  1805  DIEVTSRHK-----CDLTEQHTPKQRSPELGMRSLPKPLKITVEQPSPVSILDVAFDDDES

Query  656   PSPVRKISIVFKEDDNLSSEESHWMNKNNNLCRSIVWPESNTSLKQPDAELTEGFMEDDA
PSPVRKISIVFK+DD++ SEES WM K+NNLCRSIVWPESNTSL QPDA LTE FME+ A
Sbjct  1973  PSPVRKISIVFKDDDDHIRSEESLWMKKHNNLCRSIVWPESNTSLNQPDAVLTESFMEEGA

Query  716   EFKNGDHKYISEIMLASGLLRDIDYSMISIQLHQAHLPINPSLFFVLEQNKTSNVSLQDN
+ +NGD KYISEI+ ASGLL+DIDYSM+SIQLHQAHLPINPSLFFVLEQNKTSNV+
Sbjct  2153  DLRNGDRKYISEILSASGLLKDIDYSMLSIQLHQAHLPINPSLFFVLEQNKTSNVT----

Query  776   KHKGRGFGQQQTVNLVERSKRKLIFDTINEILAHRFAAEGCTKQPSITLSISTQRTHEKS
 H+GRGFG QQT NL+ RS+RKL+FDT+NEILA +FAAEGCTKQP IT SIS      +KS
Sbjct  2321  -HRGRGFG-QQTANLIGRSRRKLVFDTVNEILARKFAAEGCTKQPYITSSISPLMKTDKS

Query  836   SRGEELLQTLCSEIDRLQDNSKCildeddedliwedlQSHGMNWKEIEGETPGLVLDIER
SRG+ELL+ LCSEIDRLQDNS CILDEDDEDLIWEDLQS GMNWKEIEGETPGLVLDIER
Sbjct  2495  SRGKELLEALCSEIDRLQDNSNCILDEDDEDLIWEDLQSQGMNWKEIEGETPGLVLDIER

Query  896   LIFKDLIGEVVTSEFAAFP-RMLSGQPRQLFHC  927
LIFKDLI EVVTSE AAFP   LSGQPRQLFHC
Sbjct  2675  LIFKDLISEVVTSEVAAFPGNKLSGQPRQLFHC  2773
```

To verify if I have identified a novel gene, I will run a BLASTp search against the non-redundant (nr) database. This will allow me to determine if there are any matches and confirm the novelty. If there is a match with 100% identity but to a different species than the one you started with, then you have likely succeeded in finding a novel gene



Descriptions | Graphic Summary | Alignments | Taxonomy

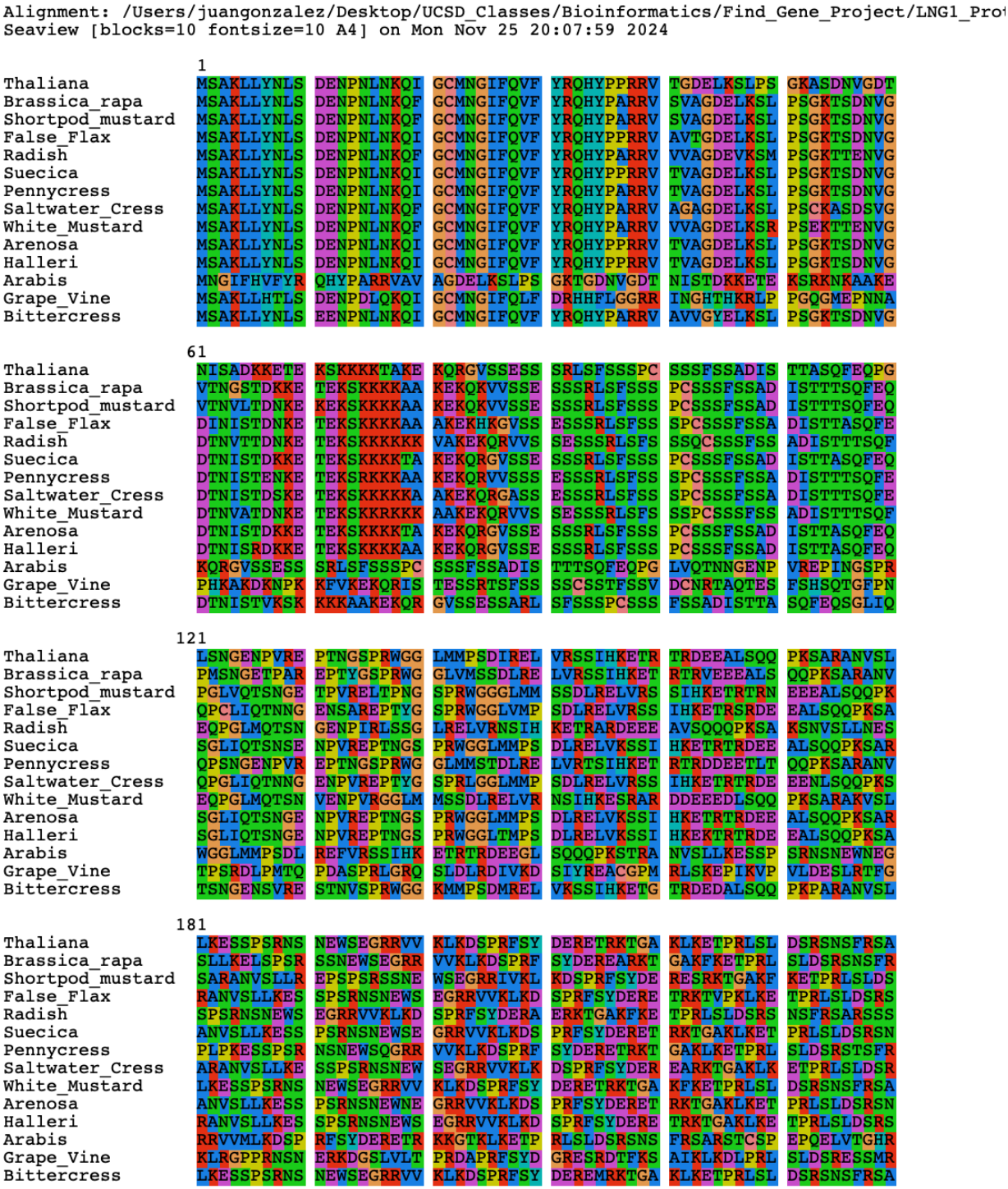
Sequences producing significant alignments Download Select columns Show 100

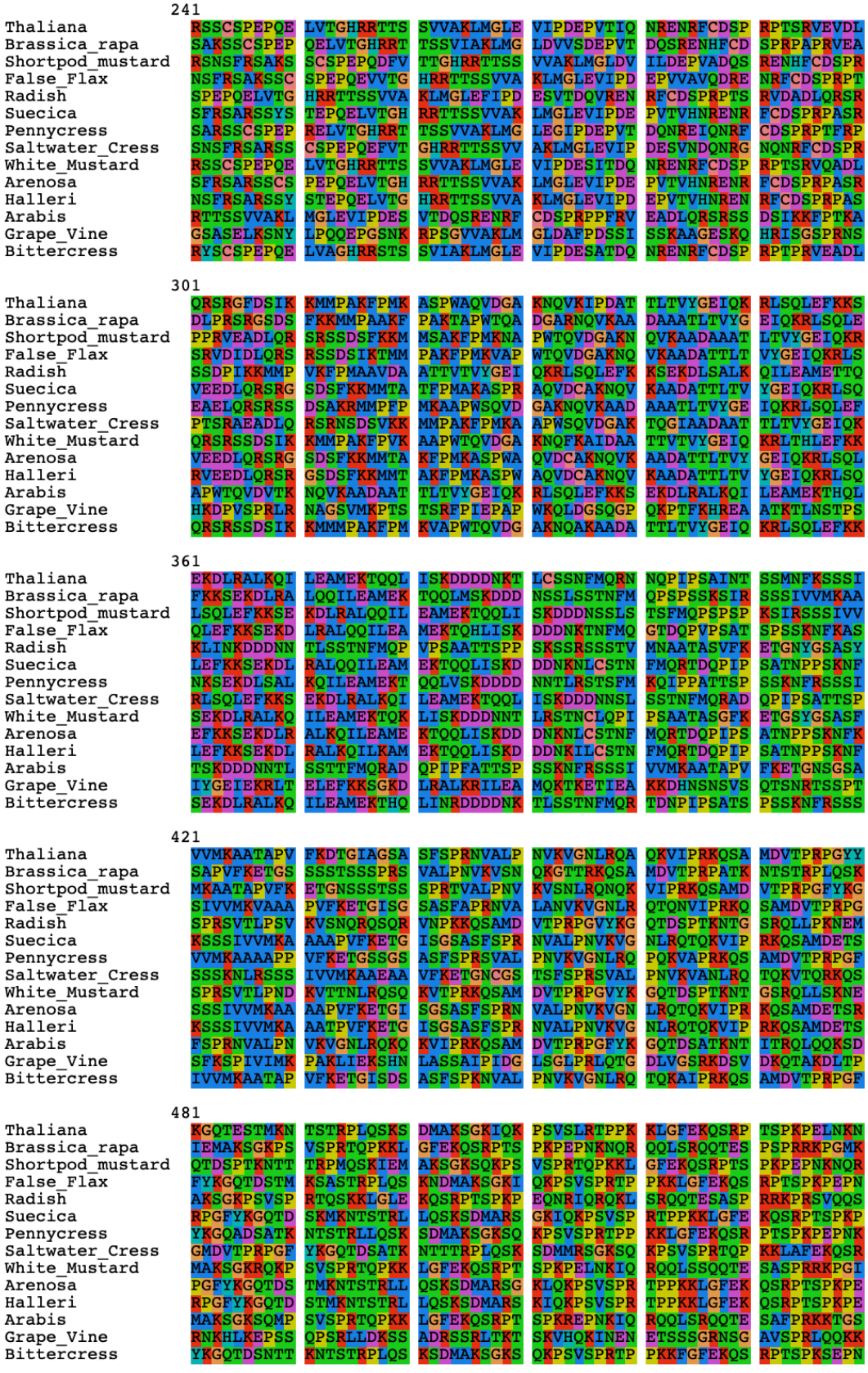
select all 100 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

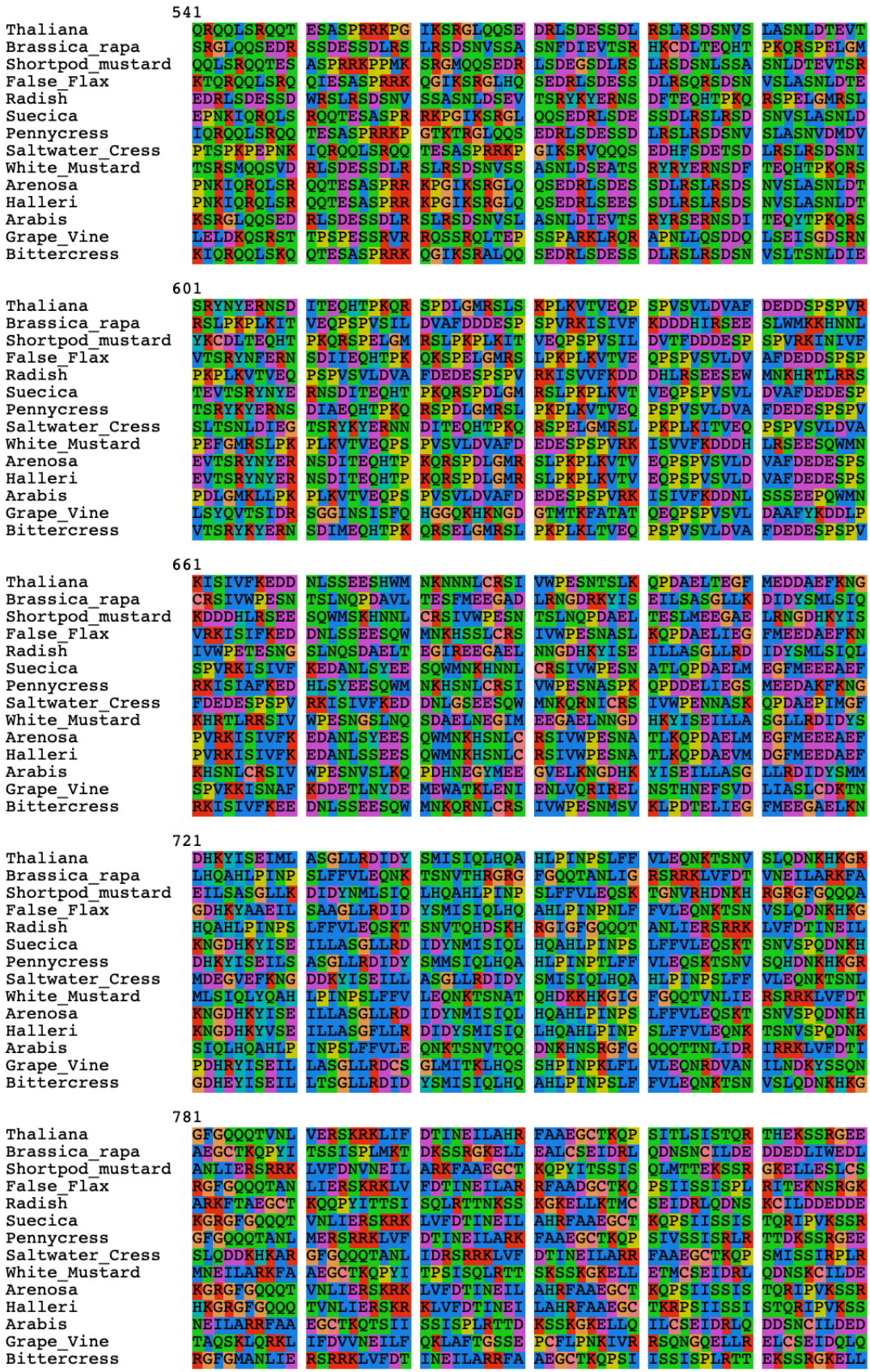
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
protein LONGIFOLIA 1 [Brassica rapa]	Brassica rapa	1832	1832	100%	0.0	100.00%	896	XP_009121563.2
hypothetical protein YC2023_123704 [Brassica napus]	Brassica napus	1809	1809	100%	0.0	99.11%	896	WZZ22317.1
unnamed protein product [Brassica napus]	Brassica napus	1806	1806	100%	0.0	98.88%	896	CAF2346203.1
protein LONGIFOLIA 1-like [Brassica napus]	Brassica napus	1771	1771	100%	0.0	97.10%	897	XP_013667780.2
hypothetical protein Bca101_033555 [Brassica carinata]	Brassica cari…	1551	1551	100%	0.0	92.25%	902	KAL0751552.1
PREDICTED: protein LONGIFOLIA 1-like [Brassica oleracea var. oleracea]	Brassica oler…	1548	1548	100%	0.0	92.25%	902	XP_013611590.1
unnamed protein product [Brassica oleracea]	Brassica oler…	1547	1547	100%	0.0	92.54%	896	VDD33669.1
Protein LONGIFOLIA 1 [Hirschfeldia incana]	Hirschfeldia i…	1536	1536	100%	0.0	89.06%	921	KAJ0243573.1
hypothetical protein IGI04_041384 [Brassica rapa subsp. trilocularis]	Brassica rap…	1526	1526	86%	0.0	96.65%	776	KAG5376788.1
hypothetical protein HID58_039127 [Brassica napus]	Brassica napus	1515	1673	93%	0.0	96.63%	842	KAH0907300.1
BnaC09g42280D [Brassica napus]	Brassica napus	1515	1515	100%	0.0	88.16%	893	CDY26440.1
hypothetical protein Bca101_011831 [Brassica carinata]	Brassica cari…	1511	1511	100%	0.0	88.67%	925	KAL0887848.1
