**Practical Assignment**

**Module topic:** Bioinformatics resources and databases

**Contact session title:** DNA sequence analysis

**Trainer:** Nicola Mulder, Shaun Aron

**DNA sequence analysis**

**Introduction**

DNA sequences can be extracted from public databases, or users may have their own piece of DNA which they have sequenced, and which requires some basic analysis. In this practical we explore basic analysis of a DNA sequence extracted from the public database. Analysis of DNA sequences usually entails finding important features encoded on the sequence. In this practical we will search for open reading frames and try to find suitable restriction enzyme sites for cloning most of the ORF.

**Tools used in this session**

For extraction of the sequence we will use the EBI website, and for analysis we will use other online tools.

**Please note**

* **Hand-in information** If you are formally enrolled in the IBT course,please upload your completed assignment to the Vula ‘Assignments’ tab. Take note of the final hand-in date, which will be indicated on Vula.

**Task 1: Extracting the DNA sequence**

We will work on the Escherichia coli DNA polymerase I gene, which is in ENA (European Nucleotide Archive) entry with accession number V00317.

**Task 1: instructions**

Go to the European Nucleotide Archive: <https://www.ebi.ac.uk/ena/browser/search> and search for the accession number “V00317”. Choose the **sequence** result with the correct accession number and open the entry.

1. What kind of nucleotide molecule type is it?

Keep the entry open for later. Save the sequence in FASTA format by clicking on the “FASTA” link next to “Download” in the right hand box on the entry page. Do not save it in MSWord, save it as a text or .fsa file.

Now go to <http://molbiol-tools.ca/> which provides a set of freely available online tools for sequence analysis. Under “DNA sequence analysis” tools select “Composition” and then the “[Genomics %G~C Content Calculator](https://www.sciencebuddies.org/science-fair-projects/references/genomics-g-c-content-calculator)”. Paste your sequence in the box (you can leave out the first line starting with >) and calculate the nucleotide composition of your sequence.

1. What are the numbers of each of the bases in your sequence and GC content?  What is the % GC content of your sequence?

**Task 2: Identify the open reading frame in the sequence**

**Task 2: instructions**

For translation we use the EMBOSS Sixpack tool. Go to: <http://www.bioinformatics.nl/emboss-explorer/> on the left under the heading “NUCLEIC GENE FINDING”, select *sixpack*. This does a six-frame translation of the nucleotide sequence.

Paste the sequence in the box or upload the fasta file, select the “bacterial” code for the “code to use”, and “Yes” for “ORF start with an M?” -this ensures the start codon is ATG which codes for methionine (M). Leave other options as the default.

In the results, at the bottom is the OUTPUT FILE which shows the various predicted ORF sequences, each starting with a line “>\_....”. This line provides information on the frame and number of amino acids. Save the longest ORF sequence (either copy and paste into a text editor or select “outseq” and edit out the other sequences).

1. Why is it important to select the bacterial codon usage?
2. What is the reading frame (+1, +2, +3, -1, -2 or -3) of the longest ORF?
3. What is the length of the protein sequence (number of amino acids (aa)) for the longest ORF?

Now go back to the original nucleotide entry of V00317 in the ENA. Click on the “EMBL” link next to “View” in the right hand box. Scroll down to the features (FT lines) to find the CDS (coding sequence) Compare the protein sequence in the entry with the one you predicted from the six frame translation (you could try aligning the sequences if you know how or just view the beginning and end of the sequence by eye).

1. Are the sequences the same?
2. What is the location (i.e. coordinates on the DNA sequence) of the coding sequence (CDS)?
3. What is the UniProtKB/Swiss-Prot accession number for the corresponding protein?
4. Paste the protein sequence that you got from Sixpack below.

**Task 3: Identifying possible restriction enzyme sites for cloning**

**Task 3: instructions**

If we want to clone the first part of the ORF we need to generate a “Restriction endonuclease map”, we will use: <http://bioinformatics.org/seqext/>. Clear the boxes, and paste your original **DNA** sequence in (be sure it is the DNA sequence (ACGTs) and not the protein sequence). Ignore the primers section (delete what is in there) and for “Translate reading frame” choose the frame you found above, then “Submit”. This generates a restriction map for the sequence and at the same time displays the predicted amino acid sequence for the selected frame.

To select restriction enzymes to use for cloning, you would need to find 2 different restriction enzymes that cut around the beginning of the predicted ORF (check the start of the ORF using the coordinates from question 7) and somewhere in the middle of the ORF to get the longest possible fragment. The enzymes should only cut the insert once each otherwise you would get multiple fragments. In the results output table (below the sequence view) the enzymes highlighted in green cut once, the location of where the enzyme will cut is shown as the nucleotide number.

1. Which restriction enzymes would you select to get the largest portion of the ORF (appropriate enzymes closest to the start and end of the ORF)? Hint: from the Restriction summary table find the green highlighted enzymes that cut closest to the beginning and middle/end of the ORF.