Published in final edited form as:
Biochem Biophys Res Commun. 2008 September 26; 374(3): 559–564. doi:10.1016/j.bbrc.2008.07.077.

Histone Methylation Marks Play Important Roles in Predicting the Methylation Status of CpG Islands

Shicai Fan[a,b], **Michael Q. Zhang**[c,a], and **Xuegong Zhang**[a,*]

[a]MOE Key laboratory of Bioinformatics and Bioinformatics Division, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

[b]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

[c]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11274, USA

Abstract

The methylation status of CpG islands are highly correlated with gene expression. Current methods for computational prediction of DNA methylation only utilize DNA sequence features. In this study, besides 35 DNA sequence features, we added 4 histone methylation marks to predict the methylation status of CpG islands, and improved the accuracy to 89.94%. Also we applied our model to predict the methylation pattern of all the CpG islands in the human genome, and the results are consistent with the previous reports. Our results imply the important roles of histone methylation marks in affecting the methylation status of CpG islands. H3K4me enriched in the methylation-resistant CpG islands could disrupt the contacts between nucleosomes, unravel chromatin and make DNA sequences accessible. And the established open environment may be a prerequisite for or a consequence of the function implementation of zinc finger proteins that could protect CpG islands from DNA methylation.

Keywords

DNA methylation; CpG island; Histone methylation; Epigenetics

Introduction

In vertebrates, DNA methylation occurs at the cytosine residue in the context of CpG dinucleotide by virtue of DNA methyltransferases [1]. DNA methylation and histone modifications are two main categories of epigenetic alterations, which are responsible for potentially stable and heritable changes in gene expression (hence in cellular phenotype) without changes of DNA [2]. These epigenetic alterations play important roles in orchestrating some key biological activities, including differentiation, imprinting and silencing chromosomal domains [3].

[*]Corresponding Author Email: zhangxg@tsinghua.edu.cn Telephone: +86−10−62794919 Fax Number: +86−10−62773552.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

About 70−80% of CpG dinucleotides are methylated in human somatic cells [4]. Unmethylated CpGs tend to reside in regions called CpG islands (CGIs), which are characterized by high CpG density [5]. According to Gardiner-Garden sequence criteria, a CGI is defined as a region ≥ 200bp with a G+C content ≥ 50% and the observed/expected CpG ratio ≥ 0.6 [6]. Over 50% of human genes include CGIs in their promoter regions [7]. In the classical viewpoint, CGIs are typically methylation-resistant [5]. However, a substantial proportion of CGIs have recently been reported to undergo methylation during imprinting, X-chromosome inactivation, and even in normal tissues [8]. The methylation status of CGIs in the promoter regions are highly correlated with the gene expression [1]. Aberrant methylation of promoter CGIs has been reported to be a key factor of some tumorigenesis [9].

Because of the biological implication of CGIs, it becomes more and more interesting to predict the methylation status of CGIs. We had constructed a computational method (MethCGI) to predict the methylation status of CGI fragments (segments of CpG islands chopped into identical lengths) based on DNA sequence features [10]. For the task of predicting the methylation status of whole CGIs, three major methods can be found in the literature [11-13]. [11] and [12] only considered DNA sequence features. In [16], Bock *et al* used predicted epigenetic state and chromatin structure features which are also inferred from DNA sequences. Although they realized the importance of epigenetic features in the prediction, there are mainly two problems in their method. Firstly, the epigenetic states and chromatin structures were inferred from DNA sequences. It is still a matter of debate about to what extent the sequence preferences of histone modifications and higher-order chromatin structure will be. Secondly, they use more than 800 attributes in the classification which makes the classifier complicated. Based on a study of the recent genome-wide high-resolution profiling of histone methylations in the human genome [14], we found 4 histone methylation marks that are highly correlated with the DNA methylation status of CGIs. This supports the previous reports that some histone modification enzymes may physically interact with DNA methylases [15-17]. In this study, we built a Support Vector Machine (SVM) model for classifying the methylation status of CGIs with 35 DNA sequence features and 4 extra features of histone methylation marks. This model was trained on CGI methylation data of the CD4 T cells extracted from the Human Epigenome Project (HEP)[18] and got an accuracy of 89.94% assessed with Leave-One-Out Cross-Validation (LOOCV), which shows a significant improvement over the accuracy (85.01%) achieved with only the DNA sequence features. It illustrates that the histone methylation features play important roles in predicting the methylation status of CGIs. We compared our model with Epigraph [13] (an online server for CGI methylation status prediction) on CGI data from human brain[19], and observed noticeable improvement. We applied the proposed classification model on the human genome and predicted the methylation status of all the CGIs.

