Chapter 1

**Introduction**

* 1. **Methylation at CpG**

DNA consists different combination of four types of nucleotides – A, T, G, C. The CpG sites or CG sites are regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along its 5' → 3' direction. Different types of epigenetic modifications can happen on nucleotides of DNA – these are -phosphorylation, acetylation, deacetylation, methylation etc. Among all types of epigenetic modification- methylation is of high importance and methylation at CpG site is of further more importance. It is involvedin processes such as regulation of gene expression, cell differentiation, genomic

imprinting, X-chromosome inactivation, transposon silencing and chromosomal stability [1].

Aberrant methylation patterns have been shown to be associated with a growing number of

conditions and disease, in particular cancer [2].

**1.2 Motivation**

By predicting the missing methylation values for CpGs, we can at first fulfill the CpG methylation data for all of the CpGs in human. Thus we can build a model using machine learning or deep learning for normal and cancer methylation dataset. That dataset will be helpful to early detection of disease condition- to say specifically- cancer. This will only need dataset of DNA methylation from the patient. Not only cancer this type of analysis would be helpful to determine a large array of diseases as most of the disease is related to abnormal protein production rate.

**1.3 Book Overview**

In Introduction, we spoke about why do we want to work on this topic, discussed our whole theme briefly. We had to learn about a lot of things clearly for conducting the study in which we had no idea. For example: Different operating system( Linux), biological facts, KNN etc. In the section- background study of our book, we tried to give some basic information on those. In Methodology, we tried to mention our whole working procedure briefly including data source and their pre-processing. In Challenges, we mentioned those interruption which we faced.InResultWe have shown some results of our project with necessary images. Accuracy for those results is shown too. In Conclusion and Future Work, wesummed up our whole project and told what we want to do with this project in future.

Chapter 2

**Background Study**

Our background study can be broadly divided into two parts-

**2.1 Biological facts**

**2.1.1CpG and CpG island**

In the introduction part, we have already described what CpG is! CpG islands (or CG islands) are regions with a high frequency of CpG sites. To say more specifically, this is a region with at least 200 bp, a GC percentage greater than 50%, and an observed-to-expected CpG ratio greater than 60%.

**2.1.2 Promoter and methylation of CpG sites**

DNA methylation is a biological process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence. When located in a gene promoter, DNA methylation typically acts to repress gene transcription. In mammals, DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, repression of transposable elements, aging, and carcinogenesis.

Two of DNA's four bases, cytosine and adenine, can be methylated. Cytosine methylation is widespread in both eukaryotes and prokaryotes.In mammalian cells, DNA methylation occurs mainly at the C5 position of CpG dinucleotides and is carried out by two general classes of enzymatic activities – maintenance methylation and de novo methylation.[60]

Maintenance methylation activity is necessary to preserve DNA methylation after every cellular DNA replication cycle. Without the DNA methyltransferase (DNMT), the replication machinery itself would produce daughter strands that are unmethylated and, over time, would lead to passive demethylation. DNMT1 is the proposed maintenance methyltransferase that is responsible for copying DNA methylation patterns to the daughter strands during DNA replication.



Figure 2.1: methylation of cytosine

In humans, about 70% of promoters located near the transcription start site of a gene (proximal promoters) contain a CpG island.Hypermethylation of a promoter leads to transcription repression of the associated gene and hypomethylation leads transcription activation. So, changing the methylation status means an altered condition of our body. Most of the time, cancer is an outcome of this altered condition.

**2.1.3 WGBS(whole genome bisulfate sequencing)& β value of methylation**

Bisulfite sequencing is the gold-standard for measuring methylation over the genomes of interest.There are roughly 28 million CpGs in the human genome, 60–80% are generally methylated. Less than 10% of CpGs occur in CG-dense regions that are termed CpG islands in the human genome.

