Journal of Advanced Engineering Trends
ISSN : 2682 - 2091

Vol.43, No.2. July 2024

http://jaet.journals.ekb.eg

Predicting DNA Methylation state of CpG Islands Using Machine Learning

Asmaa Abo Bakr Kamel(1) Esraa M. Hashem(2) Mai S. Mabrouk(3) Mohamed W. Fakhre(4)

College of Computing and Information Technology (CCIT), Arab Academy for ScienceTechnology and Maritime Transport (AASTMT) Cairo, Egypt(1)
Biomedical Engineering Department, Misr University for Science and Technology (MUST University), 6th of October, Egypt(2)
Biomedical Engineering Department, Misr University for Science and Technology (MUST University), 6th of October, Egypt(3)
Computer Engineering Department Arab Academy for Science, Technology & Maritime Transport Cairo, Egypt.(4)

ARTICLE INFO

Article history:
Received:
Accepted:
Online:

Keywords:
DNA methylation
logistic regression
support vector machine
CpG Islands
Random forest

ABSTRACT

DNA methylation is the primary and best understood epigenetic element that controls human health. It is an essential regulator of gene transcription. Methylation may be the head of some diseases like Parkinson's, cardiovascular, chronic kidney, cancer, and Alzheimer's. The implementation of models to predict DNA methylation has been concentrated by researchers in the bioinformatics area, according to the difficulties of predicting the methylation that is very sensitive to lifestyle or pollution changes. Recent improvements in methylation sequencing way permit the recognition of genome-wide methylated sites in DNA. In the represented work, computational methods are used to predict the methylation of DNA for every CpG locus and non-CpG locus in the whole genome, utilizing Illumina 450K array data within the 250bp region around every CpG site of the human embryonic stem cell with three classifiers including logistic regression, support vector machine, and random forest. The proposed classifiers have been evaluated. Results show that the best performance criteria came from the random forest approach giving an accuracy of 99.9% for a methylation status compared to the other two classifiers. Expressing more features will lead to higher prediction performance and wider detection coverage for methylation of CpG loci.

1. Introduction

Bioinformatics is applied in various applications such as biomedicine applications, microbiology, and agriculture. It is incorporating large sizes of data with tools to predict, analyze and help in clinical and preclinical results analysis, which is in a short period with low cost and low risk. While in microbiology applications, studying the micro-organism’s genome can be understanding these microbes at every fundamental level. Also, completing the micro-organisms genome sequences has provided a greater insight into the microbial world. [1- 2].

DNA methylation is an operation that adds methyl groups to the DNA molecule, especially onto a cytosine base, as shown in Figure 1, the addition of methyl group (-CH $_{2}$) to the DNA molecule, more specifically to the cytosine ring. In which methylation can modify the performance of a DNA segment without modifying the sequence. Whereas cytosine is matched with guanine, after methylation, thymine is mismatched with guanine [3- 4-5]. CpG(cytosine phosphate guanine) sites are methylated by one of three DNA methyltransferases (DNMTs), DNMT1 maintains the methylation state in the daughter strands and *De novo* DNA methylation at the cytosine in CpG dinucleotides is initiated by DNMT3.

Revised:29 June, 2022, Accepted:18 August , 2022

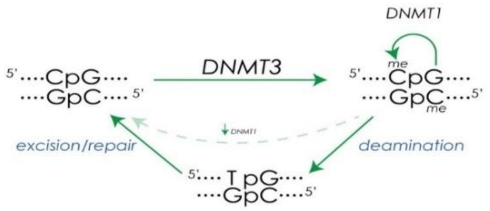


Figure 1 **Mechanism of DNA methylation**

One of the reasons that increase the difficulty of the methylation prediction is the lowest presence of CpG locus related to the whole genome, as searching for a needle inside a heap of straw.

The whole human genome sequence contains about 29 million CpG loci among 3.2 billion bases in DNA. The methylation of the CpG locus is in a range from 60% to 80% in mammals. Another reason may be an external factor where methylation is influenced by environment or lifestyle. Besides, an internal factor may be cell types or diseases [3]. DNA methylation is the most important epigenetic modification that appears shiny in humans, especially in CpG dinucleotide [6].