DNA methylation and histone modifications form a complex regulatory network that modulates chromatin structure and genome function [20]. But a mechanistic understanding of how histone modifications effect DNA methylation is still lacking. It has been shown that CGIs could be protected from DNA methylation when specific zinc finger proteins bind to their flanking sequences [21,22]. Since histone modifications can regulate TF binding by remodeling the chromatin structure [23], we predict that the extent to which the zinc finger proteins could protect CGIs from methylation must be partly affected by the intensity of methylated lysine 4 in histone H3.

Materials and Methods

Datasets

The DNA methylation dataset is from the HEP [18], which aims to identify, catalogue and interpret genome-wide DNA methylation patterns of all human genes in all major tissues.

Currently 1.9 million CpG methylation values are obtained across chromosomes 6, 20 and 22 from 12 different tissues including human CD4 T cell. We mapped the detected CpG dinucleotides to the human genome, and extracted CGIs (Gardiner-Garden sequence criteria) more than 10% of whose CpGs are with methylation value (value ranges from 0 to 100). The methylation value of CGI is calculated as the mean of detected CpGs. CGIs with methylation value larger than 50 were regarded as M-CGIs, while less than 10 were U-CGIs. We got 367 U-CGIs and 100 M-CGIs from T cell.

The histone methylation dataset was published by Barski *et al* [14]. It provides the first genome-scale high-resolution profiling of 20 histone methylations of human T cells. They detected the number of tags for each nucleosome by direct sequencing analysis of ChIP DNA samples using ChIP-Seq. We mapped these methylation tags to CGIs and treated the number of tags as the modification intensity.

For validation, we applied DNA methylation data from Rollins *et al* [19], which detects the in *vivo* DNA methylation profile of human brain. They digested the sequences with McrBC and another five restriction endonucleases, and identified 4240 methylation-resistant domains and 3518 methylation-prone domains respectively. We extracted 301 U-CGIs and 192 M-CGIs according to Gardiner-Garden *et al*'s definition.

Features used in the classifier

Previous results have indicated that many DNA sequence features are distinguishing between U-CGIs and M-CGIs. In this study, we used 3 types of DNA sequence features: (1) the CGI characteristics: the length, G+C content and CpG ratio; (2) the count of AluY repetitive elements, extracted by RepeatMasker [24]; (3) the count of Transcription Factor Binding Sites (TFBSs), extracted by MATCH [25]. TFBSs used here are the 214 non-redundant vertebrate TFBSs from TRANSFAC 11.2 [26]. We filtered some uninformative TFBSs. The overall variances of the count of these uninformative TFBSs are less than 0.01. 31 TFBSs were left.

To investigate the intensity distribution of the 20 histone methylation marks between U-CGIs and M-CGIs, we counted the number of each modification in the U- and M-CGIs and their 1000bp flanking sequences. In the 1000bp flanking regions, we counted the number of each modification in a 200bp-window sliding with 10bp offset. Inside the CGIs, we normalized the length of all the CGIs to get 200 counts (arbitrarily chosen). In each count, we got the number of each modification in a 200bp-window, and the sliding offset is adjusted according to CGIs' length. The intensity number was normalized to the counts per million tags.

Support Vector Machine

SVM has been widely used in classification problems of many fields of computational biology. Its basic principle is: given a training set of n samples, $\{x_i, y_i\}$, $i = 1,..., n$, where $x_i \in R^d$ are the feature vectors of d dimension and $y_i \in \{+1, -1\}$ are class labels. In this study, $y = +1$ is for U-CGIs and $y = -1$ for M-CGIs. SVM obtains a decision function by minimizing the predictive errors and maximizing the separation margins on training data. We used the linear SVM provided in the LibSVM package [27] to implement the algorithm. The classification performances (SP, SE, ACC and CC) were evaluated by LOOCV (See Supplementary material).

Results

Discriminating DNA sequence features and histone methylation marks between U- and M-CGIs

Recently we and other researchers found that certain DNA sequence features are highly predictive of CGI methylation [10,12,13]. In this study, we only selected 35 DNA sequence features for the discrimination of U-CGIs vs. M-CGIs, including the length, G+C content and the CpG ratio of CGIs, and the count of AluY and 31 TFBSs, after filtering the uninformative TFBSs from the original 214 non-redundant vertebrate TFBSs of TRANSFAC 11.2 [26].

