Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
NOV. 2024 VOL.3, NO.6

Predicting Yeast Chromatin Accessibility Based on DNA Sequence Features

Biyu Dong[1], Qiguo Zhang[1] & Zhi Zhang[1]

[1] School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou, China

Correspondence: Biyu Dong, School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou, China.

doi:10.56397/IST.2024.11.05

Abstract

The relationship between chromatin accessibility regions and DNA sequences represents a significant yet underexplored area of research. Supervised machine learning has emerged as an effective approach to elucidate this relationship. Most current predictions have focused on non-yeast organisms; however, in the field of synthetic biology, chromatin accessibility directly influences chromatin structure and the binding potential of regulatory proteins, which is crucial for enhancing production efficiency. In this study, we utilized ATAC-seq data from public databases specific to yeast. By combining the k-mer features of sequences from accessible regions with ensemble algorithm classifiers, we developed a predictive model for chromatin accessibility. Our model achieved an impressive AUC of 0.99, which holds promise for uncovering deeper insights into the mechanisms linking chromatin structure and DNA sequences.

Keywords: machine learning, chromatin accessibility, kmer, ATAC-seq

1. Introduction

1.1 Importance of Yeast Chromatin Structure in Synthetic Biology and Industrial Production

Saccharomyces cerevisiae, commonly known as baker's yeast, is a vital model organism in synthetic biology and serves as a key chassis in metabolic engineering. Whole-genome evolution is one of the most effective approaches for systematic modification in synthetic biology (Nielsen J & Keasling JD, 2016). Recent studies have increasingly linked chromatin accessibility to gene expression regulation and other biological processes. Understanding the factors influencing chromatin accessibility is crucial for comprehensively decoding the transcriptional regulatory network of baker's yeast and optimizing production processes. Furthermore, it provides a reference model for elucidating the complexities of cellular metabolism and regulatory networks (Rando OJ., 2012). In synthetic biology, modifying the chromatin structure of yeast can enhance the expression of specific genes, thereby improving production efficiency. For instance, altering the activity of chromatin remodeling complexes can significantly impact protein synthesis within yeast cells. Additionally, chromatin structure modifications are applied to enhance yeast performance in industrial production settings. By employing genome-scale modeling and proteomic constraints, the production of recombinant proteins in yeast can be optimized (Cheng L, et al., 2024). Research into yeast chromatin structure also encompasses genetic stability and the dynamic changes of chromosomes, which are crucial for gene editing and chromosome engineering in synthetic biology. Moreover, investigating the higher-order chromatin structures and their effects on gene expression holds significant potential for the integrated control of cellular functions within synthetic biology applications.

1.2 Relationship Between Yeast Chromatin Structure and Accessibility

The relationship between chromatin structure and accessibility in yeast is a critical aspect of gene expression

regulation. Chromatin accessibility refers to the degree of looseness in chromatin at specific regions, allowing transcription factors and other regulatory proteins to access DNA and thereby modulate gene expression. In model organisms like Saccharomyces cerevisiae, investigating chromatin structure is essential for understanding the mechanisms of gene regulation and expression. In yeast, particularly in Saccharomyces cerevisiae, chromatin structure is closely linked to the regulation of gene expression and the overall accessibility of the genome. The chromatin structure comprises nucleosomes and other protein complexes, which determine DNA accessibility and subsequently influence the activity of gene expression. Various chromatin remodeling complexes, such as SWI/SNF and RSC, play significant roles in this process by altering nucleosome positioning or removing nucleosomes to expose or re-hide DNA, thus regulating gene accessibility and expressibility. For example, the SWI/SNF complex enhances DNA accessibility by facilitating nucleosome sliding or eviction, which aids in the binding of transcription factors and the initiation of gene transcription (Jian Y, Shim WB & Ma Z., 2021). Furthermore, the positioning and density of nucleosomes in yeast are crucial for gene accessibility. Promoter regions that are nucleosome-depleted are more likely to be recognized and bound by transcription factors, promoting gene expression. Conversely, if the promoter region is covered by nucleosomes, it can inhibit gene activation (Kaplan N, et al., 2009). Histone modifications, such as acetylation, methylation, and phosphorylation, also regulate chromatin structure and accessibility. For instance, acetylation of lysine 9 and 14 on histone H3 (H3K9ac and H3K14ac) is commonly associated with an open chromatin structure and active gene expression. In contrast, deacetylation of histones can lead to chromatin compaction, suppressing gene expression (Pokholok DK, et al., 2005; Robyr D, et al., 2002). Additionally, certain DNA sequences may inherently favor a loose nucleosome configuration due to their physical properties, which could facilitate histone disassembly or attract specific transcriptional co-factors.

1.3 Significance of Studying Chromatin Accessibility in Yeast

Research on chromatin accessibility in yeast holds substantial scientific and practical significance. As a model organism, yeast has a simple genetic background and has been extensively studied, making it an ideal tool for investigating fundamental biological processes. The following points highlight the main significance of studying chromatin accessibility in yeast:

Gene Regulation: Chromatin accessibility is crucial for understanding gene regulatory mechanisms. By investigating how chromatin opens or closes under various conditions, scientists can elucidate the molecular mechanisms that control gene expression (Rando OJ & Winston F., 2012).