Whole genome bisulfite sequencing (WGBS) is considered the ‘gold standard' for assaying DNA methylation because it provides global coverage at single-base resolution. Briefly, it combines bisulfite conversion of DNA molecules with high-throughput sequencing. To perform WGBS, the genomic DNA is first randomly fragmented to the desired size (200 bp). The fragmented DNA is converted into a sequencing library by ligation to adaptors that contain 5mCs. The sequence library is then treated with bisulfite. This treatment effectively converts unmethylatedcytosines to uracil. After amplifying the library treated with bisulfite by PCR, it is sequenced using high-throughput sequencing. After the PCR, uracils will be represented as thymines. A precise recall of cytosine methylation requires not only sufficient sequencing depth, but also strongly depends on the quality of bisulfite conversion and library amplification. The benefit of this shotgun approach is that it typically reaches coverage of over 90% of the CpGs in the human genome in unbiased representation. It allows identification of non-CG methylation as well as identification of partially methylated domains and regions at distal regulatory elements with low methylation and DNA methylation valleys (DMVs) in embryonic stem cells. Despite its advantages, WGBS remains the most expensive technique and standard library prep requires relatively large quantities of DNA (100ng–5 ug); as such, it is usually not applied to large numbers of samples. To achieve high sensitivity in detecting methylation differences between samples, high sequencing depth is required which leads to significant increase in sequencing cost.

Since BS-seq changes unmethylatedcytosines (C) to thymines (T), subsequent analysis steps focus on counting the number of C to T conversions and quantifying the methylation proportion per base. This is simply done by identifying C-to-T conversions in the aligned reads and dividing number of Cs by the sum of Ts and Cs for each cytosine in the genome. [3] This proportion is then expressed as β value. So this can take up any fractional number between 0 to 1 where, 0means fully un-methylatedCpG and 1 means fully methylated.

Chapter 3

**Data Collection and Preprocessing**

In this chapter, we have tried to assemble all the information about data collection and pre-processing to make it suible for the model run.

**3.1 Data source**

The whole genome bisulfite sequencing (WGBS) dataset was generated under the Centre for Epigenomics Mapping Technologies (CEMT), Vancouver, Canada [www.epigenomes.c]. For WGBS, genomic DNA was collected from germinal center B-cells (GCB) that expressed CD19, IgD and CD38 cell surface markers. The GCB cells were collected from from Tonsil of a 3 year old male individual. The datasets were processed using the standard protocol described here [http://www.epigenomes.ca/protocols-and-standards].

We collected two files-

1. Known data file- This contains information about those CpGs of the autosomal chromosomes whose methylation status is determined by WGBS. The information provided here are- CpG location on chromosome(start and end position), number of bi-sulfite converted reads (un-methylated CpG number), number of bi-sulfite unconverted reads (methylated CpG number) and the β value (ratio of methylated and unmethylated reads).
2. Known and unknown data file- This contains information about the total CpGs of autosomal chromosomes whether the methylation status is known or not. This file only includes the start and end position of the CpGs.

**3.2 Data prossesing**

This can be divided into multiple stages-

**3.2.1 Separating data of Chromosome 15 from whole dataset**

Using Linux command-

Less input\_file\_name | grep “chr15” | output\_file\_name

We separated the data of Chromosome 15 from the known data file.

We also did the same for the Known and Unknown data file.

**3.2.2 Merging the two output data files**

Then using another Linux command we merged the two output files.

**3.2.3 Sorting the merged file**

The merge file was then sorted depending on the ascending order of the positions of the CpGs. This was done in Libre Office in Linux environment.

**3.2.4 Removing the duplicate rows**

Using python command in linux environment, at last we removed the duplicated rows that were created as a result of merging the known data file with known and unknown data file.

Thus the data was at last ready for the use to run a model with it.

Chapter 4

**Methodology**

In this chapter we have tried to describe our methodology and implementation. We have collected data with methylation level known and unknown data as mentioned in our data collection process before. For predicting beta value we have implemented supervised machine learning approaches like Linear Regression and K Nearest Neighbor Regression (KNN Regression) model.

**4.1 Known Data Implementation with Linear Regression**

First we have implemented Linear Regression on known methylation dataset to come up with a relation between methylated and un-methylated values and beta values. Before implementing we have analyzed our methylated data sequence.

From the dataset we have chosen only number of bisulfite converted reads column and number of bisulfite un-converted reads column for training and testing data as dependent variable.

We have select 855016 data for linear regression to estimate the relation. For training purpose we have select 70% data and 30% data for testing purpose.

We plot a scatter diagram between real value and test value which shows a linear relation between them.

