INVESTIGATION OF STRUCTURAL PROPERTIES OF METHYLATED HUMAN PROMOTER REGIONS IN TERMS OF DNA HELICAL RISE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BURCU YALDIZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN
BIOINFORMATICS

AUGUST 2014

INVESTIGATION OF STRUCTURAL PROPERTIES OF METHYLATED HUMAN PROMOTER REGIONS IN TERMS OF DNA HELICAL RISE

submitted by **Burcu YALDIZ** in partial fulfillment of the requirements for the degree of **Master of Science, Bioinformatics Program, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, Informatics Institute

Assist. Prof. Dr. Yeşim Aydın Son _____
Head of Department, Health Informatics, METU

Assist. Prof. Dr. Yeşim Aydın Son
Supervisor, Health Informatics, METU _____

Examining Committee Members:

Assoc. Prof. Dr. Tolga Can _____
METU, CENG

Assist. Prof. Dr. Yeşim Aydın Son _____
METU, Health Informatics

Assist. Prof. Dr. Aybar Can Acar _____
METU, Health Informatics

Assist. Prof. Dr. Özlen Konu _____
Bilkent University, Molecular Biology and Genetics

Assoc. Prof. Dr. Çağdaş D. Son _____
METU, Biology

Date: 27.08.2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Burcu Yaldız

Signature :

ABSTRACT

INVESTIGATION OF STRUCTURAL PROPERTIES OF METHYLATED HUMAN PROMOTER REGIONS IN TERMS OF DNA HELICAL RISE

Yaldız, Burcu
M.Sc. Bioinformatics Program
Advisor: Assist. Prof. Dr. Yeşim Aydın Son

August 2014, 60 pages

The infamous double helix structure of DNA was assumed to be a rigid, uniformly observed structure throughout the genomic DNA. However, the differences in physical structure of DNA in terms of local helical parameters such as twist, tilt, roll, rise and angles between adjacent base pairs in B-DNA molecule have been shown in many studies. This observed flexibility satisfies the known physical and chemical properties of DNA while providing a better model to explain how DNA fulfills its biological functions. While the relation between human promoters' methylation status and gene expression profiles in certain cancer types has been established in various studies, the structural properties of methylated promoters were rarely investigated. In this study our goal is to investigate the structural differences between human promoters due to methylation status and gene expression profiles in terms of sequence dependent DNA helical rise. The resulting structural differences have the potential to facilitate further studies to predict the methylation status of human promoters across the whole genome for the investigation of clinically relevant biomarkers in cancer.

Keywords: Promoter methylation, nucleosome occupancy, gene silencing, helical rise, DNA structure

ÖZ

METİLASYON GÖSTEREN İNSAN PROMOTÖR BÖLGELERİNİN YAPISININ DNA HELİKSEL YÜKSEKLİĞİ AÇISINDAN İNCELENMESİ

Yaldız, Burcu
Yüksek Lisans, Biyoenformatik Programı
Tez Danışmanı: Yrd. Doç. Dr. Yeşim Aydın Son

Ağustos 2014, 60 sayfa

DNA'nın çift sarmal yapısının, genom boyunca homojen olduğu kabul edilmekteydi. Ancak, pek çok çalışmada, DNA'nın fiziksel yapısında, burkulmalar, eğilmeler, kaymalar ve açılmalar gibi lokal sarmal parametrelere ve komşu baz çiftleri arasındaki açılara dayalı değişiklikler olduğu gösterilmiştir. Bu durum, DNA'nın biyolojik fonksiyonlarını nasıl yerine getirdiğini açıklayacak daha iyi bir model sunarken, DNA'nın bilinen fiziksel ve kimyasal özelliklerini de sağlar. Pek çok çalışmada, bazı kanser türlerinde insan promotörlerinin metilasyon durumlarıyla gen ekspresyonu profilleri arasındaki ilişki gösterilmiş olmasına karşın metillenmiş promotörlerin yapısal özelliklerini inceleyen çalışmalara sık rastlanmamaktadır. Bu çalışmada bizim amacımız, insan promotörlerinin, metilasyon durumlarına ve gen ekspresyon profillerine göre sekansa bağlı DNA heliksel açıklığı açısından yapısal farklılıklarını incelemektir. Elde edilen yapısal farklılıkların, genom boyunca insan promotörlerinin metilasyon durumlarını öngören klinik açıdan önemli biyo-belirteçler elde edilmesini sağlayacak diğer çalışmalarda kullanılma potansiyeli vardır.

Anahtar Kelimeler: Promotör metilasyonu, nükleozom yerleşimi, heliksel yükseklik, gen susturma, DNA yapısı

To my family

ACKNOWLEDGEMENTS

First of all I would like to express my deepest gratitude to my supervisor Assist. Prof. Dr. Yeşim Aydın Son for her guidance, patience and encouragement. She has supported and guided me throughout the programme and my research. Also, she has become a wonderful model for me as an academician and a scientist with her immense knowledge, the importance given to her job, personality, kindness and sincerity. I think I am very lucky to have had such a supervisor in my first step to my academic career.

I would like to thank Assoc. Prof. Dr. Vilda Purutçuoğlu for her recommendations on the choice of statistical methods.

I would like to thank to Assoc. Prof. Dr. Tolga Can, Assist. Prof. Dr. Aybar Can Acar, Assoc. Prof. Dr. Çağdaş D. Son and Assist. Prof. Dr. Özlen Konu for reviewing my work.

I am grateful to Gökçe Oğuz and Zelha Nil for their support, encouragement, and valuable friendship. I also thank to my friends Güngör Budak, Özlem Özkan, Bilge Sürün, Onur Erdoğan, Dr. Cengizhan Açıkel and Alper Döm for their contributions and suggestions. I would like to thank to Informatics Institute members Sibel Gülnar and Hakan Güler for their support.

I would like to thank Özlem Özmen for her editorial contributions and for her friendship during the summer periods.

My sister Feruze Birer has been always with me and supported me during my whole life. I cannot thank enough to her for her priceless friendship. Also, special thanks to my dear friend Taylan Kutlu for being in my life with his endless support, encouragement and precious friendship.

I would like to thank to my brother Özgür Yaldız, my aunts and best friends Ümit Çağlar and Buket Çağlar, my grandmother Mürüvvet Çağlar, my uncle Alp, his wife Esra and cousins Ece and Arda Çağlar for their endless support and love.

Last but not least, special thanks to my mother Çiğdem Yaldız and my father Kadir Yaldız for their endless support, love, trust and patience in every step of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

The completion of the Human Genome Project in the early 2000s was a breakthrough in many fields such as genetics, genomic tools and technologies, human health and medicine, veterinary medicine, agriculture, food, industrial biotechnology etc. Sequencing the entire human genome and developing new analysis tools led to other projects that aimed to discover more about genomic variations, complex parts of the genome and regulatory mechanisms [1] [2].The International HapMap Project proceeded after Human Genome Project to investigate the genetic variants in humans. Its aim was to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared. In parallel, The 1000 Genomes Project started with the goal of sequencing the genomes of more people to get more information about genetic variants for investigating the relationship between genotype and phenotype [3].

The Encyclopedia of DNA Elements (ENCODE) Project was another important genome-wide project whose goal was to find all functional elements including genes, transcripts and transcriptional regulatory regions, DNA binding proteins that interact with regulatory regions, different versions of histones and DNA methylation patterns. In addition to the initial goals of the project, long range chromatin interactions were also examined. Investigation of the relation between regulatory process and chromatin structure led identification of binding proteins that are localized on mRNA and transcriptional silencer elements. Also the promoter sequence architecture in a subset of the genome was exploited [4].

Despite all studies on genome function and structure there are still many aspects of genome that are not yet fully understood, such as the epigenetic regulation of the genome. We believe that further studies on promoter methylation and genome structure might provide additional insight to epigenetic regulation, but new perspectives for the analysis of the genomic DNA will be required.

Here in this study we have inspected one feature of DNA structure, the helical rise, in order to reveal its potential value for the prediction of promoter methylation through DNA sequence and structure analysis.

1.2 Goal

The main purpose of this thesis is to develop a prediction method for the promoter methylation by the examination of structural properties of promoter regions.

For this purpose, the sequence dependent helical rise value was used as a bench mark for the comparison of structural features of different promoter regions which were identified according to their locations with reference to transcription start site. Mean helical rise values of the promoter regions were examined by statistical methods for determining the significance of the structural differences in gene groups and in each promoter region individually.

Additionally, we have investigated the frequencies of helical rise values on promoter regions for each gene individually. A difference score by using areas of the histograms of helical rise values were calculated. Our observations were largely consistent with the results of statistical analysis.

1.3 Contributions

It has been represented that there is a significant difference between promoter regions that are far from and near the TSS of differentially methylated genes by statistical analysis that was applied to the gene groups seperately. This situation likely to be conciding with previous studies which represent the role of methylation status and nucleosome occupancy on gene silencing.

Besides, it is deduced that, significance of the difference between mean helical rise values of promoter regions is a good indicator for the distinction of housekeeping genes from differentially methylated genes.

Furthermore, overall results suggest a relation between helical rise values and methylation potential of promoter regions.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 DNA Structure

2.1.1 Double Helix Structure

DNA is a polymer which is made up of four types of nucleotides that consist of deoxyribose sugar, phosphate and one of four heterocyclic bases, adenine, guanine, thymine and cytosine. The first clear X-ray diffraction photographs of DNA were obtained in the early 1950s and the structure of DNA became clearer with the following studies. Pauling and Corey proposed a three helical chain structure in which the chains were related to each other [5]. According to their model phosphate groups formed the core and the bases were on the outside of the DNA molecule.

In 1953, Watson and Crick found Pauling and Corey's model unsatisfactory because of some chemical reasons. They had indicated that negatively charged phosphates in the core could not hold the molecule together and van der Waals distances were not big enough. They suggested the double helix structure which is now prevalently known. With regard to this structure, the hydrophobic bases are inside the helix and the phosphates are on the outside of it. As well there are two right handed helical strands running in opposite directions and each strand has a backbone that is formed by deoxyribose sugar molecules linked together by phosphate groups. In addition, the distance is 3.4 Å and the angle is 36° between adjacent basepairs. It gives rise to repetitions after 10 residues in each chain and the repeat distance is 34 Å [6].

Accordingly, hydrogen bonding between bases on opposite strands, hydrophobicity of the nucleic acid bases, negatively charged phosphates on the backbone, salinity and directionality are the reasons for the helical form of the DNA.

2.1.2 Different Forms of Double Helix Structure

A-form, B-form and Z-form are the three forms of DNA's double helix structure (Figure 2.1). The most common form mentioned above is known as the B-form. In the B-form, major and minor grooves are apparent; it has a wide major groove that makes it more accessible to proteins. However, there are about 10.5 base pairs per

helical turn, its helix diameter is approximately 20Å, the base pairs are nearly perpendicular to helix axis and this form can be observed at high humidity levels.

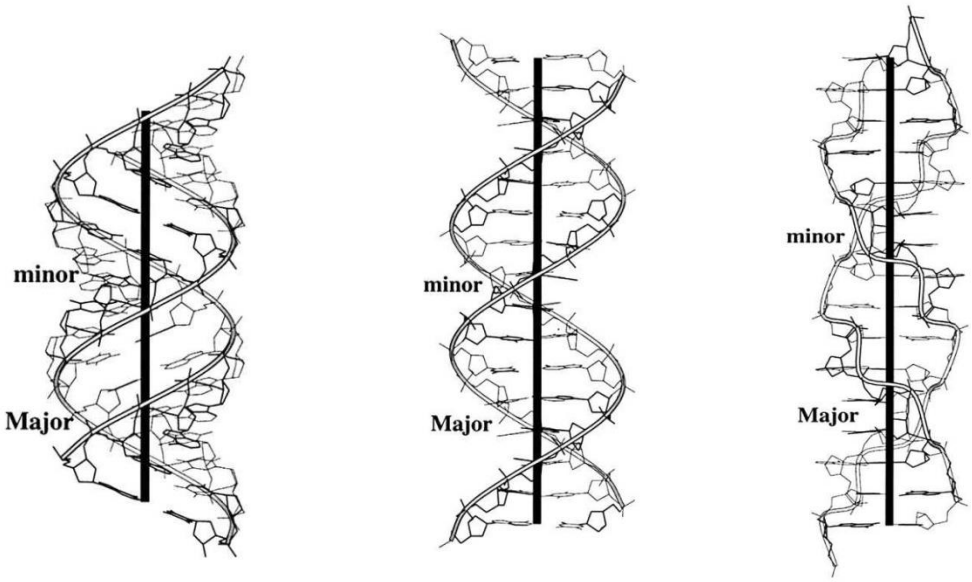

Figure 2.1 From left to right A-form, B-form and Z-form of the DNA (Adopted from [7])

In the A-form, the major groove is narrow and deep and the minor groove is shallow. There are 11 base pairs per helical turn and its helix diameter is about 26Å. It's seen that A-form is shorter and wider than B-form and as distinct from B-form, A-form can be observed at lower humidity levels. It is also right-handed as the B-form. It has been suggested that specific sites of the DNA such as promoter regions and transcription factor binding sites have A-form [8] [9].

The third form, the Z-form is distinguished from other forms by being left-handed. Also, it is thinner than other forms with approximately 18Å helix diameter and distinguished by the zigzag path of the sugar-phoshphate backbone. The minor groove of Z-form is deep and narrow but it does not have a distinct major groove [10].

2.1.3 Local Helical Parameters

Although the structures mentioned above seem to be rigid and uniformly observed throughout the genomic DNA when examined at atomic level that was not the case. The variation among angles and distance between adjacent basepairs was shown in many studies. Proteins can recognise the specific DNA sequences due to these local sequence specific structures.

In 1980, the helical structure of a self complementary dodecamer d(CGCGAATTCGCG) was investigated at atomic resolution using X-ray diffraction of single crystals [11]. Local helical parameters such as propeller twist, helix twist angle and rise per base pair were calculated. Here, propeller twist is a rotation of the two base pairs in opposite directions about their long axis and rise is the distance between adjacent basepairs. As a result of this study, it has been observed that the dodecamer molecule formed the B-DNA structure with some local sequence-dependent differences and also showed that the sequence influences the conformation of the DNA double helix structure.

Since the size of the purine and pyrimidine bases is different, propeller twist caused some uncomfortable contacts between adjacent base pairs. For this reason, it is suggested that in addition to propeller twist and local helical twist, the base roll angle is an important local helical parameter for sequence-dependent variations in the DNA helix structure [12]. This new parameter measures the angle between adjacent base pairs about their long axis.

Afterwards, many other local helical parameters were defined. In 1989, at an EMBO workshop on DNA Curvature and Bending, definitons of these local helical parameters were made and a common classification for the parameters is suggested [13]. Six base pair parameters, six base pair step parameters and global helical parameters which define the formation of the helix were described in the workshop.

Three base pair and three base pair step parameters that show the orientation are identified as rotational parameters. Other three base pair and three base pair step parameters that show the relative position are identified as translational parameters (Figure 2.2).

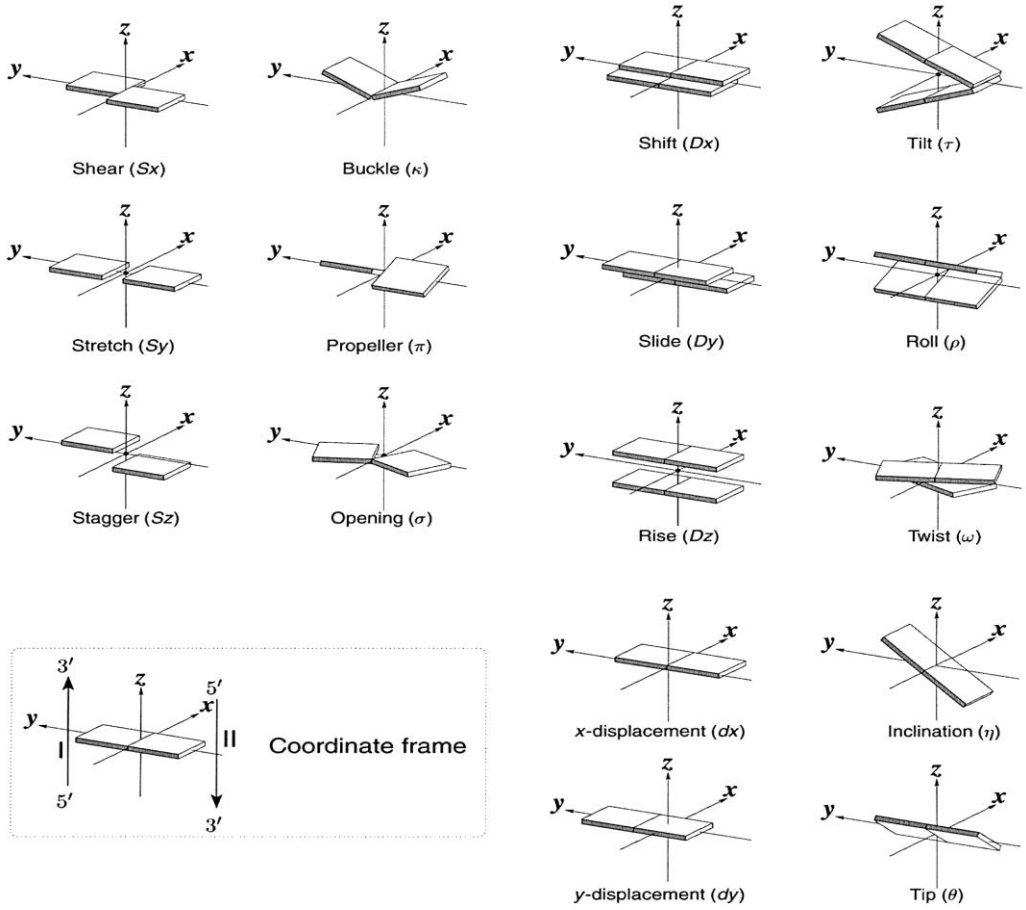

Figure 2.2 Base pair and base pair step parameters [7]. Rectangular slabs represent the base pairs.

Helical rise (h) is one of the global helical parameters. While the dimer step rise represents the vertical distance between base pairs as mentioned above, helical rise (h) represents the distance between the centers of base pairs along the helix axis [14]. Helical rise is same as dimer step rise (D_z) in B-DNA but smaller than rise in A-DNA because of squeezing of DNA via slide and tilt/roll in A-form (Figure 2.3) [7].

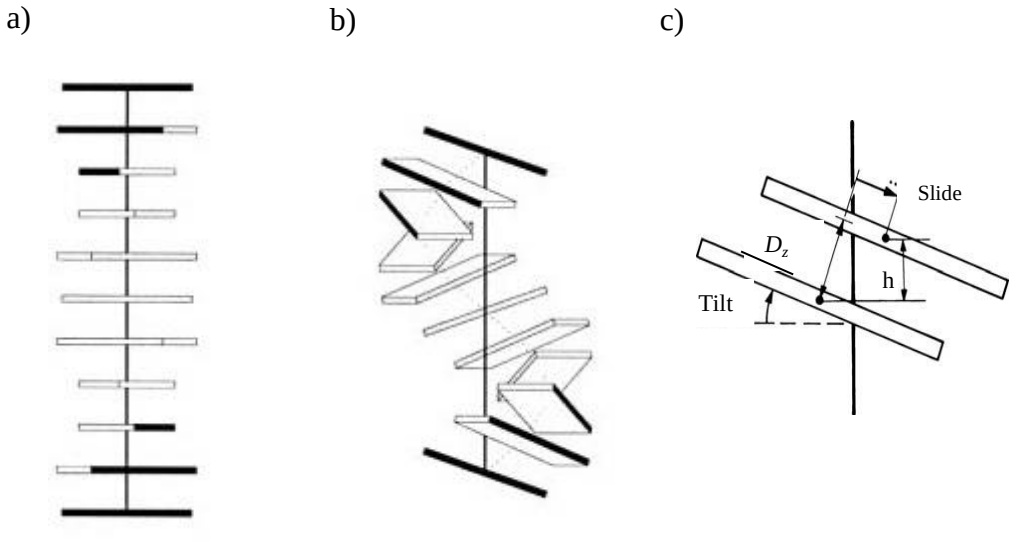

Figure 2.3 a) Representation of B-DNA as rectangular slabs b) Representation of A-DNA as rectangular slabs (Adopted from [7]) c) Effect of slide and tilt values to dimer step rise (D_z) in A-DNA (Adopted from [14])

By common classification of helical parameters they were standardized and these standard definitions have been used in algorithm development for analysing, reconstructing and visualizing the nucleic acid structures in many studies.

Structure and Conformation of Helical Nucleic Acids Analysis Program (SCHNAaP) is one of the programs that provide the local base pair and base pair step parameters, global helical parameters and according to these parameters structure can be categorized as A, B or Z form [15]. 3DNA is an other software for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Algorithm of analysing and rebuilding programs is derived from SCHNAaP [7].

Then a new structural parameter twist of supercoiling is defined. This parameter is related to global shape of the helical axis of a closed DNA and its sensitivity to chiral distortions is a distinguishing feature from step parameter twist. TwiDDL is a database that especially represents the difference between twist of supercoiling and step parameter twist. In order to establish this difference other base pair step parameters are also used. This database helps the user to see how the twisting is distributed and how the presence of proteins, drugs or other forces effect the local structural features of DNA and RNA [16].

2.1.4 Chromatin Structure

Human DNA includes 3 billion base pairs distributed between 23 pairs of chromosomes. In eukaryotic nucleus chromosomes are packaged with proteins called histones and form the chromatin structure. By this packaging mechanism, DNA can fit into the cell in a much smaller volume, is prevented from damages and gene expression and DNA replication can be controlled.

Histone proteins contain substantially the basic amino acids arginine and lysine. They are divided into five groups named H1, H2A, H2B, H3 and H4 due to their arginine/lysine content. Repeating subunits of chromatin are called nucleosomes and they consist of 147 base pairs wrapped around eight histone molecules which include two copies of each H2A, H2B, H3 and H4 (Figure 2.4.a [17]). Fifth type of histone H1 binds to linker DNA that connects two nucleosomes to each other [18]. Nucleosome structure represented in Figure 2.4.a is known as "beads on a string" state.

Histone molecules of neighboring nucleosomes interact with each other with the inclusion of linker histones and they form the higher level structure called "30nm fiber". Exact structure of 30nm fiber has not been known in detail, it's been a subject of debate. (Two possible structures of 30nm fiber can be seen in Figure 2.4.b [19]).

a) b)

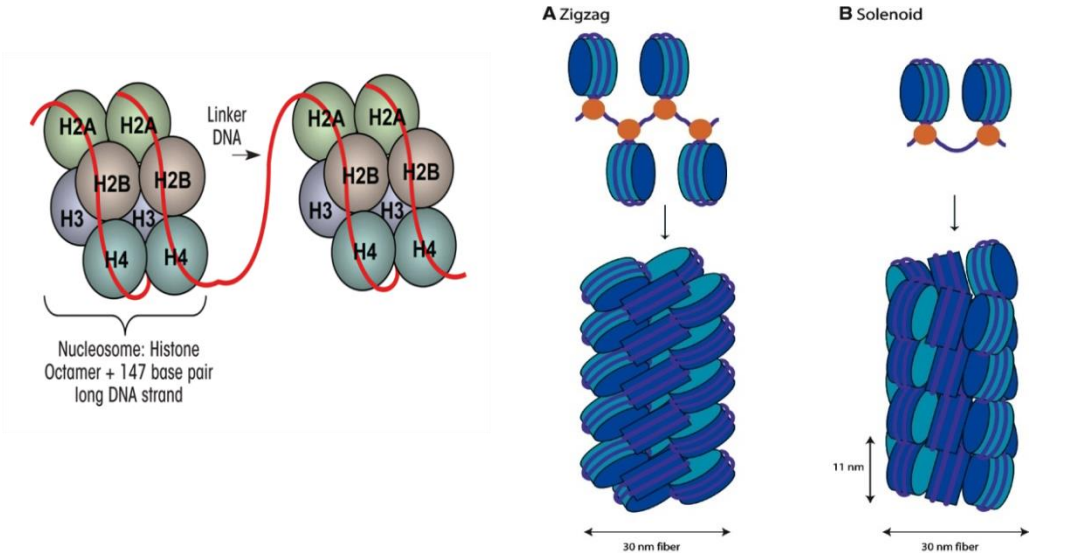

Figure 2.4 a) Schematic representation of nucleosomes, "beads on a string" state (Adopted from[17]) b)Schematic represenatation of two models of 30nm fiber structure (Adopted from[19])

In the next level of packaging, chromatin fibers coiled into the nucleus and formed supercoiled loops and domains. Spatial organization of the chromatin does not form randomly, it depends on various factors such as gene density of the region, transcriptional activity, epigenetic modifications and the stage of the cell cycle. Euchromatin and heterochromatin are two different types that are distinguished from each other according to their spatial oraganization. Euchromatin contains genes in active or inactive states, it is more flexible, accessible to transcription factors and represents an open conformation. On the other side, heterochromatin is highly condensed and inaccessible to transcription factors or other proteins.

2.2 Gene Structure

Gene is a nucleic acid sequence which serves as a physical and functional unit of heredity. Non-coding genes that encode the functional RNA molecules and the protein coding genes are two general types of genes in the human genome.

Synthesis process of proteins and functional RNA molecules is called gene expression. Synthesis process of proteins includes two steps: transcription and translation. Transcription, is the copying of RNA molecules from DNA templates and translation is the protein synthesis process after transcription.

One of the two DNA strands which is copied into mRNA is called the template strand and mRNA's sequence is complementary to this strand. Opposite strand is called the coding strand and its sequence is same as the mRNA's sequence .

Transcription process is catalyzed by an enzyme called RNA polymerase which binds to the template strand. Also, transcription is regulated by proteins called transcription factors that binds to specific DNA sequences. The upstream region of a gene which contains these specific sequences is called promoter. Both RNA polymerases and transcription factors attach to this promoter region. The first nucleotide of transcribed DNA sequence where the mRNA is began to synthesized by RNA polymerase is called the transcription start site (TSS).

mRNA consists of both coding and non-coding parts. Coding parts that are translated into proteins are called exons and non-coding parts that break up exons are called introns. After transcription process mRNA does not translated into a protein sequence directly. Introns are removed by a modification process called splicing and remaining exons are connected to each other for translation. Additionally, at the 5' and 3' ends of an mRNA there are regions called untranslated regions (UTR). They are not translated into proteins but play crucial roles in post-transcriptional regulation of gene expression (Figure 2.5).

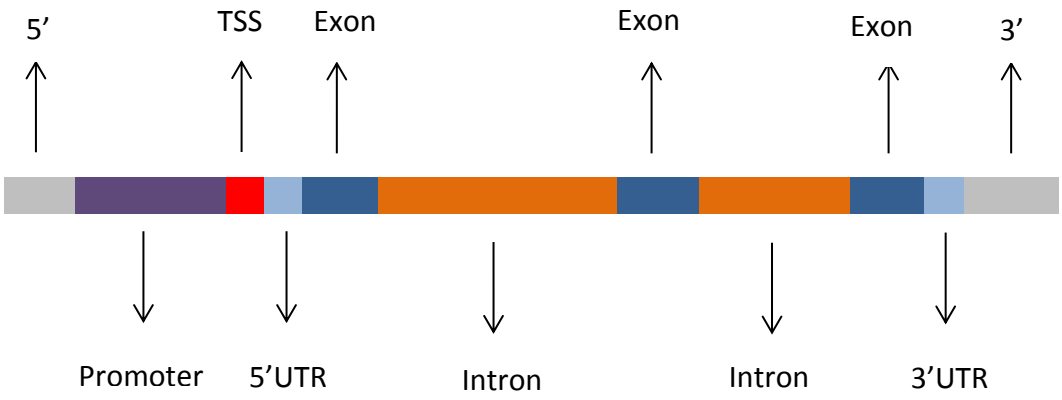

Figure 2.5 Representation of elements of a gene structure

2.3 Gene Expression Profile and Nucleosome Positioning

Nucleosomes are not stable constructions, conversely they are active from the point of composition and position. Relationship between gene expression profile and nucleosome occupancy around transcription start site has been indicated in many studies.

In a study released in 2007, nucleosome positioning in promoters was investigated with a high-resolution microarray approach combined with an analysis algorithm. Expressed genes and genes with preinitiation complexes at their promoters are shown to have nucleosome free regions at their transcription start sites and contrary to this in unexpressed genes nucleosome positioned transcription start sites are beheld [20].

In another study, in which genome-wide nucleosome position maps were generated, nucleosome positioning was found to be correlated with RNA Pol II binding near TSS. Related to this, +1 nucleosome which is just upstream of the transcription start site of the genes positioned differentially according to expressional status. In inactive genes 5’end of the +1 nucleosome was closer to TSS. Besides, -1 nucleosome depletion was observed in active gene promoters [21].

2.4 Nucleosome Occupancy and Helical Rise

Connection between local helical structure of the DNA and the nucleosome occupancy is another subject that has been an area of interest. Pedone and Santoni represented the relation between distribution of helical rising values and nucleosome stability. In their study, helical steps of known nucleosomal regions were analyzed by using helical rising values. Helical distances were calculated with reference to

nucleosomal dyad axis and a symmetric distribution was observed where the nucleosomes are stable. Also, as a supporting evidence for previous studies low affinity for nucleosome binding at the transcription start site was observed [22].

Further analysis indicates that the mean helical rise get its largest values on the nucleosome occupant regions on DNA [23]. Additionally, mean helical rise values of nucleosome occupant and nucleosome free regions around TSS were calculated as 3.05Å on promoter sequences. A meaningful difference was found between mean helical rise values of transcription start site and regions where the most stable nucleosomes named +2, +3, +4 etc. are located upstream of the TSS whose mean helical rise values exceed 3.25 Å. It is suggested that higher mean helical rise values in nucleosome occupant regions result from lower energetic cost required for DNA wrapping around histones and this case could be better provided with mean helical rise value greater than 3.2 Å.

Besides, as a result of X-ray analysis of DNA crystals it is suggested that A-form DNA has helical rise values around 2.83±0.36 Å and B-form DNA has around 3.29±0.21 [7].

2.5 Epigenetics

Epigenetics is "The study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [24]. In other words, study of heritable and reversible alterations in regulation of gene activity and expression that are not caused by changes in DNA sequence is known as epigenetics. Development, environmental chemicals, drugs, pharmaceuticals, aging and diet are the processes and factors that affect the epigenetics mechanisms [25].

Although epigenetics has a significant role in turning off genes in normal cellular processes such as differentiation of human embryonic stem cells, X chromosome inactivation and genomic imprinting, it is also responsible for some diseases including various cancer types, autoimmune diseases and mental disorders.

DNA methylation, histone modifications and non coding RNAs are three main mechanisms that cause epigenetic alterations.

2.5.1 DNA Methylation

DNA methylation is a biochemical process which involves the addition of a methyl group to the 5'-carbon of cytosine in a CpG dinucleotide. Methylation of 5'-carbon of cytosine also occurs in CpA and CpT in embryonic stem cells [26]. A small family of enzymes called DNA methyltransferases (DNMT) are responsible for controlling the addition of methyl groups in mammals. DNMT1, DNMT2, DNMT3a, DNMT3b and DNMT3L are the members of this family.

DNA methyltransferases are classified into two groups according to their role in de novo methylation and maintenance of methylation [27]. Since DNMT1 methylates hemimethylated CpGs preferentially it is defined as maintenance DNMT. Its potential role in de novo methylation of tumor suppressor gene promoters was also shown in some studies [28]. DNMT3a and DNMT3b are shown as responsible for de novo methylation.
Although DNMT2 has all of the conserved methyltransferase motifs it does not methylate DNA. But it is shown to methylate cytosine 38 in the anticodon loop of aspartic acid transfer RNA [29]. Similarly, DNMT3L contains DNA methyltransferase motifs but it is catalytically inactive. It interacts with DNMT3a and DNMT3b to methylate DNA during the differentiation of embryonic stem cells [30].

2.5.1.1 Genomic Imprinting

Genomic imprinting is a developmental process in which one of two gene copies is silenced depending on the parental origin. As a result of this process the gene copy that comes from only one parent is expressed. For instance, IGF2 is only active on the paternal chromosome, but CDKN1C is only active on the maternal chromosome.

Since differential DNA methylation is observed between maternal and paternal versions of imprinted genes and mice are unable to maintain the imprinted stage of genes in the lack of DNMT1 it is thought that imprinting is the result of DNA methylation in one of two alleles [18].

Prader-Willi syndrome, Angelman sydrome and some cancer types are the examples of the diseases that are related to deletions, uniparental disomy or mutations in the functional copy of the imprinted genes.

2.5.1.2 Methylation of Promoter CpGs

Certain levels of methylcytosine is needed in our genomes for controlling gene regulation. For instance, since females carry two X chromosomes and X chromosome includes much more genes than Y chromosome, female mammals transcriptionally silence one of their X chromosomes. DNA methylation acts as an epigenetic mark in this X inactivation event [31]. Likewise, methylation plays a significant role in differentiation of human embriyonic stem cells [32] and genomic imprinting [33].

In vertebrates, 60-90% of all CpGs are found to be methylated. In addition, CpG islands are the regions which have high frequency of CpG dinucleotide. More than

half of all human genes have CpG islands on their 5' end position and these CpG islands are non-methylated in active genes [34].

Besides, most CpG islands are unmethylated during development at all expression states. However, small amount of CpG islands become methylated related to transcriptional silencing of associated genes throughout the development. X chromosome inactivation and genomic imprinting are the processes that are related to methylation of promoter CpG islands during development [35].

Additionally, in normal somatic cells, methylation of CpG islands around transcription start site has been discovered via genome-wide studies and some of these genes are shown to be silenced in a tissue-specific manner [36]. Moreover, a few promoter CpG islands of the germline specific genes are found to be methylated in somatic tissues [37] and a few promoter CpG islands were found to be methylated during differentiation of embriyonic stem cells into neurons [38].

2.5.2 Histone Modifications

Chemical modification of the nucleosome histones is another epigenetic mechanism that regulates the gene expression and changes the structure of the chromatin (Figure 2.6). Acetylation, methylation, phosphorylation, ubiquitylation and sumoylation are the primary posttranslational modifications that plays fundamental role in gene regulation [39].

Histone acetylation is the addition of an acetyl group to the lysine residues catalyzed by histone acetyl transferases (HATs). By doing this, lysine positive charge is changed to neutral and interaction between the histone and DNA is weakened. As a result DNA becomes more accessible and transcription factors could bind the DNA easily. In this way, acetylation leads to increase in the expression of genes.

Histone deacetylation is the reverse of lysine acetylation process, catalyzed by histone deacetylases (HDACs). Contrary to acetylation, deacetylation returns the lysine charge to positive. By deacetylation event, DNA is wrapped around the histone cores more tightly and chromatin becomes condensed. This is one of the mechanisms that play a role in gene silencing.

Histone methylation is the addition of methyl groups to lysine or arginine residues of histone proteins. This process does not change the charge of the histone as acetylation and it can be related to transcriptional repression or activation.

Lysine methylation is catalyzed by histone lysine methyltransferase enzymes (HKMT). Lysines can be monomethylated, dimethylated or trimethylated. Location of different degree of methylation effects the gene regulation in different ways. For instance, H3K4me3 (trimethylation of histone H3 at lysine 4) is related to active transcription and H3K27me3 is related to transcriptional silencing. However there is

no clear distinction between effects of modifications on regulation. Same modification is also known to play a role in both transcriptional activation and repression [40].

Histone phosphorylation is another modification that alters the charge of the histone. Mostly it takes place on serine, threonine and tyrosine residues. It is controlled by kinases and phosphatases. This modification plays a crucial role in DNA damage repair. Also it is associated with regulation of transcription. For instance, phosphorylation of histone H3 is related to chromatine relaxation and activation of transcription [41].

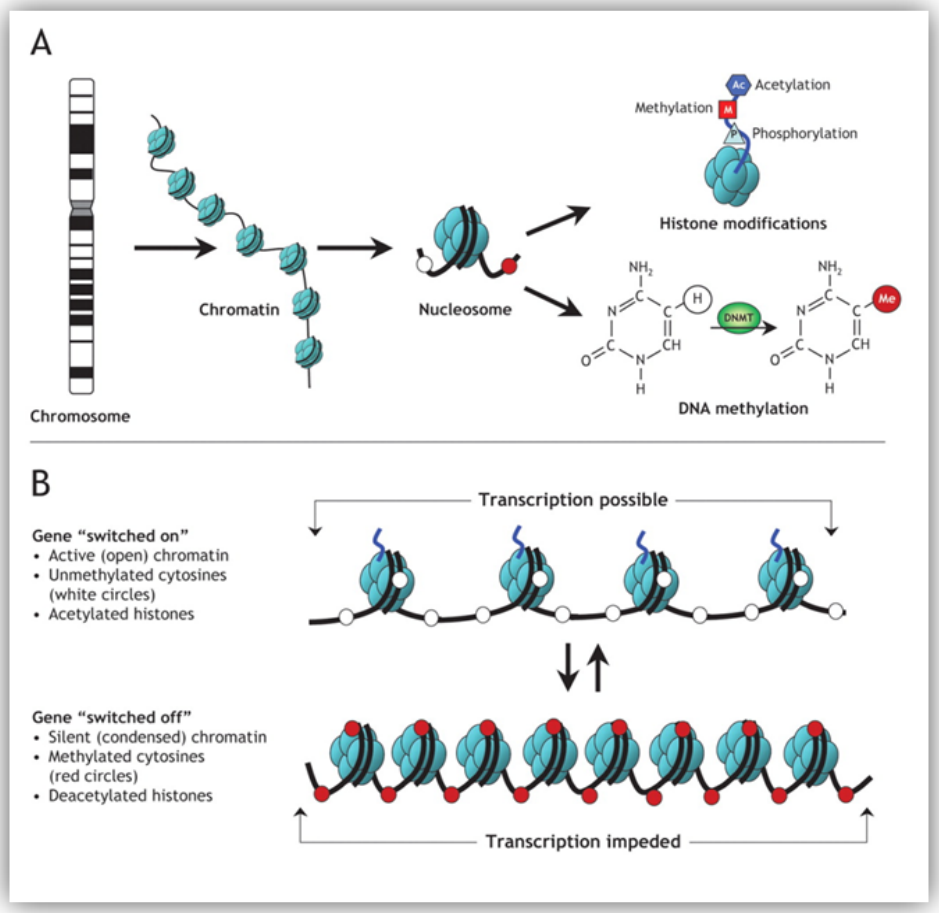

Figure 2.6 Schematic representation of epigenetic modifications (Adopted from [42])

2.6 DNA Methylation and Nucleosome Occupancy

Nucleosome positioning and DNA methylation together effect the gene regulatory mechanisms by influencing the chromosome structure. In a study, methylation patterns of nucleosome bound DNA were investigated and it is found that nucleosomal DNA was more highly methylated than flanking DNA [43].

In another study, genome-wide nucleosome positioning and DNA methylation information is combined. Relationship between nucleosome positioning around TSS and gene expression levels was observed in this study. In low expression levels, promoters with and without CpG islands are found to be nucleosome occupant around TSS consistent with previous findings. On the contrary, in high expression levels, upstream of TSS is found to be nucleosome depleted and downstream of TSS is found to be nucleosome occupant. Methylation profiles of these regions are correlated with nucleosome occupancy in different expression profiles. In expressed genes, promoters are non-methylated and they are methylated in silenced genes [44].

Effect of methylation on nucleosomal stability was also investigated by using molecular dynamics simulations and elastic deformation models. When CpG methylation occurs on the minor groove of DNA that attaches to the histone core makes the nucleosomes unstable and change the positioning of them. It is suggested that repositioning of nucleosome could cause alterations in gene activity by changing the chromatin structure and accessibility of DNA to transcription factors [45].

As epigenetic mechanisms play a crucial role in gene silencing for normal cellular processes they are also found to be associated with silencing of tumor suppressor genes in cancer. DNA methylation, histone modifications and physical changes in nucleosomal positioning also effect the tumor suppressor gene silencing [46].

CpG islands are non-methylated except CpG islands on promoters of the genes on the inactivated X chromosome and imprinted alleles as mentioned above. In addition, promoter CpG islands have low H1 histone levels, high histone acetylation (H3ac, H4ac) and methylation (H3K4me2, H3K4me3) levels, hypersensitivity to DNaseI [47]. All these features indicate an open and accessible euchromatin structure.

Relationship between promoter methylation and silencing of genes in cancer has been shown in many studies. While the active promoter of MLH1 gene lacks nucleosome just on the upstream of transcription start site, in cancer cells the same location is occupied by nucleosomes the promoter CpG island is methylated. Similarly, removal of nucleosomes from promoter regions after treatment with demethylating agent has been shown and as a result, it is suggested that the nucleosome occupancy in promoter regions is part of the epigenetic silencing of the tumor suppressor genes [48].

Also it has been shown that, after MLH1 promoter methylation has been reversed with decitabine treatment for three days, the methylation and nuclesome levels decreases and transcription is activated. Four days after withdrawal of decitabine,

first resilencing and then nucleosome reoccupation were observed while the methylation levels were still low. Accordingly, it is suggested that DNA methylation does not precede gene silencing and nucleosome occupancy can be more relevant with gene silencing [49].

Additional studies have shown the role of non-CpG island promoter methylation in gene silencing in cancer cells. Methylated promoter of the RUNX3 P1 gene was shown to be nucleosome occupant at the upstream of its TSS and unmethylated promoter was shown to be nucleosome depleted at the same region. It shows that methylation of non-CpG island and CpG island promoters effects the regulation of the transcription in a similar way [50].

CHAPTER 3

MATERIALS AND METHODS

3.1 Data

Y.-J.Kwon *et al*. published a set of hypermethylated-downregulated genes and hypomethylated-upregulated genes based on genome-wide methylation and microarray analysis results together for identifying the genes regulated by DNA methylation in SCC [51]. We have used these genes for the promoter region analysis in terms of DNA helical rise.

In the refered study, methylated CpGs were found by methylated CpG island recovery assay (MIRA) technique which is based on the high tendency of the methyl-CpG-binding domain protein-2b (MBD2b) / methyl-CpG-binding domain protein 3-like-1 (MBD3L1) complex for methylated DNA [52]. This method is similar to bisulfite sequencing method. MBD2b protein can particularly identify methylated DNA sequences and MBD3L1 protein upgrade MBD2b's affinity to methylated CpG islands.

Methylated DNA obtained from MIRA was used for high throughput sequencing and sequence tags were mapped to reference human genome (UCSC hg18 database based on the NCBI build 36.1 assembly) using the Solexa Analysis Pipeline. Then, microarray data was combined with the whole-genome DNA methylation pattern and based on the results 30 hypermethylated and down-regulated genes and 22 hypomethylated and up-regulated genes were selected. As seen in Table 3.1, for $\log_2$(T/N) above 1 genes were regarded as up-regulated and below -1 genes were regarded as down-regulated. Since DLEC1 gene was shown to be silenced by hypermethylation in a previous study [53] gene expression data does not exist in this microarray. CCDC37 is not down-regulated according to this criteria butits methylation level was extremely high. Similarly CBS, COL1A1, ELMO3 and MT1Bare not up-regulated but their methylation levels were extremely low, so they were added to the gene list.

The housekeeping genes are comprised in basic cell activities, they are expected to pursue constant expression levels in all cells and conditions and they have a non-methylated promoter which also includes a CpG island [54], [55]. Therefore, we have also selected 10 housekeeping genes as a control group for our analysis (Table 3.1)

Table 3.1 Genes that are selected based on the previous publications for the promoter region analysis (Adopted from [56])

Hypermethylated and Down-regulated		**Hypomethylated and up-regulated**		**Housekeeping**
Gene	**Fold Change (Log_2(T/N))**	**Gene**	**Fold Change (Log_2(T/N))**	**Gene**
ADAM33	*-1.221087063*	*ADSSL1*	*1.172602594*	*ACTBP2*
AGTR1	*-1.356755098*	*CBS*	*0.854992104*	*ACTBP7*
APOB48R	*-1.809420667*	*CCL7*	*1.128776661*	*ALB*
ASTN2	*-2.116127773*	*CDCA5*	*2.514666762*	*GAPDH*
CA10	*-2.829902911*	*COL1A1*	*0.84662518*	*PPIAL4E*
CCDC37	*-0.438497777*	*EDN2*	*3.577795253*	*PPIAP30*
CDO1	*-1.985933328*	*ELMO3*	*0.859621462*	*RNA18S5*
CLTCL1	*-1.023127859*	*GBX2*	*2.88765218*	*RNA28S5*
COL13A1	*-1.470238996*	*HOXD11*	*2.72750794*	*TUBA1A*
CTSE	*-1.154048118*	*KCNC1*	*2.427335724*	*TUBA3FP*
CYTL1	*-1.165674222*	*KRTCAP3*	*1.552826764*	*TUBB8P8*
DLEC1		*LASP1*	*1.026014302*	
DNM3	*-2.665170126*	*MT1B*	*0.832941395*	
GUCA2B	*-1.17527637*	*NETO2*	*1.125125323*	
HIST1H1B	*-2.52113352*	*NSDHL*	*1.037793283*	
HIST1H3A	*-1.264196556*	*PAGE4*	*3.211556392*	
HOPX	*-1.099664166*	*PDCL2*	*4.248027683*	
KANK2	*-1.182291311*	*PRIM1*	*2.185488422*	
KLRC4	*-1.451254932*	*PSMA6*	*0.964107805*	
LMO3	*-1.188834946*	*SERPINB5*	*4.121717307*	
MYH2	*-3.987329902*	*SLC35F3*	*2.02506441*	
NID1	*-1.272255513*	*SPC25*	*2.701452195*	
PAX9	*-1.108395547*			
PGC	*-2.417429465*			
PPP1R14A	*-1.440936626*			
RGS5	*-1.434346574*			
SLC1A2	*-1.129283017*			
SLIT2	*-1.817988624*			
TCAP	*-1.012605722*			
TMEM146	*-1.029065264*			

3.2 Analysis

3.2.1 Preparation of the data for the analysis

After specifying our dataset, we have identified the promoter sequences 1500bp upstream and 500bp downstream of transcription start sites (TSS) of the genes listed in Table2.1. TSSs are pointed on almost the same locations in Ensembl and UCSC Genome Browsers and presentation of TSSs is more convenient in Ensembl Genome Browser for our analysis. For these reasons, although the genes were mapped to UCSC database we have downloaded the promoter sequences from Ensemble Genome Browser.

Then, we divided promoter sequences into three regions according to their position with reference to the TSS. First region is the part between 1500bp upstream and 500bp upstream of TSS (A), second region is 500bp upstream of the TSS (B) and the third region is the 500bp downstream of the TSS (C) (Figure 3.1).

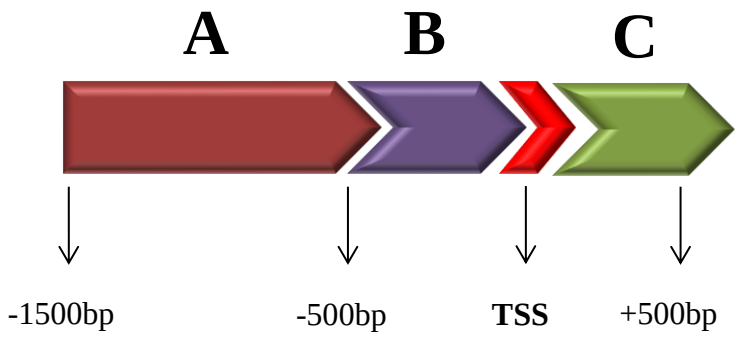

Figure 3.1 Promoter regions assigned in the study according to their position with reference to transcription start site

3.2.2 Helical Step Analysis

Pedone and Santoni constituted a tetranucleotide code by using data collected from available databases of resolved DNA structures for 136 possible tetrads. Helical rise values ranged from 2.36Å to 4.46Å [22]. (The table that represents these tetranucleotide codes can be found in Appendix A). In this tetranucleotide code, helical rising values were assigned to the central dinucleotide helical step according to the adjoint bases. For instance, the table shows that the helical rise value of TCCA is 3.33Å and this means that helical rise value of CC dinucleotide step is 3.33Å when it occurs with T and A adjoint bases. The reason for using the tetranucleotide codes rather than dinucleotide steps is the sensitivity of the dinucleotides to neighboring base pairs.

It was mentioned that since there is not any resolved structure some of the values in the table were calculated by averaging values for tetranucleotides containing the same dinucleotide step and some of them could be derived by a single DNA oligomer. Therefore it was suggested that the table might need to be refined by using new resolved structures. Since this is the most current tetranucleotide code that represents the helical rise values we have used it for our analysis.

We converted the promoter region sequences into helical rise values arrays by running the Python script that we have developed (Python script can be found in Appendix B). We started the process with a four bases window at the beginning of the sequence then, slide the window one by one base up to the end of the sequences. We compared each four bases window with the tetranucleotides in the table and we assigned the helical rise values to an array (Figure 3.2).

```
tetramers = ["136 possible tetrads"]

helicalrise = ["helical rise values of possible tetrads"]

a= [];

for i in range(0 , (len(seq)-3)):

        for j in range(0,len(tetramers)):

                if tetramers[j]==seq[i:i+4]:

                        a.append(helicalrise[j]);
```

Figure 3.2 Python script we used for helical step analysis

Number of elements in a helical rise array of a sequence is three bases less than the number of bases in that sequence. If "n" is the number of bases in the sequence, number of helical rise values is expected to be "n-1". We compare the sequence with tetranucleotide codes from the beginning of the sequence by sliding four bases windows. In the first window, comparison is between 1^{st} to 4^{th} bases and at this step helical rise value between 2^{nd} and 3^{rd} value is assigned. Similarly, in the last window we compare the "n-3"th to n^{th} bases to the tetranucleotide codes and we get the helical rise value between "n-2"th and 'n-1'th bases. We can not calculate the helical rise value in the first and last step with tetranucleotide codes. So that, there are "n-3"elements in helical rise arrays of the sequences.

3.2.3 Statistical Analysis

We have calculated the means of all the promoter sequence regions that we have converted to helical rise arrays. We aimed to find differences between the means of helical rise values at different locations. There are 6 groups of mean helical rise values include regions A, B, C, A+B, B+C and A+B+C for each promoter in hypermethylated, hypomethylated and housekeeping gene groups. There are 30 mean helical rise values in each group for hypermethylated genes, 22 mean helical rise values in each group for hypomethylated genes and 10 mean helical rise values in each group for housekeeping genes. We have analyzed the differences in the mean helical rise values of the regions defined in Figure 3.1 with the following order: A with B, A with C, B with C, A+B with C, B+C with B, B+C with C, A+B+C with B and lastly A+B+C with C regions in each gene group in terms of mean helical rise values. Since we expected to see higher mean helical rise values around TSS we did not compare A+B with A and A+B+C with A.

In order to select the statistical test for our analysis first, we have analyzed the dependency and distribution of the groups. A, B and C regions are independent groups and B+C and B, A+B and B, B+C and C, A+B+C and B, A+B+C and C are the dependent groups. Additionally, since all these groups have discrete probability distributions they could not be normally distributed. Thus, we applied Wilcoxon-Mann-Whitney test for independent groups and Wilcoxon signed-rank test for dependent groups by using R for assesing if there is any significant difference between the mean helical rise values of these regions. Additionally, we applied Bonferroni correction to the p-values for controlling the familywise error rate which arises from unequal number of genes in gene groups.

As we have observed a significant difference between mean helical rise values of promoter regions, in the next step we have individually examined the gene promoters. We have divided promoters into 40 sub-regions with 50 bp length then calculated the mean helical rise values of these sub-regions to see the differences between mean helical rise values clearly. Furthermore, we fitted a curve into mean helical rise values of sub-regions by using LOESS method for the demonstration of differences across the whole promoter regions in terms of mean helical rise values.

Next, we have grouped these mean helical rise values according to their location (A, B or C) on the promoter sequence. There were 20 sub-regions on A, 10 sub-regions on B and C. Afterwards, since these groups are independent and not normally distributed Wilcoxon-Mann-Whitney test was applied to the groups for the individual genes, which exhibit the highest significant difference in the previous analysis.

3.2.4 Comparing Frequencies of Helical Rise Values

We have examined the frequencies of helical rise values in all regions of the promoter sequences and drew their histograms for visualizing the distribution. We have adjusted the bin width as 0.05 Å in the histograms.

Afterwards, the frequency of helical rise value occurrences for each of the genes in all groups are compared by calculating the total area of the bars in the histograms. We have used the helical rise values at midpoints of bars as a weighting coefficient to calculate the area of the bars. Then, we multiplied this weighting coefficient and height of bars to find the areas of bars for A, B and C regions. However, since the number of helical rise values in region A is two times higher than regions B and C we used half of height values of bars for region A. The sum of areas described above for each region A, B and C is calculated and the difference between these scores for different regions are further investigated. By doing this, we expected to find a threshold value for differences between different regions. This threshold value could be used for the prediction of regions for that have a potential for differential promoter methylation.

CHAPTER 4

RESULTS

4.1 Investigation of Selected Genes According to Their Chromosomal Location

The preliminary analysis of the genes is done by examining them according to their locations on the genome, CpG island presence on their promoter and overlapping with other genes.

20 of the 30 hypermethylated genes have CpG islands on their promoters, CCDC37 and TMEM146 promoters are found to be overlaping with other gene exons. HIST1H1B and HIST1H3A are single exon genes and HIST1H3A has CpG island across the whole gene (Table 4.1).

Table 4.1 Results of Examination of Hypermethylated Genes

Hypermethylated Genes				
Gene	**Location**	**# of Transcripts**	**CpG island on Promoter**	**Nested**
ADAM33	20: 3,667,965-3,682,246	7	Present	No
AGTR1	3: 148,697,784-148,743,008	9	Present	No
APOB48R	16:28494649-28498970	2	Absent	No
ASTN2	9:116425225-117415070	8	Absent	No
CA10	17:51630313-52160017	9	Present	No
CCDC37	3:126394939-126436556	8	Present	Yes
CDO1	5:115804733-115816954	4	Present	No
CLTCL1	22:19179473-19291716	12	Present	No
COL13A1	10:69801931-69964275	14	Present	No
CTSE	1:206009264-206023909	4	Absent	No
CYTL1	4:5014586-5019472	3	Present	No
DLEC1	3:38039205-38124025	7	Present	No
DNM3	1:171841498-172418466	7	Present	No

Table 4.1 continued				
GUCA2B	1:42153421-42155824	1	Absent	No
HIST1H1B	6:27866849-27867529	1	Present	No
HIST1H3A	6:26020490-26020900	1	Present	No
HOPX	4:56647988-56681899	15	Present	No
KANK2	19:11164267-11197791	13	Absent	No
KLRC4	12:10407382-10409757	1	Absent	No
LMO3	12:16548373-16610594	35	Absent	No
MYH2	17:10521148-10549957	6	Absent	No
NID1	1:235975830-236065162	2	Present	No
PAX9	14:36657568-36679715	6	Present	No
PGC	6:41736711-41754109	4	Absent	No
PPP1R14A	19:38251237-38256591	5	Present	No
RGS5	1:163244505-163321894	9	Absent	No
SLC1A2	11:35251206-35420063	8	Present	No
SLIT2	4:20253260-20620561	12	Present	No
TCAP	17:39664187-39666555	2	Present	No
TMEM146	19:5720677-5778734	3	Present	Yes

13 of the 22 hypomethylated genes have CpG islands on their promoters. GBX2 have CpG island across the whole gene. CDCA5, ELMO3, KRTCAP3 and NETO2 are found to be overlapped with other gene exons (Table 4.2).

Table 4.2 Results of Examination of Hypomethylated Genes

Hypomethylated Genes				
Gene	**Location**	**# of Transcripts**	**CpG island on Promoter**	**Nested**
ADSSL1	14:104724186-104747325	12	Present	No
CBS	21:6444871-6468040	3	Present	No
CCL7	17:34270221-34272242	3	Absent	No
CDCA5	11:65066300-65084164	10	Present	Yes
COL1A1	17:50183289-50201632	13	Present	No
EDN2	1: 41478775-41484673	5	Absent	No
ELMO3	16: 67199111-67204029	5	Present	Yes

Table 4.2 continued				
GBX2	2: 236165236-236168369	3	Present	No
HOXD11	2: 176104216-176109754	3	Present	No
KCNC1	11: 17734812-17783055	4	Present	No
KRTCAP3	2: 27442366-27446481	7	Present	Yes
LASP1	17: 38869859-38921770	9	Present	No
MT1B	16: 56651899-56653204	2	Absent	No
NETO2	16: 47077703-47143997	5	Absent	Yes
NSDHL	X: 152830967-152869729	3	Absent	No
PAGE4	X: 49829260-49833973	4	Absent	No
PDCL2	4: 55556525-55592212	2	Absent	No
PRIM1	12: 56731596-56752373	8	Present	No
PSMA6	14: 35278633-35317493	15	Present	No
SERPINB5	18: 63476761-63505085	6	Absent	No
SLC35F3	1: 233904933-234324516	2	Present	No
SPC25	2: 168834132-168913371	4	Absent	No

Lastly, 4 of the 10 housekeeping genes have CpG islands on their promoters and only two them are found to be overlapped with other gene exons (Table 4.3).

Table 4.3 Results of Examination of Housekeeping Genes

Housekeeping Genes				
Gene	**Location**	**# of Transcripts**	**CpG island on Promoter**	**Nested**
ACTBP2	5: 77784881-77786003	1	Absent	No
ACTBP7	15: 43989061-43990184	1	Absent	No
ALB	4: 73397114-73421412	20	Absent	No
GAPDH	12: 6533927-6538374	11	Present	No
PPIAL4E	1: 144372875-144373659	1	Absent	No
PPIAP30	10: 15154802-15155320	1	Absent	No
RNA18S5	Y: 10198504-10199102	1	Present	No
RNA28S5	X: 109054131-109054562	1	Absent	Yes
TUBA1A	12: 49184796-49189324	9	Present	No
TUBA3FP	22: 21002895-21014292	3	Present	Yes

4.2 Helical Rising Arrays

We performed helical step analysis for all the genes we have selected for the study. Promoter sequences obtained from ENSEMBL Genome Browser were used as input for our calculations. A, B, C, A+B, B+C and A+B+C regions of the promoter sequences identified above were transformed to helical rise arrays for further analysis.

In region A of the promoter sequences there are 1000 bases and we got 997 helical rise values for this region. In A+B region there are 1500 bases and we got 1497 helical rise values, for the regions B and C there are 500 bases and we got 497 helical rise values, lastly we analyzed A+B+C and we got 1997 helical rise values for whole promoter region (Figure 4.1). Helical rise arrays of the gene promoters can be found in Appendix C and scatter plots of the gene promoters' helical rise values can be found in Appendix D.

We observed obvious differences between different regions of most of the promoters we have analyzed. For instance, in Figure 4.1 there are not many helical rise values higher than 4Å upstream to the TSS but the helical rise values are much higher at the downstream of TSS. We needed further analysis to compare these values on different regions to understand whether they were significant.

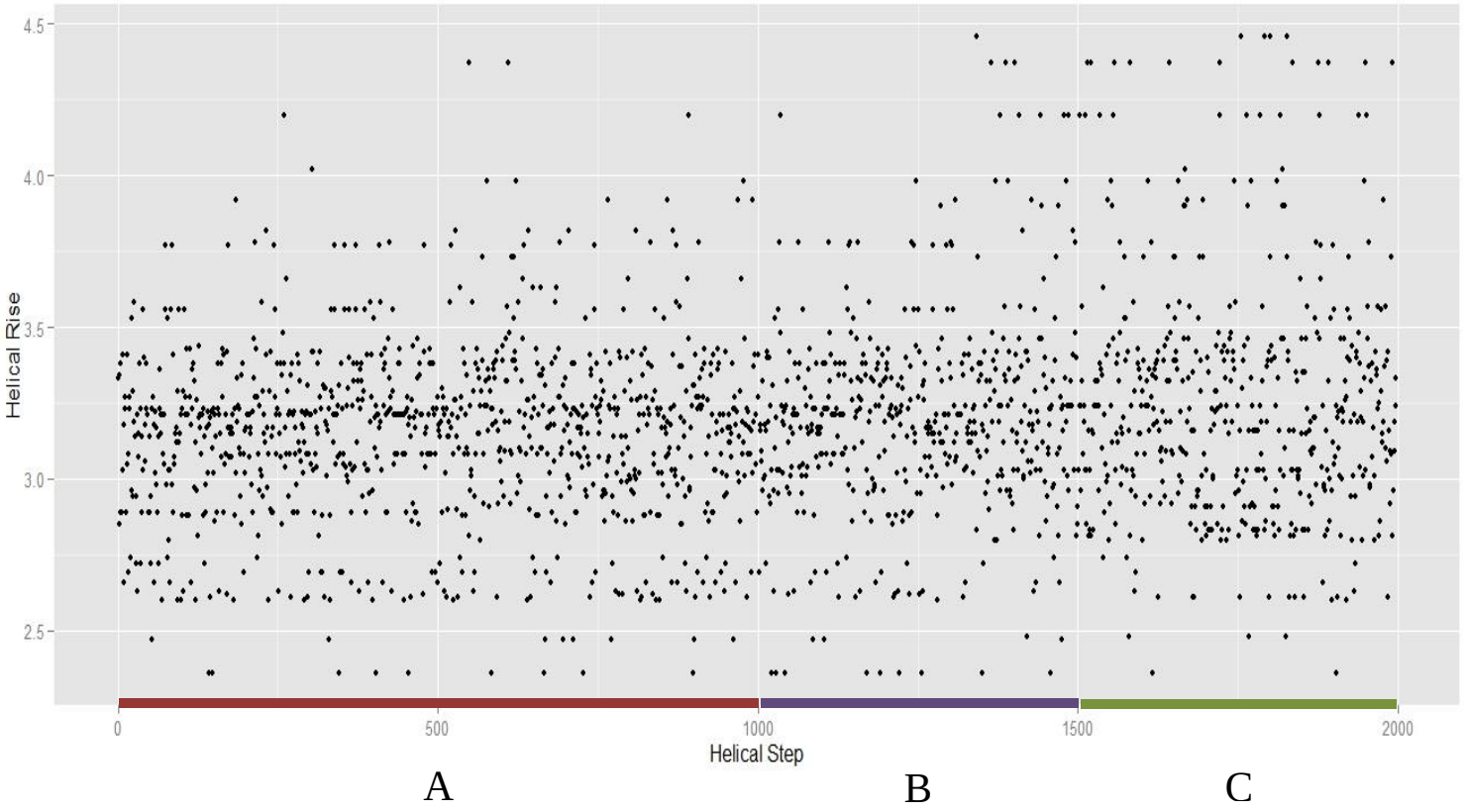

Figure 4.1 Scatter plot of helical rise values in AGTR1 gene promoter: On the x-axis 0-1000 represents region A(marked red), 1000-1500 represents region B(marked blue) and 1500-2000 represents region C (marked green) and on the y-axis helical rise values are represented

4.3 Comparison of Mean Helical Rise Values of Gene Groups

We calculated the mean helical rise values to examine the differences between the regions. Mean helical rise values for all regions of the gene promoters can be found in Appendix E. Usually promoters have A-form and their helical rise values are around 2.83±0.36 Å. As seen in Appendix E, mean helical rise values of our sequences are between ~3.13 Å and ~3.26 Å. We could suggest that our sequences also exhibit structures close to the A-form.

The differences between mean helical rise values of different promoter regions of the genes are statistically tested for three groups of genes; hypermethylated, hypomethylated and housekeeping genes as presented in Table 4.4.

When the mean helical rise value for whole promoters (A+B+C) are compared to the downstream of transcription start sites (C) for each group of genes, p-value for the hypermethylated genes is found as 9.22E-06. This shows that there is a highly significant difference in terms of mean helical rise values between these regions for the hypermethylated gene group. Difference between A and C and A+B and C is also considerably significant ($p<0.00005$) for the hypermethylated gene group. On the other hand when only the upstream regions are compared, the significance of the difference is decreased ($p>0.001$) for the hypermethylated genes.

For the hypomethylated and the housekeeping genes the distinctness between regions were not found to be significant as the hypermethylated genes for all p-values were higher than 0.001 (Table 4.4).

Since the number of genes in hypermethylated, hypomethylated and housekeeping gene groups were not equal, we also applied Bonferonni Correction to each p-value we obtained from the statistical analysis. Adjusted p-values are also represented in Table 4.4. Adjusted p-values indicate the significance of difference in hypermethylated gene group more clearly.

Table 4.4 Differences Between Mean Helical Rise Values for Regions A, B and C of Gene Promoters.

First column shows the regions of the promoters that are compared to each other. In following columns p-values obtained from Wilcoxon signed-rank or Wilcoxon-Mann-Whitney tests are shown, respectively. Adjusted p-values are reported according to the Bonferonni Correction results for all groups, which presents are clear difference between hypermethylated and other groups. Also among the hypermethylated group p-values <0.001 (in bold) presents the regions that are most significantly different from each other in terms of mean helical rise values.

	Hypermethylated		**Hypomethylated**		**Housekeeping**	
	p-value	**Adjusted p-value**	**p-value**	**Adjusted p-value**	**p-value**	**Adjusted p-value**
(A+B+C) vs C	**9.22E-06**	3.07E-07	0.03289	0.001495	0.01172	0.001172
A vs C	**2.45E-06**	8.15E-08	0.065	0.002955	0.2973	0.02973
(A+B) vs C	**3.78E-05**	1.26E-06	0.3527	0.016032	0.2581	0.02581
(A+B) vs B	0.03454	1.15E-03	0.009274	0.000422	0.8203	0.08203
(B+C) vs C	0.001232	4.1067E-05	0.7024	0.031927	0.003906	0.000391
A vs B	0.001525	5.08E-05	0.007922	0.00036	0.8633	0.08633
B vs C	0.02847	9.49E-04	0.5073	0.023059	0.1903	0.01903
(B+C) vs B	0.03454	1.15E-03	0.7024	0.031927	0.003906	0.000391
(A+B+C) vs B	0.01745	5.82E-04	0.0425	0.001932	0.25	0.025

4.4 Comparison of Mean Helical Rise Values of Individual Genes

In differentially methylated genes the presence of nucleosome occupant regions around transcription start site has been reported in different studies. Also an increase in the mean helical rise values on the nucleosome occupant DNA have been reported. Supporting these observation the differences between the mean helical rise value of regions A+B+C vs C and A vs C of hypermethylated promoters were highly significant. So we have concentrated on the A vs C regions of gene promoters for futher analysis at individual gene level, where we expect to see the most structural change and observe that the significant difference in the mean helical rise values.

For the comparison of regions at gene level, we have divided the whole promoter into 40 sub regions of 50 nucleotides and calculated the mean helical rise values of these sub-regions. Scatter plots of these sub-regions' mean helical rise values can be found in Appendix F. The Wilcoxon-Mann-Whitney test is applied to analyze difference between the mean helical rise values region A and C for individual genes in all three groups, as presented in Table 4.5.

In hypermethylated group 15 out of 30 genes, AGTR1, DNM3, CA10, KANK2, SLC1A2, CDO1, APOB48R, PPP1R14A, CTSE, DLEC1, KLRC4, COL13A1, GUCA2B, CLTCL1 and LMO3, showed a significant difference ($p<0.05$) between regions A and C of their promoters. The difference of mean helical rise values between region A and C was significant for 7 out of 22 hypomethylated genes CCL7, SLC35F3, NSDHL, CBS, SPC25, MT1B and PAGE4, selected for this study. However RNA28S5 is the only housekeeping gene out of 10 selected, which shows significant difference ($p<0.05$) between regions A and C (Table 3.2).

The distribution of the mean helical rise values for the AGTR1 gene which showed highest significant difference ($p<0.0005$) is given in Figure 4.2. While the mean helical rise value of region A was ~3.15Å, region C's mean helical rise value was ~3.26Å (Appendix E). As seen in Figure 4.2, on regions A and B mean helical rise values were lower than 3.2Å. Solely, we see an increase in mean helical rise values in the vicinity of transcription start site. This increase begins at the end of the region B near the transcription start site.

Table 4.5 P-values obtained from Wilcoxon-Mann-Whitney test applied to individual genes to comparing promoter regions A and C

Hypermethylated Genes		Hypomethylated Genes		Housekeeping Genes	
Gene	**p-value**	**Gene**	**p-value**	**Gene**	**p-value**
AGTR1	8.22E-05	**CCL7**	4.38E-05	**RNA28S5**	0.009616
DNM3	0.0002587	**SLC35F3**	0.005287	**PPIAL4E**	0.1196
CA10	0.000324	**NSDHL**	0.01273	**GAPDH**	0.1829
KANK2	0.001349	**CBS**	0.02156	**ACTBP7**	0.198
SLC1A2	0.003258	**SPC25**	0.02442	**RNA18S5**	0.4745
CDO1	0.003843	**MT1B**	0.03662	**ACTBP2**	0.53
APOB48R	0.009616	**PAGE4**	0.03925	**ALB**	0.53
PPP1R14A	0.009616	**HOXD11**	0.0673	**PPIAP30**	0.7132
CTSE	0.01108	**PDCL2**	0.267	**TUBA1A**	0.198
DLEC1	0.01108	**SERPINB5**	0.2865	**TUBA3FP**	0.8458
KLRC4	0.01273	**COL1A1**	0.333		
COL13A1	0.02442	**ELMO3**	0.3735		
GUCA2B	0.03112	**ADSSL1**	0.3789		
CLTCL1	0.03285	**EDN2**	0.448		
LMO3	0.03467	**GBX2**	0.5379		
NID1	0.06069	**LASP1**	0.5525		
ADAM33	0.1552	**PSMA6**	0.7083		
RGS5	0.2348	**KCNC1**	0.7414		
MYH2	0.3735	**CDCA5**	0.7787		
CCDC37	0.3735	**NETO2**	0.8798		
ASTN2	0.3909	**KRTCAP3**	0.895		
CYTL1	0.448	**PRIM1**	0.9483		
SLIT2	0.4745				
HOPX	0.4814				
PGC	0.6129				
TCAP	0.7132				
PAX9	0.7749				
TMEM146	0.7787				
HIST1H3A	0.8121				
HIST1H1B	1				

Additionally, DLEC1 gene which was reported to be silenced by hypermethylation [53] shows a significant difference ($p<0.05$) on its promoter between A and C regions (Table 4.5). In Figure 4.3 we have seen that on the region A mean helical rise value was about 3.19 Å however, on the C region mean helical rise value was 3.25 Å (Appendix E) for DLEC1.

As an example to housekeeping genes, GAPDH promoter is presented (Figure 4.4.) which did not show a significant difference between regions A and C in terms of mean helical rise. Mean helical rise value of region A was ~3.25Å and region C was 3.28Å (Appendix E). When we examine the mean helical rise values of 40 sub-regions of the promoter we see that almost all mean helical rise values were greater than 3.2 Å.

The only housekeeping gene that showed highly significant difference ($p<0.05$) between A and C regions was RNA28S5, but due to the high mean helical value of region A instead of the area around TSS (Figure 4.5.).

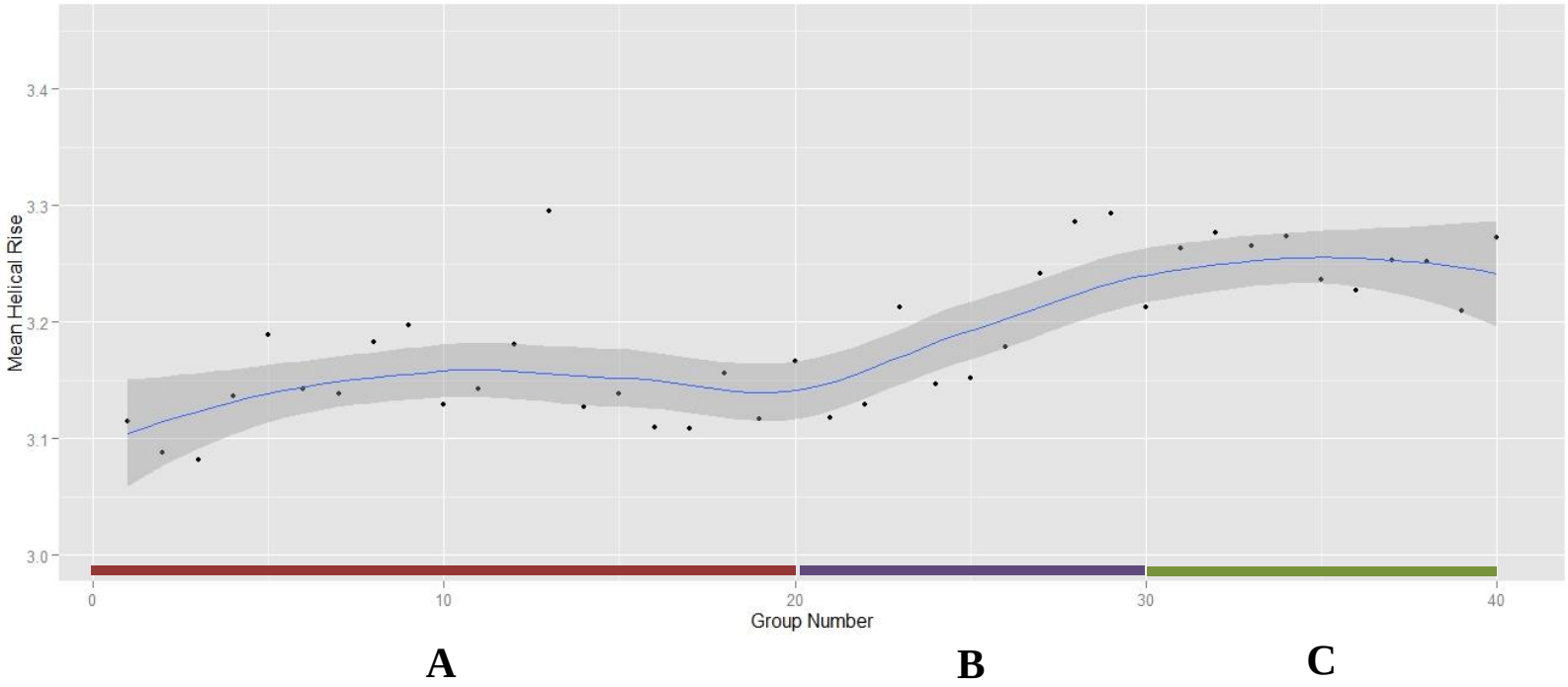

Figure 4.2 AGTR1 hypermethylated gene promoter: Curve fitting analysis of the mean helical rise values of sub-regions of the promoter are represented on the y-axis, group numbers are represented on the x-axis. On the x-axis 0-1000 represents region A(marked red), 1000-1500 represents region B(marked blue) and 1500-2000 represents region C (marked green) and on the y-axis helical rise values are represented

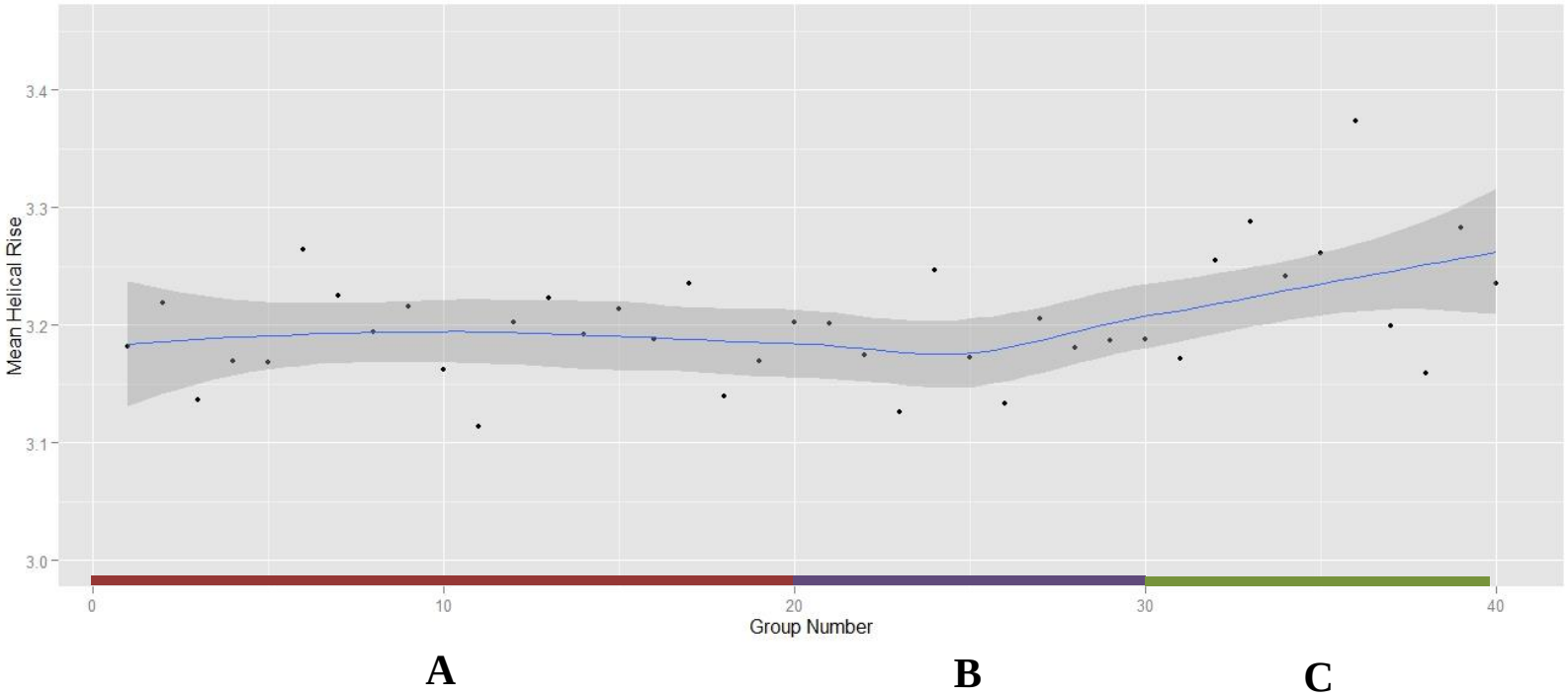

Figure 4.3 DLEC1 hypermethylated gene promoter: Curve fitting analysis of the mean helical rise values of sub-regions of the promoter are represented on the y-axis, group numbers are represented on the x-axis. On the x-axis 0-1000 represents region A(marked red), 1000-1500 represents region B(marked blue) and 1500-2000 represents region C (marked green) and on the y-axis helical rise values are represented

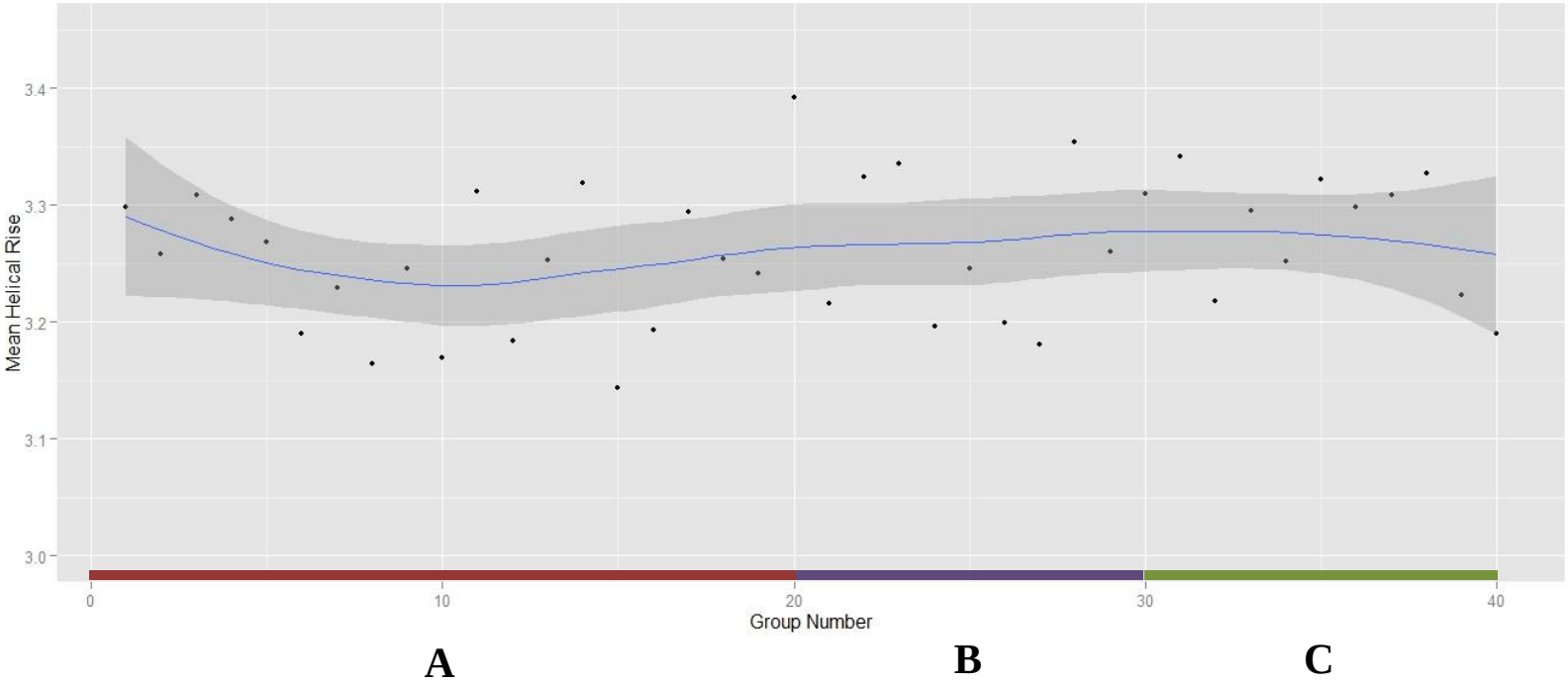

Figure 4.4 GAPDH housekeeping gene promoter: Curve fitting analysis of the mean helical rise values of sub-regions of the promoter are represented on the y-axis, group numbers are represented on the x-axis. On the x-axis 0-1000 represents region A(marked red), 1000-1500 represents region B(marked blue) and 1500-2000 represents region C (marked green) and on the y-axis helical rise values are represented

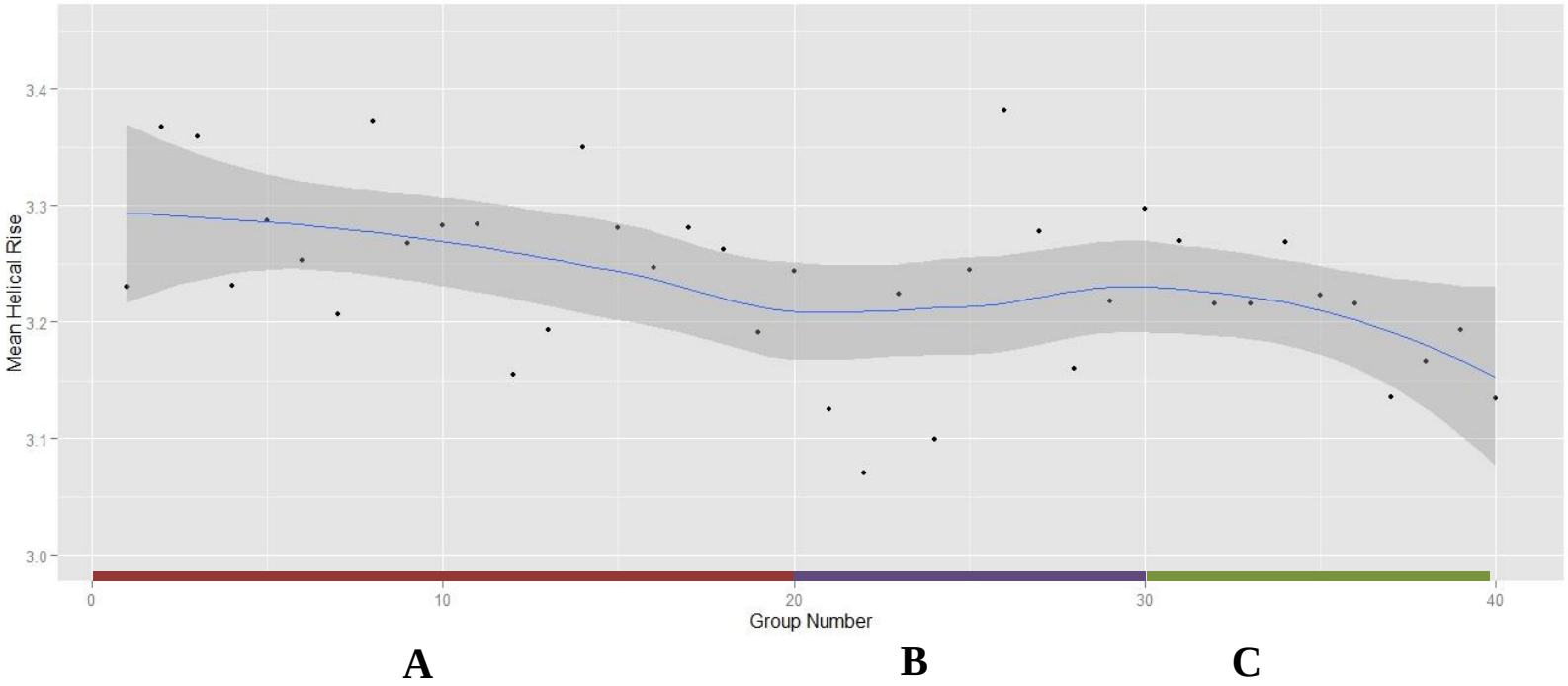

Figure 4.5 RNA28S5 housekeeping gene promoter: Curve fitting analysis of the mean helical rise values of sub regions of the promoter are represented on the y-axis, group numbers are represented on the x-axis. On the x-axis 0-1000 represents region A(marked red), 1000-1500 represents region B(marked blue) and 1500-2000 represents region C (marked green) and on the y-axis helical rise values are represented

4.5 Comparison of Frequencies

We observed significant differences between regions A and C of differentially methylated promoters as we expected in previous individual analysis of the genes. Separately, mean helical rise value of the tetranucleotide code that we used for our calculations (Appendix A) is 3.2Å and we observed the highest frequency of the helical rise values between 3.2Å and 3.25Å in the regions of most of the promoters (Figure 3.6). Histograms can be found in Appendix G. Since we expected to see nucleosome occupant regions around TSS of differentially methylated genes, helical rise values of these regions should be greater than 3.2Å as mentioned in Chapter 1. For this purpose, we compared frequencies of helical rise values greater than 3.25 Å in regions A and C.

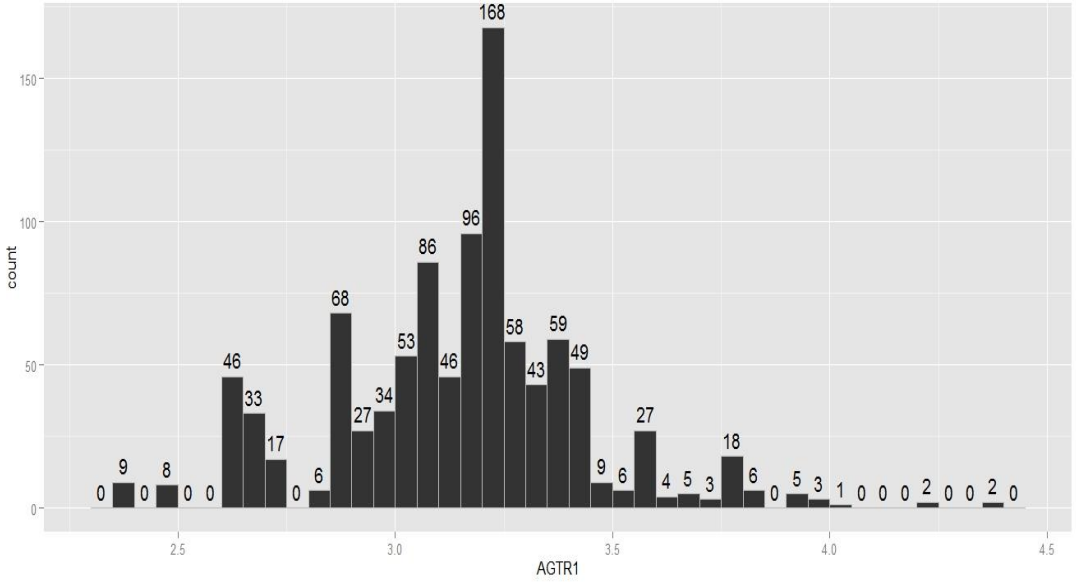

Figure 4.6 Histogram of helical rise values in region A of AGTR1 promoter

We calculated the score as we identified in Chapter 2 for regions A and C of all promoters and found the difference between A and C regions, as presented in Table 4.6.

A positive diffeence has been observed for most of the hypermethylated genes. The differences between frequency scores were above 45 for the genes which showed significant difference between regions A and C of promoters (Table 4.6). Only exception to this set of hypermethylated genes with was two genes ADAM33 and CCDC37, whose mean helical rise values didn't show a significant difference in the previous analysis.

Table 4.6 Frequency scores and score differences of the regions A and C in hypermethylated genes

Hypermethylated Genes			
Gene	**A**	**C**	**C-A**
DNM3	557.5	787.7	230.2
AGTR1	518.275	743.7	225.425
CA10	519.225	737.675	218.45
PPP1R14A	560.2125	778.35	218.1375
KANK2	582.4625	784.675	202.2125
DLEC1	625.925	767.325	141.4
ADAM33	566.4875	703.675	137.1875
GUCA2B	591.3375	713.3	121.9625
SLC1A2	575.9	688.2	112.3
LMO3	572.4375	668.375	95.9375
NID1	632.1375	724.325	92.1875
KLRC4	537.7625	624.65	86.8875
CDO1	585.0125	671.025	86.0125
CLTCL1	658.175	733.1	74.925
CCDC37	703.7	765.975	62.275
COL13A1	639.7	697.4	57.7
CTSE	661.7875	711.35	49.5625
APOB48R	619.6	666.15	46.55
MYH2	603.8	648.1	44.3
TCAP	715.4375	753.1	37.6625
HIST1H3A	609.025	636.55	27.525
PGC	538.55	564.75	26.2
TMEM146	673.6375	680.5	6.8625
RGS5	637.2375	620.7	-16.5375
HOPX	620.325	588.7	-31.625
PAX9	768.225	733.55	-34.675
ASTN2	635.7875	593.575	-42.2125
SLIT2	662.8375	619.275	-43.5625
HIST1H1B	607.2875	539.175	-68.1125
CYTL1	672.4875	513.6	-158.888

For hypomethylated genes, we again found the frequency score differences over 45 for the genes that show significant difference between A and C regions as hypermethylated genes (Table 4.7) and for housekeeping genes, a score difference over 50 is obtained only for GAPDH (Table 4.8).

Table 4.7 Frequency scores and score differences of the regions A and C in hypomethylated genes

Hypomethylated Genes			
Gene	**A**	**C**	**C-A**
NSDHL	571.7375	762.775	191.0375
CCL7	599.9125	766.775	166.8625
SLC35F3	553.5875	696.025	142.4375
CBS	634.6625	768.925	134.2625
PDCL2	573.6125	674.2	100.5875
HOXD11	606.0375	694.925	88.8875
SPC25	565.125	651.15	86.025
MT1B	624.4375	708.6	84.1625
SERPINB5	619.025	684.925	65.9
KRTCAP3	635.5125	701.3	65.7875
ELMO3	725.5	782.35	56.85
PAGE4	638.45	672.85	34.4
EDN2	651.95	683.275	31.325
NETO2	652.5125	666.4	13.8875
LASP1	693.5375	683.475	-10.0625
ADSSL1	734.3375	723	-11.3375
PRIM1	659.375	645.725	-13.65
KCNC1	675.6875	646.6	-29.0875
GBX2	668.3875	636.575	-31.8125
COL1A1	681.3	644.775	-36.525
PSMA6	585.0375	546.2	-38.8375
CDCA5	685.55	641.15	-44.4

Table 4.8 Frequency scores and score differences of the regions A and C in housekeeping genes

Housekeeping Genes			
Gene	**A**	**C**	**C-A**
GAPDH	693.825	791.25	97.425
ALB	566.95	609.525	42.575
ACTBP7	606.9625	633	26.0375
PPIAL4E	544.2	562.15	17.95
ACTBP2	618.6375	633.775	15.1375
RNA18S5	691.3125	692.5	1.1875
TUBA1A	636.525	631.425	-5.1
PPIAP30	618.2	581.275	-36.925
TUBA3FP	725.6	679.675	-45.925
RNA28S5	719.15	650.925	-68.225

CHAPTER 5

DISCUSSION

Differential positioning of the first nucleosome downstream of a TSS (+1 nucleosome) according to expression status was previously demonstrated [21]. In a similar way the role of nucleosome occupancy on promoter region just upstream of transcription start site in epigenetic silencing of tumor suppressor genes was also presented in previous studies [46]. Additionally, it was referred that the nucleosome occupant regions tend to have higher mean helical rise values and helical rise values greater than 3.2 Å provides DNA wrapping around histones with lower energetic costs [23].

The difference we have observed between mean helical values for regions A, B and C of hypermethylated genes when they were analyzed as a group also suggests a similar mechanism (Table 3.1). Differentially methylated genes should have nucleosome occupant regions around their transcription start sites. We found a significant difference between mean helical rise values of regions that are far from and next to the transcriptional start site (A and C) for the differentially methylated genes. Since the nucleosome occupant regions have higher mean helical rise values the observed difference may arise from nucleosome occupancy near transcription start sites in differentially methylated genes.

Next, we also observed significant differences between regions A and C of the differentially methylated genes as a result of individual analysis. Housekeeping genes have constant expression levels and non-methylated promoters as mentioned in Chapter 2. Therefore, epigenetic silencing mechanism is not expected to work for the regulation of expression of housekeeping genes. In consistency with this mechanism, while we have observed significant difference between regions A and C of differentially methylated gene promoters, there was not any clear distinction between regions A and C of housekeeping gene promoters.

The significant difference (p-value < 0.05) between regions A and C of differentially methylated gene promoters at individual level suggests a connection between methylation status and helical rise values, but there was a high number of false negatives in both hyper and hypomethylated genes group. As this study only focused on the mean helical rise difference between regions A and C as an initial attemp to investigate the role of this structural change on the promoter methylation, the analysis was not comprehensive enough. In this initial examination we used TSS as an artifical landmark to be able to divide the promoter sequences into regions that

would suitable for the statistical analysis. But, we believe that the structural differences that we can observe by the change in helical rise of DNA is not only limited to the downstream of TSS, but lies within an area around the TSS. Further examination of the whole promoter sequences might help us explain the false negative results in our analysis.

Additionally the exact methylation process of genomic DNA is still not known to its full extend, and there have been multiple molecular mechasims suggested. As we have selected our gene set based on the results of methylation and expression arrays, there is a possibility that the mechanism of the differential methylation observed might be different. We expect promoter sequences to have structural change and high helical rise values when methylation is facilitated by a nucleosome dependent mechanism.

Moreover, overlapping, length of CpG islands on the promoters and gene lengths could be other reasons for the false negative results. CCDC37, TMEM146, CDCA5, ELMO3, KRTCAP3 and NETO2 are the genes whose promoter regions on the upstream of the TSS are overlapping with other gene promoters and exons (Figure 5.1). The exon regions show higher nucleosome occupancy levels [57] and this might be a reason for not observing significant difference between mean helical rise values of A and C regions. On the other hand, hypermethylated HIST1H3A and hypomethylated GBX2 genes have CpG islands across the whole gene and this should give rise to mean helical rise values to get close to each other.

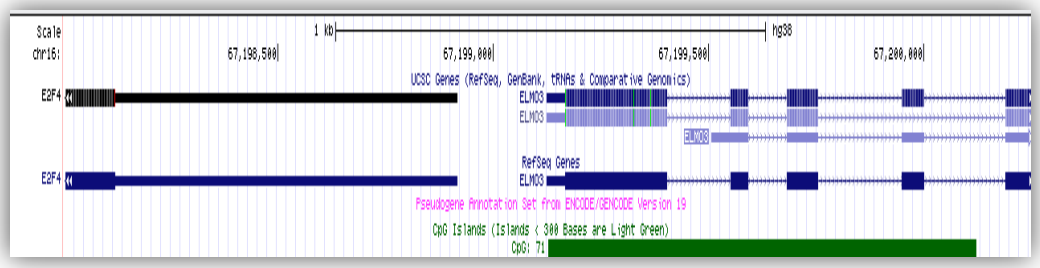

Figure 5.1 ELMO3 gene promoter is overlapping with E2F4 gene promoter.

Also as the genes we have selected for our analysis were identified with high-throughput methods, we expect to have a number of genes that might not be differentially methylated.

In paralel to significance analysis of mean helical rise differences, we have also investigated the difference between the helical rise frequencies above 3.2Å, and

analyzed whether this frequency score might represent the differential methylation of promoter sequences. The significance of differences was found to be a better indicator of identifying housekeeping genes from differentially regulated genes, as none of the housekeeping genes found to have higher helical rise values at their region C. On the other hand the higher number of differentially methylated genes have a positive difference between the frequency scores of regions A and C. Also while all differentially methylated genes with a significant difference (p-value <0.05) had a frequency difference score above 45, only one housekeeping genes had a difference at this level.

CHAPTER 6

CONCLUSION AND FUTURE STUDIES

6.1 Conclusions

In this study our goal was to investigate sequence dependent DNA structure in terms of local parameter helical rise for the prediction of human promoter methylation. In line with our goal, we selected differentially methylated genes with different expression levels in squamous cell lung carcinoma. Additionally we selected ten housekeeping genes as a control group.

Differentially methylated genes were extracted from a study in which they were identified by genome-wide DNA methylation and microarray analysis. Then, promoter sequences of these genes were obtained from ENSEMBL Genome Browser and we divided promoter sequence into three parts according to their position with reference to transcription start site.

We transformed each promoter sequence part to helical rise arrays by using tetranucleotide code which was constituted by using data collected from available databases of resolved DNA structures for 136 possible tetrads.

Since nucleosome occupancy around transcription start site changes according to expression status we expected to see differences in terms of mean helical rise values due to epigenetic silencing mechanism in differentially methylated genes. In parallel with this issue, we represented significant difference between promoter regions that are far from and near TSS of differentially methylated genes by statistical analysis we applied to gene groups.

Later on, we analyzed each gene individually for promoter regions that show significant difference in previous results. 16 of the 30 hypermethylated genes and 8 of the 22 hypomethylated genes show significant difference individually in terms of helical rise. These observations were consistent with the current literature.

As a final step, we developed a scoring mechanism by using difference between helical rise frequencies of different regions for the prediction of methylated promoters. The results that we obtained by using this scoring mechanism was confirmative with our previous findings.

In this study, we have showed the structural differences in terms of sequence dependent helical rise between regions of promoters which we specified according to their locations with reference to transcription start site. We claimed that this difference arises from different nucleosome occupancy states just upstream and downstream of TSS according to expression status as a result of epigenetic silencing mechanism. Genes that are regulated by DNA methylation are expected to present higher helical rise values around TSS due to nucleosome occupancy. Additionally, we developed a scoring method for the prediction of methylated promoters by using sequence dependent DNA structural differences. This will enable researchers to identify genes that are regulated by DNA methylation in normal development and carcinogenesis. Correspondingly, it will be possible to identify epigenetic markers for early diagnosis of cancer.

6.2 Future Work

- We made structural comparisons between promoter regions that we specified arbitrarily. New analysis methods should be developed to investigate helical rise differences all through out the promoter sequences.

- Helical rise values of all human genes could be calculated and analysis of differences in comparison to methylation and nucleosome occupancy status and other regulative elements could be done.

- Link between nucleosome positioning and splicing has been represented in recent studies [57] [58]. Nucleosome occupancy levels are shown to be higher in exons than introns. Accordingly, it is suggested that nucleosome occupancy has a role in identification of exons by splicing machinery. In parallel with this issue, helical step analysis and statistical calculations presented in this study could be investigated for identification of possible exon-intron splicing sites.

REFERENCES

[1] J. Watson, “Human Genome Project,” no. October, pp. 1–2, 2010.

[2] A. Cancer, S. Cancer, A. Network, L. G. Universities, L. Technologies, P. Cancer, P. Healthcare, T. E. Society, T. F. Scientific, and S. California, “About United for Medical Research About Battelle.”

[3] G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. a Gibbs, M. E. Hurles, and G. a McVean, “A map of human genome variation from population-scale sequencing.,” *Nature*, vol. 467, no. 7319, pp. 1061–73, Oct. 2010.

[4] T. Encode and P. Consortium, “A user’s guide to the encyclopedia of DNA elements (ENCODE).,” *PLoS Biol.*, vol. 9, no. 4, p. e1001046, Apr. 2011.

[5] B. Y. L. Pauling and R. B. Corey, “A Proposed Structure For The Nucleic Acids,” pp. 84–97, 1953.

[6] J. D. Watson and F. H. . Crick, “Molecular Structures Of Nucleic Acids.” Nature 171, pp. 737–738, 1953.

[7] X.-J. Lu, “3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures,” *Nucleic Acids Res.*, vol. 31, no. 17, pp. 5108–5121, Sep. 2003.

[8] Z. Shakked, D. Rabinovich, W. B. T. Cruse, E. Egert, O. Kennard, G. Sala, S. A. Salisbury, and M. A. Viswamitra, “Crystalline A-DNA: The X-Ray Analysis of the Fragment d(G-G-T-A-T-A-C-C),” *Proc. R. Soc. London. Ser. B. Biol. Sci.* , vol. 213 , no. 1193 , pp. 479–487, Nov. 1981.

[9] L. Fairall, S. Martin, and D. Rhodes, “The DNA binding site of the Xenopus transcription factor IIIA has a non-B-form structure.,” *EMBO J.*, vol. 8, no. 6, pp. 1809–17, Jun. 1989.

[10] A. D. Bates Maxwell, Anthony.,, *DNA topology*. Oxford [u.a.]: Oxford University Press, 2006.

[11] H. R. Drewt, R. M. Wingtt, T. Takanot, C. Brokat, S. Tanakat, K. Itakuraii, and R. E. Dickersont, “Structure of a B-DNA dodecamer : Conformation and dynamics * Biochemistry :,” vol. 78, no. 4, pp. 2179–2183, 1981.

[12] C. R. Calladine, “Mechanics of sequence-dependent stacking of bases in B-DNA,” *J. Mol. Biol.*, vol. 161, no. 2, pp. 343–352, Oct. 1982.

[13] B. Nomenclature, “Definitions and nomenclature of nucleic acid structure parameters.,” *J. Mol. Biol.*, vol. 205, no. 4, pp. 787–91, Feb. 1989.

[14] C. R. Calladine and H. Drew, *Understanding DNA: the molecule and how it works*. Academic press, 1997.

[15] X. Lu, M. A. El Hassan, and C. A. Hunter, “Structure and Conformation of Helical Nucleic Acids : Analysis Program (SCHNAaP),” pp. 668–680, 1997.

[16] L. a Britton, W. K. Olson, and I. Tobias, “Two perspectives on the twist of DNA.,” *J. Chem. Phys.*, vol. 131, no. 24, p. 245101, Dec. 2009.

[17] B. G. Starkman, A. J. Sakharkar, and S. C. Pandey, “Epigenetics-beyond the genome in alcoholism.,” *Alcohol Res.*, vol. 34, no. 3, pp. 293–305, 2012.

[18] G. Karp, *Cell and molecular biology : concepts and experiments*. Hoboken, NJ: John Wiley, 2010.

[19] S. Khorasanizadeh, “The Nucleosome : From Genomic Organization to Genomic Regulation University of Virginia Health System,” vol. 116, pp. 259–272, 2004.

[20] F. Ozsolak, J. S. Song, X. S. Liu, and D. E. Fisher, “High-throughput mapping of the chromatin structure of human promoters.,” *Nat. Biotechnol.*, vol. 25, no. 2, pp. 244–8, Feb. 2007.

[21] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, “Dynamic regulation of nucleosome positioning in the human genome.,” *Cell*, vol. 132, no. 5, pp. 887–98, Mar. 2008.

[22] F. Pedone and D. Santoni, “Sequence-dependent DNA helical rise and nucleosome stability.,” *BMC Mol. Biol.*, vol. 10, p. 105, Jan. 2009.

[23] F. Pedone and D. Santoni, “Preferential nucleosome occupancy at high values of DNA helical rise.,” *DNA Res.*, vol. 19, no. 1, pp. 81–90, Jan. 2012.

[24] V. E. A. Russo, R. A. Martienssen, and A. D. Riggs, “Epigenetic mechanisms of gene regulation,” *Cold Spring Harb. Monogr. Ser.*, 1996.

[25] “Epigenetic mechanisms,” *http://commonfund.nih.gov/epigenomics/figure#*.

[26] B. H. Ramsahoye, D. Biniszkiewicz, F. Lyko, V. Clark, a P. Bird, and R. Jaenisch, “Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 10, pp. 5237–42, May 2000.

[27] K. D. Robertson, “DNA methylation, methyltransferases, and cancer.,” *Oncogene*, vol. 20, no. 24, pp. 3139–55, May 2001.

[28] K.-W. Jair, K. E. Bachman, H. Suzuki, A. H. Ting, I. Rhee, R.-W. C. Yen, S. B. Baylin, and K. E. Schuebel, “De novo CpG island methylation in human cancer cells.,” *Cancer Res.*, vol. 66, no. 2, pp. 682–92, Jan. 2006.

[29] M. G. Goll, F. Kirpekar, K. A. Maggert, J. A. Yoder, C.-L. Hsieh, X. Zhang, K. G. Golic, S. E. Jacobsen, and T. H. Bestor, “Methylation of tRNAAsp by the DNA Methyltransferase Homolog Dnmt2,” *Sci.* , vol. 311 , no. 5759 , pp. 395–398, Jan. 2006.

[30] F. Neri, A. Krepelova, D. Incarnato, M. Maldotti, C. Parlato, F. Galvagni, F. Matarese, H. G. Stunnenberg, and S. Oliviero, “Dnmt3L Antagonizes DNA Methylation at Bivalent Promoters and Favors DNA Methylation at Gene Bodies in ESCs,” *Cell*, vol. 155, no. 1, pp. 121–134, Aug. 2014.

[31] J. Ahn and J. Lee, “X Chromosome: X Inactivation,” *Nat. Educ.*, vol. 1, no. 1, p. 24, 2008.

[32] E. Li, T. H. Bestor, and R. Jaenisch, “Targeted mutation of the DNA methyltransferase gene results in embryonic lethality,” *Cell*, vol. 69, no. 6, pp. 915–926, Jun. 1992.

[33] M. a Surani, “Imprinting and the initiation of gene silencing in the germ line.,” *Cell*, vol. 93, no. 3, pp. 309–12, May 1998.

[34] F. Antequera, “CpG Islands and Methylation,” in *eLS*, John Wiley & Sons, Ltd, 2001.

[35] M. M. Suzuki and A. Bird, “DNA methylation landscapes: provocative insights from epigenomics.,” *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 465–76, Jun. 2008.

[36] L. Shen, Y. Kondo, Y. Guo, J. Zhang, L. Zhang, S. Ahmed, J. Shu, X. Chen, R. a Waterland, and J.-P. J. Issa, “Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters.,” *PLoS Genet.*, vol. 3, no. 10, pp. 2023–36, Oct. 2007.

[37] M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler, “Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome,” *Nat Genet*, vol. 39, no. 4, pp. 457–466, Apr. 2007.

[38] F. Mohn, M. Weber, M. Rebhan, T. C. Roloff, J. Richter, M. B. Stadler, M. Bibel, and D. Schübeler, “Lineage-specific polycomb targets and de novo

DNA methylation define restriction and potential of neuronal progenitors.," *Mol. Cell*, vol. 30, no. 6, pp. 755–66, Jun. 2008.

[39] A. J. Bannister and T. Kouzarides, "Regulation of chromatin by histone modifications.," *Cell Res.*, vol. 21, no. 3, pp. 381–95, Mar. 2011.

[40] E. L. Greer and Y. Shi, "Histone methylation: a dynamic mark in health, disease and inheritance.," *Nat. Rev. Genet.*, vol. 13, no. 5, pp. 343–57, May 2012.

[41] D. Rossetto, N. Avvakumov, and J. Côté, "A chromatin modification involved in diverse nuclear events Histone phosphorylation," vol. 7, no. 10, pp. 1098–1108, 2012.

[42] Prof. Le Dinh Luong, "Basic Principles of Genetics," *OpenStax-CNX Web site*, 2009. [Online]. Available: http://cnx.org/content/m26565/1.1/.

[43] R. K. Chodavarapu, S. Feng, Y. V Bernatavichute, P.-Y. Chen, H. Stroud, Y. Yu, J. a Hetzel, F. Kuo, J. Kim, S. J. Cokus, D. Casero, M. Bernal, P. Huijser, A. T. Clark, U. Krämer, S. S. Merchant, X. Zhang, S. E. Jacobsen, and M. Pellegrini, "Relationship between nucleosome positioning and DNA methylation.," *Nature*, vol. 466, no. 7304, pp. 388–92, Jul. 2010.

[44] T. K. Kelly, Y. Liu, F. D. Lay, G. Liang, B. P. Berman, and P. A. Jones, "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules," pp. 2497–2506, 2012.

[45] G. Portella, F. Battistini, and M. Orozco, "Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations.," *PLoS Comput. Biol.*, vol. 9, no. 11, p. e1003354, Nov. 2013.

[46] P. a Jones and S. B. Baylin, "The epigenomics of cancer.," *Cell*, vol. 128, no. 4, pp. 683–92, Feb. 2007.

[47] E. Birney, J. a Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. a Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. a Navas, F. Neri, S. C. J. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J.

Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. a Hirsch, E. a Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W.-K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C.-L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaöz, A. Siepel, J. Taylor, L. a Liefer, K. a Wetterstrand, P. J. Good, E. a Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. a Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. a Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. H. Lee, P. Ng, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. a Singer, T. a Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. a Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. S. Haidar, Y. Yu, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. a Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. W. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrímsdóttir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. B. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. a Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. a Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong, “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.,” *Nature*, vol. 447, no. 7146, pp. 799–816, Jun. 2007.

[48] J. C. Lin, S. Jeong, G. Liang, D. Takai, M. Fatemi, Y. C. Tsai, G. Egger, E. N. Gal-Yam, and P. a Jones, "Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island.," *Cancer Cell*, vol. 12, no. 5, pp. 432–44, Nov. 2007.

[49] L. B. Hesson, V. Patil, M. a Sloane, A. C. Nunez, J. Liu, J. E. Pimanda, and R. L. Ward, "Reassembly of nucleosomes at the MLH1 promoter initiates resilencing following decitabine exposure.," *PLoS Genet.*, vol. 9, no. 7, p. e1003636, Jan. 2013.

[50] H. Han, C. C. Cortez, X. Yang, P. W. Nichols, P. a Jones, and G. Liang, "DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter.," *Hum. Mol. Genet.*, vol. 20, no. 22, pp. 4299–310, Nov. 2011.

[51] Y.-J. Kwon, S. J. Lee, J. S. Koh, S. H. Kim, H. W. Lee, M. C. Kang, J. B. Bae, Y.-J. Kim, and J. H. Park, "Genome-wide analysis of DNA methylation and the gene expression change in lung cancer.," *J. Thorac. Oncol.*, vol. 7, no. 1, pp. 20–33, Jan. 2012.

[52] T. Rauch and G. P. Pfeifer, "Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer.," *Lab. Invest.*, vol. 85, no. 9, pp. 1172–80, Sep. 2005.

[53] T. J. Seng, N. Currey, W. a Cooper, C.-S. Lee, C. Chan, L. Horvath, R. L. Sutherland, C. Kennedy, B. McCaughan, and M. R. J. Kohonen-Corish, "DLEC1 and MLH1 promoter methylation are associated with poor prognosis in non-small cell lung carcinoma.," *Br. J. Cancer*, vol. 99, no. 2, pp. 375–82, Jul. 2008.

[54] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes, revisited.," *Trends Genet.*, vol. 29, no. 10, pp. 569–74, Oct. 2013.

[55] J. C. Rogers, "DNA methylation: Molecular biology and biological significance," *Cell*, vol. 73, no. 3, p. 429, Jul. 2014.

[56] Y.-J. Kwon, S. J. Lee, J. S. Koh, S. H. Kim, H. W. Lee, M. C. Kang, J. B. Bae, Y.-J. Kim, and J. H. Park, "Genome-wide analysis of DNA methylation and the gene expression change in lung cancer.," *J. Thorac. Oncol.*, vol. 7, no. 1, pp. 20–33, Jan. 2012.

[57] S. Schwartz, E. Meshorer, and G. Ast, "Chromatin organization marks exon-intron structure.," *Nat. Struct. Mol. Biol.*, vol. 16, no. 9, pp. 990–5, Sep. 2009.

[58] M. Amit, M. Donyo, D. Hollander, A. Goren, E. Kim, S. Gelfman, G. Lev-Maor, D. Burstein, S. Schwartz, B. Postolsky, T. Pupko, and G. Ast,

“Differential GC content between exons and introns establishes distinct strategies of splice-site recognition.,” *Cell Rep.*, vol. 1, no. 5, pp. 543–56, May 2012.

APPENDICES

Some supplementary material of this study is provided in electronic format in one CD. Organization of the contents in the CD is shown in Appendix C, Appendix D, Appendix F and Appendix G.

8.1 APPENDIX A: HELICAL RISE VALUES FOR TETRANUCLEOTIDES (ADOPTED FROM [22])

Tetranucleotides	Helical Rise (Å)	Tetranucleotides	Helical Rise (Å)
AAAA/TTTT	3.21	TCCC/GGGA	4.20
AAAG/CTTT	3.17	CCCC/GGGG	4.37
AAAC/GTTT	3.05	GCCC/GGGC	2.81
AAAT/ATTT	3.38	ACCC/GGGT	3.32
TTTC/GAAA	3.07	CGGA/TCCG	3.15
CTTC/GAAG	3.04	CCCG/CGGG	3.03
GAAC/GTTC	3.12	CGGC/GCCG	2.91
ATTC/GAAT	3.36	ACCG/CGGT	3.23
CAAA/TTTG	2.89	TCCA/TGGA	3.33
CAAG/CTTG	2.98	CCCA/TGGG	3.48
CAAC/GTTG	3.03	TGGC/GCCA	2.94
ATTG/CAAT	3.27	ACCA/TGGT	3.37
TAAA/TTTA	3.08	ACAA/TTGT	2.89
CTTA/TAAG	3.12	ACAG/CTGT	2.99
TAAC/GTTA	3.27	ACAC/GTGT	3.19
ATTA/TAAT	2.88	ACAT/ATGT	2.61
AAGA/TCTT	3.17	TTGC/GCAA	2.74
AAGG/CCTT	2.97	CTGC/GCAG	3.73
AAGC/GCTT	3.20	GCAC/GTGC	2.86
AAGT/ACTT	3.77	ATGC/GCAT	3.10
TCTC/GAGA	3.19	CCAA/TTGG	3.78
CCTC/GAGG	3.24	CCAG/CTGG	3.19
GAGC/GCTC	2.80	CCAC/GTGG	2.86
ACTC/GAGT	2.95	ATGG/CCAT	2.85

CAGA/TCTG	3.92	TCAA/TTGA	2.66
CAGG/CCTG	3.09	CTGA/TCAG	3.06
CAGC/GCTG	3.39	TCAC/GTGA	3.01
ACTG/CAGT	3.40	ATGA/TCAT	2.36
TAGA/TCTA	3.21	AATA/TATT	3.23
CCTA/TAGG	3.26	AATG/CATT	3.34
TAGC/GCTA	2.96	AATC/GATT	3.23
ACTA/TAGT	3.29	AATT	3.43
AACA/TGTT	3.41	TATC/GATA	3.16
AACG/CGTT	3.42	CATC/GATG	3.23
AACC/GGTT	3.41	GATC	3.22
AACT/AGTT	3.56	CATA/TATG	3.31
TGTC/GACA	3.22	CATG	3.20
CGTC/GACG	3.90	TATA	2.94
GACC/GGTC	3.57	AGCA/TGCT	3.53
AGTC/GACT	3.66	AGCG/CGCT	3.42
CACA/TGTG	3.30	AGCC/GGCT	3.44
CACG/CGTG	3.46	AGCT	3.29
CACC/GGTG	3.00	TGCC/GGCA	3.08
AGTG/CACT	3.58	CGCC/GGCG	3.16
TACA/TGTA	3.14	GGCC	3.40
CGTA/TACG	3.25	CGCA/TGCG	4.46
TACC/GGTA	3.26	CGCG	3.35
AGTA/TACT	3.42	TGCA	3.63
AGAA/TTCT	3.15	ACGA/TCGT	3.12
AGAG/CTCT	3.01	ACGG/CCGT	2.81
AGAC/GTCT	3.08	ACGC/GCGT	2.48
ATCT/AGAT	3.02	ACGT	4.02
TTCC/GGAA	3.38	TCGC/GCGA	3.36
CTCC/GGAG	3.24	CCGC/GCGG	2.83
GTCC/GGAC	3.46	GCGC	2.61
ATCC/GGAT	3.24	CCGA/TCGG	3.33
CGAA/TTCG	3.15	CCGG	2.85
CTCG/CGAG	3.98	TCGA	3.44
CGAC/GTCG	3.34	ATAA/TTAT	2.69
ATCG/CGAT	3.29	ATAG/CTAT	2.63
TTCA/TGAA	3.24	ATAC/GTAT	2.72
CTCA/TGAG	3.03	ATAT	2.47
TGAC/GTCA	3.10	TTAC/GTAA	2.62
ATCA/TGAT	3.18	CTAC/GTAG	2.86
AGGA/TCCT	3.32	GTAC	2.90

AGGG/CCCT	2.96	CTAA/TTAG	2.60
AGGC/GCCT	2.92	CTAG	2.94
ACCT/AGGT	3.82	TTAA	3.21

8.2 APPENDIX B: PYTHON SCRIPT CODE FOR HELICAL STEP ANALYSIS

```
import array
import string
import json
import csv

seq=raw_input("Enter a promoter sequence");
tetramers=["AAAA","TTTT","AAAG","CTTT","AAAC","GTTT","AAAT",
"ATTT","TTTC","GAAA","CTTC","GAAG","GAAC","GTTC","ATTC","GAAT",
"CAAA","TTTG","CAAG","CTTG","CAAC","GTTG","ATTG","CAAT","TAAA",
"TTTA","CTTA","TAAG","TAAC","GTTA","ATTA","TAAT","AAGA","TCTT",
"AAGG","CCTT","AAGC","GCTT","AAGT","ACTT","TCTC","GAGA","CCTC",
"GAGG","GAGC","GCTC","ACTC","GAGT","CAGA","TCTG","CAGG","CCTG"
,"CAGC","GCTG","ACTG","CAGT","TAGA","TCTA","CCTA","TAGG","TAGC"
,"GCTA","ACTA","TAGT","AACA","TGTT","AACG","CGTT","AACC","GGTT",
"AACT","AGTT","TGTC","GACA","CGTC","GACG","GACC","GGTC","AGTC",
"GACT","CACA","TGTG","CACG","CGTG","CACC","GGTG","AGTG","CACT",
"TACA","TGTA","CGTA","TACG","TACC","GGTA","AGTA","TACT","AGAA",
"TTCT","AGAG","CTCT","AGAC","GTCT","ATCT","AGAT","TTCC","GGAA",
"CTCC","GGAG","GTCC","GGAC","ATCC","GGAT","CGAA","TTCG","CTCG",
"CGAG","CGAC","GTCG","ATCG","CGAT","TTCA","TGAA","CTCA","TGAG",
"TGAC","GTCA","ATCA","TGAT","AGGA","TCCT","AGGG","CCCT","AGGC",
"GCCT","ACCT","AGGT","TCCC","GGGA","CCCC","GGGG","GCCC","GGGC"
, "ACCC","GGGT","CGGA","TCCG","CCCG","CGGG","CGGC","GCCG",
"ACCG","CGGT","TCCA","TGGA","CCCA","TGGG","TGGC","GCCA","ACCA"
,"TGGT","ACAA","TTGT","ACAG","CTGT","ACAC","GTGT","ACAT","ATGT",
"TTGC","GCAA","CTGC","GCAG","GCAC","GTGC","ATGC","GCAT","CCAA"
,"TTGG","CCAG","CTGG","CCAC","GTGG","ATGG","CCAT","TCAA","TTGA"
,"CTGA","TCAG","TCAC","GTGA","ATGA","TCAT","AATA","TATT","AATG",
"CATT","AATC","GATT","AATT","TATC","GATA","CATC","GATG","GATC",
"CATA","TATG","CATG","TATA","AGCA","TGCT","AGCG","CGCT","AGCC",
"GGCT","AGCT","TGCC","GGCA","CGCC","GGCG","GGCC","CGCA","TGCG"
,"CGCG","TGCA","ACGA","TCGT","ACGG","CCGT","ACGC","GCGT",
"ACGT","TCGC","GCGA","CCGC","GCGG","GCGC","CCGA","TCGG","CCGG"
,"TCGA","ATAA","TTAT","ATAG","CTAT","ATAC","GTAT","ATAT","TTAC",
"GTAA","CTAC","GTAG","GTAC","CTAA","TTAG","CTAG","TTAA"];
```

```
helicalrise = [3.21, 3.21, 3.17, 3.17, 3.05, 3.05, 3.38, 3.38, 3.07, 3.07, 3.04, 3.04,
3.12, 3.12, 3.36, 3.36, 2.89, 2.89, 2.98, 2.98, 3.03, 3.03, 3.27, 3.27, 3.08, 3.08, 3.12,
3.12, 3.27, 3.27, 2.88, 2.88, 3.17, 3.17, 2.97, 2.97, 3.2, 3.2, 3.77, 3.77, 3.19, 3.19,
3.24, 3.24, 2.8, 2.8, 2.95, 2.95, 3.92, 3.92, 3.09, 3.09, 3.39, 3.39, 3.4, 3.4, 3.21, 3.21,
3.26, 3.26, 2.96, 2.96, 3.29, 3.29, 3.41, 3.41, 3.42, 3.42, 3.41, 3.41, 3.56, 3.56, 3.22,
3.22, 3.9, 3.9, 3.57, 3.57, 3.66, 3.66, 3.3, 3.3, 3.46, 3.46, 3, 3, 3.58, 3.58, 3.14, 3.14,
3.25, 3.25, 3.26, 3.26, 3.42, 3.42, 3.15, 3.15, 3.01, 3.01, 3.08, 3.08, 3.02, 3.02, 3.38,
3.38, 3.24, 3.24, 3.46, 3.46, 3.24, 3.24, 3.15, 3.15, 3.98, 3.98, 3.34, 3.34, 3.29, 3.29,
3.24, 3.24, 3.03, 3.03, 3.1, 3.1, 3.18, 3.18, 3.32, 3.32, 2.96, 2.96, 2.92, 2.92, 3.82,
3.82, 4.2, 4.2, 4.37, 4.37, 2.81, 2.81, 3.32, 3.32, 3.15, 3.15, 3.03, 3.03, 2.91, 2.91,
3.23, 3.23, 3.33, 3.33, 3.48, 3.48, 2.94, 2.94, 3.37, 3.37, 2.89, 2.89, 2.99, 2.99, 3.19,
3.19, 2.61, 2.61, 2.74, 2.74, 3.73, 3.73, 2.86, 2.86, 3.1, 3.1, 3.78, 3.78, 3.19, 3.19,
2.86, 2.86, 2.85, 2.85, 2.66, 2.66, 3.06, 3.06, 3.01, 3.01, 2.36, 2.36, 3.23, 3.23, 3.34,
3.34, 3.23, 3.23, 3.43, 3.16, 3.16, 3.23, 3.23, 3.22, 3.31, 3.31, 3.2, 2.94, 3.53, 3.53,
3.42, 3.42, 3.44, 3.44, 3.29, 3.08, 3.08, 3.16, 3.16, 3.4, 4.46, 4.46, 3.35, 3.63, 3.12,
3.12, 2.81, 2.81, 2.48, 2.48, 4.02, 3.36, 3.36, 2.83, 2.83, 2.61, 3.33, 3.33, 2.85, 3.44,
2.69, 2.69, 2.63, 2.63, 2.72, 2.72, 2.47, 2.62, 2.62, 2.86, 2.86, 2.9, 2.6, 2.6, 2.94,
3.21];


a= [];
for i in range(0 , (len(seq)-3)):
        for j in range(0,len(tetramers)):
                if tetramers[j]==seq[i:i+4]:
                        a.append(helicalrise[j]);

with open('FILE NAME', 'wb') as f:
  spamwriter = csv.writer(f, delimiter=',',
                    quotechar='|', quoting=csv.QUOTE_MINIMAL);
  spamwriter.writerow(a);
```

8.3 APPENDIX C: HELICAL RISE ARRAYS THAT ARE TRANSLATED FROM A, B, C, A+B, A+B+C REGIONS OF GENE PROMOTER SEQUENCES

Folder A: Hypermethylated gene promoters' helical rise values
Folder B: Hypomethylated gene promoters' helical rise values
Folder C: Housekeeping gene promoters' helical rise values

8.4 APPENDIX D: SCATTER PLOTS OF THE GENE PROMOTERS' HELICAL RISE VALUES

Folder D: Scatter plot of hypermethylated genes
Folder E: Scatter plot of hypomethylated genes
Folder F: Scatter plot of housekeeping genes

8.5 APPENDIX E: MEAN HELICAL RISE OF THE GENE PROMOTER REGIONS

Hypermethylated Genes					
	Mean Helical Rise (Å)				
Gene	**A**	**B**	**C**	**A+B**	**A+B+C**
ADAM33	3.19822467	3.244608	3.246149	3.21300601	3.22009514
AGTR1	3.14705115	3.197746	3.258972	3.16371409	3.18562344
APOB48R	3.20251755	3.271952	3.271149	3.2252839	3.23503756
ASTN2	3.21293882	3.289437	3.194315	3.23843687	3.22557837
CA10	3.17612839	3.217364	3.256331	3.19012692	3.20489234
CCDC37	3.21905717	3.221006	3.256028	3.22042084	3.2281973
CDO1	3.16560682	3.188511	3.241956	3.17287909	3.18852779
CLTCL1	3.19338014	3.245453	3.239496	3.21112224	3.21614422
COL13A1	3.20534604	3.243099	3.251754	3.21915832	3.22612419
CTSE	3.2105015	3.212696	3.263065	3.21085504	3.22198798
CYTL1	3.21446339	3.26332	3.203286	3.23092184	3.22238358
DLEC1	3.19134403	3.180664	3.253004	3.18765531	3.20238358
DNM3	3.16547643	3.202254	3.27125	3.17711423	3.19992989
GUCA2B	3.18462387	3.234427	3.241694	3.20229793	3.21050075
HIST1H1B	3.18532598	3.17161	3.184315	3.18057448	3.17971457
HIST1H3A	3.1869007	3.19674	3.203952	3.19021376	3.19157236
HOPX	3.18551655	3.190865	3.208468	3.18773547	3.19123686
KANK2	3.18748245	3.224889	3.291371	3.2000334	3.22142213
KLRC4	3.131334	3.139074	3.192399	3.13416166	3.14702554
LMO3	3.16945838	3.203078	3.22121	3.18050768	3.19016024
MYH2	3.17797392	3.190423	3.209435	3.18132933	3.18675013
NID1	3.20691073	3.245412	3.25748	3.21961256	3.22835253
PAX9	3.26903711	3.198833	3.273004	3.24580494	3.25135704
PGC	3.18296891	3.195231	3.184536	3.18651971	3.18434652
PPP1R14A	3.18148445	3.224085	3.261411	3.19513026	3.21018027
RGS5	3.17746239	3.227264	3.205423	3.19394122	3.19527291

SLC1A2	3.19231695	3.239598	3.26756	3.2088644	3.22164747
SLIT2	3.23054162	3.263823	3.211794	3.24124916	3.23211317
TCAP	3.26397192	3.237042	3.274315	3.25424182	3.25739109
TMEM146	3.21863591	3.208571	3.233427	3.21565798	3.21849775

Hypomethylated Genes					
Mean Helical Rise (Å)					
Gene	**A**	**B**	**C**	**A+B**	**A+B+C**
ADSSL1	3.2658170	3.2787525	3.2470624	3.2698330	3.2640711
CBS	3.2100500	3.2005433	3.2806237	3.2069873	3.2252829
CCL7	3.1711530	3.1729779	3.2582495	3.1713895	3.1929294
CDCA5	3.2217250	3.2500000	3.2304628	3.2311623	3.2305508
COL1A1	3.2478940	3.3038833	3.2293763	3.2668671	3.2572308
EDN2	3.2303510	3.2612274	3.2396378	3.2408617	3.2407261
ELMO3	3.2500500	3.2264185	3.2874648	3.2425117	3.2538157
GBX2	3.2228280	3.2294769	3.2053722	3.2248096	3.2196345
HOXD11	3.1877930	3.2597384	3.2471026	3.2120107	3.2207411
KCNC1	3.2246240	3.2370825	3.2203823	3.2286640	3.2270856
KRTCAP3	3.2201910	3.2686318	3.2267203	3.2372278	3.2348723
LASP1	3.2328890	3.2519517	3.2423742	3.2389512	3.2400601
MT1B	3.1836610	3.2334809	3.2396781	3.2002672	3.2101753
NETO2	3.1981240	3.2079074	3.2039437	3.2009686	3.2016074
NSDHL	3.1925880	3.2222938	3.2537827	3.2021376	3.2154882
PAGE4	3.1880540	3.2132998	3.2306841	3.1963794	3.2046720
PDCL2	3.1943130	3.1928370	3.2160563	3.1937074	3.1991738
PRIM1	3.2109030	3.2058551	3.1874044	3.2087308	3.2040110
PSMA6	3.1754060	3.2008048	3.1664386	3.1838477	3.1794442
SERPINB5	3.1785560	3.1804829	3.2068410	3.1787375	3.1857186
SLC35F3	3.1577930	3.1831388	3.2304829	3.1654843	3.1814522
SPC25	3.1628080	3.1848692	3.2247887	3.1699733	3.1832599

Housekeeping Genes					
Mean Helical Rise (Å)					
Gene	**A**	**B**	**C**	**A+B**	**A+B+C**
ACTBP2	3.17883651	3.164044	3.1934	3.17350701	3.178438
ACTBP7	3.184082247	3.146278	3.204245	3.17187709	3.179885
ALB	3.155827482	3.158934	3.166559	3.15664663	3.159424
GAPDH	3.249448345	3.261006	3.280181	3.25293921	3.259559
PPIAL4E	3.148184554	3.165634	3.18	3.15380094	3.160341
PPIAP30	3.187642929	3.178028	3.189215	3.18418838	3.185158
RNA18S5	3.234754263	3.207082	3.218732	3.22530394	3.223801
RNA28S5	3.268074223	3.209879	3.203883	3.24796927	3.236855
TUBA1A	3.181594784	3.201790744	3.215211268	3.188189713	3.194952429
TUBA3FP	3.249458375	3.252113	3.259537	3.25032732	3.252283

8.6 APPENDIX F: SCATTER PLOTS OF MEAN HELICAL RISE VALUES OF 40 PROMOTER SUB-REGIONS

Folder G: Mean helical rise scatterplots of hypermethylated genes
Folder H: Mean helical rise scatterplots of hypomethylated genes
Folder I: Mean helical rise scatterplots of housekeeping genes

8.7 APPENDIX G: HISTOGRAMS OF HELICAL RISE VALUES

Folder J: Histograms for hypermethylated genes
Folder K: Histograms for hypomethylated genes
Folder L: Histograms for housekeeping genes

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ☐

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : .....
Adı : .....
Bölümü : .....

TEZİN ADI (İngilizce) : .....
.....
.....
.....
.....

TEZİN TÜRÜ : Yüksek Lisans ☐ Doktora ☐

1. Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. ☐
2. Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullancılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐
3. Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

Yazarın imzası ..... Tarih .....