Enhancement on the predictive power of the prediction model for human genomic DNA methylation

Hao Zheng[1], Shi-Wen Jiang[2], and Hongwei Wu[1*]

Abstract— **DNA methylation is an important type of epigenetic modification that plays an instrumental role in organogenesis, cellular differentiation, suppression of deleterious elements, and carcinogenesis. In addition to the experiment-based approaches, computational prediction provides guidance in an effective, fast and cheap way to the genome-wide DNA methylation profiling. In this paper, we describe the development of support vector machine-based models for the prediction of the CpG island methylation. The features used for prediction include those that have been previously demonstrated effective (e.g., CpG island specific attributes, DNA sequence composition patterns, DNA structure patterns, distribution patterns of functional and evolutionarily conserved elements, and histone methylation status) as well as those that have not been extensively explored but are likely to contribute additional information from a biological point of view (e.g., nucleosome positioning propensities, gene functions, and histone acetylation status). Statistical tests were performed to identify the features that are significantly correlated with the methylation status of CpG islands, and principal component analysis was subsequently performed to decorrelate the selected features. The CpG island methylation profile data from the Human Epigenetic Project were used to train, validate and test our predictive models. Specifically, the models were trained and validated by using the data of the CD4 lymphocyte, and were then further tested for generalizability using the data of the other 11 tissues and cell types. The experiments showed that (1) an eight-dimensional feature space that was selected via the principal component analysis and that combines all categories of information was effective for predicting the CpG island methylation status, (2) by incorporating the information regarding the nucleosome positioning, gene functions, and histone acetylation, the model could achieve a higher specificity and accuracy than the existing model while maintaining a comparable sensitivity, (3) the histone modification information contributed significantly to the prediction, without which the performance of the model deteriorated, especially in terms of sensitivity, and, (4) the predictive models generalized well to different tissues and cell types, no matter whether the histone modification information was incorporated or not.**

I. INTRODUCTION

Epigenetics refers to a somatically inheritable pattern of gene expression that is determined by mechanisms other than those encoded in DNA sequences. DNA methylation is an important type of epigenetic modification, implicated in critical cellular functions including genetic imprinting, X-chromosome inactivation, suppression of retroviral elements, and carcinogenesis. DNA methylation involves the addition of a methyl group to DNA via DNA methyltransferase, and typically occurs at the cytosine residues in a CpG dinucleotide context [1]. CpG dinucleotides in human genome are relatively rare but are enriched in short DNA segments known as CpG islands [2]. Most CpG dinucleotides are methylated in human somatic cells [3], but the CpG dinucleotides residing within CpG islands tend to remain unmethylated.

DNA methylation can be determined experimentally via biochemical assays or sequencing. On the other hand, computational modeling can effectively complement the wet chemistry approach in identifying critical factors or pathways controlling DNA methylation patterns, as well as to provide valuable information when methylation data are unavailable for certain genome regions. Computational prediction of DNA methylation has been conducted at two levels – CpG dinucleotides and CpG islands, respectively. At the CpG dinucleotide level, DNA fragments of fixed length with a cytosine in the center were used for the prediction. Each nucleotide was represented by a 5-bit binary sparse code, so that each DNA fragment was represented by a series of codes and the difference between DNA fragments could be quantified. With the optimal DNA fragment length (39 nucleotides), a $\sim$75% of accuracy could be reached for predicting whether a CpG dinucleotide is methylated or not [4]. At the other level, computational models have been developed to distinguish between methylated and unmethylated CpG islands (or DNA fragments with high CpG density). For example, Feltus et al. used DNA sequence patterns to distinguish methylation-prone and methylation-resistant CpG islands under *de novo* methylation, and reached an 82% accuracy [2]. Bock et al. augmented the feature space by including DNA sequence patterns, DNA repeats and predicted DNA structure. Their experiments on the Human Epigenome Project (HEP) data set showed a $\sim$90% accuracy for predicting the methylation status of DNA fragments of high CpG density [5] [6]. The MethCGI used both the DNA sequence composition and transcription factor binding site (TFBS) features to characterize CpG islands and reached an 84% specificity and 84% sensitivity on human brain data [7]. Fan et al. augmented the feature space of the CpG island by including histone methylation information, which is highly correlated with DNA methylation, and reported a 94% sensitivity and 74% specificity on the HEP data [8].

In this study, we considered various attributes that are possibly related to the CpG island methylation. These attributes include those that have been previously investigated (e.g., the CpG island specific attributes, DNA sequence

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA. {hzheng7, hongwei.wu}@gatech.edu

[2]Department of Biomedical Sciences, Mercer University School of Medicine-Savannah Campus, GA, USA. Jiang_s@mercer.edu

[*]corresponding author

patterns, DNA structure patterns, distribution patterns of functional and evolutionarily conserved elements, and the histone methylation status) as well as those that have not been extensively investigated (e.g., nucleosome positioning propensities, gene functions, and histone acetylation status). The contribution of each individual feature was evaluated by statistical tests; and the correlation between features was reduced by principal component analysis (PCA). These DNA methylation-relevant yet non-intercorrelated features were then used to build support vector machine-based classifiers to predict the methylation status of CpG islands. The predictive models were evaluated by using the HEP data set. Specifically, the CpG island methylation profiles in the CD4 lymphocyte were used to train and validate the models, while the CpG island methylation profiles in the other 11 tissues/cell types were used to test the generalizability of the models. Through these experiments, we assessed the individual and combinational influence of the newly added features (nucleosome positioning propensities, functions of nearby genes, and the acetylation status of nearby histones) and the impact of histone modification information.

