Genomics 104 (2014) 229–233

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Prediction of CpG island methylation status by integrating DNA physicochemical properties

Pengmian Feng [a,*], Wei Chen [b], Hao Lin [c,*]

[a] School of Public Health, Hebei United University, Tangshan, 063000, China
[b] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China
[c] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE INFO

Article history:
Received 27 May 2014
Accepted 19 August 2014
Available online 27 August 2014

Keywords:
DNA methylation
CpG island
Naïve Bayes
Pseudo trinucleotide composition
DNA physicochemical property

ABSTRACT

As an inheritable epigenetic modification, DNA methylation plays important roles in many biological processes. The non-uniform distribution of DNA methylation across the genome implies that characterizing genome-wide DNA methylation patterns is necessary to better understand the regulatory mechanisms of DNA methylation. Although a series of experimental technologies have been proposed, they are cost-ineffective for DNA methylation status detection. As complements to experimental techniques, computational methods will facilitate the identification of DNA methylation status. In the present study, we proposed a Naïve Bayes model to predict CpG island methylation status. In this model, DNA sequences are formulated by "pseudo trinucleotide composition" into which three DNA physicochemical properties were incorporated. It was observed by the jack-knife test that the overall success rate achieved by the proposed model in predicting the DNA methylation status was 88.22%. This result indicates that the proposed model is a useful tool for DNA methylation status prediction.

1. Introduction

Gene expression is a complicated biological process that can be regulated not only by genetic mechanisms, but also by epigenetic mechanisms. DNA methylation is one of the major epigenetic modifications, which involves the addition of a methyl group (CH_3) to the carbon-5 position in the pyrimidine ring of the cytosines of CpG dinucleotides [1,2]. DNA methylation is essential for normal development and is associated with a number of key biological processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, aging and cancer [3–6].

Since the methylated C residues can spontaneously deaminate to form T residues, the CpG dinucleotides are underrepresented in the human genome [7]. Although most CpGs are methylated in the genome, there are also unmethylated CpGs that group into clusters called CpG islands (CGIs) and account for 1–2% of the mammalian genomes. Most of the CGIs are located in promoter regions; however, they also exist within the bodies of genes and within gene deserts [8]. Recent studies have demonstrated that the methylation status of CGIs in promoter regions differs from methylation status elsewhere [9]. For example, compared with that in promoter regions, the methylation status of CGIs in gene bodies can be extensively methylated or unmethylated [9]. Methylation in the immediate vicinity of the TSS blocks transcription initiation, while that in the gene bodies may stimulate transcription elongation, indicating that the position of the methylation in the transcriptional unit influences its relationship to gene control [9] and may correlate with tissue specific gene expression patterns [10]. The non-uniform distribution of methylated CpG sites across the genome and the important roles of methylation in biological processes imply that characterizing genome-wide DNA methylation patterns is necessary to better understand the regulatory mechanisms of DNA methylation.

In the past decade, various experimental techniques, such as methylation-specific PCR [11], bisulfite sequencing [12] and methylation microarrays [13], have been proposed to detect genome-wide DNA methylation patterns. However, experimental approaches are expensive to perform the genome-wide analysis of DNA methylation. As complements to experimental techniques, a series of computational methods have been developed to speed up genome-wide DNA methylation detections. Based on DNA composition features, Bhasin et al. proposed a method to predict the methylation of single cytosines and obtained an accuracy of ~75% [14]. By employing nearest neighbor algorithm, Lu et al. reported an accuracy of ~77% for predicting CpG methylation status by using the optimal pentamers [15]. Recently, by integrating DNA composition, DNA structure, histone modification marks, and functional annotations of nearby genes, Zheng et al. proposed a support vector machine based model to predict the methylation status of CGIs [16]. All these methods yield quite encouraging results, and each of them did play a role in stimulating the development of this area.

* Corresponding authors. Fax: +86 315 3725715.
E-mail addresses: fengpengmian@gmail.com (P. Feng), greatchen@heuu.edu.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

http://dx.doi.org/10.1016/j.ygeno.2014.08.011
0888-7543/© 2014 Elsevier Inc. All rights reserved.

