ITN MARCH RETREAT PROJECT PROPROSAL: BASECALLING OF OXFORD NANOPORE READS USING MACHINE LEARNING METHODOLOGIES

by Menno Witteveen

Since the advent of next-generation sequencing, the drop in the cost of DNA sequencing has been dramatic. The commonly referred to "$1000 dollar genome" seems attainable within certainly within the next decade, which has the potential to be a great driver for personalized medicine. Currently there is a race between several parties to deliver this $1000 dollar genome and one very promising contender in this race is Oxford Nanotechnologies, a company which has developed and distributed a very useful nanopore sequencer. Although this technology is able to sequence very long DNA reads, the performance of the base calling algorithm is currently an issue that needs addressing in order for the technology to enjoy wider applicability.

This was the inspiration for trying to develop a more powerful base calling algorithm. The perspective is that Machine Learning methodologies can be a tremendous aid for this aim, since provides a way to model the complex physical interactions of a DNA molecule with a nanopore, whilst being trans-located through it.

By working with nanopore data it became apparent that more research is needed to develop algorithms that can effectively deal with this novel data type. Through the retreat experience was gained on different potential approaches to the problem, guiding future efforts.