**BGGN-213: FOUNDATIONS OF BIOINFORMATICS**

The find-a-gene project assignment [http://thegrantlab.org/bggn213](http://thegrantlab.org/bggn213/)

Dr. Barry Grant

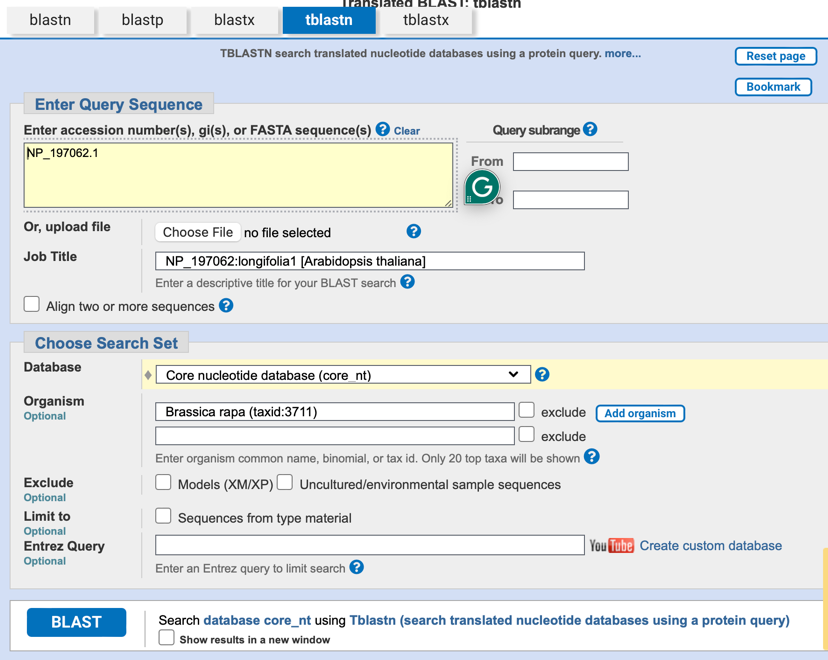
Juan Gonzalez; [jhg001@ucsd.edu](mailto:jhg001@ucsd.edu); PID: A69036681

