New England Biolabs Project Technical Report

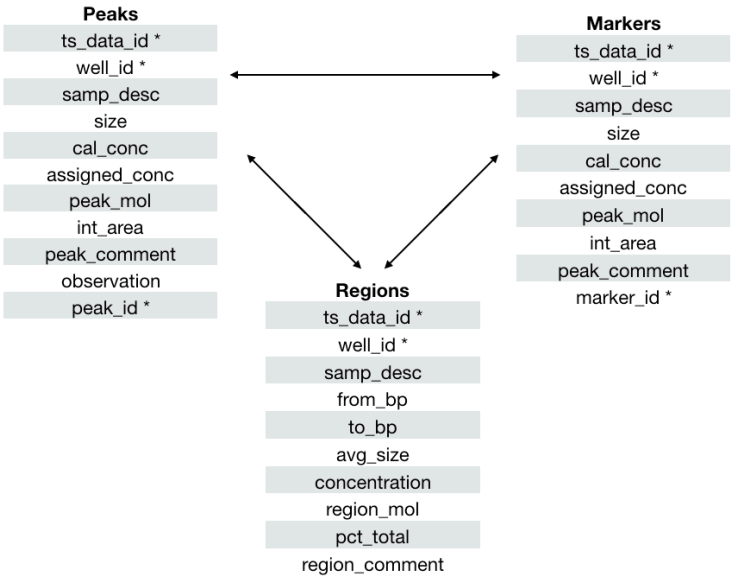
Louis Penafiel (lap279@cornell.edu)

1 Introduction

The dataset is from an Agilent TapeStation to perform automated gel electrophoresis for samples. The main goal is to easily compare different libraries to each other based on size and concentrations. We want to check the variability of the data set to check if the instrument is performing normally.

2 Datastore

A SQL database is chosen for ease of importing new peaks and regions files, as well as ease of access remotely. Since the dataset is not large, we opted to use a sqlite database, so it can be accessed locally. The peaktable files are subdivided into 'peaks' and 'markers' in the sense that the 'peaks' table does not contain any information that originate from synthetic markers. This is so a primary composite key based on the peaks itself. Below is the illustration of the SQL database structure,



where the asterisks corresponds to the primary composite key. In the 'peaks' table, the 'peak_id' column corresponds to the peaks associated with the well, sorted by integrated area in descending order. In the 'markers' table, the 'marker_id' corresponds to either an 'Upper Marker' or 'Lower Marker.'

3 Consistency of strongest ladder peaks

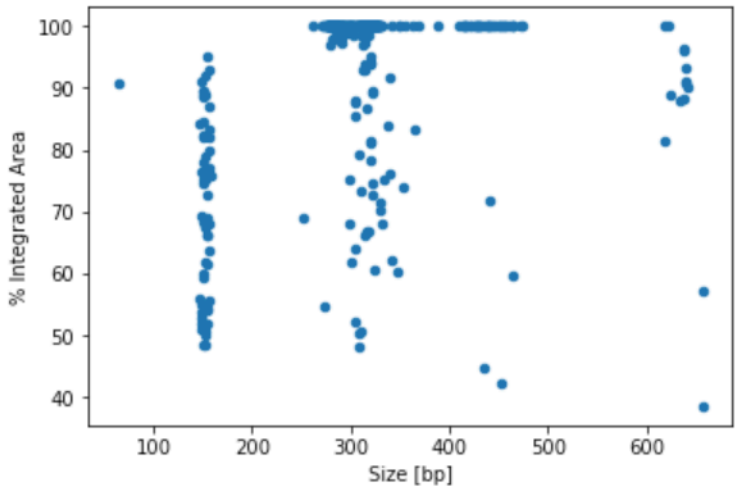
Most of the subsequent analyses are performed using Python with the pandas package.

From looking at the peaks corresponding to the synthetic ladders, they are consistent in size, concentration, peak molarity, and integrated area with no variability.

Note: The peak molarity has a linear relationship with the calibrated concentration.