Methylation can affect gene silencing, genomic imprinting, X-chromosome inactivation, the silencing of intra-genomic parasites, stem cell discrimination, embryonic development, DNA replication, inflammation, aging, and carcinogenesis as a huge number of experiments proved [7].

Many researchers exert great efforts to predict DNA methylation. The DNA methylation prediction difficulties result from the un-correlation within a genomic locus that can be caused by many influencing divergent factors [8]. One of the factors that can be impacted by methylation is a lifestyle such as nutrition, diet, obesity, tobacco smoking, psychological stress, and working night shifts [9]. Implementing computational models that may be used on any cell type or area of the genome without affecting the model's performance is a challenge.

Embryonic stem cells (ESC) are the most primitive type of stem cells. ESC can generate all types of cells in the human body. Accordingly, scientists focus on them to be used as therapies for diseases in the future [10]. The importance of studying ESC is raised especially chromosome 18 which can be utilized to extract features. Bhasin et al., [11] proposed a support vector machine (SVM) to predict binary methylation status that gives about 75% accuracy. Bock et al., [12] suggested SVM to predict binary methylation status in CpG islands, giving about 91.5% accuracy. Shicai Fan et al., [13] proposed an SVM classifier to predict the binary methylation status for the CpG locus that calculated the performance criteria for only chromosome 6 and chromosome 21, giving about 90% accuracy.

In this work, The methylation was predicted on every CpG locus by extracting specific sequence features from whole-genome bisulfite sequencing (WGBS) data and Illumina 450K array data within the 250bp region around every CpG site by using three machine learning classifiers: support SVM, Random forest (RF), and Logistic regression (LR) to expand coverage of detecting for methylation of CpG loci than experiments limited and to define and implement techniques that predict methylation for every CpG locus accurately.

RF was implemented by Breiman et al. [14] to be the best "off-the-shelf" classifier for high-dimensional data. It is one of the most common and powerful ensemble methods used today in machine learning, the idea of RF is to de-correlate the several trees which are generated on the diverse bootstrapped samples from training Data [15].

Averaging the trees reduces the variance and enhances the performance on a test set to avoid overfitting. The technique works in two stages: the first stage is a random forest creation, and then the second stage is to predict the RF classifier created in the first stage. The votes for each predicted target will be calculated. In which the highest voted predicted target is counted as the latest prediction from the RF algorithm. The main advantages of RF are that it is easy to compute, it can efficiently process data, and is fault-tolerant to missing or unbalanced data.

SVMs are a set of related supervised learning techniques, applicable to both classification and regression. SVM is based on the statistical learning theory that was developed by Vladimir Vapnik in 1995 [16]. SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space. The main idea of SVM is to find the optimal hyperplane, which maximizes the margin between two classes. Vectors that define the hyperplane are the support vectors. While the hyperplane is just a line in 2D or a plane in 3D that can separate two classes [17].

LR is a technique of statistical classification which is used to predict binary classification. It measures the relationship between a categorical attached element and one or more predictor elements. The Generalized linear model was implemented by Nelder and Wedderburn in 1972 to use linear regression in problems that were not directly suited for the application of LR. [18]. The binary categorical can pick only 2 values like 1 or 0. The objective is to determine a mathematical equation that can be utilized to predict the probability of event 1[19].

2. Material and Methods

DNA methylation is playing an important role in cell differentiation and diseases such as cancers. The block diagram of the proposed system demonstrates in figure 2. The system started by preprocessing input data that consists of both sequence data and an average value of the methylation data. Then, classify windows of various sizes around each CpG site.

Then, the features had been extracted from each window to distribute into four groups, the first group contained about 37 features, the second group contained about 111 features, then the third group contained about 125 features, and finally, the fourth group contained about 209 features. After extracting all the feature groups inside all windows that were determined on chromosome 18 of the ESC to train and test using three classifiers as RF, SVM, and LR to evaluate the system.

