Membrane Technology

ISSN (online): 1873-4049

A Data-Driven Modular Framework for Predicting Single-Cell DNA Methylation Landscapes

Prasanth Tirumalasetty

Computer Science/ Data Science

Business Analyst III @ Medical Device Company

OBJECTIVE

DNA methylation can now be measured at single cell resolution due to recent technological advancements. Current techniques, however, are hampered by insufficient CpG coverage, therefore methods to forecast missing methylation states are essential for genome – wide analysis. In this project, we use DeepCpG, a deep learning – based computational technique to predict methylation status. Here, we evaluate DeepCpG on single – cell methylation data from five cell types generated using alternative sequencing protocols. Compared to other methods DeepCpG yields more accurate results. Furthermore, we demonstrate that the model parameters can be understood, by revealing how methylation variability is influenced by sequence composition.

1. Introduction

DNA methylation is an epigenetic mechanism that occurs by the addition of a methyl group to DNA, therefore often modifying the function of the genes and affecting gene expression. It is implicated in wide range of biological processes, including chromosome instability, X chromosome inactivation, cell differentiation, gene regulation and cancer progression. Either by using genome wide bisulfite sequencing or reduced representation protocols we can predict DNA methylation at single cell resolution. Here, the first step is to predict missing methylation states to enable genome – wide analyses. The method we used here is DeepCpG to predict single-cell methylation states and for modelling the sources of DNA methylation variability. It is a computational method based on deep neural networks, it leverages association between DNA sequence patterns and methylation states as well as between neighboring CpG sites, within across and individual cells. This approach does not separate model training and extraction of informative features.

DeepCpG is based on a modular architecture and learns methylation patterns, DNA sequence in data – driven manner. Across, all cell types DeepCpG yielded more accurate predictions of methylation states.

ALGORITHMS

DeepCpG model: DeepCpG consists of a DNA module to extract features from DNA sequence, CpG module to extract features from CpG neighborhood of all cells and Joint module that takes output from both DNA and CpG modules to predict methylation states of target CpG sites for multiple cells.

DNA module: There are multiple convolutional pooling layers and one fully connected hidden layer in this convolutional neural network. CNN is a form of neural network model that uses a

series of convolutional and pooling techniques to extract key features from high-dimensional inputs while keeping the number of model parameters manageable. Unless otherwise specified, the DNA module accepts a 1001-bp DNA sequence centered on a target CpG site n as input, which is encoded as a binary matrix sn by one-hot encoding the D=4 nucleotides as binary vectors A = [1, 0, 0, 0], T = [0, 1, 0, 0], G = [0, 0, 1, 0], and C = [0, 0, 0, 1]. sn is the input matrix. is first transformed by a 1d-convolutional layer, which computes the activations anfi of multiple convolutional filters f at every position i:

$$a_{nfi} = \mathrm{ReLU}\left(\sum_{l=1}^{L}\sum_{d=1}^{D} w_{fld}\, s_{n,i+l,d}\right).$$

Here, wf are the parameters or weights of convolutional filter f of length L. These can be interpreted similarly to

position weight matrices, which are matched against the input sequence sn at every position i to recognize distinct motifs. The ReLU(x) = max (0, x) activation function sets negative values to zero, such that anfi corresponds to the evidence that the motif represented by wf occurs at position i. A pooling layer is used to summaries the activations of P adjacent neurons by their maximum value:

$$p_{nfi} = \max_{|k|<P/2}\left(a_{nf,i+k}\right).$$

Non-overlapping pooling is applied with step size P to decrease the dimension of the input sequence and hence the number of model parameters. The DNA module has multiple pairs of convolutional-pooling layers to learn higher-level interactions between sequence motifs, which are followed by one final fully connected layer with a ReLU activation function. The number of convolutional pooling layers was optimized on the validation set.