However, the accuracies are still not satisfactory and the performance of these methods should also be improved further. Therefore, there is an urgent need to develop efficient computational tools for DNA methylation predictions, which will be a great help to identify the key features and pathways that are correlated with DNA methylation.

Keeping this in mind, in the present study, a Naïve Bayes model was proposed to predict the CpG island methylation status. By using the pseudo trinucleotide composition (PseTNC) as the input feature vector of Naïve Bayes, the long-range sequence-order effects and DNA physicochemical properties were integrated together. In the jackknife test, the proposed model obtained an overall accuracy of 88.22% for identifying the CGI methylation status in the benchmark dataset. To demonstrate its effectiveness, the model was also applied to identify methylated CpG islands in different tissues/cell types and yielded encouraging results.

2. Materials and methods

2.1. Dataset

The high resolution DNA methylation profiles were obtained from the Human Epigenome Project (HEP). The methylation status was determined by bisulfite sequencing. The current dataset contains approximately 1.9 million CpG methylation values, obtained from the analysis of 2524 amplicons across human chromosomes 6, 20 and 22 in 43 samples derived from 12 different tissues/cell types [17]. The methylation values of CpG sites range from 0 to 100, where 0 corresponds to the lowest and 100 to the highest methylation intensity, respectively. Since the difference of methylation status among different cells is lower [17], the present study only used the data from the CD4$^+$ T lymphocyte cell to train the models.

According to the Gardiner-Garden criteria [18], i.e., a CpG island (CGI) should be with length of ≥200 bps, GC content ≥ 50%, and observed to expected CpG ratio ≥ 0.6, we extracted the CGIs from the UCSC genome browser. The methylation status of a CGI is determined by the average of all the methylation values of CpG sites within it. If the CGIs with the average methylation values are no less than 50 and more than 10% of whose CpG are methylated, they are regarded as methylated; while the CGIs with the average methylation values less than 10 are regarded as unmethylated.

As elaborated in [19], a dataset containing many redundant samples with high similarity would be lack of statistical representativeness. A predictor, if trained and tested by such a biased benchmark dataset, might yield misleading results with overestimated accuracy [20]. To get rid of redundancy and avoid bias, the CD-HIT software [21] was used with the cutoff threshold set at 80% to remove those DNA fragments with high sequence similarity. Finally, we obtained a benchmark dataset containing 240 methylated CpG islands (mCGI) and 210 unmethylated CpG islands (umCGI), respectively.

2.2. DNA sequence formulation

As indicated in a recent work [22], one of the keys in developing a method for identifying DNA attributes is to formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted. Suppose a DNA sequence with L nucleic acid residues, $R_1R_2R_3 \ldots R_i \ldots R_L$, where R_i is the residue at position i and it can be adenine (A), cytosine (C), guanine (G), or thymine (T). The straightforward method to formulate the DNA sequence is using its nucleic acid composition as given below.

$$\mathbf{D} = [\, f(\mathrm{A}) \quad f(\mathrm{C}) \quad f(\mathrm{G}) \quad f(\mathrm{T}) \,]^{\mathbf{T}} \tag{1}$$

where $f(\mathrm{A})$, $f(\mathrm{C})$, $f(\mathrm{G})$, and $f(\mathrm{T})$ are the normalized occurrence frequencies of adenine, cytosine, guanine, and thymine in the DNA sequence. However, it missed the sequence order information. If using the dinucleotide composition, i.e. $f(\mathrm{AA}), f(\mathrm{AC}), f(\mathrm{AG}), \ldots, f(\mathrm{TT})$ to represent the DNA sequence, although the most contiguous local sequence order information is included, none of the global sequence order information is reflected by the formulation.

