Application of the machine learning in DNA methylation research

Si Chen*

Department of Biology, McGill University, Montreal, Quebec H3A 0E7, Canada

*Corresponding author: si.chen4@mail.mcgill.ca

Keywords: Machine learning, DNA methylation, data analysis, information extraction, epigenetic changes.

Abstract: Machine learning (ML) is an artificial intelligence technique that allows systems to make decision or prediction by learning from experience independently based on large data sets. It is efficient for data analysis and information extraction when the data sets are big and complicated, which makes it an ideal tool to help study epigenetic changes like DNA methylation. DNA methylation, one of the major epigenetic processes that mediate gene expression without changing DNA sequences, is an important causative factor of many malignant transformations like cancers and aging. ML applications in studying DNA methylation patterns and identifying genomic regions susceptible to DNA methylation changes have brought great promises for disease diagnosis and personalized treatment. By reviewing the cases of ML algorithms applied in DNA methylation study, we derive a brief developmental history of how different ML techniques had been used to extract more and more useful information about DNA methylation, which provides insights for future ML application improvement.

1. Introduction

Nowadays, enormous data with great complexity has become an essential part of our life as computing power and information technologies keep advancing. Valuable information underlies the large and constantly growing data sets. To analyze such data and extract useful information efficiently, the help of certain tools and techniques is required. Machine learning (ML), with its ability to generate predictive or decision-making model based on past data records independently, is actually fitting to the job [1]. Based on the different features and training methods, ML can be divided into many different categories, with supervised learning, unsupervised learning, and reinforcement learning being the three main types [2]. Given their superiority in predictive analytics and data exploration, ML techniques have been applied in many aspects of life, including speech recognition, face detection, ads personalization, and stock forecasting. Among them, the application of ML in medical industry has become one of the most unneglectable [3].

Epigenetic mechanisms, acting as an additional layer of gene activity regulator without changing the DNA sequence, are deeply involved in the occurrence of many pathologies like cancers, pediatric syndromes, and genetic disorders [4], so recognizing regions susceptible to epigenetic alterations can be crucial. DNA methylation is an important epigenetic process that has been well studied throughout the years. It is capable of mediating gene expression by adding a methyl group to the fifth carbon of a cytosine to form 5-methylcytosine. DNA methylation status controls chromatin states and hence transcription initiation. Hypermethylation and hypomethylation are related to transcription repression and transcription promotion respectively. When aberrant DNA methylation pattern alterations, like hypermethylation of tumor repressing genes and hypomethylation of oncogenes, take place, unfavorable diseases might occur [5]. Therefore, studying DNA methylation changes can provide promising insights for disease diagnosis and treatment [6].

However, certain difficulties lie within the study of DNA methylation. Conducting high-throughput assays to search for epigenetic changes across all tissues and cells in various conditions is overly demanding for time and money [7]. Traditionally, epigenetic data sets are continuous, non-linear, interacting, and high in dimensionality. Analyzing such data sets often requires correction for multiple

hypothesis testings and dealing with multi-collinearity. How to uncover useful information hidden beneath the large and complex data sets efficiently becomes the key. In this case, incorporating ML techniques such as support vector machine (SVM), regression, and deep learning (DL) to generate predictive models and extract information has been considered as the solution, as it has been so much more efficient than previous approaches [5]. In this review, we are going to go through basic ML techniques, epigenetic mechanisms especially DNA methylation, and the applications of ML algorithms in DNA methylation study, so that we can have a brief idea of how ML has been applied to epigenetics across the years.

2. Machine learning and epigenetics

2.1. What is machine learning & Machine learning types

Machine learning, branched from artificial intelligence, serves as a powerful tool for data analytics. It is developed from pattern recognition and computational learning theory. With the input of suitable data, machine learning enables automatic building of analytical models that generate accurate predictions based on historical data trends [3]. Algorithms are used to produce such models. People are now using machine learning in various areas for task solving, including image processing, bioinformatics, information retrieval, intrusion detection, pattern recognition and so on [1].

