Title

Predicting CpG Methylation Sites in Arabidopsis Thaliana Chromosome 1

Abstract

CpG methylation is a fundamental epigenetic mechanism influencing gene expression, genome stability, and stress response in plants like Arabidopsis thaliana. Understanding CpG methylation sites provides crucial insights into epigenetic regulation and its biological consequences. This study aimed to develop machine learning models to predict CpG methylation sites on chromosome 1 of Arabidopsis thaliana. The hypothesis posits that GC content and nucleotide composition are sufficient features for accurately determining methylation patterns. To address the research question, random negative samples were generated to balance the original dataset, and two machine learning models, SGDClassifier and XGBoost, were trained and evaluated. XGBoost outperformed SGDClassifier, achieving higher scores of precision, recall, F1, and AUC. These findings highlight the potential of machine learning in advancing epigenetic research. Future studies can extend this approach to genome-wide methylation prediction and integration with broader epigenomic data.

Introduction

DNA methylation, particularly CpG methylation, plays a crucial role in regulating gene expression and maintaining genome stability in plants like Arabidopsis thaliana. It silences transposable elements and influences the plant's response to environmental stress. This study aims to answer the following research question: Can machine learning models effectively predict CpG methylation sites on Arabidopsis thaliana chromosome 1 using GC content and nucleotide composition as features? Accurate prediction of CpG methylation sites can provide valuable insights into gene regulation, genome stability, and plant responses to environmental stress. To address this question, two powerful machine learning models, SGDClassifier and XGBoost, were employed. These models were selected for their ability to handle large datasets and their potential to capture complex patterns within sequence data. By leveraging these models, we aim to deepen our understanding of CpG methylation and its impact on plant biology.

Background

CpG sites, where a cytosine nucleotide is followed by a guanine nucleotide, are key players in epigenetic regulation. Methylation of the cytosine at these sites, a process that doesn't alter the DNA sequence, significantly impacts gene expression, genome stability, and defense against transposable elements. In Arabidopsis thaliana, CpG methylation is particularly important for controlling transposable elements and modulating gene expression in response to environmental stress.

Accurate prediction of CpG methylation sites can provide valuable insights into gene regulation, genome stability, and plant responses to environmental stress. To address this question, two powerful machine learning models, SGDClassifier and XGBoost, were employed. These models were selected for their ability to handle large datasets and their potential to capture complex patterns within sequence data. By leveraging these models, we aim to deepen our understanding of CpG methylation and its impact on plant biology

Motivation for the problem addressed

While the significance of CpG methylation in biological processes is well-established, accurately identifying these methylated sites across the genome remains a major hurdle. Experimental methods like bisulfite sequencing are costly and time-consuming, limiting their applicability to large-scale studies and cross-species comparisons (Jones, 2012). This limitation restricts the understanding of methylation in complex biological processes and its potential applications in agriculture.

Beyond fundamental research, these insights have practical implications. As agriculture faces increasing pressures from climate change and food insecurity, predicting methylation patterns could guide breeding programs to develop stress-resistant crops. By identifying epigenetic modifications linked to desirable traits, researchers can implement targeted interventions to improve crop resilience and productivity (Thiebaut et al., 2019). The machine learning models explored in this study demonstrate the potential for computational tools to transform both epigenetic research and agricultural practices.

Materials

The data used in this study consists of two primary sources: a BEDGraph file and a FASTA file. The BEDGraph file provides CpG methylation data for chromosome 1 of

Arabidopsis thaliana, detailing the chromosomal coordinates and methylation levels at each site. These data were essential for identifying CpG sites and their respective methylation statuses, serving as the foundation for the positive samples in the dataset. The FASTA file contains the reference genome of Arabidopsis thaliana, enabling the extraction of sequence-based features such as GC content and nucleotide composition. Together, these files form a comprehensive dataset that facilitates the prediction of CpG methylation sites by linking sequence characteristics to methylation patterns. To address the imbalance between methylated and unmethylated CpG sites, random negative samples were generated from the reference genome, ensuring no overlap with known CpG sites. These negative samples provide a balanced representation of methylated and unmethylated regions.