To deal with this problem, the pseudo dinucleotide composition (PseDNC) was proposed to represent DNA sequences by incorporating the global sequence order information into the feature vector and has been applied for predicting the recombination spots [22] and nucleosome positioning sequences [23]. Recently, a flexible web-server, called 'pseudo k-tuple nucleotide composition (PseKNC)', was developed to generate pseudo k-tuple nucleotide compositions [24]. According to Eqs. (15)–(16) in [24], the pseudo k-tuple nucleotide composition can be defined as,

$$\mathbf{D} = \left[d_1 \quad d_2 \quad \cdots \quad d_{4^k} \quad d_{4^k+1} \quad \cdots \quad d_{4^k+\lambda} \right]^{\mathbf{T}} \tag{2}$$

where

$$d_u = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{4^k} f_i + w \sum_{j=1}^{\lambda} \theta_j} & \left(1 \le k \le 4^k\right) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w \sum_{j=1}^{\lambda} \theta_j} & \left(4^k < u \le 4^k + \lambda\right) \end{cases}. \tag{3}$$

In Eq. (3), $f_k (k = 1, 2, \cdots, 4^k)$ is the normalized occurrence frequency of the non-overlapping k-tuple nucleotides in the DNA sequence. λ is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence, and w is the weight factor. It is through the λ correlation factors that not only considerable global sequence-order effects can be incorporated but the DNA sequences with extreme difference in length can also be converted into a set of feature vectors with a same dimension [22]. The concrete values for λ and w as well as k will be further discussed in Section 3.1; while the correlation factor θ_j represents the j-tier structural correlation factor between all the j-th most contiguous k-tuple nucleotide $T_i = R_iR_{i+1} \ldots R_{i+k-1}$ and is defined as,

$$\theta_j = \frac{1}{L-j-k+1} \sum_{i=1}^{L-j-k+1} \Theta\left(T_i,\ T_{i+j}\right) \quad (j = 1,\ 2,\ \cdots,\ \lambda;\ \lambda < L). \tag{4}$$

For example, θ_1 is called the first-tier correlation factor that reflects the sequence order correlation between all the most contiguous k-tuple nucleotide along a DNA sequence (Fig. 1a); θ_2, the second-tier correlation factor between all the second most contiguous k-tuple nucleotide (Fig. 1b); θ_3, the third-tier correlation factor between all the third most contiguous k-tuple nucleotide (Fig. 1c); and so forth. The correlation function $\Theta(T_i, T_j)$ is given by

$$\Theta\left(T_i, T_j\right) = \frac{1}{v} \sum_{u=1}^{v} \left[P_u(T_i) - P_u\left(T_j\right) \right]^2 \tag{5}$$

where v is the number of DNA physicochemical properties.

2.3. DNA local structural property parameters

Since DNA methylation is closely correlated with nucleosome positioning [25], the three trinucleotide physicochemical properties, i.e. bendability [26], nucleosome-rigid [27] and nucleosome-positioning [28] that can measure nucleosome-forming abilities were selected for the construction of PseKNC, and their concrete values are given in Table 1. Therefore, in the present work, k is equal to 3 meaning that the pseudo trinucleotide composition was used, and v is also equal to 3 reflecting the number of DNA physicochemical properties considered. $P_u(T_i)$ is the numerical value of the u - th$(u = 1, 2, 3)$ property for the trinucleotide T_i at position i, and $P_u(T_j)$ is the corresponding value for the trinucleotide T_j at position j.