II. Data Sets

Methylation profiles were obtained from HEP. HEP aims to provide the high-resolution data set regarding genome-wide DNA methylation patterns in human tissues and cell lines [9]. It currently covers chromosomes 6, 20 and 22, and provides 1.9 million CpG methylation values of 2,524 amplicons derived from 12 different tissues and 43 different samples using bisulfite DNA sequencing. The methylation values of the analyzed CpGs range from 0 to 100 inclusive, where 0 corresponds to the lowest and 100 to the highest methylation intensity.

CpG islands can be defined in a number of ways, one of which is based on the Gardiner-Garden criteria: (i) with at least 200 base pairs (bp), (ii) with a GC content>50%, and (iii) with an observed/expected CpG ratio>60% [10]. When applying the Gardiner-Garden criteria on the human genome, we also excluded the repetitive sequence fragments (such as the Alu repeats, which are GC rich and with high CpG observed-to-expected ratio). The methylation intensity of a CpG island was considered as the average methylation intensities of all CpG dinucleotides contained in the island. For statistical reliability, we only considered those CpG islands with more than 10% CpG dinucleotides being measured the methylation intensity levels, and defined *unmethylated* CpG islands as those whose average methylation intensities are less than 10% while *methylated* CpG islands as those whose average methylation intensities are larger than 50% [8].

III. Methods

A. Feature Extraction

It has been shown that the CpG island methylation status is correlated with the following features: CpG island specific attributes (e.g. length, GC content, GC observed/expected ratio) [11] [12] [7], patterns of DNA sequence composition [2] [12] [5], patterns of predicted DNA structure [11] [5], patterns of repetitive elements [11] [12] [7] [5], patterns of TFBS, patterns of evolutionarily conserved elements [11], as well as the methylation status of nearby histones [8]. Computational prediction of CpG island methylation status based on the statistical properties of these features could render fairly reasonable accuracy (e.g., ∼89% [2] [8]). In this study we incorporated three more sets of attributes that have not been extensively explored, including (*i*) the nucleosome positioning propensities of the CpG island, (*ii*) the acetylation status of nearby histones, and (*iii*) the functional roles of nearby genes. These attributes are promising to add more dimensions of information, because an accumulating body of evidence has shown that DNA methylation is influenced by nucleosome positioning [13], associated with histone acetylation [14], and involved in biological processes such as gene imprinting, X chromosome inactivation, and tumor suppressor gene silencing [15] [16]. In the following paragraphs of A.1 to A.6., we describe how these features were extracted.

A.1. The CpG island specific attributes, including the GC content, length and observed/expected CpG ratio, were directly obtained from UCSC human genome browser.

A.2. We considered the DNA composition and structure of each CpG island. For the DNA compositional features, we focused on the frequencies of the tetramer oligonucleotides and their z-scores; and, for the DNA structural features, we focused on those basic characteristics capturing the DNA 3-D conformation as well as the nucleosome positioning propensities.

The z-score of a tetramer oligonucleotide fragment, $N_1N_2N_3N_4$, was calculated as:

$$Z(N_1N_2N_3N_4) = \frac{O(N_1N_2N_3N_4) - E(N_1N_2N_3N_4)}{\sigma(N_1N_2N_3N_4)} \tag{1}$$

where $O(\cdot)$ represents the observed frequency, $E(\cdot)$ and $\sigma(\cdot)$ represent the expected frequency and standard deviation. $E(N_1N_2N_3N_4)$ was estimated empirically based on a maximal-order Markov model [17]:

$$E(N_1N_2N_3N_4) = \frac{O(N_1N_2N_3)O(N_2N_3N_4)}{O(N_2N_3)} \tag{2}$$

and $\sigma(N_1N_2N_3N_4)$ was approximated as:

$$\sigma(N_1N_2N_3N_4) = E(N_1N_2N_3N_4)* \frac{[O(N_2N_3) - O(N_1N_2N_3)][O(N_2N_3) - O(N_2N_3N_4)]}{O^2(N_2N_3)} \tag{3}$$

The DNA conformation related attributes include twist, tilt, roll, shift, slide and rise, which were estimated based on a model of dinucleotide stiffness [18]. For each of these six attributes, the average value over all dinucleotides of the CpG island was used.