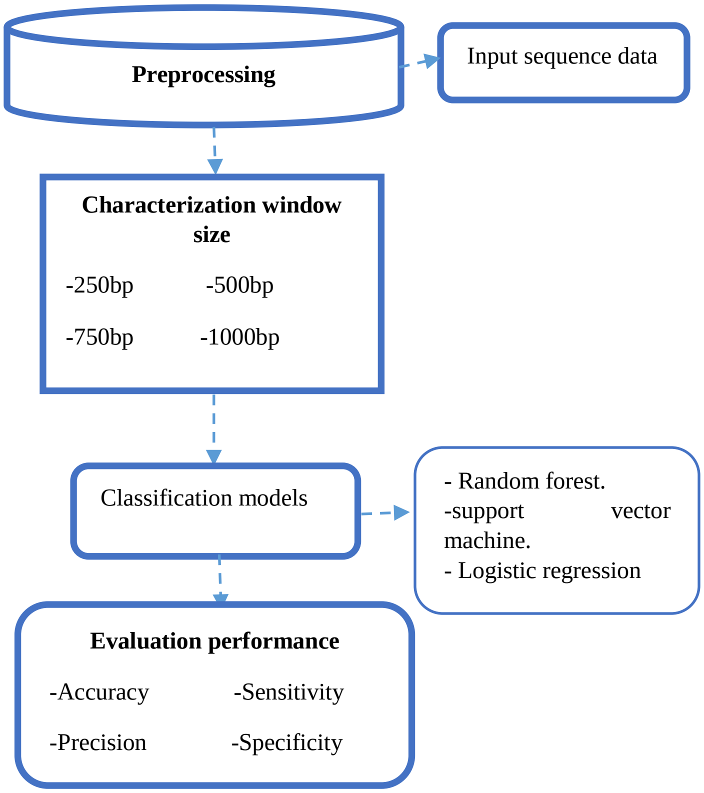


Figure 2 Block diagram of the proposed work

2.1. Dataset:

The data was downloaded from the GEO database in NCBI [20] of H1 ESC (accession number GSM432685) as data for a training model. While the H9 ESC data (accession number GSM706059) was downloaded as data for a testing model. The methylation value had been calculated from the Illumina Human Methylation 450K array data.

The data was downloaded from the GEO database in NCBI of H1 ESC (accession number GSM853420) as data for the training model. The H9 ESC data (accession number GSM853421) was downloaded as data for a testing model from WGBS. Data were normalized with internal controls according to Illumina´s standard procedures. The methylation level at each locus was calculated with the Genomestudio methylation module as a beta-value (ranging from 0 to 1) [21].

2.2. Characterization of window size and features extraction:

After loading both sequence data and an average value of the methylation data, classify many windows of various sizes starting from 250bp, 500bp, 750bp, and 1000bp throughout each CpG locus, while the three classifiers were trained on a cell line H1 of embryonic stem cell.

Then the features had been extracted from each window to distribute into four groups.

For the first group, the sequence features a recurrence of 1-mer and 2-mer DNA base pairs and has been extracted inside the window which giving 20 features [22]. Another 16 features were calculated as the NpN ratio inside each window [23] as shown by equation 1.

$$\text{NpN ratio} \ = \frac{\text{\#NpN x len(Sequence)}}{\text{\#N x \#N}} \qquad (1)$$

where (N = A/C/G/T) [23]. The last feature was calculated from the Illumina Infinium Human Methylation 450 bead array data (450k array), in which the average of the methylation beta value was calculated by collecting data of the 450k array inside a range of both 10 Kbp and 100 Kbp regions surrounding each CpG locus inside the whole-genome bisulfite sequencing data, to select the highest value from both.

For the second group, the first 110 features were extracted from the whole-genome bisulfite sequencing data by computing 1-mer, 2-mer, and 3-mer DNA base pairs that gave 84 features, then the NpN ratio was calculated that give 16 features, and the NpN content was calculated that giving 10 features. To calculate the NpN content as in equation 2.

$$\text{NpN content} \ = \frac{\text{\#N} + \text{\#N}}{\text{len(Sequence)}} \qquad (2)$$

The last feature was calculated from the Illumina Infinium Human Methylation 450 bead array data, in which the average of the methylation beta value was calculated by collecting data of the 450k array inside a range of both 10 Kbp and 100 Kbp regions surrounding each CpG locus inside the whole-genome bisulfite sequencing data, to select the highest value from both. For the third group, the first 124 features were extracted from the whole-genome bisulfite sequencing data by computing 1-mer, 2-mer, and 3-mer DNA base pairs that gave 84 features, and DNA sequence patterns that gave 14 features. then the NpN ratio was calculated that giving 16 features, then the NpN content was calculated that giving 10 features. To calculate the NpN ratio as in equation 1, and to calculate the NpN content as in equation 2. The methylation value takes 0 for un methylated value and takes 1 for the methylated value

For the fourth group, the first 208 features were extracted from the whole-genome bisulfite sequencing data by computing 1-mer, 2-mer, and 3-mer DNA base pairs that gave 84 features, and DNA sequence patterns as in table 1 and table 2 were extracted that giving 98 features, then the NpN ratio was calculated that giving 16 features, then the NpN content was calculated that giving 10 features.