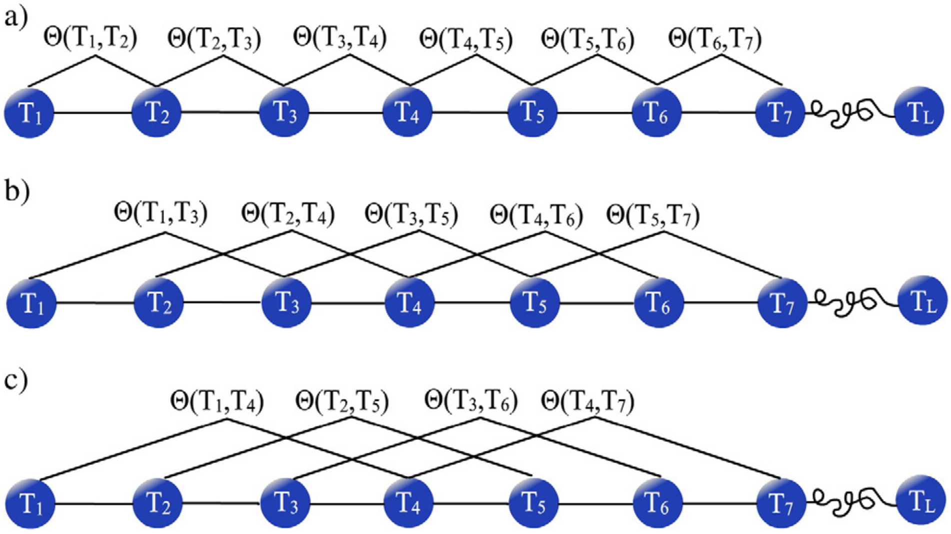

Fig. 1. A schematic illustration showing the correlations of trinucleotides along a DNA sequence. (a) The first-tier correlation reflects the sequence-order mode between all the most contiguous non-overlapping *k*-tuple nucleotide. (b) The second-tier correlation reflects the sequence-order mode between all the second-most contiguous non-overlapping *k*-tuple nucleotide. (c) The third-tier correlation reflects the sequence-order mode between all the third-most contiguous non-overlapping *k*-tuple nucleotide.

Note that before substituting them into Eq. (5), all the original values $P_u(\mathrm{T}_i)(u = 1, 2, 3)$ were subjected to a standard conversion, as described by the following equation,

$$P'_u(\mathrm{T}_i) = \frac{P_u(\mathrm{T}_i) - \langle P_u(\mathrm{T}_i)\rangle}{\mathrm{SD}(P_u(\mathrm{T}_i))} \tag{6}$$

where the symbol < > means taking the average of the quantity therein over the 64 different trinucleotides, and SD means the corresponding standard deviation. The converted values obtained in Eq. (6) will have a zero mean value over the 64 different trinucleotides, and will remain unchanged if going through the same conversion procedure again.

Table 1
Concrete values of the three physicochemical properties for trinucleotides.

Trinucleotide	$P_1(\mathrm{T}_i)$	$P_2(\mathrm{T}_i)$	$P_3(\mathrm{T}_i)$	Trinucleotide	$P_1(\mathrm{T}_i)$	$P_2(\mathrm{T}_i)$	$P_3(\mathrm{T}_i)$
AAA	0.10	7.05	−36.00	GAA	5.10	5.26	−12.00
AAC	1.60	4.86	−6.00	GAC	5.60	3.88	8.00
AAG	4.20	3.99	6.00	GAG	6.60	3.88	8.00
AAT	0.00	6.62	−30.00	GAT	3.60	3.94	7.00
ACA	5.80	3.99	6.00	GCA	7.50	3.54	13.00
ACC	5.20	3.88	8.00	GCC	8.20	1.31	45.00
ACG	5.20	3.88	8.00	GCG	4.30	2.69	25.00
ACT	2.00	3.65	11.00	GCT	6.30	2.69	25.00
AGA	6.50	5.09	−9.00	GGA	6.20	4.80	−5.00
AGC	6.30	2.69	25.00	GGC	8.20	1.31	45.00
AGG	4.70	3.88	8.00	GGG	5.70	3.54	13.00
AGT	2.00	3.65	11.00	GGT	5.20	3.88	8.00
ATA	9.70	5.38	−13.00	GTA	6.40	4.86	−6.00
ATC	3.60	3.94	7.00	GTC	5.60	3.88	8.00
ATG	8.70	3.14	18.00	GTG	6.80	3.25	17.00
ATT	0.00	6.62	−30.00	GTT	1.60	4.86	−6.00
CAA	6.20	5.09	−9.00	TAA	7.30	5.85	−20.00
CAC	6.80	3.25	17.00	TAC	6.40	4.86	−6.00
CAG	9.60	4.57	−2.00	TAG	7.80	5.73	−18.00
CAT	8.70	3.14	18.00	TAT	9.70	5.38	−13.00
CCA	0.70	3.88	8.00	TCA	10.00	3.88	8.00
CCC	5.70	3.54	13.00	TCC	6.20	4.80	−5.00
CCG	3.00	4.28	2.00	TCG	5.80	2.25	31.00
CCT	4.70	3.88	8.00	TCT	6.50	5.09	−9.00
CGA	5.80	2.25	31.00	TGA	10.00	3.88	8.00
CGC	4.30	2.69	25.00	TGC	7.50	3.54	13.00
CGG	3.00	4.28	2.00	TGG	0.70	3.88	8.00
CGT	5.20	3.88	8.00	TGT	5.80	3.99	6.00
CTA	7.80	5.73	−18.00	TTA	7.30	5.85	−20.00
CTC	6.60	3.88	8.00	TTC	5.10	5.26	−12.00
CTG	9.60	4.57	−2.00	TTG	6.20	5.09	−9.00
CTT	4.20	3.99	6.00	TTT	0.10	7.05	−36.00

