Prediction of Lab Origins From Feature

Analysis of Sequencing Read Data

Bowen Xue, Andy Lin, Fan Zhang

**Background**

In forensic settings, criminal investigators may come across sequencing data and files that may be relevant to a criminal investigation. In these situations, it may be important to determine where a DNA sample was sequenced. Datasets from different sequencing institutions may have informative features which makes it possible to use machine learning to do the classification. Therefore, we will investigate, using machine learning, whether it is possible to predict the location a DNA sample was sequenced from sequencing data.

**Data**

Our dataset consisted of 359 paired-end sequencing runs that were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). We chose these datasets because we focused on institutions that sequenced *Escherichia coli* using Illumina MiSeqs. 174 of the 359 datasets were submitted to the SRA by the Centers for Disease Control and Prevention Enteric Diseases Laboratory Branch in May 2017. The remaining datasets were submitted by the Statens Serum Institut in Denmark in May 2017. All of the datasets used Illumina MiSeqs to sequence *Escherichia coli*.

The SRA Toolkit (version 2.8.2-1) was used to download SRA files from the Sequence Read Archive and to convert the SRA files into FASTQ files. Custom Python scripts (version 3.6.0) were used to do batch preprocessing. Figure 1 shows an example FASTQ file.

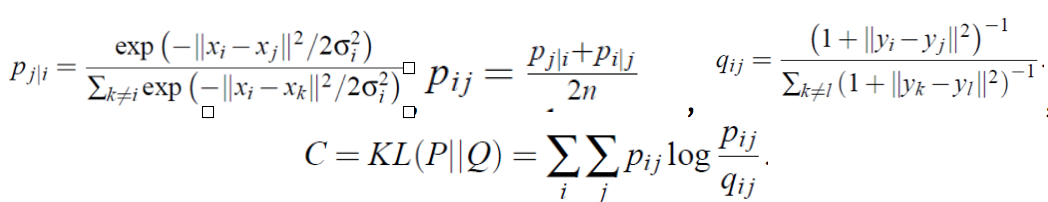
**Methodology**

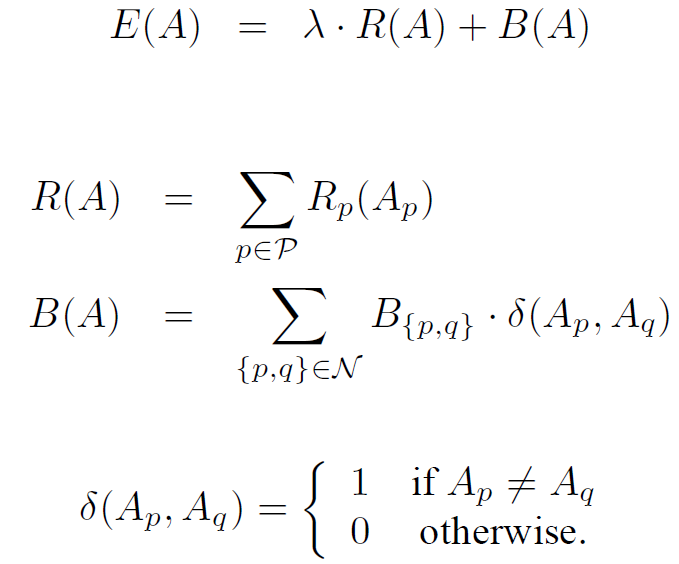
We used machine learning to determine whether genomic reads contained signatures of laboratory origin. We used both supervised and unsupervised methods to predict whether a sequencing run originated from one of two laboratories. The supervised methods we used included logistic regression and a support vector machine. Our unsupervised method involved the use of k-means clustering. Each dataset was split into three subsets (60%, 20%, 20%) for use a training, testing, and validation set.