Methods

To initiate the analysis, all the necessary libraries to process the datasets and training models were imported and the primary datasets were loaded. As the chromosome identifiers differed between the two datasets (numeric in BEDGraph, NCBI-style in FASTA), a mapping dictionary was created to harmonize them. The BEDGraph data was enriched with a new column containing mapped chromosome names, aligning it with the genomic data. To delve deeper into CpG sites, sequences flanking each site within a ±50 base window were extracted using a custom function. The extracted sequences were integrated into a new column, setting the stage for feature engineering and model training.

Starting with the predictors, GC content was computed as a key predictive feature for this study. To evaluate its relevance, the GC content was calculated both locally for each CpG region and globally for the entire genome. For CpG regions, the GC content was determined by counting the occurrences of G and C bases within the extracted sequence window surrounding each CpG site. This information was stored in a new column within the dataset, providing a localized metric for each CpG site. In addition to local calculations, the genome-wide GC content was computed for all chromosomes in the Arabidopsis thaliana reference genome. This step was essential to compare the GC content of CpG regions with the average GC content across the genome. Next, nucleotide composition was calculated for each CpG region to quantify the relative abundance of each nucleotide. It was computed by dividing the count of each nucleotide

by the total length of the sequence. For each sequence, the composition was calculated and stored in a new DataFrame, which was then merged with the original dataset.

Moving on to generating random negative samples, a custom function was designed to generate the samples which avoided overlaps with known CpG sites by comparing the coordinates of randomly selected sequences against the CpG dataset. Each negative sample matched the sequence length of the positive samples (±50 bases). Afterward, GC content and nucleotide composition were calculated for these regions using the same functions applied to the CpG samples. Once the features were extracted, the positive CpG samples were labeled as (1), and the negative samples were labeled as (0) to establish a binary classification framework. Finally, datasets were concatenated into a single combined dataset.

The next step was to train the models. Starting with SGDClassifier - the dataset was split into training and testing subsets with an 80:20 ratio. A balanced dataset was created by upsampling the negative samples to address class imbalance. Before feeding the data into the model, feature scaling was performed using a StandardScaler. This step standardized the input features, ensuring that all predictors had comparable scales. The model was initialized with a hinge loss function, suitable for binary classification tasks, and trained using the scaled training data. Next, the study ran a feature importance analysis test by extracting the weights assigned to each feature. The outcomes will be discussed in the Results section.

Moving onto XGBoost, the dataset was split into the same ratio. Feature scaling was not explicitly performed for XGBoost, as it is robust to varying feature scales. Feature scaling was not explicitly performed for XGBoost, as it is robust to varying feature scales. The model was initialized with parameters such as “**use_label_encoder=False”** and “**eval_metric="logloss"”** to ensure compatibility and effective evaluation. Once the results were produced, a feature importance analysis was performed again for this model to determine which predictor held more weight than others.

Results

SGDClassifier

SGDClassifier Classification Report on Balanced Dataset:

	precision	recall	f1-score	support
0	0.74	0.78	0.76	1102682
1	0.77	0.72	0.74	1102981
accuracy			0.75	2205663
macro avg	0.75	0.75	0.75	2205663
weighted avg	0.75	0.75	0.75	2205663

Figure 1. SGDClassifier Classification Report

Figure 1 reveals the performance metrics across two classes: methylated (label 1) and unmethylated (label 0) CpG sites. The model achieved an overall accuracy of 75%. The precision for methylated sites (0.77) indicates that 77% of predicted methylated sites were correct. However, the recall for methylated sites was slightly lower at 72%, showing room for improvement in identifying all true methylated cases. The F1 scores, which provide a harmonic mean of precision and recall, were comparable for both classes.

