RNA Classification Using Statistical/Machine Learning Models

University of Houston-Downtown

By Chukwudi Yehoshua Iberossi

October 20, 2019

Abstract

The purpose of this dissertation is to examine the random forest that can be used to best predict whether a sequence of Ribonucleic acid (RNA) is from the nucleus or cytosol. The data are two sets of sequences of nucleotides, one was designated as nuclear RNA and the other was cytosol RNA. Since this a supervised learning classification problem, certain models which would best fit the situation were chosen. Before the models could be used on the data, it had to be formatted to be properly used in the model. The model were first made into k-mer of counts 4, 6, 8 where those numbers represented the characters in each predictor, in essence there were three distinct data types that the model was applied on.

Acknowledgement

I cannot express my deepest gratitude to Dr Randy Davila for his guidance and knowledge about Statistical/Machine learning helped in the proper understanding of the algorithms that were used in this project

Dr Patrick King has to be mentioned as his introductory class in Data Science and exposing me to R programming language laid the foundation all other classes that were to follow.

To Dr Dexter Cohoy, his explanations of the Statistical aspects and the notions of building regression models made it easier to comprehend the more complex Machine learning models.

Dr Benjamin Soibam is worth mentioning as well, having provided for me the opportunity to work with him during the summer on the RNA data.

To all the members of the Department of Mathematics and Statistics University of Houston-Downtown, whom I took classes or sought advice from, all have be helpful in some capacity or the other.

Finally I would like to recognise Dr Ryan Pepper, Department Chair of Mathematics University of Houston-Downtown whose vision for a bachelors of Science in Data Science made all this happen.

Contents

Chapter 1

Introduction

1.1 Nucleic Acid (DNA, RNA) and Proteins

Nucleic acids are bio polymers that are used by living organisms to encode information vital for life. In terms of size, being that nucleic acids are polymers, there are large molecules.The chemical composition of a nucleic acid can be broken down into a phosphate, a sugars (5-carbon pentagon shape), and a mixture of organic bases. The organic bases which are known as nitrogenous bases or simply bases come in two forms, purines and pyrimidines. For the purines there are two; Adenine(A) and Guanine(G). Pyrimidines are Cytosine(C), Thymine (T) and Uracil (U). There different due to their molecular structure given that purines have a double ring structure while pyrimidines have a single ring structure. Of most importance is that these bases pair only with certain other bases via hydrogen bonding. Adenine pairs with Thymine(T) in DNA or Uracil(U) in RNA, while Guanine(G) pairs with Cytosine(C.)[7].

Nucleic acids are the main information-carrying molecules of the cell, and, by directing the process of protein synthesis, they determine the inherited characteristics of every living thing [7]. Nucleic acids naturally occur in two forms with respect to living organisms, these are Ribonucleic acid (RNA) and Deoxyribonucleic acid (DNA.) Both DNA and RNA are very similar only with very simple structural differences, yet these differences constitute lots of observable consequences in the form and nature of organisms as well as the functions of these macro-molecules.