**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: ATGTCGGCGAAGCTTTTGTATAACTTGTCGGATGAGAATCCAAATCTGAATAAACAGATTGGATGTATGAATGGGATCTTTCAGGTGTTTTACCGGCAACATTATCCACCGAGACGTGTCACCGGAGATGAGCTCAAGTCTCTTCCTTCAGGCAAAGCAAGTGACAATGTCGGTGATACCAACATTTCAGCGGACAAGAAGGAAACGGAGAAGAGTAAGAAGAAGAAGACTGCAAAGGAGAAACAGAGGGGAGTATCCTCTGAATCGTCCTCGAGGTTGTCGTTTTCTTCATCACCATGCTCCTCGAGCTTCTCATCTGCAGATATTAGCACCACGGCTTCTCAGTTTGAACAGCCCGGTTTGAGTAATGGTGAGAATCCTGTAAGAGAACCGACCAATGGGTCGCCAAGGTGGGGCGGTTTAATGATGCCAAGTGATATAAGGGAGCTTGTGAGAAGCTCTATTCATAAGGAGACAAGAACCAGAGATGAAGAAGCCTTGTCTCAGCAGCCTAAATCAGCCAGAGCTAATGTGTCTCTTCTCAAAGAATCATCACCATCTCGGAATTCTAATGAATGGAGTGAGGGACGGAGAGTAGTGAAGCTGAAAGACAGTCCTCGGTTCTCTTACGATGAGAGGGAGACGAGAAAGACAGGGGCCAAGTTGAAAGAGACCCCGAGGTTGTCATTAGACAGTAGATCGAATTCCTTTAGGAGCGCAAGATCTAGTTGTTCACCAGAGCCACAAGAGCTTGTAACAGGTCACAGAAGAACAACATCAAGCGTCGTTGCAAAACTGATGGGTCTTGAAGTAATTCCAGATGAGCCTGTGACCATTCAGAATAGAGAAAATCGCTTCTGCGACTCCCCGAGGCCAACTTCCCGAGTGGAAGTAGATCTACAAAGATCAAGAGGTTTTGATTCAATCAAAAAGATGATGCCTGCTAAGTTTCCTATGAAAGCATCTCCATGGGCGCAAGTTGATGGTGCCAAGAACCAAGTCAAAATACCAGACGCTACTACGCTGACAGTTTATGGTGAGATACAGAAGCGGCTTTCACAGCTTGAGTTCAAAAAGTCCGAGAAAGACCTCAGGGCTCTTAAGCAAATACTCGAAGCAATGGAGAAGACGCAGCAGTTGATAAGCAAAGATGATGATGACAACAAAACTCTATGTTCAAGCAATTTTATGCAGAGAAATAATCAGCCAATTCCATCTGCAATAAACACCTCGTCCATGAATTTTAAATCATCCTCTATCGTGGTTATGAAAGCAGCTACCGCTCCAGTCTTCAAAGACACAGGCATTGCAGGTTCTGCGTCTTTCTCCCCGCGGAATGTTGCTTTACCAAATGTCAAGGTTGGAAACCTGAGGCAAGCCCAGAAAGTCATTCCGAGGAAGCAGAGTGCTATGGATGTGACCCCGAGGCCAGGATATTACAAGGGCCAGACAGAATCCACAATGAAAAACACTAGTACCAGACCATTACAATCGAAAAGCGACATGGCCAAGTCAGGGAAGATCCAGAAGCCTAGTGTCAGCCTAAGAACACCACCAAAGAAGCTTGGGTTCGAGAAGCAGTCTCGGCCAACATCCCCAAAACCAGAACTGAACAAGAACCAGAGACAACAACTCAGTAGGCAACAGACAGAGTCAGCCTCCCCGAGAAGAAAGCCAGGGATAAAATCTCGTGGCCTACAGCAATCTGAAGACCGTTTAAGTGATGAAAGCAGTGACTTGAGAAGTCTAAGATCTGACAGCAACGTAAGCTTGGCCTCTAACCTTGATACTGAGGTTACAAGCAGGTATAACTATGAGAGGAACAGCGACATTACGGAGCAGCACACCCCGAAACAAAGGAGTCCAGACTTGGGAATGAGGTCGCTGTCAAAACCTCTGAAAGTTACAGTGGAGCAGCCCAGCCCGGTGTCAGTTCTTGATGTAGCTTTCGACGAAGATGATTCACCATCCCCTGTGCGGAAGATATCCATTGTCTTTAAAGAGGACGACAATCTAAGTTCTGAAGAGTCCCATTGGATGAACAAGAACAACAACTTATGTAGATCGATTGTGTGGCCTGAGAGTAACACGAGTCTAAAGCAACCTGATGCTGAACTTACGGAGGGTTTCATGGAAGACGATGCCGAATTCAAAAATGGTGACCACAAGTACATCTCAGAGATAATGTTGGCATCAGGGCTTCTACGAGATATCGACTACAGCATGATAAGCATCCAGCTGCACCAAGCACACCTACCGATCAACCCGAGCCTTTTCTTTGTACTGGAACAGAACAAGACAAGCAATGTGAGTCTACAGGACAACAAGCACAAAGGCAGAGGATTTGGACAACAACAAACGGTGAACCTGGTTGAGAGAAGTAAGAGGAAGCTCATATTTGACACCATCAACGAGATCTTAGCTCACAGATTCGCTGCAGAAGGGTGTACAAAGCAACCATCCATAACATTATCAATCAGCACGCAAAGGACACACGAAAAAAGTTCAAGAGGGGAAGAGCTTCTGCAAACTCTGTGTTCAGAGATTGATCGATTACAAGATAACTCAAAGTGTATCTTGGATGAGGACGATGAAGACCTCATTTGGGAGGATCTGCAAAGCCATGGCATGAACTGGAAGGAGATTGAAGGAGAGACACCAGGGTTAGTGTTAGACATTGAGAGGCTAATCTTCAAAGACTTGATTGGTGAAGTTGTGACAAGCGAGTTTGCAGCTTTTCCAAGGATGCTCAGTGGGCAACCAAGGCAGCTTTTTCATTGCTAA**