CpG module: The CpG module consists of a non-linear embedding layer to model dependencies between CpG sites within cells, which is followed by a bidirectional gated recurrent network (GRU) [36] to model dependencies between cells. Inputs are 100d vectors x1, …, xT, where xt represents the methylation state and distance of K = 25 CpG sites to the left and to the right of a target CpG site in cell t. Distances were transformed to relative ranges by dividing by the

$$\bar{x}_t = \mathrm{ReLU}(W_{\bar{x}} \cdot x_t + b_{\bar{x}}).$$

maximum genome-wide distance. The embedding layer is fully connected and transforms xt into a 256d vector xt, which allows learning possible interactions between methylation states and distances within cell t:

The sequence of vectors xt is then fed into a bidirectional GRU [36], which is a variant of recurrent neural network (RNN). RNNs have been successfully used for modelling long range dependencies in natural language [58, 59], acoustic signals [60] and, more recently, genomic sequences [61, 62]. A GRU scans input sequence vectors x1; …; xT from left to right and encodes them into fixed-size hidden state vectors h1, …, hT.

$$r_t = \mathrm{sigmoid}(W_{r\bar{x}} \cdot \bar{x}_t + W_{rh} \cdot h_{t-1} + b_r)$$

$$u_t = \mathrm{sigmoid}(W_{u\bar{x}} \cdot \bar{x}_t + W_{uh} \cdot h_{t-1} + b_u)$$

$$\tilde{h}_t = \tanh\left(W_{\tilde{h}\bar{x}} \cdot \bar{x}_t + W_{\tilde{h}h} \cdot (r_t \odot h_{t-1}) + b_{\tilde{h}}\right)$$

$$h_t = (1-u_t) \odot h_{t-1} + u_t \odot \tilde{h}_t.$$

The reset gate rt and update gate ut determine the relative weight of the previous hidden state ht−1 and the current input xt for updating the current hidden state ht. The last hidden state hT summarizes the sequence as a fixed-size vector. Importantly, the set of parameters W and b are independent of the sequence length T, which allows summarizing the methylation neighborhood independent of the number of cells in the training dataset. To encode cell-to-cell dependencies independently of the order of cells, the CpG module is based on a bidirectional GRU. It consists of a forward and backward GRU with 256d hidden state vectors ht, which scan the input sequence from the left and right, respectively. The last hidden state vector of the forward and backward GRU are concatenated into a 512d vector, which forms the output of the CpG module.

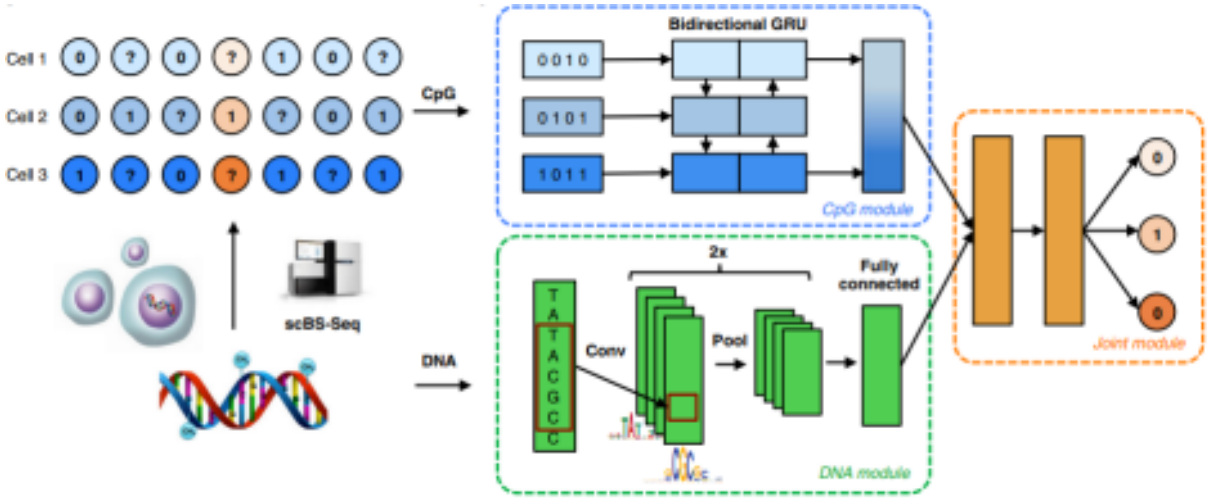
Joint module: The Joint module takes as input the concatenated last hidden vectors of the DNA and CpG module and models interactions between the extracted DNA sequence and CpG neighbourhood features via two fully connected hidden layers with 512 neurons and ReLU activation function. The output layer contains T sigmoid

