Prediction of Genomic Methylation Status on CpG Islands Using DNA Sequence Features

YOICHI YAMADA†, KENJI SATOU †

† Graduate School of Natural Science and Technology
Kanazawa University
Kakuma-machi, Kanazawa 920-1192
JAPAN
youichi@is.t.kanazawa-u.ac.jp http://bioinfo.ec.t.kanazawa-u.ac.jp

Abstract: -. In mammals, cytosines of most CpG dinucleotides in their genomes except gene promoters are subject to modification by methyl group (methylation). A number of genes in a mammal are regulated developmental-specifically or tissue-specifically by the methylation. Mammalian DNA methylation contributes to regulation of gene expression, repression of parasitic sequences, inactivation of X chromosome in female, genomic imprinting, etc. Aberrant methylation results in a part of cancers and genetic diseases in human. Therefore it is required that methylation status on human genome is comprehensively revealed in each kind of cells. However, since comprehensive methylation analyses require a lot of times and large labor, methylation status on only a part of genomic regions is revealed in mammals. Because of this, machine learning using already known methylation data and prediction of methylation status on other genomic regions are important. Moreover, since sequence differences between DNA regions showing different methylation status also remain unclear, those differences should be also determined. Therefore we conducted machine learning by support vector machine using our previously reported methylation data, and predicted methylation status on DNA sequences using DNA sequence features. Furthermore we explored different sequence features among four types of methylation using random forest. Consequently high methylation prediction accuracies were observed between two different methylation status pairs. Moreover it was revealed that sequences containing CG, CT or CA were important for discrimination between them.

Key-Words: - CpG island, DNA methylation, Human chromosome 11, Human chromosome 21, Support vector machine, Random forest

1 Introduction

Human genome project has enabled us to obtain the entire sequences of human genomic DNA. It has also revealed that human genomic DNA has genes of about 25,000. When the genes function, the genetic information of them is transcribed to messenger RNA (mRNA) and is generally translated to proteins. The expression of genes is regulated by transcription factors, DNA methylation, histone phosphorylation, acetylation and methylation, etc [1]-[4]. These chemical modifications of DNA or histone proteins are called epigenetic modification. The Epigenetic modification can lead to heritable changes of gene function without any changes of DNA sequence.

Mammalian genomes have less CpG dinucleotides than expected from the numbers of C and G on their DNA sequences [5]. This results from modification by methylation at 5-position of cytocine of CpG dinucleotides, because methylated CpG dinucleotides but not unmethylated CpG dinucleotides tend to be converted to TpG dinucleotides [5],[6].

DNA methylation contributes to regulation of gene expression [7], inactivation of X-chromosome in female [8], repression of parasitic sequences, genomic imprinting [9] and chromosome stabilization [10],[11]. When cytocines of CpG dinucleotides in the promoter region are methylated, transcription factors can not directly or indirectly bind to the promoter region. Therefore DNA methylation in promoter regions generally contributes to suppression of genes.

The vanishment or disorder of DNA methylation system results in embryonic lethality or an incidence of a part of cancers: correct maintenance of DNA methylation system is critical for mammals [12]-[14].

Mammals have enzymes that add a methyl group at 5-position of cytocine of CpG dinucleotides, and therefore most CpG dinucleotides are subject to methylation [15]. However CpG-rich sequences (i.e.,

CpG island) in promoter regions are exceptional DNA regions because they are exempt from methylation in all developmental stages and adult tissues [16]-[18]. Approximate half of genes in a mammal have the CpG islands on their promoter regions [19],[21]. However a part of CpG islands are allele-specifically methylated or tissue-specifically methylated. For example, CpG islands on the X-chromosome in female and in the vicinity of imprinted genes are methylated in an allele-specific manner [8],[21]. Moreover, a part of tissue-specifically expressed genes are biallelically methylated in their repressed tissues.

Genome projects in some organisms have revealed their almost complete genomic sequences, but DNA methylation modification of their genomic sequences has remained unclear. We therefore previously developed HpaII-McrBC PCR (HM-PCR) method to comprehensively examine methylation status of CpG islands on chromosome 21 and 11 [22],[23]. In HM-PCR method, CpG islands are identified *in silico*, and each of those CpG islands is classified into one of four kinds of methylation status (i.e., null, complete, incomplete and composite) using two restriction enzymes with complementary methylation sensitivity. Several other methods have also been developed to comprehensively analyze methylation status on genomic sequences. These methods have revealed methylation status in a part of human genome sequences.