**LNG1 *Arabidopsis thaliana* protein sequence: MSAKLLYNLSDENPNLNKQIGCMNGIFQVFYRQHYPPRRVTGDELKSLPSGKASDNVGDTNISADKKETEKSKKKKTAKEKQRGVSSESSSRLSFSSSPCSSSFSSADISTTASQFEQPGLSNGENPVREPTNGSPRWGGLMMPSDIRELVRSSIHKETRTRDEEALSQQPKSARANVSLLKESSPSRNSNEWSEGRRVVKLKDSPRFSYDERETRKTGAKLKETPRLSLDSRSNSFRSARSSCSPEPQELVTGHRRTTSSVVAKLMGLEVIPDEPVTIQNRENRFCDSPRPTSRVEVDLQRSRGFDSIKKMMPAKFPMKASPWAQVDGAKNQVKIPDATTLTVYGEIQKRLSQLEFKKSEKDLRALKQILEAMEKTQQLISKDDDDNKTLCSSNFMQRNNQPIPSAINTSSMNFKSSSIVVMKAATAPVFKDTGIAGSASFSPRNVALPNVKVGNLRQAQKVIPRKQSAMDVTPRPGYYKGQTESTMKNTSTRPLQSKSDMAKSGKIQKPSVSLRTPPKKLGFEKQSRPTSPKPELNKNQRQQLSRQQTESASPRRKPGIKSRGLQQSEDRLSDESSDLRSLRSDSNVSLASNLDTEVTSRYNYERNSDITEQHTPKQRSPDLGMRSLSKPLKVTVEQPSPVSVLDVAFDEDDSPSPVRKISIVFKEDDNLSSEESHWMNKNNNLCRSIVWPESNTSLKQPDAELTEGFMEDDAEFKNGDHKYISEIMLASGLLRDIDYSMISIQLHQAHLPINPSLFFVLEQNKTSNVSLQDNKHKGRGFGQQQTVNLVERSKRKLIFDTINEILAHRFAAEGCTKQPSITLSISTQRTHEKSSRGEELLQTLCSEIDRLQDNSKCILDEDDEDLIWEDLQSHGMNWKEIEGETPGLVLDIERLIFKDLIGEVVTSEFAAFPRMLSGQPRQLFHC**

**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:**

**A screenshot of a computer

Description automatically generated**

[**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) MSAKLLYNLSDENPNLNKQFGCMNGIFQVFYRQHYPARRVSVAGDELKSLPSGKTSDNVGVTNGSTDKKETEKSKKKKAAKEKQKVVSSESSSRLSFSSSPCSSSFSSADISTTTSQFEQPMSNGETPAREPTYGSPRWGGLVMSSDLRELVRSSIHKETRTRVEEEALSQQPKSARANVSLLKELSPSRSSNEWSEGRRVVKLKDSPRFSYDEREARKTGAKFKETPRLSLDSRSNSFRSAKSSCSPEPQELVTGHRRTTSSVIAKLMGLDVVSDEPVTDQSRENHFCDSPRPAPRVEADLPRSRGSDSFKKMMPAAKFPAKTAPWTQADGARNQVKAADAAATLTVYGEIQKRLSQLEFKKSEKDLRALQQILEAMEKTQQLMSKDDDNSSLSSTNFMQPSPSSKSIRSSSIVVMKAASAPVFKETGSSSSTSSSPRSVALPNVKVSNQKGTTRKQSAMDVTPRPATKNTSTRPLQSKIEMAKSGKPSVSPRTQPKKLGFEKQSRPTSPKPEPNKNQRQQLSRQQTESPSPRRKPGMKSRGLQQSEDRSSDESSDLRSLRSDSNVSSASNFDIEVTSRHKCDLTEQHTPKQRSPELGMRSLPKPLKITVEQPSPVSILDVAFDDDESPSPVRKISIVFKDDDHIRSEESLWMKKHNNLCRSIVWPESNTSLNQPDAVLTESFMEEGADLRNGDRKYISEILSASGLLKDIDYSMLSIQLHQAHLPINPSLFFVLEQNKTSNVTHRGRGFGQQTANLIGRSRRKLVFDTVNEILARKFAAEGCTKQPYITSSISPLMKTDKSSRGKELLEALCSEIDRLQDNSNCILDEDDEDLIWEDLQSQGMNWKEIEGETPGLVLDIERLIFKDLISEVVTSEVAAFPGNKLSGQPRQLFHC

**Name: protein LONGIFOLIA 1  
Species:** [**Brassica**](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=81985) **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%A screenshot of a computer code

Description automatically generatedTo 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**

**A screenshot of a computer

Description automatically generated**

[**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.

A screenshot of a computer screen

Description automatically generated

A screenshot of a chart

Description automatically generated

A screenshot of a chart

Description automatically generated

A group of colorful letters

Description automatically generated

[**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.

A chart of different types of food

Description automatically generated

[**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.

A diagram of a heatmap

Description automatically generated

[**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: A screenshot of a computer

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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](https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb) 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](http://pfam.xfam.org/) 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:A screenshot of a computer

Description automatically generated

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

A close-up of a dna molecule

Description automatically generated

[**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]”.

A screenshot of a computer

Description automatically generated

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)

Evidence of ChEMBEL searches 1