neurons to predict the methylation rate ŷnt ∈ [0; 1] of CpG site n in cell t:

$$\hat{y}_{nt}(x) = \text{sigmoid}(x) = \left(\frac{1}{1+e^{-x}}\right).$$

Below is diagram for DeepCpG architecture:

$$\text{NLL}_w(\hat{y}, y) = -\sum_{n=1}^{N}\sum_{t=1}^{T} o_{nt}[y_{nt}\log(\hat{y}_{nt}) + (1-y_{nt})\log(1-\hat{y}_{nt})].$$



$$L(w) = \text{NLL}_w(\hat{y}, y) + \lambda_2 \|w\|_2$$

Model Training: Model parameters were learnt on the training set by minimizing the following loss function:

Here, the weight-decay hyper-parameter λ2 penalizes large model weights quantified by the L2.

$$L(w) = \text{NLL}_w(\hat{y}, y) + \lambda_2 \|w\|_2.$$

norm, and NLLw(ŷ, y) denotes the negative log-likelihood, which measures how well the predicted methylation rates ŷnt fit to observed binary methylation states ynt ∈ {0, 1}:

The binary indicator ont is set to one if the methylation state ynt is observed for CpG site n in

$$\text{NLL}_w(\hat{y}, y) = -\sum_{n=1}^{N}\sum_{t=1}^{T} o_{nt}[y_{nt}\log(\hat{y}_{nt}) + (1-y_{nt})\log(1-\hat{y}_{nt})].$$

cell t, and zero otherwise. Dropout with different dropout rates for the DNA, CpG and Joint module was used for additional regularization. Model parameters were initialized randomly following the approach in Glorot et al. The loss function was optimized by mini-batch stochastic gradient descent with a batch size of 128 and a global learning rate of 0.0001. The learning rate was adapted by Adam [65] and decayed by a factor of 0.95 after each epoch. Learning was terminated if the validation loss did not improve over ten consecutive epochs (early stopping). The DNA and CpG module were pre-trained independently to predict methylation from the DNA sequence (DeepCpG DNA) or the CpG neighborhood (DeepCpG CpG). For training the Joint

module, only the parameters of the hidden layers and the output layers were optimized, while keeping the parameters of the pre-trained DNA and CpG module fixed. Training DeepCpG on 18 serum mESCs using a single NVIDIA Tesla K20 GPU took approximately 24 h for the DNA module, 12 h for the CpG module and 4 h for the Joint module. Model hyper-parameters were optimized on the validation set by random sampling. DeepCpG is implemented in Python using Theano 0.8.2 and Keras 1.1.2.

Random forest models (RF, RF Zhang): Features of the RF model were i) the methylation state and distance of 25 CpG sites to the left and right of the target site (100 features) and ii) k mer frequencies in the 1001-bp genomic sequence centered on the target site (256 features). The optimal parameter value for k (k = 4) was found using holdout validation (Additional file 1: Figure S21a). The features for the RF Zhang model included

i. the methylation state and distance of two CpG neighbors to the left and right of the target site (eight features)

ii. annotated genomic contexts (23 features),

iii. transcription factor binding sites (24 features)

iv. histone modification marks (28 features) and

v. DNaseI hypersensitivity sites (one feature).

These features were obtained from the ChipBase database and UCSC Genome Browser for the GRCm37 mouse genome and mapped to the GRCm38 mouse genome using the liftOver

$$\hat{y}_{nt} = \text{mean}_{t' \neq t}(y_{nt'}).$$

tool from the UCSC Genome Browser. We trained a separate random forest model for each individual cell, as a pooled multi-cell model performed worse. Hyper-parameters, including the number of trees and the tree depth, were optimized for each cell separately on the validation set by random sampling. Random forest models were implemented using the Random Forest Classifer class of the scikit-learn v0.17 Python package.