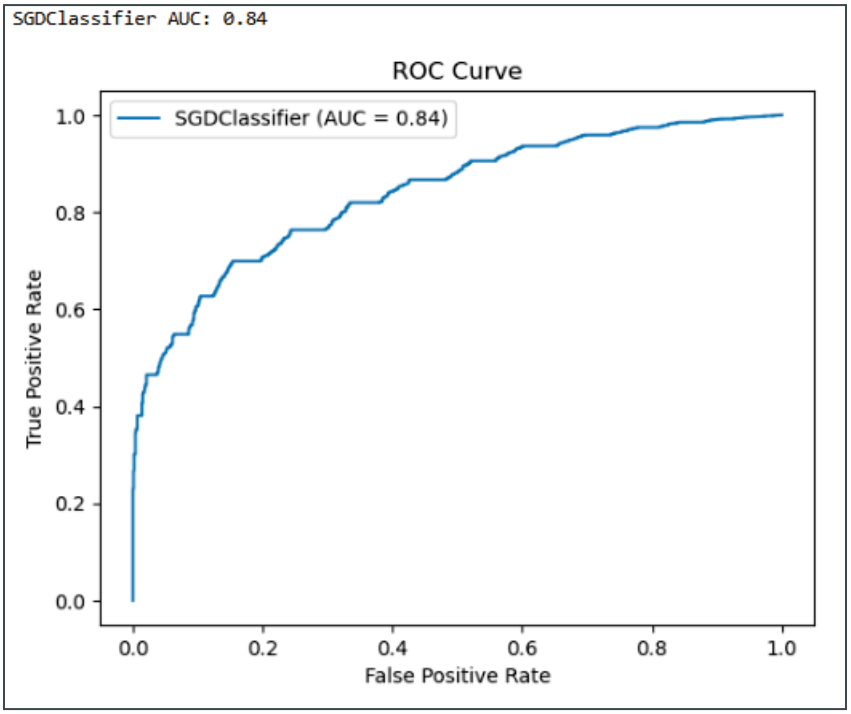

Figure 2. SGDClassifier AUC Score & ROC Graph

According to Figure 2., the model's area under the ROC curve (AUC) was calculated as 0.84, which signifies good discriminatory power between methylated and unmethylated sites. The ROC curve further illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate across different thresholds. The steep rise at the beginning of the curve suggests that the model performs well in distinguishing between the two classes in most scenarios.