However little is known about the relationship between DNA sequences and their methylation status: we do not know how DNA methylation enzymes discriminate among sequences to which add null, complete, incomplete, or composite methylation. Therefore, to reveal the relationship between DNA sequence composition and methylation status, it is required that computational approaches are applied to these comprehensively analyzed methylation data.

In addition, machine learning using these methylation data and prediction of methylation status on other genomic regions are also important because comprehensive methylation analyses are time-consuming and labor-intensive.

In this study, we therefore conducted machine learning and prediction of methylation status of CpG islands on human chromosome 11 and 21 by support vector machine (SVM) using short DNA sequence features. Moreover, we explored different DNA sequence features among CpG islands with null, complete, incomplete, or composite methylation using random forest.

2 Materials and methods

2.1. Methylation data of CpG islands on human chromosome 11 and 21

The 656 and 149 CpG islands analyzed were identified in the human chromosome 11q and 21q sequences as the regions having the following features: length >200bp, GC content >50%, expected CpG frequency >0.6 for chromosome 11; length >400bp, GC-content >50%, expected CpG frequency >0.6 for chromosome 21 [22],[23]. Methylation status of computationally identified CpG islands was investigated in human peripheral blood leucocytes by HpaII-McrBC PCR (HM-PCR) method. The HM-PCR method can classify CpG islands into 4 classes according to their methylation status (i.e., null, complete, incomplete and composite) [22],[23]. Fig.1 describes the principal of HM-PCR. In Fig.1, each edge-rounded square shows methylation status of two alleles in a single cell. Open circles and squares in edge-rounded squares depict unmethylated HpaII- and McrBC-recognition sites, respectively. In contrast, closed circles and squares in edge-rounded squares depict methylated HpaII- and McrBC-recognition sites, respectively. HpaII is a restriction enzyme which can recognize CCGG sequence and digest unmethylated CCGG sequence but not methylated one. In contrast, McrBC is also a restriction enzyme which can cut methylated $R^mCN_{40\text{-}60}R^mC$ sequence but not unmethylated one. If CpG island is biallelically methylated (i.e., complete methylation), HpaII will not digest this CpG island but McrBC will digest. Therefore we can obtain amplification products from HpaII-digested genome by PCR using the CpG island-specific primer pairs (see right black panel in Fig.1). In contrast, no amplification can be obtained from McrBC-digested genome because template genome for PCR is completely digested by McrBC. By contrast, since CpG island with null methylation is completely digested by HpaII but not by McrBC, PCR specific to this CpG island yields amplification products from McrBC-digested genome but not from HpaII-digested one (see right black panel in Fig.1). The CpG island consisting of completely methylated and unmethylated alleles (i.e., composite methylation) can be completely digested by neither HpaII nor McrBC. Therefore PCR amplification occurs in both HpaII- and McrBC-digested genomes (see right black panel in Fig.1). The CpG island methylated partially in both alleles (i.e., incomplete methylation) is digested by both HpaII and McrBC. Consequently no amplification is obtained from both enzymes-digested

genomes.

645 null methylation- (542 and 103 CpG islands on chromosome 11 and 21, respectively), 116 complete methylation- (87 and 29 CpG islands on chromosome 11 and 21, respectively), 14 composite methylation- (7 and 7 CpG islands on chromosome 11 and 21, respectively), and 25 incomplete methylation-CpG islands (17 and 8 CpG islands on chromosome 11 and 21, respectively) were used in support vector machine and random forest.