Nucleosome positioning propensities of the CpG islands were estimated based on the genome-wide prediction of the nucleosome organization map [19]. There were two types of predictions, one at the nucleotide level, and the other at the DNA fragment level. The nucleotide level prediction regards the probability of each nucleotide being covered by any nucleosome, based on which we calculated the mean and standard deviation over the entire CpG island. The fragment level prediction regards the nucleosome positioning potential of each 147 bp (the typical length of a nucleosome) DNA fragment, based on which we calculated the mean and standard deviation over all fragments overlapping with the CpG island.

A.3. We also considered the distribution patterns of the functional or evolutionarily conserved elements in the chromosomal region flanking the CpG island, where the functional elements refer to the TFBS that are conserved in human, mouse and rat genomes [20], and the evolutionarily conserved elements are those that are conserved across vertebrate, insect, worm and yeast genomes [21]. To account for both the short- and long-range association between these elements and CpG islands, we considered flanking regions of various lengths, ranging from 100 bps to 2,000 bps (with step size of 100 bps) upstream and downstream of the CpG island. Each TFBS or evolutionarily conserved element is characterized by a score quantifying its degree of conservativeness across genomes. We counted the number of these elements overlapping with the CpG island, and calculated their average score.

A.4. We examined whether a CpG island's nearby genes are involved in any cancer-related biological processes. A CpG island's nearby genes refer to those whose promoter region (from the 1,000 bps upstream to the 200 bps downstream of the transcription start site) overlaps with the CpG island. 37 biological processes (30 oncogene related, 11 tumor suppressor related, and 4 common) were determined through gene ontology enrichment analysis of the genes retrieved from the Cancer Gene Census [22]. If the gene ontology annotations of a gene include one or more of these processes, the corresponding gene function feature is 1 and 0 otherwise.

A.5. We considered the methylation and acetylation statuses of each CpG island's nearby histones. The histone methylation information was obtained from Barkski et al's data set, which characterizes the genome wide distribution of 20 histone methylations as well as histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF in CD4 lymphocytes [23]. The histone acetylation information was obtained from Wang et al.'s data set [24], which characterizes the genome-wide patterns of 18 histone acetylations in CD4 lymphocytes. In both data sets, a nucleotide is tagged if its nearby histone undertakes a methylation or acetylation modification; hence, the number of tags at each nucleotide can be interpreted as being proportional to the modification level of nearby histones. We used the average and standard deviation of the number of tags over all nucleotides of a CpG island to represent the methylation (or acetylation) level of the CpG island's nearby histones.

B. Feature Selection through Statistical Tests and Principal Component Analysis

A total number of 841 features were extracted for each CpG island, including three CpG island-specific attributes, 512 DNA compositional features and 10 DNA structural features of the CpG island, 230 about the distribution of TFBS and two about the distribution of the evolutionarily conserved elements in the flanking chromosomal region, two about the involvement of the neighboring genes in oncogene or tumor-suppressor related processes, and 82 about the methylation and acetylation status of nearby histones. The extraction of these features was biologically motivated. However, from a statistical point of view, the correlations of these features to the CpG island methylation status vary from one feature to another. For instance, it was reported that DNA sequence composition patterns, distribution of repeat elements, and DNA structure properties are highly or moderately correlated with the CpG island methylation status; whereas the distribution of genes, single nucleotide polymorphism, and CpG island distribution are only weakly correlated with the CpG island methylation status [5]. To screen out the features of predictive power, we performed various statistical tests, including the Fisher's exact test [25], Chi-squared test [26], and Kolmogorov-Smirnov (KS) test [27]. The Fisher's exact test was used for functional roles of nearby genes, for which the feature variable is categorical and some expected values in the contingency tables are extremely small (<5); the Chi-squared test with Yates corrections [28] was used for the other categorical features (i.e., the number of functional and evolutionarily conserved elements in the flanking chromosomal region); and, the KS test was used for those features whose values are continuous, including CpG island specific attributes, tetramer frequencies and z-scores, DNA structural features, scores of functional and evolutionarily conserved elements, and scores of histone methylation and acetylation. For each of these statistical tests, a feature was considered to be statistically significantly correlated with the methylation status of CpG islands if its p-value was less than 0.05.

Besides their correlations with the CpG island methylation status, these features might be inter-correlated. For example, the histone methylation and acetylation status are likely to be correlated, because some acetylation and methylation (e.g. histone H3 at lysine 9) play opposite roles in gene activity [29]; DNA sequence and structure properties are likely to be correlated, because most DNA structures are predicted based on DNA sequences; and, the distribution of functional/evolutionarily conserved elements in a short flanking neighborhood (e.g., +/- 200 bps) is likely to be correlated with the distribution in a longer flanking neighborhood (e.g., +/- 2000 bps). The correlation between features makes the feature space unnecessarily high-dimensional. To minimize the redundancy in the features, we performed the PCA on those CpG island methylation-related features that were selected via the above statistical tests.