hypothetical protein N665_0145s0007 [Sinapis alba]	Sinapis alba	1510	1510	100%	0.0	87.92%	924	KAF8106186.1
hypothetical protein Bca4012_046135 [Brassica carinata]	Brassica cari…	1502	1502	100%	0.0	90.37%	884	KAL0647844.1
protein LONGIFOLIA 1 [Raphanus sativus]	Raphanus sa…	1494	1494	100%	0.0	86.62%	913	XP_018445962.1
unnamed protein product [Eruca vesicaria subsp. sativa]	Eruca vesica…	1456	1456	100%	0.0	87.19%	905	CAH8327969.1
hypothetical protein F2Q68_00022715 [Brassica cretica]	Brassica cret…	1441	1441	94%	0.0	91.72%	845	KAF2535731.1
unnamed protein product [Thlaspi arvense]	Thlaspi arve…	1418	1418	100%	0.0	81.12%	923	CAH2072673.1
protein LONGIFOLIA 1 [Capsella rubella]	Capsella rub…	1403	1403	100%	0.0	81.50%	931	XP_006287012.1
unnamed protein product [Arabidopsis thaliana]	Arabidopsis t…	1385	1385	100%	0.0	80.60%	927	CAA0402799.1
protein LONGIFOLIA 1 isoform X2 [Eutrema salsugineum]	Eutrema sals…	1384	1384	99%	0.0	79.96%	936	XP_006400082.1
hypothetical protein AALP_AA8G158400 [Arabis alpina]	Arabis alpina	1383	1383	100%	0.0	79.70%	932	KFK25775.1
unnamed protein product [Arabidopsis thaliana]	Arabidopsis t…	1383	1383	100%	0.0	79.94%	926	VYS66927.1
longifolia1 [Arabidopsis thaliana]	Arabidopsis t…	1382	1382	100%	0.0	80.49%	927	NP_197062.1

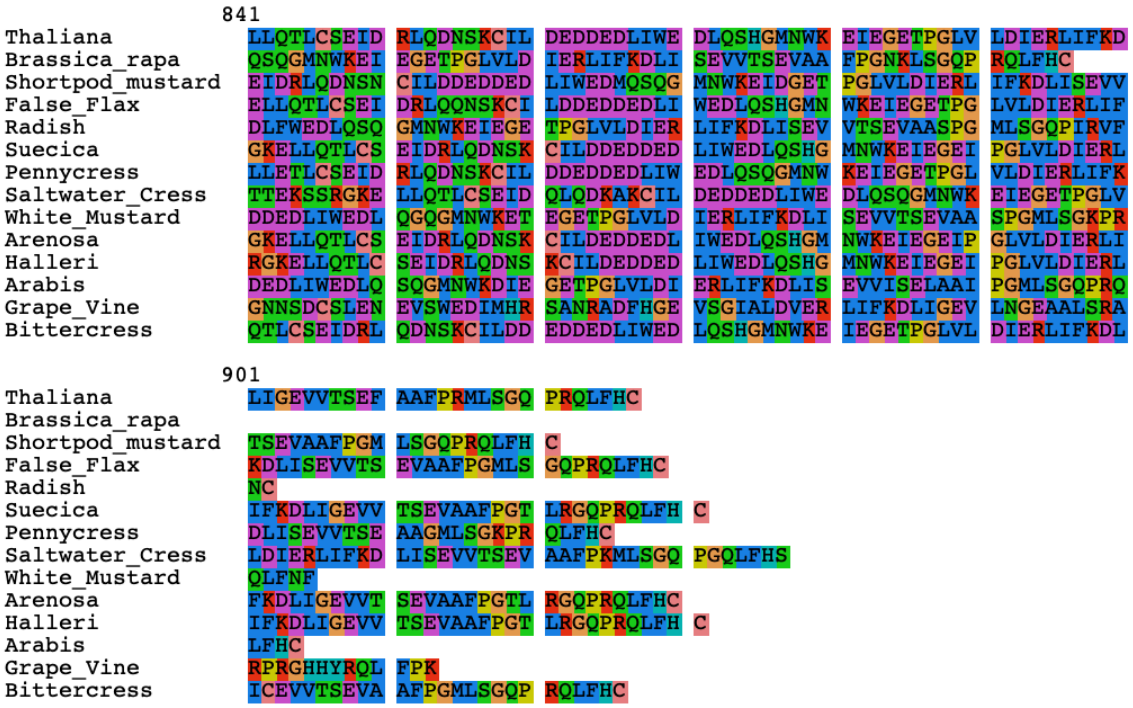
[**Q5**] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

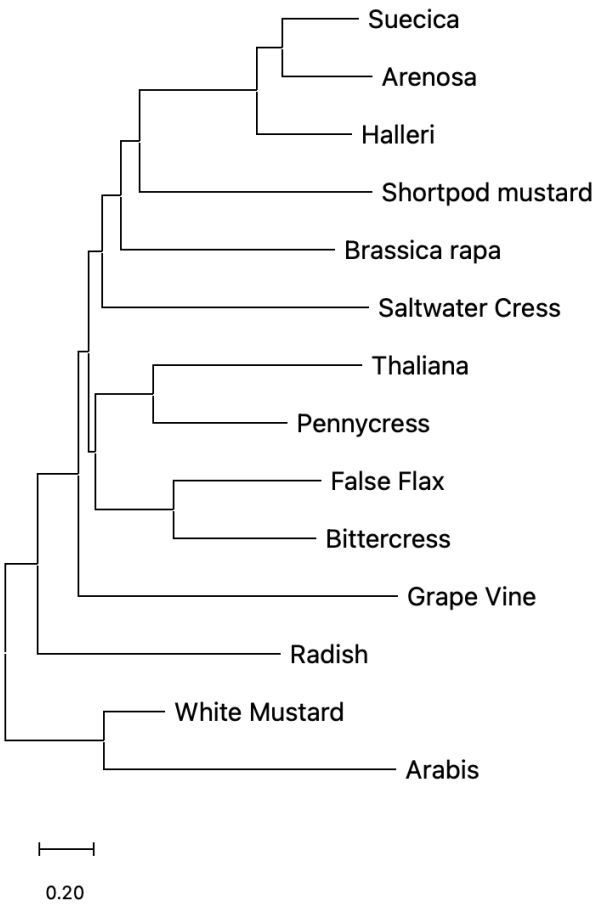

Alignment: /Users/juangonzalez/Desktop/UCSD_Classes/Bioinformatics/Find_Gene_Project/LNG1_Prot
Seaview [blocks=10 fontsize=10 A4] on Mon Nov 25 20:07:59 2024
1
Thaliana MSAKLLYNLS DENPNLNKQI GCMNGIFQVF YRQHYPPRRV TGDELKSLPS GKASDNVGDT
Brassica_rapa MSAKLLYNLS DENPNLNKQF GCMNGIFQVF YRQHYPARRV SVAGDELKSL PSGKTSDNVG
Shortpod_mustard MSAKLLYNLS DENPNLNKQF GCMNGIFQVF YRQHYPARRV SVAGDELKSL PSGKTSDNVG
False_Flax MSAKLLYNLS DENPNLNKQI GCMNGIFQVF YRQHYPPRRV AVTGDELKSL PSGKTSDNVG
Radish MSAKLLYNLS DENPNLNKQF GCMNGIFQVF YRQHYPARRV VVAGDEVKSM PSGKTTENVG
Suecica MSAKLLYNLS DENPNLNKQI GCMNGIFQVF YRQHYPPRRV TVAGDELKSL PSGKTSDNVG
Pennycress MSAKLLYNLS DENPNLNKQI GCMNGIFQVF YRQHYPARRV TVAGDELKSL PSGKTSDNVG
Saltwater_Cress MSAKLLYNLS DENPNLNKQF GCMNGIFQVF YRQHYPARRV AGAGDELKSL PSCKASDSVG
White_Mustard MSAKLLYNLS DENPNLNKQF GCMNGIFQVF YRQHYPARRV VVAGDELKSR PSEKTTENVG
Arenosa MSAKLLYNLS DENPNLNKQI GCMNGIFQVF YRQHYPPRRV TVAGDELKSL PSGKTSDNVG
Halleri MSAKLLYNLS DENPNLNKQI GCMNGIFQVF YRQHYPPRRV TVAGDELKSL PSGKTSDNVG
Arabis MNGIFHVFYR QHYPARRVAV AGDELKSLPS GKTGDNVGDT NISTDKKETE KSRKNKAAKE
Grape_Vine MSAKLLHTLS DENPDLQKQI GCMNGIFQLF DRHHFLGGRR INGHTHKRLP PGQGMEPNNA
Bittercress MSAKLLYNLS EENPNLNKQI GCMNGIFQVF YRQHYPARRV AVVGYELKSL PSGKTSDNVG
61
Thaliana NISADKKETE KSKKKKTAKE KQRGVSSESS SRLSFSSSPC SSSFSSADIS TTASQFEQPG
Brassica_rapa VTNGSTDKKE TEKSKKKKAA KEKQKVVSSE SSSRLSFSSS PCSSSFSSAD ISTTTSQFEQ
Shortpod_mustard VTNVLTDNKE KEKSKKKKAA KEKQKVVSSE SSSRLSFSSS PCSSSFSSAD ISTTTSQFEQ
False_Flax DINISTDNKE TEKSKKKKAA AKEKHKGVSS ESSSRLSFSS SPCSSSFSSA DISTTASQFE
