JOURNAL OF
EMERGING INVESTIGATORS

Prediction of Age-Related Gene Expression in Human Hematopoietic Stem and Progenitor Cells using Machine Learning

Hannah Thomas[1], Aristeidis G. Telonis[2,3]

[1] Nashua High School South, Nashua, NH.
[2] Department of Biochemistry and Molecular Biology, University of Miami Miller School of Medicine, Miami, FL.
[3] Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL.

Student Authors
Hannah Thomas, High School

KEYWORDS: Gene Expression, Machine Learning, Leukemia, Prediction

OVERVIEW: The human aging process involves extensive epigenetic reprogramming, leading to significant changes in gene expression, which can predispose to leukemia. Our method to predict the direction of expression change with age provides a novel perspective on age-related gene regulation by identifying motifs that can be used as biomarkers to evaluate biological aging and potentially the risk of developing leukemia.

SUMMARY

The human aging process involves extensive epigenetic reprogramming, leading to significant changes in gene expression, which can predispose to leukemia. To enhance the understanding of aging at the molecular level, we examined the potential of motifs within gene bodies to predict age-related expression changes in human hematopoietic stem and progenitor cells (HSPCs). We extracted all k-mer motifs present in the differentially expressed genes between young and aged individuals and trained a support vector classifier (SVC) to predict the direction of expression change with age. Our analysis showed that datasets with 5,000 8-mer motifs predicted gene regulation with an accuracy of 81.8%. Our findings provide a novel perspective on age-related gene regulation by identifying motifs that can be used as biomarkers to evaluate biological aging and potentially the risk of developing leukemia.

INTRODUCTION

Aging is a natural process involving gradual changes throughout the body, and it is known to have several negative impacts on health, including an increased risk of disease, physical decline, sensory impairments, and cognitive deterioration. From a cellular perspective, the human aging process leads to a decline in the function of hematopoietic stem and progenitor cells (HSPCs). Adelman et al. investigated the epigenomic and transcriptomic alterations in HSPCs during aging (Adelman, 2019). They collected bone marrow mononuclear cells from young and aged donors without a history of hematologic cancer and performed Chromatin immunoprecipitation sequencing (ChIP-seq) to find statistically significant changes in gene regulation. Their study led to several new findings. Notably, they discovered that differentially methylated regions observed in aged HSPCs were also present in both young and elderly patients with acute myeloid leukemia (AML). This suggests that the differential gene expression in aged HSPCs may contribute to a predisposition to myeloid malignancies.

Identifying specific architectural parameters that characterize the genes undergoing epigenetic reprogramming could enhance our understanding of the role genetics plays in aging. Existing literature has linked gene length to a transcriptome imbalance, with some studies exploring this relationship in genes associated with life expectancy (Stoeger, 2022). Our research builds upon these findings by exploring additional characteristics driving these changes in aged HSPCs. In particular, we examined the correlation between certain gene motifs and changes in gene expression in HSPCs, which has not yet been explored in the scientific literature. Moreover, we demonstrated that this correlation could be used to predict the regulation of genes differentially expressed with age.

Machine learning has become an essential tool in biology, with broad applications in genomics, proteomics, and other data analysis fields. In this study, we used a support vector machine (SVM) algorithm that takes specific motifs as features and uses them to classify a differentially expressed gene as up-regulated or down-regulated in aged HSPCs. This research offers new insights into the relationship between gene architecture and age-related gene regulation.

RESULTS

Testing for Different Motif Lengths

We hypothesized that the genes differentially expressed with aging contain motifs that can be used to predict if they are up- or down-regulated. Tables 1 and 2 summarize the results from the experiments using the input datasets for each of the seven tested motif lengths. These analyses were conducted with the top 500 motifs sorted by the difference in percentage occurrence between up-regulated and down-regulated genes. We calculated the mean accuracy from 10-fold cross-validation, the area under the receiver operating characteristic (ROC) curve, and additional performance metrics, such as sensitivity, specificity, and false discovery rate (FDR). Accuracy represents the overall correctness of the model. Sensitivity is a measure of how well the model identifies true positives. Specificity is a measure of how well the test identifies true negatives. False Discovery Rate (FDR) measures the rate at which false positives are classified among all the positive results returned by the test. The ROC curve is a graphical representation of the trade-off between sensitivity (Y-axis) and specificity (X-axis). The area under the ROC curve (AUC) represents the probability that the model will rank a randomly chosen positive data point higher than a randomly chosen negative data point.

