You have received a list with short DNA sequences (Primers.txt), these primers will be used to amplify fragments of human DNA using a multiplex PCR. The order of the primers in the list is random. We are interested to know exactly what amplicons can be formed if we were to combine all primers in one reaction, based on GRCh38.

**Assignment 1:**

- Provide a report containing the exact locations on the GRCh38 reference genome of all amplicons (max size 1000 bp) that can be formed. Also indicate which combination of primers was used to form each amplicon.
- Provide a fasta file with all sequences that were found (in GRCh38) using the primer combinations as headers. Exclude the primer sequences from the amplicon sequence.

Most of the sequences that were found are expected to contain repetitive elements (STRs), an example could be:

AAATAAATAAATAAATAAATAAATAAATAAAT

To make it easy to compare the number of the repeats we can use a short-hand annotation, e.g., [AAAT]8 for the sequence given above. However, because of the repetitive nature you can also use different starting points and annotate this sequence as:

[N]1[AATA]7[N]3 ; [N]2[ATAA]7[N]2 ; [N]3[TAAA]7[N]1

If we also consider the complementary strand we have 4 additional possibilities:

[TTTA]8 ; [N]1[TTAT]7[N]3 ; [N]2[TATT]7[N]2 ; [N]3[ATTT]7[N]1

We want to use an algorithm that reads fasta files and that transforms sequences containing STRs to a short-hand annotation. We provided you with a list of motifs (Motifs.txt), only these motifs should be considered to avoid the aforementioned ambiguity.

**Assignment 2:**

- Use the fasta files from assignment 1 and write an algorithm that can transform these sequences to a short-hand annotation using only the motifs that were provided. To be considered an STR at least 2 repetitions of the motif needs to be detected. All non-repeated sequences can be grouped using [N]n (where n represents the number of bases in that sequence). Both strands need to be considered, e.g., TTTATTTATTTATTTATTTA should result in [AAAT]5.

Please provide the generated output files, plus the script(s) that were used to produce these files (including a help function) in the week before the interview at latest. You can also add other descriptive files if you, for example, want to explain why you made certain decisions. In case of questions, please contact (Arwin Ralf (a.ralf@erasmusmc.nl) and Diego Montiel González (d.montielgonzalez@erasmusmc.nl).

Speed is good, quality is better, but accuracy is the most important to us.