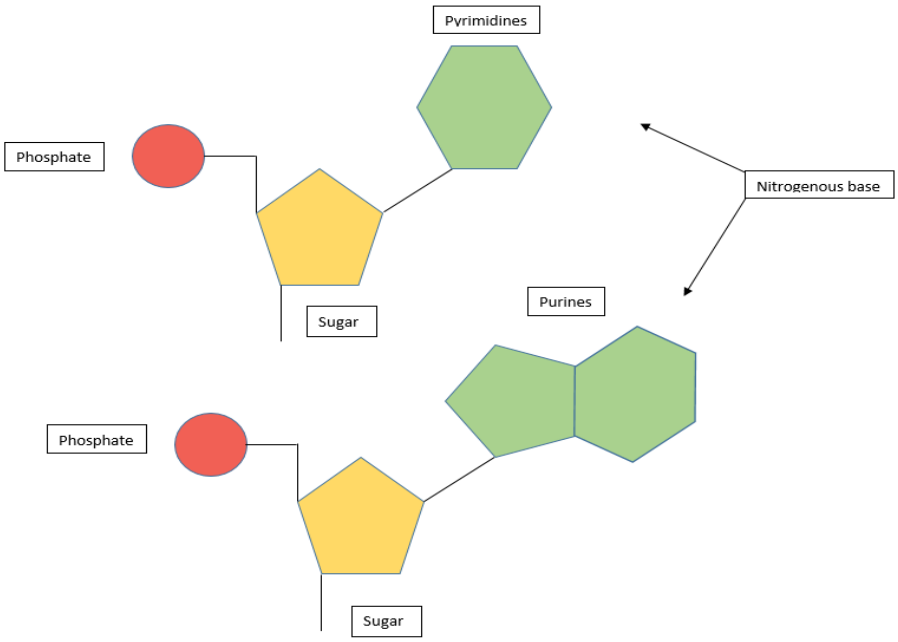

Figure 1.1: Nucleotide:

1.2 DNA

Deoxyribonucleic acid (DNA) consists of a phosphate and sugar backbone, the sugar is ribose but without an oxygen atom at the 2' carbon atom in the ribose pentagon. The is why it has the 'deoxy' indicating it lost an oxygen atom.The nitrogenous bases of DNA are Adenine(A),Guanine(G), Cytosine(C) and Thymine (T). Also noticeable is the exclusion of the nitrogenous base Uracil(U.) With just four bases all the genetic code of an organism can be stored in the form of DNA. so as humans all expressions of what we can physiologically observe including our height, eye color, skin tone, and even genetic mutations that cause defects are encoded in the DNA.

DNA is located in the Nucleus of a cell in Eukaryotes and a nucleoid region in Prokaryotes [5]. In the nucleus it is in a compressed form bounded by proteins and this form are called chromosomes[4]. The chemical structure of DNA is typically in a double helix pattern that was propounded by James Watson and Francis Crick in 1957. However there have been observed viral DNA as a single strand. Even though DNA is responsible for all the coding for Amino acids which are the building blocks of proteins, it cannot leave the nucleus of the cell. Rather it is coded into a single

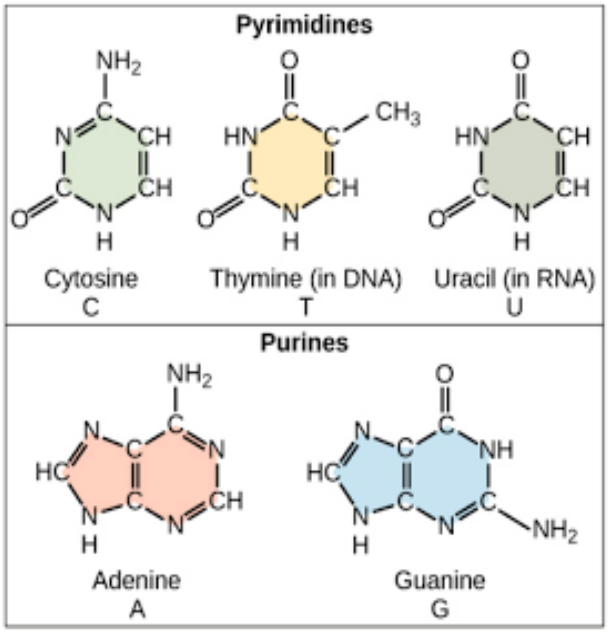

Figure 1.2: Nitrogenous bases
[7]

form called RNA to be released from the cell to the Ribosome (another organelle of a cell) which then builds the proteins based on the instructions in the RNA from the DNA.

1.3 RNA and Proteins

Ribonucleic acid (RNA) similar very similar to the structure of DNA in that it has a phosphate and sugar backbone, the sugar is ribose but there is an oxygen in 2' carbon. the nitrogenous bases Adenine(A), Guanine(G), Cytosine(C) are present but Thymine (T) is replaced by Uracil(U) in RNA. RNA is usually seen as a single stranded but occasionally observed in double stranded form in viruses. Given that RNA seem to be objective specific it is shorter in length compared to DNA which is coiled and wrapped as chromosomes [4]. There are several types of RNA, and these include messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA) just to name a few. mRNA is the encoded information from the nucleus to the Ribosome on how a protein should be synthesized. tRNA in finds amino acids in