Transcription Regulation: The state of chromatin directly influences the accessibility of transcription factors and RNA polymerase to genes, thereby affecting mRNA production and protein synthesis (Rando OJ & Winston F., 2012).

Genetic Engineering: By manipulating chromatin accessibility, researchers can more precisely regulate the expression of target genes in genetically modified yeast. This capability is vital for the production of pharmaceuticals, enzymes, and other industrially relevant products (Rando OJ & Winston F., 2012).

Fermentation Engineering: In the production of alcohol, biofuels, and other biochemical products, optimizing the chromatin state in yeast can enhance metabolic capabilities and improve production efficiency (Rando OJ & Winston F., 2012).

Environmental Adaptation Studies: Yeast can adjust its chromatin structure in response to environmental changes, such as variations in temperature, pressure, and nutrient availability. This adaptability provides valuable insights into how environmental stressors influence cellular function.

Genetic Diversity and Evolution: Investigating how chromatin accessibility affects gene expression and adaptive evolution can help scientists understand how different species adapt to their environments and how these adaptive traits evolve within populations (Galdieri L, Mehrotra S, Yu S & Vancura A., 2010).

The study of chromatin accessibility in yeast not only deepens our understanding of the complex mechanisms of gene regulation within cells but also fosters advancements across multiple fields, from basic science to applied technology. This research illustrates the broad connections between model organisms and their implications for human health and disease treatment.

2. Research Background

2.1 Significance of Chromatin Accessibility in the Study of Saccharomyces Cerevisiae

Saccharomyces cerevisiae, commonly known as baker’s yeast, is a crucial microorganism in the fields of brewing and biotechnology. Chromatin accessibility is particularly significant in yeast research, as it relates to how yeast regulates gene expression in response to various environmental stresses, thereby influencing its growth and metabolic performance. Optimization of Fermentation Performance: Chromatin accessibility can impact the expression of specific genes involved in critical steps of the brewing process, such as sugar utilization

and the production of alcohol and carbon dioxide. By understanding and controlling the expression of these genes, scientists and brewers can optimize brewing processes to enhance yield and product quality. For instance, modifying the activity of chromatin remodeling complexes can boost the expression of targeted genes, leading to increased production efficiency. Moreover, primary metabolites produced by yeast during fermentation, such as glycerol, succinate, acetate, and lactate, significantly influence the quality of wine. The production of these metabolites is closely linked to yeast metabolic pathways. Therefore, employing synthetic biology techniques, such as the CRISPR-Cas9 system, allows for efficient gene editing to optimize yeast metabolic pathways, enhancing the efficiency and quality of the brewing process (Tirosh I, Reikhav S, Levy AA & Barkai N., 2009). During fermentation, yeast encounters various stress conditions, including high osmotic pressure, low oxygen environments, and alcohol toxicity. The dynamic adjustment of chromatin structure enables yeast to rapidly adapt to these environmental changes, modulating its physiological state to ensure survival and sustained metabolic activity. Studying this adaptive process is invaluable for enhancing the environmental resilience and robustness of yeast. Furthermore, chromatin accessibility is associated with genetic stability. Open chromatin is more susceptible to mutations and recombination events, which can lead to genetic variation. Understanding these mechanisms can aid in controlling undesirable variations while maintaining the desired genetic traits, thereby ensuring the consistency and predictability of the brewing process.

2.2 Applications of Machine Learning in Predicting Chromatin Accessibility

The application of machine learning in predicting chromatin accessibility within biological processes is a vibrant research direction in the fields of genomics and bioinformatics (DiCarlo JE, et al., 2013; Libbrecht MW & Noble WS., 2015; Singh A, Ganapathysubramanian B, Singh AK & Sarkar S., 2016). Researchers at Tsinghua University have developed machine learning methods for predicting chromatin accessibility. They introduced a random forest method, kmerForest, which integrates genomic sequences with evolutionary conservation, and a mixed convolutional neural network model, Deopen. These methods achieve binary classification and continuous regression of chromatin accessibility signals, demonstrating superior predictive performance compared to existing methods and facilitating the analysis of genetic data (Leung MK, Xiong HY, Lee LJ & Frey BJ., 2014). In another study, a dense convolutional network model, DeepCAGE, was proposed to predict chromatin accessibility across different cell lines by integrating genomic annotations and transcriptomic data. This model leverages existing biological priors to effectively enhance predictive performance and further establishes a method for analyzing complex phenotypic associations with genetic factors based on chromatin accessibility (Chen, X., Chen, S., Song, S. et al., 2022). Additionally, a smart prediction model called SMOC (Smart Model for Open Chromatin region prediction) was developed to predict open chromatin regions within the rice genome. This model trains and predicts using chromatin accessibility data through machine learning techniques, offering new insights for the identification and information mining of open chromatin regions (Liu Q, Hua K, Zhang X, Wong WH & Jiang R., 2022). C. Origami has introduced a novel multimodal machine learning model aimed at predicting chromatin conformations specific to certain cell types. Based on the principles of genetic screening, this model also proposes a new high-throughput computational genetic screening (in silico genetic screening, ISGS) method to identify cell-type-specific functional genomic elements, thereby facilitating the discovery of new mechanisms regulating chromatin conformations (Guo W, et al., 2022). These studies highlight the increasingly important role of machine learning technologies in predicting chromatin accessibility, contributing to a deeper understanding of the regulatory mechanisms governing gene expression and providing new tools and methodologies for biomedical research.