Radish DTNVTTDNKE TEKSKKKKKK VAKEKQRVVS SESSSRLSFS SSQCSSSFSS ADISTTTSQF
Suecica DTNISTDKKE TEKSKKKKTA KEKQRGVSSE SSSRLSFSSS PCSSSFSSAD ISTTASQFEQ
Pennycress DTNISTENKE TEKSRKKKAA KEKQRVVSSS ESSSRLSFSS SPCSSSFSSA DISTTTSQFE
Saltwater_Cress DTNISTDSKE TEKSKKKKKA AKEKQRGASS ESSSRLSFSS SPCSSSFSSA DISTTTSQFE
White_Mustard DTNVATDNKE TEKSKKRKKK AAKEKQRVVS SESSSRLSFS SSPCSSSFSS ADISTTTSQF
Arenosa DTNISTDKKE TEKSKKKKTA KEKQRGVSSE SSSRLSFSSS PCSSSFSSAD ISTTASQFEQ
Halleri DTNISRDKKE TEKSKKKKAA KEKQRGVSSE SSSRLSFSSS PCSSSFSSAD ISTTASQFEQ
Arabis KQRGVSSESS SRLSFSSSPC SSSFSSADIS TTTSQFEQPG LVQTNNGENP VREPINGSPR
Grape_Vine PHKAKDKNPK KFVKEKQRIS TESSRTSFSS SSCSSTFSSV DCNRTAQTES FSHSQTGFPN
Bittercress DTNISTVKSK KKKAAKEKQR GVSSESSARL SFSSSPCSSS FSSADISTTA SQFEQSGLIQ
121
Thaliana LSNGENPVRE PTNGSPRWGG LMMPSDIREL VRSSIHKETR TRDEEALSQQ PKSARANVSL
Brassica_rapa PMSNGETPAR EPTYGSPRWG GLVMSSDLRE LVRSSIHKET RTRVEEEALS QQPKSARANV
Shortpod_mustard PGLVQTSNGE TPVRELTPNG SPRWGGGLMM SSDLRELVRS SIHKETRTRN EEEALSQQPK
False_Flax QPCLIQTNNG ENSAREPTYG SPRWGGLVMP SDLRELVRSS IHKETRSRDE EALSQQPKSA
Radish EQPGLMQTSN GENPIRLSSG LRELVRNSIH KETRARDEEE AVSQQQPKSA KSNVSLLNES
Suecica SGLIQTSNSE NPVREPTNGS PRWGGLMMPS DLRELVKSSI HKETRTRDEE ALSQQPKSAR
Pennycress QPSNGENPVR EPTNGSPRWG GLMMSTDLRE LVRTSIHKET RTRDDEETLT QQPKSARANV
Saltwater_Cress QPGLIQTNNG ENPVREPTYG SPRLGGLMMP SDLRELVRSS IHKETRTRDE EENLSQQPKS
White_Mustard EQPGLMQTSN VENPVRGGLM MSSDLRELVR NSIHKESRAR DDEEEDLSQQ PKSARAKVSL
Arenosa SGLIQTSNGE NPVREPTNGS PRWGGLMMPS DLRELVKSSI HKETRTRDEE ALSQQPKSAR
Halleri SGLIQTSNGE NPVREPTNGS PRWGGLTMPS DLRELVKSSI HKEKTRTRDE EALSQQPKSA
Arabis WGGLMMPSDL REFVRSSIHK ETRTRDEEGL SQQQPKSTRA NVSLLKESSP SRNSNEWNEG
Grape_Vine TPSRDLPMTQ PDASPRLGRQ SLDLRDIVKD SIYREACGPM RLSKEPIKVP VLDESLRTFG
Bittercress TSNGENSVRE STNVSPRWGG KMMPSDMREL VKSSIHKETG TRDEDALSQQ PKPARANVSL
181
Thaliana LKESSPSRNS NEWSEGRRVV KLKDSPRFSY DERETRKTGA KLKETPRLSL DSRSNSFRSA
Brassica_rapa SLLKELSPSR SSNEWSEGRR VVKLKDSPRF SYDEREARKT GAKFKETPRL SLDSRSNSFR
Shortpod_mustard SARANVSLLR EPSPSRSSNE WSEGRRIVKL KDSPRFSYDE RESRKTGAKF KETPRLSLDS
False_Flax RANVSLLKES SPSRNSNEWS EGRRVVKLKD SPRFSYDERE TRKTVPKLKE TPRLSLDSRS
Radish SPSRNSNEWS EGRRVVKLKD SPRFSYDERA ERKTGAKFKE TPRLSLDSRS NSFRSARSSS
Suecica ANVSLLKESS PSRNSNEWSE GRRVVKLKDS PRFSYDERET RKTGAKLKET PRLSLDSRSN
Pennycress PLPKESSPSR NSNEWSQGRR VVKLKDSPRF SYDERETRKT GAKLKETPRL SLDSRSTSFR
Saltwater_Cress ARANVSLLKE SSPSRNSNEW SEGRRVVKLK DSPRFSYDER EARKTGAKLK ETPRLSLDSR
White_Mustard LKESSPSRNS NEWSEGRRVV KLKDSPRFSY DERETRKTGA KFKETPRLSL DSRSNSFRSA
Arenosa ANVSLLKESS PSRNSNEWNE GRRVVKLKDS PRFSYDERET RKTGAKLKET PRLSLDSRSN
Halleri RANVSLLKES SPSRNSNEWS EGRRVVKLKD SPRFSYDERE TRKTGAKLKE TPRLSLDSRS
Arabis RRVVMLKDSP RFSYDERETR KKGTKLKETP RLSLDSRSNS FRSARSTCSP EPQELVTGHR
Grape_Vine KLRGPPRNSN ERKDGSLVLT PRDAPRFSYD GRESRDTFKS AIKLKDLPRL SLDSRESSMR
Bittercress LKESSPSRNS NEWSEGRRVV KLKDSPRFSY DEREMRKTGA KLKETPRLSL DSRSNSFRSA


241
Thaliana RSSCSPEPQE LVTGHRRTTS SVVAKLMGLE VIPDEPVTIQ NRENRFCDSP RPTSRVEVDL
