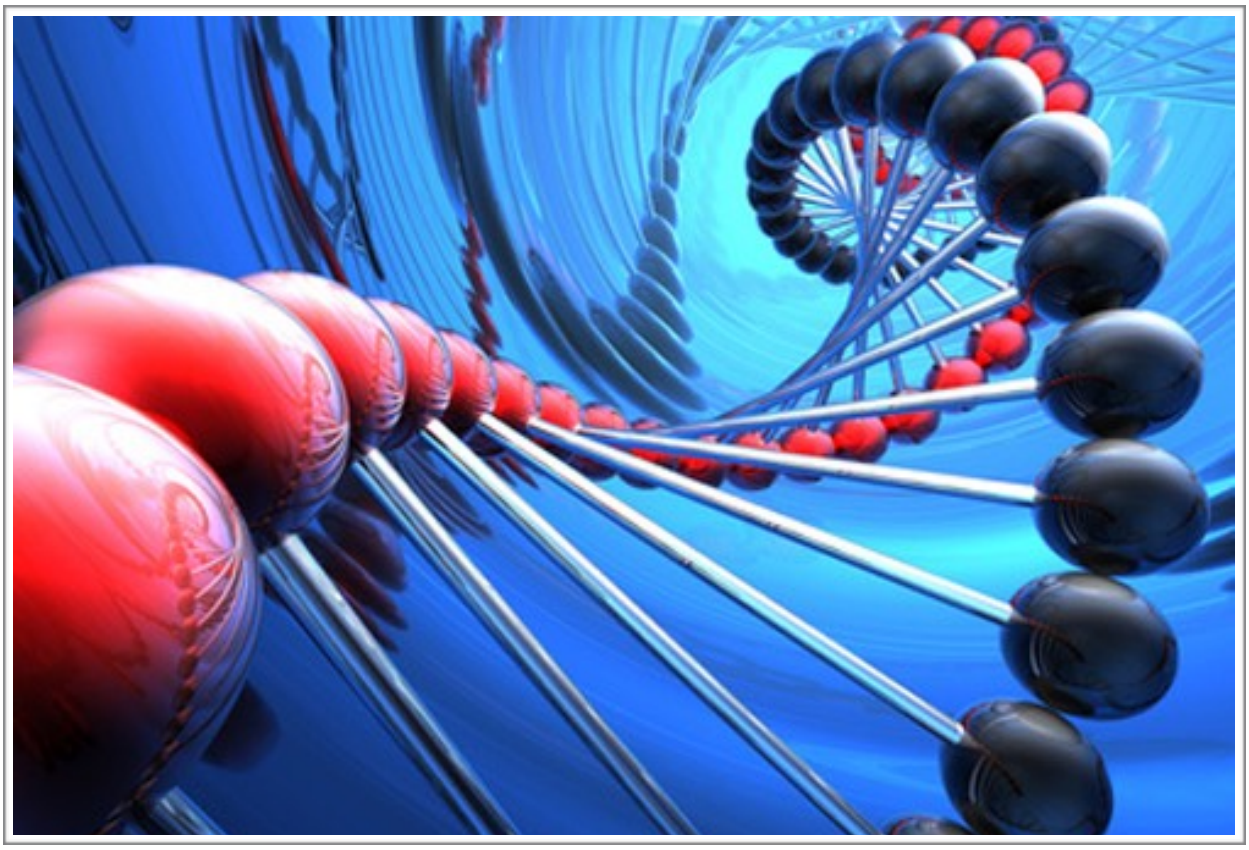
DNA Classification Report

Bio 111, Group #34 DNA Sequencing and Computer Science



Andy He, Andy Tang

Winter 2019, Term 1

Introduction

The program we have written are mostly base on an online GitHub project. We are lucky that we find some data along with this project so that we be able to play around the data and add some of our own stuff. The main purpose of the program is to predict what family does each DNA sequence chunk belong to.

Gene family	Number	Class label
G protein coupled receptors	531	0
Tyrosine kinase	534	1
Tyrosine phosphatase	349	2
Synthetase	672	3
Synthase	711	4
Ion channel	240	5
Transcription factor	1343	6

Gene family

In the picture above, we separate the sample gene data by their family and label them from class 0 to 6. A gene family is a set of several similar genes, formed by duplication of a single original gene, and generally with similar biochemical functions. Yet genes are not exactly identical, therefore we need to classify them

Training Data

As we shown on the picture above, we have sample data that has already been label with the correct label. These are the training data for our machine learning model so that the machine learning model being able to "learn" from these data and predict new data that hasn't been labeled in the future.

Result

As a result, the accuracy for the model being able to reach around 98.4 % for the human gene data. A accuracy of 92% for the chimpanzee gene data.

```
human gene classcification accuracy
accuracy = 0.984

chimp gene classcification accuracy
accuracy = 0.920
```

Play Around the Data

After we have trained our model, we put the chimpanzee data and the dog data into our human gene classifier model. Also, we do the same for the chimpanzee model where we switch the chimpanzee data into human data.

Conclusion

As result, the prediction of human data in a chimpanzee model and vice versa, reach a very high accuracy.

```
gene similarity test
human model vs chimp test data
accuracy = 0.993

chimp model vs human test data
accuracy = 0.934
```

On the other hand, the dog gene data on both model are perform not as good as the other data set.

```
human model vs dog test data
accuracy = 0.926
chimp model vs dog test data
accuracy = 0.910
```

Base on the result, we came up a hypothesis that human's genes and chimpanzee's genes are more similar and their genes have similar biochemical functions. Where on the other hand, dog's genes are slightly different to human's genes and chimpanzee's genes. Yet this result can because of the randomness in data or because the dataset is simply not large enough. At the end, we said that our model are pretty good at classifying genes base on families.