Accuracy = (True Positives + True Negatives)/(Total Predictions)
Sensitivity = (True Positivies)/(True Positives + False Negatives)
Specificity = (True Negatives)/(True Negatives + False Positives)
FDR = (False Positivies)/(True Positives + False Positives)

We observed that experiments with motif lengths of 8 nucleotide bases produced the best accuracy and AUC metrics.

Testing for Different Numbers of Motifs

We conducted additional tests using motifs of length 8 nucleotide bases, as they produced the highest accuracy among the seven tested lengths. We hypothesized that model accuracy depends on the number of motif features used in the model. We ran experiments with 200, 500, 1,000, 5,000, and 10,000 motifs, and the results are tabulated in Tables 3 and 4. We observed that experiments with 5,000 motifs produced the best accuracy and AUC metrics.

As shown in Tables 2 and 3, experiments with motif features indicating the presence/absence of the motif compared to the frequency of occurrence of the motif produced similar results. We hypothesized that this was due to the low frequency of occurrence of motifs in each gene. We found that the average number of motif occurrences per gene across all motif lengths is 1.14, which supports our hypothesis.

Tables 5, 6 and 7 list the top 25 motifs by the absolute difference of occurrence percentage between up-regulated and down-regulated genes, of motif lengths 8, 9 and 10 respectively.

SHAP Summary Plots

Figure 1 shows summary plots from the SHAP analysis we conducted. These plots display the distribution of SHAP values for each of the top 10 most heavily weighted motifs in both models with 5000 8-nucleotide motifs. These motifs are the most significant in classifying age-related differential gene expression.

DISCUSSION

In this study, we observed that the presence/absence of certain motifs in a differentially expressed gene's cDNA can accurately predict its change in regulation with respect to age. The best accuracy was observed using 5,000 motifs that are 8 nucleotide bases long. Additionally, we identified the ten most significant motifs in predicting the direction of a differentially expressed gene's regulation in HSPCs. To the best of our knowledge, no similar studies have been reported in the literature that explore the correlation between gene motifs and age-related gene regulation in human HSPCs, highlighting the novelty of our work.

We conclude that the presence/absence of these motifs in cDNA sequences may influence gene expression. Additionally, we infer that running our model on the entire human genome could find other genes that may be differentially expressed with age in HSPCs. We plan to do that in future work. These conclusions can be further tested in future laboratory research and, if corroborated, have the potential for significant applications in oncology and human aging. Previous studies have linked age-related epigenetic reprogramming to a predisposition to myeloid leukemia. Given that our SVC model is able to predict epigenetic changes in HSPCs, the motif features we identified through SHAP analysis could serve as biomarkers for gene regulation that may possibly lead to myeloid malignancies. Further, these findings could contribute to the development of cancer-preventative treatments, such as gene therapy, that target the motifs most heavily weighted in our model.

Our findings invite further investigation into the relationship between motifs in protein-coding regions and changes in HSPC gene regulation. Future research could explore whether aging impacts the occurrence of motifs in cDNA sequences, as well as look into other gene architectural parameters that are correlated with or can be used in conjunction to predict age-related gene regulation in HSPCs and other cell types.

MATERIALS AND METHODS

Age-related Differential Gene Expression Data

We obtained Table S6 from the supplementary data of Adelman et al., which listed the gene identifiers of all differentially expressed genes in aged HSPCs compared to young HSPCs (Adelman, 2019). It also included additional details, such as the base-2 logarithm of the fold change (log2FoldChange), which quantifies the amount of change in gene expression level with aging. A positive log2FoldChange value indicates up-regulation, while a negative value indicates down-regulation. The dataset contained 517 up-regulated genes and 616 down-regulated genes. The sign of the log2FoldChange, reflecting the direction of regulation, served as the target variable for classification in our SVM model. Specifically, up-regulated genes were assigned a value of "1", while down-regulated genes were assigned a value of "0".