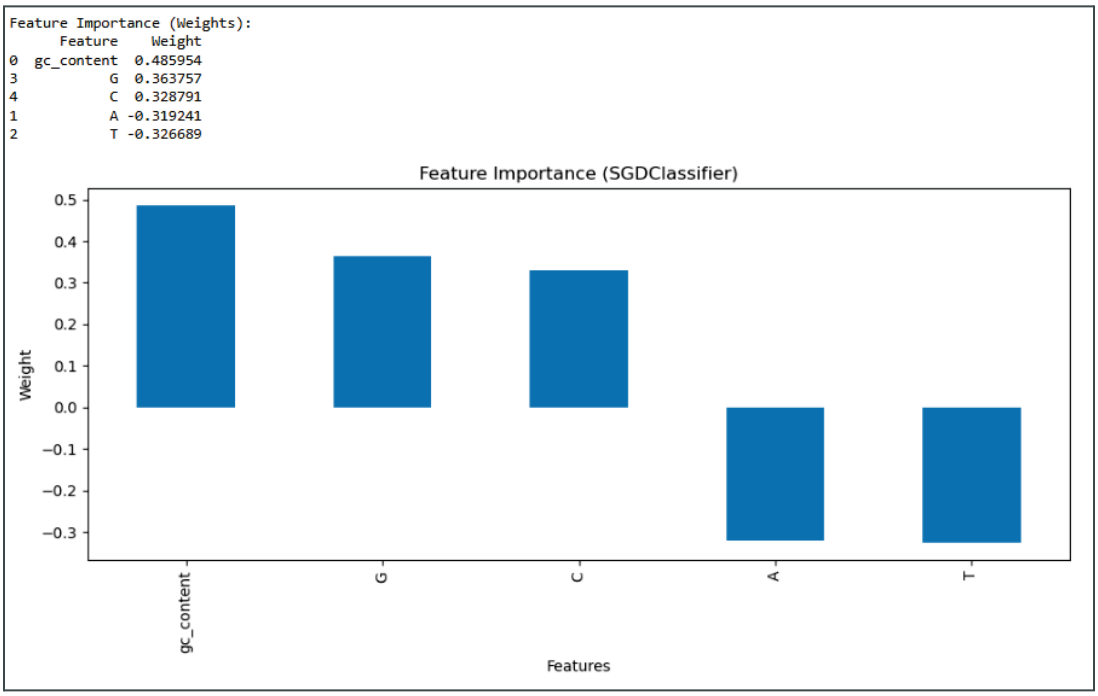

Figure 3. SGDClassifier Feature Importance Analysis

Figure 3. showcases the weights assigned to different predictors by the model. GC content emerged as the most influential predictor with a weight of approximately 0.49. This finding supports the hypothesis that GC content is critical in CpG methylation prediction. Among nucleotide composition features, guanine (G) and cytosine (C) exhibited higher weights compared to adenine (A) and thymine (T), further validating their relevance to the biological context of methylation.

Overall, the SGDClassifier provided a strong starting point for CpG methylation prediction, with reasonable accuracy and robust feature insights. However, the moderate recall and precision metrics highlight the need for further optimization or exploration of alternative models to improve predictive performance.

XGBoost

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.98	0.97	0.97	1102682
1	0.97	0.98	0.97	1102981
accuracy			0.97	2205663
macro avg	0.97	0.97	0.97	2205663
weighted avg	0.97	0.97	0.97	2205663

XGBoost AUC: 1.00

Figure 4. XGBoost Classification Report

Figure 4. demonstrates the exceptional performance by XGBoost - achieving a precision of 97% for both classes, indicating a high level of accuracy in predicting methylated and unmethylated CpG sites. Recall and precision scores of positive and negative samples culminated in an F1 score of 97%. The AUC of 1.0 indicates a perfect distinction between the positive and negative classes, underscoring the model's ability to assign higher probabilities to true positives than to false positives.

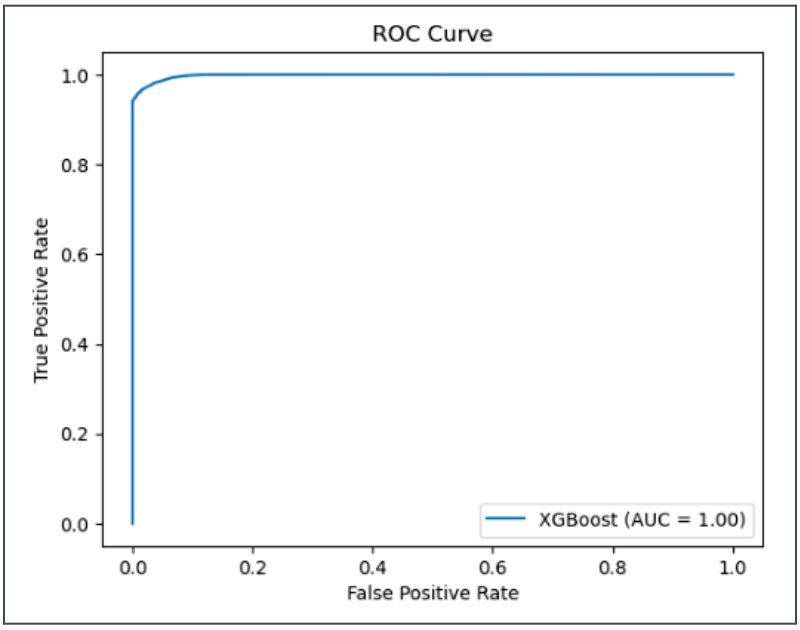

Figure 5. XGBoost ROC Graph

The ROC curve further illustrates the efficacy of the XGBoost model. It approaches the upper-left corner, a hallmark of highly accurate predictive performance. This comprehensive

evaluation highlights XGBoost's capability in accurately predicting CpG methylation sites based on sequence-based features.

