**Find a Gene Project**

**Part 1**

Student: Yushi Li

Email: yul012@ucsd.edu

PID: A15639705

**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: CLOCK

Accession: AAB83969

Species: Homo sapiens

Functions: Involved in circadian regulation of gene expression through a variety of cellular processes such as protein acetylation and DNA damage checkpoint signaling.

[**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).

Method: tBLASTn against frog TSA.

Database: Transcriptome Shotgun Assembly (TSA)

Organism: Frogs (taxid:8342)

Also 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 or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called Screen Shot [].png in your Desktop directory). It is **not** necessary to print out all of the blast results if there are many pages.

A screenshot of a computer

Description automatically generated

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. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

Chosen match: Accession GIVH01185592.1, a 2514bp segment of a piece of transcribed RNA from *Xenopus tropicalis* obtained through transcriptome shotgun assembly (TSA).

A screenshot of a computer

Description automatically generated See below for alignment:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Alignment details:

>TSA: Xenopus tropicalis TRINITY\_DN3887\_c0\_g1\_i2, transcribed RNA sequence

Sequence ID: GIVH01185592.1 Length: 5403

Range 1: 309 to 2822

Score:1151 bits(2978), Expect:0.0,

Method: Compositional matrix adjust.,

Identities:662/852(78%), Positives:725/852(85%), Gaps:20/852(2%)