CpG averaging (CpGAvg): For CpG averaging, the methylation rate ynt of CpG site n in cell t was estimated as the average of the observed methylation states ynt′ across all remaining cells t′≠ t:

ynt was set to the genome-wide average methylation rate of cell t if no methylation states were observed in any of the other cells.

Window averaging (WinAvg): For window averaging, the methylation rate ŷnt of CpG site n and cell t was estimated as the mean of all observed CpG neighbor's yn+k,t in a window of length W = 3001 bp centered on the target CpG site n

ŷnt was set to the mean genome-wide methylation rate of cell t if no CpG neighbours were observed in the window.

RESULTS

In accurate prediction of DNA methylation states, we assessed the ability of DeepCpG to predict single-cell methylation states and compared the model to existing imputation strategies for DNA methylation ("Methods"). As a baseline approach, we considered local averaging of the observed methylation states, either in 3-kbp windows centered on the target site of the same cell (WinAvg) or across cells at the target site (CpGAvg). Additionally, we compared DeepCpG to random forest classifiers trained on individual cells using the DNA sequence information and neighboring CpG states as input (RF). Finally, we evaluated a recently proposed random forest model to

predict methylation rates for bulk ensembles of cells, which takes comprehensive DNA annotations into account, including genomic contexts, and tissue-specific regulatory annotations such as DNase1 hypersensitivity sites, histone modification marks, and transcription factor binding sites (RF Zhang). All methods were trained, selected, and tested on distinct chromosomes via holdout validation ("Methods"). Since the proportion of methylated versus unmethylated CpG sites can be unbalanced in globally hypo- or hypermethylated cells, we used the area under the receiver operating characteristics curve (AUC) to quantify the prediction performance of different models. We have also considered a range of alternative metrics, including precision-recall curves, F1 score and Matthew's correlation coefficient, resulting in overall consistent conclusions. Initially, we applied all methods to 18 serum-cultured mouse embryonic stem cells, profiled using whole-genome single-cell bisulfite sequencing (scBS-seq) [5].

DeepCpG yielded more accurate predictions than any of the alternative methods, both genome-wide and in different genomic contexts. Notably, DeepCpG was consistently more

$$\hat{y}_{nt} = \text{mean}_{|k|<\frac{W}{2},k\neq 0}\left(y_{n+k,t}\right)$$

accurate than RF Zhang, a model that relies on genomic annotations. These results indicate that DeepCpG can automatically learn higher-level features from the DNA sequence. To investigate this, we tested for associations between the activity of convolutional filters in the DNA module and known sequence annotations ("Methods"), finding both positive and negative correlations with several annotations, including DNase1 hypersensitive sites, histone modification marks, and CpG-rich genomic contexts. The ability to extract higher level features from the DNA sequence is particularly important for analyzing single-cell datasets, where individual cells may be of different cell types and states, making it difficult to derive appropriate annotations.

To assess the relative importance of DNA sequence features compared to neighboring CpG sites, we trained the same models, however, either exclusively using DNA sequence features (DeepCpG DNA, RF DNA) or neighboring methylation states (DeepCpG CpG, RF CpG). Consistent with previous studies in bulk populations [12], methylation states were more predictive than DNA features, and models trained with both CpG and DNA features performed best. Notably, DeepCpG trained with CpG features alone outperformed random forest classifiers trained with both CpG and DNA features. A likely explanation for the accuracy of the CpG module is its recurrent network architecture, which enables the module to effectively transfer information from neighboring CpG sites across different cells.

