Bi1X Spring 2013
Bioinformatics assignment

Computer requirements: To complete this exercise you will need to install Geneious (www.geneious.com), which is available for both Unix, Mac and Windows. Activate the 14 free day full trial, to open up locked functionality that will be used in this assignment.

Part I: Learn a little more about the 16S gene

The *rrsA* gene of *E. coli* codes for an RNA (rRNA) molecule that is part of the ribosomal small subunit. Let's learn a little more about this gene. Go to the BioCyc microbial database (the default organism being *E. coli*)

http://ecocyc.org/

In the search box type *rrsA* and click **rrsA** under **genes**.

1. What is the length of this gene?

2. Since you used 'all bacterial primers' for your PCR your primers should also match this rRNA gene. Find out the length of the amplicon you generated. The primer sequences you used were

   357F (16S forward primer) 5' CTCCTACGGGAGGCAGCAG 3'
   492RL2D (16S reverse primer) 5' TACGGYTACCTTGTTACGACTT 3'

   In the reverse primer 'Y' stands for either 'C' *or* 'T'. Here is a tool to help you reverse complement a nt sequence
   http://www.geneinfinity.org/sms/sms_reversecomplement.html

   Hint: *To get E. coli's 16S sequence click on the* ***rrsA*** *hyperlink under* ***Genes*** *and then* press the **Nucleotide Sequence** button*. To convert this sequence into a long string of letters either copy this sequence into your favorite text editor and use the "search-replace" function to remove white spaces and line breaks. Another option is to copy the sequence into the Geneious -> Sequence -> New sequence dialog, and select the whole gene the full sequence with your mouse.*

3. Click on the **Genome Browser** button. You will see that *rrsA* is part of an operon. What other genes are present in this operon? What do they do? Is

their function related to *rrsA*?

4. Does *E. coli* have more than one copy of this gene? In the EcoCyc website go to **Search → BLAST** and BLAST the *nucleotide* sequence of the rrsA gene against *E. coli's* genome (the default database). How many hits do you get? What is the name of the homologous genes that you found?

5. How about other organisms? Do other organisms have multiple copies of this gene? What is range of copy numbers in nature? To find out the answer take the forward primer above and BLAST it against other organisms. You can change the organism database in BioCyc by clicking on the **change** hyperlink at the upper right corner. Try this for a few prokaryotes, including the spore forming bacterium *Bacillus subtilis* (a different strain of which causes anthrax), the highly abundant cyanobacterium *Prochlorococcus marinus,* the human pathogen *Streptococcus pyogenes* and finally the extreme thermophile *Thermococcus kodakarensis*. ***Hint:*** *Use "BLAST" from the "Search" menu and remember to compare the primer nucleotide sequence with a nucleotide database*.

6. How do you think the rRNA copy number can affect the PCR reaction you carried out? Do all prokaryotes get equal representation?

Part II: Analyzing your 16S rRNA sequences

1. Open up your sequencing chromatograms (.ab1 files) using Geneious and locate the poorly sequenced regions. Delete these regions form the beginning and the end of your sequences and save the trimmed sequences. If less than three of your sequencing reactions were successful complement with sequence numbers {32,40,78} from the course webpage (these reactions are known to be successful) for a total of three sequences. State clearly in your report which sequences you use.

2. Use HOMD (www.homd.org) to classify your three (trimmed) sequences. Find phylum, class, order, family, genus and species of the *closest* match in each case. Write a paragraph about what is known about these bacteria. If the score of the closest match is less than 80% your sequencing reaction was likely not successful, and you should use three sequences from the course webpage mentioned above.

3. Predict the (most stable) secondary structure for your 16S rRNA sequences using Geneious' RNA-fold function. Attach the corresponding figure in your report.

4. Import the already trimmed sequences gathered by the whole class from the course webpage. Align these sequences. Are there certain 16S rRNA areas that have far fewer mutations than average? Compare with the folded 16S rRNA structure and see where these conserved sequences are located. ***Hint:*** *If you select a base pair sequence it will appear highlighted in the RNA-fold structure. Include figure in your report. Limit yourself to two conserved areas.*

5. Using the data gathered by the whole class construct an unrooted phylogenetic tree with the "Tamura-Nei/Neighbor joining method". What phylum, class and order do the branches containing sequence numbers 32-40-78 and 34-60-62 correspond to respectively?

6. In Geneious select the tree you created and click "Distances". The distances are reported as number of nucleotide mismatches per total length. What is a typical sequence distance between the 32-40-78 branch and the 60-62 branch? What is a typical distance *within* the two branches? If the mutation rate was 0.1 % per million years, how long have the two branches been diverging? ***Note:*** *If one truly wants to find out the evolutionary origin of these organisms this strategy is a gross oversimplification!*

7. Translation of a gene begins at a *start* codon and ends at a *stop codon*. The sandwiched sequence between these two codons is called an *Open Reading Frame (ORF).* Does the *E. coli* rDNA (X80722) sequence contain any open reading frames? How many? Only count ORFs starting with the codons ATG/GTG, which essentially are the only two start codons used in *E. coli*. If you found ORFs, do you think these actually produce proteins? If not, why? ***Hint 1:****Use Geneious ORF finder function ("Annotate and predict menu").* ***Hint 2:*** *Google "Shine-Dalgarno_sequence".*

<u>Part III:</u> Yeast (CAN1) mutation experiment

Yeast with an intact CAN1 gene will import *canavanine,* which is a toxic chemical that will substitute argine in polypeptide chains leading to the ultimate death of the cell. In the presence of canavanine therefore only yeast cells with a mutated non-functional CAN1 can survive.

1. Import your CAN1 sequences to Geneious. Since CAN1 is too long to sequence in one single reaction the sequencing was performed using five primers labeled A, B, C, D, E. Moreover the experiment was performed in two different strains of yeast, one with a corrupted DNA repair system which should affect the mutation rate. Hence you should in total have **ten** sequences. If your sequencing reactions were not successful use sequences {10,20}. State your sequence numbers in your report. Also

import the functional CAN1 gene (NM_001178878).

2. Select the five sequences corresponding to one of the strains plus the functional CAN1 sequence, right click and choose "map to reference". In the dialog choose the functional CAN1 (NM_001178878) as reference. This step will assemble your five sequences and align the result with the functional CAN1 gene. Repeat this step for the five sequences corresponding to the other strain.

3. How many mutations in the CAN1 gene do you observe and where are they located? Do you find the results surprising? Are the mutations point mutations (single base substitution) or insertions/deletions? Is the mutation type the same in both strains? If the mutation is a point mutation will it result in an amino acid substitution? Only count mutations that at least *two* of the sequencing reactions agree upon. Remember that sequencing results are poor at the beginning and the end of a read. ***Hint:*** *To find out in what codon a point mutation is located open up the functional CAN1 sequence and in "Sequence view" choose "Colors - By translation".*