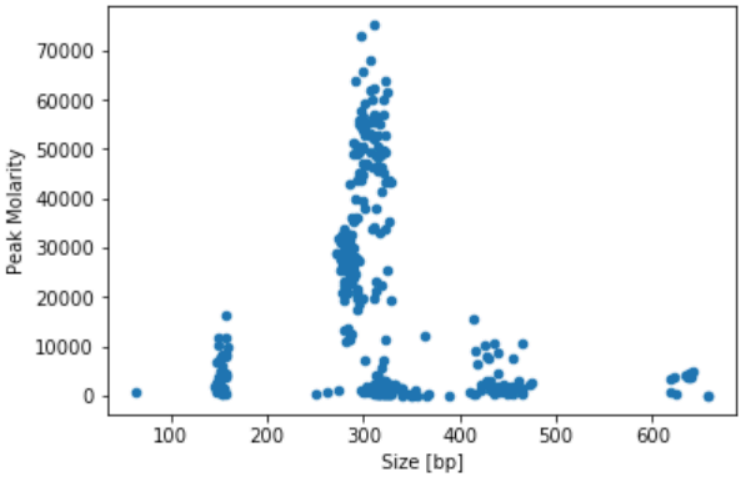
Looking at the peaks from the samples, some samples have multiple peaks. From looking at researchgate questions[1], this could be associated with contamination of samples or issues with the primer or amplification.

Either way, we can filter out those samples by using the % Integrated Area of each peak. The plot below shows the first peak's size against it's % Integrated area.



We find that if we only choose samples with 100% Integrated Area, then the sizes will be in the 300-500 bp range. However, since we want to analyze short fragments ($\sim$ 120 bp), we should look into why there are multiple peaks.

Analyzing the peak molarity shown below and focusing on the neighborhood of the short fragments ($\sim$ 120 bp), we find a large variability with a range of 285 - 16,200 pmol/l, with mean and standard deviation of 3610 ± 3522 pmol/l.

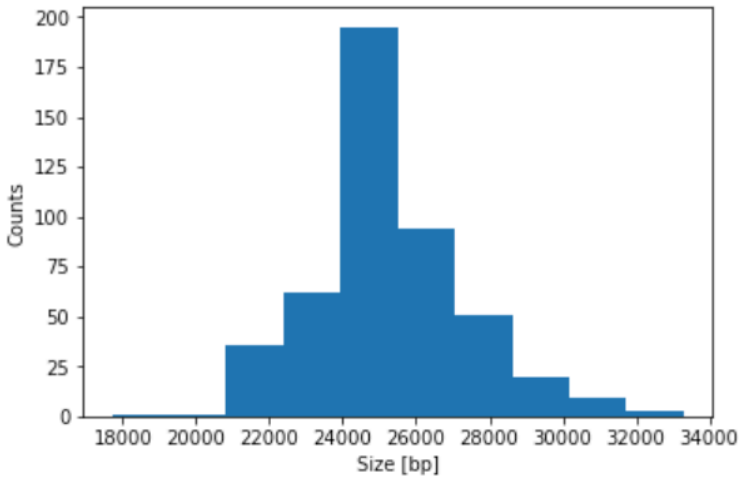


[1] Example 1 and Example 2

4 Consistency of Markers

Starting with the upper markers, their size, calibrated concentration, assigned concentration, and peak molarity are consistent with no variability.

For the lower markers, the size is consistent. However, the calibrated concentration and, hence the peak molarity have some variability, with the peak molarity plotted below.