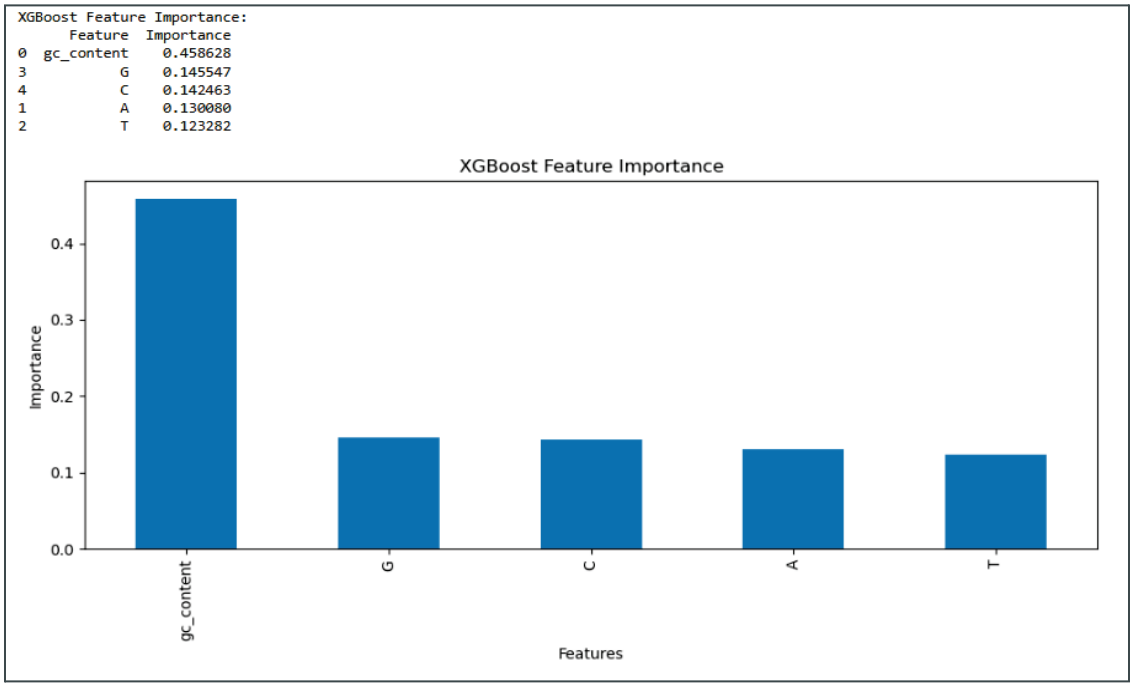

Figure 6. XGBoost Feature Importance Analysis

Figure 6 displays the most weighted predictors evaluated by XGboost. GC content emerged as the most influential feature, with an importance score of approximately 0.46, far exceeding the contributions of nucleotide composition. Among individual nucleotide frequencies, guanine (G) had the second-highest importance, followed by cytosine (C), adenine (A), and thymine (T).

