**The Sequence Aligner**

In this exercise you will pretend to be a sequence aligner. Your goal is to align the below sequences (all with sequence length of 12bp) from a hypothetical RNA-Seq experiment to its corresponding reference genome/trancriptome. You can use the search function in word (**ctrl-F**) to search for the sequences in the reference genomes and trancriptomes.

**Task 1: Alignment to reference genome:**

Open the reference genome (3000 bp including three known genes highlighted in different colors) in the file Genome\_1.docx

Try to align the below sequences to the reference genome, using **ctrl-F.**

Sequences:

CCAACATAGTGG

AAACGGAACAAC

ACCAGAGAGTAG

TCAATGTGCACT

TCAGTGATCGGA

**Q1:** Did the sequences align where you would expect?

**Q2:** Did all sequences align to the reference genome?

**Q3:** What could be the reasons that not all sequences aligned?

**Task 2: Alignment to reference transcriptome**

Try to re-align all the sequences above to the reference transcriptome in the file Transcritome.docx

**Q1.** Did all sequences align?

**Q2.** Did the same sequences align to the transcriptome and genome?

**Q3:** Why would sequences align to the transcriptome and not genome, and the other way around?

**Q4:** Did any reads not align at all? What could be the reason for this? How would you try to align this sequence?

**Task 3: Unaligned sequence**

The sequence TCAATGTGCACT which aligned only to the genome can potentially be a new transcript not previously annotated. This transcript is shown in the file Genome\_2\_extra\_gene.docx

The sequence TCAGTGATCGGA did not align to the genome or transcriptome. Try if you can align the unaligned sequence TCAGTGATCGGA by aligning sub-parts of the sequence-.

**Q1**: What is the maximum part/length of the sequence that can be aligned uniquely to the reference genome? Where does this part align?

**Q2:** Try to align the remaining part of sequence. What do you observe?

**Q3:** What information can you use to select the most likely correct alignment of the remaining part?

**Task 4 – Count Table**

Align the below sequences to the reference genome/transcriptome. Count the number of sequences that belongs to each gene in the table below. Count both these sequences and the sequences that was used in the previous tasks, and write the total for each gene in the table below.

AACATAGTGGCC

TGCACTTCAGTG

TATAGATTTACT

ATAGGCACGATA

CTGTGGCTCGGC

CAACTGTTGGCC

GTCAAAAAGTGC

TGGTGTAATAAG

GGAATTTCAACG

|  |  |
| --- | --- |
| GENE | COUNT |
| Gene 1 |  |
| Gene 2 |  |
| Gene 3 |  |
| New gene |  |

NOTE: The sequence length of 12bp is much shorter than what you would sequence in reality. Todays standard for sequencing 75-150 bp. In the exercise exons introns are depicted to be of similar length. In real eukaryotic genomes exons are much shorter than introns.