Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.31276

Regression Prediction of Household Gene Expression Levels Based on Machine Learning Algorithms

Yuanyuan Li[1], Xiaohui Huang[1*]

[1]College of Computer Science and Technology, Xinjiang University, Urumqi, China
**Corresponding Author. Email: hxhdemail@sina.com*

Abstract. Household genes are the core genes that maintain the basic life activities of cells. The stable regulation of their expression levels has a significant impact on key physiological processes such as cell metabolism and signal transduction. Once their expression is abnormal, it may cause cell functional disorders. Transcription factors, as the core molecules for regulating gene expression, form specific binding patterns in the promoter regions of housekeeping genes, construct complex regulatory networks, and directly determine the transcriptional efficiency and expression level of genes. In view of the deficiencies of the existing models, this paper proposes the DE-BiLSTM-Transformer regression algorithm. This algorithm integrates the parameter optimization ability of the differential evolution algorithm, the advantage of local temporal feature extraction of BiLSTM, and the long-distance dependency capture ability of Transformer. To construct a model that is more suitable for predicting the expression of household genes. The study first conducted correlation analysis and violin graph analysis, and then compared this algorithm with multiple machine learning algorithms. The results showed that all its evaluation indicators demonstrated significant advantages. MSE 95.028, RMSE 9.748, MAE 8.013 and MAPE 8.017 were all the lowest values among all the compared algorithms. R^2 0.488 is the maximum value. Although classic algorithms such as XGBoost, GBDT and Random Forest have certain predictive performance, their error indicators are higher and R^2 are 0.371, 0.369 and 0.356 respectively. AdaBoost, ExtraTrees and CatBoost have similar performances. The error index is between 110 and 118, and the R^2 is within the range of 0.297-0.339. The Decision tree algorithm performs the worst, with the highest error index in all aspects and an R^2 of -0.363, which cannot effectively fit the data. This algorithm significantly enhances the accuracy of gene expression prediction and data fitting by integrating the feature extraction capabilities of deep learning with the sequence modeling advantages of Transformer. It provides a more reliable prediction tool for related research and is of great significance for in-depth analysis of the physiological mechanisms and functional regulation of cells.

Keywords: Household genes, gene expression, BiLSTM, Transformer

1. Introduction

Housekeeping genes are core genes that maintain the basic life activities of cells. The stable regulation of their expression levels directly affects key physiological processes such as cell

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.31276

metabolism and signal transduction. Once their expression is abnormal, it may cause cell functional disorders [1]. Transcription factors, as the core molecules for regulating gene expression, directly determine the transcriptional efficiency and expression level of genes by forming specific binding patterns in the promoter regions of housekeeping genes and constructing complex regulatory networks [2]. With the development of high-throughput sequencing technology, a large amount of ChIP-seq data of transcription factors in the ENCODE database, motif data of JASPAR, and CAGE gene expression quantitative data of FANTOM5 have been accumulated, providing data support for analyzing the association between the regulatory mechanism of transcription factors and the expression of domestic genes [3]. However, traditional statistical analysis methods are difficult to handle redundant information in high-dimensional features and cannot accurately capture the nonlinear association between transcription factor binding patterns and gene expression, which limits the prediction accuracy of gene expression levels. There is an urgent need for a more efficient data analysis framework to break through this bottleneck [4].

Machine learning algorithms, with their ability to model nonlinear relationships and their advantage in processing high-dimensional features, have become a key technology for predicting the expression levels of household genes. Traditional linear models cannot fit the complex correlations among the number of transcription factor bindings, sequence coding features and gene expression, while deep learning models such as BiLSTM can capture the local temporal dependencies of sequence data, and Transformer can mine long-distance feature associations, which to a certain extent improves the prediction performance [5].

In view of the deficiencies of the existing models, this paper proposes the DE-BiLSTM-Transformer regression algorithm, aiming to integrate the parameter optimization ability of the differential evolution algorithm [6], the advantage of local temporal feature extraction of BiLSTM and the long-distance dependency capture ability of Transformer, and construct a model that is more suitable for the prediction of household gene expression.

