SLE111 - Bioinformatics Assignment

Gene Identifcation: An introduction to bioinformatics and databases

Assessment

This task is worth 40 marks, and is worth 10% of your overall grade.

It will be due at the estart of week 12, and will be submitted on Moodle.

Background

You are a member of a laboratory. Some unknown samples and data files have been unearthed. Your supervisor has asked you to analyse the sequences, determine which organism/s the data comes from and other information about the sequences. A website you need to familiarize yourself with is at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) and, within that site, you will be using NCBI BLAST: http://blast.ncbi.nlm.nih.gov/.

BLAST stands for Basic Local Alignment Search Tool. It is a program that compares an unknown sequence of nucleotides or amino acids (depending on which type of BLAST that is used) against all the nucleotide and amino acid sequences that have been lodged in genetic databases worldwide. The 'alignment' refers to how the comparisons are made: that is, by aligning your query sequence against all sequences in the database and looking for the best match. A BLAST search is the standard way for biologists to identify sequences of genes or proteins, or their closest relatives.

Unknown Sequences

The sequences are available in a separate document. This is available in the bioinformatics assignment folder on Moodle.

Assignment Instructions

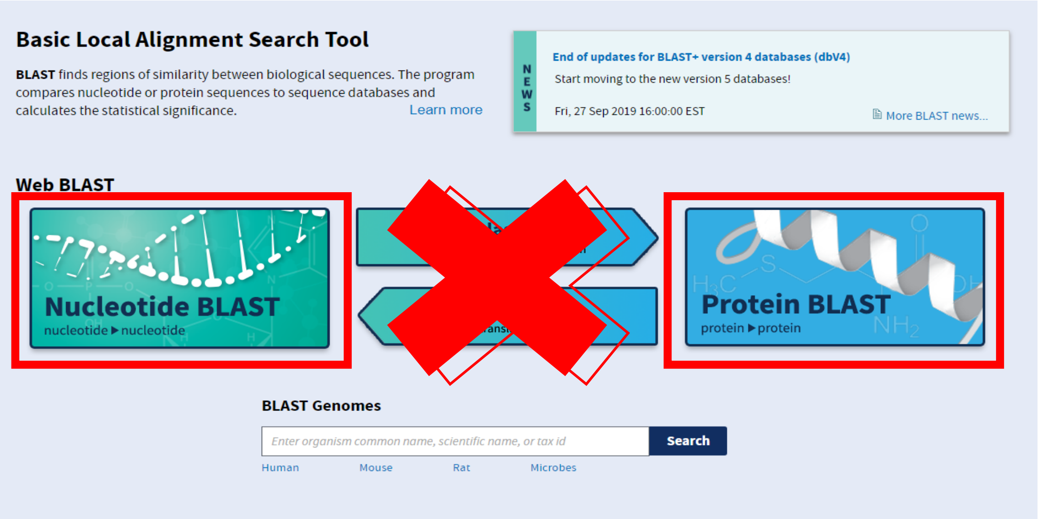
Before attempting this assignment, please view the lecture slides on this assignment. This provides a step by step guide on how to run a BLAST search and how to find the necessary information. There is also a discussion forum available on Moodle if you are still confused. **This discussion forum is monitored and should not be used to swap answers with fellow students. Students that are caught misusing this forum will receive a mark of 0 for this assignment.**

How to run a BLAST search

Please note that the images included are an example that is not the same as your unknown sequences. If you use this information in your own assignment, it will be incorrect.

1. Go to http://blast.ncbi.nlm.nih.gov/.

2. Select the relevant BLAST search.
   You will need to choose out of **Nucleotide or Protein** BLAST.
   (DO NOT use Blastx or Tblastn).



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS
End of updates for BLAST+ version 4 databases (dbV4)
Start moving to the new version 5 databases!
Fri, 27 Sep 2019 16:00:00 EST
More BLAST news...