Brassica_rapa SAKSSCSPEP QELVTGHRRT TSSVIAKLMG LDVVSDEPVT DQSRENHFCD SPRPAPRVEA
Shortpod_mustard RSNSFRSAKS SCSPEPQDFV TTGHRRTTSS VVAKLMGLDV ILDEPVADQS RENHFCDSPR
False_Flax NSFRSAKSSC SPEPQEVVTG HRRTTSSVVA KLMGLEVIPD EPVVAVQDRE NRFCDSPRPT
Radish SPEPQELVTG HRRTTSSVVA KLMGLEFIPD ESVTDQVREN RFCDSPRPTS RVDADLQRSR
Suecica SFRSARSSYS TEPQELVTGH RRTTSSVVAK LMGLEVIPDE PVTVHNRENR FCDSPRPASR
Pennycress SARSSCSPEP RELVTGHRRT TSSVVAKLMG LEGIPDEPVT DQNREIQNRF CDSPRPTFRP
Saltwater_Cress SNSFRSARSS CSPEPQEFVT GHRRTTSSVV AKLMGLEVIP DESVNDQNRG NQNRFCDSPR
White_Mustard RSSCSPEPQE LVTGHRRTTS SVVAKLMGLE VIPDESITDQ NRENRFCDSP RPTSRVQADL
Arenosa SFRSARSSCS PEPQELVTGH RRTTSSVVAK LMGLEVIPDE PVTVHNRENR FCDSPRPASR
Halleri NSFRSARSSY STEPQELVTG HRRTTSSVVA KLMGLEVIPD EPVTVHNREN RFCDSPRPAS
Arabis RTTSSVVAKL MGLEVIPDES VTDQSRENRF CDSPRPPFRV EADLQRSRSS DSIKKFPTKA
Grape_Vine GSASELKSNY LPQQEPGSNK RPSGVVAKLM GLDAFPDSSI SSKAAGESKQ HRISGSPRNS
Bittercress RYSCSPEPQE LVAGHRRSTS SVIAKLMGLE VIPDESATDQ NRENRFCDSP RPTPRVEADL
301
Thaliana QRSRGFDSIK KMMPAKFPMK ASPWAQVDGA KNQVKIPDAT TLTVYGEIQK RLSQLEFKKS
Brassica_rapa DLPRSRGSDS FKKMMPAAKF PAKTAPWTQA DGARNQVKAA DAAATLTVYG EIQKRLSQLE
Shortpod_mustard PPRVEADLQR SRSSDSFKKM MSAKFPMKNA PWTQVDGAKN QVKAADAAAT LTVYGEIQKR
False_Flax SRVDIDLQRS RSSDSIKTMM PAKFPMKVAP WTQVDGAKNQ VKAADATTLT VYGEIQKRLS
Radish SSDPIKKMMP VKFPMAAVDA ATTVTVYGEI QKRLSQLEFK KSEKDLSALK QILEAMETTQ
Suecica VEEDLQRSRG SDSFKKMMTA TFPMAKASPR AQVDCAKNQV KAADATTLTV YGEIQKRLSQ
Pennycress EAELQRSRSS DSAKRMMPFP MKAAPWSQVD GAKNQVKAAD AAATLTVYGE IQKRLSQLEF
Saltwater_Cress PTSRAEADLQ RSRNSDSVKK MMPAKFPMKA APWSQVDGAK TQGIAADAAT TLTVYGEIQK
White_Mustard QRSRSSDSIK KMMPAKFPVK AAPWTQVDGA KNQFKAIDAA TTVTVYGEIQ KRLTHLEFKK
Arenosa VEEDLQRSRG SDSFKKMMTA KFPMKASPWA QVDCAKNQVK AADATTLTVY GEIQKRLSQL
Halleri RVEEDLQRSR GSDSFKKMMT AKFPMKASPW AQVDCAKNQV KAADATTLTV YGEIQKRLSQ
Arabis APWTQVDVTK NQVKAADAAT TLTVYGEIQK RLSQLEFKKS EKDLRALKQI LEAMEKTHQL
Grape_Vine HKDPVSPRLR NAGSVMKPTS TSRFPIEPAP WKQLDGSQGP QKPTFKHREA ATKTLNSTPS
Bittercress QRSRSSDSIK KMMMPAKFPM KVAPWTQVDG AKNQAKAADA TTLTVYGEIQ KRLSQLEFKK
361
Thaliana EKDLRALKQI LEAMEKTQQL ISKDDDDNKT LCSSNFMQRN NQPIPSAINT SSMNFKSSSI
Brassica_rapa FKKSEKDLRA LQQILEAMEK TQQLMSKDDD NSSLSSTNFM QPSPSSKSIR SSSIVVMKAA
Shortpod_mustard LSQLEFKKSE KDLRALQQIL EAMEKTQQLI SKDDDNSSLS TSFMQPSPSP KSIRSSSIVV
False_Flax QLEFKKSEKD LRALQQILEA MEKTQHLISK DDDNKTNFMQ GTDQPVPSAT SPSSKNFKAS
Radish KLINKDDDNN TLSSTNFMQP VPSAATTSPP SKSSRSSSTV MNAATASVFK ETGNYGSASY
Suecica LEFKKSEKDL RALQQILEAM EKTQQLISKD DDNKNLCSTN FMQRTDQPIP SATNPPSKNF
Pennycress NKSEKDLSAL KQILEAMEKT QQLVSKDDDD NNTLRSTSFM KQIPPATTSP SSKNFRSSSI
Saltwater_Cress RLSQLEFKKS EKDLRALKQI LEAMEKTQQL ISKDDDNNSL SSTNFMQRAD QPIPSATTSP
White_Mustard SEKDLRALKQ ILEAMEKTQK LISKDDDNNT LRSTNCLQPI PSAATASGFK ETGSYGSASF
Arenosa EFKKSEKDLR ALKQILEAME KTQQLISKDD DNKNLCSTNF MQRTDQPIPS ATNPPSKNFK
Halleri LEFKKSEKDL RALKQILKAM EKTQQLISKD DDNKILCSTN FMQRTDQPIP SATNPPSKNF
Arabis TSKDDDNNTL SSTTFMQRAD QPIPFATTSP SSKNFRSSSI VVMKAATAPV FKETGNSGSA
Grape_Vine IYGEIEKRLT ELEFKKSGKD LRALKRILEA MQKTKETIEA KKDHNSNSVS QTSNRTSSPT