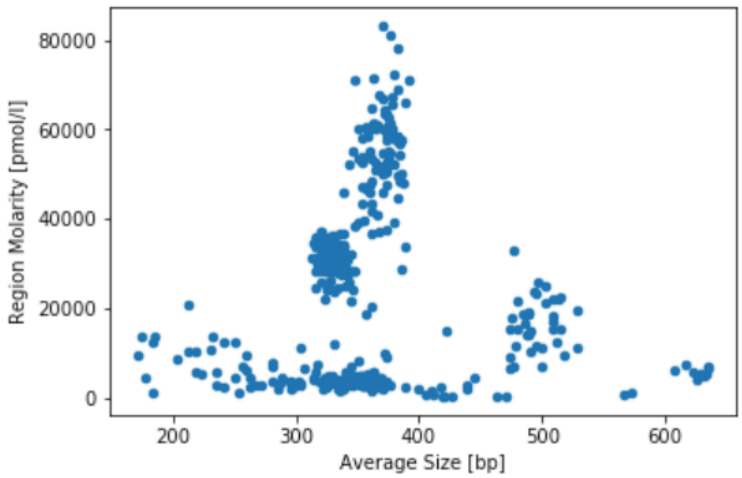


5 Information contained in the other peaks

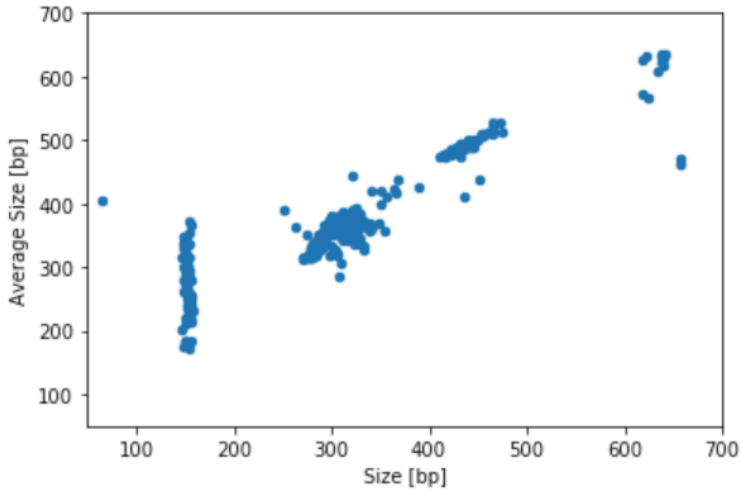
For completeness, the secondary peaks's, a peak that does not have a majority of the % Integrated Area, Integrated Area range from 0.68% all the way to 49.78% with a mean of 16.77%. They also tell us the amount of other peaks there are, which again possibly correspond to sample contamination or issues with primers or amplification. There are some samples that have up to 7 peaks.

6 Regions

We next look at the average sizes and region molarities of the region data, which are plotted below. Almost all of the region data has a region from 100 to 1000 bp. Comparing the below plot to the size vs peak molarity of the strongest peaks plot, there are not any sharp features. The peaks in the plot mainly correspond to the amount and variability in the samples we have.



Furthermore, we can check to see if the peaks overlap with the regions by plotting the size against the average size per well, shown below. From this plot, we see that the average size of the region slightly overestimates the location of the peak. Note that the size, we are plotting is for the strongest peak. Hence, why there is a large variability of the average size for the short fragments ($\sim$ 120 bp) since they have multiple peaks of varying intensities.



7 Interface

7.1 Information

The interface is created using the python package Bokeh, so that the user can interact with the data sets, primarily to filter the data being plotted. The interface is hosted using Flask.

