**Minimal Data Classification of Sequencing Strategy [using Machine Learning]**

The project will proceed in two main phases. First, we will test the viability of full-data classification of reads (in other words, map a read to probabilities that it came from each of N strategies). In the second phase, we will implement some sort of stopping algorithm.

**Phase 1: Model building and proof of concept**

Use “SRA Advanced Search Builder” to select for the bisulfite seq strategy and a particular organism (start with mouse). We will also play around with ways to specify the tissue type. Examining the data, we will import to Python and get a feel for the raw reads. We will also consider other metadata to include in the model. The first model will classify whether a read comes from the standard whole genome sequencing strategy or from the bisulfite sequencing (whole genome) strategy. We’ll consider in-sample and out-of-sample analysis, and consider different models, feature engineering, etc..

We will see if the results reasonably imply a chance of success for phase 2, for instance if the model maps Bisulfite-Seq reads to higher probability that the read came from Bisulfite-Seq rather than RNA-seq.

We will then see about using other features of the reads as labels [metadata].

I expect to run into memory and file storage issues immediately, so will either host on CAIS++ server or use HPC. Jupyter Notebooks and Python being suitable for this—to be determined.

Tasks:

1. NCBI/SRA queries
   1. Given a set of search terms, for instance strategy=bisulfite-seq, return fastq download links.
2. Load fastq reads into python with reasonable efficiency.
   1. Pipeline:
      1. Given: Web link to fastq
      2. Download to local
      3. Load reads into interactive python instance
3. K-mer counts
   1. Given a read ACCTGACGTAGCT, give k-mer counts (k=1,…,6)
   2. Use the principles of counting to make [stream]
   3. Do the equivalent for sets of reads [full data problem]
   4. How can we save this file format efficiently, with labels, to be suitable for training i.e. step 4 below.
4. Model Building:
   1. Input: K-mer counts of a read
      1. How should we organize this?
   2. Predict:
      1. Probability of coming from std. or bisulfite

**Phase 2: Minimal data problem**

We will use the “stream” capability of fastq-dump to analyze reads one by one and propose some sort of dynamic belief system with a stopping condition. Key words: Bayesian updating, an “emission” style model as in HMM, belief filter, Kalman filter, control, optimal stopping.

We will talk about the best use case for an app using our new classification technology – possibilities:

1. Web extension to tag SRA with new metadata
2. Separating out types of reads in blood samples

**Center for AI in Society blurb:**

The increasing amounts of genetic data becoming available due to the rise of Next Generation Sequencing (NGS) technologies are revolutionizing the studies of evolution, pharmaceuticals, and clinical medicine. However, NGS data is unique because it is *huge* data -- millions of reads in each experiment make it slow to analyze as a whole dataset. One open source for these sequenced reads from biological experiments is the Sequence Read Archive (SRA), which can stream the reads one at a time. Unfortunately, the SRA uploads are under-labeled which poses a massive problem for researchers mining the genetic data. Our group will be using machine learning, probabilistic models (like Hidden Markov Models), statistics, optimal stopping theory, and biology to analyze the reads to help researchers advance biomedical science.

**Title of Project: Minimal Data Machine Learning Classification of Genetic Sequencing Type**

In the field of genetic sequence analysis, which includes the use of Genome-Wide Association Studies (GWAS) using sequencing data, there is a need for a tool which determines the type and protocol used to generate the genetic sequencing data, as data is often “under-labeled” or even mislabeled on public databases and critical associated metadata is lost. These data are critical to recover so that new medical cohort studies can be performed accurately, but since the number of reads sequenced can reach the scale of millions, it is important to develop a technique which is capable of intelligent stopping after the minimal number of reads needed to determine the classes is reached. The project will proceed in three main phases. Firstly, we will utilize supervised machine learning techniques, such as random forests and recurrent neural networks, on the clearly separable binary classification problem of identifying bisulfite sequencing data. The second phase will expand the technique to a multi-class classification problem, in which each sequence generates a posterior probability vector of class labels such as RNA-seq or ChIP-seq. Finally, we will utilize the opportunity to stream the data from a single, unknown experiment on NCBI-Sequence Read Archive to dynamically update the class probabilities for that experiment and use an optimal stopping rule to stop analysis after convergence via the theories of probability and mathematical statistics. The result will be incorporated into an open-source application for general application to efficiently enhance sequence read archives vis a vis consistent automated tagging, greatly enhancing the capability for researchers to perform data mining on cohort studies. In addition, researchers can use the tool to verify the quality and type of their data when it is mislabeled or under-labeled. After implementation of the theory, the same technique can be used to identify tissue types based on gene data and epigenome profiles using minimal data for instance on the open problem of efficient classification of circulating free DNA in the blood.

**Plan:**