Barski *et al*'s [14] profiling of 20 histone methylations in human T cells is the first and was then the only such genome-scale high-resolution data available. We investigated the intensity distribution of these histone methylation marks between U-CGIs and M-CGIs by counting the number of each modification in the U- and M-CGIs and their flanking sequences. Among the 20 histone methylation marks, H3K4me1, H3K4me2, H3K4me3 and H3K9me1are differentially distributed between U-CGIs and M-CGIs. Therefore, we adopted the counts of these 4 histone methylation marks in CGIs and their flanking regions as the 4 extra histone modification features in our classification model. Figure 1 shows the intensity distribution of H3K4me (H3K4me1, H3K4me2, H3K4me3) and H3K9me1 in U- and M-CGIs, and in their 1000bp flanking regions. One can see that the intensities of H3K4me1 (Figure 1(A)), H3K4me2 (Figure 1(B)) and H3K9me1 (Figure 1(D)) are much higher in the flanking regions of U-CGIs than the flanking regions of M-CGIs, while the plateaus of H3K4me3 (Figure 1(c)) modification is much more pronounced within the U-CGI regions.

Prediction of methylation status of CGIs

We constructed an SVM classifier for the U-CGIs vs. M-CGIs with the 35 informative DNA sequence features and 4 histone methylation intensities. In order to investigate the effect of the flanking sequence length in the prediction, we experimented the SVM classifier with features extracted from flanking sequences of different lengths and found that the best performance in LOOCV is reached when the length was set to 500bp (Figure 2). The AluY count, the count of TFBSs and histone methylation intensities were extracted in CGIs and their 500bp flanking regions in our final classifier. The LOOCV accuracy corresponding to this flanking length is 89.94% , with specificity of 94.28% and sensitivity of 74%.

In order to check the contribution of histone methylation marks in prediction, we compared this result with the result of the same method using only the 35 DNA sequence features. Figure 3 shows the ROC curves of the SVM classifiers with and without the 4 histone methylation features. We can see that the histone methylation features have substantially improved the prediction accuracy.

Performance comparison with other methods

Currently there are three published methods for predicting the methylation status of whole CGIs. Feltus *et al* constructed a model to predict the methylation status of CGIs based on *in vitro* experimental data using SVM [11]. We are unable to get their program or data to do any comparison. Using 918 DNA sequence related features, Bock *et al* applied SVM on a dataset of 132 CGIs on chromosome 21 measured in human peripheral blood lymphocytes [12]. Then adding the predicted epigenetic state and chromatin structure features, they also applied SVM on the same dataset with 847 sequence based features and provided an online server named Epigraph [13] . In order to make a fair performance comparison, we also built the Epigraph model with the same CGIs from T cells and compared our performance with theirs on an independent data. The data are from human brain including 301 U-CGIs and 192 M-CGIs [19]. The predictive results of our model (both with and without histone methylation features) and Epigraph are shown in Table 1. One could see that the 4 extra histone methylation marks could significantly improve the accuracy and correlation coefficient, and using the epigenetic features directly from biological experiment is more reliable than the predicted epigenetic states.

Methylation status profiling of all CGIs on the human genome

We predicted the methylation status of all CGIs on the human genome using our classification model. The CGIs were downloaded from UCSC browser (Hg18). We got 27,639 CGIs after filtering the CGIs located in clones that are not yet finished or cannot be placed with certainty at a specific place on the chromosome. The distributions of the number of CGIs in chromosomes and in promoters, intragenic and intergenic regions are shown in Figure 4. Promoter regions are defined as the regions located between 1kb upstream of Transcription Start Site (TSS) and 200bp downstream of TSS. The predictive results are available at http://bioinfo.au.tsinghua.edu.cn/member/sfan/MethStateCGI.html, one can also access the results via the UCSC browser from that link. Based on this predicted profile, 34.22% of the CGIs are prone to methylation, which is consistent with Yamada *et al*'s observation that almost a third of CGIs undergo DNA methylation[8]. Also we showed the proportion of methylation-prone CGIs in each chromosome (red bar in Fig 4(A)), and in promoters, intragenic and intergenic regions (red bar in Fig 4(B)). Around 60% of CGIs located in chrX and chrY are methylation-prone. Only ~13 % of the CGIs located in promoters are methylated. Such results are consistent with current reports that many genes are repressive in sex chromosomes [28] and CGIs located in promoter regions are seldom methylated [1].

Discussion

Takai and Jones proposed another definition of CGIs as a region ≥ 500bp with a G+C content ≥ 55% and the observed/expected CpG ratio ≥ 0.65 [29]. In order to check whether our conclusions are sensitive to these thresholds, we used the same procedures on CGIs with this definition (See Supplementary material). One could also see the important roles of the 4 histone marks in the accuracy increase.

The tissue specificity of histone modifications

Currently it is unclear to what extent the histone modification profiles differ in various tissues. In the ENCODE project, it is indicated that there are modest to strong correlations between the modification data from 5 cell lines for some modifications, such as H3K4me2 and H3K4me3 [30]. In our analysis, we could get satisfactory predicting results on CGIs from human brain (Rollins *et al*'s data) by using the histone methylation features derived from human T cells, which also suggests that the histone methylation profiles of CGIs in different tissues may be highly correlated.