2.3. Classification:

By implementing R programming, RF has been created by an RF package using the RF glm function [24], the model was trained where the last column was the predictor column, and all other columns were the features.

For SVM, the model was trained where the last column was the predictor column, and all other columns were the features. SVM has been created by an SVM, e1071 package [24].

After specifying the data frame, the kernel was specified as a radial for nonlinearity, then specifying a value for the cost parameter to know the error budget C in the model. With a radial kernel, the gamma parameter should be assigned a positive value. A small gamma value defines a Gaussian function with a large variance and vice-versa [19]. The radial kernel is outlined as Kernel $_{\text{radial}}$(x-x') = exp {-γ $\|$x-x'$\|^2$} where γ is a parameter that sets the "spread" of the kernel [25]. Next, the evaluation metrics of the test data were examined to calculate the evaluation metrics of the model.

For LR, the binary classification is achieved by contacting the log odds of the event ln(P/1−P), where P is the probability of an event. So, P always lies between 0 and 1. Equation 3.

$$\text{Z}_\text{i} = \ln\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n \qquad (3)$$

2.4. Evaluation Matrices:

The evaluation metrics for the models was evaluated using Sensitivity (%), Specificity (%), Precision (%) and Accuracy (%) calculated by the following equations 4-7:

$$\text{Sensitivity}(\%) = \frac{TP}{TP + FN} \text{X } 100 \quad (4)$$

$$\text{Specificity}(\%) = \frac{TN}{TN + FP} \text{X } 100 \quad (5)$$

$$\text{Precision}(\%) = \frac{TP}{TP + FP} \text{X } 100 \quad (6)$$

$$\text{Accuracy}(\%) = \frac{TP + TN}{TN + TP + FN + FP} \text{X } 100 \quad (7)$$

Where TP is True Positive, which means that the value of actual class is yes, and the value of predicted class is also yes. TN is True Negative, which means that the value of actual class is no, and value of predicted class is also no. FP is False Positive, when actual class is no and predicted class is yes, and FN is False Negative, when actual class is yes but predicted class in no. In which methylation has occurred when methylation value is greater than 0.5 were represented by 1, but if the methylation value is less than 0.5, un-methylation occurs and represented by 0.

3. Results and discussions:

The DNA methylation prediction is challenging to predict because methylation can be changed according to the cell's genomic regions or type. The methylation of DNA can also be influenced by the environment, aging, lifestyle, and pollution. As a challenge, to implement computational models that can be applied to any cells or any genomic region without changing the model's performance. DNA methylation is the most researched epigenetic mark involved in various processes in the human cell, including gene regulation and development of diseases, such as cancer. An early study in cancer systems, at candidate regions, quickly revealed that losses and gains of DNA CpG methylation were remarkably detected at the sites of tumor suppressors and oncogenes. Alterations in DNA methylation are known now to interact with genetic events and to be involved in human carcinogenesis. Therefore, knowing the concept of DNA methylation is critical for the analysis of disease processes.

At the presented work, the WGBS data of chromosome 18 in the cell line H1 of the human ESC was selected to define each CpG site. Every CpG site was chosen as a central position inside a window determined around. Various sizes of windows were including each window, many different sequence features had been extracted and combining with the chosen of the highest value of average beta in 10kbp and 100kbp from 450K array data. All features were divided into four groups to facilitate the comparison between them to choose the best group. The four groups were illustrated as following: The first group of features contained 37 features that consist of 1mer, 2mer, NpN ratio (equation1). The second group of features contained 111 features that consist of the above 37 features in the first group, added to them 3-mer and NpN content (equation 2).

The third group of features contained 125 features that consist of the above 111 features in the second group, added to them 14 frequency patterns [7] as shown in Table 1.