Web BLAST

BLAST Genomes

Enter organism common name, scientific name, or tax id | Search

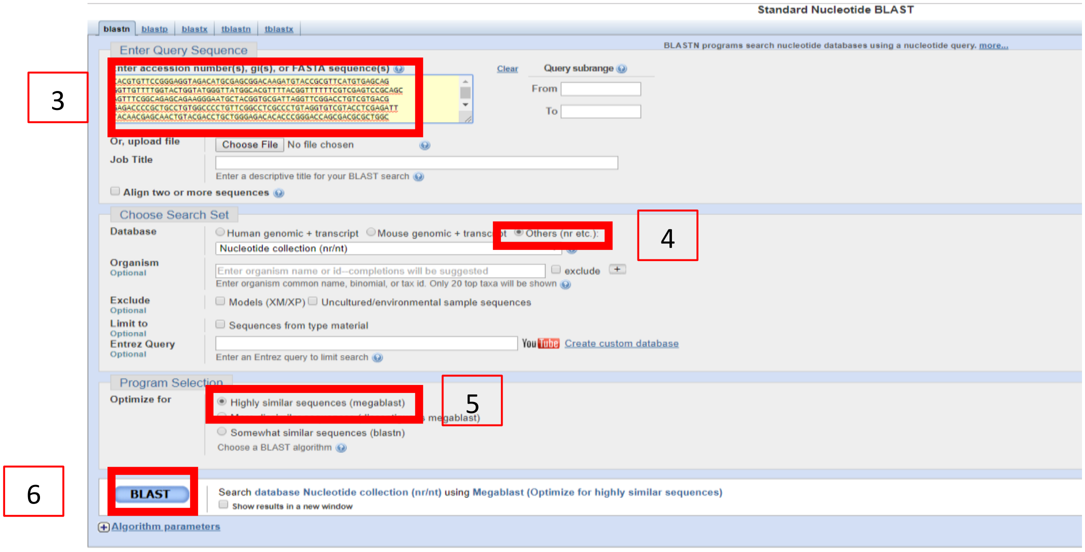
Human Mouse Rat Microbes

3. Copy and paste the relevant sequence into the box that says "**Enter Query Sequence"**

4. Select **standard databases** under **"Choose Search Set"**

5. Make sure the program is optimised for:
   "Highly similar sequences (megablast)" or **"blastp (protein-protein BLAST)"**

6. Click "**BLAST"**



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

nter accession number(s), gi(s), or FASTA sequence(s)

```
ACGTGTTCCGGGAGGTAGACATGCGAGCGGACAAGATGTACCGCGTTCATGTGAGCAG
GTTGTTTTGGTACTGGTATGGGTTATGGCACGTTTTACGGTTTTTTCGTCGAGTCCGCAGC
GTTTCGGCAGAGCAGAAGGGAATGCTACGGTGCGATTAGGTTCGGACCTGTCGTGACG
AGACCCCGCTGCCTGTGGCCCCTGTTCGGCCTCGCCCTGTAGGTGTCGTACCTCGAGATT
ACAACGAGCAACTGTACGACCTGCTGGGAGACACACCCGGGACCAGCGACGCGCTGGC
```

Clear

Query subrange
From
To

Or, upload file: Choose File No file chosen

Job Title: Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):
Nucleotide collection (nr/nt)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude +
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional): YouTube Create custom database
Enter an Entrez query to limit search

Program Selection

Optimize for: ● Highly similar sequences (megablast)
megablast)
○ Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST

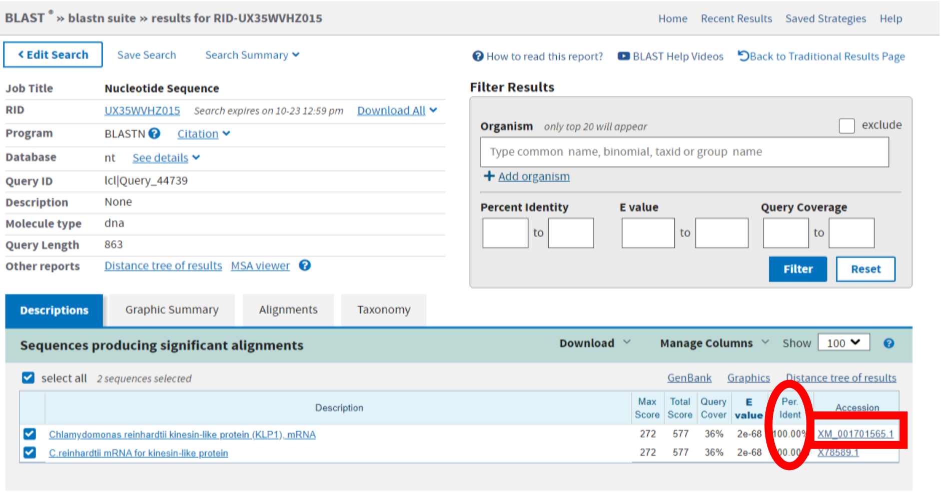
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
□ Show results in a new window