There are three main types of problems that are especially suitable for machine learning to solve: structure adjustment when a large amount of data is input into the system, data mining problem, and highly active tasks requiring consistent upgrade of present machine designs to keep up with the change of environment [1, 8]. Depending on the input and output data types, features of problems to be solved, and the learning approach, the machine learning algorithms can be roughly classified into three main types: supervised learning, unsupervised learning, and reinforcement learning [1].

To be more precise, supervised learning is featured by having example input and corresponding output so that the algorithm can learn how to map the input to the output from the labeled data. Classification and regression are two major supervised learning technique categories popular in use [2].

Unsupervised learning algorithms, such as clustering, on the other hand, are provided with unlabeled datasets and have only input data but no related output variables. They learn the features of the datasets by uncovering hidden patterns or data groupings, which makes them useful in many ways, including exploratory data analysis [1, 2].

Reinforcement learning is a sequential decision-making technique allowing the computer to reach a specific goal in the absence of input or supervision. To apply this technique, a model is trained by a reward-penalty mechanism based on its actions. The model aims for reward maximization, so it will manage to learn from experience [1, 2].

Besides the three main types, there also exist other useful ML techniques: hybrid approaches such as semi-supervised learning, self-supervised learning, and self-taught learning; and other common approaches like multi-task learning, active learning, online learning, transfer learning and so on [2].