Relationship between enriched histone methylations and U-CGIs

H3K4me2 has been reported to be elevated in CpG-rich promoters in the human genome [17]. In our data, we found that all H3K4me are enriched in U-CGIs or their flanking regions. H3K4me are positively correlated with gene expression [31]. The co-enrichment of the three forms of histone H3 lysine 4 methylation provides another evidence that some modifications may combine redundantly to ensure robust chromatin activation [32]. The other enriched histone methylation mark-H3K9me1, has been reported to implicate in the transcription repression in some literatures [33]. On the other hand, it was also reported to be associated with transcription activation in Barski *et al*'s genome-wide histone methylation data of human T cells [14]. H3K9me1 may offer a potential mechanism for genes to shift between transcription repression and activation in different environmental or physiological conditions.

Certain zinc finger proteins (such as Sp1 and CTCF) have been reported to protect CGIs from methylation by actively binding to CGIs' flanking sequences [21,22]. Also it is known that histone modifications such as H3K4me could regulate TF binding [23]. Based on these understanding, we propose the hypothesis that H3K4me can recruit some remodeling proteins to modify chromatin structure and provide DNA access, then the zinc finger proteins bind to DNA sequences to block the spreading of DNA methylation and protect CGIs from methylation. The extent to which the zinc finger proteins could protect CGIs from methylation can be partly affected by the intensity of H3K4me. This may antagonize DNMT3L and BHC80/LSD1 that only recognize H3 tails that are unmethylated at lysine 4 [34,35].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (60540420569 to M.Q.Z. and X.Z., 30625012 and 60721003 to X.Z.), the National Basic Research Program of China (2004CB518605 to X.Z. and M.Q.Z.), the Hi-tech Research and Development Program of China (2006AA02Z325 to X.Z.), National Institutes of Health (HG01696 to M.Q.Z.) and the Changjiang Scholar Award of China (to M.Q.Z.).

References

1. Bird AP. CpG islands as gene markers in the vertebrate nucleus. Trends Genet 1987;3:342–347.
2. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. Cell 2007;128:635–638. [PubMed: 17320500]
3. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. Science 2001;293:1089–1093. [PubMed: 11498579]
4. Melanie E, Miguel AG-S, Lan-Hsiang H, Rose Marie M, Kenneth CK, Roy AM, Charles G. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res 1982;10:2709–2721. [PubMed: 7079182]
5. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev 2002;16:6–21. [PubMed: 11782440]
6. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol 1987;196:261–282. [PubMed: 3656447]
7. Larsen F, Gundersen G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. Genomics 1992;13:1095–1107. [PubMed: 1505946]
8. Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, Iwasaka T, Mukai T, Sakaki Y, Ito T. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res 2004;14:247–266. [PubMed: 14762061]
9. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70. [PubMed: 10647931]
10. Fang F, Fan S, Zhang X, Zhang MQ. Predicting methylation status of CpG islands in the human brain. Bioinformatics 2006;22:2204–2209. [PubMed: 16837523]
11. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. Predicting aberrant CpG island methylation. Proc Natl Acad Sci U S A 2003;100:12253–12258. [PubMed: 14519846]
12. Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet 2006;2:e26. [PubMed: 16520826]
13. Bock C, Walter J, Paulsen M, Lengauer T. CpG island mapping by epigenome prediction. PLoS Comput Biol 2007;3:e110. [PubMed: 17559301]
14. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell 2007;129:823–837. [PubMed: 17512414]

15. Roh TY, Cuddapah S, Zhao K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev 2005;19:542–552. [PubMed: 15706033]

16. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 2006;125:315–326. [PubMed: 16630819]

17. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 2007;39:457–466. [PubMed: 17334365]

18. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 2006;38:1378–1385. [PubMed: 17072317]

19. Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH. Large-scale structure of genomic methylation patterns. Genome Res 2006;16:157–163. [PubMed: 16365381]

20. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell 2007;128:669–681. [PubMed: 17320505]

21. Turker MS. The establishment and maintenance of DNA methylation patterns in mouse somatic cells. Semin Cancer Biol 1999;9:329–337. [PubMed: 10547341]

22. Fan S, Fang F, Zhang X, Zhang MQ. Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome. PLoS ONE 2007;2:e1184. [PubMed: 18030324]

23. Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall' Olio V, Zardo G, Nervi C, Bernard L, Amati B. Myc-binding-site recognition in the human genome is determined by chromatin context. Nat Cell Biol 2006;8:764–770. [PubMed: 16767079]

