OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Commentary

Machine Learning-Based Methodologies for DNA Methylation Investigation

Tao Dong*

Department of Biostatistics, Shanghai Maritime University, Shanghai, China

DESCRIPTION

Epigenetic markers, such as DNA (Deoxyribonucleic Acid) or chromatin changes are essential for creating and preserving cellular identity during development. Specific memory systems that control epigenetically regulated gene expression patterns have developed during that developmental phase. Specific epigenetic alteration patterns that either make chromatin accessible to transcription factors and activators or result in a closed chromatin structure that prohibits activated transcription are used to classify the active and inactive states of gene expression. The methylation of the cytosine's carbon-5 (5mC), which results from a covalent attachment of the methyl group to the carbon-5 of the cytosine ring, is the most noticeable of these markings. The majority of letter changes in DNA are typically accepted within the cell without changing behaviour, pathophysiology, or fate, although a genetic disease caused by a single point mutation does appear to be deterministic (DNA change=disease), and we can predictably change the sequence-determined outcome by genetic surgery. In this regard, either the majority of the genome's primary sequence is an evolutionary by-product made up of non-coding sequences that evolved over time from repetitive virus-like depositors of "junk" DNA, or one must take into account that the majority of genome regulation changes occur at the systems level [1].

To find 5mC sites, a number of conventional high-throughput sequencing methods have been created, including bisulfite sequencing, oxidative bisulfite sequencing, TET-Assisted Pyridine Borane Sequencing (TAPS), and Aza-IP. However, given the rapid expansion of nucleotide sequences produced in the post-genomic age, these experimental approaches are frequently time-consuming and labor-intensive. Investigating efficient computational techniques to locate 5mC locations is so crucial. To date, a number of machine learning-based prediction models, such as Methylator, MethCGI, iDNA-Methyl, and others, have been developed to solve this problem to identify cytosine methylation in CpG dinucleotides from the MethDB database, where each nucleotide was represented by utilising the traditional binary sparse encoding, Bhasin et al. created a Support Vector Machine (SVM) model dubbed Methylator. MethCGI is an SVM-based classifier that uses nucleotide sequence content and transcription factor-binding sites as characteristics to predict the methylation state of CpG islands in human brain tissues [2].

Later, the iDNA-Methyl predictor was built, and based on the pseudo-trinucleotide composition and the SVM classifier; it produced amazing gains in annotating the DNA methylation sites. Since then, several computational predictors, including RNAm5Cfinder, iRNAm5C-PseDNC, RNAm5CPred, iRNA-PseTNC, m5CPred-SVM, iRNAm5C SVM, and others, have been developed to identify 5mC sites in RNA sequences. A online application called RNAm5Cfinder uses the one-hot encoding and random forest technique to locate RNA 5mC locations in eight different rodent and human tissue/cell types. Comparing the effectiveness of various feature extraction techniques and classification algorithms, we gathered experimentally verified 5mC data from *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*, and developed an optimal predictor called iRNA-m5C for the identification of 5mC sites [3].

A loss function can be optimised with the use of the generalised boosting approach known as XGBoost. Its foundations lie in the boosting approach, an iterative ensemble method that trains models one after the other. Since these models (which are typically decision trees) follow simple prediction rules and only perform marginally better than a random guess, they might be referred to as "weak learners." The fundamental idea of boosting is to focus on the "hard" cases, or the examples that the model is unable to accurately predict. An ensemble method built on several decision trees is called Random Forest (RF). It inherits decision tree model advantages including good scalability to bigger datasets and resistance to irrelevant characteristics. Additionally, it enhances performance by lowering variance, which is one of the drawbacks of decision trees. The sampling procedure from the training data was used to build the "random" component of the Random Forest algorithm. A unique machine learning technique called Deep Forest (DF) draws inspiration from both ensemble learning and deep neural networks [4].

Correspondence to: Tao Dong, Department of Biostatistics, Shanghai Maritime University, Shanghai, China, E-mail: dong@tao.edu.cn

Received: 03-Jan-2023, Manuscript No. JDMGP-23-19624; **Editor assigned:** 06-Jan-2023, JDMGP-23-19624 (PQ); **Reviewed:** 20-Jan-2023, QC No. JDMGP-23-19624; **Revised:** 27-Jan-2023, Manuscript No. JDMGP-23-19624 (R); **Published:** 03-Feb-2023, DOI: 10.4172/2153-0602.23.14.273

Citation: Dong T (2023) Machine Learning-Based Methodologies for DNA Methylation Investigation . J Data Mining Genomics Proteomics. 14:273

Layer-by-layer processing and model diversity are the greatest principles from these two methodologies that Deep Forest incorporates. Deep Forest uses a cascade structure for layer-by-layer processing, where each level of the cascade gets feature information from the layer below and passes the results of its processing to the layer above. Additionally, each layer transforms the entire model into an ensemble of ensembles using a decision tree forest. BiLSTM-5mC is a deep learning-based method for precisely locating 5mC sites in genome-wide DNA promoters in SCLC cell line models. Finding 5-Methylcytosine (5mC) locations in genome-wide DNA promoters, which can aid scientists in better understanding a variety of critical biological processes, was the specific issue we have decided to address[5].

REFERENCES

1. Greenberg MV, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nature reviews Molecular cell biology. 2019;20(10):590-607.
2. Shi H, Wei J, He C. Where, when, and how: context-dependent functions of RNA methylation writers, readers, and erasers. Molecular cell. 2019;74(4):640-50.
3. Breiling A, Lyko F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. Epigenetics & chromatin. 2015;8(1):1-9.
4. Michalak EM, Burr ML, Bannister AJ, Dawson MA. The roles of DNA, RNA and histone methylation in ageing and cancer. Nature reviews Molecular cell biology. 2019;20(10):573-89.
5. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339(6121): 819-23.