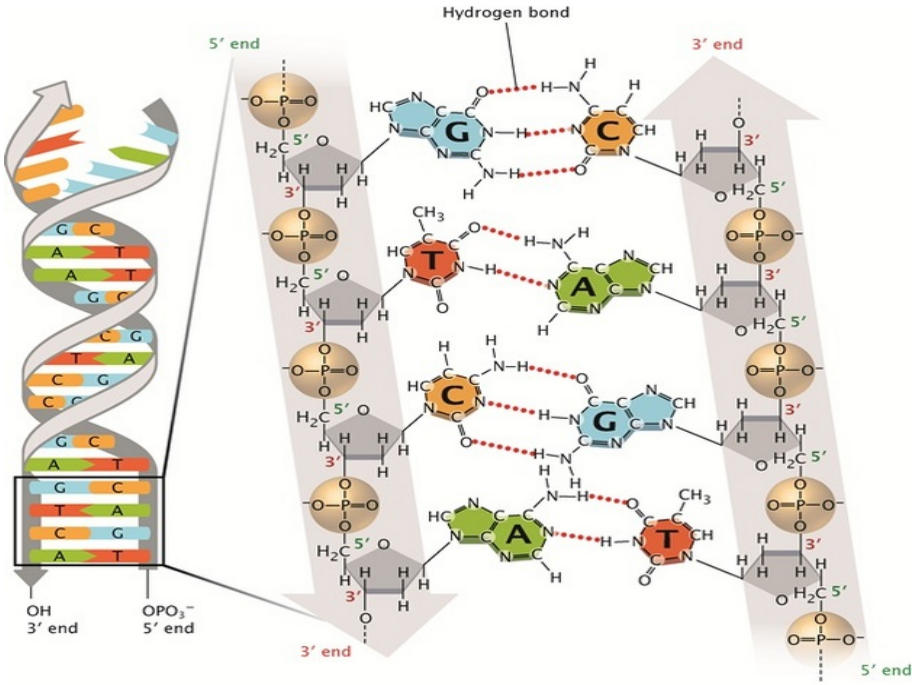

Figure 1.3: Double helical form of DNA
[6]

the cell to be brought to the ribosome so they can be put in the right order to form the proteins. rRNA are coded from portions of DNA called rDNA and then bind with proteins to form the ribosome units which are further brought together to make the Ribosomes. Scientists generally accept that RNA is a precursor to DNA. The focus of this project was on classifying RNA between those that are from the nucleus and those that are from the cytosol or cytoplasm.

Chapter 2

Data

2.1 The Raw RNA data

The data that was used to build the statistical models were two sets designated as nuclear data and cytosol data. This made is easier to attribute the two data sets to certain outputs. Typically nucleic acid data comes in its raw form as a text file with row of sequences. These sequences are a list of the nitrogenous bases abbreviated with their leading letters, so Adenine is A Guanine is G, so are Cytosine C, Thymine T, and Uracil U. The sizes of the files were 4.0 megabytes(4.0MB) for the nuclear data and 4.4 megabytes (4.4MB) for cytosol. For nuclear data there are five thousand nine hundred and fifty 2 lines (5952) while cytosol data was eight thousand five hundred and seventy lines (8570). The programming language used for the whole process was R [2].

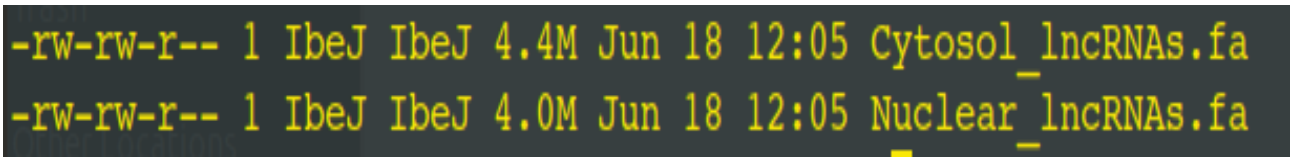

Figure 2.1: Raw Data in megabytes

2.2 Preparing the Data

Exploring both nuclear and cytosol data in its raw form, the Thymine (T) was not yet replaced with Uracil (U), and this was part of the effort to clean up the data. This was achieved by the use of the sed operator on a Linux server. This change would be cosmetic in purpose given that it would not affect the nature of the models since all that changed was a character representation and not something like a deletion which would affect the count. The response variable that will be use for the statistical models had to be built into the data frame. Since this was a classification problem or could be described as qualitative in nature, 1 was chosen as the designation for any data that was nuclear in origin and 0 was given to the all cytosol data. The column of 0s and ones were attached to data frame as the response variable.