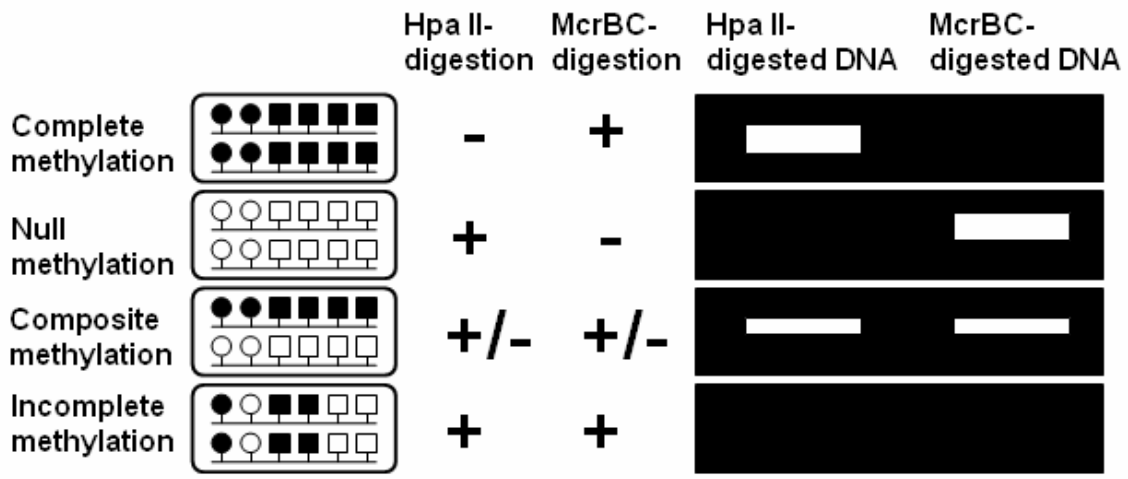

Fig. 1: Principal of HpaII-McrBC PCR method

2.2 Counting of DNA sequence patterns on each CpG island and preparation of vector data.

All the combinations of 4 characters (A, C, G, and T) in length 2-5 were generated. For instance, DNA sequence patterns in the length of 2 are 16 combinations of AA, AC, AG, AT, CA, …, GT, TA, TC, TG, TT. Then K-grams were counted for each CpG island using each of window size 2-5. The counting numbers of DNA sequence patterns in each CpG island were divided by the length of the CpG island. Note that counting of zero in each DNA sequence pattern was also considered as zero. Then each CpG island was converted into a vector having dimensions of the number of DNA sequence patterns. For instance, in DNA sequence patterns of length 2, a vector has 16 dimensions of attributes: AA, AC, AG, AT, CA, …, GT, TA, TC, TG, TT.

2.3 Feature ranking by random forest

2.3.1 Ranking of Important features for discrimination between null- and complete-methylated CpG islands on chromosome 11 and 21

Random forest [24] was applied to determining important features for classification of CpG islands into correct methylation classes (i.e., null and complete). Two matrix data were used in this analysis: one is the vector dataset from CpG islands on chromosome 11 and the other is the vector dataset from CpG islands on chromosome 21. In each matrix data, a row corresponded to a vector (i.e., a CpG island) and a column was an attribute (i.e., a DNA sequence pattern). Since the total number of attributes (i.e., DNA sequence patterns) was 1,360, total number of columns in each matrix data was also 1,360. RandomForest function [25],[26] of R program package was performed for generating a value called MeanDecreaseGini for each attribute. The number of constructed decision trees was 500 and the size of a decision tree was the square root of 1,360. Based on the MeanDecreaseGini, attributes were ranked in the order of importance (discriminative power).

2.3.2 Ranking of Important features for discrimination among CpG islands showing four methylation status on chromosome 11 and 21

Random forest [24] was applied to determining important features for classification of CpG islands into correct methylation classes (i.e., null, complete, incomplete and composite). Compared to CpG islands showing null and complete methylation , incompletely methylated CpG islands (17 and 8 CpG islands on chromosome 11 and 21, respectively) and compositely methylated CpG islands (7 and 7 CpG islands on

chromosome 11 and 21, respectively) are too few. Accordingly, to conduct two class discriminations among four methylations (i.e., null, complete, composite and incomplete) by random forest, we combined CpG islands showing null, complete, incomplete and composite methylation on chromosome 11 with those on chromosome 21, respectively. RandomForest function [25],[26] of R program package was performed for generating a value called MeanDecreaseGini for each attribute. The number of constructed decision trees was 500 and the size of a decision tree was the square root of 1,360. Moreover since large bias of data number between training datasets leads to class misprediction of test data, the sampsize option of randomforest function in R was used, and smaller number of CpG islands between two methylation status was adopt as each training dataset size. Based on the MeanDecreaseGini, attributes were ranked in the order of importance (discriminative power).