Algorithm parameters

7. WAIT! The BLAST program has to compare your sequence with every single sequence they have on their database. This can sometimes take a long time.

How to get the results of your BLAST search

1. Click on the accession number of the entry that has the highest **percentage identity** and **query cover.** If there are a few that have the same percentage identity, then click on the one at the top of the list. You also want to look for the most complete **description**. Any descriptions that include terms such as: *hypothetical; unnamed; predicted* should be avoided. If you are still unsure of which sequence to select after using this information, please discuss this with your teacher or demonstrator.



Job Title	Nucleotide Sequence
RID	UX35WVHZ015 *Search expires on 10-23 12:59 pm* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_44739
Description	None
Molecule type	dna
Query Length	863
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity □ to □ E value □ to □ Query Coverage □ to □

Filter Reset

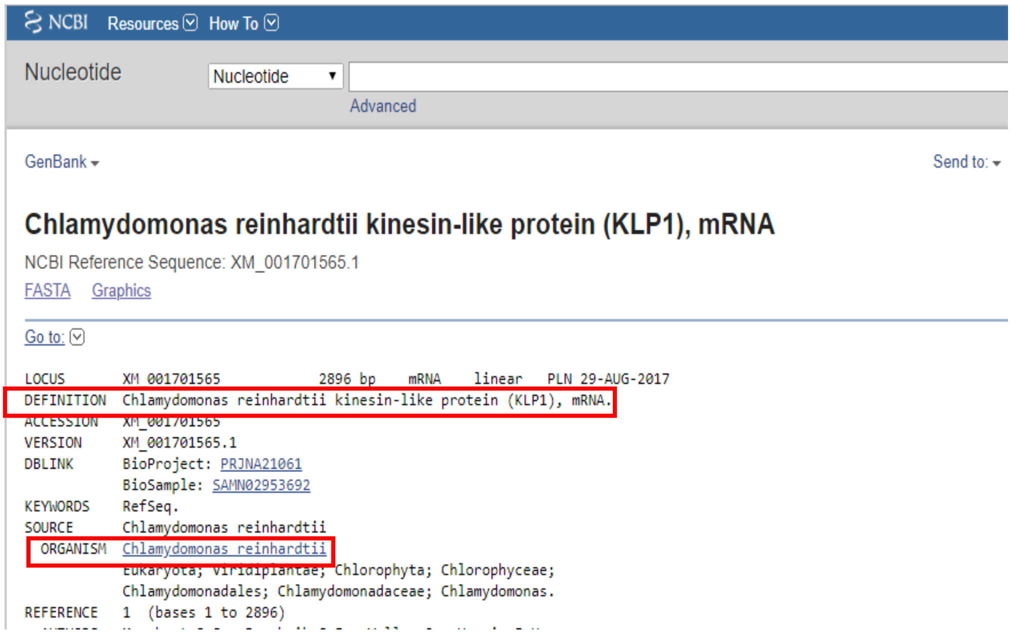
Sequences producing significant alignments

Show 100

☑ select all *2 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Chlamydomonas reinhardtii kinesin-like protein (KLP1), mRNA	272	577	36%	2e-68	100.00%	XM_001701565.1
☑	C.reinhardtii mRNA for kinesin-like protein	272	577	36%	2e-68	100.00%	X78589.1

2. Once the sequence information opens, you will need to look for a few pieces of information. Firstly, look at the **definition** to find out what the gene actually is. Next, look at the **organism** to find out which species the gene is from. Make sure you write this on the answer sheet with correct binomial nomenclature.