Table 1 Patterns for 14 frequencies

Window pattern frequency	Feature's number
AAWGGR	4
TGRAAT	2
ATGVAA	3
CCGC	1
CCCC	1
CGCC	1
CTCC	1
AAAG	1

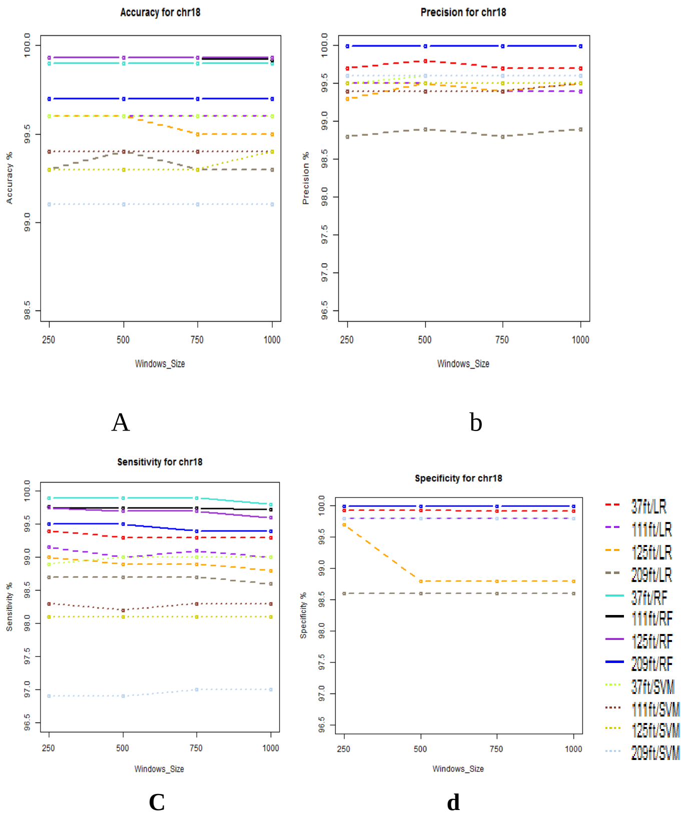
The fourth group of features contained 209 features that consist of the above 125 features in the third group, added to them 84 frequency patterns [23] as shown in Table 2.

Table 2 Patterns for 84 frequencies

Window pattern frequency	Feature's number
CCGSSC	4
TCCSSG	4
SGMGCC	4
CCDGGV	9
BCCCWG	6
GGVCCH	9
CCCWGH	6
GGSCTB	6
CCTGMV	6
GMCCCN	8
SCCWCR	8
WGCCCH	6
CKGSCM	8

The three proposed classifiers had been implemented to predict the methylation and unmethylation for each CpG site. The evaluation metrics for every model were measured by testing the models on the data of chromosome 18 of the cell line H9 of the human ESC. The curves in Figure 3 (a) were given to show the relation between the various window sizes and the percentage of accuracy for the four feature groups when applied on classifiers. The RF model was observed that giving the highest accuracy of the first three groups of features and slightly decrease by window size increase (table 3). While in Figure 3 (b) shows the relation between the various window sizes and the percentage of precision for the four feature groups when applied on the classifiers.

The RF model was observed that giving the highest precision at all feature groups and without any change while the window size increases. Whilst in Figure 3 (c) explained the relevance between the various window sizes and the percentage of sensitivity for the four feature groups. The RF model was observed that giving the highest sensitivity of the first group feature and decreases by the increase of the features and window size. Finally, in Figure 3 (d) revealed that the relevance between the various window sizes and the percentage of specificity for the four feature groups. The RF model was observed that give the highest specificity of all feature groups and without any change while window size increases. Hence, the results show that RF model had the highest evaluation metrics with the first group of 37 features and a window of size 250bp. Consequently, the first group of 37 features was extracted inside a window of size 250bp to apply to the rest of all chromosomes. The RF, SVM, and LR classifiers were trained on the data of all chromosomes of the cell line H1 of the human ESC. The evaluation metrics for every model were measured by testing the models on the data of all chromosomes of the cell line H9 of the human ESC given curves in the following figures 3.



A

b

C

d

Figure 3 The performance criteria for chromosome 18.