**Unsupervised method**

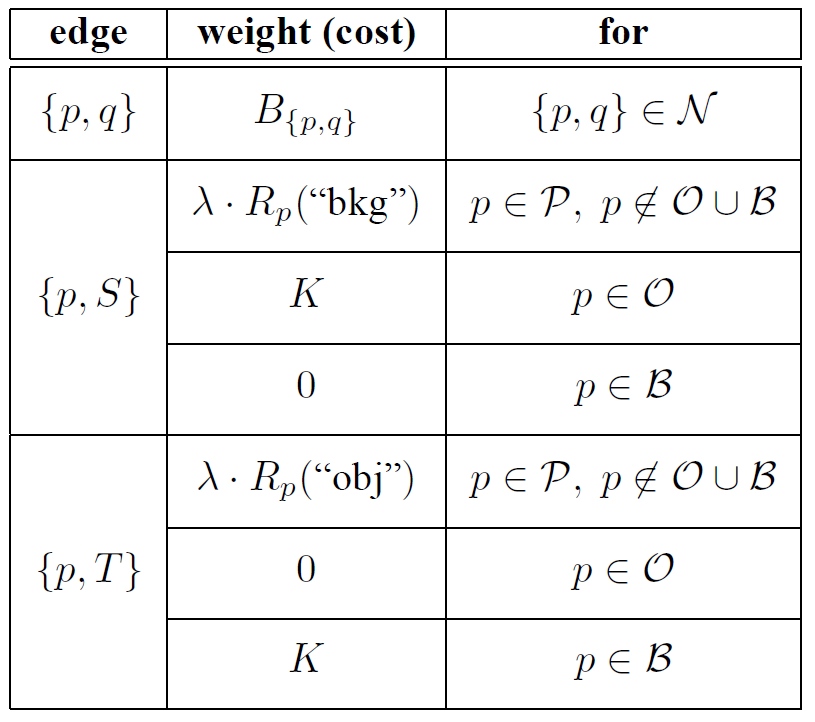
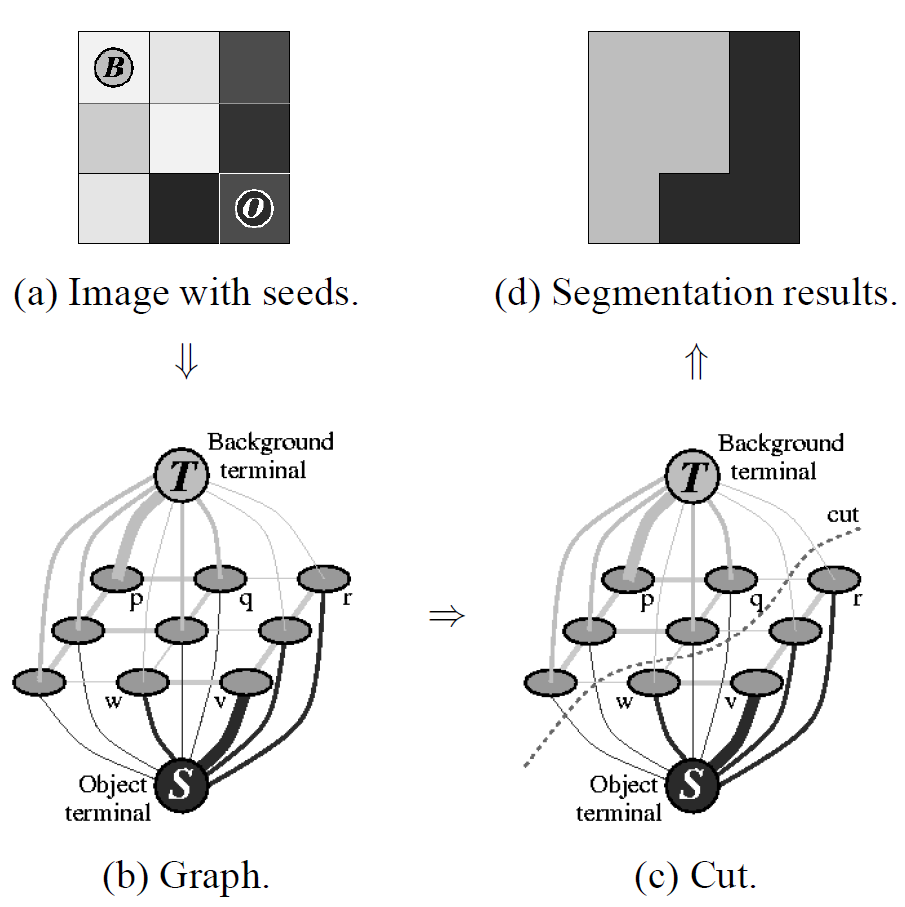
 For unsupervised learning, since the dimension of the features can be up to several hundred, we first tried dimension reduction using two methods: T-SNE (t-distributed stochastic embedding) and SIMLR (Single-cell Interpretation via Multi-Kernel Learning). SIMLR uses multiple kernels to give the best estimate for the dimension reduction result.