C. Prediction Test

The features selected through statistical tests and PCA were used to build support vector machine-based models to predict the CpG island methylation status. To examine the contribution of the newly added features as well as the impact of the inhibitive-to-acquire histone modification information, we established the following predictive models, (1) M_1: a model with all information being incorporated, (2) M_2: a model with all but the histone modification information being incorporated, (3) M_3–M_9: seven models with individual or combinations of the newly added features being excluded, and (4) M_{10}–M_{16}: seven models with individual or combinations of the newly added features as well as the histone methylation information being excluded. We used the CD4 lymphocyte data for training and validating the models, while the data of the other 11 tissues/cell types for generalizability testing.

Training/Validation (based on the CD4 lymphocyte data): All these models were trained and validated by using a 10-fold cross validation scheme. That is, all CpG islands were partitioned randomly into 10 approximately equally-sized folds, each of which was used in turn for validation while the remaining folds were used for training. The performance of the classifiers was assessed by using three metrics defined in Eqns. (4)–(6), namely, sensitivity (SE), specificity (SP), and accuracy (ACC). This partition-training-and-validation procedure was repeated for 20 times, and the classifier performance was averaged over the 200 validation folds.

$$SP = \frac{\#\text{correctly classified unmethylated CpG islands}}{\#\text{unmethylated CpG islands}} \tag{4}$$

$$SE = \frac{\#\text{correctly classified methylated CpG islands}}{\#\text{methylated CpG islands}} \tag{5}$$

$$ACC = \frac{\#\text{correctly classified CpG islands}}{\#\text{CpG islands}} \tag{6}$$

For fair comparisons with the existing method, a leave-one-out cross-validation (LOOCV) scheme was also used. That is, each CpG island was in turn used for validation while the remaining CpG islands were used for training. The performance of the model in the LOOCV scheme was also assessed by the three metrics averaged over all validation CpG islands.

Generalizability testing (based on data of other tissue/cell types): Two predictive models built on the CD4 lymphocyte data were tested for generalizability using the data of the other 11 tissues and cell types: one (M_1) relying on all information, while the other (M_2) relying on all but the histone modification information. For the former model, because the genome-wide histone methylation and acetylation profiles are not available for these 11 tissues and cell types, we used the genome-wide histone modification profiles in the CD4 lymphocytes, assuming that histone modifications in various cell types are moderately or even highly correlated [41].

IV. Results and Discussions

A. Statistical Tests and PCA

Out of a total number of 841 features, 342 features were retained whose p-values in the statistical tests were less than 0.05. These features include two of the CpG island specific attributes, 217 DNA compositional and eight DNA structural features, 35 functional element features and two evolutionarily conserved element features, two features regarding the functional roles of the neighboring genes, and 76 features related to the modification status of nearby histones. Particularly, among the newly added features, two out of the four nucleosome positioning features, all of the 36 histone acetylation features, and both of the features regarding the functional roles of the neighboring genes were retained after statistical tests.

PCA was performed to decorrelate these 342 selected features. Table I summarizes the number of principal components that must be retained to keep a certain percentage of the variance of the original feature space. Observe that the first eight principal components together can account for the ~99.90% of the variance in the original feature space and were therefore used to build the predictive models. Fig. 1 depicts the contribution of each of the 342 original feature dimensions to the eight principal components. Observe from Fig. 1 that each of the following eight categories of features, (*i*) the CpG island specific attributes, (*ii*) DNA sequence patterns, (*iii*) DNA structure patterns, (*iv*) distribution of TFBS, (*v*) distribution of the evolutionarily conserved elements, (*vi*) gene functions, (*vii*) histone methylation and (*viii*) histone acetylation status, makes substantial contributions to one or more principal components, suggesting that these categories of information, though correlated, are complementary to a certain extent for predicting the CpG island methylation.

TABLE I

Number of principal components (PCs) required to retain a certain percentage (Pcnt) of the total variance.

Pcnt	100%	99.99%	99.90%	99.00%
PCs	342	10	8	6
Pcnt	95.00%	90.00	75.0%	50.00%
PCs	5	4	3	2

B. Performance of the Predictive Models Based on the CD4 Lymphocyte Data

The specificity, sensitivity, and accuracy measures of our predictive model M_1 that incorporates all information are summarized in Table II. Observe that both cross-validation schemes rendered similar results, indicating that these measures can reliably characterize our model. The performance of our classifier was compared to that of Fan et al.'s [8] method. Note that both models incorporated the histone modification information. Observe that our model showed an improved specificity and accuracy than Fan et al.'s model while maintaining a comparable sensitivity. Furthermore, it was reported in [8] that when evaluated on the human brain data, Fan et. al.'s method could outperform Epigraph [6].

We could argue that the improvement of our model M_1 over the existing model was partly due to the incorporation of the three new types of features – nucleosome positioning propensities, gene functions, and histone acetylation status.