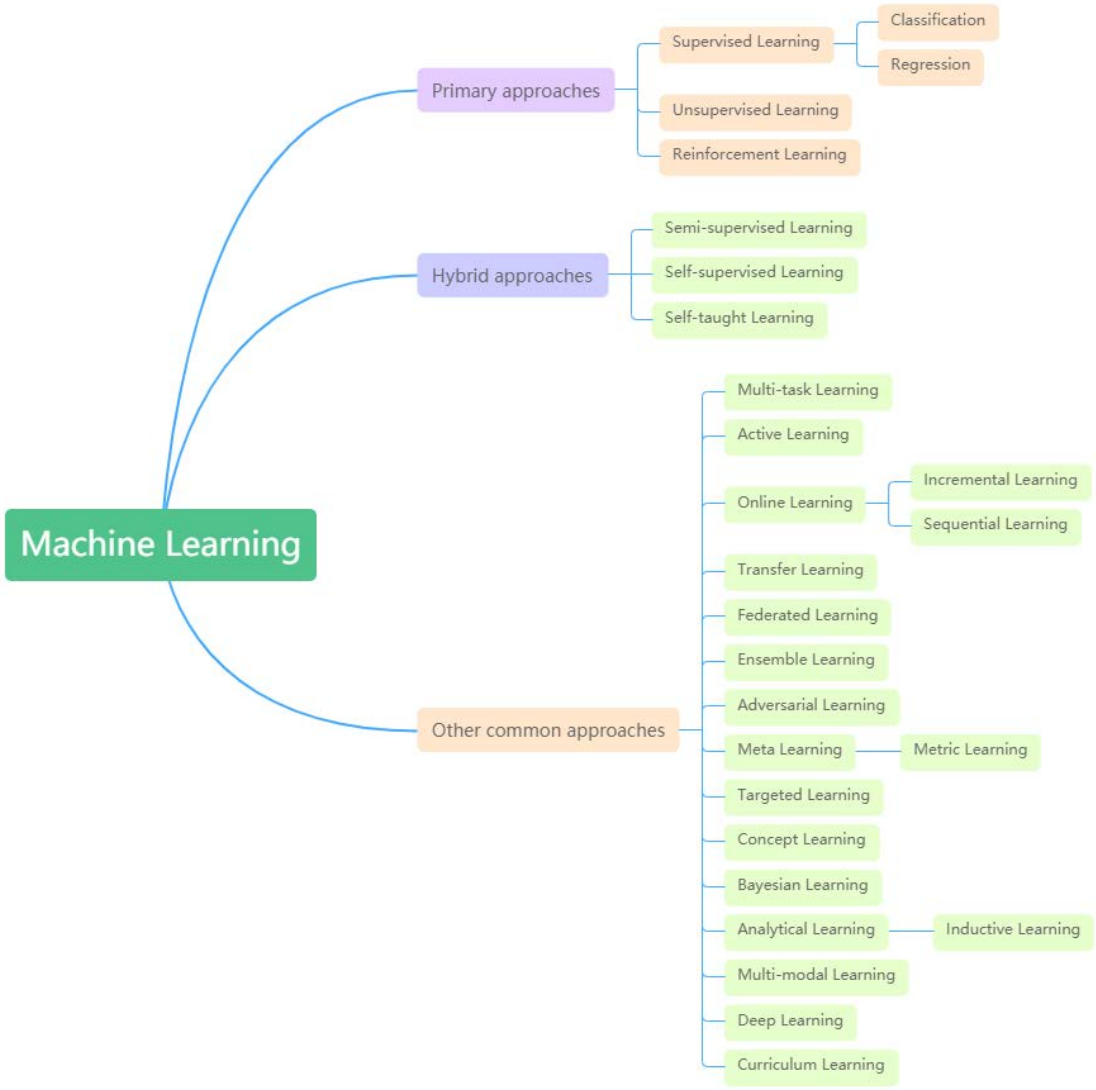

Figure 1. ML approaches [2]

2.2. Epigenetics and DNA Methylation

2.2.1. What is epigenetics.

Epigenetics is the study of heritable changes in gene activity that are independent from DNA sequence variations. Epigenetic mechanisms can alter the structure of DNA as a result of post-translational modification in proteins or post-replication modification in DNA. Such changes are usually reversible and take place dynamically [9]. There are a variety of epigenetic processes affecting gene functions, including methylation, acetylation, phosphorylation, ubiquitylation and so on [10]. Among them, DNA methylation has been one of the major topics that are well-studied.

Epigenetic mechanisms play an essential role in regulating the variated gene expression profiles in different tissues and cells [11]. For instance, epigenetic mechanisms are critical for regulating nervous system. DNA methylation and histone modification mediate neural cell differentiation, neural behavior, and neural plasticity changes in the CNS [12]. Epigenetic processes also contribute to stem cell development regulation. Jumonji domain-containing protein-3 (JMJD3), being a histone demethylase controlling H3K27me3-related gene expression, is a typical example of epigenetic regulator that determines cell fate of multipotent and pluripotent stem cells [13].

If epigenetic processes alter improperly, there can occur substantial health problems, including cancers, mental retardation, autoimmune diseases, pediatric syndromes, aging, and syndromes

involving chromosomal instabilities can be resulted from disruptions in epigenetic mechanisms [4, 6]. Take DNA methylation and cancers, which is under the spotlight nowadays, as an example. On the one hand, when the cytosines 5' in CpG islands from the promoter region mediating suppressor genes are hypermethylated, the gene expression is silenced and the DNA methylation patterns are disturbed, which often relates to tumorigenesis. When abnormal hypomethylation occurs in the promoter region of oncogenes in tumors, on the other hand, transcriptional activities are activated repeatedly, instablizing chromosomes. Prostate cancer, one of the most widespread diagnosed cancers in the world, is actually deeply related to aberrant epigenetic changes like frequent hypermutation in CpG islands, which severely disrupts regular gene activities and causes disease [4, 9].

Since epigenetics is so deeply connected to gene activity and diseases, having a thorough understanding of epigenetic mechanisms can provide enormous help for the diagnosis and treatment of such diseases. Leukemias and myelodysplastic syndrome, which used to be known for hard to cure, have found breakthroughs via epigenetic study in recent years. It has been identified that DNA hypermethylation makes unneglectable contribution to the diseases. As a result, nucleoside analogues that are capable of methylation inhibition and gene reactivation, like azacitidine, have been incorporated into DNA replication and tested to be effective in clinical trials treating the two diseases [4]. Hence, epigenetic therapies that mediate epigenetic modifications have become a promising research area for health industry [6].