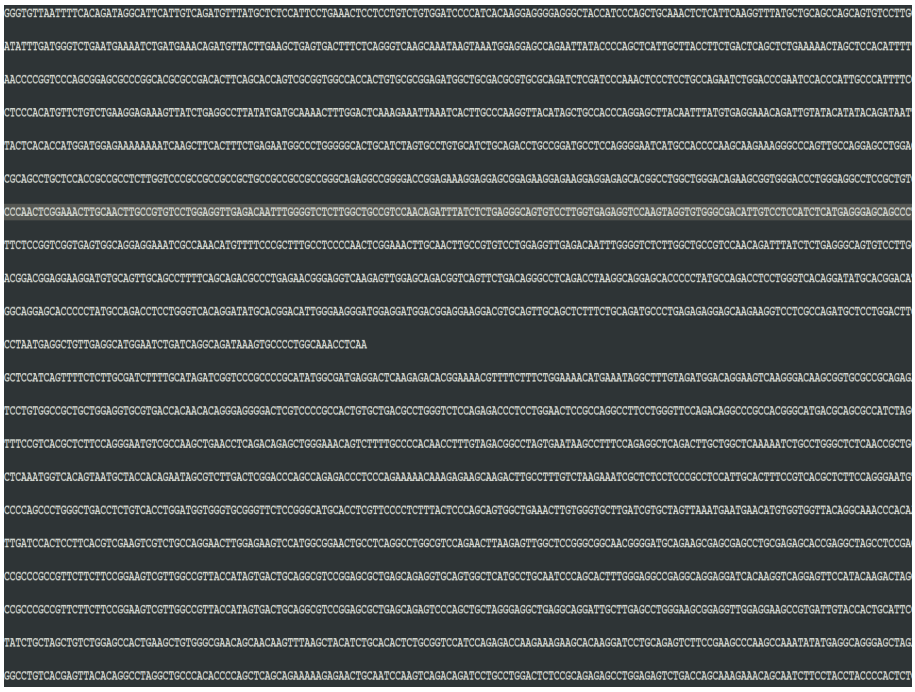

Figure 2.2: Raw data as strings of nucleotides

2.3 k-mer Sequencing

Since the RNA data when first received as raw data is text with rows of strings of nucleotides, example would be AUCCG, it needs to be made into a data frame to be used in statistical models. This data frame should contain elements of thew raw data. An efficient way achieve both of these is the use of k-mer sequencing. k-mer is simply splitting each row into strings of k-count characters. These strings of k amount become the predictors. k-mer of 4-count, 6-count and 8-count were used. The main benefit of k-mer sequencing is to help analysis the data as fixed chunks rather than as a whole string [1]. The construction of the different data frames were done by first using the library in R named Biostrings from the package Boiconductor.

```
#-------------------------------------------------------------------------------#
#            Summer Research Classifying RNA Into Cytosol and Nuclear RNA       #
#-------------------------------------------------------------------------------#

# Start with downloading the module to obtain Biostrings via biocLite
source("http://bioconductor.org/biocLite.R")

#---- Getting Biostrings via biocLite ----#
biocLite("Biostrings")

#---- Running Biostrings library ----#
library('Biostrings')

#---- Reading in Cytosol RNA sample into variable Cytosol ----#
Cytosol = readDNAStringSet('Cytosol_lncRNAs.fa')

#head(Cytosol)

#---- Reading in Nuclear RNA sample into variable Nuclear ----#
Nucleus = readDNAStringSet('Nuclear_lncRNAs.fa')

#head(Nuclear)
```

Figure 2.3: Code using Biostrings library

In the Biostrings package is a function called oligonucleotideFrequency which can take a parameter as an integer. This parameter would be the desired k-mer. Since count of 4 , 6, and 8 were used this function was run three times and attributed to three different variables. After this a column with the corresponding number for the qualitative numbers of 1 and zero were attached to each new predictor variables.