Bittercress SEKDLRALKQ ILEAMEKTHQ LINRDDDDNK TLSSTNFMQR TDNPIPSATS PSSKNFRSSS
421
Thaliana VVMKAATAPV FKDTGIAGSA SFSPRNVALP NVKVGNLRQA QKVIPRKQSA MDVTPRPGYY
Brassica_rapa SAPVFKETGS SSSTSSSPRS VALPNVKVSN QKGTTRKQSA MDVTPRPATK NTSTRPLQSK
Shortpod_mustard MKAATAPVFK ETGNSSSTSS SPRTVALPNV KVSNLRQNQK VIPRKQSAMD VTPRPGFYKG
False_Flax SIVVMKVAAA PVFKETGISG SASFAPRNVA LANVKVGNLR QTQNVIPRKQ SAMDVTPRPG
Radish SPRSVTLPSV KVSNQRQSQR VNPKKQSAMD VTPRPGVYKG QTDSPTKNTG SRQLLPKNEM
Suecica KSSSIVVMKA AAAPVFKETG ISGSASFSPR NVALPNVKVG NLRQTQKVIP RKQSAMDETS
Pennycress VVMKAAAAPP VFKETGSSGS ASFSPRSVAL PNVKVGNLRQ PQKVAPRKQS AMDVTPRPGF
Saltwater_Cress SSSKNLRSSS IVVMKAAEAA VFKETGNCGS TSFSPRSVAL PNVKVANLRQ TQKVTQRKQS
White_Mustard SPRSVTLPND KVTTNLRQSQ KVTPRKQSAM DVTPRPGVYK GQTDSPTKNT GSRQLLSKNE
Arenosa SSSIVVMKAA AAPVFKETGI SGSASFSPRN VALPNVKVGN LRQTQKVIPR KQSAMDETSR
Halleri KSSSIVVMKA AATPVFKETG ISGSASFSPR NVALPNVKVG NLRQTQKVIP RKQSAMDETS
Arabis FSPRNVALPN VKVGNLRQKQ KVIPRKQSAM DVTPRPGFYK GQTDSATKNT ITRQLQQKSD
Grape_Vine SFKSPIVIMK PAKLIEKSHN LASSAIPIDG LSGLPRLQTG DLVGSRKDSV DKQTAKDLTP
Bittercress IVVMKAATAP VFKETGISDS ASFSPKNVAL PNVKVGNLRQ TQKAIPRKQS AMDVTPRPGF
481
Thaliana KGQTESTMKN TSTRPLQSKS DMAKSGKIQK PSVSLRTPPK KLGFEKQSRP TSPKPELNKN
Brassica_rapa IEMAKSGKPS VSPRTQPKKL GFEKQSRPTS PKPEPNKNQR QQLSRQQTES PSPRRKPGMK
Shortpod_mustard QTDSPTKNTT TRPMQSKIEM AKSGKSQKPS VSPRTQPKKL GFEKQSRPTS PKPEPNKNQR
False_Flax FYKGQTDSTM KSASTRPLQS KNDMAKSGKI QKPSVSPRTP PKKLGFEKQS RPTSPKPEPN
Radish AKSGKPSVSP RTQSKKLGLE KQSRPTSPKP EQNRIQRQKL SRQQTESASP RRKPRSVQQS
Suecica RPGFYKGQTD SKMKNTSTRL LQSKSDMARS GKIQKPSVSP RTPPKKLGFE KQSRPTSPKP
Pennycress YKGQADSATK NTSTRLLQSK SDMAKSGKSQ KPSVSPRTPP KKLGFEKQSR PTSPKPEPNK
Saltwater_Cress GMDVTPRPGF YKGQTDSATK NTTTRPLQSK SDMMRSGKSQ KPSVSPRTQP KKLAFEKQSR
White_Mustard MAKSGKRQKP SVSPRTQPKK LGFEKQSRPT SPKPELNKIQ RQQLSSQQTE SASPRRKPGI
Arenosa PGFYKGQTDS TMKNTSTRLL QSKSDMARSG KLQKPSVSPR TPPKKLGFEK QSRPTSPKPE
Halleri RPGFYKGQTD STMKNTSTRL LQSKSDMARS KIQKPSVSPR TPPKKLGFEK QSRPTSPKPE
Arabis MAKSGKSQMP SVSPRTQPKK LGFEKQSRPT SPKREPNKIQ RQQLSRQQTE SAFPRKKTGS
Grape_Vine RNKHLKEPSS QPSRLLDKSS ADRSSRLTKT SKVHQKINEN ETSSSGRNSG AVSPRLQQKK
Bittercress YKGQTDSNTT KNTSTRPLQS KSDMAKSGKS QKPSVSPRTP PKKFGFEKQS RPTSPKSEPN


541
Thaliana QRQQLSRQQT ESASPRRKPG IKSRGLQQSE DRLSDESSDL RSLRSDSNVS LASNLDTEVT
Brassica_rapa SRGLQQSEDR SSDESSDLRS LRSDSNVSSA SNFDIEVTSR HKCDLTEQHT PKQRSPELGM
Shortpod_mustard QQLSRQQTES ASPRRKPPMK SRGMQQSEDR LSDEGSDLRS LRSDSNLSSA SNLDTEVTSR
False_Flax KTQRQQLSRQ QIESASPRRK QGIKSRGLHQ SEDRLSDESS DLRSQRSDSN VSLASNLDTE
Radish EDRLSDESSD WRSLRSDSNV SSASNLDSEV TSRYKYERNS DFTEQHTPKQ RSPELGMRSL
Suecica EPNKIQRQLS RQQTESASPR RKPGIKSRGL QQSEDRLSDE SSDLRSLRSD SNVSLASNLD
Pennycress IQRQQLSRQQ TESASPRRKP GTKTRGLQQS EDRLSDESSD LRSLRSDSNV SLASNVDMDV
Saltwater_Cress PTSPKPEPNK IQRQQLSRQQ TESASPRRKP GIKSRVQQQS EDHFSDETSD LRSLRSDSNI
White_Mustard TSRSMQQSVD RLSDESSDLR SLRSDSNVSS ASNLDSEATS RYRYERNSDF TEQHTPKQRS
Arenosa PNKIQRQLSR QQTESASPRR KPGIKSRGLQ QSEDRLSDES SDLRSLRSDS NVSLASNLDT