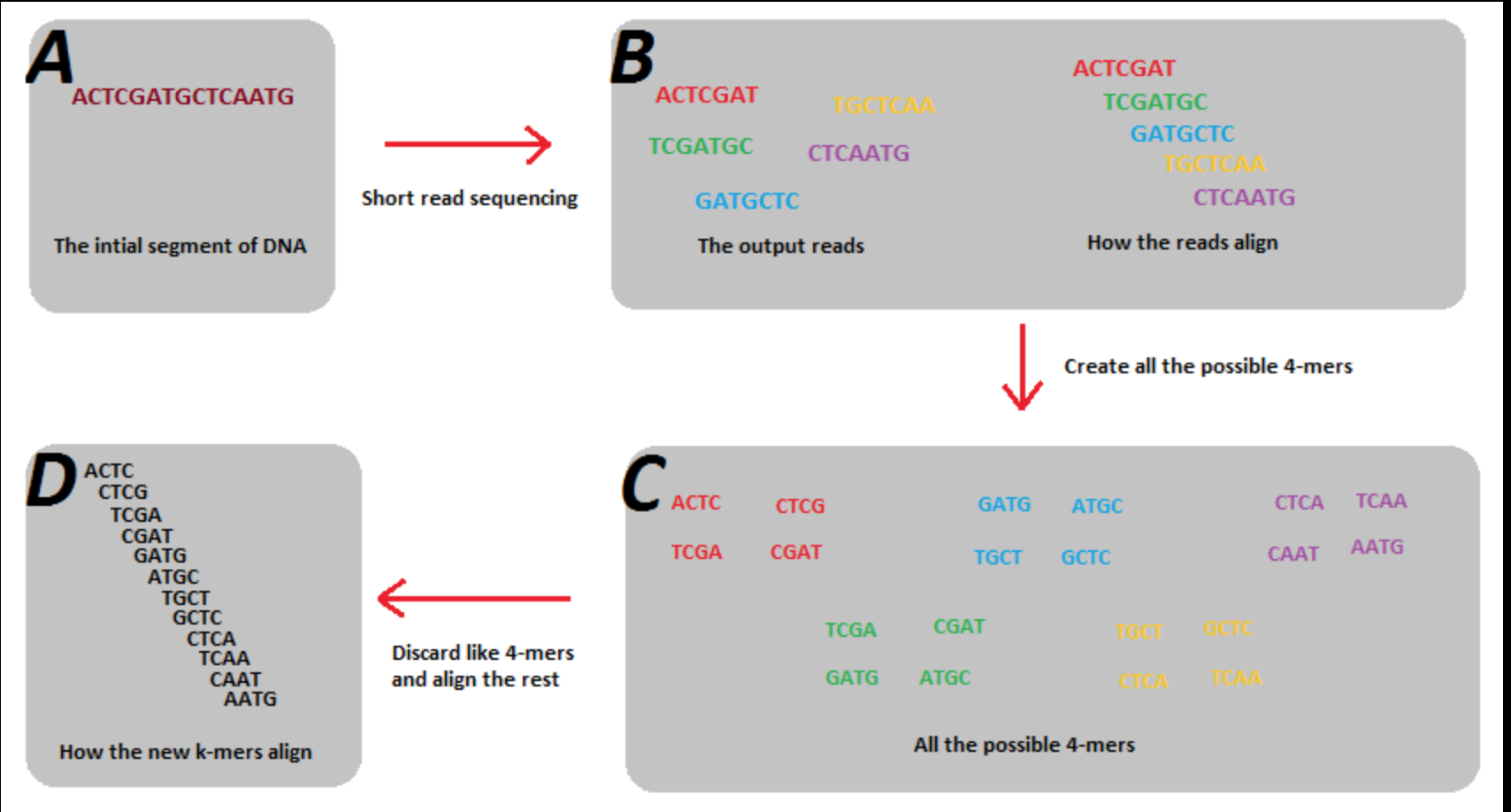
Figure 1: Screen shot from the NCBI SRA denoting what a FASTQ file looks like.