Table 3The Percentage of Accuracy inside Window of Size 250 bp

	L R	RF	SVM
Group 1	99.7%	99.9%	99.6%
Group 2	99.6%	99.9%	99.4%
Group 3	99.6%	99.9%	99.3%
Group 4	99.3%	99.7%	99.1%

By comparing the percentage of sensitivity for all 22 pairs of chromosomes added to the X and Y chromosomes, conclude that the RF gave about 99.9% for all chromosomes except in chromosome 4 gave 100%, while chromosome 13 and chromosome 18 gave 99.8%, moreover chromosome Y gave 99.1%. Whereas the LR model gave more than 99.5% for all chromosomes except chromosome 18 and chromosome 21 gave 99.2%, while chromosome X gave 98.6%, and chromosome Y gave 71.3%. Whilst the SVM model gave more than 97.7% for all chromosomes except chromosome 19 gave 96.4%, chromosome X gave 94.7%. All chromosomes can be found in the supplementary file.

The study's present limitations: In order to fairly compare the various algorithms with related studies, it is necessary to use the same input data, the same performance criteria for the same output of the methylation prediction, and the same point of view when searching. SVM is very sensitive to the selection of the kernel parameters, which may be tested for a variety of possible values to determine the correct kernel parameters.

4. Conclusions

Methylation of the DNA is one of the main essential epigenetic modifications in the genome, with profound consequences on the structure and the activity of the DNA molecule. Defining the specific methylated sites in the DNA can only be accomplished through laborious and time-consuming experiments. As a challenge, to implement computational models that can be applied to any type of cells or any genomic region without changing the performance of the model. The highest performance is an aiming with a low cost in a short time. The chromosome 18 of human embryonic stem cell was chosen to extract about 209 features that divided into four groups. Around each CpG locus in the sequence data of WGBS was determined windows of four different sizes. A three machine learning models were created to predict the methylated or unmethylated for each CpG locus. In each window, the four features’ groups were extracted to train on chromosome 18 of cell line H1 inside the human embryonic stem cell. Whereas the cell line H9 was used for testing the machine learning models. In the final analysis, an average accuracy about 99.9% was resulted by random forest algorithm inside the window of size 250bp with the features of the first group.

In this study, some features had been extracted as 1mer, 2 mer, NpN ratio from the sequence data of whole-genome bisulfite, and the highest value of the average beta value in 10 kbp and 100 kbp from 450K array data inside a window of size 250 bp around each CpG site. The RF model was implemented to train on the data of cell line H1 of embryonic stem cells. An accuracy of about 99.9% was given as testing the model on the data of cell line H9 of embryonic stem cells. Whereas the logistic regression model gave an accuracy of about 99.3%, while the SVM gave an accuracy of about 98.5% in most of the chromosomes. So, the algorithm can predict the human-specific DNA methylation regions with accepted accuracy due to expressing more features, and thus the prediction performance will be improved and the coverage of detecting methylation of CpG sites will be expanded beyond the limits of trials.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Kumar and N. Chordia, ‘’Role of Bioinformatics in Biotechnology’’. Research & Reviews in BioSciences, 12(1), 116, 2017.

[2] U. S. Mehmood.,r and A. Niaz, ‘’Use of Bioinformatics Tools in Different Spheres of Life Sciences’’. Journal of Data Mining in Genomics & Proteomics. 2014

[3] W. P. Laird,. ‘’Principles and challenges of gengenome-wideA methylation analysis’’. Nature Reviews | Genetics, 11, 2010.

[4] G. Ficz, ‘’New insights into mechanisms that regulate DNA methylation patterning . The Company of Biologists Ltd’’, The Journal of Experimental Biology, 218, 14-20,2015.

[5] W. Zhang, T. Spector, P. Deloukas, J. Bell, and B. E Engelhardt, ‘’Predicting genome-wide DNA methylation using methylation marks, genomic position,and DNA regulatory elements’’. Genome Biology, 16, 14, 2015.

[6] C. Wu, S. Yao, X. Li, C. Chen and X. Hu,. ‘’Genome-Wide Prediction of DNA Methylation Using DNA Composition and Sequence’’, Complexity in Human. Int. J. Mol. Sci, 18, 420. 2017.

[7] Y. Wang, T. Liu, D. Xu, H. Shi, C. Zhang, Y-Y Mo and Z. Wang, ‘’Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks’’. Sci Rep. , 6: 19598,2016.