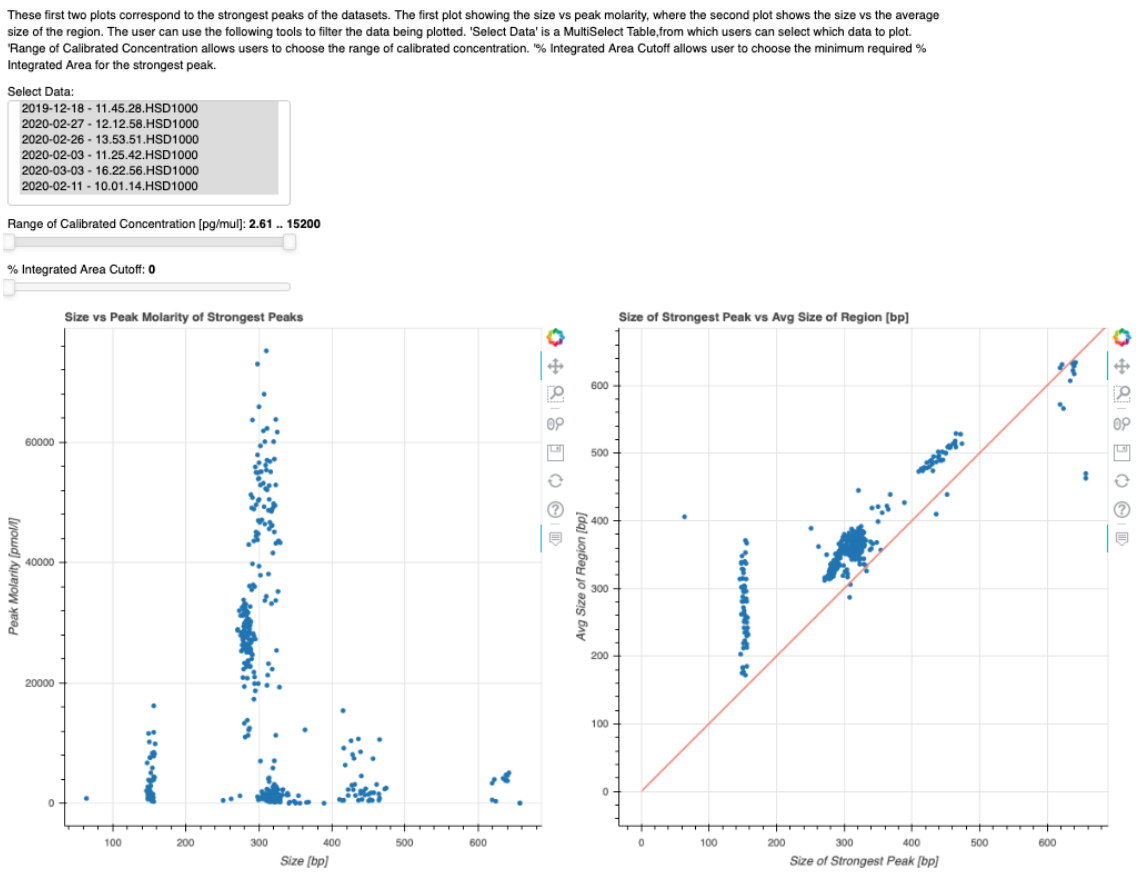
7.2 Instructions to Run

1. The project is located in `/home/ubuntu/project/`. To run the interface on your machine. Go to `/home/ubuntu/project/py_code/`.
2. Then input `python3 neb_interface.py`. This should run a server on port=5000.

3. Now in another terminal, use the following code to access the server, `ssh -L 8000:localhost:5000 ubuntu@52.91.114.198`
4. In a browser input `http://127.0.0.1:8000` then you should be able to see the interface. You can change the localhost port, 8000, to something else if necessary.

7.3 Interface Page

What you should see on the page, should be



As the paragraph on the top of the interface says. These plots show data for the strongest peaks of the data sets. The first plotting its size against its peak molarity, while the second plots the size against the average size of the region. The user can filter the data in multiple ways.

- A MultiSelect Tool in which the user can choose which data sets to plot
- A RangeSlider to choose the range of Calibrated Concentration
- A Slider to determine the threshold for % Integrated Area

Moreover, the HoverTool from bokeh is taken advantage of, so when the user hovers over points. For the first plot, information about that points' Dataset, Well, Size, Peak Molarity, Calibrated Concentration, and % Integrated Area are shown. For the second plot, information about the points' Dataset, Well, Size, and Avg Size are shown.

8 Conclusion

We analyzed the dataset from an Agilent TapeStation. The data is stored in a sqlite database for ease of local access, as well as future data insertion. Most of the analyses is done using Python. The electronic ladders and the upper synthetic markers are consistent with zero variability. The lower synthetic markers have variability in the calibrated concentration. Meanwhile, for the actual peaks themselves, there is a wide range of sizes. Narrowing this down to the short fragments, we still see a large variability in its intensities, due to the presence of other peaks. The regions dataset does not have necessarily sharp features, except those associated with the amount and actual variability of the peaks themselves. Also, there is much overlap with the size of the strongest peak and the average size of the region, except for data that has multiple peaks, e.g. the short fragments. Lastly an interface is created with the Python package Bokeh, to allow users to filter the data plotted, and hosted via Flask.