**Prediction of Lab Origins From Feature Analysis of Sequencing Read Data**

Bowen Xue/ Andy Lin/ Fan Zhang

**Background**

For forensic investigation, it’s 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；

We will focus on institutions that have sequenced E.coli using Illumina MiSeqs；

**Data**

359 paired-end sequencing runs were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). 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. The same toolkit was used to convert the SRA files into FASTQ files. Custom Python scripts (version 3.6.0) were used to preprocess the data into a useable format as input into the machine learning algorithms.

Reference where the experiment data came from

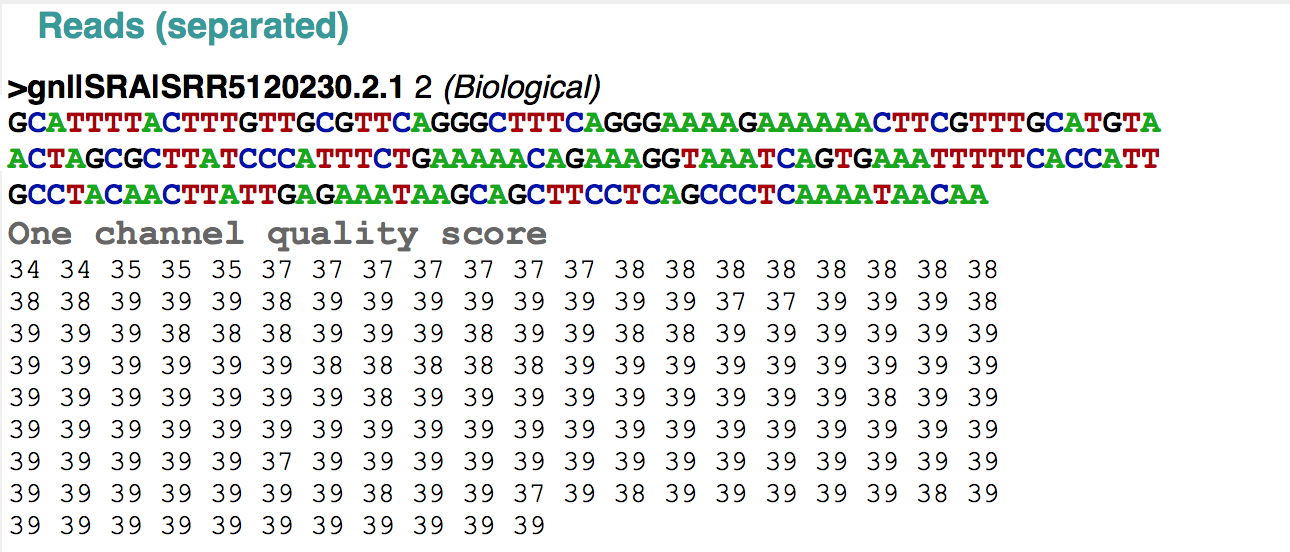
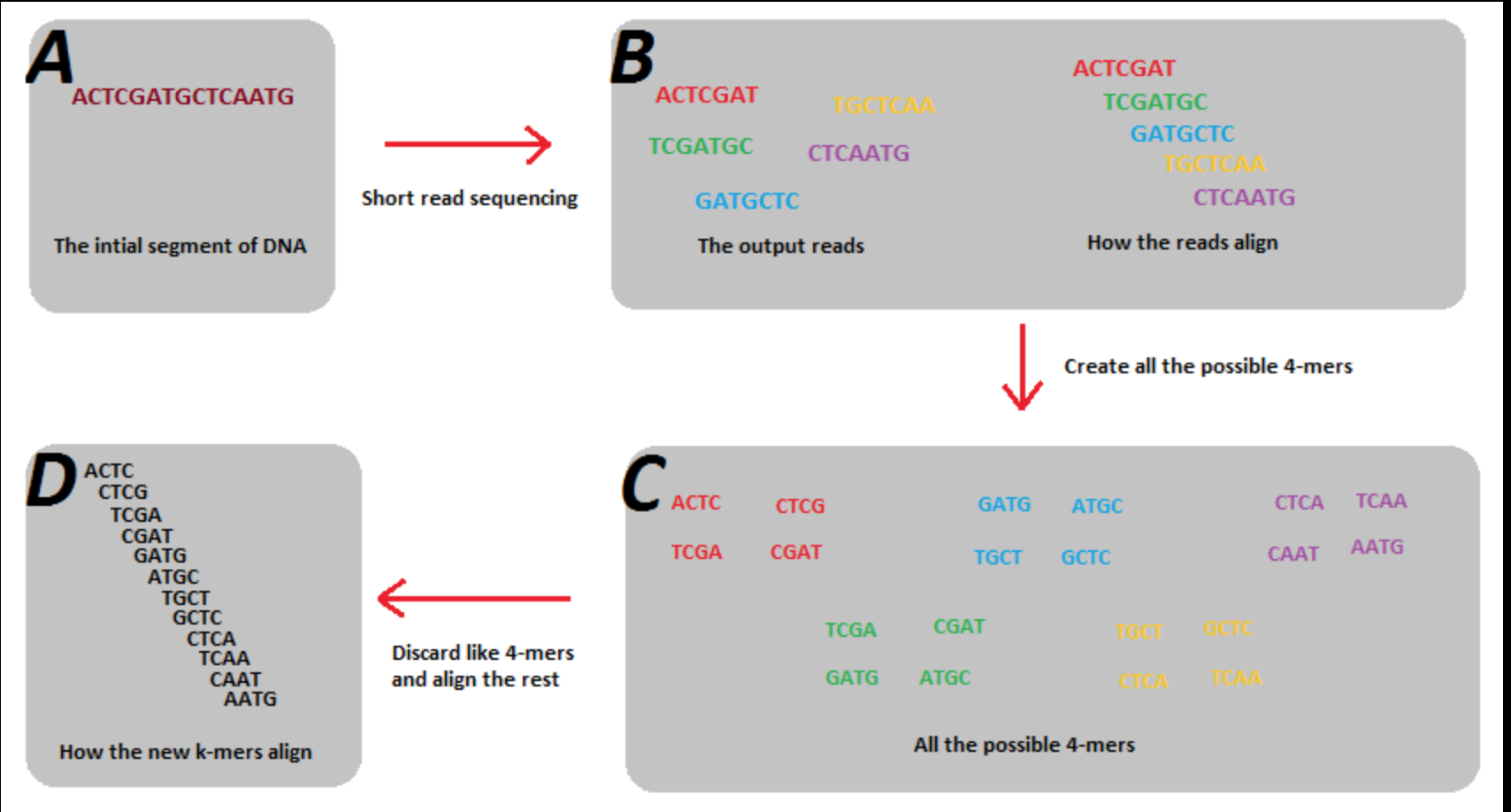
**Methodology**