2.3 Limitations and Challenges in Current Research

Current research on the relationship between chromatin accessibility and the genome predominantly involves specific wet-lab experiments. Although several studies utilizing machine learning or deep learning approaches to investigate chromatin accessibility have been cited, they are largely limited to human cells or species such as rice. In particular, while there are existing studies predicting chromatin accessibility based on human cells (Tan J, et al., 2023), there has been a notable absence of similar research in microorganisms. This gap is especially critical given the importance of yeast as a cell factory in industrial production and synthetic biology. The lack of comprehensive research on yeast hinders scientists' understanding of whole-genome evolution and systematic modification strategies for this crucial synthetic biology chassis.

2.4 Innovation and Objectives of the Research

This study's innovation lies in the scarcity of research predicting chromatin accessibility based on DNA sequences in yeast, with most existing studies relying heavily on experimental methods. Regions such as promoters and enhancers contain numerous transcription factor binding sites, which regulate gene expression through interactions with specific transcription factors. As such, these areas are potential candidates for chromatin accessibility. The relationship between chromatin open regions and various biological processes, including gene expression, is complex and multifaceted. By applying advanced machine learning algorithms to

biological sequence data, this research aims to uncover intricate patterns and relationships that traditional methods struggle to identify. Utilizing large-scale genomic datasets that encompass chromatin states under various conditions will enhance the accuracy of predictions. In our approach, we will develop novel feature extraction methods that transform DNA sequence information into formats amenable to machine learning models, thus improving model performance. Furthermore, enhancing the interpretability of these models will allow researchers to understand which sequence features significantly influence chromatin accessibility, thereby facilitating deeper biological insights. We have constructed a machine learning model based on the ATAC-seq peak interval sequences of Saccharomyces cerevisiae to predict chromatin accessibility in yeast. This model leverages information about chromatin accessibility and incorporates sequence features to achieve more efficient and accurate predictions of gene expression. Through this approach, we aim to establish a reliable predictive model that accurately forecasts the open state of chromatin in yeast based on DNA sequence information and identifies critical DNA sequence features associated with chromatin accessibility, providing a foundation for further experimental investigations. By analyzing the prediction results, we seek to gain deeper insights into the regulatory mechanisms influencing chromatin accessibility. Ultimately, we aim to elucidate how chromatin structure impacts chromatin activity and provide a powerful tool for future functional genomics research, as well as offer new methodologies and perspectives for subsequent genomic regulation studies. This research not only contributes to a more profound understanding of yeast biology but also aids in the development of general predictive models that can advance research in other areas.

3. Materials and Methods

We obtained the ATAC-seq dataset for the wild-type Saccharomyces cerevisiae strain CEN, generated and processed by Gowans and colleagues, from a public database (GEO accession number GSE101290) (Natarajan A, et al., 2012). This dataset includes measurements of chromatin accessibility at different developmental stages. Samples were collected at two time points for each stage: (i) early and mid-rc, (ii) late ox, and (iii) early and late rb. Each time point has two biological replicates, and we used the average of the replicates for subsequent analyses. Initially, we performed an analysis of the open chromatin data based on these sequencing results to identify open regions within the chromatin and to determine the DNA sequences of these regions, preparing the data for further modeling.

3.1 Pre-Modeling Analysis of ATAC-seq Data

We began by installing the necessary software for ATAC-seq analysis and conducting a quality assessment of the downloaded sequencing data. This assessment included checks for sequencing error rates, duplicate content, and GC content, which are essential for removing low-quality reads to enhance the accuracy and reliability of the data analysis. Next, we mapped the sequencing reads to the reference genome and aligned the reads, allowing us to determine the genomic locations of each read. This alignment is crucial for analyzing chromatin accessibility. We employed the peak detection tool MACS3 to analyze the aligned data, identifying regions of read enrichment that typically correspond to open chromatin regions.

3.2 Data Preparation

The core input for the predictive model is based on k-mer analysis of sequences. In bioinformatics, a k-mer refers to a contiguous subsequence of length k bases (or amino acids) within a biological sequence. For a nucleic acid sequence of length L, sliding a window of one base can generate a total of (L-k+1) k-mers. Additionally, the reverse complement of the nucleic acid sequence can be used to generate a second set of k-mers. This algorithm is widely utilized in genomics and proteomics, serving as a fundamental analytical unit that offers various advantages (Gowans GJ, et al., 2018; Breitwieser FP, Lu J & Salzberg SL., 2019) and is particularly effective in capturing local features.