In the T-SNE framework, each single point has a distribution of potential neighbors on all other points, defined as Pj|i which translates as probability that j is i’s neighbor, which means each data has its internal view about all other points.

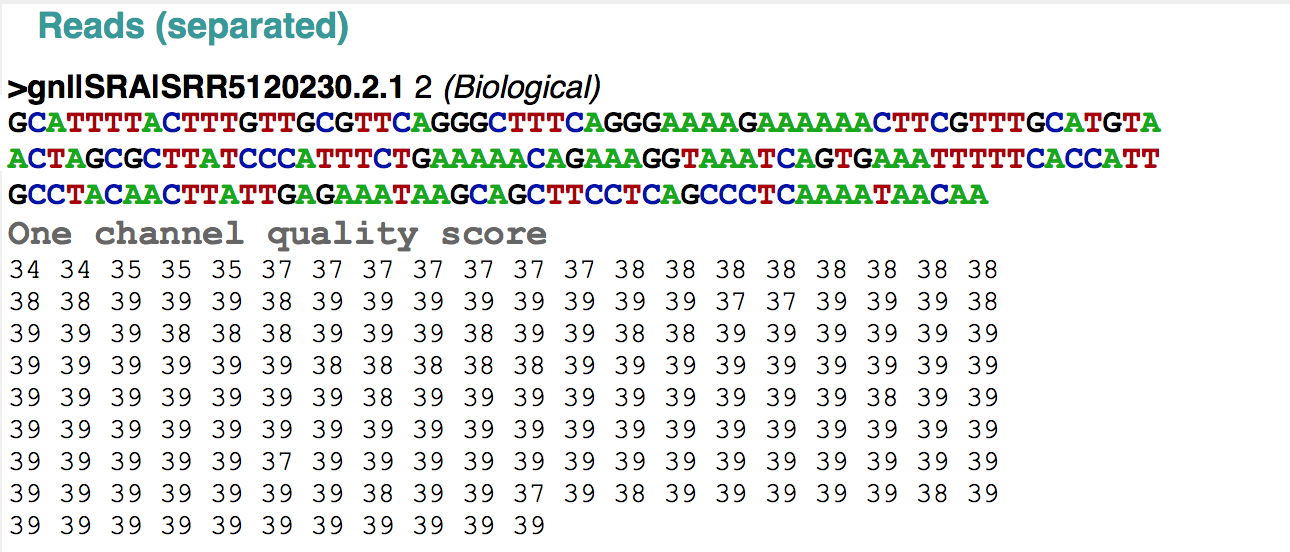
After some initial attempts using the K-means++ algorithm with the two dimensionality reduction techniques, the accuracy was just 0.573 with T-SNE and 0.615 with SIMLR. Due to the low classification accuracy, we simplified this unsupervised part to “semi-supervised learning”. Inspired by the idea from image segmentation, the Graphcut Algorithm does the following cost minimization:

Where the R(A) represents the data term which try to fit the optimal model for the measurement; B(A) represents the smoothness term, which comes from some prior knowledge. A is a vector with each element being a node, and the node is either an object or a background; P denotes all pixels in an image; R is a metrics which measures the fitness of a node Ap if we know the category of the node; B is a metrics which measures the similarity between nodes.