Sequence Data

The complementary DNA (cDNA) sequences were downloaded on August 30, 2024, from the Human Genes (GRCh38.p14) dataset in the “Ensembl Genes 113” database using the Ensembl BioMart data mining tool (https://useast.ensembl.org/biomart/martview). The sequences were organized by Gene Stable ID and Gene Stable ID version, which indicates the most recent version of the sequences. From this dataset, we extracted the sequences for 1,103 differentially expressed genes. This step was done with a program written in Python. We processed the sequences further and identified all motifs of a specific length (such as 7, 8, 9, etc.) present in the cDNA of the differentially expressed genes. For each motif, we calculated the number of occurrences across the 1,103 genes, the number of genes in which the motif appeared, and the distribution of these occurrences in up-regulated versus down-regulated genes. These metrics enabled us to determine for each motif, the percentage of genes the motif appeared in that were up-regulated and down-regulated, respectively (data included in Appendix). A subset of these motifs were chosen as input features for the SVM model.

Motif Features

According to the motif discovery algorithm, MEME, and findings from various genomic studies in the literature, motifs related to gene regulation can range from at least six nucleotides in length (Bailey, 2009; Hashim, 2019). Therefore, we tested the SVM model with different input datasets having motif feature lengths of 6, 7, 8, 9, 10, 12, and 15 nucleotide bases. When choosing motifs to include as features for each input dataset, we first excluded those found in fewer than 5% of all differentially expressed genes, since these likely were not biologically significant. We then sorted the remaining motifs by the difference in their percentage of occurrence between up-regulated and down-regulated genes and selected the top ones for the analyses.

Predictive Model for Classification

We used SVM as our classification model algorithm. An SVM is a classification algorithm that finds the hyperplane that best separates data into different classes. We chose this algorithm because it works well with high dimensional data, is memory efficient, and is good at generalization, all of which are beneficial to handling the datasets in this study. It is also popular in bioinformatics and other biological research for the same reasons (Yang, 2004). We performed 10-fold cross-validation using a Support Vector Classifier (SVC) with an 80-20 train-test split, implemented through scikit-learn (version 1.6.1). We tested several kernel functions, including the Radial Basis Function (RBF) and Sigmoid kernels; however, we decided on the linear kernel for its computational efficiency and accuracy.

We constructed two distinct dataset types to input into our SVM. The first dataset consisted of binary values indicating the presence or absence of each motif feature in the gene's cDNA, with a value of '1' for presence and '0' for absence. The second dataset represented the frequency of each motif feature appearing within the gene's cDNA.

In addition to calculating classification accuracy by cross-validation, we plotted receiver operating characteristic (ROC) curves and calculated different evaluation metrics, such as sensitivity, specificity, and false discovery rate (FDR), for each experiment. The classification and analyses were conducted with the scikit-learn (version 1.6.1), pandas (version 2.2.2), matplotlib (version 3.10.0), and NumPy (version 1.26.4) libraries in Google Colab.

SHAP Analysis

SVMs are commonly labeled as “black box” algorithms due to the complexity involved in creating decision boundaries and the difficulty in interpreting such mechanisms. To determine exactly how the SVCs were able to predict the change in regulation of each differentially expressed gene, we performed SHAP (SHapley Additive exPlanations) analysis on the models with 5000 8-nucleotide motifs. This method calculates the SHAP value, inspired by Shapely values in game theory, of each motif feature and instance, which indicates the feature's contribution to the final prediction. This would allow us to identify the most significant motifs in age-related HSPC gene regulation.