We calculated k-mer frequencies ranging from single nucleotides to six nucleotides, specifically for k values from 1 to 6. To systematically process and analyze this data, we compiled all k-mer frequency statistics into a well-structured dataset. This dataset not only contains frequency counts for k-mers of varying lengths but also categorizes data types. Specifically, we extracted 1-6 mer information separately from open regions and non-open regions to construct the dataset. The k-mer frequency information from open regions was defined as the positive set, while that from non-open regions was defined as the negative set.

k-mer Frequency Calculation Method:

K-mers can be understood as representing the relative dependencies of symbols within different sequences (Leggett RM, et al., 2013). For a sequence $S = (s_1, s_2, \dots s_n)$, of length n, generated from an alphabet {A, T, C, G}, the overall probability of the sequence is expressed by starting with the initial probability of the first character and progressively calculating the conditional probabilities of the subsequent characters. This approach relies on the first-order Markov model formula:

$$P(x_i|x_{i-1}, x_{i-2}, \dots x_1) = P(x_i|x_{i-1}) \quad (1)$$

Calculating the frequencies of all distinct k-mers within biological sequences is a critical step in many bioinformatics applications.

3.3 Dataset Partitioning

To construct a sample set for model training, we developed rigorous criteria for the classification of positive and negative samples. Positive samples were defined based on known open chromatin regions identified in the ATAC-seq analysis. Given the variability in peak widths of accessible chromatin regions, we selected the peak point of open intervals as the center and extracted sequences of 300 bp upstream and downstream, resulting in DNA sequences of 600 bp labeled as positive samples (1). This strategy is supported by extensive prior research, indicating that sequences within this range are representative of local or concentrated chromatin accessibility signals. Consequently, we collected a total of 15,762 positive samples corresponding to each identified open chromatin region.

For the construction of the negative sample set, we employed a scientifically sound approach to ensure the authenticity and diversity of the negative samples, which is critical for developing a binary classification model for chromatin accessibility prediction. Negative samples were primarily sourced from regions of the genome identified as non-open chromatin based on ATAC-seq data. Similar to the selection of positive samples, we extracted 600 bp sequences from these regions, labeling them as negative samples (0) to represent a non-open state. Importantly, we ensured that these negative regions were at least 500 bp away from known open regions to minimize potential signal overlap and confusion. This approach yielded an equal number of negative samples, creating a balanced learning environment for the model.

To maintain the balance between positive and negative samples and ensure the effectiveness of model training, we determined that the number of negative samples would be 15,808, achieving an approximate 1:1 ratio with positive samples. Throughout the sample selection process, we prioritized maintaining this balance to avoid skewed learning in the model. Additionally, all sequences underwent stringent quality filtering to eliminate low-complexity sequences, duplicate sequences, and potential contaminants, thereby ensuring the purity and reliability of the final dataset.

Through these carefully designed data processing steps, we ensured that the input data for the model was of high quality and representative, laying a solid foundation for subsequent predictive model development. These efforts not only deepen our understanding of yeast gene regulatory mechanisms but also provide a transferable methodological framework for predicting similar expression regulations in other organisms.

3.4 Selection of Classifiers

Based on prior knowledge and data sensitivity, we selected two classifiers — Random Forest (RF) and Support Vector Machine (SVM) — for modeling. Each of these classifiers offers distinct advantages, providing a diverse approach to prediction. Random Forest (RF) is an ensemble learning method based on decision trees, known for its robustness against overfitting, especially in handling high-dimensional data and capturing non-linear relationships. RF enhances classification accuracy through a voting mechanism across multiple decision trees, making it less sensitive to noise in the data. Given the complex and potentially non-linear features within chromatin accessibility data, RF's ensemble characteristics allow it to better capture these intricate patterns, providing robust predictive performance. Support Vector Machine (SVM), on the other hand, finds an optimal hyperplane that maximizes the margin between classes, making it effective for classification in high-dimensional spaces, particularly when dealing with limited high-dimensional data. With kernel methods, SVM can effectively model non-linear relationships. In chromatin accessibility prediction, complex boundaries may exist between categories, and SVM's ability to find an optimal separating hyperplane allows it to handle such boundary complexities effectively. SVM is particularly suited for medium to small datasets, aligning well with the sample size in this study.

4. Results and Discussion

To investigate the characteristics of DNA sequences within chromatin-accessible regions, we developed a method based on ATAC-seq data from wild-type yeast strains at six different time points during the yeast metabolic cycle (YMC). We conducted chromatin accessibility analysis (peak calling) at each time point, followed by an examination of k-mer sequence frequencies within specific ranges across different categories. This approach aims to reveal potential associations between varying chromatin accessibility states and k-mer sequence frequencies. To explore the global relationship between chromatin accessibility and DNA sequence features, we aggregated the k-mer feature data from the different time points. We then used stratified sampling to split the data into two parts: 80% was allocated to the training set for model construction and parameter optimization, while the remaining 20% served as the test set for evaluating the predictive performance of the model. To ensure high predictive accuracy and practical robustness, we employed 5-fold cross-validation.