P_1, bendability [26]; P_2, nucleosome-rigid [27] and P_3, nucleosome-positioning [28].

2.4. Naïve Bayes

Naïve Bayes is an effective statistical classification algorithm and has been successfully used in the realm of bioinformatics [29–31]. Naïve Bayes assumes the attribute variables to be independent from each other given the outcome. This assumption greatly simplifies the calculation of conditional probabilities.

In the Naïve Bayes framework, a classification problem can be seen as the problem of finding the outcome with maximum probability given a set of observed variables. Given an example, described by its feature vector $F = (f_1, f_2, \ldots, f_\mathrm{n})$, we are looking for a class $\mathbf{C}$ that maximizes the likelihood $\mathbf{P}(\boldsymbol{F}|\boldsymbol{C}) = \mathbf{P}(f_1, f_2, \ldots, f_\mathrm{n}\,|\boldsymbol{C})$. Since the current work is intended to classify mCGI and umCGI, a binary class $\mathbf{C} \in \{0,1\}$ was generated, where 1 denotes that the sample was predicted as mCGI and 0 denotes umCGI. For the binary classification, the class for the DNA sample could be determined by comparing two posteriors as

$$\frac{P(C=1|F=f_1,f_2,\ldots,f_\mathrm{n})}{P(C=0|F=f_1,f_2,\ldots,f_\mathrm{n})} = \frac{P(C=1)\prod_{i=1}^{n} P_i(f_i|C=1)}{P(C=0)\prod_{i=1}^{n} P_i(f_i|C=0)}. \tag{7}$$

Taking the logarithm of Eq. (7), we obtain

$$\log\frac{P(C=1|F=f_1,f_2,\ldots,f_\mathrm{n})}{P(C=0|F=f_1,f_2,\ldots,f_\mathrm{n})} = \log\frac{P(C=1)}{P(C=0)} + \sum_{i=1}^{n}\log\frac{P_i(f_i|C=1)}{P_i(f_i|C=0)}. \tag{8}$$

Hence the sample will be predicted as 1 (mCGI) if

$$\log\frac{P(C=1|F=f_1,f_2,\ldots,f_\mathrm{n})}{P(C=0|F=f_1,f_2,\ldots,f_\mathrm{n})} \geq \xi \tag{9}$$

and 0 (umCGI) otherwise. ξ is the threshold determining the trade-off between sensitivity and specificity, and can be trained on the training dataset to maximize the prediction performance.

2.5. Performance evaluation

The performance of the proposed model was evaluated using sensitivity (*Sn*), specificity (*Sp*), Matthew's correlation coefficient (*MCC*) and accuracy (*Acc*), which are expressed as

$$Sn = \frac{TP}{TP + FN} \tag{10}$$

$$Sp = \frac{TN}{TN + FP} \tag{11}$$

$$Mcc = \frac{TP \times TN - FP \times FN}{(TP + FN) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FP})} \tag{12}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}. \tag{13}$$

TP, TN, FP and FN represent the number of the correctly recognized mCGI, the number of the correctly recognized umCGI, the number of umCGI recognized as mCGI and the number of mCGI recognized as umCGI, respectively.

2.6. Cross-validation

Three cross-validation methods, i.e., independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test, are often used to evaluate the anticipated success rate of a predictor. Among the three methods, however, the jackknife test is deemed the least arbitrary and most objective one as demonstrated in Eqs. (28)–(32) of [19], and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors [32–36]. Accordingly, the jackknife test was also used to examine the performance of the model proposed in the current study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified.

