Lab 8: Guess the Animal

Due date: Wednesday, December 8, midnight.

Overview

In this assignment you will use knowledge gained throughout the course to analyze a unique and peculiar dataset and attempt to help on-going research in the field of bioinformatics.

Assignment Preparation

This is a pair programming assignment. I will also allow small teams.

Data

You will be analyzing a single dataset, **PYROPRINTS**. The dataset describes the results of DNA sequencing for 92 *e.coli* bacteria strains.

E.coli bacteria is an important environmental contaminant, which exists in nature in a wide array of different strains. Because some of the strains are pathogenic on one hand, and because *e.coli* is highly pervasive in the environment, on the other hand, identification of *e.coli* strains is an important task for environmental biologists.

e.coli strains used in the creation of our dataset, were grown in the laboratory *from a collection of samples obtained from multiple species.* In particular, the dataset includes *e.coli* strains collected from chickens, cows, humans, pigeons and seagulls. A traditional way to determine if two strains of *e.coli* are the same or different is to sequence the DNA of the *e.coli* strains and compare the DNA sequences.

One of the modern DNA sequencing methods is pyrosequencing. In a nutshell, a pyrosequencing process works as follows. A strand of DNA that we want to sequence is first *amplified* via a specially designed laboratory process, i.e., copied many-many times. The strand copies (all copies of the same DNA sequence) are placed in a well, a special location designed to hold the DNA and the reagents during the sequencing process. A collection of wells is supplied to the pyrosequencing machine.

Each DNA strand can be viewed as a string in an alphabet {A, T, C, G}, where the letters A, T, C and G stand for the four nucleotides. For each of the four nucleotides, the pyrosequencing machine stocks a special reagent, that reacts only with it and with no other nucleotide. The reaction yields emission of light (hence the term pyrosequencing), which can be captured and measured inside the pyrosequencing machine. Only the nucleotides at the "top" of the DNA strand can join the reaction with the appropriate reagent.

The pyrosequencing process works as follows. The reagents are introduced into the well according to a specific dispensation sequence: a preset sequence of A,T, C and G. After each reagent is introduced the emitted light is measured until the reaction stops and the light goes away. The next reagent is introduced then.

Example. Consider a DNA sequence `AATCGGGT`. Consider a pyrosequencing process in which we use the dispensation sequence `ATCGAT`.

The pyrosequencing process will proceed as follows:

1. On step 1, the reagent `T` will be introduced. It will react with the `A` nucleotides at the front of the DNA strand. There are two `A` nucleotide, and so, they will be consumed by the reaction, and the emitted light will be measured.
2. On step 2, the reagent `T` will react with a single nucleotide `T` to produce about half of the light that was produced on step 1.
3. On step 3, the reagent `C` will react with a single nucleotide `C` to produce about the same amount of light as on step 2.
4. On step 4, the reagent `G` will react with three nucleotides `G` to produce about three times the light than on steps 2 and 3.
5. On step 5, the reagent `A` will not be able to react with the front of the remaining DNA sequence (which is `T`), and so (almost) no light will be emitted.
6. On step 6, the reagent `T` will react with a single nucleotide `T` to produce about the same amount of light as on steps 2 and 3.

Pyrosequencing allows biologists to sequence relatively short (up to 100-200 nucleotides long) fragments of DNA *very fast and inexpensively.* Each sequencing process produces a histogram that associates with each position in the dispensation sequence a light intensity. Such a histogram is called a pyrogram or a pyroprint.

The **PYROPRINTS** dataset consists of pyroprints of the **same region of DNA** for 92 *e.coli* strains. All pyroprints were obtained using the same dispensation sequence, and all of them are 104 positions long (i.e., the dispensation sequence contained 104 reagents in it). For each sequence and each position in the dispensation sequence *the exact raw light intensity* as reported by the pyrosequencing equipment is shown in the dataset.

The dataset consists of a single CSV file, `PeakData-92poopisolates.csv`[1]. The first two rows in this file contain column information, rows three and beyond contain individual pyroprints. The format of the data is described below.

- **Row 1.** This row represents the dispensation sequence. The first two columns of this row are empty. The next 104 columns contain one character from the A,T,C,G list each. The character indicates the reagent used in the appropriate position in the dispensation sequence. Note, that the dispensation sequence starts with two dispensations of the `A` reagent. The second dispensation should be redundant - all `A` nucleotides should enter a reaction with the first `A` reagent.

- **Row 2.** This row contains column names. The first column, Well is the unique identifier of each pyroprint. The second column, textsfSource describes the species from which the sequenced sample was collected. There are five different species in the dataset: Chicken, Cow, Human, Pigeon and Seagull. The remaining columns have names `Disp. 1`, `Disp. 2` and so on until `Disp. 104` and represent the positions in the dispensation sequence.

- **Rows 3–94.** Each row contains one pyroprint. The first two columns identify the pyroprint origin and the species, the next 104 columns contain (floating point) light intensity values observed at each position in the dispensation sequence.

Assignment

We are primarily interested in one question:

> Is it possible to classify the 92 sequences in the dataset **by the species of *e.coli* origin**?

Additionally, other questions are of interest as well:

- **Clusters.** The pyrograms given to you have been classified using CLUSTAL, a clustering tool used in computational biology. However, we are interested in seeing if other ways to cluster the given data exist.

[1] I kept the word "poop" in the name of the file to make sure you have no illusions where the data really came from.

- **Interesting patterns.** We are mostly concerned with finding what makes *e.coli* from the same species similar to each other and different from *e.coli* from other species. However, any other interesting patterns you may detect (certain features that persist from species to species; features that appear only in birds, and so on) are useful as well.

To perform your analyses you are allowed:

- to use any software you have written during the course;
- to use any publically available data mining/machine learning/KDD software, as long as you understand the method you are planning on using;
- to develop new software and/or modify any code you created for the lab.
- to use any host language libraries for parsing data.

Deliverables and submission instructions

There is a hardcopy deliverable and a group of softcopy deliverables.

Hardcopy. By the deadline, submit a short report describing:

- What questions you were trying to study.
- What methods you used.
- What software you developed/modified/used.
- What results you obtained.
- Whether results you obtained appear to be interesting/insightful to you and whether, you believe, your results solved the main problem.

Softcopy. Submit the soft copy of your presentation and any code that you wrote in the course of the quarter that you used for analyzing the data. Provide a README file describing how to run the programs to obtain your results.

Submit all electronic deliverables as a single zip of gzipped tar archive (`lab08.zip` or `lab08.tar.gz`). Use the following command

```
$ handin dekhtyar-grader lab08 lab08.<ext>
```