➢ Assignment 2

1-Base calling :

Base calling is a crucial step in any sequencing method. It is a process of transforming a raw signal obtained from a sequencer into a string of nucleotides. In the case of nanopore sequencing, it is a computational processing of electric signal collected from an ONT instrument (MinION, GridION, or PromethION). The accuracy of base calling is influenced by two factors. First, the chemistry used can affect a signal-to-noise ratio. If the ratio is low, determination of underlying DNA sequence may not be possible . The second factor is how well the signal can be interpreted by a software used for base calling. To discriminate between signal and noise a specific training dataset is used, which may not be optimal for interpretation of real DNA molecule if the latter has, for instance, strong nucleotide composition bias. For example, the genome of malaria-causing parasite is 80% AT rich and nanopore base calling of reads from this genome is usually far from optimal, especially within homopolymers stretches.

2- Mapping :

Obtaining a sequence is only the beginning of the analysis. Depending on a biological question we ask or why the sequencing was done at the first place, there are several avenues that one may take. However, aligning raw reads to existing sequences is often the first task on the to do list. This is especially true if a sequencing project involves organisms for which genomes has been already decoded. Often, the task of aligning raw-sequencing reads to already determined sequences, for instance a genome, is called mapping raw reads to the target. Aligning is such a basic task in the molecular sequence analysis that several algorithms that deal with the problem were developed long before bioinformatics field existed .

Although these early algorithms are very elegant and guarantee optimal solution to the problem, they are computational heavy and not very practical when one has to deal with the vast number of sequences. Consequently, heuristic algorithms have been developed to conquer speed limitation of exhaustive algorithms. Probably, the most successful and the best known is the BLAST algorithm . It is worth to mention that BLAST was developed for database similarity searches and its goal is to find all the instances of similar sequences in a database.