Query 1 MLFTVSCSKMSSIVDRDDSSIFDGLVEEDDKDKAKRVSRNKSEKKRRDQFNVLIKELGSM 60

+L ++S KMSS DRDD SIFDGLVEEDDKDKAKRVSRNKSEK+RRDQFN+LIKELGSM

Sbjct 309 LLLSISIHKMSSTADRDDGSIFDGLVEEDDKDKAKRVSRNKSEKRRRDQFNILIKELGSM 488

Query 61 LPGNARKMDKSTVLQKSIDFLRKHKEITAQSDASEIRQDWKPTFLSNEEFTQLMLEALDG 120

LPGNAR+MDKSTVLQKSIDFLRKHKEI+AQSDASEIRQDWKPTFLSNEEFTQLMLEALDG

Sbjct 489 LPGNARRMDKSTVLQKSIDFLRKHKEISAQSDASEIRQDWKPTFLSNEEFTQLMLEALDG 668

Query 121 FFLAIMTDGSIIYVSESVTSLLEHLPSDLVDQSIFNFIPEGEHSEVYKILSTHLLESDSL 180

FFLA+MTDG+IIYVSESVTSLLEHLPSDLVDQSIFNF+PEGEHSEVYKILST +LES SL

Sbjct 669 FFLAVMTDGNIIYVSESVTSLLEHLPSDLVDQSIFNFVPEGEHSEVYKILSTRMLESGSL 848

Query 181 TPEYLKSKNQLEFCCHMLRGTIDPKEPSTYEYVKFIGNFKSLNSVSSSAHNGFEGTIQRT 240

+ EYLK+KN+LEFCCHMLRGT DPKEPSTYE+VKFIGNFKSLN+V SS HNGF+G +QR+

Sbjct 849 SSEYLKTKNELEFCCHMLRGTADPKEPSTYEFVKFIGNFKSLNNVPSSTHNGFDGALQRS 1028

Query 241 HRPSYEDRVCFVATVRLATPQFIKEMCTVEEPNEEFTSRHSLEWKFLFLDHRAPPIIGYL 300

RP YE+RVCFVATVRLATPQFIKEMCTVEE EEFTSRHSLEWKFLFLDHRAPPIIGYL

Sbjct 1029 LRPPYEERVCFVATVRLATPQFIKEMCTVEESTEEFTSRHSLEWKFLFLDHRAPPIIGYL 1208

Query 301 PFEVLGTSGYDYYHVDDLENLAKCHEHLMQYGKGKSCYYRFLTKGQQWIWLQTHYYITYH 360

PFEVLGTSGYDYYHVDDLENLAKCHEHLMQYGKGKSCYYRFLTKGQQWIWLQT YYITYH

Sbjct 1209 PFEVLGTSGYDYYHVDDLENLAKCHEHLMQYGKGKSCYYRFLTKGQQWIWLQTRYYITYH 1388

Query 361 QWNSRPEFIVCTHTVVSYAEVRAERRRELGIEESLPETAADKSQDSGSDNRINTVSLKEA 420

QWNSRPEFIVCTHTVVSYAEV AERRRE GIE+S P ADK+QDS SDN +NTVSLKEA

Sbjct 1389 QWNSRPEFIVCTHTVVSYAEVGAERRRERGIEDSPPAITADKNQDSVSDNHMNTVSLKEA 1568

Query 421 LERFDhsptpsassrssrkssh-tavsDPSSTPTKIPTDTSTPPRQHLPAHEKMVQRRss 479

LERFD S TPS SS+SS KSS TAVSDPSSTPTKI T+TSTPPRQ + +K RRSS

Sbjct 1569 LERFDDSRTPSPSSKSSIKSSSHTAVSDPSSTPTKITTETSTPPRQAITGLDK---RRSS 1739

Query 480 fssqsinsqsvgssLTQPVMSQATNLPIPQGMSQ--FQFSAQLGAMQHLKDQLEQRTRMI 537

SSQS++SQSV +L+Q VM Q ++ + QGM+Q FQF+AQ GAM+HLKDQLEQRTR+I

Sbjct 1740 VSSQSMSSQSVSQALSQSVMKQTASMQLQQGMAQPMFQFTAQFGAMKHLKDQLEQRTRII 1919

Query 538 EANIHRQQEELRKIQEQLQMVHGQGLQMFLQQSNPGLNFGSVQLSSGNSSNIQQLAPINM 597

E NI RQQEELRKIQEQL MVHGQG+QMFLQQ PGLNFG VQ+SS NS++IQQL + M

Sbjct 1920 EENIQRQQEELRKIQEQLHMVHGQGIQMFLQQPGPGLNFGPVQVSSANSASIQQLPQLTM 2099

Query 598 QGQVVPTNQIQSGMNTGHIGTTQHMIqqqtlqststqsqqNVLSGHsqqtslpsqtqstl 657

QGQVV TNQ+Q+GMNTGH+G TQH++QQQ LQ+TS Q QQN+ GHSQQTSL SQT TL

Sbjct 2100 QGQVVQTNQLQAGMNTGHVG-TQHIMQQQQLQTTSQQGQQNIHGGHSQQTSLSSQTSGTL 2276

Query 658 tAPLYNTMVISQPAAGSMVQIPSSMPQNSTQSAAVTTFTQDRQIRFSQGQQLVTKLVTAP 717

T+PLYNTMVISQP +GSMVQ+PS++ Q S Q A+VTTF QDRQIRFSQ QQ+VTKLVT P

Sbjct 2277 TSPLYNTMVISQPPSGSMVQMPSNI-QQSNQGASVTTFPQDRQIRFSQAQQIVTKLVTTP 2453

Query 718 VACGAVMVPSTMLMGQVVTAYPTFatqqqqsqtlsvtqqqqqqssqeqqltsv---qqps 774

VACG VMVPSTM MG VVTAYPTFATQQQQ Q T Q +Q S+ QQP+

Sbjct 2454 VACGTVMVPSTMFMGPVVTAYPTFATQQQQQQQQPQTLSITQHQPNQQDQQSMPTVQQPA 2633

Query 775 qaqltqppqqflqtsRLLHGNPSTQLILSAAFPLqqstfpqshhqqhqsqqqqqlsrhRT 834

QAQL Q PQQFLQTSRL+HGN STQLILS FP+ QQQLS HRT

Sbjct 2634 QAQLAQQPQQFLQTSRLVHGNQSTQLILS-PFPV--------QQNTFAPSHQQQLSHHRT 2786

Query 835 DSLPDPSKVQPQ 846

DS+ DPSKVQ Q

Sbjct 2787 DSMSDPSKVQQQ 2822

In general, [Q2] 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 [Q4]), and a non-homologous result.

If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[**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.

Chosen sequence:

>GIVH01185592.1:309-2822 TSA: Xenopus tropicalis TRINITY\_DN3887\_c0\_g1\_i2, transcribed RNA sequence

LLLSISIHKMSSTADRDDGSIFDGLVEEDDKDKAKRVSRNKSEKRRRDQFNILIKELGSMLPGNARRMDKSTVLQKSIDFLRKHKEISAQSDASEIRQDWKPTFLSNEEFTQLMLEALDGFFLAVMTDGNIIYVSESVTSLLEHLPSDLVDQSIFNFVPEGEHSEVYKILSTRMLESGSLSSEYLKTKNELEFCCHMLRGTADPKEPSTYEFVKFIGNFKSLNNVPSSTHNGFDGALQRSLRPPYEERVCFVATVRLATPQFIKEMCTVEESTEEFTSRHSLEWKFLFLDHRAPPIIGYLPFEVLGTSGYDYYHVDDLENLAKCHEHLMQYGKGKSCYYRFLTKGQQWIWLQTRYYITYHQWNSRPEFIVCTHTVVSYAEVGAERRRERGIEDSPPAITADKNQDSVSDNHMNTVSLKEALERFDDSRTPSPSSKSSIKSSSHTAVSDPSSTPTKITTETSTPPRQAITGLDKRRSSVSSQSMSSQSVSQALSQSVMKQTASMQLQQGMAQPMFQFTAQFGAMKHLKDQLEQRTRIIEENIQRQQEELRKIQEQLHMVHGQGIQMFLQQPGPGLNFGPVQVSSANSASIQQLPQLTMQGQVVQTNQLQAGMNTGHVGTQHIMQQQQLQTTSQQGQQNIHGGHSQQTSLSSQTSGTLTSPLYNTMVISQPPSGSMVQMPSNIQQSNQGASVTTFPQDRQIRFSQAQQIVTKLVTTPVACGTVMVPSTMFMGPVVTAYPTFATQQQQQQQQPQTLSITQHQPNQQDQQSMPTVQQPAQAQLAQQPQQFLQTSRLVHGNQSTQLILSPFPVQQNTFAPSHQQQLSHHRTDSMSDPSKVQQQ

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.

Name: Kermit’s CLOCK

Species: *Xenopus tropicalis*

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amphibia; Batrachia; Anura; Pipoidea; Pipidae; Xenopodinae; Xenopus; Silurana

[**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 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 the top match reported has 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 likely succeeded in finding a novel gene.
* If there are no database matches to the original query from [Q1], 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.

Details:

BLASTP against NR database yielded a top hit from *X. tropicalis* with less than 100% identity.

Graphical user interface, text, application, email

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence