5-2008

DNA Methylation and Promoter Sequence Analysis of Colon Cancer Genes

Fang Wang
Clemson University, fangw@clemson.edu

DNA METHYLATION AND PROMOTER SEQUENCE ANALYSIS OF COLON CANCER GENES

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Genetics

By
Fang Wang
May 2008

Accepted by
Dr. Chin-Fu Chen, Committee Chair
Dr. Feng Luo
Dr. Jeff Tomkis
Dr. Weiguo Cao

ABSTRACT

Aberrant DNA methylation in the genome is associated with human cancers. Tumor cells can undergo an overall loss of DNA methylation in non-coding repetitive regions (including the Alu elements) and at the same time maintain hypermethylation in the CpG islands in promoter region of multiple genes. We are interested in understanding the pattern of DNA methylation in the promoters of genes that may mediate the tumor genesis in colon cells. Recent literature survey suggested that about two dozens of genes, when mutated or undergone changes in DNA methylation, can directly promote the development of colon cancers. We downloaded the promoter sequences of these possible colon cancer causing genes from the Human Genome Browser for bioinformatics analysis and investigation of the sequence features on their Alu elements, transcription factor binding sites, and CpG islands. Our results suggest that sequence conservation of the flanking transcription factor binding sites plays an important role in protecting their flanking CpG islands from hypermethylation. In addition, colon cancer genes may harbor a lower density Alu element when compared to control random genes. We selected 4 characteristics of DNA sequences for machine learning and prediction of colon cancer genes.

DEDICATION

I would like to dedicate this work to my parents, Yan-Ning Wang and Ping Li. This work is not my own, but ours.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Chin-Fu Chen, for his continuous support throughout my graduate years. You are the definition of compassion and understanding. I am honored to be in your group to pursue my master study.

I would like to thank Dr. Feng Luo, who provides me a lot of intelligent insight and guidance especially at the beginning stage of this project. Thanks for your support and help always, thank you for serve as my committee advisor.

To Dr. Weiguo Cao and Dr. Jeff Tomkins, for serving as my committee advisors. Your valuable suggestions were vital to this project.

To MS. Rooksie Noorai and Mr. Chun-huai Cheng, for providing me a lot of help throughout this project.

To MS. Debbie Botki, for training my English.

TABLE OF CONTENTS

Table of Contents (Continued)

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

List of Figures (Continued)

Figure Page

CHAPTER ONE

LITERATURE REVIEW

Introduction

Cancer results from an accumulation of genetic changes within a cell that allow uncontrolled cell growth. Not associated with DNA sequences change, DNA methylation represents an epigenetic means of inheritance. Cytosine methylation is the most common epigenetic modification of DNA in human cells. In humans, within tandem and interspersed repeats in CpG rich region, lies most of the methylated cytosines. Alu elements are the most common repeat sequence family. Demethylation of Alu occurs often in aging and cancer development [1]. It is thought that the frequent hypomethylation of repetitive elements is responsible for the global hypomethylation seen in diverse human cancers [2]. CpG islands mostly locate in promoter region of the genes, and are hypermethylated in human cancers [9].

Alu subfamilies' distribution in colon cancer genes

The term "repetitive element" refers to DNA sequence, with multiple copies of the same repetitive nucleotides present in the genome. Repetitive elements are categorized into two groups. One group is tandemly arrayed sequence such as microsatellites, minisatellites, and telomeres. The other group is interspersed sequence such as mobile element and pseudo genes. Interspersed sequence could be divided on the basis of size. SINE (short Interspersed element) belongs to interspersed sequence, and is less than 500 bp long [3]. Alu SINEs were identified almost 30 years ago [4]. The Alu sequence was so named because it contains a recognition site for the restriction enzyme AluI [4].

A series of Alu subfamilies of different ages have evolved from the propagation of Alu elements to more than one million copies during the past 65 million years. Each Alu element consists of a ~300 base pair long sequence, with 5' half (containing an RNA polymerase III promoter) and 3' terminus (containing a stretch of As which sometimes interspersed with C, G or Ts) [3].

It is thought that a processed 7SL RNA provides the ancestral element named "FAM" (fossil Alu monomer) for Alu element [19]. FAM later evolved into two sequence variants named FLA (free left Alu monomer) and FRA (free right Alu monomer). With the fusion of an FRA and an FLA, the progenitor of the Alu family appears (Figure 1.1). During the fusion, the FLA remains its internal RNA polymerase III promoter. Each Alu element monomers are separated by a stretch A region with its 3' terminus a poly-A tail [19].

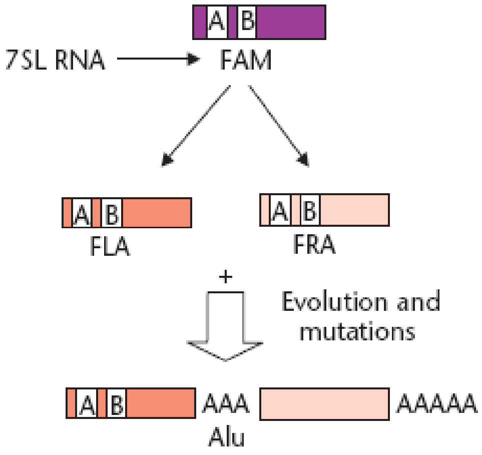


Figure1.1: The origin of Alu elements (adopted from [19]). Alu element originates from a processed 7SL RNA, which evolved to FAM. FLA and FRA are two variants of FAM. FLA and FRA fuse to form Alu sequences.

Alu element amplification is believed to be a result of retrotransposition. The process includes: (1) producing transcripts of RNA polymerase III in Alu and (2) reverse transcription of the transcripts. In order for amplification to exist, Alu element needs to "borrow" the factors which are necessary in amplification since Alu element does not have open reading frames. Usually Alu element borrows such factors from long interspersed elements (LINEs) [6].

Only a very few human Alu elements have the ability of retrotransposition, so they are called source genes or "master" [7] genes. Because it lacks the appropriate flanking sequences, the internal RNA polymerase III promoter in Alu copies does not transcribe actively in vivo [8]. As a result, usually new Alu copies in human genome are not functional in retrotransposition, except when they happened to land in a region where the active ability to the incomplete RNA polymerase III promoter is provided. Even when Alu copies are transposed in the "lucky" region, they also have a short life. There are two possible reasons for this: (1) individual Alu copy contains more than 24 CpG dinucleotides, which are easy to mutate because of the deamination of 5methy-cytosine residue[10]. The fact that mutations often occurred in the CpG dinucleotide leads to elimination of the new Alu element's retrotransposition ability, (2) a defect in A-rich tail of Alu element could also contribute to the failing of the retrotransposition of the new Alu element. This is because after a new Alu element integrates into a new region, the stretch-A tail of individual Alu copy (which has been proven important in the amplification process) may also easily become mutated and change to simple sequence [14-16]. In summary, it is very hard for individual Alu copy to expand the copy number of Alu element [7].

Only in the "master" gene, the accumulating mutations are passed on to their copies. Hence, characterized through a series of hierarchical mutations, several distinct Alu subfamilies, on different genetic age, comprise the human Alu element family. There are three major Alu subfamilies: AluY, AluJ and AluS. Under each major Alu subfamily, there exist small subfamilies (Figure 1.2).

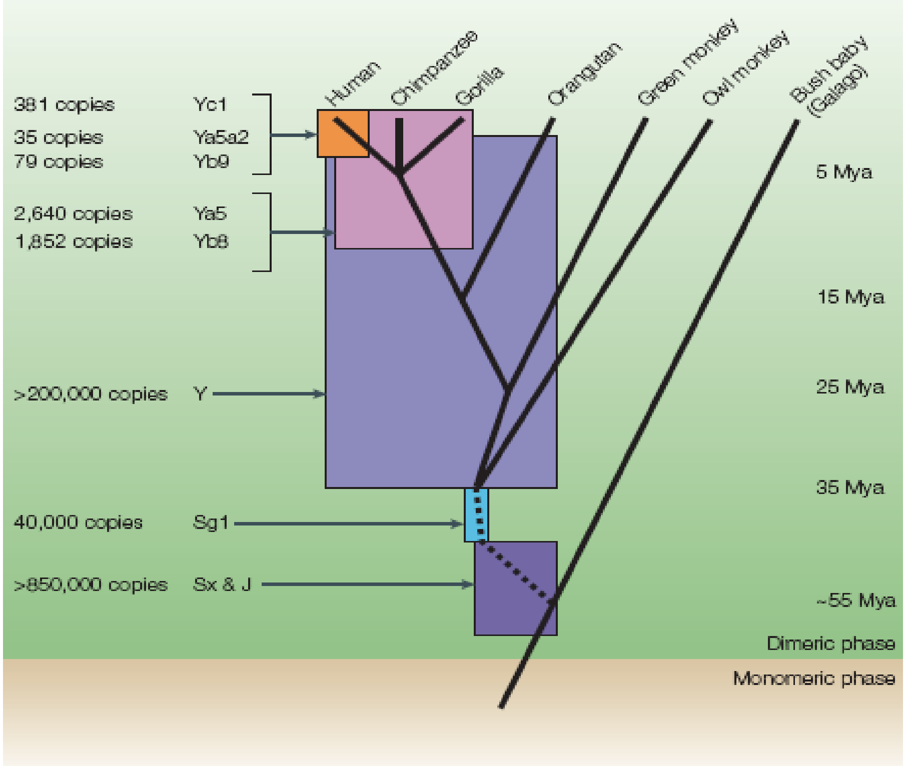


Figure 1.2: Alu elements expansion in primates (adopted from [3]). There are mainly three major Alu subfamilies: AluY, AluS, AluJ. Each Alu subfamily is colored according to the times of amplification peak. Dark purple: AluJ, light blue: AluS, light purple: AluY. Each Alu subfamily's copy number is also listed according to the time.

Among different Alu sequences, there are commonly high proportions of correlated base occurrences in some certain positions. This can be used to separate Alu element into different Alu subfamilies [5]. The fact that many human Alu elements share common diagnostic sequences has been proven in some laboratories [64]. In addition, different Alu subfamily also contains different consensus sequence (Figure 1.3).

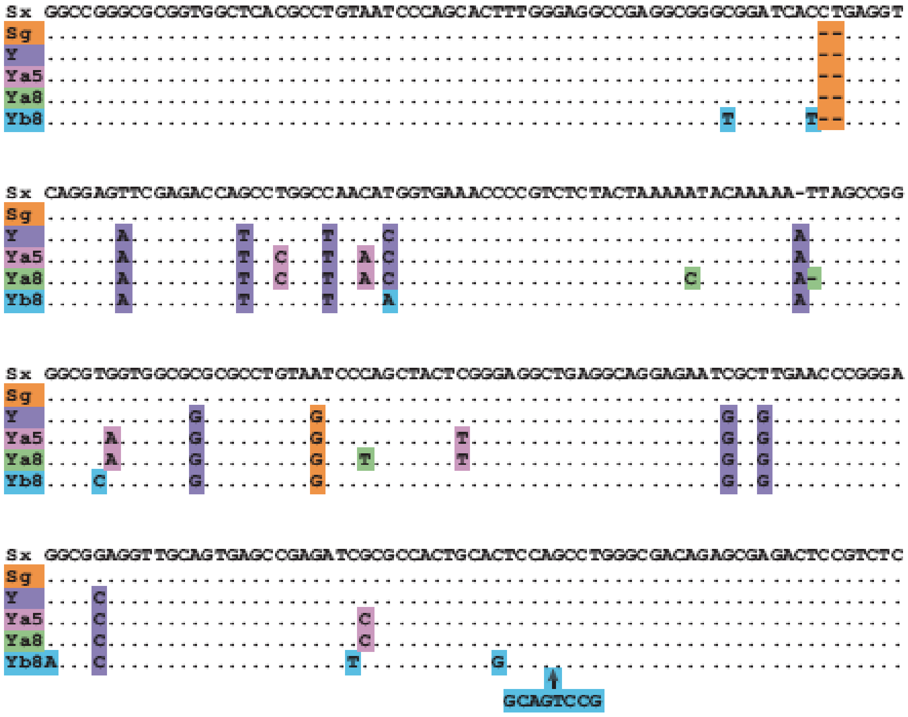

Sx GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTGAGGT
Sg
Y
Ya5
Ya8
Yb8
Sx CAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAA-TTAGCCGG
Sg
Y
Ya5
Ya8
Yb8
Sx GGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGA
Sg
Y
Ya5
Ya8
Yb8
Sx GGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTC
Sg
Y
Ya5
Ya8
Yb8A
GCAGTCCG

Figure 1.3. Consensus sequence in the alignment of Alu subfamilies [64]. The consensus sequence for Alu small subfamily Alusx is shown at the top. The younger Alu subfamily is progressively underneath the Alusx sequence. Dot refers to the same base as it is in the consensus sequence. Dash refers to deletion comparing to the consensus sequence. Colored box refers to mutation. The newer subfamily (such as Ya5) not only has all the mutations that belong to the ancestral Alu element, but also has several extra mutations which are the diagnostic positions for the particular Alu subfamily.

Previous studies showed that there is an inverse relationship among Alu families based on the age of the Alu family and its methylation status. The younger Alu family usually exhibits higher methylation rates, compared to the older Alu family, which implies that there is a stronger silencing pressure put on the younger Alu elements [18].

Alu element acts as de novo methylation center

It has been suggested that the high level of methylation of Alu elements inhibit gene transcription [20, 21]. It is also suspected that Alu elements might induce de novo methylation. SINEs genomic region were found in some cases to induce de novo methylation of the genomic sequences nearby [32]. In neoplasia, tumor suppressor genes, such as TP53 gene [23, 24] play an important role in tumor genesis. Alu human sequences were believed to be the potential de novo methylation centers in these tumor suppressor genes [32].

Depending on the distance between nearest Alu element to the transcription start site in human genome, there is a linear or inverse correlation between the size of CpG island and density of the Alu elements [17]. If the distance between the transcription start site and nearest Alu element is less than 2000 base pairs, the longer the CpG island, the lower density the Alu element has; if the distance between transcription start site and nearest Alu element is longer than 2000 base pairs, the longer the CpG island, the higher density the Alu element has [17]. There also is a linear relationship between CpG island length and the distance from the transcription start site to the nearest retroelements [17]. A transitional area is located between retro-elements and the CpG island. Methylation can be expended from retro-element to the CpG island through transitional area (Figure 1.4).

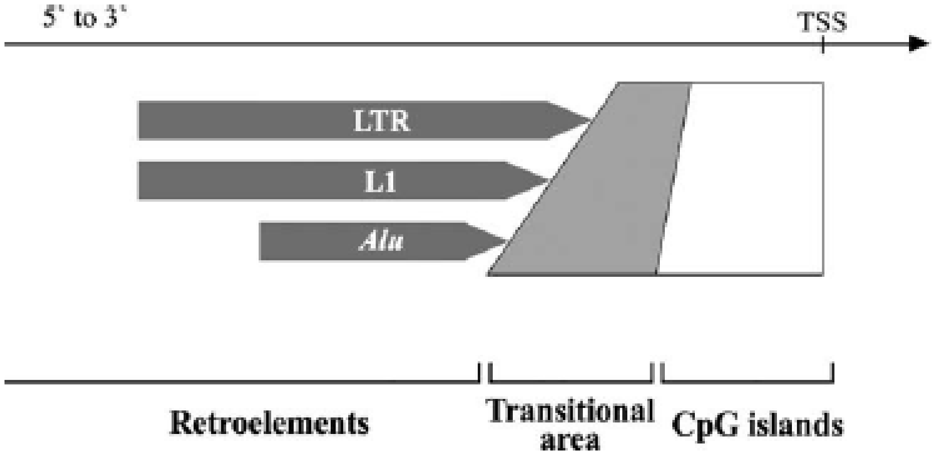


Figure 1.4. Methylation is extended from retrotransposons to transitional area and CpG islands near transcription start site [17]. The extent of the methylation spreading could be predicted by the length of the CpG island. The longer the CpG island, the further away the retroelement is from the transcription start site.

CpG island protection model

CpG islands in the promoter regions are the only regions in genome where the frequency of CpG dinucleotides is close to the expectations [17]. In other regions, the frequency of CpG dinucleotides is much lower than in CpG islands. CpG islands contain several characteristics [17]: (1) A sequence at least 200 base pair long, (2). A GC percentage greater than 50%, (3) Observed/expected CpG ratio greater than 60%. Most CpG islands are usually located very near to the transcription start site of housekeeping genes in mammals. One possible mechanism for silencing tumor suppressor genes in cancer cells is the aberrant hypermethylation of CpG islands in the promoter regions of these tumor suppressor genes [25]. DNA methylation, an epigenetic modification, gives rise to the 5-methylated cytosine by adding a methyl-group to the fifth carbon of the cytosine. In normal cells, CpG dinucleotide in CpG islands usually is unmethylated in the promoter region, but methylated in coding region of the gene.

It has been proposed that Sp1 transcription factor binding sites can behave as the boundaries which could prevent the CpG islands from methylation from these methylation centers [31]. Subsequently, the methylated CpG dinucelotides could not diffuse their methylation to the unmethylated nearby CpG dinucleotides [29]. Tumor genesis could be stopped by the Sp1 protection to the CpG dinucleotide in the promoter region from methylation (Figure 1.5).

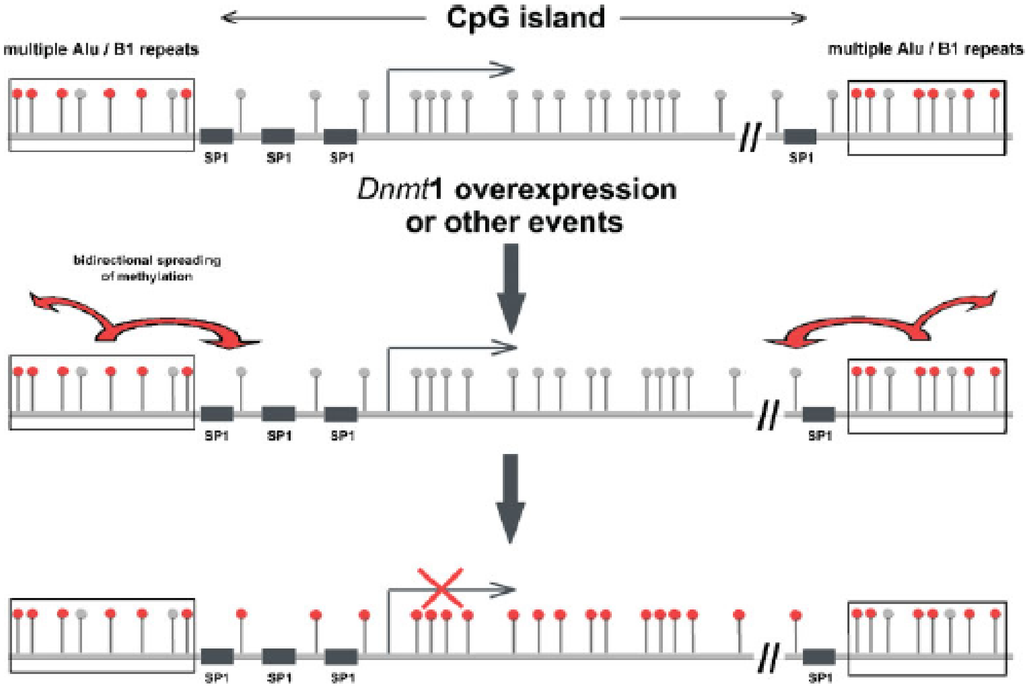


Figure 1.5. Sp1 protection model (adopted from [31]). Alu elements are "de novo methylation centers" which pass its methylation to the nearby CpG islands. Sp1 transcription factor binding sites flank the CpG island, protecting it from the expanded methylation from the methylation centers. In this diagram, grey circles refer to the unmethylated cytosines, red circles refer to the methylated cytosines.

The Sp1 transcription factor binding sites are not present in the promoter region of all cancer-related genes, so it is reasonable to predict that other transcription factor binding site consensus sequence which flanks the CpG island in promoter region could also behave as boundaries to protect the CpG island from methylation from the nearby methylation centers (Figure 1.6). In the experiment on T24 bladder carcinoma cell line, after demethylation was achieved by adding 5-aza 20-deoxycytidine, the time for remethylation is prolonged, which proves the protective effect on the CpG island by the flanking of the transcription factor binding sites [30]. In this T24 bladder carcinoma cell line experiment, after demethylation, the remethylation of p16 exon 2 CpG island is more rapidly than the p16 promoter CpG island, Sp1 sites within the CpG island in p16 promoter region is thought to play an important role in protecting the CpG island from de novo methylation.

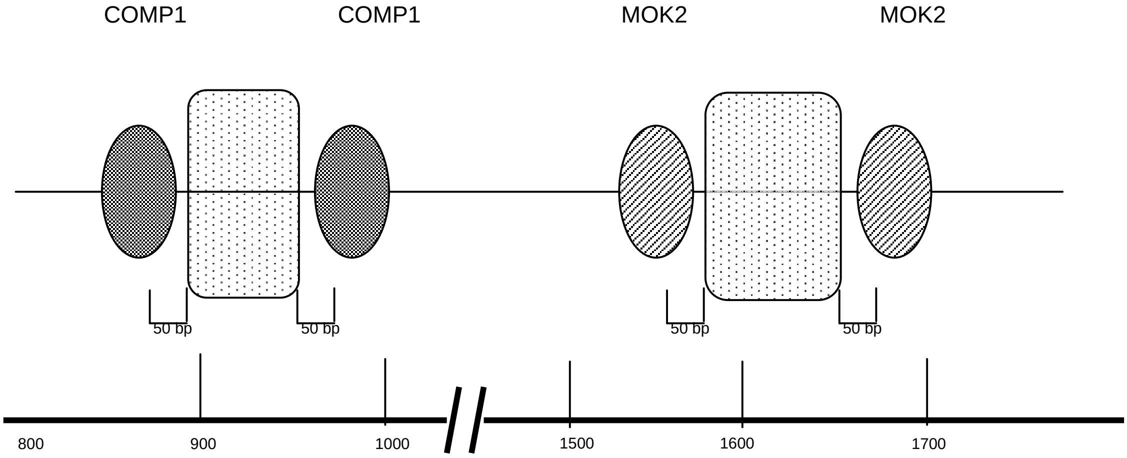


Figure 1.6: A protection model for gene MAL. Squares refer to the two CpG islands found in the promoter region of the MAL gene. The paired circles refer to two pairs of transcription factor binding sites which flank the CpG island within 50 base pairs, protecting the CpG island from methylation. One pair of transcription factor binding sites is COMP, the other pair of transcription factor binding sites is MOK2.

CpG and TpG dinucleotides density

A guanine following a cytosine is not common in vertebrate DNA sequence. During evolution, through deamination, the methylated cytosine is transformed into thymine, and the unmethylated cytosine changed into uracil. Subsequently, the DNA repair mechanism treats uracil as extraneous base in DNA and substitute it, because uracil only occurs in RNA. 5-methylated cytosine (5-meC) deamination produces thymine (T), which creates C.G to T.A transition. CpG becomes TpG. This transition will remain in the DNA sequence without the recognition and repair by the repair enzyme. Since the transition C->T creates "damage" T which still belongs to a normal DNA base, it requests higher level recognition by the repair enzyme [33, 43].

CHAPTER 2

METHODS AND DATA

Data

Experimental evidence shows that hypermethylation of the CpG island in the promoter region of multiple genes is closely associated with colon cancer happening [34-39]. We have selected 24 colon cancer genes from literature: TAC1, TBXA2R, PTGS2, MLH1, FAT, SFRP1, ENG, DKK1, CALCA, RBP1, STK11, GATA5, AKAP12, EPHB2, GATA4, LMX1B, TMEFF2, CDKN2A, WIF1, SST, WRN, NELL1, MAL, RARB (Appendix table 1), we will call these 'colon cancer genes' in this study.

We also selected 166 genes with hypermethylated promoters that are associated with colorectal cancer [40]. These 166 colorectal cancer genes will be used as independent controls to investigate the difference between colon cancer related genes and normal genes. We will call these 'colon cancer related genes' in this study.

For normal, non-cancer genes, 90 genes in chromosome 21 were obtained from the study by Yamada and co-workers [41]. 75 genes with unmethylated promoter, 15 genes with methylated promoter. The genomic DNA in the experiment was received from human peripheral blood leukocytes and placental tissue. In current study, promoter sequences will be called ‘Yamada methylated genes’ and ‘Yamada unmethylated genes’.

Weber and co-workers studied the methylation ratio of about 16,000 promoters in fibroblast cells [42]. 23 out of 24 of our colon cancer genes could be found in Weber study. As a negative control, based on the respective methylation ratio of the hypermethylated 23 colon cancer genes, we randomly selected 50 genes which contain the similar methylation ratio as the 23 colon cancer genes from Weber’s data. In addition, 48 random genes with hypermethylation in their promoters from the Weber data were also selected for evaluation of the performance of the classifiers in machine learning during the later stage of this study.

Finally, as part of the performance evaluation for classifiers in machine learning, we selected 147 genes that are labeled as colon cancer related genes in NCI cancer gene database [69].

In total, five groups of data were used for this study: (1) 24 colon cancer genes (2) 166 hypermethylated ‘colon cancer related genes’. (3) 75 Yamada unmethylated genes (4) 15 Yamada methylated genes (5) 50 control random genes.

Methods

Alu sequence prediction and assembly

Our operational definition of promoter region is the genomic DNA sequence 2000 bp upstream of the transcription start site (TSS) of a gene. Promoter sequences upstream the transcription start sites are downloaded from Ensembl Biomart genome browser [65].

Alu element prediction was performed by the on-line software "Censor" [44, 66], using option of masking "N" symbols, and reporting the classification of the repetitive sequences. The data assembly was achieved by Perl programming.

Transcription factor binding site prediction

From a computational biology view, transcription factor binding sites (TFBSs) can be represented by position weight matrices PWM [47], which describe the TFBS by giving the complete nucleotide distribution for each single position in the TFBS. We use the software MatInspector [46, 68] for TFBS prediction. MatInspector, based on a large database of matrix for different transcription factor binding sites, locates matches in the query DNA sequences and assigns percentage of similarity to matches to evaluate the matches.

CpG island prediction and assembly

CpG island prediction was achieved through CpGPlot predictor [45, 67]. The criteria for CpG islands were: length longer than 50 bp, GC content larger than 50% and CpG Expected/Observed ratio greater than 60%.

Classification, Feature selection, Prediction and Gene Ontology

Classification

The procedure for classification is usually based on a number of characteristics. The training procedure is to find a decision rule which could explain the data set well.

Because we didn't have a large collection of data (i.e. we have only 24 colon genes), as in most biological studies, using independent data for test set is impractical. One common technique to partially increase the robustness of the classification is to use the method of cross validation [50].

Cross-validation is a statistical practice which partition a data sample into subsets, then the analysis is initially performed on a single subset, the other subsets are kept for subsequent practice in volition of the initial analysis. We used the method of 10 folds cross-validation in which the original data sample is partitioned into 10 sub-samples. One sub-sample is kept as validation data for testing the model, called testing set, the remaining nine sub-samples are used as training data. The process of cross-validation is repeated ten times (10 folds), each of the 10 sub-sample is used once as the testing set. Then the ten results are averaged to create the final estimation [50].

We have used two classifiers and compared the performance of the two methods.

Classifier I: Naïve Bayes

The Naïve Bayes classification is based on Bayes theorem. Naïve Bayes method combines the probability of each feature on the class label, assuming the independence between the features, calculates the probability of correctness of hypothesis. For example, the Naïve Bayes classifier will hypothesize that: X1, X2 are cancer genes, then it calculates all the probabilities and chooses the highest probability [59].

Classifier II: Decision tree

A decision tree is the algorithm which takes the input situation by a list of attributes, and produces a yes/no decision, displaying graphically the relationships underlying the data. Decision tree works by recursively partitioning the training set until each partition is composed of examples entirely from one class. Non-leaf nodes in the tree represent a splitting point which determines how the data is partitioned. The partition proceeds until inside each partition all the instances belong to the same class [60].

We used the software Weka (Waikato Environment for Knowledge Analysis) [49] for implementation of Naïve Bayes and Decision Tree classifiers. We used the programs NaiveBayes for Naïve Bayes and J48 for decision tree, respectively, in WEKA.

Feature selection

Feature selection could help improve the performance of the classification model by getting rid of some unnecessary redundant features. It is usually achieved by building classification models by choosing some more important features out of a list of features.

Gene Ontology

The Gene Ontology [22] is a controlled biological vocabularies (i.e. ontologies) that describe gene products based on their functionalities in the cell. In order to obtain more insights, we use gene ontology (GO) to annotate the gene list and separate the predicted colon cancer gene candidates from the non-colon cancer gene candidates. We used the on-line software "FatiGO" [64, 70] for GO annotation.

CHAPTER 3

ALU ELEMENT STUDY

Alu subfamily distribution is significantly different between colon cancer and other normal genes

Our results suggest that in the 24 colon cancer genes, AluS, AluJ and AluY subfamilies all occurred, and the Alu subfamily distribution proportion is similar to Yamada unmethylated genes (Figure 3.1). In Yamada methylated genes, AluY subfamily is absent for every single gene, but enriched with AluS subfamily (88% of genes). In random genes, Alu distribution is different from other groups of genes. They contain all the three major Alu subfamilies with the AluS subfamily having the highest proportion (Figure 3.1).

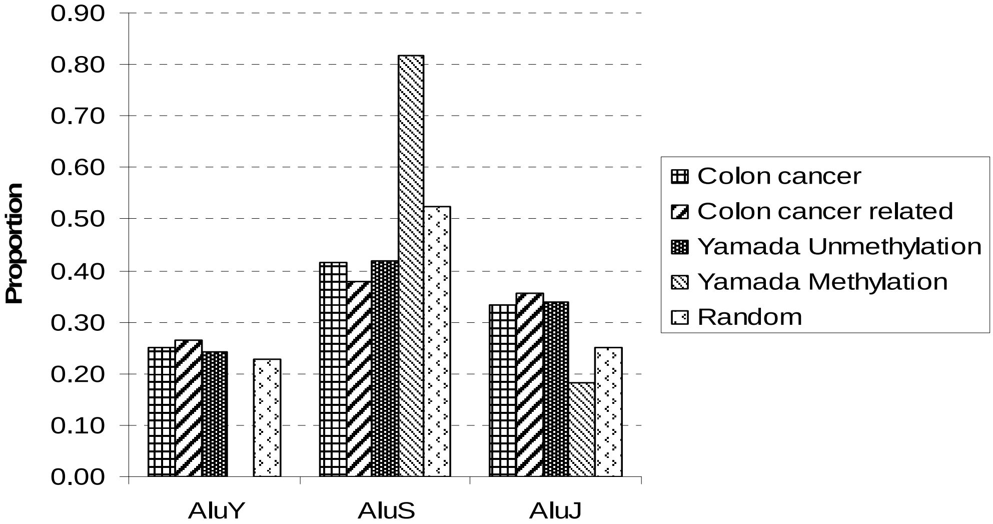


Figure 3.1: Alu family members' distribution. The average proportion of each Alu subfamily in each group is calculated by the value: (number of Alu element belonging to one subfamily) / (total number of Alu elements).

For the average number of Alu element per gene, there is no significant difference between colon cancer genes and any other group of genes (Figure 3.5).

The average number of AluY subfamily per gene in colon cancer genes is the lowest compared to Yamada unmethylated (*P* <0.05), colon-related (*P* <0.05) and Yamada methylated genes (*P* <0.01) (results from a parewise t-test, respectively). The difference in AluY subfamily distribution between colon cancer genes and Yamada methylated genes is the largest (Figure 3.2). For the average number of AluJ subfamily, there is a significant difference between colon cancer genes and Yamada methylated genes, between colon cancer genes and colon cancer related genes (Figure 3.3). The average number of AluJ subfamily per gene in colon cancer genes is as low as Yamada methylated genes, both being lower than other groups of genes (Figure 3.3). For AluS subfamily, there is a significant difference between colon cancer genes and Yamada unmethylated genes, between colon cancer genes and colon cancer related genes (Figure 3.4). The average number of AluS subfamily per gene in colon cancer is lower than in other groups (Figure 3.4).

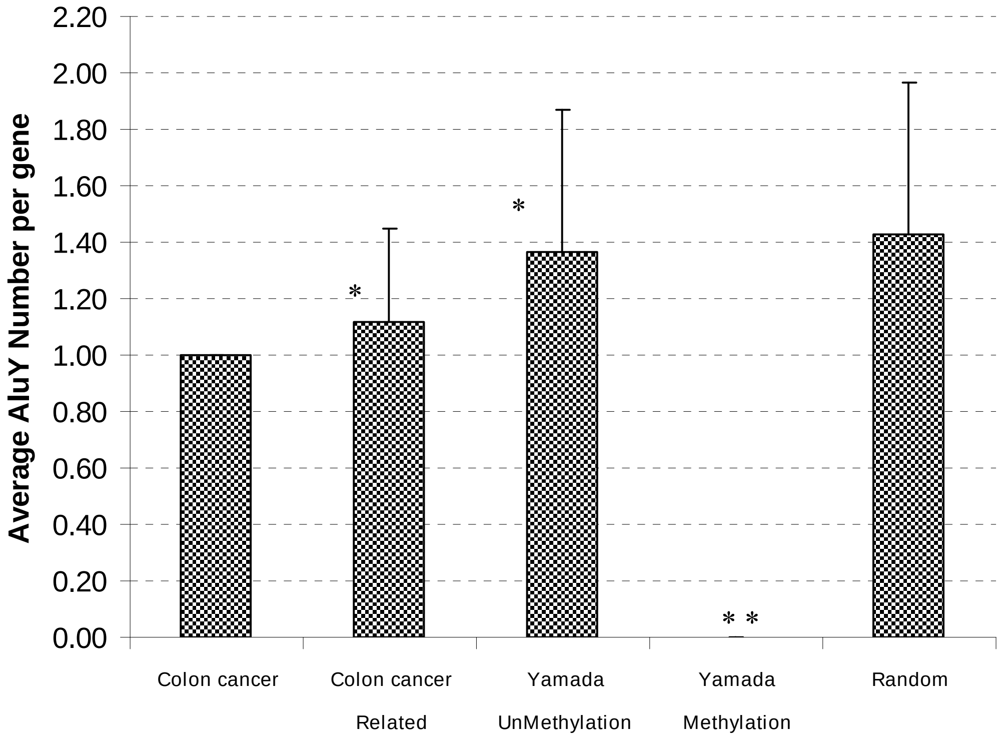


Figure 3.2: Average number of AluY elements. The average number of AluY element per gene in each group is calculated by the value: (total number of AluY elements) / (total number of genes containing AluY elements). Pairwise t-tests were performed between

colon cancer genes and each of the other four groups of genes. * denotes $P < 0.05$; ** denotes $P < 0.01$.

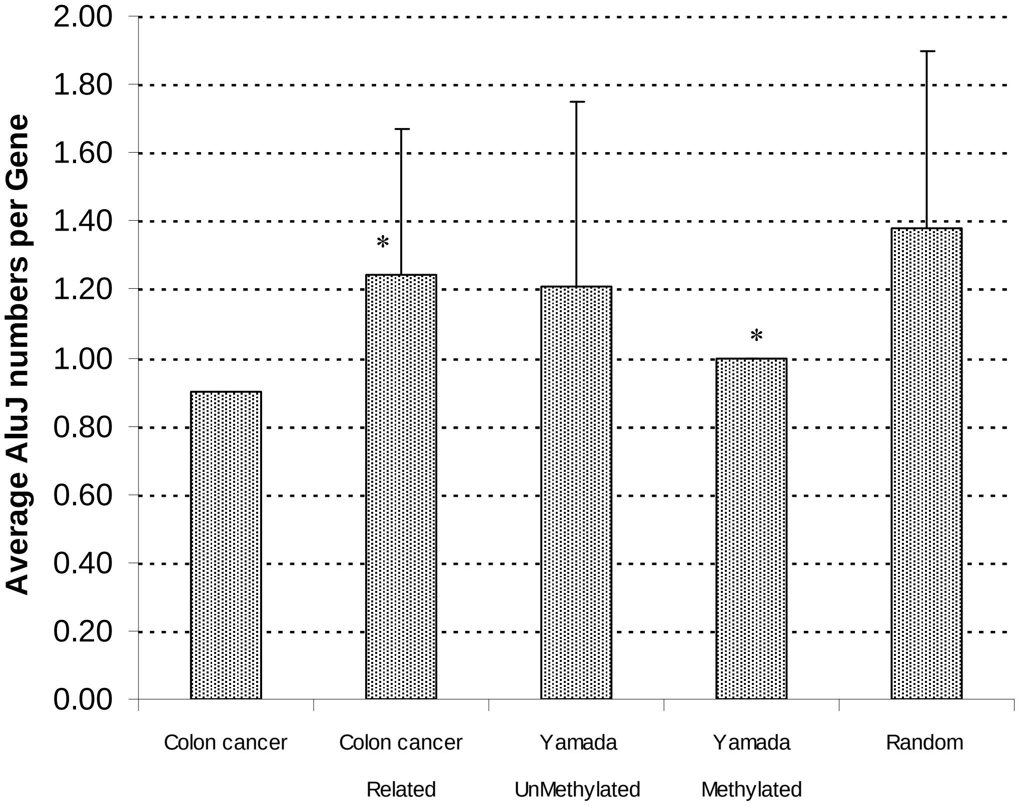


Figure 3.3: Average number of AluJ elements. The average number of AluJ element per gene in each group is calculated by the value: (total number of AluJ elements) / (total number of genes containing AluJ elements). Pairwise t-tests were performed between colon cancer genes and each of the other four groups of genes. * denotes $P < 0.05$.

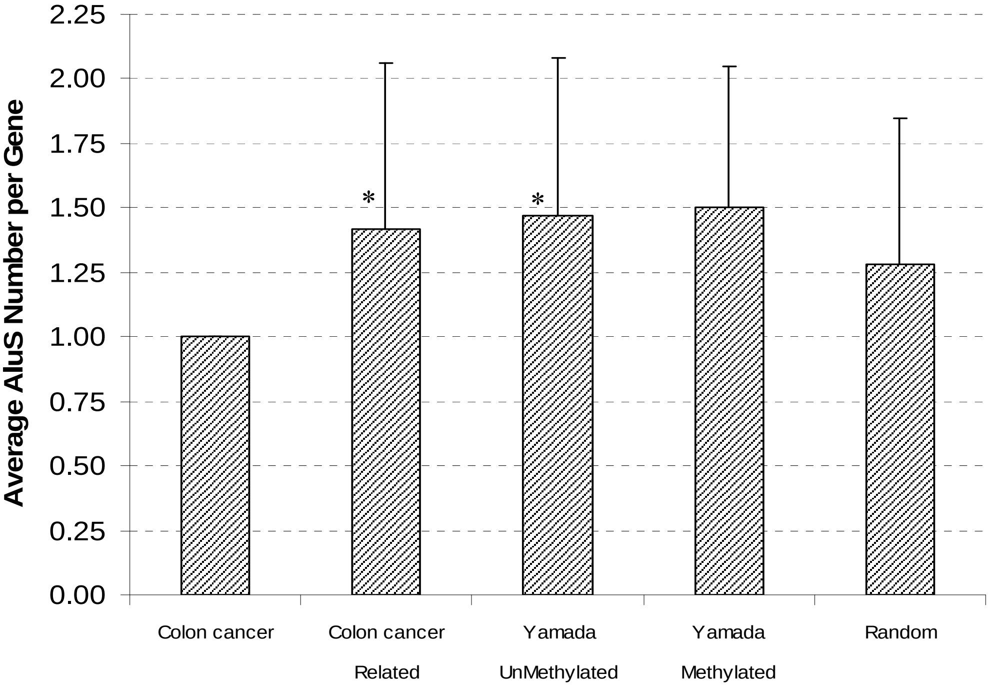


Figure 3.4: Average number of AluS elements. The average number of AluS element per gene in each group is calculated by the value: (total number of AluS elements) / (total number of genes containing AluS elements). Pairwise t-tests were performed between colon cancer genes and each of the other four groups of genes. * denotes $P < 0.05$.

Colon cancer genes harbor a lower density Alu element close to transcription start site

Our results suggest that each group of genes has more than 50% genes that do not contain any Alu element (Table 3.1). Colon cancer genes contain the lowest proportion of genes containing Alu elements and also have the lowest Alu density among all of the five groups of genes. Statistical test results, however, do not support the difference in Alu element density being significant (pairwise t-tests, data not shown); possibly due to the large standard deviation in each group (Figure 3.5).

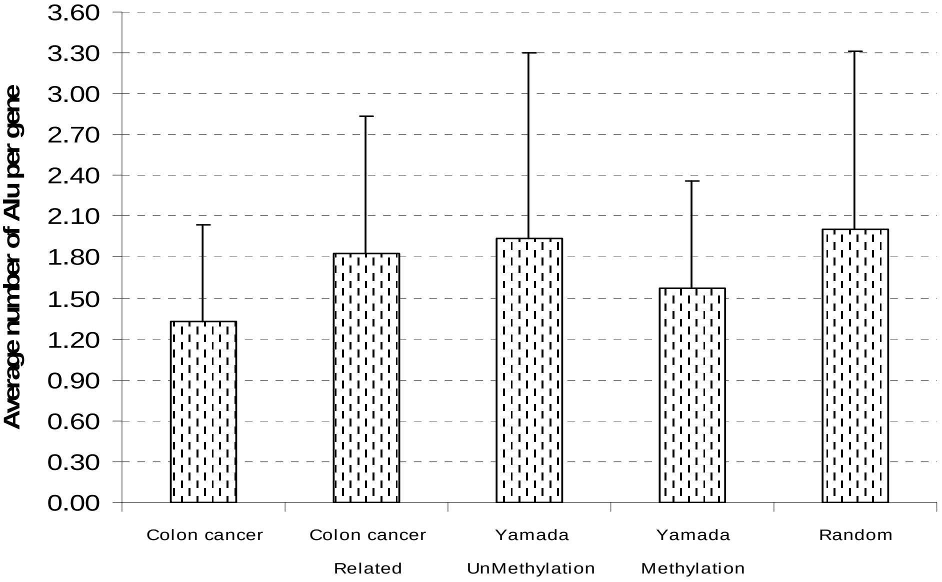
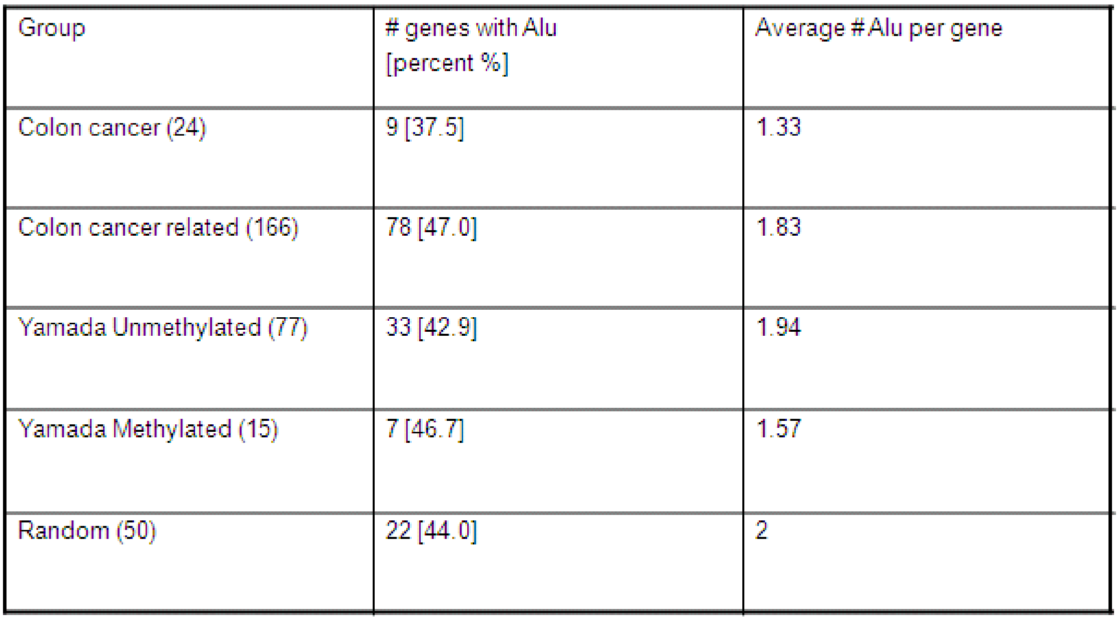


Figure 3.5: Average number of Alu elements. The average number of Alu element per gene in each group is calculated by the value: (total number of Alu elements) / (total number of genes containing all Alu elements). Pairwise t-tests were performed between colon cancer genes and each of the other four groups of genes.

Table 3.1: Percent of genes contains Alu element in each group.



Group	# genes with Alu [percent %]	Average # Alu per gene
Colon cancer (24)	9 [37.5]	1.33
Colon cancer related (166)	78 [47.0]	1.83
Yamada Unmethylated (77)	33 [42.9]	1.94
Yamada Methylated (15)	7 [46.7]	1.57
Random (50)	22 [44.0]	2

Our results indicate that the distance from the nearest Alu element to the transcription start site in colon cancer genes in average is shorter than that in other groups of genes. Again, possibly due to the large standard deviation in each group, the difference in distance was not significant (pairwise t-tests, data not shown), (Figure 3.6).

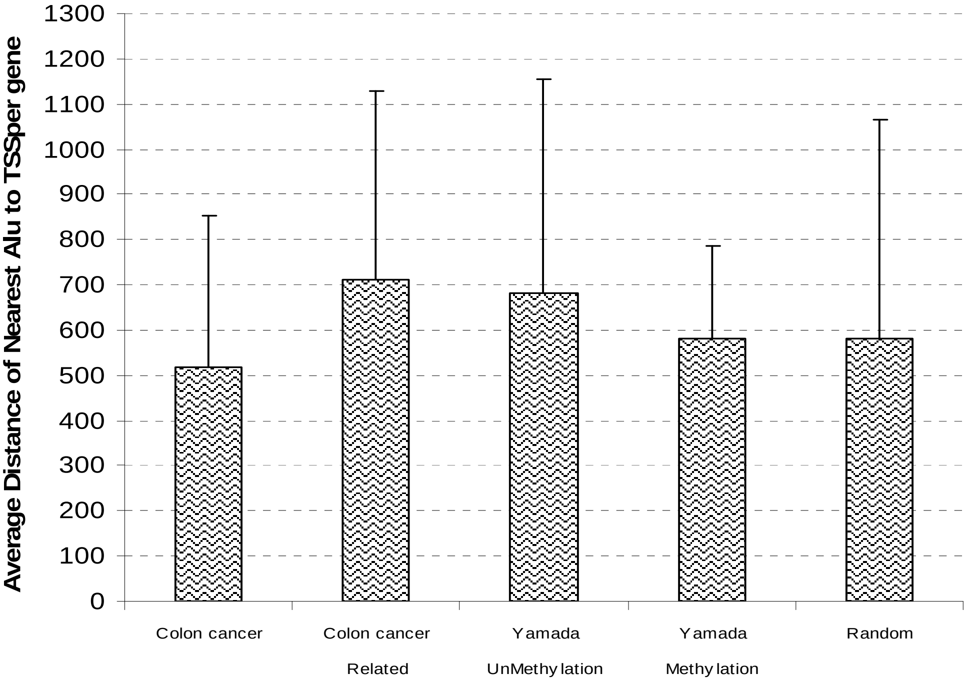


Figure 3.6: Average distance between the nearest Alu element to transcription start site per gene. Average distance between the nearest Alu element and transcription start site in colon cancer is the lowest. This diagram is based only on the genes containing Alu elements. T-test is done between colon cancer genes and each of the other four groups of genes. T-test shows there is no significant difference between colon cancer genes and other groups of genes.

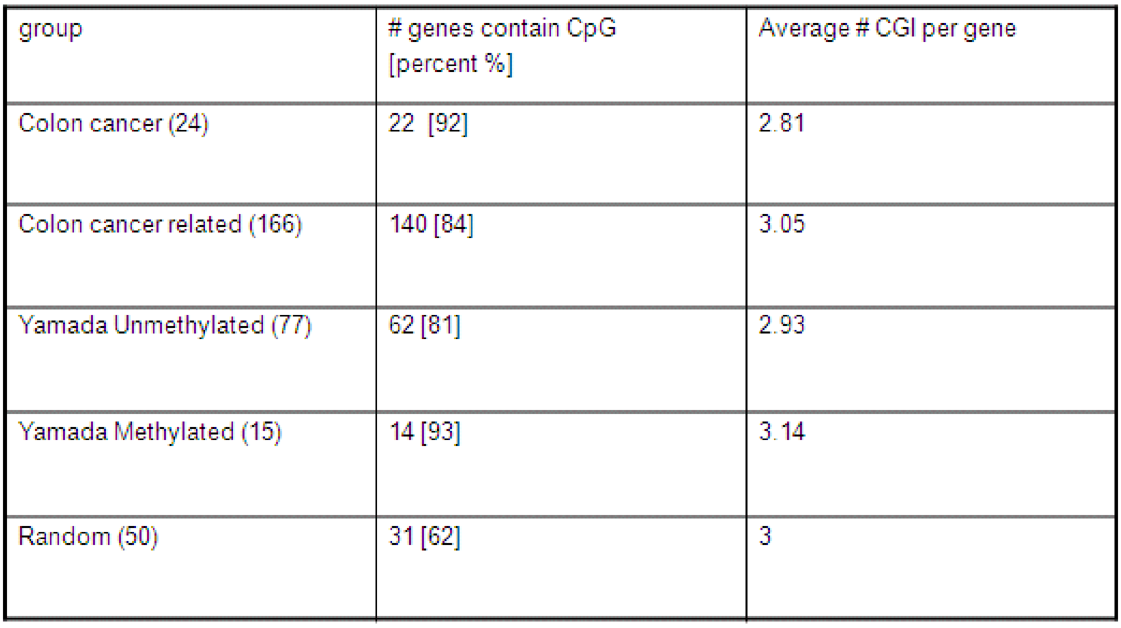
CHAPTER 4

CpG ISLAND METHYLATION STUDY

Sp1 protection model

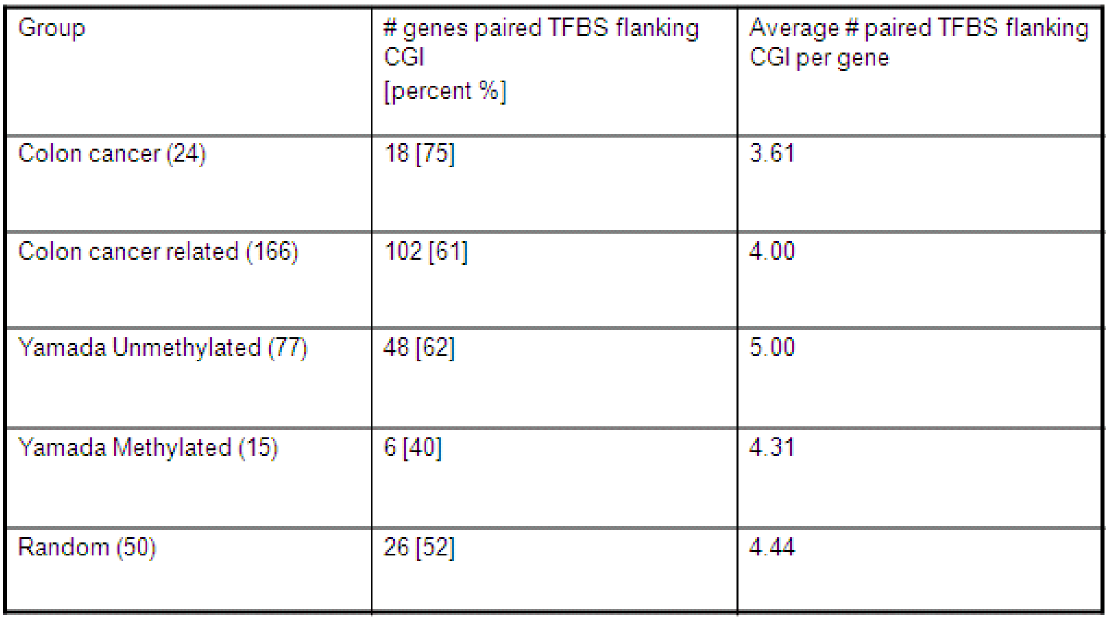
22 out of 24 colon genes contain predicted CpG islands (Table 4.1). 18 out of these 22 colon cancer genes have paired TF flanking CpG island within a 50bp range (Table4.2), with average transcription factor binding sites number per gene of 3.6.

Table 4.1 Percent of genes contains CpG islands in each group.



group	# genes contain CpG [percent %]	Average # CGI per gene
Colon cancer (24)	22 [92]	2.81
Colon cancer related (166)	140 [84]	3.05
Yamada Unmethylated (77)	62 [81]	2.93
Yamada Methylated (15)	14 [93]	3.14
Random (50)	31 [62]	3

Table 4.2: Percent of genes contains paired TFBS flanking CpG island within 50bp in each group.



Group	# genes paired TFBS flanking CGI [percent %]	Average # paired TFBS flanking CGI per gene
Colon cancer (24)	18 [75]	3.61
Colon cancer related (166)	102 [61]	4.00
Yamada Unmethylated (77)	48 [62]	5.00
Yamada Methylated (15)	6 [40]	4.31
Random (50)	26 [52]	4.44

The proportion of genes containing paired TFBS flanking CpG island within 50 base pairs in colon cancer genes is the highest, 0.75, and that in Yamada methylated genes is the lowest. However, for the density of paired TFBS flanking CpG island in the genes which contains them in each group, colon cancer genes have the lowest number, while Yamada unmethylated genes have the highest number. However, these differences were not sufficient to be statistically significant (pairwise t-test, results not shown) (Figure 4.1).

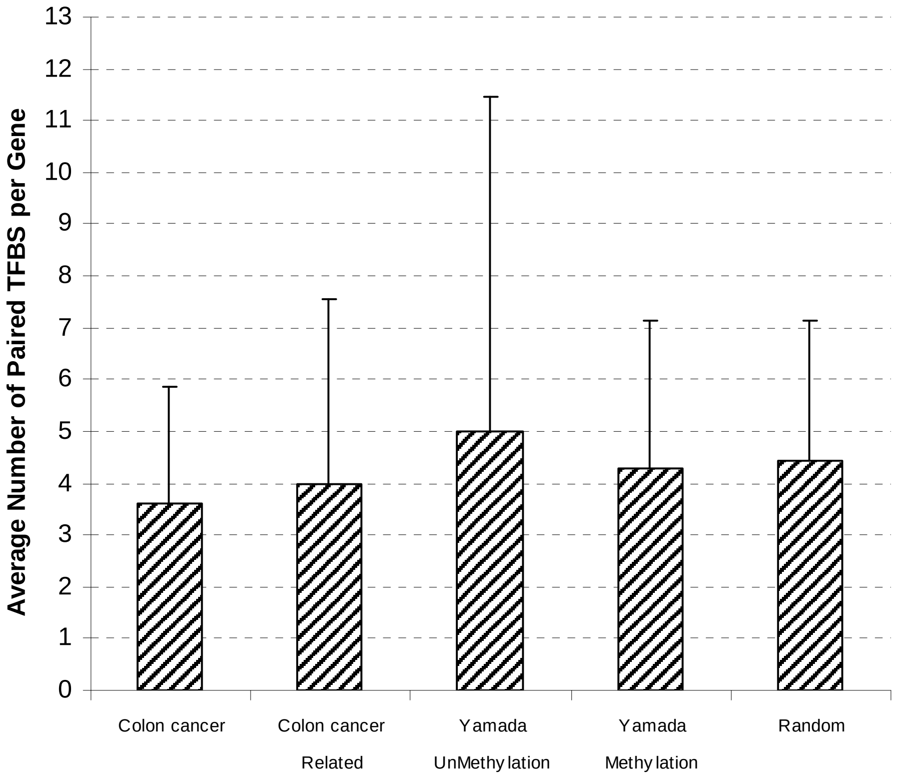


Figure 4.1: Comparison of average number of paired TFBS flanking CpG island within 50 bp per gene among different groups. The average number of paired TFBS flanking the CpG island within 50bp per gene in each group is calculated by the value: (total number of paired TFBS flanking CpG island within 50bp for all the genes) / (total number of genes containing paired TFBS flanking CpG island within 50bp). Pairwise t-tests were performed between colon cancer genes and each of the other four groups of genes.

We hypothesize that TFBS sequence conservation could play a role in protection of CpG islands against the methylation originated from Alu elements: when the degree of matched sequence of TFBS is low with respect to the consensus, the binding of transcription factor (TF) will be weak and thus could not prevent the spreading of methylation into the nearby CpG islands. To test this hypothesis, we calculated the sequence conservation of the flanking TFBS. Our results suggest that the average sequence conservation of the TFBS flanking the CpG islands in colon cancer genes and Yamada methylated genes are lower than other groups of genes (Figure 4.2).

There is a significant difference in the sequence conservation between colon cancer genes and Yamada unmethylated genes, also between colon cancer genes and random genes (Figure 4.2).

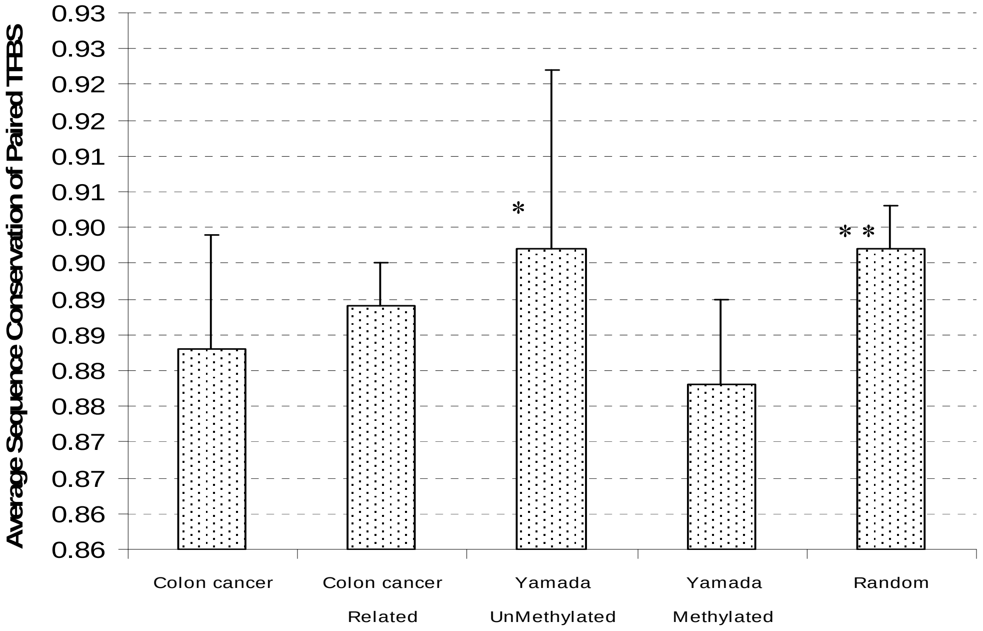


Figure 4.2: Comparison of sequence conservation of paired TFBS flanking CpG islands. Pairwise t-tests were performed between colon cancer genes and each of the other four groups of genes. * denotes $P < 0.05$, ** denotes $P < 0.01$.

CpG island density is related to colon cancer

Our results indicate that the proportion of genes containing CpG islands in colon cancer gene group is higher than other groups of genes except for Yamada methylated genes. However, the average density of CpG islands per gene in colon cancer groups is lower than in other groups. (Figure4.3). However, no significant difference can be demonstrated by statistical means ((pairwise t-test, results not shown).

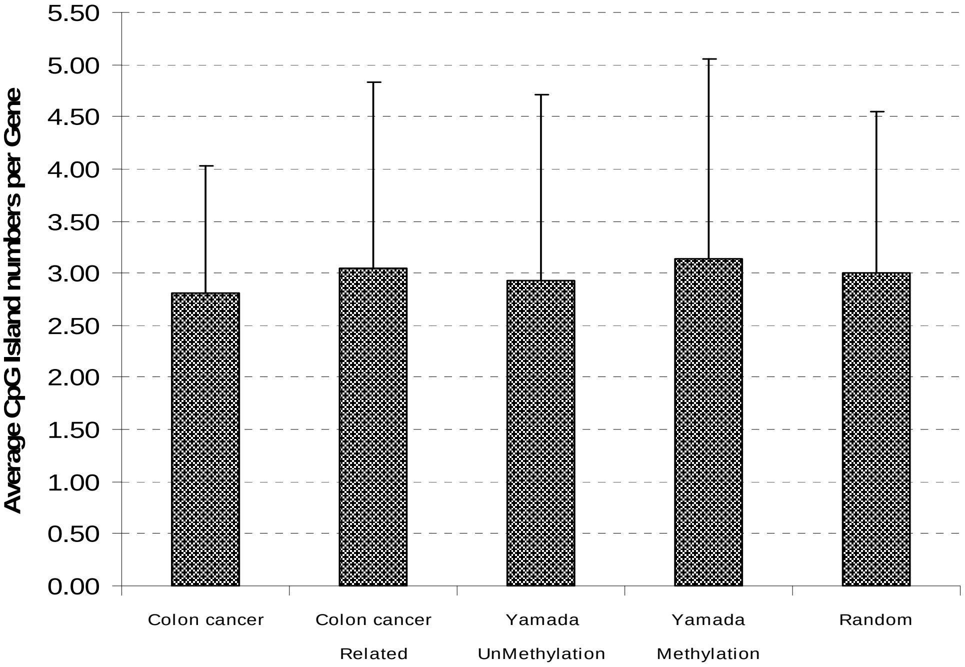


Figure 4.3: Number of CpG islands among different groups of genes. This diagram is based only on the genes containing CpG islands. The average number of CpG island per gene is calculated by the value: (total number of CpG islands for all the genes) / (total number genes containing CpG island).

CpG dinucleotide and TpG dinucleotide

CpG dinucleotides in CpG island in promoter region are the main targets for enzyme transferases involved in DNA methylation [31]. We thus inspected the average percent of CpG dinucleotide in CpG islands for each group. Our findings suggest that the CpG dinucleotide in CpG islands in colon cancer is indeed slightly higher than in other groups of genes (*$P<0.05$* in pairwise t-test between colon cancer genes and random genes) (Figure 4.4).

Throughout evolution, methylated CpG could become TpG because of deamination [33]. Our results suggest that the average percent of TpG dinucleotide in CpG islands in colon cancer genes and colon cancer related genes and random genes is lower than other groups of genes, although these differences are not statistically significant (data not shown) (Figure 4.5). We reasoned that the dinucleotide environment in the entire promoter region may play a role in determining the methylation in cancer genes. When we inspected promoter region in colon cancer genes in comparison with other groups of genes, the results showed that the average percent of CpG dinucleotide in promoter region in colon cancer genes is the highest among all of the five groups of genes (P values <0.05 Figure 4.6). On the other hand, the average TpG dinucleotide percent in promoter region in colon cancer genes is the lowest among all the five groups of genes (p-values < 0.05, Figure 4.7).

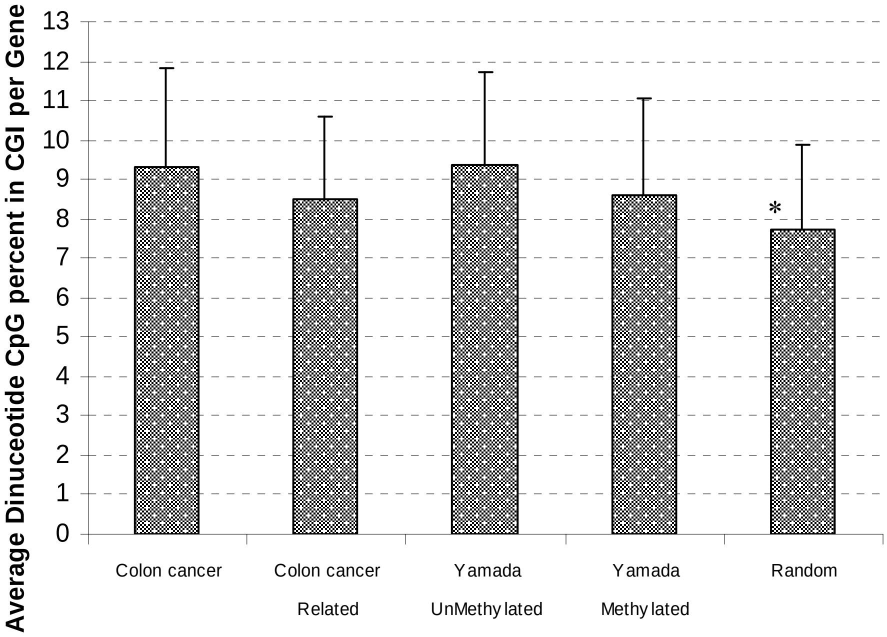


Figure 4.4: Comparison of average CpG dinucleotide percent in the CpG island (CGI) per gene. This diagram is only based on the CpG island region of the genes containing CpG islands. The average dinucleotide CpG per gene is calculated by the value: (total number of dinucleotide CpG) / (number of all possible dinucleotides). * denotes $P < 0.05$ (pairwise *t*-test)

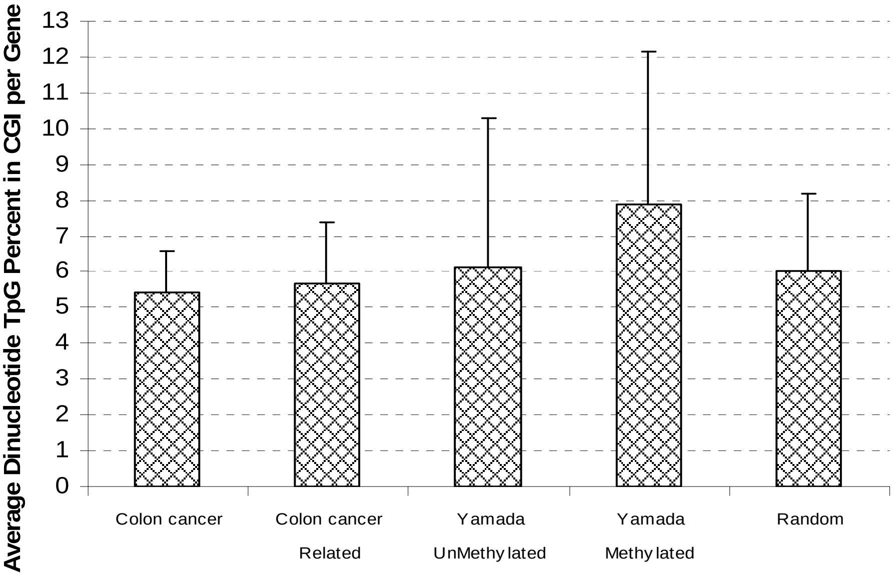


Figure 4.5: TpG dinucleotide in CpG islands among different groups of genes. This diagram is only based on the CpG island region of the genes containing CpG islands. The average dinucleotide TpG per gene was calculated by the value: (total number of dinucleotide TpG) / (number of all possible dinucleotides).

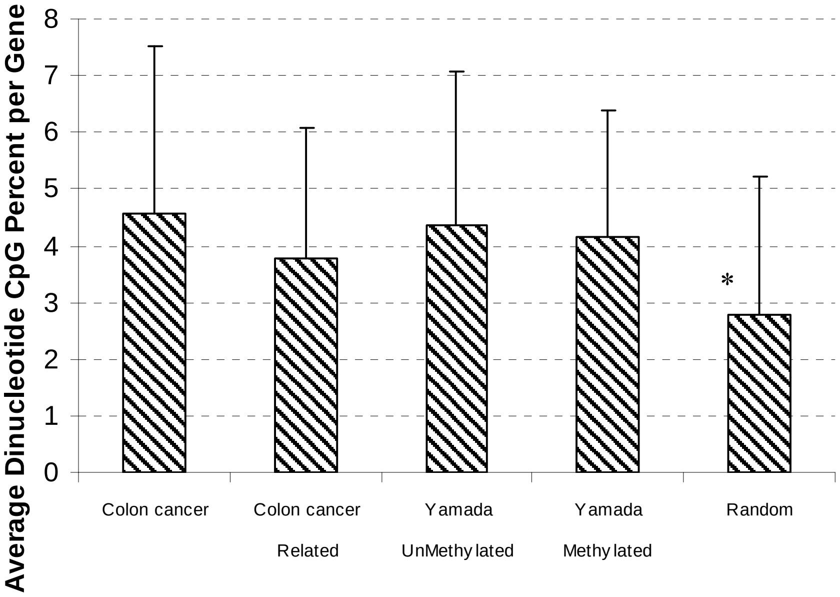


Figure 4.6: Comparison of the average dinucleotide CpG percent for the entire promoter region. This diagram is based on the whole promoter region. The average dinucleotide CpG percent per gene is calculated by the value: (number of dinucleotide CpG) / (number of all possible dinucleotide). * denotes $P < 0.05$ (pairwise t-test).

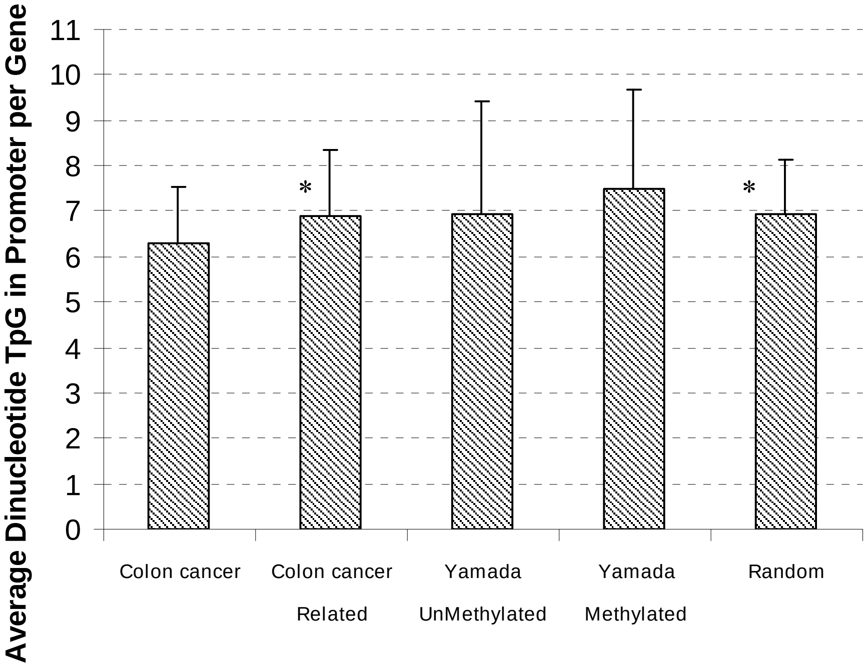


Figure 4.7: Comparison of the dinucleotide TpG percent in entire promoter region per gene. This diagram is based on the whole promoter region. The average dinucleotide TpG percent per gene is calculated by the value: (number of dinucleotide TpG) / (number of all possible dinucleotides).

G+C content is different between colon cancer genes and normal genes

To determine whether G+C content can be used to distinguish colon cancer genes from other groups of genes, we compute the percentage of G+C in the CpG islands for each gene group (Figure 4.8). Our results suggest that the average GC content in random genes is the lowest and there is a significant difference only between colon cancer genes and random genes (*P*< 0.05, pairwise t-test).

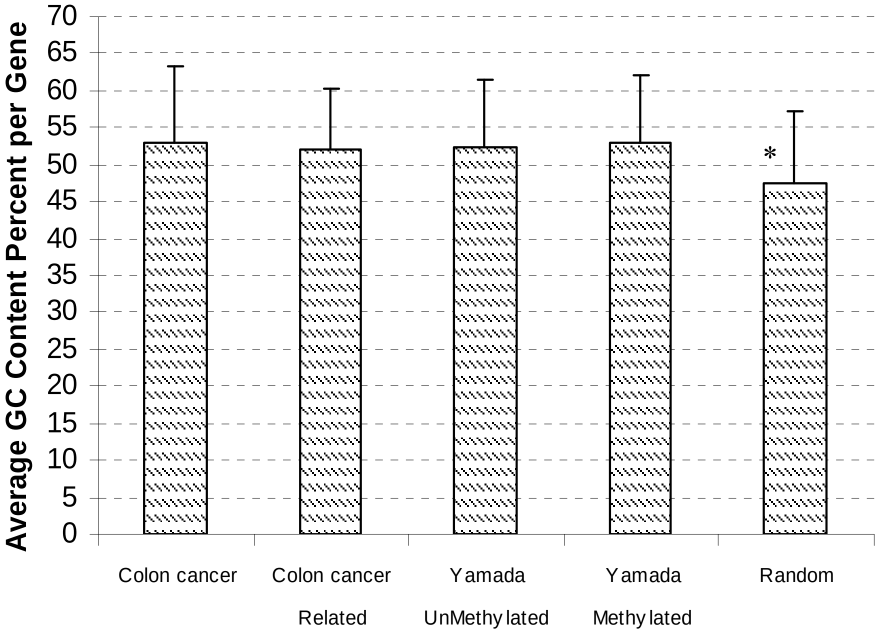
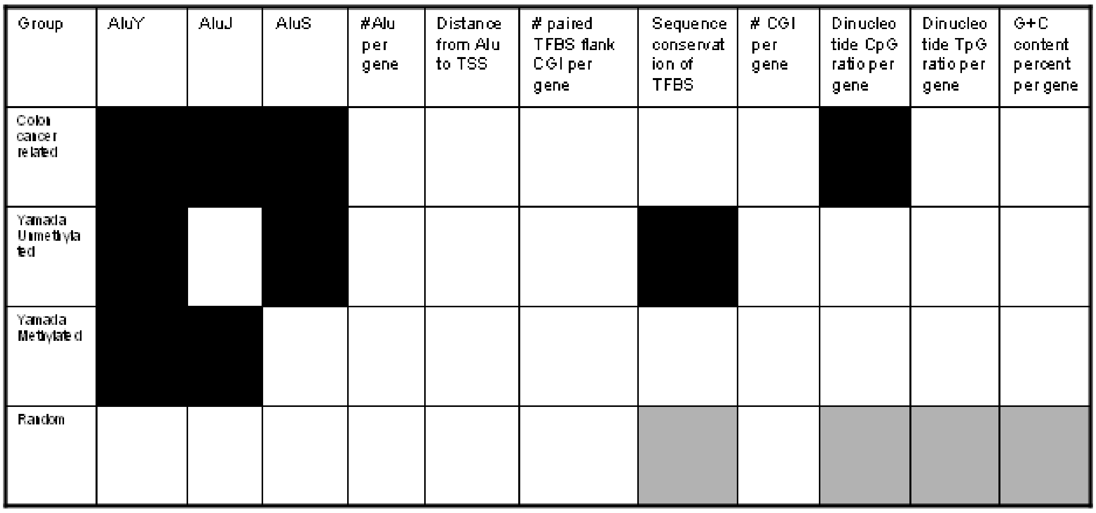


Figure 4.8: Comparison of average G+C content percent in promoter. The average GC content percent per gene is calculated by the value: (number of nucleotide G + number of nucleotide C) / (total number of nucleotides in promoter.). * denotes *P* < 0.05 (pairwise *t*-test).

CHAPTER 5

FEATURE SELECTION, CLASSIFICATION, PREDICTION AND GENE ONTOLOGY

In total, we have studied the following 11 features for comparison between colon cancer genes and four other groups of genes: (1) average number of Alu elements per gene; (2) average number of AluY subfamily elements per gene; (3) average number of AluJ subfamily elements per gene; (4) average number of AluS subfamily element per gene; (5) average distance from the nearest Alu element to TSS per gene; (6) average number of CpG island per gene; (7) average number of paired TFBS flanking CpG island within 50bp per gene; (8) average sequence conservation for paired TFBS flanking CpG island within 50bp per gene; (9) average percent of TpG dinucleotide in promoter region per gene; (10) average percent of CpG dinucleotide in promoter region per gene; (11) average percent of GC content in promoter region per gene (Figure 5.1).



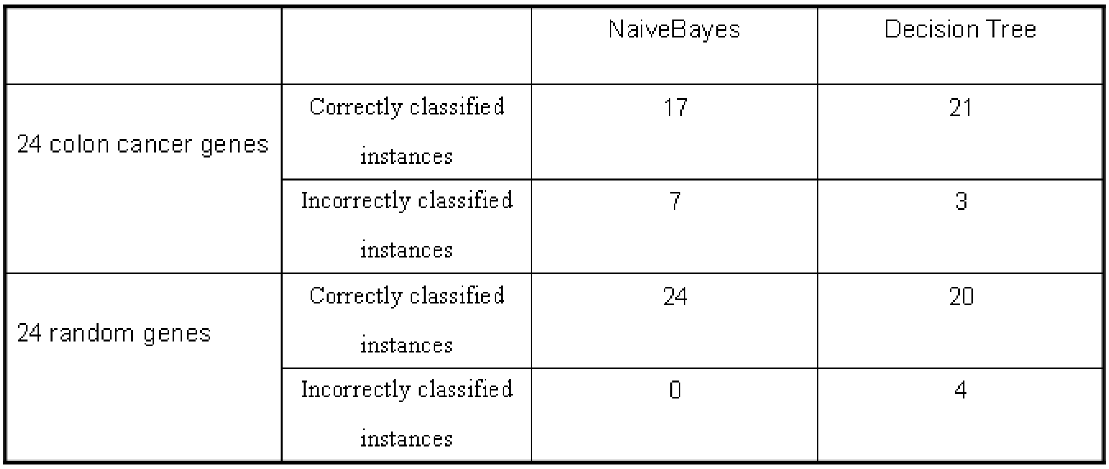
Group	AluY	AluJ	AluS	#Alu per gene	Distance from Alu to TSS	# paired TFBS flank CGI per gene	Sequence conservation of TFBS	# CGI per gene	Dinucleotide CpG ratio per gene	Dinucleotide TpG ratio per gene	G+C content percent per gene
Colon cancer related	■	■	■						■		
Yamada Unmethylated	■		■				■				
Yamada Methylated	■	■									
Random							▨		▨	▨	▨

Figure 5.1: Significant difference distribution of 11 features. The black or gray box depicts the difference between colon cancer genes and the genes in the group. Four gray boxes denote four features in which significant difference between colon cancer genes and the random genes exist.

Feature selection and classification

Based on the statistical result, four features are used to train the classifiers for separation of colon cancer genes from random genes: (1). average sequence conservation for paired TFBS Flanking CpG island within 50bp per gene, (2). average percent of TpG dinucleotide in promoter region per gene, (3). average percent of CpG dinucleotide in promoter region per gene, (4). average percent of GC content in promoter region per gene. The training set used 24 colon cancer genes and 24 random genes. 10 folds cross-validation is used to improve the robustness of the classifiers. For 24 colon cancer genes, Naïve Bayes classifier correctly classified 17 instances, incorrectly classified 7 instances; in comparison, the decision tree method J48 correctly classified 21 instances, incorrectly classified 3 instances. For 24 random cancer genes, Naïve bayes classifier correctly classified 24 of 24 instances, and incorrectly classified 0 instances; Decision Tree classifier correctly classified 20 instances, incorrectly classified 4 instances (Table 5.1). To sum up, Naïve Bayes performs well in the random gene set while the decision tree method predicts similar number of correct cases for both colon cancer genes and random genes.

Table 5.1: Comparison between two classifiers on 24 colon cancer genes and 24 random genes.



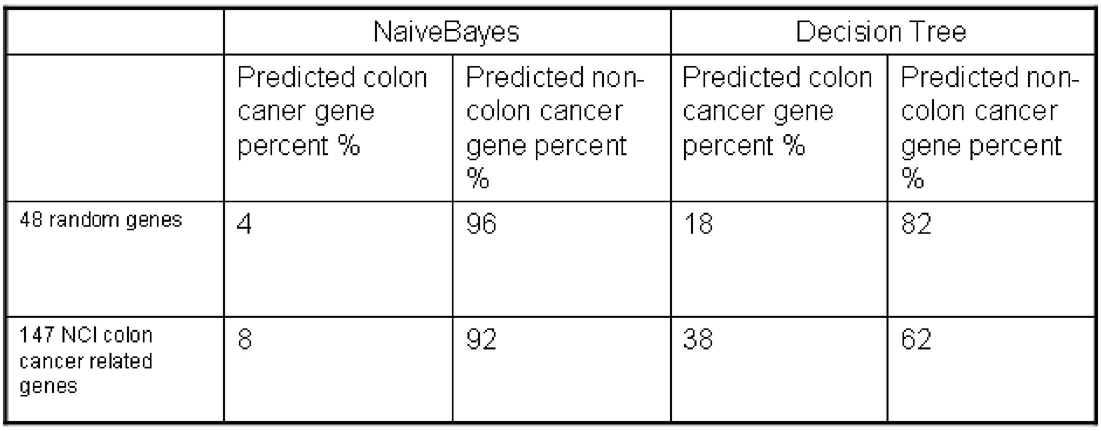
		NaiveBayes	Decision Tree
24 colon cancer genes	Correctly classified instances	17	21
	Incorrectly classified instances	7	3
24 random genes	Correctly classified instances	24	20
	Incorrectly classified instances	0	4

Prediction

It is of great interest to apply classifiers to unknown genes and provide basis for future experimental verification. To this end, we selected an independent set of 48 random genes from the study by Weber and co-workers [42] and 147 NCI colon cancer related genes [69]. NCI cancer gene database is composed of 1500 experimental identified cancer genes. The 147 NCI colon cancer related genes were obtained through National Center Institute cancer database by searching the key word 'colon cancer'.

For 48 random genes, the Naïve Bayes method predicted 2 colon cancer genes (4%) and 46 non-colon cancer genes (96%). The decision tree method predicted 9 colon cancer genes (18%) and 39 non-colon cancer genes (39). For 147 NCI colon cancer related genes, the Naïve Bayes method predicted 4 colon cancer genes (8%) and 44 non-colon cancer genes (92%). The decision tree method predicted 18 colon cancer genes (38%) and 30 non-colon cancer genes (62%) (Table 5.2).

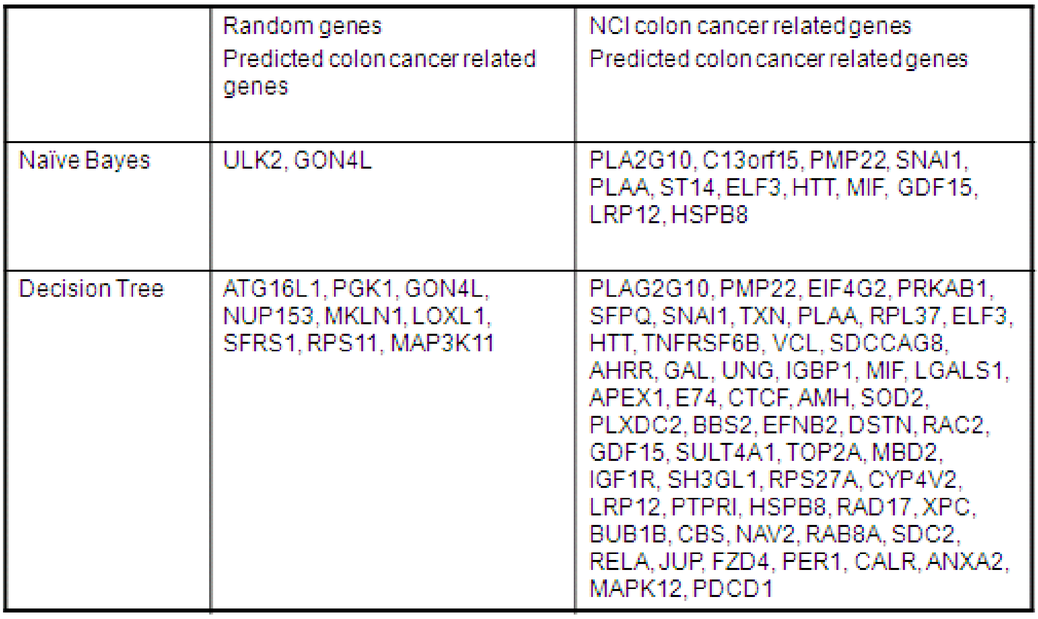
Table 5.2: Comparison between two prediction methods on 48 random genes and 147 NCI colon cancer related genes.



	NaiveBayes		Decision Tree	
	Predicted colon caner gene percent %	Predicted non-colon cancer gene percent %	Predicted colon cancer gene percent %	Predicted non-colon cancer gene percent %
48 random genes	4	96	18	82
147 NCI colon cancer related genes	8	92	38	62

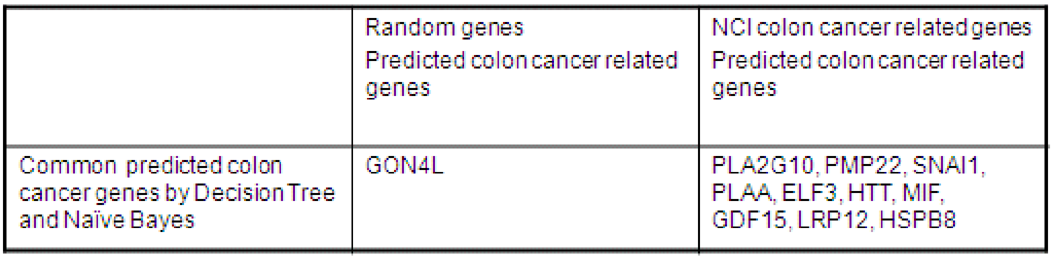
The predicted colon cancer genes from each of the two groups of genes are listed in Table 5.3. We hypothesized that the genes which were predicted by both classifiers have a high probability of being true colon cancer genes based on the four features. The common predicted colon cancer genes by both Naïve Bayes and Decision tree are listed in Table 5.4. and Table 5.5

Table 5.3: Predicted colon cancer genes by two methods (Naïve Bayes and Decision tree) on two groups of genes (48 random genes and 147 NCI colon cancer related genes).



	Random genes Predicted colon cancer related genes	NCI colon cancer related genes Predicted colon cancer related genes
Naïve Bayes	ULK2, GON4L	PLA2G10, C13orf15, PMP22, SNAI1, PLAA, ST14, ELF3, HTT, MIF, GDF15, LRP12, HSPB8
Decision Tree	ATG16L1, PGK1, GON4L, NUP153, MKLN1, LOXL1, SFRS1, RPS11, MAP3K11	PLAG2G10, PMP22, EIF4G2, PRKAB1, SFPQ, SNAI1, TXN, PLAA, RPL37, ELF3, HTT, TNFRSF6B, VCL, SDCCAG8, AHRR, GAL, UNG, IGBP1, MIF, LGALS1, APEX1, E74, CTCF, AMH, SOD2, PLXDC2, BBS2, EFNB2, DSTN, RAC2, GDF15, SULT4A1, TOP2A, MBD2, IGF1R, SH3GL1, RPS27A, CYP4V2, LRP12, PTPRI, HSPB8, RAD17, XPC, BUB1B, CBS, NAV2, RAB8A, SDC2, RELA, JUP, FZD4, PER1, CALR, ANXA2, MAPK12, PDCD1

Table 5.4: Common predicted colon cancer genes by two methods (Naïve bayes and Decision tree) on 48 random genes and 147 NCI colon cancer related genes.



	Random genes Predicted colon cancer related genes	NCI colon cancer related genes Predicted colon cancer related genes
Common predicted colon cancer genes by Decision Tree and Naïve Bayes	GON4L	PLA2G10, PMP22, SNAI1, PLAA, ELF3, HTT, MIF, GDF15, LRP12, HSPB8

Table 5.5: Annotation of common colon cancer genes predicted by Naïve Bayes and Decision tree on 48 random genes and 147 NCI colon cancer related genes. * refer to the predicted colon cancer gene from 48 random genes, the other predicted colon cancer genes are all predicted from 147 NCI colon cancer related genes.

Gene symbol	Gene name	Description
GON4L*	gon-4-like (C. elegans)	
PLA2G10	Phospholipase A2, group X	
PMP22	Peripheral myelin protein 22	This gene encodes an integral membrane protein that is a major component of myelin in the peripheral nervous system
SNAI1	snail homolog 1 (Drosophila)	The Drosophila embryonic protein snail is a zinc finger transcriptional repressor which downregulates the expression of ectodermal genes within the mesoderm. The nuclear protein encoded by this gene is structurally similar to the Drosophila snail protein, and is also thought to be critical for mesoderm formation in the developing embryo. At least two variants of a similar processed pseudogene have been found on chromosome 2
PLAA	phospholipase A2-activating protein	
ELF3	E74-like factor 3	
HTT	huntingtin	Huntingtin is a disease gene linked to Huntington's disease, a neurodegenerative disorder characterized by loss of striatal neurons. This is thought to be caused by an expanded, unstable trinucleotide repeat in the huntingtin gene, which translates as a polyglutamine repeat in the protein product.

MIF	macrophage migration inhibitory factor (glycosylation-inhibiting factor)	This gene encodes a lymphokine involved in cell-mediated immunity, immunoregulation, and inflammation. It plays a role in the regulation of macrophage function in host defense through the suppression of anti-inflammatory effects of glucocorticoids. This lymphokine and the JAB1 protein form a complex in the cytosol near the peripheral plasma membrane, which may indicate an additional role in integrin signaling pathways.
GDF15	Growth differentiation factor 15	Bone morphogenetic proteins (e.g., BMP5; MIM 112265) are members of the transforming growth factor-beta (see TGFB1; MIM 190180) superfamily and regulate tissue differentiation and maintenance. They are synthesized as precursor molecules that are processed at a dibasic cleavage site to release C-terminal domains containing a characteristic motif of 7 conserved cysteines in the mature protein.
LRP12	low density lipoprotein-related protein 12	This gene was identified by its differential expression in cancer cells. The product of this gene is predicted to be a transmembrane protein. The level of this protein was found to be lower in tumor derived cell lines compared to normal cells. This gene was thus proposed to be a candidate tumor suppressor gene.
HSPB8	heat shock 22kDa protein 8	

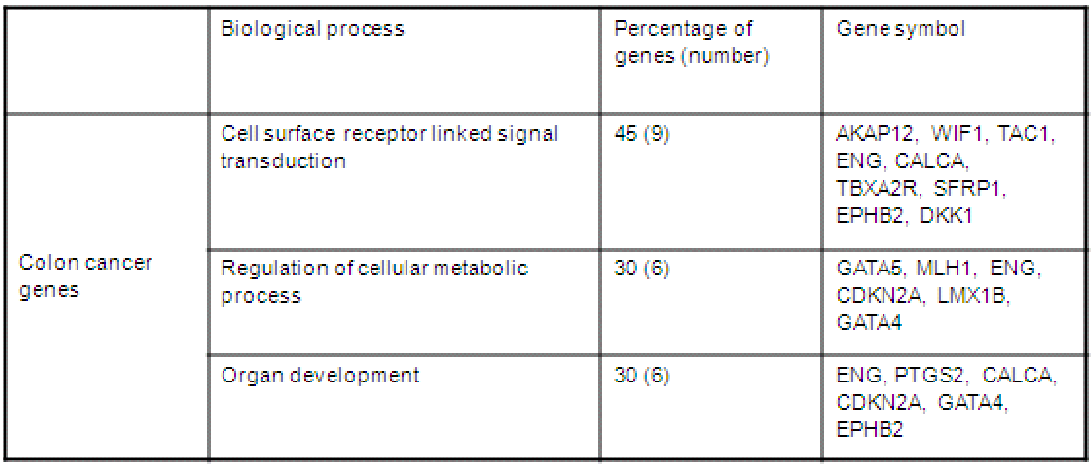
To search for the candidate colon cancer genes, we compared the genes on the 4 features used in prediction, plus an additional 7 features (see chapter 4) that we used in this study: (1) average number of Alu elements per gene; (2) average number of AluY subfamily element per gene; (3) average number of AluJ subfamily element per gene; (4) average number of AluS subfamily element per gene; (5) average distance from the nearest Alu element to TSS per gene; (6) average number of CpG islands per gene; (7) average number of paired TFBS flanking CpG island within 50bp per gene.

We have identified LRP2 as a good candidate for colon cancer gene. First, the number of CpG island in LRP2 gene is as high as 4, consistent with the hypothesis: being more substrate for methylation enzyme. Secondly, the sequence conservation of paired TFBS flanking CpG island within 50 base pairs is relatively low in LRP2, compared to the mean sequence conservation of paired TFBS in 24 colon cancer genes, which is also consistent with our hypothesis: lower sequence conservation could decrease the protection ability of TFBS to CpG island. Third, from RefSeq, LRP2 is referred as a tumor suppressor gene because the level of LRP2 coded protein is found to be lower in tumor derived cell lines compared to normal cells (Table 5.5).

Gene Ontology

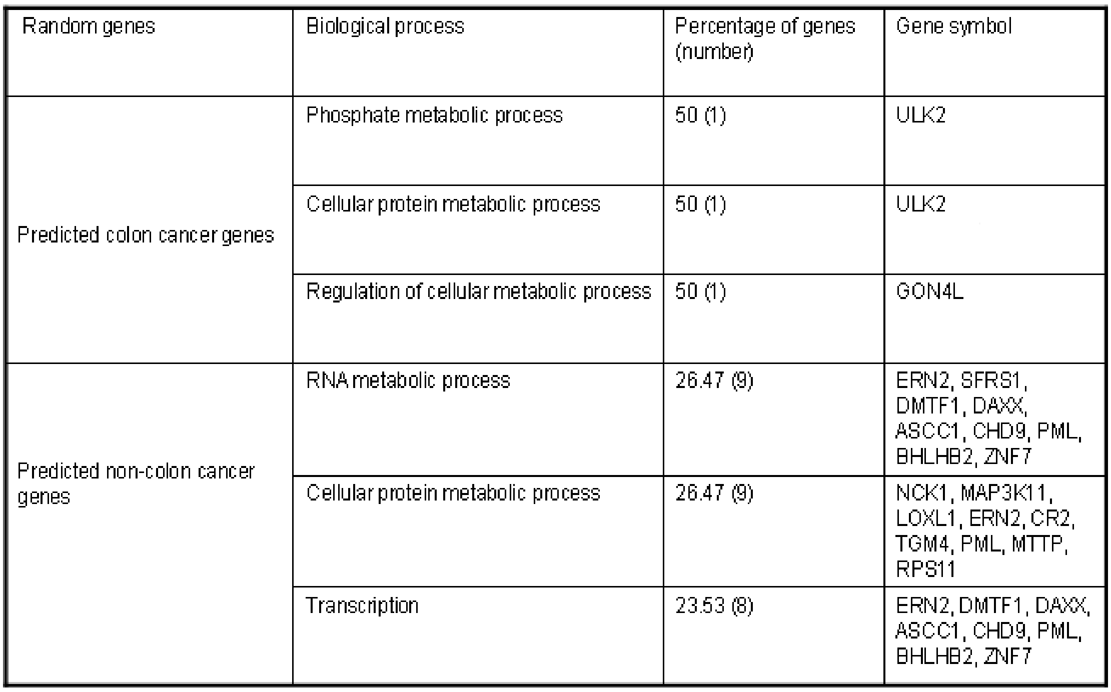
Our results suggest that for 24 colon cancer genes, the top three biological process in which colon cancer genes are involved are cell surface receptor linked signal transduction, regulation of cellular metabolic process and organ development (Table 5.6). For predicted colon cancer genes from 48 random genes by Naïve Bayes method, the top three biological processes are phosphate metabolic process, cellular protein metabolic process and regulation of cellular metabolic process (Table 5.7); the colon cancer genes predicted by Decision tree method are involved in cellular protein metabolic process, cell development and phosphate metabolic process (Table 5.9). For predicted non-colon cancer genes from 48 random genes by Naïve Bayes method, the top three biological processes are RNA metabolic process, cellular protein metabolic process and transcription (Table 5.7); the non-colon cancer genes predicted by Decision tree are involved in G1 Biopolymer metabolic process, protein metabolic process and signal transduction (Table 5.9). For predicted colon cancer genes from 147 NCI colon cancer genes by Naïve Bayes method, the top three biological processes are cellular lipid metabolic process, inflammatory response and cell surface receptor linked signal transduction (Table 5.8); for predicted non-colon cancer genes from 147 NCI colon cancer genes by Naïve Bayes method, the top three biological processes are cellular protein metabolic process, regulation of cellular metabolic process and RNA metabolic process (Table 5.8). For predicted colon cancer genes from 147 NCI colon cancer genes by Decision Tree method, the top three biological processes are cellular protein metabolic process, cell development, regulation of cellular metabolic process (Table 5.10); For predicted non-colon cancer genes from 147 NCI colon cancer genes by Decision Tree method, the top three biological processes are cellular protein metabolic protein, cell development and regulation of cellular metabolic process (Table 5.10).

Table 5.6: Annotation of 24 colon cancer genes on biological process.



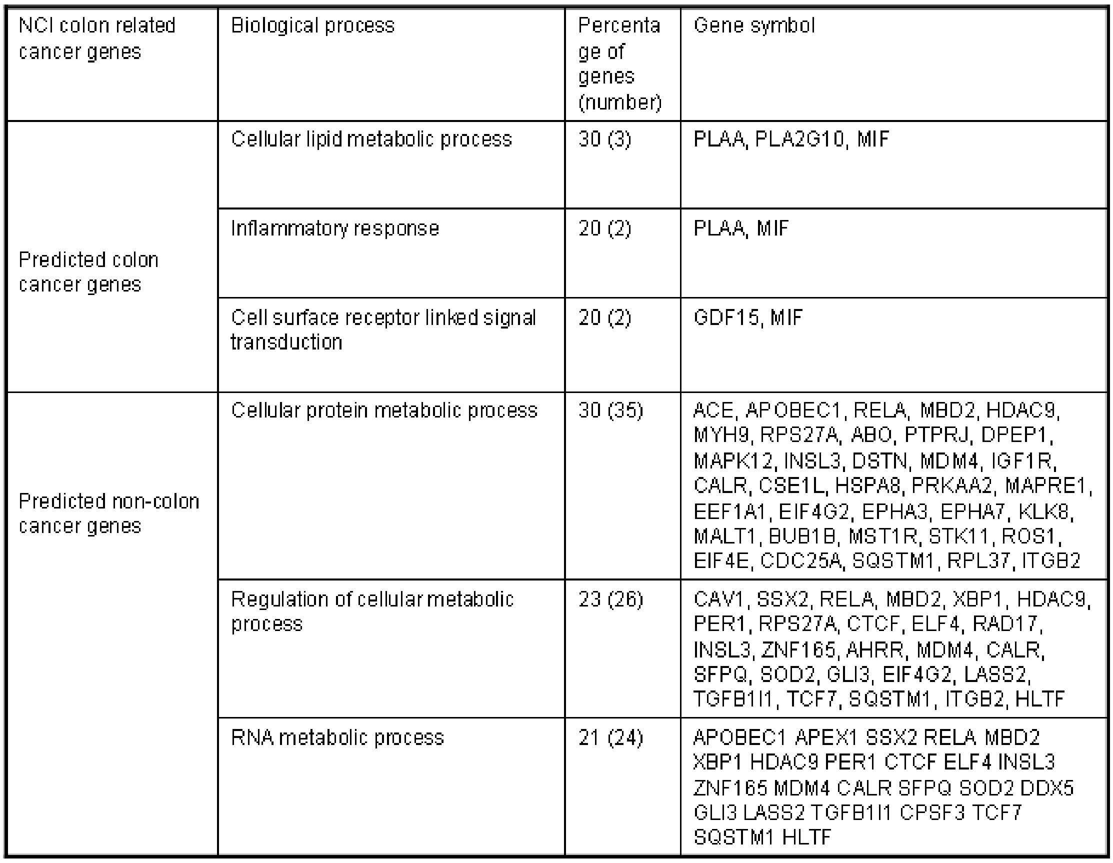
	Biological process	Percentage of genes (number)	Gene symbol
Colon cancer genes	Cell surface receptor linked signal transduction	45 (9)	AKAP12, WIF1, TAC1, ENG, CALCA, TBXA2R, SFRP1, EPHB2, DKK1
	Regulation of cellular metabolic process	30 (6)	GATA5, MLH1, ENG, CDKN2A, LMX1B, GATA4
	Organ development	30 (6)	ENG, PTGS2, CALCA, CDKN2A, GATA4, EPHB2

Table 5.7: Annotation of predicted colon cancer genes and non-colon cancer genes by Naïve Bayes method from 48 random genes.



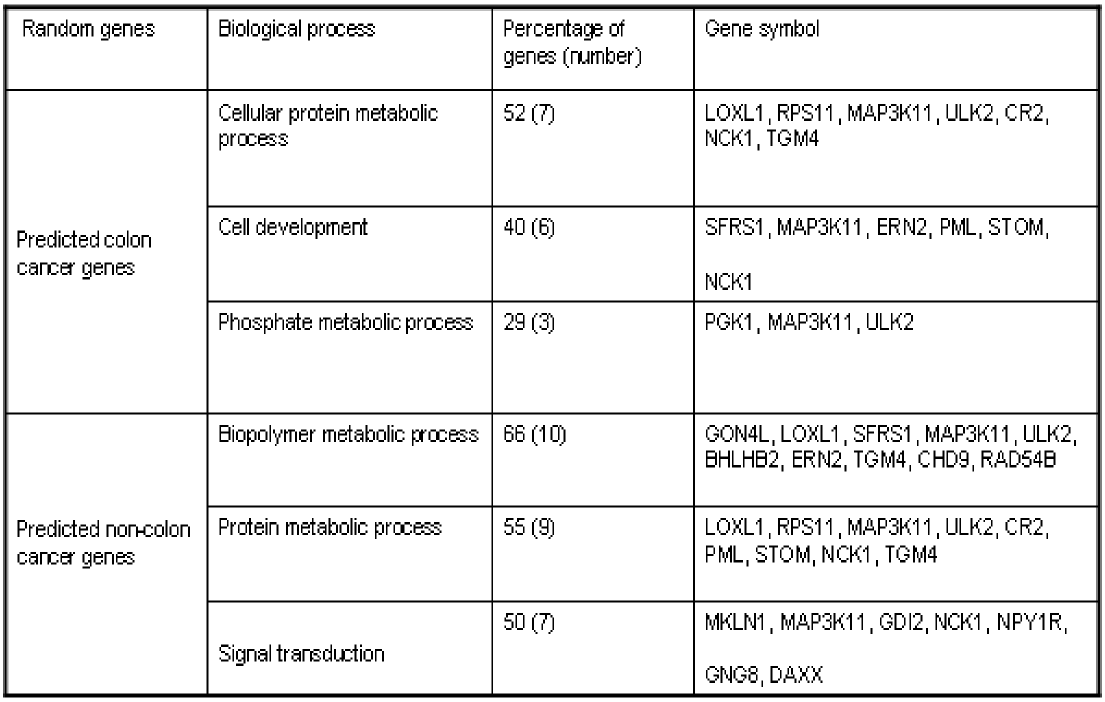
Random genes	Biological process	Percentage of genes (number)	Gene symbol
Predicted colon cancer genes	Phosphate metabolic process	50 (1)	ULK2
	Cellular protein metabolic process	50 (1)	ULK2
	Regulation of cellular metabolic process	50 (1)	GON4L
Predicted non-colon cancer genes	RNA metabolic process	26.47 (9)	ERN2, SFRS1, DMTF1, DAXX, ASCC1, CHD9, PML, BHLHB2, ZNF7
	Cellular protein metabolic process	26.47 (9)	NCK1, MAP3K11, LOXL1, ERN2, CR2, TGM4, PML, MTTP, RPS11
	Transcription	23.53 (8)	ERN2, DMTF1, DAXX, ASCC1, CHD9, PML, BHLHB2, ZNF7

Table 5.8: Annotation of predicted colon cancer genes and non-colon cancer genes by Naïve Bayes from 147 NCI colon cancer related genes.



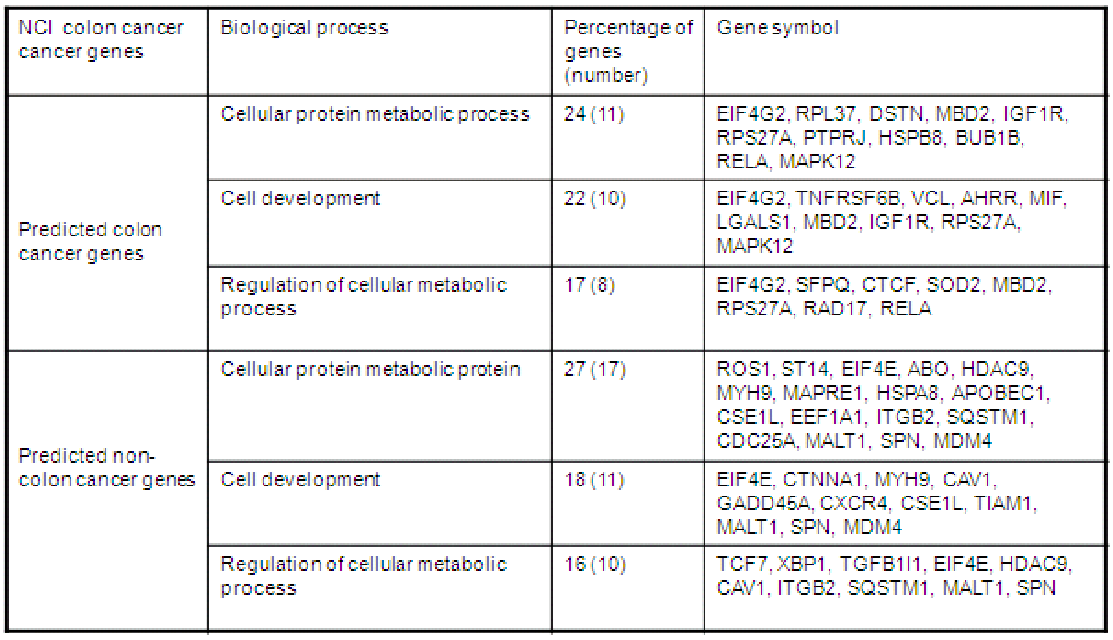
NCI colon related cancer genes	Biological process	Percentage of genes (number)	Gene symbol
Predicted colon cancer genes	Cellular lipid metabolic process	30 (3)	PLAA, PLA2G10, MIF
	Inflammatory response	20 (2)	PLAA, MIF
	Cell surface receptor linked signal transduction	20 (2)	GDF15, MIF
Predicted non-colon cancer genes	Cellular protein metabolic process	30 (35)	ACE, APOBEC1, RELA, MBD2, HDAC9, MYH9, RPS27A, ABO, PTPRJ, DPEP1, MAPK12, INSL3, DSTN, MDM4, IGF1R, CALR, CSE1L, HSPA8, PRKAA2, MAPRE1, EEF1A1, EIF4G2, EPHA3, EPHA7, KLK8, MALT1, BUB1B, MST1R, STK11, ROS1, EIF4E, CDC25A, SQSTM1, RPL37, ITGB2
	Regulation of cellular metabolic process	23 (26)	CAV1, SSX2, RELA, MBD2, XBP1, HDAC9, PER1, RPS27A, CTCF, ELF4, RAD17, INSL3, ZNF165, AHRR, MDM4, CALR, SFPQ, SOD2, GLI3, EIF4G2, LASS2, TGFB1I1, TCF7, SQSTM1, ITGB2, HLTF
	RNA metabolic process	21 (24)	APOBEC1 APEX1 SSX2 RELA MBD2 XBP1 HDAC9 PER1 CTCF ELF4 INSL3 ZNF165 MDM4 CALR SFPQ SOD2 DDX5 GLI3 LASS2 TGFB1I1 CPSF3 TCF7 SQSTM1 HLTF

Table 5.9: Annotation of predicted colon cancer genes and non-colon cancer genes by Decision tree method from 48 random genes.



Random genes	Biological process	Percentage of genes (number)	Gene symbol
Predicted colon cancer genes	Cellular protein metabolic process	52 (7)	LOXL1, RPS11, MAP3K11, ULK2, CR2, NCK1, TGM4
	Cell development	40 (6)	SFRS1, MAP3K11, ERN2, PML, STOM, NCK1
	Phosphate metabolic process	29 (3)	PGK1, MAP3K11, ULK2
Predicted non-colon cancer genes	Biopolymer metabolic process	66 (10)	GON4L, LOXL1, SFRS1, MAP3K11, ULK2, BHLHB2, ERN2, TGM4, CHD9, RAD54B
	Protein metabolic process	55 (9)	LOXL1, RPS11, MAP3K11, ULK2, CR2, PML, STOM, NCK1, TGM4
	Signal transduction	50 (7)	MKLN1, MAP3K11, GDI2, NCK1, NPY1R, GNG8, DAXX

Table 5.10: Annotation of predicted colon cancer genes and non-colon cancer genes by Decision tree from 147 NCI colon cancer related genes.