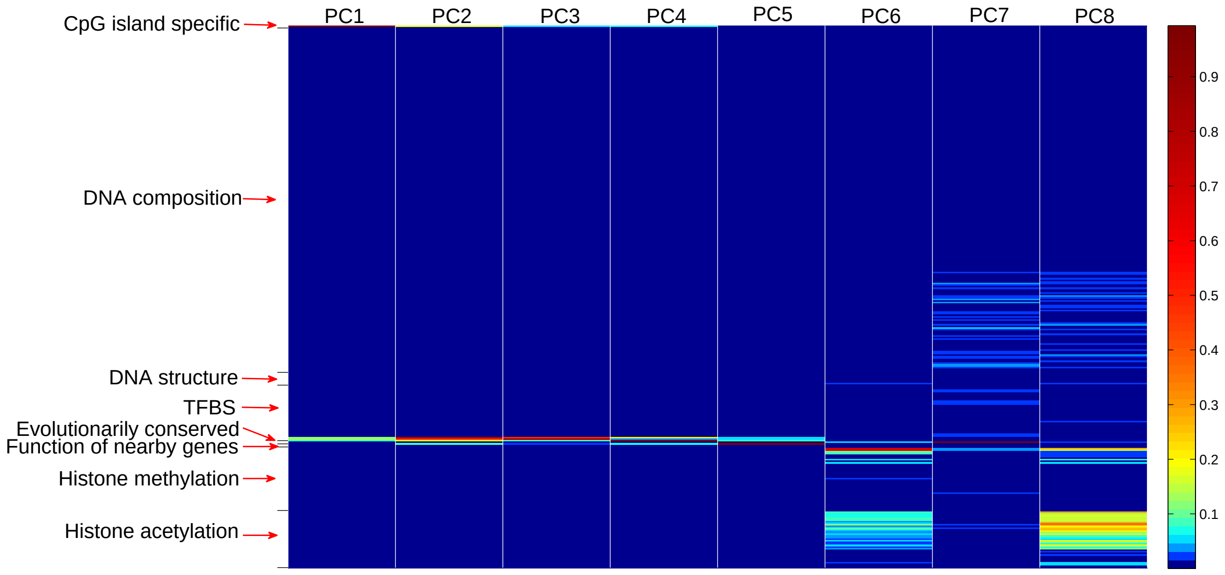

Fig. 1. Contribution of the 342 features to the eight principal components. Each column corresponds to a principal component, and each row corresponds to an original feature dimension.

The performance of our models M_3 through M_9, each with an individual or a combination of the new types of features being excluded, are summarized in Table III. Observe that the performance of the predictive model deteriorated to different extents when individual or combinations of the newly added features were excluded. Specifically, the models without histone acetylation information (M_3, M_6, M_7, and M_9) deteriorated more than those models with histone acetylation information but without the other two types of newly added features (M_4, M_5, and M_8). Therefore, histone acetylation appears to be the most influential feature to the performance of the predictive model among the newly added features.

We suspected that the information carried by the histone methylation features was too dominant to fairly assess the influence of these newly added features; and therefore excluded the histone methylation features and repeated the above experiments excluding individual or combinations of the newly added features. The resultant models were M_{10} through M_{16}, and their performance was summarized in Table III. Similarly, the models without an individual or a combination of the newly added features deteriorated. It is noteworthy that (1) the histone methylation and acetylation information greatly affected the sensitivity of the models, and (2) the loss of histone methylation information could largely be made up by including the histone acetylation information. This is not surprising, given that these two forms of histone modifications are closely related as repeatedly observed in various tissues and cell types [29].

C. Classifier Generalizability

The two predictive models, one with the histone modification information (M_1) and the other without (M_2), that were both built on the human CD4 lymphocyte data were tested on the data of the other 11 tissue and cell types for their generalizability. The sensitivity, specificity, and accuracy of M_1 and M_2 during these testing experiments are summarized in Tables IV and V.

TABLE II
PERFORMANCE OF OUR CLASSIFIERS M_1 ON CD4 LYMPHOCYTES WITH COMPARISON TO THE EXISTING METHOD.

Method	SP	SE	ACC
M_1 (10-fold)	0.9405	0.9257	0.9313
M_1 (LOOCV)	0.9429	0.9307	0.9403
Fan et al.'s [8]	0.7400	0.9428	0.8994

When the histone modification information was incorporated, the classifier model built on the CD4 lymphocyte data can be applied to most of the other tissues and cell types (except for sperm) with little or no performance deterioration. When the histone modification information was not used, the performance of the predictive model on the data of the other tissues and cell types deteriorated substantially, especially in terms of the sensitivity. However, if compared to the validation results where the histone modification information was not used, the performance on the testing data was not unexpected. Therefore, with or without the histone modification information, the predictive model established on the CD4 lymphocyte data can well generalize to the other tissue or cell type data.

Considering that DNA methylation is heavily involved in cellular differentiation, our results in Tables IV and V look suspicious. We therefore calculated the correlations of the CpG island methylation levels between different tissue and cell types, as depicted in Fig. 2. Observe that the correlation coefficients between the somatic/placenta cells are very high (mean: 0.9455, standard deviation: 0.0229), where the correlation coefficients between the somatic/placenta and sperm

TABLE III

PERFORMANCE OF THE PREDICTIVE MODELS (M_3 THROUGH M_{16}), EACH WITH AN INDIVIDUAL OR A COMBINATION OF THE NEWLY ADDED CATEGORIES OF FEATURES BEING EXCLUDED.