These were then saved as a comma separated value file (csv.) The dimensions for k-mer 4 csv was seven thousand two hundred and fifty seven rows by two hundred and fifty seven columns (7257 x 257 .) For k-mer 6 csv the dimensions were seven thousand two hundred and fifty seven rows by four thousand and ninety seven columns (7257 x 4097 .) And finally k-mer 8 csv was seven thousand two hundred and fifty seven rows by sixty five thousand five hundred and thirty seven columns.

2.4 Classification Problem

The main focus of this project is to establish a relationship between the created k-mers and the outcomes of either nuclear or cytosol RNA which are represented as one (1) and zero (0) respectively. Since these number are just for designation and not necessarily the numbers 1 and 0, there qualitative not quantitative. The problem can then be best described as a classification problem [3]. This relation would be better described as a function given that we would only want one tuple of a row of k-mers to either 1 or 0.

$$f : X \to Y \tag{2.1}$$

where X is a tuple of predictors being mapped to Y which is a set of two elements.

$$X = (x_1, x_2, x_3, \cdots, x_n), x_i \in \mathbf{R} \tag{2.2}$$

$$Y = \{0, 1\} \tag{2.3}$$

$$Y = F(X) + \epsilon \tag{2.4}$$

The function is thus Y as a function of X plus an error terms since there is a change in vector space. ϵ is an error term that is random but has an average or expectation of zero (0) with a constant variance σ^2. Since the error term ϵ has an expectation

of zero Y can be best predicted with just the function on X

$$\hat{Y} = \hat{f}(X) \tag{2.5}$$

$\hat{f}$ is typically described as a black box given that the exact form of $\hat{f}$ as long as accurate results are obtained [3]. The use of ordinary least squares will be employed to access how best $\hat{f}$ predicts the model. This involves how small the error is between the predicted value of $\hat{f}$ and the actual value of Y. These errors are averaged which in turn is called the mean squared error M.S.E.

$$M.S.E = \frac{1}{n}\Sigma_{i=1}^{n}((Y_i - \hat{Y}_i)^2) \tag{2.6}$$

From the equation we can see that if the predicted value is the same then the particular observed error will be zero.

2.5 Splitting the Data

Henceforth the term model will be used to describe the hypothesized function $\hat{f}$. As with all statistical models a training data and was needed to expose model while a testing or validation data which would not have been exposed to the model were both needed. The model trains on the training data trying to best approximate parameters or coefficients for a linear combination of the predictors . Since we had an ample amount of data all the k-mers were split into a seventy (70) percent thirty (30) percent ratio. Seventy (70) percent going to the training data while thirty (30) percent for the validation or testing data.

2.6 Statistical Models

The statistical models used were Random Forests and Neurological networks.Both of these models work with supervised learning, which RNA classification would be considered as. Also the models function both in regression problems and classifica-

tion, which made for their ability be applied to this particular function that is being elucidated. These two models were chosen given the large nature of the data and their tendency be good predictive models.

2.6.1 Random Forests

A random forest is built similarly to a Decision tree. Decision tree is a flow chart type of model that is built like a tree. It has a root node which in turn has branches and finally at the leaves of the tree are the desired outcomes which are the class labels, in this case the leaves should be a choice or 1 for nuclear RNA or 0 for cytosol RNA. Each branch of the tree represents a test on a particular predictor which lead to another test of another predictor. The mode of the observation is what is used to make the decision since this is a classification problem. A random forest is actually a forest of many Decision trees which increases the accuracy [3]. With the Random Forests, each Decision tree in the forest outputs a class of the outcomes which are our 1 for nuclear RNA or 0 for cytosol RNA, thew the outcome that is is the most observed which is still a mode, is the classification that is chosen. The advantage of the Random Forest over the Decision tree is that the Random Forest is significantly less prone to overfitting, and it is very useful for large data like the the RNA data being handled with higher number of dimensions. The central idea is to take weak learning algorithms, combine them to form a strong learning algorithm.