REFERENCES

1. Adelman, E. (2019, May 13). Aging human hematopoietic stem cells manifest profound epigenetic reprogramming of enhancers that may predispose to leukemia. *Cancer discovery. https://pubmed.ncbi.nlm.nih.gov/31085557/*
2. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research, 37(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335*
3. Gyenis, A. (2023, January 19). Genome-wide RNA polymerase stalling shapes the transcriptome during aging. *Nature genetics. https://pubmed.ncbi.nlm.nih.gov/36658433/*
4. Hashim, F. A., Mabrouk, M. S., & Walid Al-Atabany. (2019). Review of Different Sequence Motif Finding Algorithms. *Avicenna Journal of Medical Biotechnology, 11(2), 130. https://pmc.ncbi.nlm.nih.gov/articles/PMC6490410/*
5. Stoeger, T. (2022, December 9). Aging is associated with a systemic length-associated transcriptome imbalance. *Nature aging. https://pubmed.ncbi.nlm.nih.gov/37118543/*
6. Yang, Z. R. (2004). Biological applications of support vector machines. *Briefings in Bioinformatics, 5(4), 328–338. https://doi.org/10.1093/bib/5.4.328*

Figures and Figure Titles/Captions

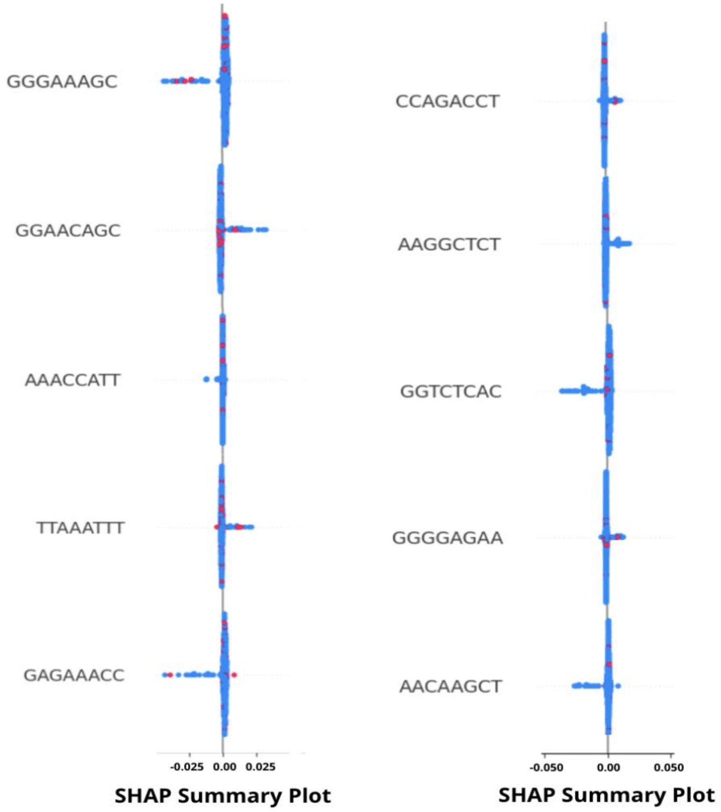

Figure 1. Summary plots from the SHAP Analysis

Tables with Titles/Captions

Motif Length (number of bases)	Mean Accuracy (from 10 fold cross validation)	Sensitivity	Specificity	False Discovery Rate	Area under ROC
6	0.393	0.375	0.394	0.611	0.408
7	0.406	0.337	0.496	0.628	0.425
8	**0.688**	**0.673**	**0.702**	**0.321**	**0.658**
9	0.588	0.535	0.636	0.390	0.572
10	0.535	0.402	0.548	0.589	0.565
12	0.529	0.529	0.529	0.529	0.529
15	0.550	0.177	0.872	0.485	0.496

Table 1: Model performance for different motif lengths. The input features indicate whether the motif occurs in the gene. The best performance was observed with 8-nucleotide motifs.

Motif Length (number of bases)	Mean Accuracy (from 10 fold cross validation)	Sensitivity	Specificity	False Discovery Rate	Area under ROC
6	0.460	0.364	0.504	0.673	0.475
7	0.436	0.422	0.429	0.613	0.420
8	**0.658**	**0.641**	**0.720**	**0.333**	**0.617**
9	0.589	0.559	0.531	0.536	0.577
10	0.535	0.454	0.513	0.529	0.498
12	0.529	0.396	0.696	0.500	0.508
15	0.550	0.190	0.860	0.472	0.506

Table 2: Model performance for different motif lengths. The input features indicate the count of occurrences of the motif in the gene. The best performance was observed with 8-nucleotide motifs.

