Py4Life Project

Overview:

The main context of my project is about absorption spectroscopy of DNA nucleotides. In our lab we try to Construct an efficient in vitro expansion methods of DNA repeats, is paramount for understanding the structures and functions of such sequences. In addition, expansions and deletion generally begins with realignment of the primer/template complex leading to slipped-strand mispairing during DNA replication. Studies that used different polymerases in vitro in order to understand the expansion mechanism. Long products of a polyG:polyC and polyA:polyT duplex has been synthesized using Klenow fragment of DNA polymerase 1 by our lab. For this field of research we rely a lot on spectroscopy absorption measurement of DNA nucleotide for concentration evaluating at wavelength 260 nm in particular, as result of the huge reports and exact calculations of molar absorption coefficient factor at this wavelength.

Considering this, it is also important to have efficient methods for processing; parsing and analyzing this absorption spectra data, with this python program can help a lot for faster and better building of the needed primers.

Beer–Lambert law is been use in my program to extract the concentration of the desired nucleotide from its absorption spectra, the law is as follows:

Where, C is the molar concentration of the molecule [M], is the absorption molar coefficient [M-1 cm-1] and L is the path length.

The main goal of the proposed python script is to give a right prediction of the type of given DNA nucleotide absorption spectra by relying on fed reference absorptions spectra.

Data:

Two types of data for the analysis introduced to the program using python:

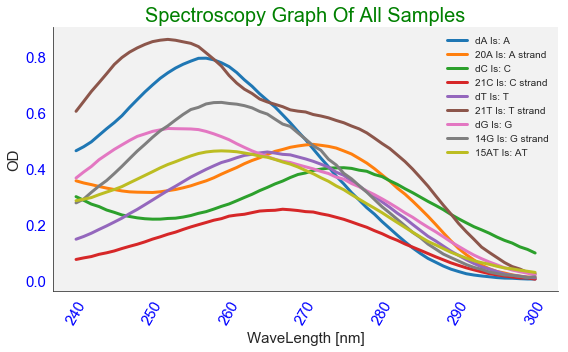
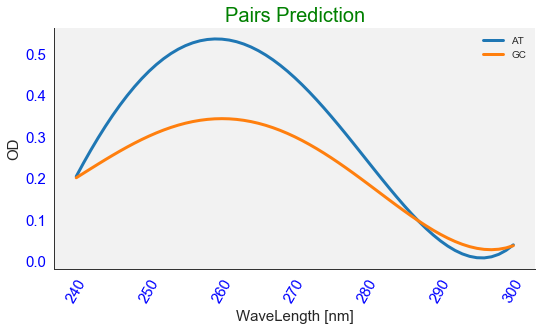
1. References data: which represent our predictive data, that we try to fit our sample to it and compare the similarity to our samples data, in other words to try to find the nearest reference to our sample.
2. Samples data: which represent our true data, the data we are trying to predict and calculating its concentration. In addition, finding the possible pairings of double strand in our data, which can construct polyApolyT or polyGpolyC primer.

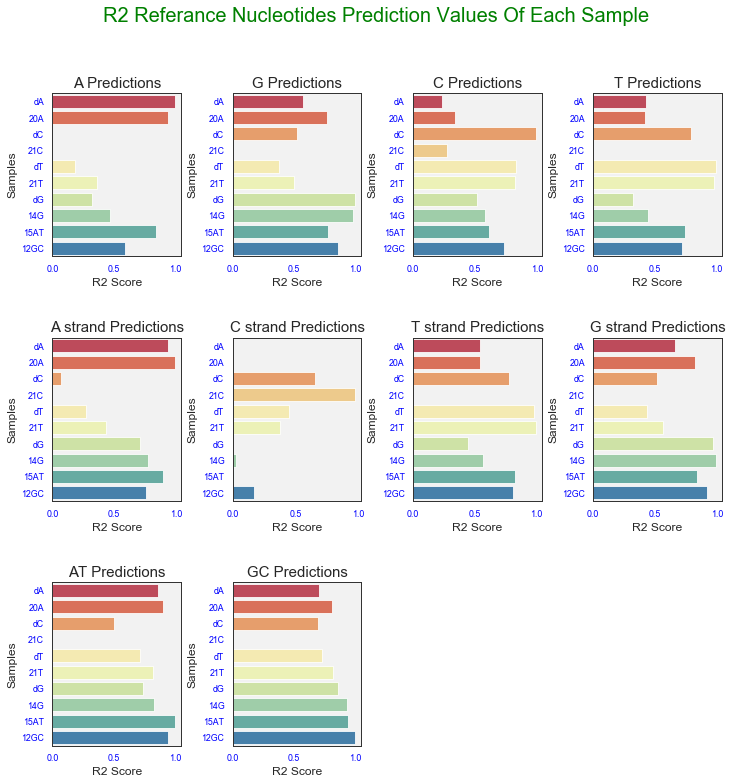
Methods:

* Reading of both the references and samples with pandas read csv.
* Building three function to help speed up the process and have a bitter arrangement:

1. Linear plotting function
2. Polynomial fitting function: which can fit a given continues data (absorption spectra) to Polynomial function and returns the Polynomial coefficients.
3. Base\_ref\_to\_sample\_fit function: gets the desired sample for the comparison to all the references, one for loop that pass on all the references, calculates concentration and R2 score of the polynomial fitting by linear regression prediction between both of the true data and the predicted one from the references fitting. In addition, compressing all the calculation in dictionary and returns it.

* For loop: which is responsible for sending the samples data one by one to polynomial fit function and joining the returned dictionary of the analysis together.
* Converting the analysis dictionary to dataframe.
* Plotting all R2 score of the prediction in seaborn bar plot.
* Finding the firs pairs of nucleotides (AT or GC) in the samples data for constructing polyApolyT or polyGpolyC primer, calculates its volume ratio for getting equivalent concentration in 1 mL volume and plotting the predicted absorption spectra of the product.
* Exporting the analysis dictionary to csv file dataframe to csv function.

Results:



3

2

1

1. Spectroscopy absorption chromatography of all given samples, containing a legend with prediction of each sample.
2. seaborn bar plots: each plot represents one of the reference nucleotides R2 score prediction for each one of the samples nucleotides.
3. Calculating the volume fraction to get equivalent concentration of nucleotides for best primer creation and annealing and plot prediction for the annealing product.