2.4 Prediction of DNA methylation status on CpG islands by support vector machine

2.4.1 Prediction of DNA methylation status (null and complete) of CpG islands on chromosome 11 and 21 by support vector machine

We used support vector machine (SVM) for learning and prediction of DNA methylation status (null and complete) on CpG islands. 8 matrix data were used in these analyses: four matrix data (four vector datasets with attributes of DNA sequence patterns of each length) from CpG islands on chromosome 11 and four matrix data (four vector datasets with attributes of DNA sequence patterns of each length) from CpG islands on chromosome 21. In each matrix data, a row corresponded to a vector (i.e., a CpG island) and a column was an attribute (i.e., a DNA sequence pattern). The ksvm function included in the kernlab package [27] for R was used for learning. We used RBF kernel and σ parameter of 0.05 for learning by SVM.

For chromosome 11, we randomly selected 86 from 542 unmethylated CpG islands and 86 from 87 methylated ones as learning data. After learning, remained data were used for prediction. These processes were repeated 87 times and average of their prediction accuracies was calculated. In addition, 28 and 28 out of 103 unmethylated and 29 methylated CpG islands on chromosome 21 were randomly selected for learning by SVM, respectively. Then, remained data were used for prediction. These processes were repeated 29 times and average of their prediction accuracy was calculated. The prediction accuracy was computed as follows:

$$\text{Prediction accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

2.4.2 Prediction of DNA methylation status (null, complete, incomplete and composite) of CpG islands on chromosome 11 and 21 by support vector machine

Compared to CpG islands showing null and complete methylation, incompletely methylated CpG islands (17 and 8 CpG islands on chromosome 11 and 21, respectively) and compositely methylated CpG islands (7 and 7 CpG islands on chromosome 11 and 21, respectively) are too few. Accordingly, to conduct two class discriminations among four methylation status (i.e., null, complete, composite and incomplete) by SVM, we combined CpG islands showing null, complete, incomplete and composite methylation on chromosome 11 with those on chromosome 21, respectively. Moreover since large bias of data number between training datasets leads to class misprediction of test data, training datasets of the same number were sampled from CpG islands showing each of two methylation status. Remained CpG islands from each methylation status were used as test datasets. Fig. 2 describes the detailed preparation process of training and test datasets. In Fig. 2, there are vector data (i.e., CpG islands) of N and M showing methylation status A and B, respectively. Here suppose that M is much larger than N. For preparation of test data, one and M-(N-1) vector data from methylation status A and B were randomly extracted, respectively. Remained vector data of N-1 from each methylation status (i.e., methylation status A and B) were used as training data. Then in next sample preparation, different one vector data from previous times was selected as test data from methylation status A, and M-(N-1) vector data were randomly sampled as test data from methylation status B. Remained N-1 vector data from methylation status A and B were used as training data. These processes were repeated N times and average of their prediction accuracies was calculated.

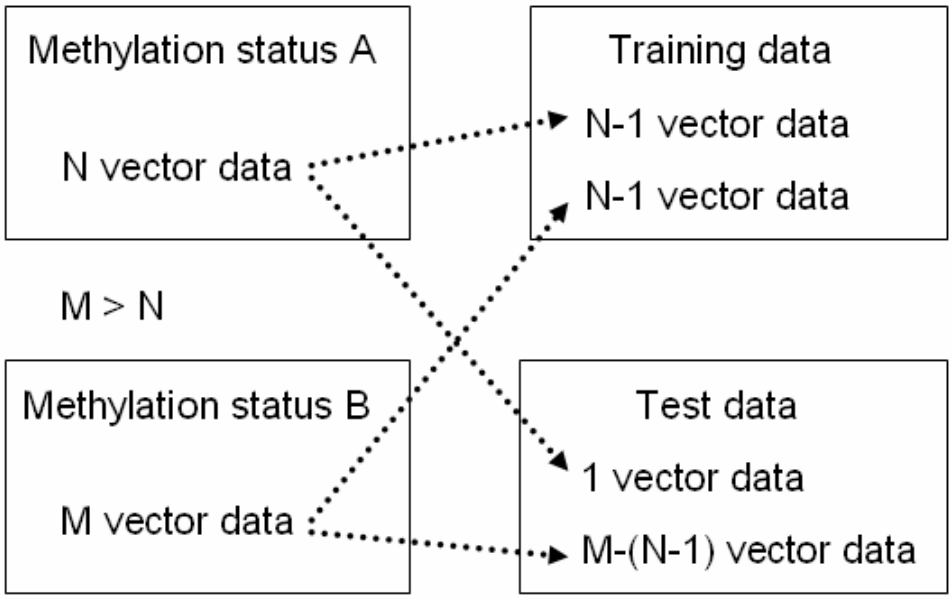

Fig. 2: Preparation process of training and test datasets for support vector machine

Here the ksvm function included in the kernlab package [27] for R was used for learning. We used RBF kernel and σ parameter of 0.05 for learning by SVM.