Number of Motifs	Mean Accuracy (from 10 fold cross validation)	Sensitivity	Specificity	False Discovery Rate	Area under ROC
200	0.681	0.632	0.642	0.466	0.656
500	0.688	0.673	0.702	0.321	0.658
1000	0.733	0.804	0.645	0.361	0.737
5000	**0.818**	**0.711**	**0.806**	**0.258**	**0.775**
10000	0.813	0.762	0.819	0.208	0.781

Table 3: Model performance for different numbers of motifs. The input features indicate whether the motif occurs in the gene. The best performance was observed for 5000 motifs.

Number of Motifs	Mean Accuracy (from 10 fold cross validation)	Sensitivity	Specificity	False Discovery Rate	Area under ROC
200	0.667	0.656	0.656	0.419	0.637
500	0.658	0.641	0.720	0.333	0.617
1000	0.714	0.734	0.693	0.361	0.688
5000	**0.807**	**0.765**	**0.772**	**0.272**	**0.769**
10000	0.798	0.733	0.879	0.154	0.734

Table 4: Model performance for different numbers of motifs. The input features indicate the count of occurrences of the motif in the gene. The best performance was observed for 5000 motifs.

Motif	Number of genes the motif occurs in			Percentage of total occurrence in		Absolute difference of occurrence percentage between up and down regulated genes
	Up regulated	Down regulated	Total	Up regulated genes	Down regulated genes	
GAGCGGCG	19	65	84	22.6%	77.4%	54.8%
CCCTAATG	43	13	56	76.8%	23.2%	53.6%
CCCCCCGC	17	54	71	23.9%	76.1%	52.1%
CCGAGCCG	14	43	57	24.6%	75.4%	50.9%
CCGGCCCG	18	54	72	25.0%	75.0%	50.0%
GCGCCCGC	17	51	68	25.0%	75.0%	50.0%
GCAAATGG	16	48	64	25.0%	75.0%	50.0%
CGCCACCC	16	47	63	25.4%	74.6%	49.2%
CGGAGCGG	15	42	57	26.3%	73.7%	47.4%
CGCCGCCA	21	58	79	26.6%	73.4%	46.8%
GGCTCCGC	15	41	56	26.8%	73.2%	46.4%
CGCCGGAG	15	41	56	26.8%	73.2%	46.4%
ACACCTGC	21	56	77	27.3%	72.7%	45.5%
ACAAGTTG	17	45	62	27.4%	72.6%	45.2%
CGGCGGCC	30	79	109	27.5%	72.5%	45.0%
CGCCGAGG	16	42	58	27.6%	72.4%	44.8%
CGGCAGGG	16	42	58	27.6%	72.4%	44.8%
AGCCGGGA	16	42	58	27.6%	72.4%	44.8%
CATGTAAT	21	55	76	27.6%	72.4%	44.7%
CGGGCCGC	18	47	65	27.7%	72.3%	44.6%
GCCGCGGG	23	60	83	27.7%	72.3%	44.6%
TTGCACCA	49	19	68	72.1%	27.9%	44.1%
TAAGTTTA	19	49	68	27.9%	72.1%	44.1%
CTTTAGTG	16	41	57	28.1%	71.9%	43.9%
GTGATAAT	18	46	64	28.1%	71.9%	43.8%

Table 5: Top 25 motifs of length 8 by the absolute difference of occurrence percentage between up-regulated and down-regulated genes

