**Acknowledgements**

At first thanks to Almighty Allah. Then we would first like to thank our thesis advisor Ms Ayesha Tasnim, Assistant Professor of the department of Computer Science & Engineering at Shahjalal University of Science and Technology. We are grateful to our respected teachers of Department of CSE for their inspiration and support. We also want to give thanks to all the senior and junior of CSE, SUST.

**Abstract**

Methylation on CpG is of a great importance for the analysis of various development and disease condition. But, 80-90% of the total CpG methylation status is determined by WGBS whereas other methods cover even less CpGs. Again, the methylation of a given CpG site is affected by the methylation of the surrounding CpG sites-i.e. by methylation status of the sequential neighbors. First we have chosen chromosome 15 out of the whole genome and used this fact of co-methylation to run KNN, a machine learning approach to find out the missing values of CpGs. KNN can be used for data that are continuous, discrete, ordinal and categorical which makes it particularly useful for dealing with all kind of missing data. Our dataset deals with discrete data and affected by nearest neighbors that's why we chose KNN to find out the missing values

**Table of Contents**

Acknowledgements..............................................................................................................i

Abstract................................................................................................................................ii

Table of contents..................................................................................................................iii

List of Figure........................................................................................................................v

Chapter 1: Introduction........................................................................................................1

1.1 Methylation at CpG............................................................................................1

1.2 Motivation..........................................................................................................3

1.3 Book Overview..................................................................................................4

Chapter 2: Background Study..............................................................................................5

2.1 Biological facts…………..................................................................................5

2.1.1 CpG and CpG Island…......................................................................6

2.1.2 Promoter and Methylation of CpG Sites............................................6

2.1.3WGBS (whole genome bisulfate sequencing) & β value of methylation………………………………………………………………..8

Chapter 3: Data Collection and Preprocessing……..…………………………………….21

Chapter 4: Methodology………………………………………………………………….21

4.1 Known Data Implementation with Linear Regression….…………………….21

4.2 Known Data Implementation with KNN………………..…………………….22

4.3 Known and Unknown Data Implementation with KNN……………………...22

Chapter 5: Results…………………………………………………………………………31

5.1 Known Data with KNN……………………………………………………….22

5.3 Known and Unknown Data with KNN…………………………….………...22

Chapter 5: Limitations........................................................................................................34

Chapter 6: Future Work .....................................................................................................35

Chapter 7: Conclusion.........................................................................................................34

References………………………………..………………………………………………36

**List of Figure**

Figure 2.1 Methylation of Cytosine…...…………………………………………………….07

Figure 4.1: Scatter Diagram of Linear Regression……….…………………………………07

Figure 4.2: Histogram of Linear Regression………………………………….…………….07

Figure 4.3: Methylated value vs Beta value Scatter Diagram………………………………09

Figure 4.4: Heat map for known data………………………………………………………09

Figure 5.1: Scatter Diagram with KNN for Known Dataset……………………………….10

Figure 5.2: Histogram of predicted Beta value and real beta value………………………..12

Figure 5.3: Scatter Diagram with KNN for Known and Unknown Dataset……………….12

Figure 5.4: Histogram of predicted Beta value and real beta value………………………..13