The prediction accuracy was computed as follows:

$$\text{Prediction accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

3 Results

3.1. Prediction of DNA methylation status (null and complete) of CpG islands on chromosome 11 and 21

Epigenetic modifications including DNA methylation are subject to erasure in early development. After that, epigenetic modifications regenerate in DNA or histone proteins during later development. In this reprogramming process, how modification enzymes discriminate sequences to which add chemical modifications from those to which do not add? The answer remains unclear but Bock et al. [28] reported the differences of sequences between unmethylated and methylated CpG islands on chromosome 21. In that report, unmethylated CpG islands had more CpG dinucleotides and less TpG or CpA dinucleotides than methylated CpG islands because methylated CpG dinucleotides are prone to convert to TpG dinucleotides. Therefore we examined whether other sequence features contribute to discrimination between unmethylated and methylated CpG islands using methylation data of chromosome 11. In addition, we examined whether there are any common DNA sequence features between chromosome 11 and 21 for the classification of CpG islands into correct methylation classes.

Table 1 shows the ranking top ten of important attributes (i.e., DNA sequence patterns) based on MeanDecreaseGini of random forest for discriminating between unmethylated and methylated CpG islands on chromosome 11 and 21. In the ranking top ten of both chromosomes, DNA sequence patterns containing CG, TG, or CA were important attributes and there were not any important attributes that do not contain those dinucleotides. These results supported the report by Bock et al. [28] in which frequencies of CpG, TpG, and CpA are different between unmethylated and methylated CpG islands. As was pointed out by Bock et al., differential mutation rates of CpG → TpG between unmethylated and methylated CpG islands may lead to these differential frequencies of sequence patterns containing CG, TG, or CA [28].

Table 2 shows the prediction accuracy of methylation status of CpG islands on each chromosome. In Table 2, “Dataset” means vector data (i.e., CpG islands) and their attributes (i.e., DNA sequencing patterns) used in SVM. For example, “Attributes_of_length2_chromosome11” describes CpG islands on chromosome 11 and DNA sequence

patterns of length 2. As shown in Table 2, high prediction accuracies over 80% were obtained in both chromosome data. Moreover, when CpG islands on chromosome 21 were used as training and test data, better prediction accuracies were obtained. These results seem to result from that methylated CpG islands on chromosome 21 but not unmethylated ones frequently have tandem repeat sequences. Since tandem repeat sequences tend to have biased frequencies of relatively long DNA sequence patterns, it may be easier to separate unmethylated CpG islands from methylated ones on chromosome 21. This interpretation was also supported by that chromosome 21 displayed longer DNA sequence patterns as important attributes in Table 1 compared to chromosome 11. Moreover longer DNA sequence patterns and mixtures of DNA sequence patterns of length 2-5 displayed better prediction accuracy in both chromosomes (see Table 2). These results suggest that more attributes lead to better prediction accuracy because they give each vector more information.

3.2. Prediction of DNA methylation status (null, complete, incomplete and composite) of CpG islands on chromosome 11 and 21

Although HM-PCR method can classify CpG islands into four methylation status (i.e., null, complete, incomplete and composite), previous reports including Bock et al. analyzed only discrimination between null and complete methylation. Therefore we tried two class discriminations among these four methylation status by support vector machine. Table 3 shows the results of two class discriminations among four methylation status for CpG islands on human chromosome 11 and 21. (A)-(F) in Table 3 describe each result of prediction accuracies for all combinations among null, complete, incomplete and composite methylation.

