Simue' Rose Isabel

Anna Ritz- Bio 131- Final Project Report

May 10, 2017

Using Excel and Pylab, my final python project visualized data from "The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA" a Forensic Science Institute article by Susan Walsh, et al. Their original data included hair and eye color probability predictions, actual phenotype records, and HIrisPlex accuracy percentages. Since eye color can be qualified into more or less three categories, I focused on manipulating and visualizing that data.

In order for Python to read the data, their original .txt file needed to be converted to a .csv file. From there I isolated the corresponding eye color variables and information. This new file contained only eye color probabilities, actual eye phenotypes, and HIrisPlex predictions. The actual eye phenotypes (Column 3) collected from the subjects had inconsistent language (i.e. blue green, hazel, etc.) To visualize all predictions, I had to narrow the eye color categories to three: blue, brown, and green/intermediate (Column 5).

From there, I could implement a scatterplot in python that would visualize the actual phenotypes (Column 5) in Pylab. We first tried this in lab with a test data, and 4 by 4 figure. Each dot would correspond with an actual phenotype and probability prediction from the HIrisplex system.

I then created a function to pull the information from the excel file. I focused the function on the probabilities of each color prediction, the percent accuracy of the final HIrisPlex prediction, and the newly defined column of actual phenotypes. I did not pull information from the original and inconsistent column of actual phenotypes (Column 3).

Within the function, Mina helped me implement a reader to go through the file and re-categorize the columns into colorized phenotypic probability list. I created a forward loop read the information and add it to the Data list. From there, I created the new probability lists: ProbGreen, ProbBlue, and ProbBrown for the .csv to be converted into colorized python probability lists. I added PhenoList (Column 5), to convert the newly categorized data into a python data list. Each line of information in the .csv file is appended to the new Data list.

I created a forward loop to convert the newly indexed data lists from lists of strings into lists of floats. Each list of floats appends to the corresponding probability list, replacing the strings from the .csv to python list conversion. These floats can be understood and visualized by python into a 6 by 6 Pylab figure. I tried to extend the figure to get a more detailed view of the clumped end data. However, it required me to go beyond a 100 by 100 figure-perhaps it was code but my computer did not like that. The data did not adjust well to the 10 by 10 figure, which still clumped the end data.

My final python code prints list of blue, brown, and green eye color and the list of actual eye color phenotypes. It returns ProbBlue, ProbBrown, and Phenolist. These are called as inputs for the final scatterplot of actual eye color phenotypes (Column 5). The visual results summed to 1, and extend the test model to include the Probability of Green in terms of blue and brown eye color probability. Predictions with a high probability of Blue rarely transform into actual phenotypic expressions of brown or green eye color. Actual phenotypes with high probabilities of brown, express themselves as blue and green more frequently. The frequency of green eye color is present within the population, when probBlue and probBrown are lower. Green eye color appears more frequently when the probability of brown is higher. The large number of clumps could also be a distortion. A visual product of my categorical restrictions on the original list of actual eye color phenotypes (Column 5) into a list of three colors (Column 3).