NCI colon cancer cancer genes	Biological process	Percentage of genes (number)	Gene symbol
Predicted colon cancer genes	Cellular protein metabolic process	24 (11)	EIF4G2, RPL37, DSTN, MBD2, IGF1R, RPS27A, PTPRJ, HSPB8, BUB1B, RELA, MAPK12
	Cell development	22 (10)	EIF4G2, TNFRSF6B, VCL, AHRR, MIF, LGALS1, MBD2, IGF1R, RPS27A, MAPK12
	Regulation of cellular metabolic process	17 (8)	EIF4G2, SFPQ, CTCF, SOD2, MBD2, RPS27A, RAD17, RELA
Predicted non-colon cancer genes	Cellular protein metabolic protein	27 (17)	ROS1, ST14, EIF4E, ABO, HDAC9, MYH9, MAPRE1, HSPA8, APOBEC1, CSE1L, EEF1A1, ITGB2, SQSTM1, CDC25A, MALT1, SPN, MDM4
	Cell development	18 (11)	EIF4E, CTNNA1, MYH9, CAV1, GADD45A, CXCR4, CSE1L, TIAM1, MALT1, SPN, MDM4
	Regulation of cellular metabolic process	16 (10)	TCF7, XBP1, TGFB1I1, EIF4E, HDAC9, CAV1, ITGB2, SQSTM1, MALT1, SPN

CHAPTER 6: DISCUSSION AND CONCLUSIONS

Alu sequence study

Alu subfamily distribution

Previous studies indicate that the younger the Alu subfamily is the higher rate of methylation in the Alu sequences [18]. Our findings suggest that the proportion of AluY subfamily of colon cancer genes is lower than other groups of genes. On the surface this would suggest that overall methylation in the Alu elements of colon cancer genes is lower than other genes. Experimental verification will be critical to determine if this is the case.

Alu elements serve as de novo methylation center

Our results suggest that the average Alu density in colon cancer genes is lower than in other groups of genes. The significance of this is unclear at this time. The average distance between the nearest Alu element to the transcription start site in colon cancer genes is shorter than in other groups of genes. We postulate if Alu sequence is the *de novo* methylation center of a promoter, the closer the Alu element is to the transcription start site, the easier its methylation could be expanded to the nearby CpG islands, affecting the transcription initiation. It is conceivable that during colon tumor genesis, the CpG islands near the transcription start site of the colon tumor suppressor genes are prone to become methylated, leading to inhibition of transcription and the subsequent development of cancers.

CpG island methylation study

Sp1 protection model

One interesting hypothesis to account for the differential pattern of DNA methylation, conjectured by some researchers is that Alu sequences can be regarded as the *de novo* methylation centers because of their high methylation status and the transcription factor binding sites (TFBSs) flanking the CpG islands behave as boundaries to impede the methylation spreading from the methylation centers [31]. Our results show that it is common to identify a pair of TFBSs flanking the CpG islands in all of the five groups of genes (Appendix table 4). This observation suggests that TFBSs cannot be the sole factor for protection from methylation. Our results show that the average sequence conservation of these paired TFBS flanking CpG islands in colon cancer genes is lower than in Yamada unmethylated genes and random genes. It is tempting to speculate that low degree of sequence conservation in TFBS might decrease the binding ability of transcription factors, hence leads to a lower degree of protection from methylation by TFBS.

CpG and TpG

Our results show that the percentage of CpG dinucleotides in the promoter region of colon cancer genes is higher than in random genes with a significant difference (Figure 4.6), but the percentage of TpG dinucleotides is lower than in random genes with significant difference (Figure 4.7). It is currently unknown why this is so. However, it would be consistent with the hypothesis that when the CpG dinucleotides in the promoters of colon cancer genes are at the non-diseased condition they are usually unmethylated.

Feature selection and classification

Our results imply that decision tree method in general perform better or at least as well as NaiveBayes method (Table 5.4). However, our training data was limited to a very small set of genes. It is unclear whether the performance of the two methods would be consistently the same for a larger data in the future.

Future work

Using the common genes predicted by two classifiers, we identified LRP12 as a candidate gene for colon cancers. Experimentation on the methylation status of the promoter and functional characterization of LRP12 in both normal cells and colon cancer cells are pressing.

Work in the future should also include identification of more features which could separate colon cancer genes from normal genes. In addition, different classification methods should be compared.

APPENDICES

Appendix table 1: Annotation for 24 colon cancer genes.

Gene symbol	Gene name	Description
TAC1	tachykinin, precursor 1	These encoded hormones are thought to function as neurotransmitters which interact with nerve receptors and smooth muscle cells.
TBXA2R	thromboxane A2 receptor	
PTGS2	prostaglandin-endoperoxide synthase 2	PTGS, known as cyclooxygenase, is the key enzyme in prostaglandin biosynthesis, and acts both as a dioxygenase and as a peroxidase.
MLH1	mutL homolog 1, colon cancer, nonpolyposis type 2	This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC).
RARB	retinoic acid receptor, beta	It is thought that this protein limits growth of many cell types by regulating gene expression.
FAT	FAT tumor suppressor homolog 1	This gene encodes a tumor suppressor essential for controlling cell proliferation during Drosophila development.
SFRP1	secreted frizzled-related protein 1	It is involved in determining the polarity of photoreceptor cells in the retina.
ENG	Endoglin	Endoglin is a component of the transforming growth factor beta receptor complex as it binds TGFB1 and TGFB3 with high affinity.
DKK1	dickkopf homolog 1	It is a secreted protein with two cysteine rich regions and is involved in embryonic development through its inhibition of the WNT signaling pathway.
CALCA	calcitonin-related polypeptide alpha	Calcitonin causes reduction in serum calcium, an effect opposite to that of parathyroid hormone
RBP1	Retinol binding protein 1, cellular	RBP1 is the carrier protein involved in the transport of retinol (vitamin A alcohol) from the liver storage site to peripheral tissue.
STK11	Serine/threonine kinase 11	This gene, which encodes a member of the serine/threonine kinase family, regulates cell polarity and functions as a tumor suppressor.
GATA5	GATA binding	The protein encoded by this gene is a transcription

	protein 5	factor that contains two GATA-type zinc fingers.
AKAP12	A kinase (PRKA) anchor protein (gravin) 12	The encoded protein is a cell growth-related protein.
EPHB2	EPH receptor B2	Ephrin receptors and their ligands, the ephrins, mediate numerous developmental processes, particularly in the nervous system.
GATA4	GATA binding protein 4	The encoded protein is thought to regulate genes involved in embryogenesis and in myocardial differentiation and function.
LMX1B	LIM homeobox transcription factor 1, beta	
TMEFF2	transmembrane protein with EGF-like and two follistatin-like domains 2	
CDKN2A	cyclin-dependent kinase inhibitor 2A	This ARF product functions as a stabilizer of the tumor suppressor protein p53 as it can interact with, and sequester, MDM1, a protein responsible for the degradation of p53.
WIF1	WNT inhibitory factor 1	WNT proteins are extracellular signaling molecules involved in the control of embryonic development.
SST	Somatostatin	The encoded hormone is an important regulator of the endocrine system through its interactions with pituitary growth hormone, thyroid stimulating hormone, and most hormones of the gastrointestinal tract.
WRN	Werner syndrome	This gene encodes a member of the RecQ subfamily and the DEAH (Asp-Glu-Ala-His) subfamily of DNA and RNA helicases. DNA helicases are involved in many aspects of DNA metabolism, including transcription, replication, recombination, and repair.
NELL	NEL-like 1	This gene encodes a cytoplasmic protein that contains epidermal growth factor (EGF) -like repeats.
MAL	mal, T-cell differentiation protein	The encoded protein has been localized to the endoplasmic reticulum of T-cells and is a candidate linker protein in T-cell signal transduction.

Appendix table 2: Annotation of 147 NCI colon cancer related genes

(available at http://www.chenlab.clemson.edu/staticpages/index.php?page=Datadownload)

Appendix table 3: Annotation of 48 random genes

(available at http://www.chenlab.clemson.edu/staticpages/index.php?page=Datadownload)

Appendix table 4: Comparison of TFBS in protection model in five groups of genes

(available at http://www.chenlab.clemson.edu/staticpages/index.php?page=Datadownload)

REFERENCES

1, Rodriguez, J. (2007). Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. Nucleic Acids Res, 2007, 1–15.

2, Ehrlich, M. (2002). DNA methylation in cancer: too much, but also too little. Oncogene, 21, 5400–5413.

3, Batzer, A. M. (2002). Alu repeats and human genomic diversity. Nat Rev Genet, 3, 370-380.

4, Houck, C. M. (1979).A ubiquitous family of repeated DNA sequences in the human genome. J. Mol. Biol, 132, 289–306.

5, Jurka, J. (1988). A fundamental division in the Alu family of repeated sequences. Proc Natl Acad Sci U S A, 85, 4775-4778.

6. Mathias, S. L. (1991). Reverse transcriptase encoded by a human transposable element. Science, 254, 1808–1810.

7. Deininger, P. L. (1992). Master genes in mammalian repetitive DNA amplification. Trends Genet, 8, 307–311.

8. Paulson, K. E. (1986). Transcriptional inactivity of Alu repeats in HeLa cells. Nucleic Acids Res, 14, 6145–6158.

9. Esteller, M. (2002). Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J Pathol, 196, 1–7.

10. Batzer, M. A. (1990). Structure and variability of recently inserted Alu family members. Nucleic Acids Res, 18, 6793–6798.

11. Arcot, S. S. (1995). Alu repeats: a source for the genesis of primate microsatellites. Genomics, 29, 136–144.

12. Economou, E. P. (1990). The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. Proc. Natl Acad. Sci USA, 87, 2951–2954.

13. Jurka, J. (1995). Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol, 40, 120–126.

14. Zuliani, G. (1990). A high frequency of length polymorphisms in repeated sequences adjacent to Alu sequences. Am. J. Hum. Genet, 46, 963–969.

15. Toth, G. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res, 10, 967–981.

16. Boeke, J. D. (1997). LINEs and Alus — the polyA connection. Nature Genet, 16, 6–7.

17. Moo-Il Kang. (2006). The length of CpG islands is associated with the distribution of Alu and L1 retroelements. Genomics, 87, 580–590.

18. Rodriguez, J. (2007). Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. Nucleic Acids Res, 2007, 1–15.

19, Roy-Engel, M. A. (2005). Retrosequences and Evolution of Alu Elements. Encyclopedia of life sciences, 1-4.

20. Goubely, C. (1999). S1 SINE retroposons are methylated at symmetrical and non-symmetrical position in Brassica napus: identification of a new methylation site in plants. Plant Mol. Biol, 39, 243–255.

21. Kochanek, S. (1993). DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J, 12, 1141– 1151.

22. Ashburner, M. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 1, 25-9.

23. Graff, J. R. (1997). Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in de novo methylation. J. Biol. Chem, 272, 22322–22329.

24. Magewu, A. N. (1994). Ubiquitous and tenacious methylation of the CpG site in codon 248 of the p53 gene may explain its frequent appearance as a mutational hot spot in human cancer. Mol. Cell. Biol, 14, 4225–4232.

25. Anselmo, P. N. (2006). Epigenetic alterations in human brain tumors in a Brazilian population. Genet Mol Biol, 29, 413-422.

26. Macleod D. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev, 8, 2282–2292.

27. Takai D. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA, 99, 3740–3745.

28. Turker MS. (2002). Gene silencing in mammalian cells and the spread of DNA methylation. Oncogene, 21, 5388–5393.

29. Lindsay H. (1996). Spreading of methylation along DNA. Biochem J, 320, 473–478.

30. Bender CM. (1999). Roles of cell division and gene transcription in the methylation of CpG islands. Mol Cell Biol, 19, 6690–6698.

31. Caiafa, P. (2005). DNA Methylation and Chromatin Structure: The Puzzling CpG Islands. J Cell Biochem, 94, 257–265.

32. Arnaud, P. (2000). SINE Retroposons Can Be Used In Vivo as Nucleation Centers for De Novo Methylation. Mol Cell Biol, 5, 3434–3441.

33. Corta′ zar, D. (2007). The enigmatic thymine DNA glycosylase. DNA repair, 6, 489–504.

34. Liddle, A. R. (2006). Epigenetic Silencing of Genes in Human Colon Cancer. Gastroenterology, 131, 960-962.

35. Paz MF. (2003). Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in DNA methyltransferases. Hum Mol Genet, 12, 2209-2219.

36. Alazzouzi H. (2005). Mechanisms of inactivation of the receptor tyrosine kinase EPHB2 in colorectal tumors. Cancer Res, 65, 10170-3.

37. Xu XL. (2004). Methylation profile of the promoter CpG islands of 31 genes that may contribute to colorectal carcinogenesis. World J Gastroenterol, 10, 3441-3454.

38. Eads CA. (1999). CpG island hypermethylation in human colorectal tumors is not associated with DNA methyltransferase overexpression. Cancer Res, 59, 2302-2306.

39. Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modeification maps. Nat Rev Genet, 8, 286-298.

40. Widschwendter, M. (2007). Epigenetic stem cell signature in cancer. Nat Genet, 39, 157-158.

41. Yamada, Y. (2004). A Comprehensive Analysis of Allelic Methylation Status of CpG Islands on Human Chromosome 21q. Genome Res, 14, 247-266.

42. Weber, M. (2005). Chromsome-wide and promter specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet, 37, 853-862.

43. Wiebauer, K. (1989). In vitro correction of G/T mispairs to G/C pairs in nuclear extracts from human cells. Nature, 339, 234–236.

44. Kohany O. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics, 25, 467-474.

45. Rice P. (1997). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet, 16, 276-277.

46. Cartharius K. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics, 21, 2933-42

47. Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics**,** 16**,** 16–23**.**

48. Noble, S.W. (2006). What is a support vector machine? Nat Biotechnol, 24, 1565-1567.

49. Holmes, G. (1994). Weka: A machine learning workbench. Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia.

50. Pirooznia, M. (2008). A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics, 9, 1-13.

51. Cho, N-Y. (2007). Hypermethylation of CpG island loci and hypomethylation of LINE-1 and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features. J Pathol, 211, 269–277.

52. Baylin SB. (2001). Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Hum Mol Genet, 10, 687–692.

53. Brandeis M. (1994). Sp1 elements protect a CpG island from de novo methylation. Nature, 371, 435–438.

54. Bender CM. (1999). Roles of cell division and gene transcription in the methylation of CpG islands. Mol Cell Biol, 19, 6690–6698.

55. Deininger, P. L. (1993). Evolution of retroposons. Evol. Biol, 27, 157–196.

56. Paz, M. F. (2003). Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in DNA methyltransferases. Hum Mol Genet, 12, 2209-2219.

57. Alazzouzi, H. (2005). Mechanisms of inactivation of the receptor tyrosine kinase EPHB2 in colorectal tumors. Cancer Res, 65, 10170-10173.

58. Jones, A. P. (2002). The fundamental role of epigenetic events in cancer. Nat Rev Genet, 3, 415–428.

59. Domingos, P. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. MachineLearning, 29, 103–130.

60. Yuan, Y. F. (1995). Induction of fuzzy decision trees. Fuzzy Sets and Systems, 69,125-139.

61. Croker, B. (2003). Cancer and programmed cell death. Genome Biol, 4, 318-320.

62. Jonathan L. (2006). Metabolic regulation of Akt: roles reversed. J Cell Biol, 175, 845–847.

63. Cox PM. (1991). Transcription and cancer. Br J Cancer, 63, 651-662.

64. Al-Shahrour. (2005). Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments. Nucleic Acids Res, 33, 460-464.

Web sites used in this project.

65. http://www.ensembl.org/Multi/martview

66. http://www.girinst.org/censor/index.php

67. http://www.ebi.ac.uk/emboss/cpgplot/

68. http://www.genomatix.de/products/MatInspector

69. ftp://ftp1.nci.nih.gov/pub/cacore/Cancer%20Gene%20Data%20Curation/

70. http://fatigo.bioinfo.cipf.es/