Table 1: Important attributes for classification between null and complete methylation classes by random forest

Ranking	Chromosome11	Chromosome21
1	CG	CGCGG
2	TG	CGCCG
3	GCG	CCGC
4	CGG	CGCG
5	GCGC	GGGCG
6	CGGGA	CCCGC
7	CCGG	GCGCA
8	GGCG	CCCG
9	CGGA	CCCGG
10	CCCG	GCGCG

Table 2: Prediction accuracy of methylation status of CpG islands by support vector machine

Dataset	Prediction_accuracy (%)
Attributes_of_length2_chromosome11	82.05
Attributes_of_length3_chromosome11	84.82
Attributes_of_length4_chromosome11	86.7
Attributes_of_length5_chromosome11	87.1
Attributes_of_length2,3,4,5_chromosome11	87.58
Attributes_of_length2_chromosome21	85.16
Attributes_of_length3_chromosome21	88.29
Attributes_of_length4_chromosome21	91.79
Attributes_of_length5_chromosome21	90.75
Attributes_of_length2,3,4,5_chromosome21	91.7

"Dataset" in Table 3 describes length of DNA sequence patterns (i.e., attributes) and CpG islands (i.e., vectors) used in methylation prediction by support vector machine. For example, "Attributes_length2_chromosome11_21" describes DNA sequence patterns of length 2 and CpG islands on chromosome 11 and 21. As (A) in Table 3 shows, prediction accuracies between null and complete methylation were over 80% in all length of DNA sequence patterns. These findings corresponded to separately analyzed results of CpG islands on each chromosome (see Table 2). In addition, classification between null and composite methylation also displayed high prediction accuracies over 80% in DNA sequence patterns of length 4, 5 and mixture of length 2-5. However over-predictions of null methylation were observed in these attributes (see Table 3). Table 4 dedicates details of methylation prediction between null and composite methylation. As shown in Table 4, although attributes of length 2 and 3 in methylation prediction of compositely methylated CpG islands showed accuracies over 50%, attributes of length 4, 5, and mixture of length 2-5 showed accuracies under 50% (see rows of "composite" in (A)-(E) of Table 4). However the attribute of length 3 in "null vs composite" showed better accuracy than results of the other methylation predictions except "null vs complete" (see (A)-(F) in Table 3). In addition, classifications between null or incomplete and complete or composite displayed relatively higher accuracies (see Table 3). In contrast, prediction accuracies were relatively lower between null and incomplete or between complete and composite (see C and D in Table 3). These may be because there are any relationships between null and incomplete or between complete and composite.

Table 3: Prediction accuracy of methylation status (null, complete, composite and incomplete) of CpG islands by support vector machine

(A) Null vs Complete

Dataset	Prediction_accuracy (%)
Attributes_length2_chromosome11_21	84.06
Attributes_length3_chromosome11_21	87.1
Attributes_length4_chromosome11_21	88.98
Attributes_length5_chromosome11_21	89.99
Attributes_length2,3,4,5_chromosome11_21	89.98

(B) Null vs Composite

Dataset	Prediction_accuracy (%)
Attributes_length2_chromosome11_21	71.36
Attributes_length3_chromosome11_21	79.46
Attributes_length4_chromosome11_21	87.15
Attributes_length5_chromosome11_21	90.53
Attributes_length2,3,4,5_chromosome11_21	87.36

(C) Null vs Incomplete

Dataset	Prediction_accuracy (%)
Attributes_length2_chromosome11_21	60.43
Attributes_length3_chromosome11_21	63.94
Attributes_length4_chromosome11_21	65.59
Attributes_length5_chromosome11_21	64.51
Attributes_length2,3,4,5_chromosome11_21	66.99

(D) Complete vs Composite

Dataset	Prediction_accuracy (%)
Attributes_length2_chromosome11_21	50.34

Attributes_length3_chromosome11_21	49.18
Attributes_length4_chromosome11_21	46.84
Attributes_length5_chromosome11_21	53.64
Attributes_length2,3,4,5_chromosome11_21	51.72

(E) Complete vs Incomplete

Dataset	Prediction_accuracy (%)
Attributes_length2_chromosome11_21	67.4
Attributes_length3_chromosome11_21	69.46
Attributes_length4_chromosome11_21	71.87
Attributes_length5_chromosome11_21	76.34
Attributes_length2,3,4,5_chromosome11_21	75.01

(F) Incomplete vs Composite

Dataset	Prediction_accuracy (%)
Attributes_length2_chromosome11_21	59.89
Attributes_length3_chromosome11_21	65.38
Attributes_length4_chromosome11_21	72.53
Attributes_length5_chromosome11_21	75.27
Attributes_length2,3,4,5_chromosome11_21	74.18

Table 4: Detailed results of prediction accuracies between null and composite methylation