	Features	SP		SE		ACC	
		LOOCV	10-fold	LOOCV	10-fold	LOOCV	10-fold
Histone Methylation Retained	**All retained**	0.9429	0.9405	0.9307	0.9257	0.9403	0.9313
	Acetylation (M_3)	0.9048	0.9012	0.9010	0.8965	0.9175	0.9046
	Functional roles (M_4)	0.9319	0.9302	0.9315	0.9265	0.9362	0.9210
	Nucleosome (M_5)	0.9285	0.9270	0.9276	0.9250	0.9205	0.9205
	Acetylation + Functional roles (M_6)	0.8876	0.8791	0.8912	0.8903	0.8915	0.8897
	Acetylation + Nucleosome (M_7)	0.8805	0.8698	0.8815	0.8835	0.8902	0.8826
	Functional roles + Nucleosome (M_8)	0.9208	0.9186	0.9107	0.9116	0.9202	0.9186
	All three (M_9)	0.8775	0.8685	0.8810	0.8822	0.8806	0.8786
Histone Methylation Excluded	**All but histone methylation**	0.9321	0.9318	0.5941	0.5932	0.8593	0.8575
	Acetylation (M_{10})	0.9701	0.9670	0.2277	0.2247	0.8102	0.8001
	Functional roles (M_{11})	0.9109	0.9092	0.5720	0.5670	0.8369	0.8312
	Nucleosome (M_{12})	0.9088	0.9078	0.5682	0.5660	0.8298	0.8296
	Acetylation + Functional roles (M_{13})	0.9402	0.9320	0.2289	0.2279	0.7885	0.7862
	Acetylation + Nucleosome (M_{14})	0.9381	0.9266	0.2302	0.2304	0.7752	0.7641
	Functional roles + Nucleosome (M_{15})	0.9012	0.8990	0.5520	0.5519	0.8252	0.8232
	All three (M_{16})	0.9098	0.8972	0.2341	0.2338	0.7406	0.7352

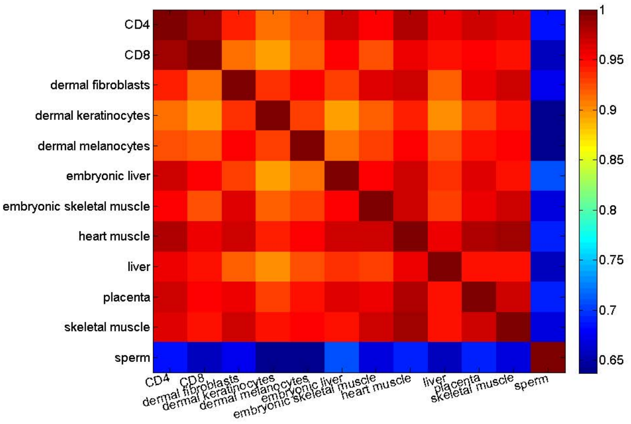

Fig. 2. Correlation coefficients of the CpG island methylation levels across different tissues and cell types.

TABLE IV

PERFORMANCES OF THE CLASSIFIER MODEL BUILT ON THE DATA OF 11 DIFFERENT TISSUES AND CELL TYPES: WITH HISTONE MODIFICATION.

Procedure	Tissue/Cell Type	SP	SE	ACC
Validation	CD4 (10-fold)	0.9405	0.9257	0.9313
	CD4 (LOOCV)	0.9429	0.9307	0.9403
Testing	CD8	0.9608	0.8932	0.9448
	liver	0.9680	0.8762	0.9465
	heart muscle	0.9462	0.9479	0.9466
	skeletal muscle	0.9542	0.9451	0.9524
	embryonic skeletal	0.9395	0.9367	0.9389
	embryonic liver	0.9259	0.9342	0.9277
	placenta	0.9695	0.9130	0.9571
	dermal melanocytes	0.9663	0.8785	0.9446
	dermal fibroblasts	0.9525	0.9239	0.9467
	dermal keratinocytes	0.9385	0.9341	0.9376
	sperm	0.8459	0.9778	0.8617

TABLE V

PERFORMANCES OF THE CLASSIFIER MODEL ON THE DATA OF 11 DIFFERENT TISSUES AND CELL TYPES: WITHOUT HISTONE MODIFICATION.