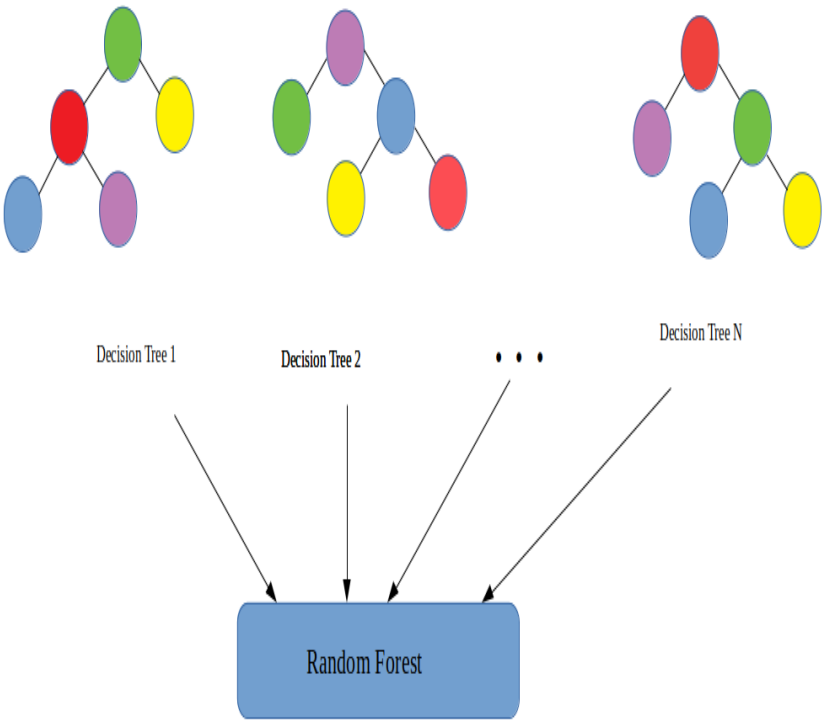

Figure 2.4: Random Forest with N fold trees

Chapter 3

Results and Conclusion

3.1 Random Forests Model

For the k-mer of 4, that is with predictors of 4 characters the random forest model achieved an accuracy of 69.99 percent. For cytosol, 1059 where accurately predicted while 459 were inaccurately predited. Theses results are of those from the test data which the model has not been exposed to. This can be interpreted that it could determine whether a given sequence of RNA data is from the nucleus or cytosol 7 out of 10 times. The error then would be 30 percent. The result also gave a McNemar's test of $2.2e-16$, this tests validates our null hypothesis and we do reject the outcome that the probabilities are equal.

For the k-mer of 6, that is with predictors of 6 characters had an accuracy of about 64.5 percent, this gives an error of 35 percent.Cytosol was predicted accurately was in number of 1106 while 623 were inaccurately predicted. With nucleus 141 were accurately predicted while 286 were inaccurately predicted. As with k-mer 4 this was the test data. This indicated that the k-mer 6 model would predict 65 percent of the time correctly. The McNemar'test was $2.2e-16$ which made for a failure to reject the null hypothesis indicating that the parameters were valid.

For the k-mer of 8, which was 8 predictor characters, the accuracy of 99.86 percent was gotten. This can be considered quite an accurate model. The error is 0.014 percent. This was test data as well. Given that 1 indicates nucleus and 0 the cytosol,