3. Results and discussions

3.1. Parameter optimization

Accordingly, by using PseTNC, each sequence in the benchmark dataset was represented by a discrete vector as defined in Eq. (2). The Naïve Bayes was used to perform classification. As we can see from Eqs. (2)–(3), the predictive accuracy of the present model depends on the two parameters *w* and λ. *w* is the weight factor usually within the range from 0 to 1, and λ is the global order effect. Generally speaking, the greater the λ is, the more global sequence-order information the model contains. However, if λ is too large, it would reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy due to over-fitting or "high dimension disaster" problem [37]. Therefore, our search for the optimal values of the two parameters is in the range of $w \in [0, 1]$ and $\lambda \in [1, 10]$ with the steps of 0.1 and 1, respectively.

In order to reduce the computational time, the 5-fold cross-validation approach was used to optimize the two parameters. We found that when $w = 0.3$ and $\lambda = 5$, a peak was observed for the *Acc* (Fig. 2). Accordingly, the two numerical values were used for the two uncertain parameters in the following analysis.

3.2. Prediction quality

The prediction quality measured by the four metrics defined in Eqs. (10)–(13) of the present model in identifying mCGI in the benchmark dataset via the rigorous jackknife test was listed in Table 2. As shown in Table 2, an accuracy of 88.22% was obtained for identifying mCGI in the CD4$^+$ T lymphocyte cell. As a comparison, the corresponding results obtained by using the trinucleotide compositions on the same benchmark dataset are also reported in Table 2. We found that the predictive results obtained by using pseudo trinucleotide composition are higher than that obtained by using trinucleotide composition, indicating that the pseudo trinucleotide composition can provide additional information for CGI methylation status predictions.

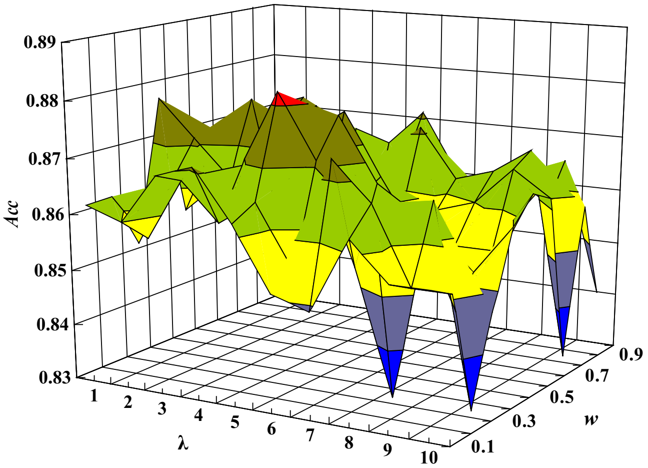

Fig. 2. The 3D graph showing the accuracies obtained in the 5-fold cross-validation with different values of *w* and λ.

3.3. Demonstration on other tissues/cell lines

Furthermore, the model trained on the data from the CD4$^+$ T lymphocyte cell was also applied to identify mCGI in the following human tissues/cell types, namely CD8$^+$ T lymphocyte cell, heart muscle and plancata. According to the Human Epigenome Project [17], we obtained 246, 252 and 258 mCGIs, and 201, 199, and 226 umCGIs in the above tissues/cell types, respectively. The predictive accuracies obtained by the present model are all over 80% for identifying mCGI in these tissues/cell types, which are also higher than those obtained by using trinucleotide composition (Table 2). All these results indicate that the proposed method is a promising approach for identifying mCGI, or at least can play a complementary role to the existing method in this area.

4. Conclusions

Genome-wide studies of the methylome have demonstrated that DNA methylation is associated with a number of key biological processes and plays important roles in cell development and differentiation. Although our understanding of DNA methylation is still in its infancy, it is clear that the methylation status of a CGI is highly correlated with transcription. Therefore, the determination of methylation status of CGIs will facilitate our understanding of the regulatory roles of DNA methylation in transcription.