Since there is not a natural edge definition in the general graph, we define the neighbor edge, n-link, to be the neighbor probability in t-SNE section and data used here is high dimensional data. To reduce, the computation and noise, we only take highest 20 probable neighbors for each node. Hence the function B(p,q) has been defined. For data fitness part, Rp(), it is unreasonable to fit a histogram in high-dimension data due to curse of dimensionality and we don’t have enough data. Therefore R(p) is built from dimension reduced data, i.e. compute a histogram based on labelled object/ background data, and when encountered a new data, we fit the location info into that histogram and generated the scores.

**Features**

We tried three different feature types to infer the laboratory origin of sequencing runs. The first set of features was the proportion of every 4-mer in the reads of a sequencing run. Prior to sequencing, genomic DNA must randomly sheared into smaller pieces. There are several types of methods for shearing DNA, which include physical, enzymatic, and chemical. These biases may be ascertained by comparing the proportion of different 4-mers across sequencing runs. To calculate the proportion of 4-mers in a sequencing run, we slid a four DNA base size window down each read. At every position in the reads, we then would increment the corresponding 4-mer. Following normalization, this results in the proportion of each 4-mer in each sequencing run. There are either 256 or 625 features that result from this process, depending on whether ‘N’ is considered a unique nucleotide or not.

The second set of features we used is related to the quality of the sequencing reads. Depending on the protocol used and the skill of the technician processing the sample, there can be differences in the quality of the sequencing run. Each base in a sequencing read has an associated quality score with it. The distribution of quality scores and the drop in quality as a function of read length may differ between laboratories as a result of technician idiosyncrasies and protocol differences.

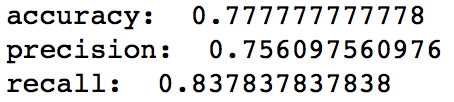
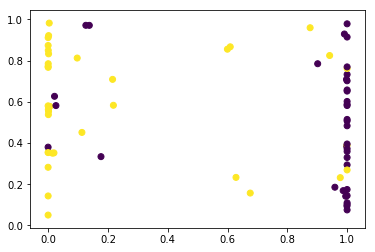
The third set of features used as input into the machine learning algorithms is the fragment size distribution of each sequencing run. Due to the various methods for fragmenting DNA prior to sequencing, it is possible that there is a signature each laboratory imparts on fragment length distribution. To determine whether fragment length distribution can differentiate between laboratory sources, we aligned paired-end reads together against a reference genome to determine fragment length. Each fragment length bin contains fragment lengths in a 200 base-pair window. The smallest bin size is -1,000 and the largest bin size is 1,000. Any fragment length that falls into a bin smaller than -1,000 is assigned to the -1,000th bin. Fragments with a length that fall into a bin larger than 1,000 is assigned to the 1,000th bin.

**Results**

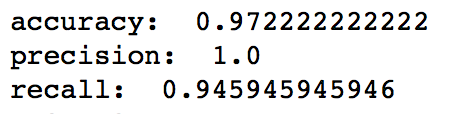
**Supervised Learning**

We found that logistic regression was successfully able to distinguish datasets that originated from different laboratories. How successful the logistic regression was at predicting the origin of a sample was depending the features given to the algorithm. Giving the logistic regression only the 4-mer feature resulted in a classification accuracy of ~77.8%. Precision was 75.6% and recall was ~83.8%. Figure 2 shows the probability that each sample in the hold out set originated from the CDC.

Figure 2: Each dot represents the probability that a dataset originated from the CDC. Yellow dots represent datasets that originated from Denmark and purple dots represent datasets that originated from the CDC. X-axis is probability.



Yellow dots represent datasets that originated from Denmark and purple dots represent datasets that originated from the CDC. The logistic regression classifier was very confident of its prediction, except for 4 datasets that originated from Denmark. It incorrectly classified them as coming from the CDC, however the logistic regression estimated that there was only a 60% that they originated from the CDC.