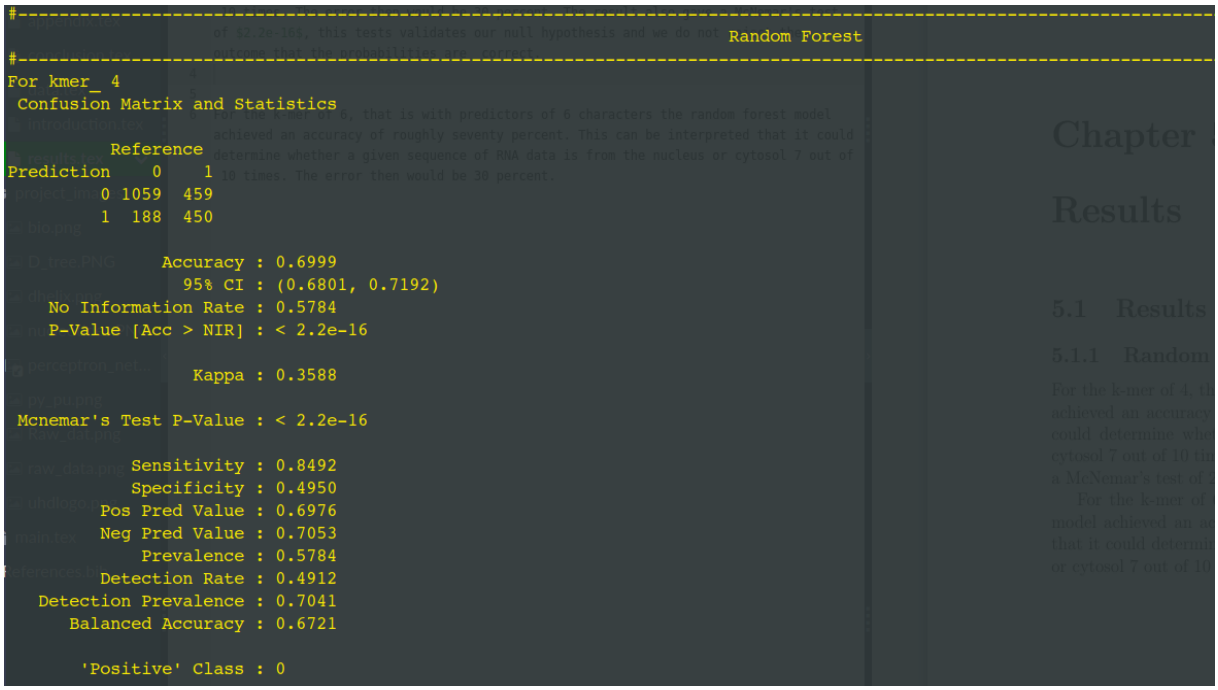

Figure 3.1: Results of Random Forests for k-mer 4

the prediction the results were all 1252 of the cytosol been accurately predicted. 901 out of 904 for the nucleus were accurately predicted. For The McNemar's test was 0.2482 which is still less than 0.50 which validates the null hypothesis so we fail to reject that our parameters are equal.

3.2 Conclusion

The k-mer 4 sequence gave about a 70 percent accuracy it would still be an acceptable model that can be used. The least accurate was the k-mer 6 model while the most accurate that of k-mer 8. k-mer 8 achieved a 99.8 percent accuracy.

```
For kmer_6

Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1106  623
         1  141  286

               Accuracy : 0.6456
                 95% CI : (0.625, 0.6659)
    No Information Rate : 0.5784
    P-Value [Acc > NIR] : 1.03e-10

                  Kappa : 0.2172

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8869
            Specificity : 0.3146
         Pos Pred Value : 0.6397
         Neg Pred Value : 0.6698
             Prevalence : 0.5784
         Detection Rate : 0.5130
   Detection Prevalence : 0.8019
      Balanced Accuracy : 0.6008

       'Positive' Class : 0
```

Figure 3.2: Results of Random Forests for k-mer 6

```
For kmer_8
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1252   0
         1    3  901

               Accuracy : 0.9986
                 95% CI : (0.9959, 0.9997)
    No Information Rate : 0.5821
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9971

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 0.9976
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9967
             Prevalence : 0.5821
         Detection Rate : 0.5807
   Detection Prevalence : 0.5807
      Balanced Accuracy : 0.9988

       'Positive' Class : 0
```

Figure 3.3: Caption

Bibliography

[1] Bernardo Calvigo. k-mer counting, part I: Introduction. *BioInfoLogics*, 2018. URL `https://bioinfologics.github.io/post/2018/09/17/k-mer-counting-part-i-introducti`

[2] Ross Ihaka and Robert Gentleman. The R Project for Statistical Computing. 1993. URL `https://www.r-project.org`.

[3] Gareth James, Daniela Witten, Hastie Trevor, and Robert Tibshrini. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

[4] Ruairi Mackenzie. DNA vs. RNA – 5 Key Differences and Comparison. *Technology Networks Group Genomics Research*, January 2018. URL `https://www.technologynetworks.com/genomics/lists/what-are-the-key-differences-b`

[5] Traci Pedersen. Prokaryotic vs. Eukaryotic Cells: What's the Difference? *Live Science*, July 2019. URL `https://www.livescience.com/65922-prokaryotic-vs-eukaryotic-cells.html`.

[6] Leslie A Pray. Discovery of DNA Structure and Function: Watson and Crick. *Scitable*, 2008. URL `https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-functio`

[7] Richard Roberts J. Nucleic Acid: Chemical Compound. *Encyclopedia Britannica*, October 1998. URL `https://www.britannica.com/science/nucleic-acid`.