EE595: Data Science for Sequencing

Project Proposal

Andy Lin / Bowen Xue / Fan Zhang

Introduction

In forensic settings, criminal investigators will often collect biological evidence. This evidence could be in the form of biological material such a DNA or protein. Biological evidence could also come in the form of data. Data in sequencing and mass spectrometry files may be useful for determining facts in a forensic scenario.

One of the questions a forensic investigator may ask is where a biological sample originated from. Determining the source of biological material is important because it can be helpful in determining who perpetrated a crime. By determining where a sample came from, forensic investigators can use this information to whittle down suspect lists.

Data related to the sequencing of deoxyribonucleic acids (DNA) is one type of biological data that may be collected as part of a forensic investigation. Since sequencing data is easily transportable, it is feasible for a sample to be sequenced in one location and analyzed in another location. As a result, there is a need for a method to determine where a sample was sequenced. Therefore we will investigate whether it is possible to establish where a sample was sequenced based on its sequencing file. For this study, we plan to use machine learning on sequencing data to determine whether we can predict which laboratory sequenced which samples.

Data

We will gather our data from the Sequence Read Archive (SRA) database hosted by the National Center for Biotechnology Information (NCBI). The SRA is a publically available database containing sequencing read data as well as their associated experimental metadata. We plan to download sequencing datasets associated with two to four different institutions. The datasets from each institution will be split into a training, test, and hold out set. A single dataset will consist of data from a single sequencing run. We will focus on institutions that have sequenced *Escherichia coli* using Illumina MiSeqs.

Methodology

One of the potential problems we expect to face is to think of informative features that will separate laboratories from each other to use as input in our supervised and unsupervised training algorithms. We currently plan to use three type of features as inputs in our machine learning algorithms. The first feature type is the proportion of all 4-mers (DNA sequences of length 4) in a sequencing run. There are 256 different possible 4-mers. We can use the proportion of these 256 4-mers in a sequencing run as 256 features.

Another feature we plan on using is average fragment size. When DNA is prepared for sequencing, it is randomly sheared into smaller pieces. As a result, the DNA molecules that are sequenced have a length distribution associated with it. It could be that different laboratory sample handling protocols create DNA fragment size distributions.

Finally the third set of features we will use is the average quality score for each base in the sequencing run. In a sequencing run, each base has a different quality score associated with it. The distribution of these quality scores could differ between different laboratories.

For our supervised machine learning approach, we plan to use a logistic regression classifier. If our supervised machine learning approach looks promising, we also plan to try unsupervised approaches using the feature set described above.