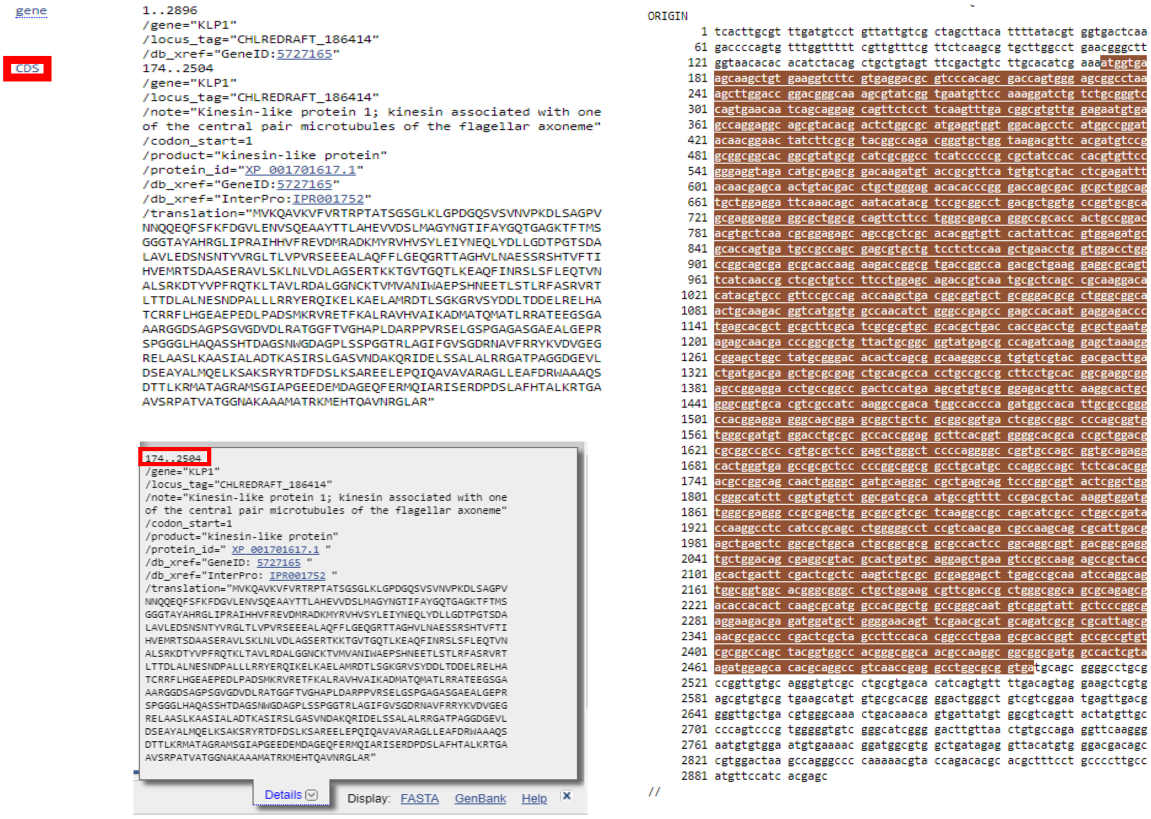
Chlamydomonas reinhardtii kinesin-like protein (KLP1), mRNA

NCBI Reference Sequence: XM_001701565.1

```
LOCUS       XM_001701565            2896 bp    mRNA    linear   PLN 29-AUG-2017
DEFINITION  Chlamydomonas reinhardtii kinesin-like protein (KLP1), mRNA.
ACCESSION   XM_001701565
VERSION     XM_001701565.1
DBLINK      BioProject: PRJNA21061
            BioSample: SAMN02953692
KEYWORDS    RefSeq.
SOURCE      Chlamydomonas reinhardtii
  ORGANISM  Chlamydomonas reinhardtii
            Eukaryota; Viridiplantae; Chlorophyta; Chlorophyceae;
            Chlamydomonadales; Chlamydomonadaceae; Chlamydomonas.
REFERENCE   1  (bases 1 to 2896)
```

3. Scroll down until you find **CDS.** This is your coding sequence. If you click on this, it will do 2 things. Firstly, it will highlight the coding sequence in brown *(the regions that do not get highlighted are untranslated regions, such as promotors, and will not code for the gene you are trying to identify).* Clicking on CDS will also bring up a small window in the bottom right hand corner of your screen.