(A) Attributes_length2_chromosome11_21

	composite	non
composite	0.64	0.36
non	180.93	451.07

(B) Attributes_length3_chromosome11_21

	composite	non
composite	0.64	0.36
non	129.64	502.36

(C) Attributes_length4_chromosome11_21

	composite	non
composite	0.43	0.57
non	80.79	551.21

(D) Attributes_length5_chromosome11_21

	composite	non
composite	0.36	0.64
non	59.29	572.71

(E) Attributes_length2_5_chromosome11_21

	composite	non
composite	0.43	0.57
non	79.43	552.57

Here one possible interpretation is that mutation rate of CG → TG is different between null or incomplete and complete or composite. If this interpretation is correct, completely and compositely methylated but not incompletely and unmethylated CpG islands will be methylated in germ line cells. This is because only conversion of CG → TG in germ line cells but not in somatic cells is transmitted to next generation.

Furthermore since “non vs complete” and “non vs composite” displayed relatively higher prediction

accuracies, important DNA sequence patterns for discrimination between non and complete or between non and composite were examined using MeanDecreaseGini of random forest (see Table 5). Table 5 shows the ranking top ten of important attributes (i.e., DNA sequence patterns of length 2-5) based on MeanDecreaseGini of random forest. Left (non vs complete) and right (non vs composite) columns in Table 5 show that DNA sequence patterns containing CG, TG, or CA are important for discrimination between non and complete or between non and composite. This result suggests that compositely methylated CpG islands are methylated in human germ line cells.

Table 5: Important attributes for classification of CpG islands into correct methylation classes by random forest

Ranking	ch11_ch21_non_complete_2_5letters	ch11_ch21_non_composite_2_5letters
1	TG	CGGAC
2	CATG	TCCCG
3	CCCGG	GACGT
4	CGCG	CCTCG
5	CCCG	TCCC
6	CCG	GCCTC
7	CCGG	CTCCT
8	GCG	CA
9	CGGGA	CGCGG
10	CCGGG	ACACG

4 Conclusion

In this study, we examined whether sequence features on CpG islands are important for their methylation status using random forest and SVM. The learning and prediction by SVM showed that DNA sequence features could distinguish unmethylated CpG islands from completely methylated ones in high prediction accuracies of over 80%. This result was consistent with previous report of Bock et al. for CpG islands on human chromosome 21. In addition, our result suggested that frequencies of sequences containing CG, CT or CA are different between unmethylated and methylated CpG islands on chromosome 11 as well as chromosome 21. Furthermore, when CpG islands on chromosome 21 were used as training and test data, better prediction accuracies were obtained compared to chromosome 11. These results may result from higher frequency of methylated CpG islands with tandem repeat sequences on chromosome 21 than chromosome 11. In addition, metylation prediction of two classes among four methylation status by SVM revealed relatively high prediction accuracy between unmethylated and compositely methylated CpG islands on chromosome 11 and 21. This result may result from different mutation rate of CG → TG not only between unmethylated and methylated CpG islands but also between unmethylated and compositely methylated CpG islands.

References:

[1] O. Hobert, Gene regulation by transcription factors and microRNAs, *Science*, Vol.319, 2008, pp.1785-1786.

[2] R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, *Nat. Genet.*, Vol.33, 2003, pp.245-254.

[3] Y. Yamada, Y. Miyata, M. Higashihara, K. Satou, Comparison of Cluster Identification Methods for Selection of GO Terms related to Gene Clusters, *WSEAS Transactions on Biology and Biomedicine*, Vol.5, 2008, pp.54-63.

[4] D.A. Charlebois, A.S. Ribeiro, A. Lehmussola, J. Lloyd-Price, O. Yli-Harja, S.A. Kauffman, Effects of microarray noise on inference efficiency of a stochastic model of gene networks, *WSEAS Transactions on Biology and Biomedicine*, Vol.4, 2007, pp.15-21.

[5] L. Ponger, L. Duret, D. Mouchiroud, Determinants of CpG islands: expression in early embryo and isochore structure, *Genome Res.*, Vol.11, 2001, pp.1854-1860.