Using the logistic regression classifier with features resulting from the 4-mer analysis greatly improved the classification accuracy. There was a classification accuracy of ~97.2%. The precision was 100% and the recall was 94.5%. Using the 4-mer feature set, the only datasets that were misclassified were two sequencing runs from the CDC. Unfortunately, the classifier very confidently incorrectly predicted the origin of these samples. Further work will need to be conducted understand why this happened. Figure 3 shows the probability that each sample in the hold out set originated from the CDC using the set of 4-mer features.

Figure 3: Each dot represents the probability that a dataset originated from the CDC. Yellow dots represent datasets that originated from Denmark and purple dots represent datasets that originated from the CDC. X-axis is probability.

The logistic regression classifier just using the fragment length set of features had a classification accuracy of ~83.3%, Its precision was ~85.7% and its recall is ~81.1%. Figure 4 shows the probability that each sample in the hold out set originated from the CDC using the set of fragment length features.

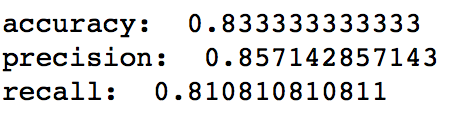
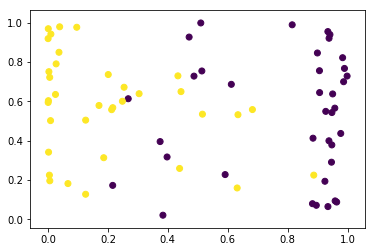


Figure 4: Each dot represents the probability that a dataset originated from the CDC. Yellow dots represent datasets that originated from Denmark and purple dots represent datasets that originated from the CDC. X-axis is probability.

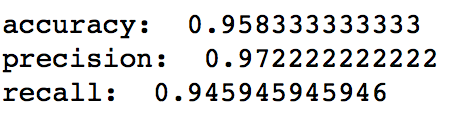
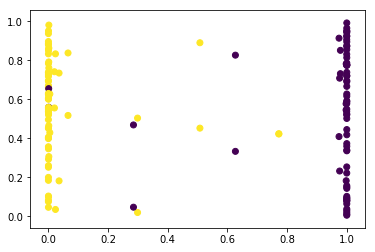
Finally, a logistic regression was built by using all three sets of features together. The combined classifier had an accuracy of ~83.3%. Its precision is ~85.7% and its recall is ~81.1%. The logistic regression classifier using all features performs worse than the classifier that just use the set of 4-mer features, which implies that the quality score feature set and the fragment length feature set may be uninformative or noise. Figure 5 shows the probability that each sample in the hold out set originated from the CDC using all features

Using a SVM with the combined set of features turned out to have better performance than a logistic regression using all features. However, the SVM did not have better performance than the logistic regression that just used the 4-mer set of features. The accuracy was ~95.8%. The precision was ~97.2%. The recall was ~94.5%.

**Comparison of semi-supervised algorithms**

Figure 5 Each dot represents the probability that a dataset originated from the CDC. Yellow dots represent datasets that originated from Denmark and purple dots represent datasets that originated from the CDC. X-axis is probability.

We compared two semi-supervised learning algorithms. We saw that the developed algorithm did not perform well and algorithms like dimension reduction did not improve the accuracy compared to simple supervised learning methods. Figure 6 shows the accuracy of the two semi-supervised algorithms as compared to the logistic regression.