In the present study, we proposed a Naïve Bayes based model for predicting methylation status of CGIs by using pseudo trinucleotide compositions and found that the jackknife predictive results of pseudo trinucleotide composition are better than those of trinucleotide composition. This is due to the addition of the DNA physicochemical properties of trinucleotides, namely bendability, nucleosome-rigid and nucleosome-positioning. In mammalian cells, DNA methylation is catalyzed by DNA methyl-transferases (DNMTs) that need to recognize and bind with specific genomic regions [2]. As an intrinsic property of DNA sequences, bendability contains the information depicting DNA structures [26], which plays important roles in DNA–protein interactions. The other two properties (nucleosome-rigid and nucleosome-positioning) may

Table 2
Comparison of different methods for identifying mCGI by the jackknife test on different tissues/cell types.

Tissue/cell type	Method	Sn (%)	Sp (%)	Acc (%)	MCC
CD4	Pseudo trinucleotide composition	89.17	87.14	88.22	0.76
	Trinucleotide composition	81.25	80.95	81.11	0.62
CD8	Pseudo trinucleotide composition	81.38	82.09	81.70	0.63
	Trinucleotide composition	76.42	84.08	79.87	0.60
Plancata	Pseudo trinucleotide composition	79.05	83.00	80.79	0.62
	Trinucleotide composition	76.19	85.93	80.49	0.62
Heart muscle	Pseudo trinucleotide composition	80.38	80.84	80.59	0.61
	Trinucleotide composition	75.19	83.19	78.93	0.58

provide clues about the interaction between DNA methylation and nucleosome positioning that is also a key factor for DNA methylation regulations.

Although the model is trained based on the data from the $CD4^+$ T cell, it is encouraging to see that the predictive results of the model for predicting methylation status of CGIs from other tissues/cell types are also quite good, which agrees with the fact that the difference of methylation status among different tissues/cell types is lower [17]. These results also demonstrated that our model is helpful for DNA methylation detections.

As an epigenetic modification, DNA methylation is a complicate progress. Besides sequence context and DNA physicochemical properties, it is also affected by other factors, such as transcription factors, histone methylation modifications and histone acetylation modifications. For better understanding of the biological function of DNA methylation, in the future work, we will combine all these factors and develop new models to identify the methylation status of CGI.

Acknowledgments

We wish to express our gratitude to the editor and anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this article. This work was supported by the National Nature Scientific Foundation of China (Nos. 61100092 and 61202256), the Nature Scientific Foundation of Hibei Province (No. C2013209105) and the Science and Technology Department of Hebei Province (No. 132777133).