2.2.2 DNA methylation.

DNA methylation occurs with the help of a family of DNA methyltransferases (Dnmts). Catalyzed by Dnmts, a methyl group is added to the 5'-carbon of cytosine in CpG dinucleotide sequences [9, 11]. The main DNMT family members popular in study are Dnmt1, Dnmt3a, Dnmt3b, which are critical for embryo development. Dnmt1 is responsible for the maintenance of DNA methylation, especially in neural progenitor cells undergoing cell division, which fits its main effecting location of embryonic nervous system. Dnmt3a and Dnmt3b are together recognized as de novo Dnmt, as they usually express themselves in a complementary way to establish a new methylation pattern to unmodified DNA. [11, 12]

DNA methylation patterns serve as an additional means of gene expression regulation without changing DNA sequences. DNA methylation can prevent gene expression. To take effect, it can change chromatin to be repressive with the help of methyl-CpG binding domain (MBD) or it can simply prevent transcription factors from DNA binding [4, 12]. Depending on the genomic regions where DNA methylation occurs, gene activities can vary. Retroviral elements silencing, tissue-specific gene expression mediating, genomic imprinting, and X chromosome inactivation and other gene activities are all regulated by DNA methylation [11].

2.3. ML algorithm applications

While the study of epigenetics is promising, researching on such topic can be challenging. Biological data sets often possess high dimensionality but relatively few interested cases. Extracting and analyzing the epigenetic information of interest, like regions susceptible to epigenetic alterations in genomes, directly from the enormous data require a lot of time and resources, so developing tools capable of analyzing genomic data efficiently and providing substitute solutions, like predicting possible epigenetic sites based on DNA sequences, to meet the need is crucial. Machine learning, with its efficient data exploration and epigenetic feature identification ability, is suitable for the task [14].

There have been many cases involving various ML approaches to help study epigenetics across the years. Here, we are going to list a few that focus on applying ML techniques on studying DNA methylation, as methylation has been an essential topic of epigenetics with great promise.

Since DNA methylation can insert great influence on gene expression, DNA methylation prediction has, unsurprisingly, been in the spotlight for researchers for decades.

Early genomic methylation predictions were more sequence- and structure-based, and the classifying approaches were more of a binary manner, like identifying methylation patterns by dividing the genomes into hypomethylated or hypermethylated categories [7].

Prior to 2005, few methods were published to predict DNA methylation sites. Bhasin et al. raised a ML approach for cytosine methylation prediction in CpG dinucleotides on the basis of SVM. They used the SVM module developed from human data to analyzed genome-wide methylation sites, finding that in UTRs CpGs are more methylated than exonic and intronic regions in human [15].

In 2006, Das et al. also developed computational pattern recognition approaches for methylation prediction in human. After testing various ML approaches such as logistic regression, K means clustering, linear discriminant analysis, and SVM on a selected genomic sequence, they as well derived certain predictions of methylation pattern in human genomes and concluded that SVM at that time was the best-performing classifier [16].

However, such methylation predictions made solely from sequences were restricted since they considered insufficiently about the remaining unchanged underlying sequence features. To further investigate methylation patterns, some researchers studied the correlations between transcription factor (TF)-binding profiles, chromatin landscapes, and DNA methylation, which later facilitated the improvement of methylation prediction techniques incorporating other epigenetic features like histone modification [7].

In 2014, Ma et al. established a model using support vector regression for locus-specific cross-tissue methylation prediction. This time, instead of binary, methylation at CpG sites is predicted in a continuous manner. They were able to predict epigenetic information of tissues whose data is hard to acquire based on tissues with more accessible data [17].