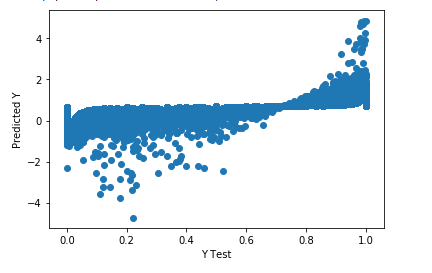


Figure 4.1: Scatter Diagram of Linear Regression

Then we plot a histogram of the residuals and make sure it looks normally distributed.

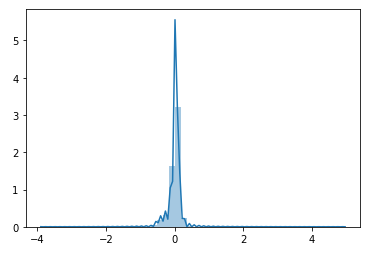


Figure 4.2: Histogram of Linear Regression

We plot a scatter diagram between methylated value and beta value.

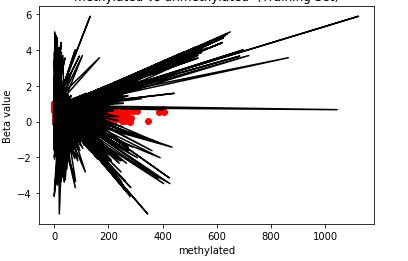


Figure 4.3: Methylated value vs Beta value Scatter Diagram

Here is the heat map which shows the relation between the features of the dataset.

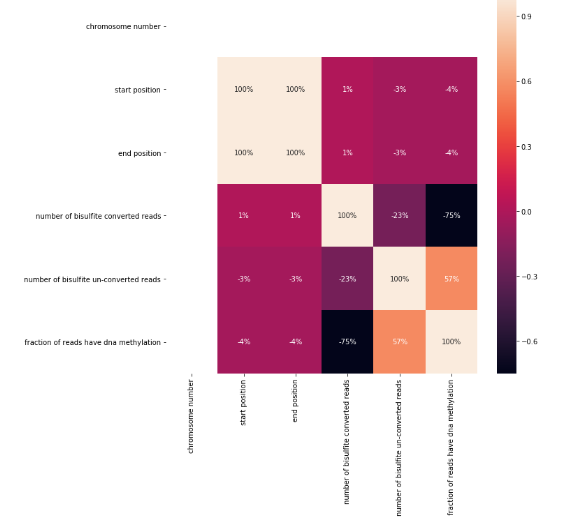


Figure 4.4: Heat map for known data

**4.2 Known Data Implementation with KNN**

Since we want to predict missing methylation data, we need to use those method which can predict missing values from related features of neighbors. KNN is one of them. Therefor we first build a method using KNN to predict the methylation level.

We have select 855016 data for KNN and select K value 3 to estimate the methylation level. For training purpose we have select 60% data, 20% data for testing purpose and 20% for validation.

We plot a scatter diagram between real value and predicted value from KNN model for known dataset.

Then we plot a histogram of the predicted Beta value and real beta value to realize the accuracy.

**4.3 Known and Unknown Data Implementation with KNN**

Finally we load model with a dataset which have both known methylation level and unknown methylation level. We use there KNN algorithm to predict unknown methylation values.

From the known and unknown mixed up dataset we have chosen 897765 data for training and testing data where 855016 are methylation known data and rest of the data are methylation unknown data. For training purpose we have select 67% data and 33% data for testing purpose.

Chapter 5

**Results**

The analysis of DNA methylation has become of considerable interest in biomedical research, as epigenetic studies have shown numerous associations between methylation levels and diseases such as cancer and cardiovascular disease [11–15].

**5.1 Known Data with KNN**

Using KNN on methylation known dataset we get 73% precision in our model. Here we see that our prediction method performance is lower, it because we could not implement our procedure properly. We need further testing process for better performance.

The errors is less.



Here the scatter diagram between real value and predicted value from KNN model for known dataset.

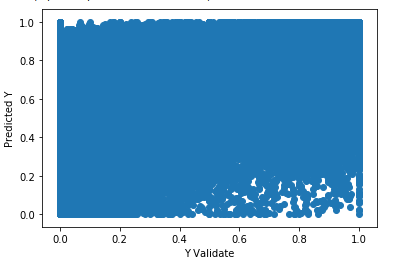


Figure 5.1: Scatter Diagram with KNN for Known Dataset

Here a histogram of the predicted Beta value and real beta value to realize the accuracy.