**supervised learning: logistic regression & sim**

**semi-supervised learning: kmean++& graphcut**

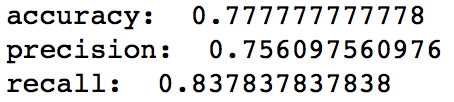
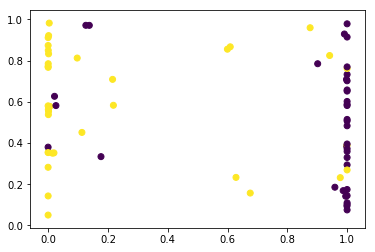
**????? <some introduction about the dimension reduction(two appreaches), new problem settings, algorithms(kmean++&graphcut) of unsupervised/semi-supervised learnning>**

**Features**

We used 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. Since these methods fragment DNA differently, it is possible these methods have biases at where they fragment DNA. In addition different sequencing methods may preferentially sequenced particular subset of the generated DNA fragment. 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. To investigate whether quality scores can be used as a signature of laboratory origin, we calculated the proportion of quality score over all the reads in a sequencing run to use as features.

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. The sequencing reads in each dataset was aligned the reference genome using Bowtie2 (version 2.3.2). We used Illumina’s iGenome *E. coli* MH1655 as the reference genome, which was built in October of 2001. After determining the length of all the sequenced fragments, we calculated the proportion of fragments lengths that fall into fragment length bins. 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. Any fragment length that falls into a bin larger than 1,000 is assigned to the 1,000th bin.



**Machine Learning**

60% train, 20% test, 20% validation

Logistic regression – 97% accuracy

Feature set

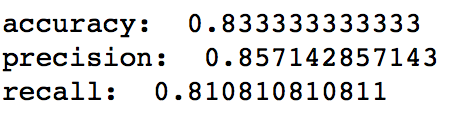
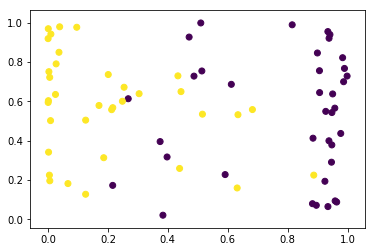
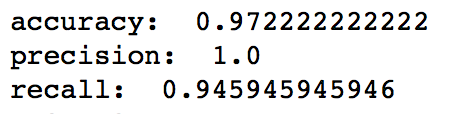
Do you include N as a valid base or throw out

**Results**

1. **logistic regression**：

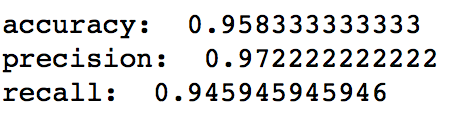
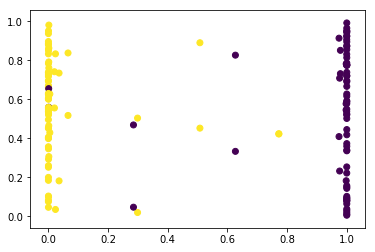
**single feature: quality score distribution:**

**4-mer features:**



**combined features**：

**analysis**：**combined features accuracy is less than just 4-mer features which implies that the quality score distribution feature maybe noise. For unsupervised learning, it may be more reasonable to just consider about 4-mer features.**



1. **svm method with combined features:**

**<Unsupervised learning>**

**?????**

**Future Work**

Additional datasets

Datasets from different experiments from the same laboratory

Datasets with different organisms

Quality scores: Mean, Sd, Full distribution, Average decrease in quality as a function of base

**References**

<http://www.nature.com/nrg/journal/v11/n10/full/nrg2825.html>

<https://academic.oup.com/nar/article/42/21/e161/2903156/svaseq-removing-batch-effects-and-other-unwanted>

<https://www.nature.com/articles/srep39921>

<https://genomemedicine.biomedcentral.com/articles/10.1186/gm208>

<https://cancergenome.nih.gov/researchhighlights/tcgainaction/AkbaniBatchEffectsCaseStudy>

<https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-014-0087-z>