[8] M. Ehrlich and M. Lacey,. ‘’DNA methylation and differentiation: silencing, upregulation and modulation of gene expression’’. Epigenomics. 5(5): 10.2217/epi.13.43. 2013.

[9] X. Gao, Y. Zhang, L. Philipp ‘’Breitling and Hermann Brenner. Tobacco smoking and methylation of genes related to lung cancer development’’. Oncotarget.; 7(37): 59017–59028. 2016.

[10] K. Raj,. ‘’The Epigenetic Clock and Aging; Book:Epigenetics of Aging and Longevity.Chapter 4’’, Elsevier; Translational Epigenetics Vol 4, Pages 95–118, 2018.

[11] G. de Wert and C. Mummery, ‘’Human embryonic stem cells: research, ethics and policy’’, Human Reproduction, Volume 18, Issue 4, 1 , Pages 672–682. 2003.

[12] M. Bhasin, H. Zhang, EL. Reinherz and PA. Reche, ‘’Prediction of methylated CpGs in DNA sequences using a support vector machine’’. FEBS Lett., 579, 4302–8. 2005.

[13] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer and J. Walter, ‘’CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure’’. PLoS Genet., 2, e26. 2006

[14] S. Fan, K. Huang, R. Ai, M. Wang and W. Wang,’’ Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data’’. Elsevier Inc.YGENO-08802; No. of pages: 6; 4C. 2016.

[15] L. Breiman,. ‘’Random Forests’’ . Machine Learning, 45(1), 5–32. 2001.

[16] V. Vapnik,’’Statistical Learning Theory’’. Wiley, New York. 1998

[17] S.H .Esraa., S.M Mai., ‘’A study of support vector machine algorithm for liver disease diagnosis’’, American Journal of Intelligent Systems, vol.4(1),pp. 9-14, 2014.

[18] C.-Y. Joanne Peng, K. L. Lee and G. M. Ingersoll,. ‘’An Introduction to Logistic Regression Analysis and Reporting’’. The Journal of Educational Research, 96(1), 3-14, 2002.

[19] C. Cortes and V. Vapnik,. ;;Support-Vector Networks’’. Machine Learning, 20, 273-297,1995.

[20] Y. Suzuki, J. Korlach, W. S. Turner, T. Tsukahara, J. Taniguchi, Wei Qu, K. Ichikawa,and J. Yoshimura, ‘’AgIn: measuring the landscape of CpG methylation of individual repetitive elements’’. Bioinformatics, 32(19), 2911–219.,2016.

[21]http://genboree.org/java/bin/project.jsp?projectName=XML%20Submissions%2FEDACC%2FANALYSIS%2FEDACC.5121.

[22] F. C. Grandi, J. M. Rosser, S. J. Newkirk, J. Yin, X. Jiang, Z. Xing, Leanne Whitmore, S. Bashir, Z.n Ivics, Z. Izsvák, P. Ye, Y. Eugene Yu and W. An, ‘’Retrotransposition creates sloping shores: a graded influence of hypomethylated CpG islands on flanking CpG sites’’ . Genome Res; 25(8): 1135–1146, 2015.

[23] S. Fan, C. Li, Rizi Ai, M., G. S. Firestein and W. Wang, ‘’Computationally expanding infiniumHumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis’’, .Bioinformatics,1–6. 2016.

[24] T. M. Davies,. The Book of R: A First Course in Programming and Statistics 1st Edition. No Starch Press San Francisco. 2016.

[25] R. Das, N. Dimitrova, Z. Xuan, R. A. Rollins, F. Haghighi,, J. R. Edwards, J.Ju, Timothy, ‘’Computational prediction of methylation status in human genomic sequences’’. Proc Natl Acad Sci U S A. Jul 11; 103(28): 10713–10716, 2006.

Abbreviation and symbols

CPG	Cytosine Phosphate Guanine

DNMTs	DNA methyltransferases
ESC	Embryonic stem cells.
FN	False negative.
FP	False Positive
GEO	Gene Expression Omnibus
LR	logistic regression
P	probability
RF	Random Forest
SVM	support vector machine
TN	True negative.
TP	True Positive
WGBS	whole-genome bisulfite sequencing