In 2015, Zhang et al. developed a random forest classifier to predict CpG methylation levels with higher accuracy and better robustness. The new model, unlike the old ones, were made on the basis that was not restricted to neighboring CpG sites or locus specificity, but took other epigenetic features into consideration as well, such as CpG islands, genomic position, histone modifications, co-localized DNase I hypersensitive sites, TF-binding sites, and so on [18].

After that, epigenetic and genetic traits in various genomic regions across cell types related to DNA methylation patterns had been systematically identified and studied, which benefited later study of methylation [19]. In 2016, Wang et al. employed a DL autoencoder called DeepMethyl to make DNA methylation prediction based on 3D genome topological features and DNA sequences patterns. The DL algorithm was tested to generate more accurate predications than traditional ML approaches (SVMs) in some circumstances, which presents the value of DL application in epigenetic study [20].

As techniques used to study epigenetic features keep evolving, analyzing the enormous data generated across different tissues and cell types are no longer as resource-consuming as the past. Novel ML approaches different from the past are being modified and tested on various biological data sets, so that scientists are now able to utilize the large data sets to train computational models and predict the epigenetic state of interest with higher accuracies and lower cost [21].

Based on the idea that signatures of DNA methylation are mainly tissue type-specific, P. Jurmeister et al., in 2019, developed three ML classifiers that can distinguish pulmonary metastases of head and neck squamous cell carcinoma (HNSC) from a second squamous cell carcinoma of the lung (LUSC) with high accuracies: artificial neural network, SVM, and random forest (RF). They trained the three algorithms with DNA methylation profiles and offered them probability scores of the classification results. Fivefold cross-validation on the reference cohort was used to help tune the algorithms as well. The models they derived were tested to be successful on independent validation datasets, with the artificial neural network performing the best. The artificial neural network has a classification correction rate of 96.4% when applied to a validation cohort of 279 patients with HNSC, LUSC, and normal lung controls while SVM and RF only correctly classified 95.7% and 87.8% of the cases respectively. The neural network also possessed the highest proportion (92.1%) of cases whose prediction accuracies were over 99%, outperforming that of SVM (90%) and RF (43%). Thus, they concluded that the artificial neural network was more favorable than the traditional ML approaches SVMs and RFs for methylation prediction under certain conditions [22].

3. Conclusion

In this review, we provide a brief introduction of ML algorithms and epigenetics as well as a brief overview of the applications of ML techniques in DNA methylation research. The features of ML and DNA methylation have made them an exceptionally compatible pair of research tool and subject. ML can generate predictive models and find hidden patterns in large data sets. DNA methylation, with its enormous data repositories like ENCODE and the BLUEPRINT consortium and its chemical stability, is ideal for ML application since bigger training data sets lead to more accurate predictive results [23].

The algorithms that are most popular in use in the field of DNA methylation is supervised learning, especially SVMs and RFs. In recent years, though, researchers have started to explore the application of DL algorithms like artificial neural network and other novel approaches to provide new insights for DNA methylation study, as they are more advantageous than traditional ML algorithms under certain circumstances. However, while enjoying the accuracy and convenience offered by ML for epigenetic study, we should also not forget that certain limitations still lie within ML techniques. For supervised learning, to generate reliable predictions, enormous data that are correctly labelled is necessary as training classes, and when applied on other external data sets there is a chance that the algorithms will not work as well as on the training data sets. For DL algorithms, which can also be applied in supervised ML, the black-boxing mechanism prevents DL from medical-related application, since the outcome-determination process is unknown. Besides, there exist other unsolved problems for all ML like prediction bias and the constantly growing need for more data [23, 24].