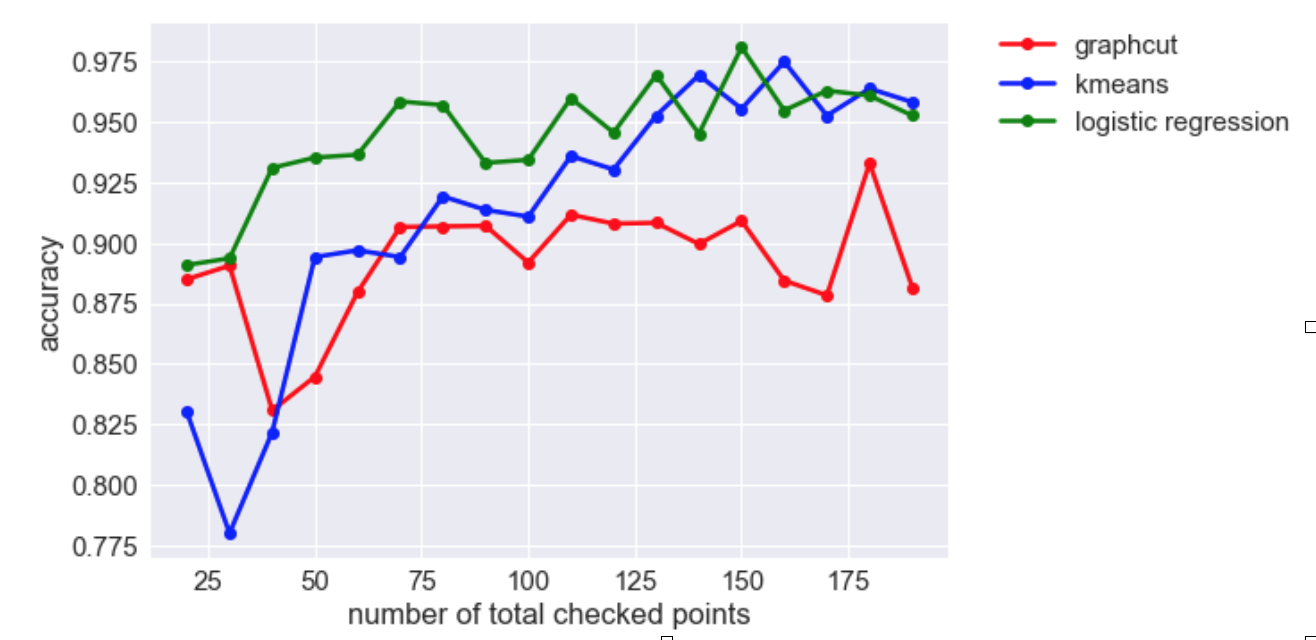
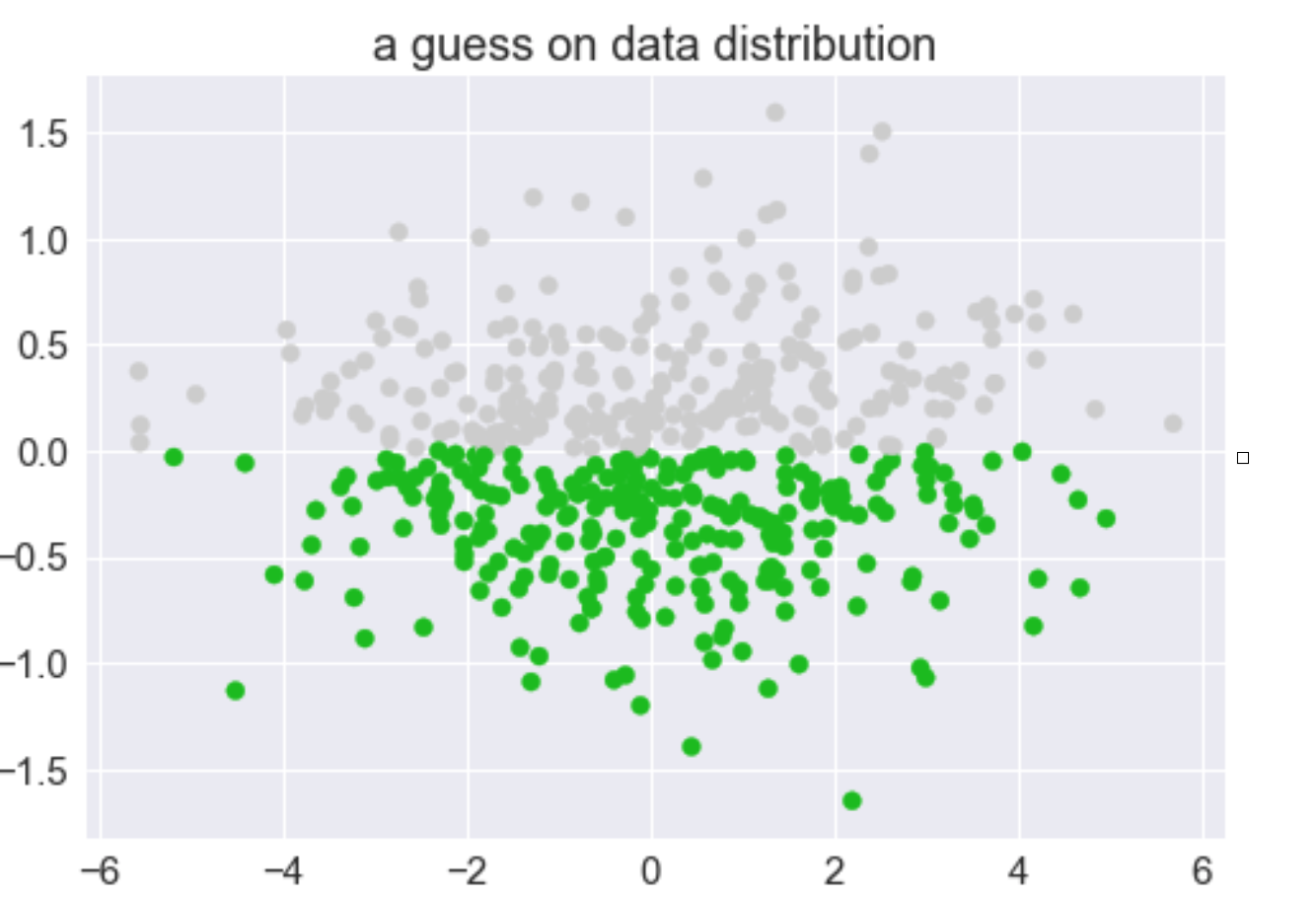
****Since logistic regression can make fairly accurate prediction based on a few points. Data in high dimension space is easily separable and there is only a few mixing there. On the other hand, in the unsupervised part, it turned out it is hard to cluster them without a few known points.