4. The two numbers that appear in the top of the window that has just popped up are very important. These numbers will tell you the position of the nucleotides within the coding sequence. (In the example provided, the coding sequence starts at the 174th nucleotide, and ends at the 2504th nucleotide.)



```
gene            1..2896
                /gene="KLP1"
                /locus_tag="CHLREDRAFT_186414"
                /db_xref="GeneID:5727165"
CDS             174..2504
                /gene="KLP1"
                /locus_tag="CHLREDRAFT_186414"
                /note="Kinesin-like protein 1; kinesin associated with one
                of the central pair microtubules of the flagellar axoneme"
                /codon_start=1
                /product="kinesin-like protein"
                /protein_id="XP_001701617.1"
                /db_xref="GeneID:5727165"
                /db_xref="InterPro:IPR001752"
                /translation="MVKQAVKVFVRTRPTATSGSGLKLGPDGQSVSVNVPKDLSAGPV
                NNQQEQFSFKFDGVLENVSQEAAYTTLAHEVVDSLMAGYNGTIFAYGQTGAGKTFTMS
                GGGTAYAHRGLIPRAIHHVFREVDMRADKMYRVHVSYLEIYNEQLYDLLGDTPGTSDA
                LAVLEDSNSNTYVRGLTLVPVRSEEEALAQFFLGEQGRTTAGHVLNAESSRSHTVFTI
                HVEMRTSDAASERAVLSKLNLVDLAGSERTKKTGVTGQTLKEAQFINRSLSFLEQTVN
                ALSRKDTYVPFRQTKLTAVLRDALGGNCKTVMVANIWAEPSHNEETLSTLRFASRVRT
                LTTDLALNESNDPALLLRRYERQIKELKAELAMRDTLSGKGRVSYDDLTDDELRELHA
                TCRRFLHGEAEPEDLPADSMKRVRETFKALRAVHVAIKADMATQMATLRRATEEGSGA
                AARGGDSAGPSGVGDVDLRATGGFTVGHAPLDARPPVRSELGSPGAGASGAEALGEPR
                SPGGGLHAQASSHTDAGSNWGDAGPLSSPGGTRLAGIFGVSGDRNAVFRRYKVDVGEG
                RELAASLKAASIALADTKASIRSLGASVNDAKQRIDELSSALALRRGATPAGGDGEVL
                DSEAYALMQELKSAKSRYRTDFDSLKSAREELEPQIQAVAVARAGLLEAFDRWAAAQS
                DTTLKRMATAGRAMSGIAPGEEDEMDAGEQFERMQIARISERDPDSLAFHTALKRTGA
                AVSRPATVATGGNAKAAAMATRKMEHTQAVNRGLAR"
```

```
174..2504
/gene="KLP1"
/locus_tag="CHLREDRAFT_186414"
/note="Kinesin-like protein 1; kinesin associated with one
of the central pair microtubules of the flagellar axoneme"
/codon_start=1
/product="kinesin-like protein"
/protein_id=" XP_001701617.1 "
/db_xref="GeneID: 5727165 "
/db_xref="InterPro: IPR001752 "
/translation="MVKQAVKVFVRTRPTATSGSGLKLGPDGQSVSVNVPKDLSAGPV
NNQQEQFSFKFDGVLENVSQEAAYTTLAHEVVDSLMAGYNGTIFAYGQTGAGKTFTMS
GGGTAYAHRGLIPRAIHHVFREVDMRADKMYRVHVSYLEIYNEQLYDLLGDTPGTSDA
LAVLEDSNSNTYVRGLTLVPVRSEEEALAQFFLGEQGRTTAGHVLNAESSRSHTVFTI
HVEMRTSDAASERAVLSKLNLVDLAGSERTKKTGVTGQTLKEAQFINRSLSFLEQTVN
ALSRKDTYVPFRQTKLTAVLRDALGGNCKTVMVANIWAEPSHNEETLSTLRFASRVRT
LTTDLALNESNDPALLLRRYERQIKELKAELAMRDTLSGKGRVSYDDLTDDELRELHA
TCRRFLHGEAEPEDLPADSMKRVRETFKALRAVHVAIKADMATQMATLRRATEEGSGA
AARGGDSAGPSGVGDVDLRATGGFTVGHAPLDARPPVRSELGSPGAGASGAEALGEPR
SPGGGLHAQASSHTDAGSNWGDAGPLSSPGGTRLAGIFGVSGDRNAVFRRYKVDVGEG
RELAASLKAASIALADTKASIRSLGASVNDAKQRIDELSSALALRRGATPAGGDGEVL
DSEAYALMQELKSAKSRYRTDFDSLKSAREELEPQIQAVAVARAGLLEAFDRWAAAQS
DTTLKRMATAGRAMSGIAPGEEDEMDAGEQFERMQIARISERDPDSLAFHTALKRTGA
AVSRPATVATGGNAKAAAMATRKMEHTQAVNRGLAR"
```

```
ORIGIN
        1 tcacttgcgt ttgatgtcct gttattgtcg ctagcttaca ttttatacgt ggtgactcaa
       61 gaccccagtg tttggttttt cgttgtttcg ttctcaagcg tgcttggcct gaacgggctt
      121 ggtaacacac acatctacag ctgctgtagt ttcgactgtc ttgcacatcg aaaatggtga
      181 agcaagctgt gaaggtcttc gtgaggacgc gtcccacagc gaccagtggg agcggcctaa