2. Data sources

The dataset used in this article contains 629 domestic gene sample data, covering transcription factor binding features, sequence coding features and biological-related features. Among them, the transcription factor binding features include the number of binding sites of four core transcription factors, namely TBP, SP1, YY1 and POLR2A, in the gene promoter region. The sequence-coding features include the GC content in the promoter region and the Motif matching score generated by the FIMO tool. The biological features involve the existence of CpG islands, the average distance between the transcription factor binding site and the transcription start site, and the number of transcription factor interactions. The predictor variable is the TPM value of the gene expression level based on the FANTOM5 database standard. Correlation analysis was conducted on each variable, and a correlation heat map was drawn, as shown in Figure 1.

DOI: 10.54254/2755-2721/2026.31276

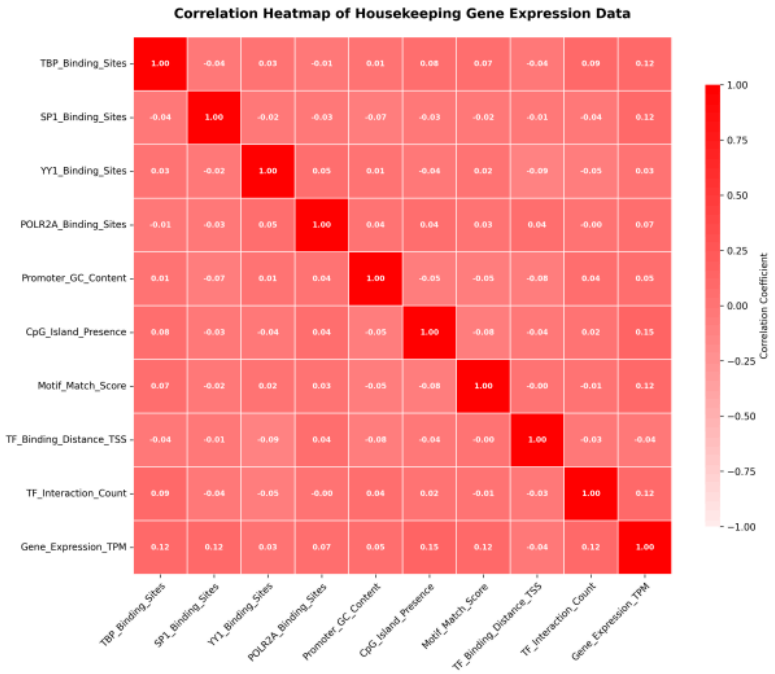

Figure 1. The correlation heat map

Violin graph analysis was conducted on each variable, and the results are shown in Figure 2.

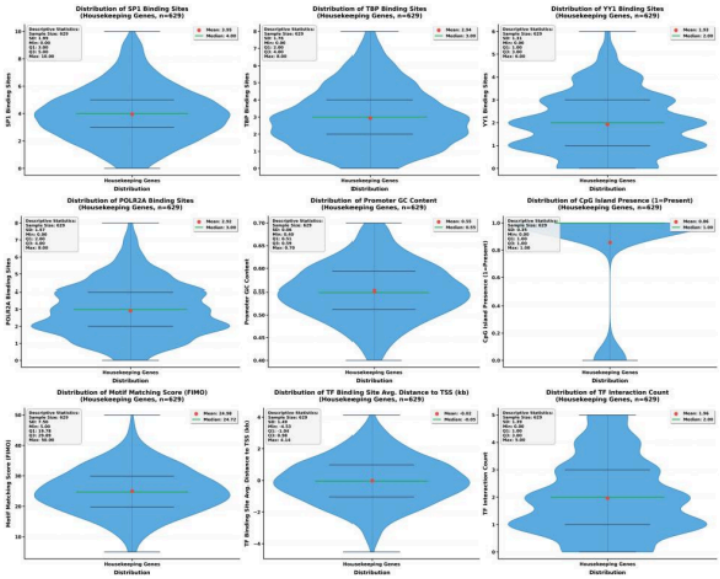

Figure 2. The violin diagrams of each variable

3. Method

3.1. Differential evolution algorithm

Differential evolution algorithm is an evolutionary optimization algorithm based on swarm intelligence. Its core idea is to achieve global search for the optimal solution by simulating mutation, crossover and selection operations in biological evolution. The algorithm initializes the random

DOI: 10.54254/2755-2721/2026.31276

generation of population individuals. The mutation operation generates mutation vectors through the linear combination of different individuals in the population [7].