Procedure	Tissue/Cell Type	SP	SE	ACC
Validation	CD4 (10-fold)	0.9670	0.2247	0.8001
	CD4 (LOOCV)	0.9701	0.2277	0.8102
Testing	CD8	0.9722	0.2108	0.8104
	liver	0.9678	0.2143	0.8122
	heart muscle	0.9562	0.2386	0.8186
	skeletal muscle	0.9594	0.2364	0.8306
	embryonic skeletal	0.9425	0.2298	0.8100
	embryonic liver	0.9389	0.2306	0.8054
	placenta	0.9655	0.2184	0.8276
	dermal melanocytes	0.9700	0.2186	0.8156
	dermal fibroblasts	0.9605	0.2200	0.8237
	dermal keratinocytes	0.9425	0.2204	0.8095
	sperm	0.8524	0.2365	0.7625

cells are only moderate (mean: 0.6706, standard deviation: 0.0225). This suggests that the methylation status of CpG islands are highly correlated in various somatic/placenta cells, and therefore do not represent tissue-specific differentially methylated regions. Our observations are consistent with recent studies [30] [31] that there are few variance in methylation levels of autosomal CpG island promotersa, and there is only a relatively small fraction of CpG islands with tissue-specific methylation. The difference between the somatic/placenta and sperm cells, as reflected by their moderate cross-correlations and the performance deteriorations of our prediction models being applied to the sperm cell data, suggests that gametes are epigenetically more deviated from somatic cells than somatic cells themselves. This difference is likely related to the meiotic process, the special conditions and gene expression required for gamete production [32].

V. CONCLUSIONS AND FUTURE WORKS

The establishment of DNA methylation pattern is a crucial part of cell differentiation and organ development, suppression of viral genes and deleterious elements, and carcinogenesis. Computational prediction of DNA methylation levels provides an effective, fast and cheap alternative approach for studying the DNA methylation patterns. In this study, we performed the computational prediction of the CpG island methylation by incorporating additional features and effectively selecting and decorrelating the features. We incorporated the information regarding the nucleosome positioning propensity, acetylation status of nearby histones, and the functional roles of nearby genes. These features were first screened through statistical tests and PCA. The most DNA methylation-relevant yet non-intercorrelated features were subsequently used to build computational models to predict the methylation status of CpG islands. Our experiments on the HEP data set demonstrated that (1) an eight-dimensional feature space, which combines all the eight categories of information, was effective in predicting the methylation status of CpG islands; (2) by incorporating the information regarding the nucleosome positioning propensities, gene functions, and histone acetylation, our predictive model achieved a higher specificity and accuracy than the existing model while maintaining a comparable sensitivity; (3) the histone modification attributes carry a weight of information for the prediction, without which the performance of the predictive model deteriorated substantially in terms of sensitivity; (4) with or without the histone modification information the performance of the predictive models are consistent on the validation and testing data. This computational model, with its evidently high specificity and sensitivity, provides an effective tool for identification of new methylation targets and therefore lays foundation for our future endeavors in the regulation mechanisms of DNA methylation.

References

[1] A. Bird, “CpG-rich islands and the function of DNA methylation,” *Nature*, vol. 321, pp. 209–213, 1986.

[2] F. Feltus, E. Lee, J. Costello, C. Plass, and P. Vertino, “Predicting aberrant CpG island methylation,” *Proceedings of the National Academy of Sciences USA*, vol. 100, pp. 12 253–12 258, 2003.

[3] M. Ehrlich, M. Gama-Sosa, L. Huang, R. Midgett, K. Kuo, R. McCune, and C. Gehrke, “Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells,” *Nucleic Acids Research*, vol. 10, pp. 2709–2721, 1982.

[4] M. Bhasin, H. Zhang, E. Reinherz, and P. Reche, “Prediction of methylated CpGs in DNA sequences using a support vector machine,” *FEBS Lett*, vol. 579, pp. 4302–8, 2005.

[5] 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 Genetics*, vol. 2, p. e26, 2006.

[6] C. Bock, J. Walter, M. Paulsen, and T. Lengauer, “CpG island mapping by epigenome prediction,” *PLoS Computational Biology*, vol. 3, p. e110, 2007.

[7] F. Fang, S. Fan, X. Zhang, and M. Zhang, “Predicting methylation status of CpG islands in the human brain,” *Bioinformatics*, vol. 22, pp. 2204–2209, 2006.

[8] S. Fan, M. Zhang, and X. Zhang, “Histone methylation marks play important roles in predicting the methylation status of CpG islands,” *Biochemical and Biophysical Research Communications*, vol. 374, pp. 559–564, 2008.

[9] F. Eckhardt, J. Lewin, R. Cortese, V. Rakyan, J. Attwood, M. Burger, J. Burton, T. Cox, R. Davies, T. Down, C. Haefliger, R. Horton, K. Howe, D. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck, “DNA methylation profiling of human chromosomes 6, 20 and 22,” *Nature Genetics*, vol. 38, pp. 1378–1385, 2006.

[10] M. Gardiner-Garden and M. Frommer, “CpG islands in vertebrate genomes,” *Journal of molecular biology*, vol. 196, pp. 261–282, 1987.

[11] C. Previti, O. Harari, I. Zwir, and C. del Val, “Profile analysis and prediction of tissue-specific CpG island methylation classes,” *BMC Bioinformatics*, vol. 10, p. 116, 2009.