Figure 6: Comparison of the accuracies of two semi-supervised algorithms with a supervised algorithm

We also tried ICA to try to cluster based on “non-gaussianity”, but the result is also poor. The data might follow the similar distribution as the following graph, which data are easily classified once some points are known, but in general hard to cluster. Although in the following case it would be easy, by considering the second PCA component.

**Future Work**

There are several paths forward that we could pursue to improve and validate these classification methods. One way we could improve the method is to add additional features. For example, we could add the mean and standard deviation of the quality score distribution as features. In addition, we could model the average decrease in base call quality as a function of read length and add those as features.

To further validate this method, additional datasets will need to be downloaded and analyzed. Adding additional datasets from the same laboratory will create a more robust signature. In addition, it will be important to add datasets from organisms other than *E. coli* to prevent overfitting to a single species. Finally, the addition of data from additional laboratories will allow us to determine whether this method is generalizable beyond discriminating between two laboratories.

**References**

Langmead et. al., “Fast gapped-read alignment with Bowtie 2.” Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923.

Leek et. al., “Tackling the widespread and critical impact of batch effects in high-throughput data.” Nat Rev Genet. 2010 Oct;11(10):733-9. doi: 10.1038/nrg2825. Epub 2010 Sep 14.

Leek et. al., “svaseq: removing batch effects and other unwanted noise from sequencing data.” Nucleic Acids Res 2014; 42 (21): e161. doi: 10.1093/nar/gku864

Tung et. al., “Batch effects and the effective design of single-cell gene expression studies.” Sci Rep. 2017 Jan 3;7:39921. doi: 10.1038/srep39921.