24. Smit, A.; Hubley, R.; Green, P. RepeatMasker Open-3.0, 1996−2004.

25. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 2003;31:3576–3579. [PubMed: 12824369]

26. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 2006;34:D108–110. [PubMed: 16381825]

27. Chang C, Lin C. LIBSVM : a library for support vector machines. 2001

28. Bernardino J, Lombard M, Niveleau A, Dutrillaux B. Common methylation characteristics of sex chromosomes in somatic and germ cells from mouse, lemur and human. Chromosome Res 2000;8:513–525. [PubMed: 11032321]

29. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 2002;99:3740–3745. [PubMed: 11891299]

30. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Weng Z, Birney E, Carter NP, Vetrie D, Dunham I. The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res 2007;17:691–707. [PubMed: 17567990]

31. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet 2004;36:900–905. [PubMed: 15247917]

32. Schreiber SL, Bernstein BE. Signaling network model of chromatin. Cell 2002;111:771–778. [PubMed: 12526804]

33. Huang Y, Greene E, Murray Stewart T, Goodwin AC, Baylin SB, Woster PM, Casero RA Jr. Inhibition of lysine-specific demethylase 1 by polyamine analogues results in reexpression of aberrantly silenced genes. Proc Natl Acad Sci U S A 2007;104:8023–8028. [PubMed: 17463086]

34. Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, Cheng X, Bestor TH. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature 2007;448:714–717. [PubMed: 17687327]

35. Lan F, Collins RE, De Cegli R, Alpatov R, Horton JR, Shi X, Gozani O, Cheng X, Shi Y. Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. Nature 2007;448:718–722. [PubMed: 17687328]