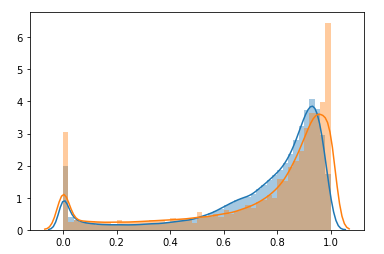
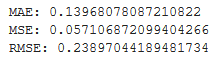


Figure 5.2: Histogram of predicted Beta value and real beta value

**5.2 Known and Unknown Data with KNN**

Using KNN on methylation known and unknown dataset we get 90% precision in our model. And errors



Here the scatter diagram between real value and predicted value from KNN model for known and unknown dataset.

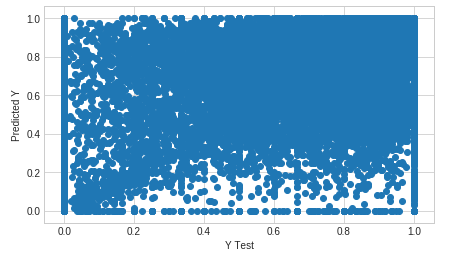


Figure 5.3: Scatter Diagram with KNN for Known and Unknown Dataset

Here a histogram of the predicted Beta value and real beta value to realize the accuracy.

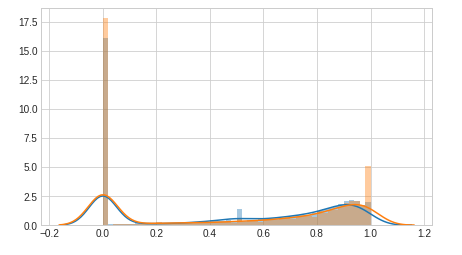


Figure 5.4: Histogram of predicted Beta value and real beta value

Chapter 6

**Limitations**

As the total CpG number in the whole autosomal chromosome is too high to load with the limited capacity of our computer, we couldn’t work with the whole chromosome & decided to work with only chromosome 15.

As we had to learn a lot of new things to conduct the project, we thought to keep it simple in the first place. That’s why we couldn’t consider enough important feature to get more accurate prediction and we couldn’t analyze the predicted results with enough graphical representation.

Chapter 7

**Future Work**

Through the future work, we will try to overcome delimit our current limitations and improve it further to turn it into its best useful form. In future, we will consider the features- such as- topological neighbor, CpG island specific methylation analysis, promoter specific methylation analysis.

We will try to work with the whole autosomal chromosome CpGs. We will also work with the sex chromosomes that have a distinct methylation pattern from the autosomal pattern.

We will also improve our data visualization techniques to make the informations more useful.

Chapter 8

**Challenges**

Learning a lots of new things and combining all the learning outcome was the greatest challenge for us. We had to be clear about different biological issues, statistical terms and varieties of machine learning and deep learning approaches. Learning different approaches helped us to find out the perfect model that is to be done for our target. Again, we had to work on a new operating system. We are comfortable with using WINDOWS – as operating system but we had to work on LINUX- that was completely new to us. Having some limitations we finally had some outcome overcoming all the challenges.

**References**

[1] PhD, Alexei Gratchev. "Review on DNA Methylation". [www.methods.info](http://www.methods.info).

[2]KatarzynaWreczycka, AlexanderGosdschan, DilmuratYusuf, BjörnGrüning, YassenAssenov and AltunaAkalin “**Strategies for analyzing bisulfite sequencing data”**

10.1016/j.jbiotec**,** Journal of Biotechnology2017

[11] Bird A. DNA methylation patterns and epigenetic memory. Gene Dev. 2002;16:6–21.

[12] Shames DS, Minna JD, Gazdar AF. DNA methylation in health, disease, and cancer. Curr Mol Med. 2007;7:85–102.

[13] Sarkar FH. Epigenetics and Cancer. Dordrecht: Springer; 2013.

[14] Varley K, Gertz J, Bowling K, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013;23:555–67.

[15] Zhang N, Wu HJ, Zhang W, et al. Predicting tumor purity from methylation microarray data. Bioinformatics. 2015;31:3401–5.