[6] D.F. Schorderet and S.M. Gartler, Analysis of CpG suppression in methylated and nonmethylated species, *Proc. Natl. Acad. Sci. USA*, Vol.89, 1992, pp.957-961.
[7] R. Jaenisch and A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, *Nat. Genet*., Vol.33, 2003, pp.245-254.
[8] D.P. Norris, N. Brockdorff, S. Rastan, Methylation status of CpG-rich islands on active and inactive mouse X chromosomes, *Mamm. Genome*, Vol.1, 1991, pp.78-83.
[9] I.M. Morison, and A.E. Reeve, A catalogue of imprinted genes and parent-of-origin effects in humans and animals, *Hum. Mol. Genet*., Vol.7, 1998, pp.1599-1609.
[10] A.F. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes, *Curr. Opin. Genet. Dev.*, Vol.9, 1999, pp.657-663.
[11] R.Z. Chen, U. Pettersson, C. Beard, L. Jackson-Grusby, R. Jaenisch, DNA hypomethylation leads to elevated mutation rates, *Nature*, Vol.395, 1998, pp.89-93.
[12] E. Li, T.H. Bestor, R. Jaenisch, Targeted mutation of the DNA methyltransferase gene results in embryonic lethality, *Cell*, Vol.69, 1992, pp.915-926.
[13] M. Okano, D.W. Bell, D.A. Haber, E. Li, DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development, *Cell*, Vol.99, 1999, pp247-257.
[14] B.B. Stephen, E. Manel, R.R. Michael, E.B. Kurtis, S. Kornel, G.H. James, Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer, *Hum. Mol. Genet*., Vol.10, 2001, pp.687-692.
[15] J. Turek-Plewa, P.P. Jagodziñski, The role of mammalian DNA methyltransferases in the regulation of gene expression, *Cell Mol.Biol.Lett*., Vol.10, 2005, pp.631-647.
[16] C. Grunau, W. Hindermann, A. Rosenthal, Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes, *Hum. Mol. Genet.*, Vol.9, 2000, pp.2651-2663.
[17] D. Macleod, R.R. Ali, A. Bird, An alternative promoter in the mouse major histocompatibility complex class II I-Aâ gene: implications for the origin of CpG islands, *Mol. Cell. Biol*., Vol.18, 1998, pp.4433-4443.
[18] I.P. Ioshikhes and M.Q. Zhang, Large-scale human promoter mapping using CpG islands, *Nat. Genet*., Vol.26, 2000, pp.61-63.
[19] M. Gardiner-Garden and M. Frommer, CpG islands in vertebrate genomes, *J. Mol. Biol*., Vol.196, 1987, pp.261-282.
[20] F. Antequera, and A. Bird, Number of CpG islands and genes in human and mouse, *Proc. Natl. Acad. Sci. USA*, Vol.90, 1993, pp.11995-11999.
[21] R. Holmes and P.D. Soloway, Regulation of imprinted DNA methylation. *Cytogenet Genome Res.*, Vol.113, 2006, pp.122-129.
[22] Y. Yamada, T. Shirakawa, T. Taylor, K. Okamura, H. Soejima, M. Uchiyama, T. Iwasaka, T. Mukai, K. Muramoto, Y. Sakaki, T. Ito, A comprehensive allelic methylation analysis of CpG islands on human chromosome 11q, *DNA Sequence*, Vol.17, 2006, pp.300-306.
[23] Y. Yamada, H. Watanabe, F. Miura, H. Soejima, M. Uchiyama, T. Iwasaka, T. Mukai, Y. Sakaki, T. Ito, A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q, *Genome Research*, Vol.14, 2004, pp.247-266.
[24] L. Breiman, Random Forests, *Machine Learning*, Vol.45, No.1, 2001, pp.5-32.
[25] A. Liaw and M. Wiener, Classification and Regression by randomForest, *Rnews 2002*, Vol.2, 2002, pp.18-22.
[26] M. Higashihara, J.D. Rebolledo-Mendez, Y. Yamada, K. Satou, Application of a Feature Selection Method to Nucleosome Data: Accuracy Improvement and Comparison with Other Methods, *WSEAS Transactions on Biology and Biomedicine*, Vol.5, 2008, pp.95-104.
[27] A. Karatzoglou, A. Smola, K. Hornik and A. Zeileis, kernlab – An S4 Package for Kernel Methods in R, *Journal of Statistical Software*, Vol.11, No.9, 2004, pp.1-20.
[28] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, J. Walter, CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure, *PLoS Genet*., Vol.2, 2006, e26.