Additionally, upon comparing input data for k-mers of lengths 1 to 6, we observed that the model performed best when using 5-mer sequences.

4.1 Chromatin Accessibility Analysis of Wild-Type Strains at Different Stages (Peak Calling)

Peak calling analysis of ATAC-seq data from wild-type strains at six distinct time points of the yeast metabolic cycle (YMC) revealed that open chromatin regions are predominantly concentrated around transcription start sites (TSS) (Figure 1). This finding underscores the central role of open chromatin regions near transcriptional start sites in gene regulation and provides important insights for further functional genomics studies.

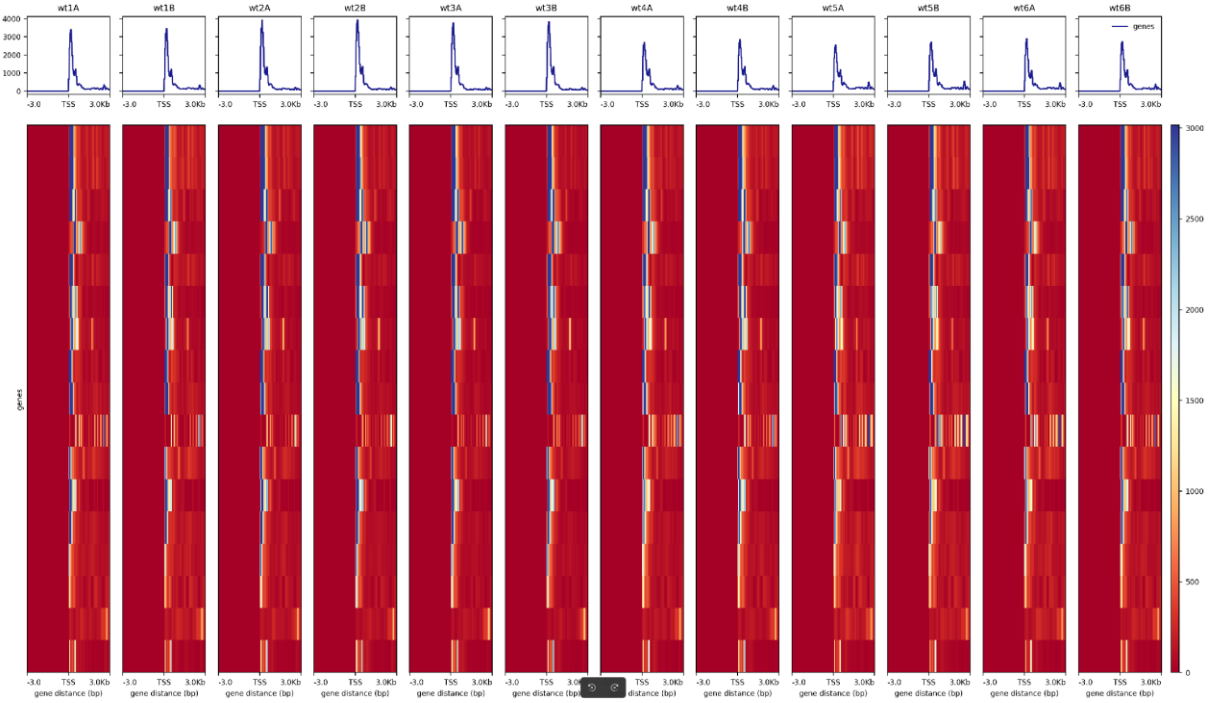

Figure 1.

Figure 1. The heatmap and line plots illustrate the significant enrichment of open chromatin regions (peak calling) at the six time points. Each time point is represented by two biological replicates.

4.2 Prediction of Chromatin Accessibility Based on DNA Sequences

Through chromatin accessibility analysis, we identified open and non-open regions across the genome. Subsequently, we analyzed the observed frequencies of k-mer sequences within defined ranges for each category, aiming to uncover potential associations between different chromatin states and k-mer sequence frequencies. To construct and evaluate the predictive model, we combined the positive (open regions) and negative (non-open regions) datasets and performed a random sampling split: 80% of the data was allocated to the training set for model construction and parameter optimization, while the remaining 20% was reserved as a test set to assess the model's predictive performance. To ensure high predictive accuracy and practical robustness, we employed a 10-fold cross-validation approach. Additionally, after comparing k-mer inputs ranging from 1 to 6, we found that the model achieved the best predictive performance with 5-mer sequences.