To address the problems and predict epigenetic phenomena more efficiently in the present environment where new data types keep emerging and data availability keeps increasing, researchers need to advance ML algorithms and explore new possibilities for ML application in epigenetics. One future approach for improvement is to combine the characteristics of different ML algorithms together as a new technique. Holder et al. established a novel algorithm by combining DL, active classification learning (ACL), and imbalanced class learning (ICL) together so that each of their advantages are combined as well. The new combinatorial approach was capable of feature generation, optimal feature selection, and imbalanced class problem improvement. It was tested to predict genomic sites susceptible to methylation alterations with higher accuracy [14]. Another effective approach to improve epigenetic prediction is having more complete genome-wide epigenetic data sets available to the public, so that researchers can test their ML models more comprehensively and find the more appropriate model for the study. This need for larger data access can be met by collaboration of organizations who possess large biological datasets, like the sequencing read archive (SRA) database from the National Center for Biotechnology Information (NCBI) [23]. In short, the increasing availability of biological data sets and the constant advancement of ML techniques are enabling the improvement of ML applications in DNA methylation research.

References

[1] Udousoro, I.C., Machine Learning:A Review. Semiconductor Science and Information Devices, 2020. 2(2).

[2] Sah, S., Machine Learning: A Review of Learning Types. 2020.

[3] Angra, S. and S. Ahuja, Machine learning and its applications: A review, in 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). 2017. p. 57-60.

[4] Rodenhiser, D. and M. Mann, Epigenetics and human disease: translating basic biology into clinical applications. CMAJ, 2006. 174(3): p. 341-8.

[5] Levy, J.J., et al., MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC Bioinformatics, 2020. 21(1): p. 108.

[6] Egger, G., et al., Epigenetics in human disease and prospects for epigenetic therapy. Nature, 2004. 429(6990): p. 457-63.

[7] Pavlovic, M., et al., DIRECTION: a machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. Bioinformatics, 2017. 33(19): p. 2986-2994.

[8] Lavecchia, A., Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today, 2015. 20(3): p. 318-331.

[9] Nowacka-Zawisza, M. and E. Wisnik, DNA methylation and histone modifications as epigenetic regulation in prostate cancer (Review). Oncol Rep, 2017. 38(5): p. 2587-2596.

[10] Weinhold, B., Epigenetics: the science of change. Environ Health Perspect, 2006. 114(3): p. A160-7.

[11] Moore, L.D., T. Le, and G. Fan, DNA methylation and its basic function. Neuropsychopharmacology, 2013. 38(1): p. 23-38.

[12] Feng, J., S. Fouse, and G. Fan, Epigenetic Regulation of Neural Gene Expression and Neuronal Function. Pediatric Research, 2007. 61(5 Part 2): p. 58R-63R.

[13] Ding, Y., et al., JMJD3: a critical epigenetic regulator in stem cell fate. Cell Commun Signal, 2021. 19(1): p. 72.

[14] Holder, L.B., M.M. Haque, and M.K. Skinner, Machine learning for epigenetics and future medical applications. Epigenetics, 2017. 12(7): p. 505-514.

[15] Bhasin, M., et al., Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Letters, 2005. 579(20): p. 4302-4308.

[16] Das, R., et al., Computational prediction of methylation status in human genomic sequences. Proceedings of the National Academy of Sciences, 2006. 103(28): p. 10713-10716.

[17] Ma, B. and et al., Predicting DNA methylation level across human tissues. Nucleic Acids Research, 2014. 42(6): p. 3515.

[18] Zhang, W., et al., Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biology, 2015. 16(1).

[19] Yan, H., et al., Chromatin modifications and genomic contexts linked to dynamic DNA methylation patterns across human cell types. Scientific Reports, 2015. 5(1): p. 8410.

[20] Wang, Y., et al., Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. Scientific Reports, 2016. 6(1): p. 19598.

[21] Fan, S. and et al., Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. Genomics, 2016. 107(4): p. 132.

[22] Jurmeister, P., et al., Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. Science Translational Medicine, 2019. 11(509): p. eaaw8513.

[23] Rauschert, S., et al., Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. Clin Epigenetics, 2020. 12(1): p. 51.

[24] Phillips, P.J., et al., An other-race effect for face recognition algorithms. ACM Transactions on Applied Perception, 2011. 8(2): p. 1-11.