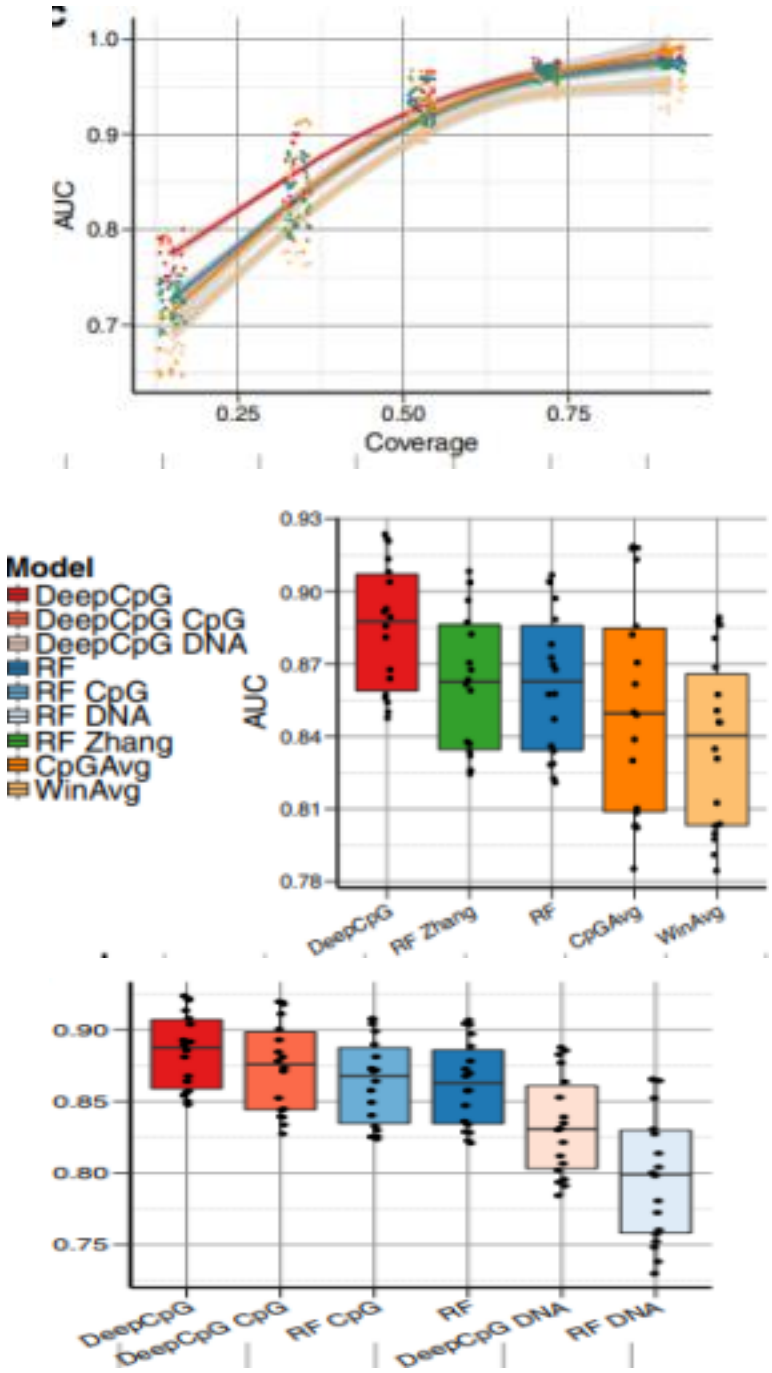
The largest relative gains between RF and DeepCpG were observed when training both models with DNA sequence information only. This demonstrates the strength of the DeepCpG DNA module to extract predictive sequence features from large DNA sequence windows of up to 1001 bp, which is particularly critical for accurate predictions from DNA in uncovered genomic regions, for example when using reduced representation sequencing data [6–8]. Consistent with this, the relative performance gain of DeepCpG compared to other methods was highest in contexts with low CpG coverage.

Next, we explored the prediction performance of all models in different genomic contexts. In line with previous findings [12, 13], all models performed best in GC-rich contexts. However, DeepCpG offered most advantages in GC-poor genomic contexts, including non CpG island promoters, enhancer regions, and histone modification marks (H3K4me1, H3K27ac)—contexts that are known to be associated with higher methylation variability between cells.