4.3 Prediction Results from Two Classifiers

To comprehensively evaluate the performance of different algorithms in our prediction task, we trained and tested models using two classifiers: Random Forest (RF) and Support Vector Machine (SVM). Among these, the Random Forest classifier demonstrated the best performance, achieving an Area Under the Curve (AUC) value of 0.99 on the test set, indicating high classification accuracy and excellent predictive ability. The SVM classifier also performed well, with an AUC value of 0.97. Figure 2A provides a detailed comparison of the classifiers' performance in predicting chromatin accessibility, showing strong results for both. To further assess model performance, we computed key evaluation metrics on both the training and test sets. A thorough evaluation of binary classification models is critical to ensure generalization, robustness, and reliable predictive power. By examining multiple performance metrics — such as accuracy, sensitivity, specificity, and F1-score — we gain a

comprehensive understanding of the models' capabilities. Overall, both RF and SVM models performed particularly well on the test set.

Specifically, the Random Forest classifier exhibited excellent and consistent performance across all metrics on the test set. Its accuracy reached 97%, with F1-score, precision, sensitivity, and specificity values of 97%, 96%, 98%, and 96%, respectively (Figure 2B). This strong performance likely stems from RF's ability to handle complex feature interactions effectively, particularly for capturing non-linear relationships in high-dimensional data. In this study, we used features such as k-mer frequencies related to gene expression, which are inherently high-dimensional and non-linear, aligning well with the strengths of the RF model. Additionally, RF mitigates overfitting risks by aggregating numerous decision trees, thus enhancing the model's generalizability.

In comparison, the Support Vector Machine classifier, while performing slightly below RF, also achieved a high AUC of 0.97 and performed well on other metrics, albeit slightly lower than those of RF. SVM excels at finding an optimal hyperplane to maximize the margin between classes, making it particularly effective when data are approximately linearly separable in the feature space. Even in cases where the data are not fully linearly separable, SVM can achieve effective non-linear separation through kernel methods, such as the Radial Basis Function (RBF) kernel. This allows SVM to perform well in high-dimensional spaces with non-linear relationships, though it may not capture complex feature interactions as effectively as RF in this study.

Furthermore, the confusion matrix provides a clear visualization of the model's classification performance, enabling the assessment of various classification metrics and the identification of any class imbalance issues. It serves as an indispensable evaluation tool in binary classification models, offering valuable insights for model optimization (Figure 2C).

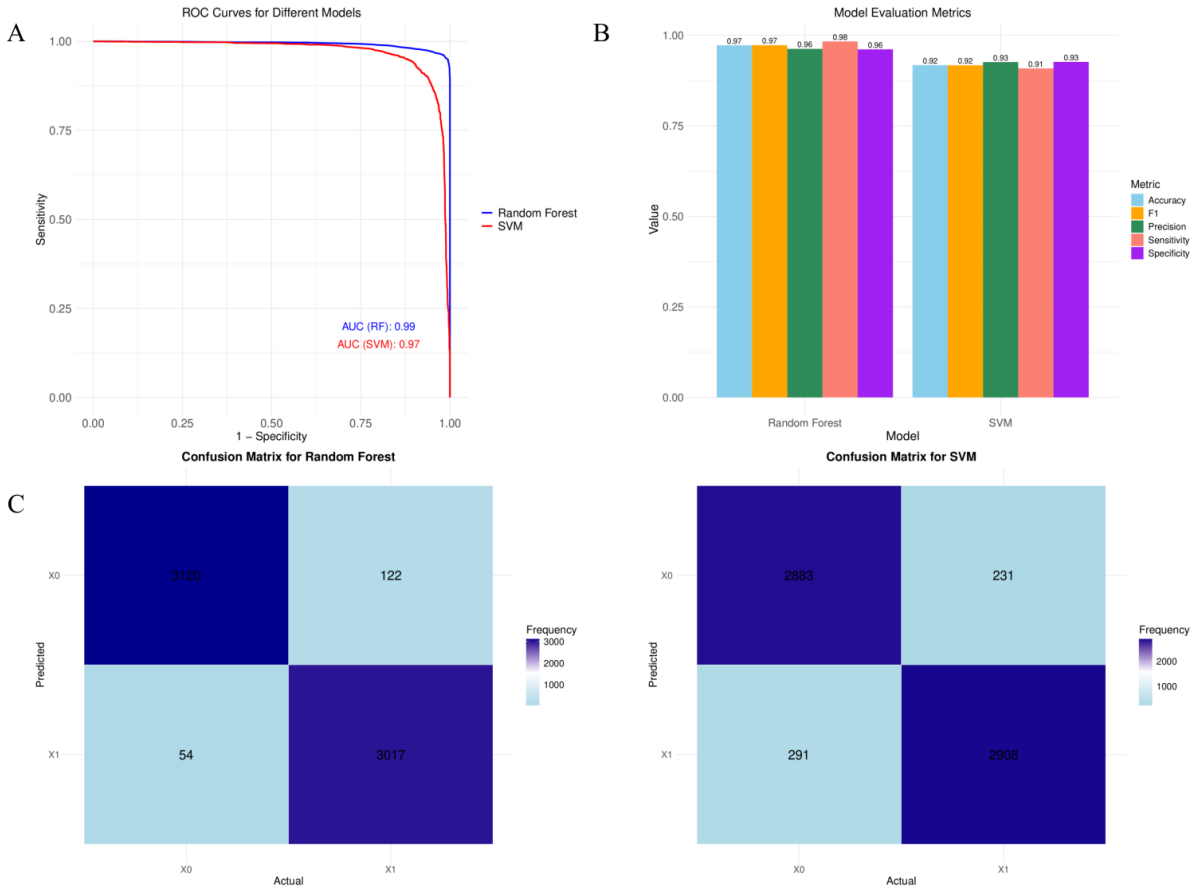

Figure 2.