      241 agcttggacc ggacgggcaa agcgtatcgg tgaatgttcc aaaggatctg tctgcgggtc
      301 cagtgaacaa tcagcaggag cagttctcct tcaagtttga cggcgtgttg gagaatgtga
      361 gccaggaggc agcgtacacg actctggcgc atgaggtggt ggacagcctc atggccggat
      421 acaacggaac tatcttcgcg tacggccaga cgggtgctgg taagacgttc acgatgtccg
      481 gcggcggcac ggcgtatgcg catcgcggcc tcatcccccg cgctatccac cacgtgttcc
      541 gggaggtaga catgcgagcg gacaagatgt accgcgttca tgtgtcgtac ctcgagattt
      601 acaacgagca actgtacgac ctgctgggag acacacccgg gaccagcgac gcgctggcag
      661 tgctggagga ttcaaacagc aatacatacg tccgcggcct gacgctggtg ccggtgcgca
      721 gcgaggagga ggcgctggcg cagttcttcc tgggcgagca gggccgcacc actgccggac
      781 acgtgctcaa cgcggagagc agccgctcgc acacggtgtt cactattcac gtggagatgc
      841 gcaccagtga tgccgccagc gagcgtgctg tcctctccaa gctgaacctg gtggacctgg
      901 ccggcagcga gcgcaccaag aagaccggcg tgaccggcca gacgctgaag gaggcgcagt
      961 tcatcaaccg ctcgctgtcc ttcctggagc agaccgtcaa tgcgctcagc cgcaaggaca
     1021 catacgtgcc gttccgccag accaagctga cggcggtgct gcgggacgcg ctgggcggca
     1081 actgcaagac ggtcatggtg gccaacatct gggccgagcc gagccacaat gaggagaccc
     1141 tgagcacgct gcgcttcgca tcgcgcgtgc gcacgctgac caccgacctg gcgctgaatg
     1201 agagcaacga cccggcgctg ttactgcggc ggtatgagcg ccagatcaag gagctaaagg
     1261 cggagctggc tatgcgggac acactcagcg gcaagggccg tgtgtcgtac gacgacttga
     1321 ctgatgacga gctgcgcgag ctgcacgcca cctgccgccg cttcctgcac ggcgaggcgg
     1381 agccggagga cctgccggcc gactccatga agcgtgtgcg ggagacgttc aaggcactgc
     1441 gggcggtgca cgtcgccatc aaggccgaca tggccaccca gatggccaca ttgcgccggg
     1501 ccacggagga gggcagcgga gcggctgctc gcggcggtga ctcggccggc cccagcggtg
     1561 tgggcgatgt ggacctgcgc gccaccggag gcttcacggt ggggcacgca ccgctggacg
     1621 cgcggccgcc cgtgcgctcc gagctgggct ccccaggggc cggtgccagc ggtgcagagg
     1681 cactgggtga gccgcgctcc cccggcggcg gcctgcatgc ccaggccagc tctcacacgg
     1741 acgccggcag caactggggc gatgcagggc cgctgagcag tcccggcggt actcggctgg
     1801 cgggcatctt cggtgtgtct ggcgatcgca atgccgtttt ccgacgctac aaggtggatg
     1861 tgggcgaggg ccgcgagctg gcggcgtcgc tcaaggccgc cagcatcgcc ctggccgata
     1921 ccaaggcctc catccgcagc ctgggggcct ccgtcaacga cgccaagcag cgcattgacg
     1981 agctgagctc ggcgctggca ctgcggcgcg gcgccactcc ggcaggcggt gacggcgagg
     2041 tgctggacag cgaggcgtac gcactgatgc aggagctgaa gtccgccaag agccgctacc
     2101 gcactgactt cgactcgctc aagtctgcgc gcgaggagct tgagccgcaa atccaggcag
     2161 tggcggtggc acgggcgggc ctgctggaag cgttcgaccg ctgggcggca gcgcagagcg
     2221 acaccacact caagcgcatg gccacggctg gccgggcaat gtcgggtatt gctcccggcg
     2281 aggaagacga gatggatgct ggggaacagt tcgaacgcat gcagatcgcg cgcattagcg
     2341 aacgcgaccc cgactcgcta gccttccaca cggccctgaa gcgcaccggt gccgccgtgt
     2401 cgcggccagc tacggtggcc acgggcggca acgccaaggc ggcggcgatg gccactcgta
     2461 agatggagca cacgcaggcc gtcaaccgag gcctggcgcg gtgatgcagc ggggcctgcg
     2521 ccggttgtgc agggtgtcgc ctgcgtgaca catcagtgtt ttgacagtag gaagctcgtg
     2581 agcgtgtgcg tgaagcatgt gtgcgcacgg ggactgggct gtcgtcggaa tgagttgacg
     2641 gggttgctga cgtgggcaaa ctgacaaaca gtgattatgt ggcgtcagtt actatgttgc
     2701 cccagtcccg tgggggtgtc gggcatcggg gacttgttaa ctgtgccaga ggttcaaggg
     2761 aatgtgtgga atgtgaaaac ggatggcgtg gctgatagag gttacatgtg ggacgacagc
     2821 cgtggactaa gccagggccc caaaaacgta ccagacacgc acgctttcct gccccttgcc
     2881 atgttccatc acgagc
//
```