We also applied DeepCpG to 12 2i-cultured mESCs profiled using scBS-seq [5] and to data from three cell types profiled using scRRBS-seq [8], including 25 human hepatocellular carcinoma cells (HCCs), six human heptoplastoma-derived (HepG2) cells, and an additional set of six mESCs. Notably, in contrast to the serum cells, the human cell types are globally hypomethylated. Across all cell types, DeepCpG yielded substantially more accurate predictions than alternative methods, demonstrating the broad applicability of the model, including to hypo- and hypermethylated cells, as well as to data generated using different sequencing protocols.

ISSN (online): 1873-4049

First graph is result for genome-wide prediction performance for imputing CpG sites in 18 serum-grown mouse embryonic stem cells (mESCs) profiled using scBS-seq [5]. Performance is measured by the area under the receiver-operating characteristic curve (AUC), using holdout validation. Considered were DeepCpG and random forest classifiers trained either using DNA sequence and CpG features (RF) or using additional annotations from corresponding cell types (RF Zhang [12]). Additionally, two baseline methods were considered, which estimate methylation states by averaging observed methylation states, either across consecutive 3-kbp regions within individual cells (WinAvg [5]) or across cells at a single CpG site (CpGAvg), second is Performance breakdown of DeepCpG and RF, comparing the full models to models trained using either only methylation features (DeepCpG CpG, RF CpG) or only DNA features (DeepCpG DNA, RF DNA) and third graph is AUC of the methods in first graph by genomic contexts with increasing CpG coverage across cells.



Above graph is result for genome – wide prediction performance on 12 2i-grown mESCs profiled using scBS-seq [5], as well as three cell types profiled using scRRBS-seq [8], including 25 human hepatocellular carcinoma

cells (HCC), six HepG2 cells, and six additional mESCs. CGI CpG island, LMR low-methylated region, TSS transcription start site.

CONCLUSION

DeepCpG is a convolutional neural network-based computational technique for modeling low-coverage single-cell methylation data. We show that DeepCpG accurately predicts missing methylation states and detects sequence motifs associated with changes in methylation levels and cell-to-cell variability using mouse and human cells. We've shown that our model can accurately impute missing methylation levels, making genome-wide downstream analysis easier. DeepCpG has significant advantages in shallow sequenced cells as well as poorly covered sequence contexts with higher methylation variability. In single cell bisulfite sequencing research, more accurate imputation approaches may also aid to minimize the needed sequencing depth, allowing for the analysis of a larger number of cells at a lower cost

DeepCpG may also be used to find known and de novo sequence motifs that predict DNA methylation levels or methylation variability, as well as to assess the influence of single nucleotide alterations. Several of the motifs found by DeepCpG may be related to recognized motifs involved in DNA methylation regulation. The motifs that can be discovered are fundamentally limited to those that account for changes in each dataset, and so are dependent on the cell type and latent factors that drive methylation variability. Computational approaches such as DeepCpG can also be used to discern pure epigenetic effects from variations that reflect DNA sequence changes. Although we have not considered this in our work, it would also be possible to use the model residuals for studying methylation variability that is independent of DNA sequence effects.

Finally, we annotated areas with enhanced methylation variability using additional data generated from parallel methylation–transcriptome sequencing procedures. Integrating numerous data modalities assessed in the same cells utilizing parallel-profiling technologies, which are becoming increasingly available for diverse molecular levels, will be a key topic of future research.

REFERENCES:

[1]. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1189-z

[2]. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. 2010;11:191–203.

[3]. https://pubmed.ncbi.nlm.nih.gov/12855470/

[4]. https://pypi.org/project/deepcpg/

[5]. https://github.com/cangermueller/deepcpg#getting-started

[6]. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11:817–20.

[7]. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep. 2015;10:1386–97.

[8]. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. 2015;12:265–72

[9]. Zhou X, Li Z, Dai Z, Zou X. Prediction of methylation CpGs and their methylation degrees in human DNA sequences. Comput Biol Med. 2012;42:408–13.

[10] Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. 2015;33:364–76