Discussion/Conclusion

The findings of this study highlight the potential of machine learning models, specifically SGDClassifier and XGBoost, in predicting CpG methylation sites based on sequence-based features. The results demonstrate that simple features like GC content and nucleotide composition can predict methylation patterns effectively. While both SGDClassifier and XGBoost provided meaningful predictions, XGBoost significantly outperformed SGDClassifier. XGBoost's ability to capture complex patterns in data and handle class imbalances made it a better fit for this task. Additionally, the feature importance analysis provided valuable insights,

showing that GC content played a central role in predictions, which aligns with its known biological relevance to CpG regions. The study underscores the biological significance of GC content in predicting CpG methylation. CpG sites are known to be GC-rich, and the results reflect this, with GC content emerging as the most important predictor in both models. Nucleotide composition also contributed, albeit to a lesser extent. These findings support the hypothesis that these sequence-based features can provide meaningful insights into methylation patterns.

Limitations and Alternative Interpretations

During the study's execution, a primary limitation arose in feature selection. While k-mer features were initially considered alongside GC content and nucleotide composition, their computational expense and memory requirements for a dataset of this scale necessitated their exclusion. This limitation may have hindered the models' ability to detect significant sequence patterns associated with methylation.

Memory constraints also posed a significant challenge. The combination of the original dataset and the addition of negative samples resulted in a large dataset that proved demanding for certain models, including Random Forest and SVM. Despite various optimization techniques, such as PCA and dataset downsizing, these models failed to produce satisfactory results.

Finally, the class imbalance between positive and negative samples presented another obstacle. To address this, upsampling techniques were employed to balance the dataset. However, it is important to note that such techniques can potentially introduce bias and impact the overall performance of the models.

Future Research

Firstly, exploring additional features beyond GC content and nucleotide composition, such as biologically relevant motifs or DNA structural features, could provide valuable insights into methylation patterns. Secondly, advanced machine learning models, including convolutional or recurrent neural networks, hold promise for capturing intricate relationships within sequence data, especially long-range dependencies. These models could potentially outperform traditional methods like SGDClassifier and XGBoost.

Furthermore, expanding the analysis to a genome-wide scale would enable a comprehensive investigation of methylation patterns across diverse genomic regions, potentially

uncovering regulatory hotspots and elucidating genome-wide methylation dynamics. Integrating CpG methylation data with other epigenomic data, such as histone modifications, chromatin accessibility, and RNA-seq data, could provide a more holistic understanding of epigenetic regulation. Finally, exploring the impact of environmental factors on CpG methylation patterns in plants like Arabidopsis thaliana could reveal stress-responsive methylation mechanisms and inform strategies for developing more resilient crops.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. Retrieved from https://github.com/dmlc/xgboost

European Nucleotide Archive (ENA). (n.d.). Project PRJEB53882: Arabidopsis thaliana methylation data. Retrieved December 5, 2024, from https://www.ebi.ac.uk/ena/browser/view/PRJEB53882

Jones, P. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13, 484–492 (2012). https://doi.org/10.1038/nrg3230

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Willing, C. (2016). Jupyter Notebooks – A publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas (pp. 87–90). IOS Press. Retrieved from https://jupyter.org/

Law, J., Jacobsen, S. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11, 204–220 (2010). https://doi.org/10.1038/nrg2719

National Center for Biotechnology Information (NCBI). (n.d.). Genome assembly: Arabidopsis thaliana TAIR10.1. Retrieved December 5, 2024, from https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001735.4/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from https://scikit-learn.org/stable/

Thiebaut, F., Hemerly, A. S., & Ferreira, P. C. G. (2019). A role for epigenetic regulation in the adaptation and stress responses of non-model plants. Frontiers in Plant Science, 10, 246. https://doi.org/10.3389/fpls.2019.00246

Valente, A., Vieira, L., Silva, M. J., & Ventura, C. (2023). The Effect of Nanomaterials on DNA Methylation: A Review. Nanomaterials, 13(12), 1880. https://doi.org/10.3390/nano13121880

Zhang, L., & Li, Y. (2022). Structural insights into methylated DNA recognition by the methyl-CpG binding domain of Arabidopsis MBD5. ACS Omega, 7(1), 49-59. https://doi.org/10.1021/acsomega.1c04917