Halleri PNKIQRQLSR QQTESASPRR KPGIKSRGLQ QSEDRLSEES SDLRSLRSDS NVSLASNLDT
Arabis KSRGLQQSED RLSDESSDLR SLRSDSNVSL ASNLDIEVTS RYRSERNSDI TEQYTPKQRS
Grape_Vine LELDKQSRST TPSPESSRVR RQSSRQLTEP SSPARKLRQR APNLLQSDDQ LSEISGDSRN
Bittercress KIQRQQLSKQ QTESASPRRK QGIKSRALQQ SEDRLSDESS DLRSLRSDSN VSLTSNLDIE
601
Thaliana SRYNYERNSD ITEQHTPKQR SPDLGMRSLS KPLKVTVEQP SPVSVLDVAF DEDDSPSPVR
Brassica_rapa RSLPKPLKIT VEQPSPVSIL DVAFDDDESP SPVRKISIVF KDDDHIRSEE SLWMKKHNNL
Shortpod_mustard YKCDLTEQHT PKQRSPELGM RSLPKPLKIT VEQPSPVSIL DVTFDDDESP SPVRKINIVF
False_Flax VTSRYNFERN SDIIEQHTPK QKSPELGMRS LPKPLKVTVE QPSPVSVLDV AFDEDDSPSP
Radish PKPLKVTVEQ PSPVSVLDVA FDEDESPSPV RKISVVFKDD DHLRSEESEW MNKHRTLRRS
Suecica TEVTSRYNYE RNSDITEQHT PKQRSPDLGM RSLPKPLKVT VEQPSPVSVL DVAFDEDESP
Pennycress TSRYKYERNS DIAEQHTPKQ RSPDLGMRSL PKPLKVTVEQ PSPVSVLDVA FDEDESPSPV
Saltwater_Cress SLTSNLDIEG TSRYKYERNN DITEQHTPKQ RSPELGMRSL PKPLKITVEQ PSPVSVLDVA
White_Mustard PEFGMRSLPK PLKVTVEQPS PVSVLDVAFD EDESPSPVRK ISVVFKDDDH LRSEESQWMN
Arenosa EVTSRYNYER NSDITEQHTP KQRSPDLGMR SLPKPLKVTV EQPSPVSVLD VAFDEDESPS
Halleri EVTSRYNYER NSDITEQHTP KQRSPDLGMR SLPKPLKVTV EQPSPVSVLD VAFDEDESPS
Arabis PDLGMKLLPK PLKVTVEQPS PVSVLDVAFD EDESPSPVRK ISIVFKDDNL SSSEEPQWMN
Grape_Vine LSYQVTSIDR SGGINSISFQ HGGQKHKNGD GTMTKFATAT QEQPSPVSVL DAAFYKDDLP
Bittercress VTSRYKYERN SDIMEQHTPK QRSELGMRSL PKPLKLTVEQ PSPVSVLDVA FDEDESPSPV
661
Thaliana KISIVFKEDD NLSSEESHWM NKNNNLCRSI VWPESNTSLK QPDAELTEGF MEDDAEFKNG
Brassica_rapa CRSIVWPESN TSLNQPDAVL TESFMEEGAD LRNGDRKYIS EILSASGLLK DIDYSMLSIQ
Shortpod_mustard KDDDHLRSEE SQWMSKHNNL CRSIVWPESN TSLNQPDAEL TESLMEEGAE LRNGDHKYIS
False_Flax VRKISIFKED DNLSSEESQW MNKHSSLCRS IVWPESNASL KQPDAELIEG FMEEDAEFKN
Radish IVWPETESNG SLNQSDAELT EGIREEGAEL NNGDHKYISE ILLASGLLRD IDYSMLSIQL
Suecica SPVRKISIVF KEDANLSYEE SQWMNKHNNL CRSIVWPESN ATLQPDAELM EGFMEEEAEF
Pennycress RKISIAFKED HLSYEESQWM NKHSNLCRSI VWPESNASPK QPDDELIEGS MEEDAKFKNG
Saltwater_Cress FDEDESPSPV RKISIVFKED DNLGSEESQW MNKQRNICRS IVWPENNASK QPDAEPIMGF
White_Mustard KHRTLRRSIV WPESNGSLNQ SDAELNEGIM EEGAELNNGD HKYISEILLA SGLLRDIDYS
Arenosa PVRKISIVFK EDANLSYEES QWMNKHSNLC RSIVWPESNA TLKQPDAELM EGFMEEEAEF
Halleri PVRKISIVFK EDANLSSEES QWMNKHSNLC RSIVWPESNA TLKQPDAEVM DGFMEEDAEF
Arabis KHSNLCRSIV WPESNVSLKQ PDHNEGYMEE GVELKNGDHK YISEILLASG LLRDIDYSMM
Grape_Vine SPVKKISNAF KDDETLNYDE MEWATKLENI ENLVQRIREL NSTHNEFSVD LIASLCDKTN
Bittercress RKISIVFKEE DNLSSEESQW MNKQRNLCRS IVWPESNMSV KLPDTELIEG FMEEGAELKN
721
Thaliana DHKYISEIML ASGLLRDIDY SMISIQLHQA HLPINPSLFF VLEQNKTSNV SLQDNKHKGR
Brassica_rapa LHQAHLPINP SLFFVLEQNK TSNVTHRGRG FGQQTANLIG RSRRKLVFDT VNEILARKFA
Shortpod_mustard EILSASGLLK DIDYNMLSIQ LHQAHLPINP SLFFVLEQSK TGNVRHDNKH RGRGFGQQQA
False_Flax GDHKYAAEIL SAAGLLRDID YSMISIQLHQ AHLPINPNLF FVLEQNKTSN VSLQDNKHKG
Radish HQAHLPINPS LFFVLEQSKT SNVTQHDSKH RGIGFGQQQT ANLIERSRRK LVFDTINEIL
Suecica KNGDHKYISE ILLASGLLRD IDYNMISIQL HQAHLPINPS LFFVLEQSKT SNVSPQDNKH
Pennycress DHKYISEILS ASGLLRDIDY SMMSIQLHQA HLPINPTLFF VLEQSKTSNV SQHDNKHKGR