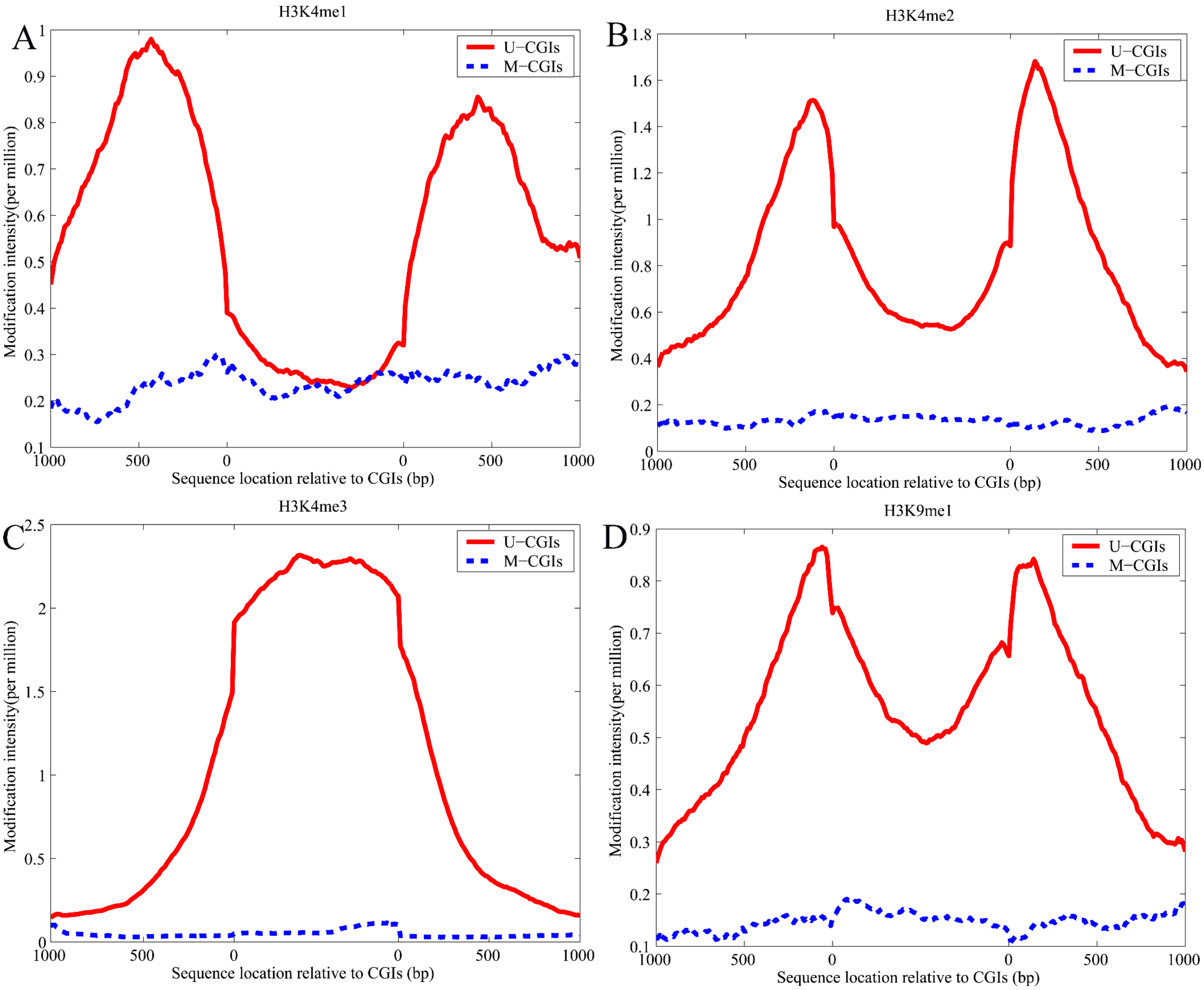

Figure 1. The intensity distribution of four histone marks in U- and M-CGIs, and in their 1000bp flanking regions

On the x-axis, fragments inside the two ‘0’s correspond to the CGIs, and other coordinates indicate the location in the flanking sequences. The y-axis measures the intensity of a specific histone modification. One can see that H3K4me1 (A), H3K4me2 (B), H3K4me3 (C) and H3K9me1 (D) are all differentially distributed between U-CGIs and M-CGIs.

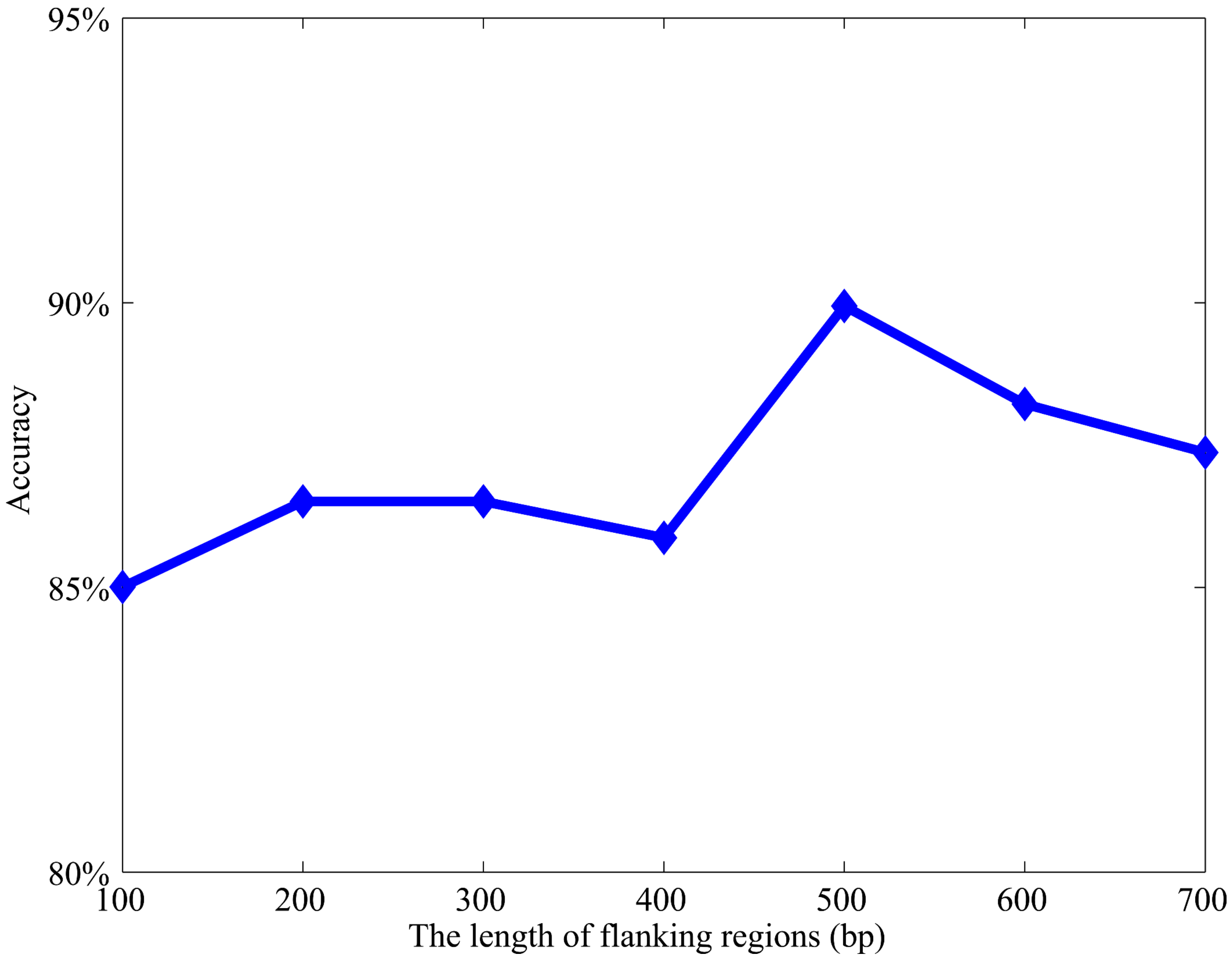

Figure 2. The prediction accuracies with different length of the flanking regions

We extracted the count of AluY, the count of TFBSs and the intensity of histone methylation marks within different flanking regions and found that the best performance is reached when the length of the flanking region is 500bp.