Motif	Number of genes the motif occurs in			Percentage of total occurrence in		Absolute difference of occurrence percentage between up and down regulated genes
	Up regulated	Down regulated	Total	Up regulated genes	Down regulated genes	
GGCGGCGGC	34	121	155	21.9%	78.1%	56.1%
CTTTTGTTT	15	51	66	22.7%	77.3%	54.5%
GCGGCGGCC	15	50	65	23.1%	76.9%	53.8%
GCGGCGGGG	14	43	57	24.6%	75.4%	50.9%
GAGGCGGCG	17	52	69	24.6%	75.4%	50.7%
CGGCGGCGG	36	108	144	25.0%	75.0%	50.0%
CGGCGGCAG	18	53	71	25.4%	74.6%	49.3%
TGGCGGCGG	16	45	61	26.2%	73.8%	47.5%
GAGAAACCT	50	18	68	73.5%	26.5%	47.1%
GGCCGCCGC	18	49	67	26.9%	73.1%	46.3%
AGAGAAACC	48	18	66	72.7%	27.3%	45.5%
GCCCCGGCC	17	45	62	27.4%	72.6%	45.2%
CCGCTGCTG	16	42	58	27.6%	72.4%	44.8%
CCAGCACCA	18	46	64	28.1%	71.9%	43.8%
CCCCGGCCC	21	53	74	28.4%	71.6%	43.2%
GGGCGGCGG	22	55	77	28.6%	71.4%	42.9%
TTACTTTTT	20	50	70	28.6%	71.4%	42.9%
GGCGGGCGG	17	42	59	28.8%	71.2%	42.4%
GCCGCCTCC	17	42	59	28.8%	71.2%	42.4%
TAATTTATT	18	44	62	29.0%	71.0%	41.9%
TGTTTGTTT	32	78	110	29.1%	70.9%	41.8%
GCGGCGGGC	19	46	65	29.2%	70.8%	41.5%
GTTGTTTTT	19	46	65	29.2%	70.8%	41.5%
CGGCGGCTG	17	41	58	29.3%	70.7%	41.4%
ATTTTTACT	17	41	58	29.3%	70.7%	41.4%

Table 6: Top 25 motifs of length 9 by the absolute difference of occurrence percentage between up-regulated and down-regulated genes

Motif	Number of genes the motif occurs in			Percentage of total occurrence in		Absolute difference of occurrence percentage between up and down regulated genes
	Up regulated	Down regulated	Total	Up regulated genes	Down regulated genes	
GCGGCGGCGG	27	91	118	22.9%	77.1%	54.2%
TTTTTGTTTG	13	43	56	23.2%	76.8%	53.6%
GGCGGCGGCG	25	81	106	23.6%	76.4%	52.8%
CGGCGGCGGC	25	81	106	23.6%	76.4%	52.8%
TTGTTTGTTT	16	50	66	24.2%	75.8%	51.5%
TTTGTTTGTT	20	53	73	27.4%	72.6%	45.2%
AGCGGCGGCG	16	40	56	28.6%	71.4%	42.9%
TTTTTAAAAG	18	42	60	30.0%	70.0%	40.0%
TGTTTGTTTT	21	48	69	30.4%	69.6%	39.1%
TTGTTTTTTT	27	58	85	31.8%	68.2%	36.5%
CCCCACCCCC	22	47	69	31.9%	68.1%	36.2%
AGAGGAGGAG	19	40	59	32.2%	67.8%	35.6%
TTTTGTTTGT	22	46	68	32.4%	67.6%	35.3%
TTTTTTGTTT	36	73	109	33.0%	67.0%	33.9%
GATTTTTTTT	32	61	93	34.4%	65.6%	31.2%
GAGGAGGAGG	31	59	90	34.4%	65.6%	31.1%
TGTTTTTTTT	29	55	84	34.5%	65.5%	31.0%
ACTTTTTTTT	25	47	72	34.7%	65.3%	30.6%
TTTTTTTGTT	31	58	89	34.8%	65.2%	30.3%
AACTCCTGAC	21	39	60	35.0%	65.0%	30.0%
TTTTTTTCCT	27	50	77	35.1%	64.9%	29.9%
AAAATTAAAA	20	37	57	35.1%	64.9%	29.8%
AAAAATATTT	20	37	57	35.1%	64.9%	29.8%
TTTTTTTTTA	65	120	185	35.1%	64.9%	29.7%
TTTAATTTTT	26	48	74	35.1%	64.9%	29.7%

Table 7: Top 25 motifs of length 10 by the absolute difference of occurrence percentage between up-regulated and down-regulated genes