Saltwater_Cress MDEGVEFKNG DDKYISEILL ASGLLRDIDY SMISIQLHQA HLPINPSLFF VLEQNKTSNL
White_Mustard MLSIQLYQAH LPINPSLFFV LEQNKTSNAT QHDKKHKGIG FGQQTVNLIE RSRRKLVFDT
Arenosa KNGDHKYISE ILLASGLLRD IDYNMISIQL HQAHLPINPS LFFVLEQSKT SNVSPQDNKH
Halleri KNGDHKYVSE ILLASGFLLR DIDYSMISIQ LHQAHLPINP SLFFVLEQNK TSNVSPQDNK
Arabis SIQLHQAHLP INPSLFFVLE QNKTSNVTQQ DNKHNSRGFG QQQTTNLIDR IRRKLVFDTI
Grape_Vine PDHRYISEIL LASGLLRDCS GLMITKLHQS SHPINPKLFL VLEQNRDVAN ILNDKYSSQN
Bittercress GDHEYISEIL LTSGLLRDID YSMISIQLHQ AHLPINPSLF FVLEQNKTSN VSLQDNKHKG
781
Thaliana GFGQQQTVNL VERSKRKLIF DTINEILAHR FAAEGCTKQP SITLSISTQR THEKSSRGEE
Brassica_rapa AEGCTKQPYI TSSISPLMKT DKSSRGKELL EALCSEIDRL QDNSNCILDE DDEDLIWEDL
Shortpod_mustard ANLIERSRRK LVFDNVNEIL ARKFAAEGCT KQPYITSSIS QLMTTEKSSR GKELLESLCS
False_Flax RGFGQQQTAN LIERSKRKLV FDTINEILAR RFAADGCTKQ PSIISSISPL RITEKNSRGK
Radish ARKFTAEGCT KQQPYITTSI SQLRTTNKSS KGKELLKTMC SEIDRLQDNS KCILDEDDDE
Suecica KGRGFGQQQT VNLIERSKRK LVFDTINEIL AHRFAAEGCT KQPSIISSIS TQRIPVKSSR
Pennycress GFGQQQTANL MERSRRKLVF DTINEILARK FAAEGCTKQP SIVSSISRLR TTDKSSRGEE
Saltwater_Cress SLQDDKHKAR GFGQQQTANL IDRSRRKLVF DTINEILARR FAAEGCTKQP SMISSIRPLR
White_Mustard MNEILARKFA AEGCTKQPYI TPSISQLRTT SKSSKGKELL ETMCSEIDRL QDNSKCILDE
Arenosa KGRGFGQQQT VNLIERSKRK LVFDTINEIL AHRFAAEGCT KQPSIISSIS TQRIPVKSSR
Halleri HKGRGFGQQQ TVNLIERSKR KLVFDTINEI LAHRFAAEGC TKRPSIISSI STQRIPVKSS
Arabis NEILARRFAA EGCTKQTSII SSISPLRTTD KSSKGKELLQ ILCSEIDRLQ DDSNCILDED
Grape_Vine TAQSKLQRKL IFDVVNEILF QKLAFTGSSE PCFLPNKIVR RSQNGQELLR ELCSEIDQLQ
Bittercress RGFGMANLIE RSRRKLVFDT INEILARRFA AEGCTKQPSI ISSISPLRTT EKSSRGKELL

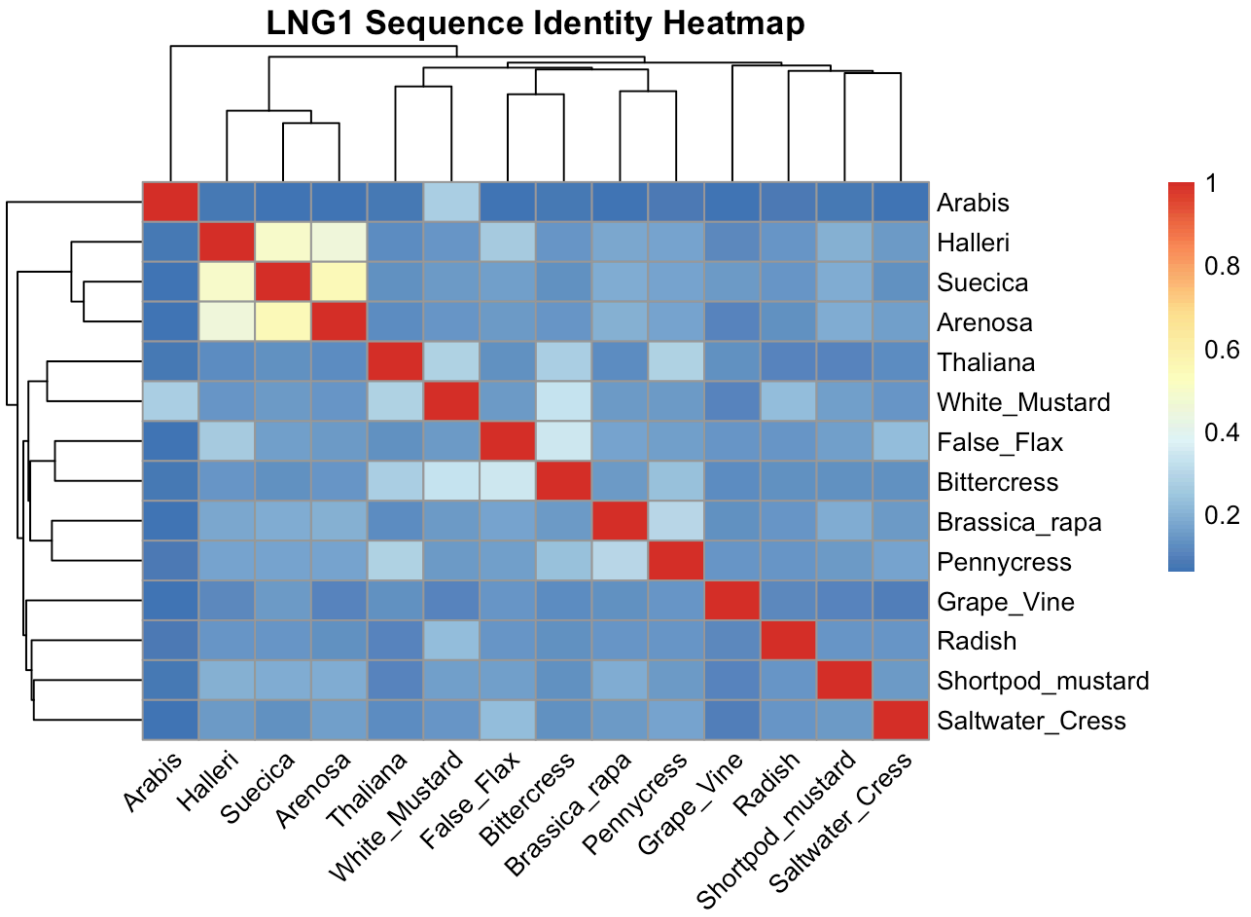


[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.



[Q7] Generate a sequence identity based **heatmap** of your aligned sequences using R.

If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions in the **Bio3D package**. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.



[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.

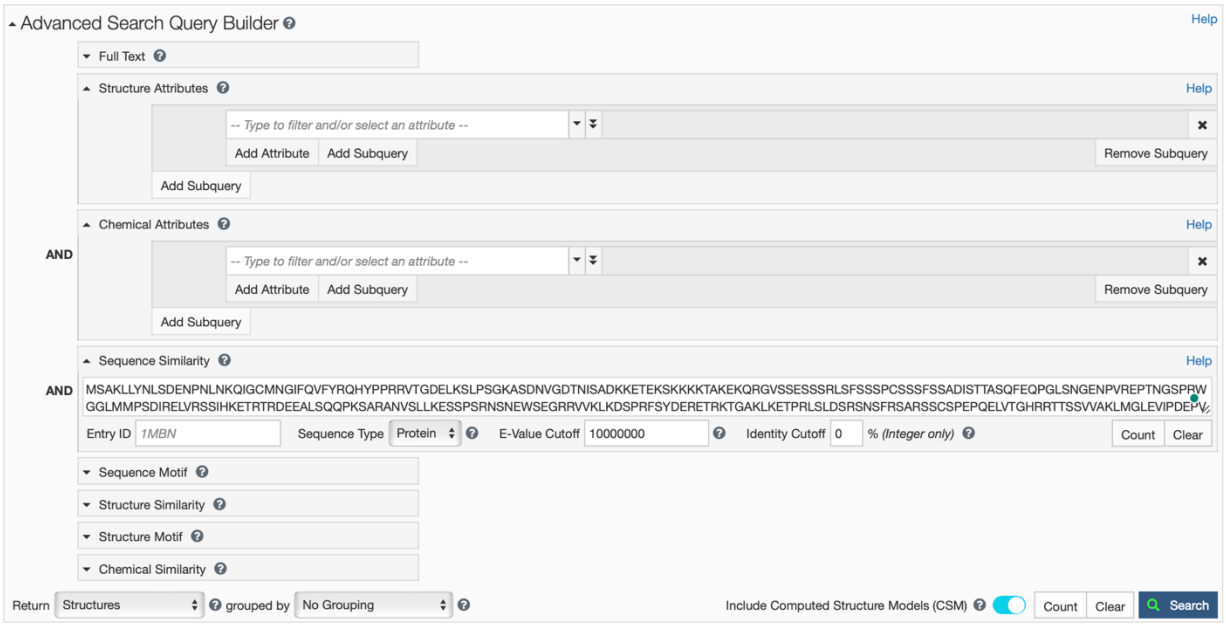
List the top 3 *unique* hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above.

Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case

you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

I went to PDB, used one sequence (A. thaliana) and expanded the e-value. This is what I received:



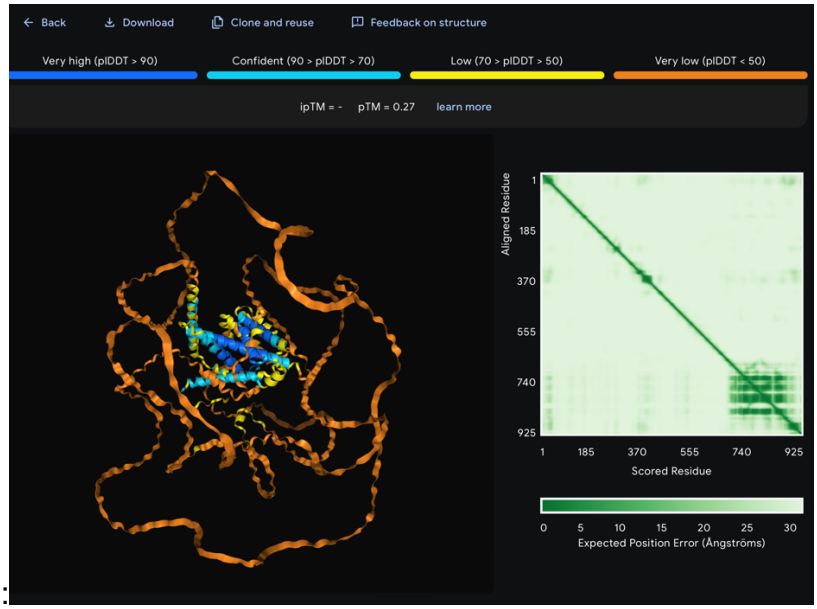
PDB Identifier	Organism	Macromolecule	Technique	Resolution	E-value	Sequence Identity (%)
AF-K7MW93F1	Glycine max	DUF4378 domain-containing protein	AlphaFold Predicted Structure	N/A	48.5	80.2
AF-Q7XIU5F1	Oryza sativa Japonica Group	Os07g0109400 protein	AlphaFold Predicted Structure	N/A	48.48	78.1
AF-A0A1D6HZT8F1	Zea mays	Protein LONGIFOLIA 2	AlphaFold Predicted Structure	N/A	47.93	77.6

[Q9] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

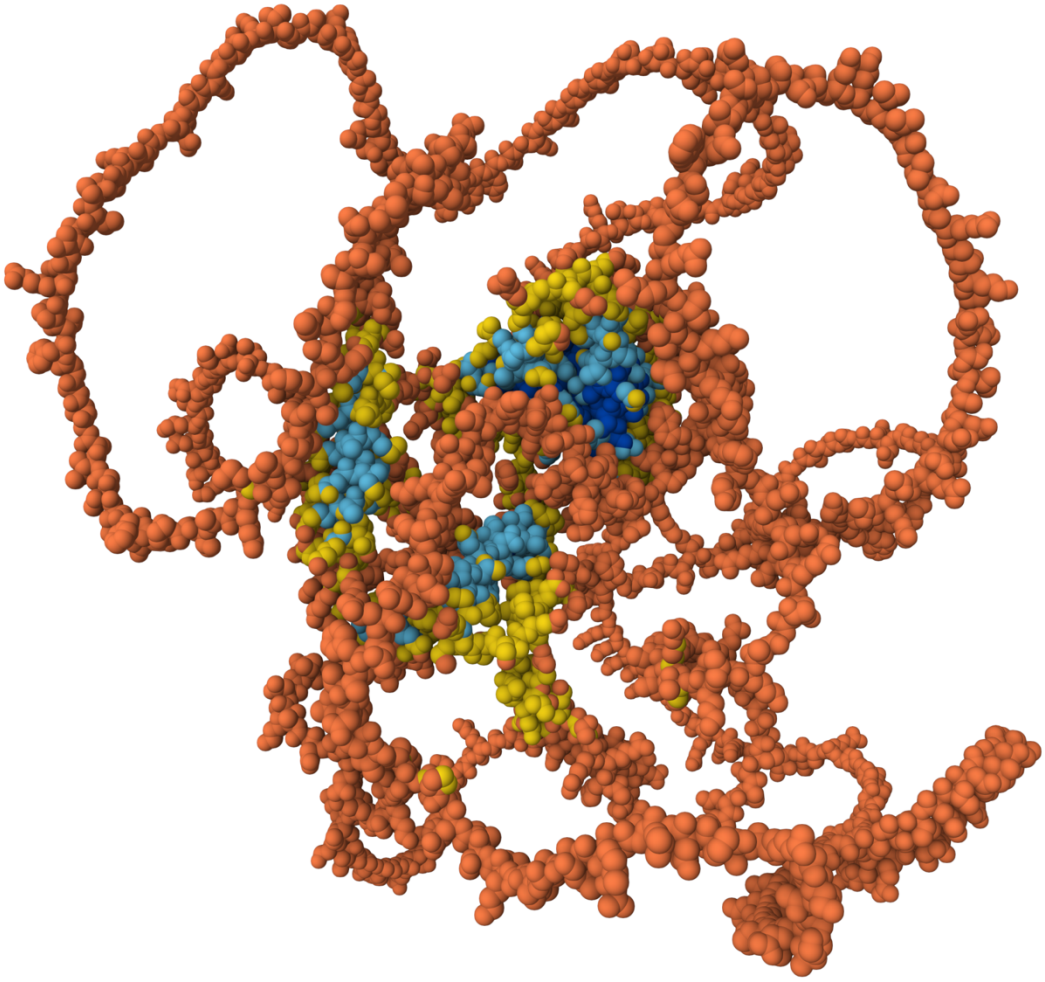
> Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a “too many amino acids” (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.