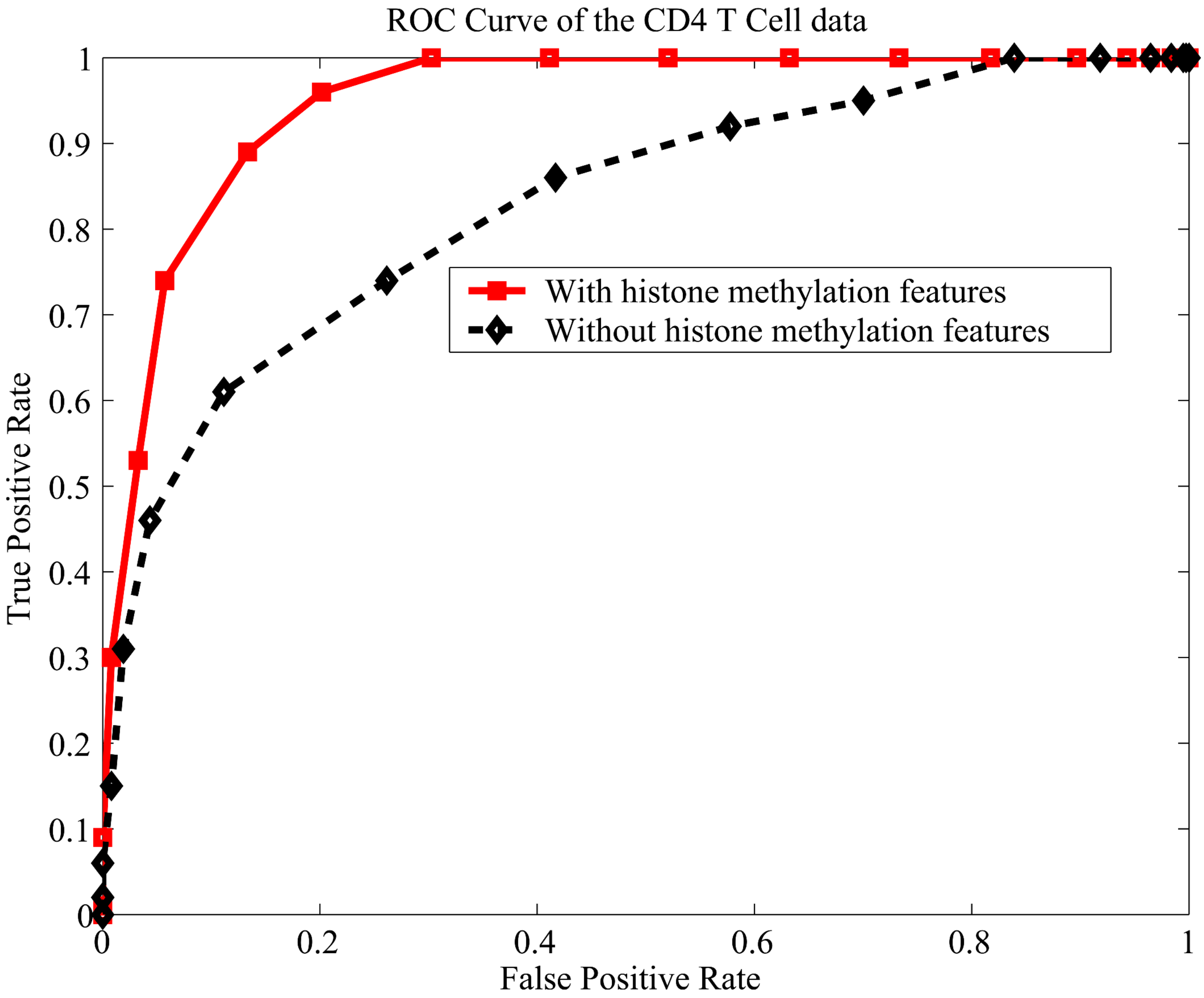

Figure 3. The ROC prediction results with and without histone methylation features

The solid red line: prediction results with both DNA sequence features and the 4 extra histone methylation features; the dashed dark line: prediction results with only DNA sequence features. One can see that the histone methylation features could largely improve the prediction accuracy.