A: Presentation of Prediction Results from Two Classifier Models

B: Display of Evaluation Metrics for Two Classification Models

C: Confusion Matrix Visualization for the Two Models

4.4 Feature Importance Analysis

While complex models like Random Forest exhibit high predictive accuracy, they are often regarded as "black-box" models due to the difficulty in directly interpreting their internal mechanisms. In this study, our goal extends beyond achieving accurate predictions; we aim to understand how specific features contribute to the model's predictions, as this insight can help uncover underlying biological mechanisms.

Feature importance analysis allows us to identify which sequence features make the most substantial contributions to the prediction of chromatin accessibility regions (Figure 3A). This not only aids in selecting the most biologically meaningful features but also facilitates the generation of new biological hypotheses.

Furthermore, the use of Local Interpretable Model-Agnostic Explanations (LIME) provides interpretability at both the individual and global prediction levels (Figure 3B). LIME helps elucidate why the model makes certain predictions, ensuring that the model's predictions are not only accurate but also understandable to researchers. This transparency enhances trust in the model by providing insights into the reasoning behind its predictions.

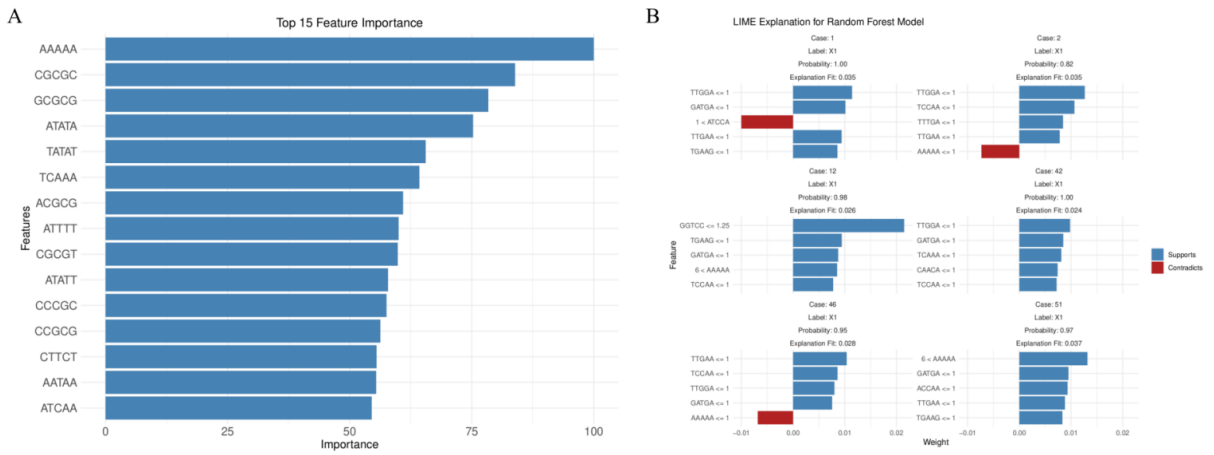

Figure 3.

A: Display of the Top 15 Features in Model Importance Analysis

B: Explanation of the Top 6 Test Data Examples Using LIME Analysis

4.5 Discussion

In our modeling approach, we balanced multiple factors, including model predictive performance and training time. After comprehensive consideration, we selected 5-fold cross-validation for model validation, set the ntree parameter of the Random Forest (RF) model to 130, and configured the tuneLength parameter of the Support Vector Machine (SVM) model to 5. In scenarios where the dataset size is smaller, or if further improvement in predictive performance is desired without concern for increased training time, a 10-fold cross-validation could be employed. Additionally, increasing the training parameters for both models may further optimize predictive accuracy. These adjustments could potentially enhance the overall model performance.

References

Breitwieser FP, Lu J, Salzberg SL., (2019). A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform*, *20*(4), 1125-1136. doi: 10.1093/bib/bbx120. PMID: 29028872; PMCID: PMC6781581.

Chen, X., Chen, S., Song, S. et al., (2022). Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. *Nat Mach Intell*, *4*, 116-126. https://doi.org/10.1038/s42256-021-00432-w

Cheng L, Zhao S, Li T, Hou S, Luo Z, Xu J, Yu W, Jiang S, Monti M, Schindler D, Zhang W, Hou C, Ma Y, Cai Y, Boeke JD, Dai J., (2024). Large-scale genomic rearrangements boost SCRaMbLE in Saccharomyces cerevisiae. *Nat Commun*, *15*(1), 770. doi: 10.1038/s41467-023-44511-5. PMID: 38278805; PMCID: PMC10817965.

DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM., (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res*, *41*(7), 4336-43. doi: 10.1093/nar/gkt135. Epub 2013 Mar 4. PMID: 23460208; PMCID: PMC3627607.

Galdieri L, Mehrotra S, Yu S, Vancura A., (2010). Transcriptional regulation in yeast during diauxic shift and stationary phase. *OMICS*, *14*(6), 629-38. doi: 10.1089/omi.2010.0069. Epub 2010 Sep 23. PMID: 20863251; PMCID: PMC3133784.

Gowans GJ, Schep AN, Wong KM, King DA, Greenleaf WJ, Morrison AJ., (2018). INO80 Chromatin Remodeling Coordinates Metabolic Homeostasis with Cell Division. *Cell Rep*, *22*(3), 611-623. doi: 10.1016/j.celrep.2017.12.079. PMID: 29346761; PMCID: PMC5949282.

Guo W, Liu H, Wang Y, Zhang P, Li D, Liu T, Zhang Q, Yang L, Pu L, Tian J, Gu X., (2022). SMOC: a smart model for open chromatin region prediction in rice genomes. *J Genet Genomics*, *49*(5), 514-517. doi: 10.1016/j.jgg.2022.02.012. Epub 2022 Feb 28. PMID: 35240305.

Jian Y, Shim WB, Ma Z., (2021). Multiple functions of SWI/SNF chromatin remodeling complex in plant-pathogen interactions. *Stress Biol*, *1*(1), 18. doi: 10.1007/s44154-021-00019-w. PMID: 37676626; PMCID: PMC10442046.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E., (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, *458*(7236), 362-6. doi: 10.1038/nature07667. Epub 2008 Dec 17. PMID: 19092803; PMCID: PMC2658732.

Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, Waite D, Davey RP., (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front Genet*, *4*, 288. doi: 10.3389/fgene.2013.00288. PMID: 24381581; PMCID: PMC3865868.

Leung MK, Xiong HY, Lee LJ, Frey BJ., (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics*, *30*(12), i121-9. doi: 10.1093/bioinformatics/btu277. PMID: 24931975; PMCID: PMC4058935.

Libbrecht MW, Noble WS., (2015). Machine learning applications in genetics and genomics. *Nat Rev Genet*, *16*(6), 321-32. doi: 10.1038/nrg3920. Epub 2015 May 7. PMID: 25948244; PMCID: PMC5204302.

Liu Q, Hua K, Zhang X, Wong WH, Jiang R., (2022). DeepCAGE: Incorporating Transcription Factors in Genome-wide Prediction of Chromatin Accessibility. *Genomics Proteomics Bioinformatics*, *20*(3), 496-507. doi: 10.1016/j.gpb.2021.08.015. Epub 2022 Mar 12. PMID: 35293310; PMCID: PMC9801045.

Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U., (2012). Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res*, *22*(9), 1711-22. doi: 10.1101/gr.135129.111. PMID: 22955983; PMCID: PMC3431488.

Nielsen J, Keasling JD, (2016). Engineering Cellular Metabolism. *Cell*, *164*(6), 1185-1197. doi: 10.1016/j.cell.2016.02.004. PMID: 26967285.

Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA., (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, *122*(4), 517-27. doi: 10.1016/j.cell.2005.06.026. PMID: 16122420.

Rando OJ, Winston F., (2012). Chromatin and transcription in yeast. *Genetics*, *190*(2), 351-87. doi: 10.1534/genetics.111.132266. PMID: 22345607; PMCID: PMC3276623.

Rando OJ., (2012). Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr Opin Genet Dev*, *22*(2), 148-55. doi: 10.1016/j.gde.2012.02.013. Epub 2012 Mar 20. PMID: 22440480; PMCID: PMC3345062.

Robyr D, Suka Y, Xenarios I, Kurdistani SK, Wang A, Suka N, Grunstein M., (2002). Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell*, *109*(4), 437-46. doi: 10.1016/s0092-8674(02)00746-8. PMID: 12086601.

Singh A, Ganapathysubramanian B, Singh AK, Sarkar S., (2016). Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends Plant Sci*, *21*(2), 110-124. doi: 10.1016/j.tplants.2015.10.015. Epub 2015 Dec 1. PMID: 26651918.

Tan J, Shenker-Tauris N, Rodriguez-Hernaez J, Wang E, Sakellaropoulos T, Boccalatte F, Thandapani P, Skok J, Aifantis I, Fenyö D, Xia B, Tsirigos A., (2023). Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat Biotechnol*, *41*(8), 1140-1150. doi: 10.1038/s41587-022-01612-8. Epub 2023 Jan 9. PMID: 36624151; PMCID: PMC10329734.

Tirosh I, Reikhav S, Levy AA, Barkai N., (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, *324*(5927), 659-62. doi: 10.1126/science.1169766. PMID: 19407207.

Yoon BJ., (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics*, *10*(6), 402-15. doi: 10.2174/138920209789177575. PMID: 20190955; PMCID: PMC2766791.