Once complete save the resulting PDB format file for your records. Finally, generate a molecular figure of your generated PDB structure using the **Mol* viewer** online (or VMD/PyMol/Chimera if you prefer). To complete your analysis you can optionally highlight *conserved residues* that are likely to be functional as **spacefill** and the protein as **cartoon** colored by local alpha fold *pLDDT quality score*. This score is contained in the B-factor column of your PDB downloaded file. Please use a white or transparent background for your figure (i.e. not the default black in PyMol/VMD/Chimera etc.).

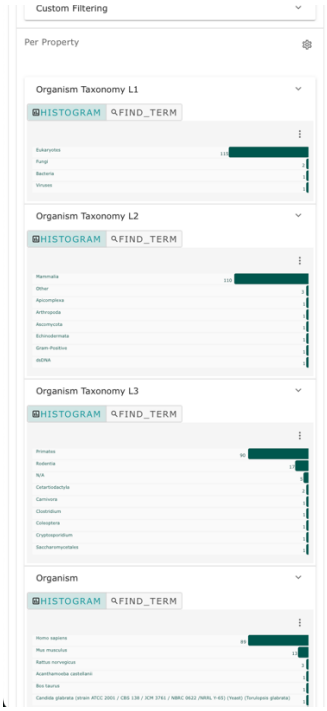
AlphaFold Server Results:



Molviewer Image (UltraHD Downloaded Figure) with space fill and pLDDT validation coloring.



[Q10] Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that may be useful starting points for exploring potential inhibition of your novel protein? If there are no assays listed here simply list “non available as of [date]”.



20 Showing 1-20 out of 119 records

#	E-Value	Positives %	Identities %	Score (bits)	Score	Length	ChEMBL ID	Name	UniProt Accessions	Type	Organism	Compounds	Activities
1.	0.00057	35.2	23.3	43.1	100	1007	CHEMBL1908382	Serine/threonine-protein kinase PRP4 homolog	Q13523	SINGLE PROTEIN	Homo sapiens	251	270
2.	0.0025	35.9	20.3	40.8	94	568	CHEMBL4295761	Protein AF-9	P42568	SINGLE PROTEIN	Homo sapiens	49	111
3.	0.004	35.3	18.4	40.4	93	1334	CHEMBL4879481	WASH complex subunit 2	Q6PGL7	SINGLE PROTEIN	Mus musculus	2	2
4.	0.011	36	19.3	38.9	89	754	CHEMBL3707467	Peptidyl-prolyl cis-trans isomerase G	Q13427	SINGLE PROTEIN	Homo sapiens	10	10
5.	0.013	39.2	20.5	38.5	88	602	CHEMBL4296025	Cytoskeleton-associated protein 4	Q07065	SINGLE PROTEIN	Homo sapiens	5	7
6.	0.014	40.8	28.2	38.5	88	838	CHEMBL5291560	Transforming acidic coiled-coil-containing protein 3	Q9Y6A5	SINGLE PROTEIN	Homo sapiens	1	2
7.	0.021	36	19.1	38.1	87	2461	CHEMBL3217383	Microtubule-associated protein 1B	P15205	SINGLE PROTEIN	Rattus norvegicus	1	1
8.	0.027	37.8	25.6	37.7	86	1309	CHEMBL2417352	Microtubule-associated serine/threonine-protein kinase 3	Q60307	SINGLE PROTEIN	Homo sapiens	30	49
9.	0.039	40.7	22.1	37	84	565	CHEMBL2366583	Ecdysone receptor	Q4W6D0, Q4W6C8	PROTEIN COMPLEX	Leptinotarsa decemlineata	3	3
10.	0.045	34.1	17.1	37	84	1220	CHEMBL4105852	Eukaryotic translation initiation factor 5B	O60841	SINGLE PROTEIN	Homo sapiens	170	170
11.	0.061	37.6	22.7	36.6	83	1616	CHEMBL1993	DNA (cytosine-5)-methyltransferase 1	P26358	SINGLE PROTEIN	Homo sapiens	767	978
12.	0.082	38	23.3	36.2	82	2564	CHEMBL3108647	Histone-lysine N-methyltransferase SETD2	Q9BYW2	SINGLE PROTEIN	Homo sapiens	90	119
13.	0.1	42.2	24.4	35.8	81	1096	CHEMBL4295672	Lysine-specific demethylase PHF2	O75151	SINGLE PROTEIN	Homo sapiens	16	16
14.	0.11	37.8	23.6	35.8	81	2446	CHEMBL4523214	Transcription factor HIVEP2	P31629	SINGLE PROTEIN	Homo sapiens	1	1
15.	0.11	43.8	26.5	35.8	81	3433	CHEMBL4523230	Utrophin	P46939	SINGLE PROTEIN	Homo sapiens	81	89
16.	0.15	40.9	22.1	35	79	589	CHEMBL2176776	Sentrin-specific protease 2	Q9HC62	SINGLE PROTEIN	Homo sapiens	74	79
17.	0.18	43.7	20.9	35	79	2320	CHEMBL3351192	Bromodomain and WD repeat-containing	Q9NSI6	SINGLE PROTEIN	Homo sapiens	52	55

No target-associated assays or ligand efficiency data for plant-specific proteins are available as of 11/30/24.

Scoring Rubric: [50 total points available]

Q1 (4 points)

Protein name	1
Species	1
Accession number	1
Function known	1

Q2 (6 points)

Blast method	1
Database searched	1
Limits applied	1
Search output list (top hits)	1
Alignment of choice	1

Evalue and other alignment stats	1
Q3 (3 points)	
Protein sequence of choice matches Subject above	1
Name in header	1
Species	1
Q4 (3 point)	
Blastp output list with identities & Evalue	1
Top alignment shown with alignment statistics	1
Results indicates a “novel” gene found	1
Q5 (3 points)	
MSA labeled with useful names	1
MSA trimmed appropriately (i.e. no gap overhangs)	1
Pasted MSA fits report page width (i.e. font, format)	1
Q6 (1 point)	
Figure illustrates sequence clustering pattern	1
Q7 (10 points)	
Heatmap figure included in report	5
Heatmap is legible (i.e. no labels obscured)	5
Q8 (9 points)	
PDB identifiers from multiple species reported	5
Annotation of PDB source, resolution and technique	4
Annotation of Evalue and Sequence Identity	1
Q9 (10 points)	
Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1
Conserved residues as spacefill	3
Protein cartoon colored by pLDDT quality score	3
Q10 (1 point)	