Questions that need to be answered

All answers will need to be entered into the answer sheet that is available on Moodle in the bioinformatics assignment folder.

Question 1. (2 marks)

What is the process by which information is transferred from a DNA to an mRNA?

Question 2. (2 marks)

What is the process by which polypeptide molecules are made from an mRNA?

There are 4 unknown sequences. You will need to answer the following questions for each sequence. These answers will also need to be entered into the answer sheet that is available on Moodle in the bioinformatics assignment folder.

Question 1. (2 marks per sequence)

Which type of BLAST search did you choose and why?

Question 2. (2 marks per sequence)

What is the protein (coded for by the gene for a nucleic acid sequence) and what is its function?

Question 3. (1 mark per sequence)

Which organism has the unknown sequence come from? *Make sure the name is written with the correct binomial nomenclature, which may not be what is written correctly on BLAST.*

Question 4. (1 mark per sequence)

Is this organism prokaryotic or eukaryotic? *Where the organism is listed, this will also include all the relevant taxonomic classifications. This should answer this question.*

Question 5. (1 mark per sequence)

What is the STOP codon for the gene? *This will be at the end of the coding sequence. Make sure you list it as it is in the sequence.* If the sequence is an amino acid sequence, then you will not be able to discern the STOP codon. If this is the case, write "NA" as your answer.

Question 6. (1 mark per sequence)

How many nucleotides are there in the coding sequence? *The numbers that appear in the window when you click on CDS will help you figure this out. Remember that the number must be divisible by 3.*

Question 7. (1 mark per sequence)

How many amino acids are there in the entire protein (number of amino acids in the translated sequence)? *Remember that one amino acid is coded for by 3 nucleotides.*