**A new approach to Base-calling for Oxford Nanopore reads.**

**Section 1:**

*DNA sequencing* is the process of determining the precise order of nucleotides within a DNA molecule. Even though the field of DNA sequencing has improved significantly since the development of second-generation sequencing technologies,it still suffers from significant limitations. One such limitation is the read length. Nanopore based DNA sequencing is a 3rd generation sequencing method enabling long reads. Oxford Nanopore’s MinION uses this technology for sequencing long reads. It performs base-calling(determining the sequence of nucleotides) from the ionic current levels produced by the DNA fragment using proprietary algorithms(based on Hidden Markov Models).

The base calling technique used by Oxford nanopore is suboptimal with a low accuracy rate(~ 32%) and is not useful in a significant number of cases.Oxford Nanopore data has been recently made available which consists of the reads generated by the MinION instrument with thousands of base pairs as the average length. Our approach would be to use this data to train other machine learning models for obtaining better accuracy levels in base calling.

**Section 2:**

We will first process the Oxford nanopore data obtained in FAST5 format using the Poretools software and obtain the events data. The event data contains the columns such as the mean,start, standard deviation,length and the model states.The Oxford Nanopore base-calling software employs a Hidden Markov Model to predict a fragment’s read based on this event data. Similarly, we plan to use Support Vector Machine to predict the fragment’s read based on the events data.The feature vectors that we currently plan to use will be a subset of the column values corresponding to a time frame in the event data(something that is sufficient to uniquely identify a single signal event). We will update the feature vector accordingly and take the locality factors into consideration later, depending on how our model performs. The label that we plan to use for the model is a 5-mer. Hence the cardinality of the label space is 4^5. To train the data we will first align the reads generated by the Nanopore to the actual DNA sequence(for Yeast or E.coli). Then we will use the aligned string as our ‘raw data’ and train the SVM using that.

**Section 3:**

Oxford Nanopore uses a HMM based modified viterbi algorithm to obtain the reads from the raw signal. The paper ‘Alta-Cyclic: a self-optimizing base caller for next-generation sequencing ’ mentions an improved base calling technique which makes use of Support Vector Machines (SVM) to sequence the raw signal obtained from the Illumina genome analyzer. Another paper ‘naiveBayesCall: An efficient model-based base-calling algorithm for high-throughput sequencing’ introduces an algorithm naiveBayesCalls which utilizes approximation and optimization methods to achieve scalability while maintaining a reasonably high level of accuracy but is used mostly for short reads.