References

[1] A.P. Bird, CpG-rich islands and the function of DNA methylation, Nature 321 (1986) 209–213.
[2] R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, Nat. Genet. 33 (2003) 245–254 (Suppl.).
[3] M.J. Barrero, S. Boue, J.C. Izpisua Belmonte, Epigenetic mechanisms that regulate cell identity, Cell Stem Cell 7 (2010) 565–570.
[4] M.I. Scarano, M. Strazzullo, M.R. Matarazzo, M. D'Esposito, DNA methylation 40 years later: its role in human health and disease, J. Cell. Physiol. 204 (2005) 21–35.
[5] Y. Tao, S. Xi, J. Shan, A. Maunakea, A. Che, V. Briones, E.Y. Lee, T. Geiman, J. Huang, R. Stephens, R.M. Leighty, K. Zhao, K. Muegge, Lsh, chromatin remodeling family member, modulates genome-wide cytosine methylation patterns at nonrepeat sequences, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 5626–5631.
[6] P.M. Das, R. Singal, DNA methylation and cancer, J. Clin. Oncol. 22 (2004) 4632–4642.
[7] F. Lienert, C. Wirbelauer, I. Som, A. Dean, F. Mohn, D. Schubeler, Identification of genetic elements that autonomously determine DNA methylation states, Nat. Genet. 43 (2011) 1091–1097.
[8] P.A. Jones, The DNA methylation paradox, Trends Genet. 15 (1999) 34–37.
[9] P.A. Jones, Functions of DNA methylation: islands, start sites, gene bodies and beyond, Nat. Rev. Genet. 13 (2012) 484–492.
[10] F. Song, J.F. Smith, M.T. Kimura, A.D. Morrow, T. Matsuyama, H. Nagase, W.A. Held, Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 3336–3341.
[11] J.G. Herman, J.R. Graff, S. Myohanen, B.D. Nelkin, S.B. Baylin, Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 9821–9826.
[12] S.J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C.D. Haudenschild, S. Pradhan, S.F. Nelson, M. Pellegrini, S.E. Jacobsen, Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning, Nature 452 (2008) 215–219.
[13] M. Weber, J.J. Davies, D. Wittig, E.J. Oakeley, M. Haase, W.L. Lam, D. Schubeler, Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells, Nat. Genet. 37 (2005) 853–862.
[14] M. Bhasin, H. Zhang, E.L. Reinherz, P.A. Reche, Prediction of methylated CpGs in DNA sequences using a support vector machine, FEBS Lett. 579 (2005) 4302–4308.
[15] L.Y. Lu, K. Lin, Z.L. Qian, H.P. Li, Y.D. Cai, Y.X. Li, Predicting DNA methylation status using word composition, J. Biomed. Sci. Eng. 3 (2010).
[16] H. Zheng, H. Wu, J. Li, S.W. Jiang, CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome, BMC Med. Genomics 6 (Suppl. 1) (2013) S13.
[17] F. Eckhardt, J. Lewin, R. Cortese, V.K. Rakyan, J. Attwood, M. Burger, J. Burton, T.V. Cox, R. Davies, T.A. Down, C. Haefliger, R. Horton, K. Howe, D.K. 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, S. Beck, DNA methylation profiling of human chromosomes 6, 20 and 22, Nat. Genet. 38 (2006) 1378–1385.
[18] M. Gardiner-Garden, M. Frommer, CpG islands in vertebrate genomes, J. Mol. Biol. 196 (1987) 261–282.
[19] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.
[20] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers, J. Proteome Res. 5 (2006) 1888–1897.
[21] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.
[22] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[23] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, K.C. Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[24] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, K.C. Chou, PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 50–63.
[25] P. Minary, M. Levitt, Training-free atomistic prediction of nucleosome occupancy, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 6293–6298.
[26] M.G. Munteanu, K. Vlahovicek, S. Parthasarathy, I. Simon, S. Pongor, Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena, Trends Biochem. Sci. 23 (1998) 341–347.
[27] D.S. Goodsell, R.E. Dickerson, Bending and curvature calculations in B-DNA, Nucleic Acids Res. 22 (1994) 5497–5503.
[28] S.C. Satchwell, H.R. Drew, A.A. Travers, Sequence periodicities in chicken nucleosome core DNA, J. Mol. Biol. 191 (1986) 659–675.
[29] M. Yousef, S. Jung, A.V. Kossenkov, L.C. Showe, M.K. Showe, Naive Bayes for microRNA target predictions—machine learning for microRNA targets, Bioinformatics 23 (2007) 2987–2992.
[30] P.M. Feng, H. Lin, W. Chen, Identification of antioxidants from sequence information using naive Bayes, Comput. Math. Methods Med. 2013 (2013) 567529.
[31] P.M. Feng, H. Ding, W. Chen, H. Lin, Naive Bayes classifier with feature selection to identify phage virion proteins, Comput. Math. Methods Med. 2013 (2013) 530696.
[32] H. Lin, W. Chen, H. Ding, AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes, PLoS One 8 (2013) e75726.
[33] H. Lin, W. Chen, L.F. Yuan, Z.Q. Li, H. Ding, Using over-represented tetrapeptides to predict protein submitochondria locations, Acta Biotheor. 61 (2013) 259–268.
[34] W. Chen, H. Lin, Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information, Biochem. Biophys. Res. Commun. 401 (2010) 382–384.
[35] W. Chen, P. Feng, H. Lin, Prediction of replication origins by calculating DNA structural properties, FEBS Lett. 586 (2012) 934–938.
[36] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 281 (2011) 18–23.
[37] T. Wang, J. Yang, H.B. Shen, K.C. Chou, Predicting membrane protein types by the LLDA algorithm, Protein Pept. Lett. 15 (2008) 915–921.