[12] R. Das, N. Dimitrova, Z. Xuan, R. Rollins, F. Haghighi, J. Edwards, J. Ju, T. Bestor, and M. Zhang, “Computational prediction of methylation status in human genomic sequences,” *Proc Natl Acad Sci USA*, vol. 22, pp. 10 713–10 716, 2006.

[13] R. Chodavarapu, S. Feng, Y. Bernatavichute, P. Chen, H. Stroud, Y. Yu, J. Hetzel, F. Kuo, J. Kim, S. Cokus, D. Casero, M. Bernal, P. Huijser, A. Clark, U. Kramer, S. Merchant, X. Zhang, S. Jacobsen, and M. Pellegrini, “Relationship between nucleosome positioning and DNA methylation,” *Nature Letter*, vol. 466, pp. 388–392, 2010.

[14] J. Dobosy and E. Selker, “Emerging connections between DNA methylation and histone acetylation,” *Cell Mol Life Sci*, vol. 58, pp. 721–727, 2001.

[15] Y. Yamada, H. Watanabe, F. Miura, H. Soejima, M. Uchiyama, T. Iwasaka, T. Mukai, Y. Sakaki, and T. Ito, “A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q,” *Genome Research*, vol. 14, pp. 247–266, 2004.

[16] D. Hanahan and R. Weinberg, “The hallmarks of cancer,” *Cell*, vol. 100, pp. 57–70, 2000.

[17] S. Schbath, B. Prum, and E. Turckheim, “Exceptional motifs in different markov chain models for a statistical analysis of DNA sequences,” *Journal of Computational Biology*, vol. 2, pp. 417–437, 1995.

[18] J. Goñi, A. Pérez, D. Torrents, and M. Orozco, “Determining promoter location based on DNA structure first-principles calculations,” *Genome Biology*, vol. 8, p. R263, 2007.

[19] N. Kaplan, I. Moore, Y. Fondufe-Mittendorf, A. Gossett, D. Tillo, Y. Field, E. LeProust, T. Hughes, J. Lieb, J. Widom, and E. Segal, “The DNA-encoded nucleosome organization of a eukaryotic genome,” *Nature Letter*, vol. 458, pp. 362–366, 2009.

[20] D. Karolchik, R. Baertsch, M. Diekhans, T. Furey, A. Hinrichs, Y. Lu, K. Roskin, M. Schwartz, C. Sugnet, D. Thomas, R. Weber, D. Haussler, and W. Kent, “The UCSC genome browser database,” *Nucleic Acids Res*, vol. 31, pp. 51–54, 2003.

[21] A. Siepel, G. Bejerano, J. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. Hillier, S. Richards, G. Weinstock, R. Wilson, R. Gibbs, W. Kent, W. Miller, and D. Haussler, “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes,” *Genome Research*, vol. 15, pp. 1034–1050, 2005.

[22] P. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. Stratton, “A census of human cancer genes,” *Nature Reviews Cancer*, vol. 4, pp. 177–183, 2004.

[23] A. Barski, S. Cuddapah, K. Cui, T. Roh, D. Schones, Z. Wang, G. Wei, I. Chepelev, and Z. K., “High-resolution profiling of histone methylations in the human genome,” *Cell*, vol. 129, pp. 823–837, 2007.

[24] Z. Wang, C. Zang, J. Rosenfeld, D. Schones, A. Barski, S. Cuddapah, K. Cui, T. Roh, W. Peng, M. Zhang, and K. Zhao, “Combinatorial patterns of histone acetylations and methylations in the human genome,” *Nature Genetics Letter*, vol. 40, pp. 879–903, 2008.

[25] A. Agresti, “A survey of exact inference for contingency tables,” *Proceedings of the National Academy of Sciences USA*, vol. 7, pp. 131–153, 1992.

[26] N. Turner, “Chi-squared test,” *Journal of Clinical Nursing*, vol. 9, p. 93, 2000.

[27] G. Marsaglia, W. Tsang, and J. Wang, “Evaluating kolmogorov’s distribution,” *Journal of Statistical Software*, vol. 8, pp. 1–4, 2003.

[28] J. Freeman and S. Julious, “The analysis of categorical data,” *Scope*, vol. 16, pp. 18–21, 2007.

[29] K. Zhang, J. Siino, P. Jones, P. Yau, and E. Bradbury, “A mass spectrometric western blot to evaluate the correlations between histone methylation and histone acetylation,” *Proteomics*, vol. 4, pp. 3765–3775, 2004.

[30] M. Weber, I. Hellmann, M. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler, “Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome,” *Nature Genetics*, vol. 39, pp. 457–466, 2007.

[31] R. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. Potash, S. Sabunciyan, and A. Feinberg, “The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores,” *Nature Reviews Cancer*, vol. 41, pp. 178–186, 2009.

[32] P. Nawapen, S. Junpen, H. Dion, D. Michael, C. Bernie, and T. Mongkol, “Different DNA methylation patterns detected by the Amplified Methylation Polymorphism Polymerase Chain Reaction (AMP PCR) technique among various cell types of bulls,” *Acta Veterinaria Scandinavica*, vol. 52, p. 18, 2010.