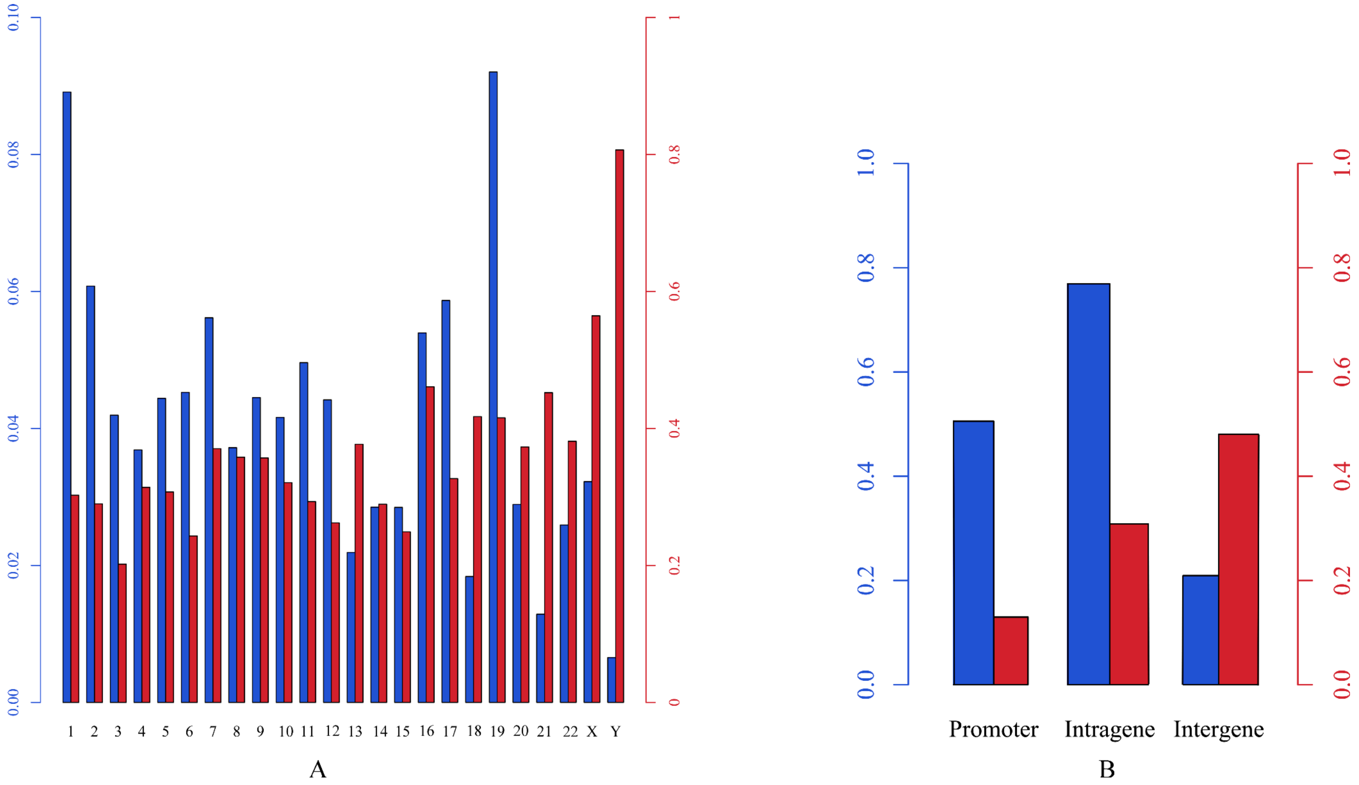

Figure 4. The distribution of CGIs and methylation-prone CGIs

(A) The distribution of the number of CGIs in chromosomes (blue bar) and the proportion of methylation-prone CGIs in each chromosome (red bar). One can see that there are the most CGIs in chr19 and the least CGIs in chrY, and more than 80% of the CGIs located in chrY are prone to DNA methylation, while 33.16% in autosome are methylation-prone. (B) The distribution of the number of CGIs in promoters, intragenic and intergenic regions (blue bar), and the proportion of methylation-prone CGIs located in promoters, intragenic and intergenic regions (red bar). One can see that less than 13% of the CGIs located in promoter regions are prone to DNA methylation.

Table 1.

The predictive results of our model and Epigraph on the human brain data. One could see that our model with the histone marks could get much better results.

	SP (%)	SE (%)	ACC (%)	CC
Our method (with histone marks)	82.39	76.56	80.12	0.59
Our method (without Histone marks)	95.35	21.35	66.53	0.26
Epigraph	94.68	38.54	72.82	0.43