3.2. Transformer

The Transformer algorithm is a sequence modeling method based on the self-attention mechanism. It abandons the serial computing mode of recurrent neural networks and adopts a parallelized self-attention mechanism to capture the dependencies at different positions in the sequence. The algorithm consists of two parts: the encoder and the decoder. The encoder includes a multi-head self-attention layer and a feedforward neural network layer. The decoder, on this basis, adds a mask self-attention layer and an encoder-decoder attention layer. Meanwhile, the timing information of the sequence is introduced through position encoding [8]. The network structure of the Transformer is shown in Figure 3.

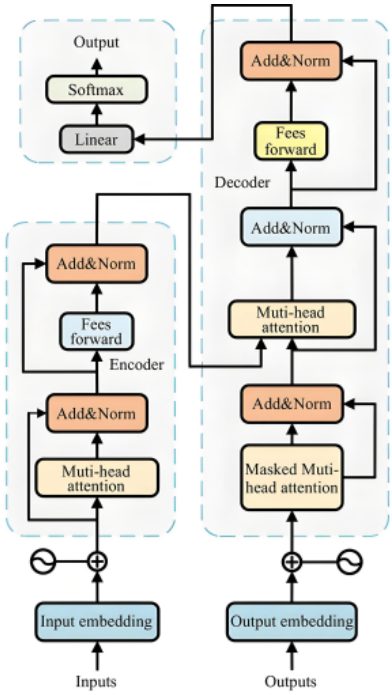

Figure 3. The network structure of the transformer

3.3. BiLSTM

The bidirectional long short-term memory network is a bidirectional extension form of the long short-term memory network, aiming to capture the bidirectional dependency relationship between the context and the background of sequence data. LSTM controls the information transmission of cell states through input gates, forgetting gates and output gates, effectively solving the vanishing gradient problem of traditional recurrent neural networks. BiLSTM is composed of forward LSTM and backward LSTM. The forward LSTM processes information in the forward sequence, while the backward LSTM processes information in the reverse sequence [9]. The network structure of BiLSTM is shown in Figure 4.

DOI: 10.54254/2755-2721/2026.31276

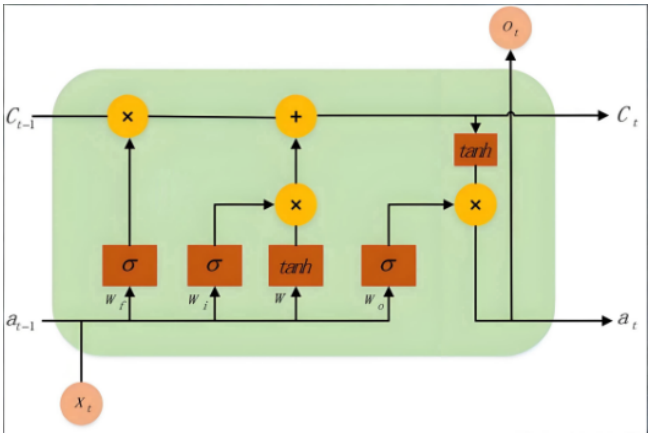

Figure 4. The network structure of BiLSTM

3.4. DE-BiLSTM-Transformer

The DE-BiLSTM-Transformer regression algorithm is a hybrid regression model that integrates the differential evolution algorithm, bidirectional long short-term memory network and Transformer, aiming to combine the advantages of the three to improve the regression prediction performance of sequential data. The algorithm first uses the DE algorithm to globally optimize the key parameters of BiLSTM and Transformer, including the number of hidden layer nodes and learning rate of BiLSTM, as well as the number of multi-head attention heads and the dimension of the feedforward neural network of Transformer, to solve the problems of low efficiency and local optimality of manual parameter adjustment [10].

4. Result

The parameter Settings of the project are as follows: The number of search agents in the optimization algorithm is 8, the maximum number of iterations is 6, the dimension of the optimization parameters is 3, the lower bounds of the parameters are 1e-3, 10, 1e-4, and the upper bounds are 1e-2, 30, 1e-1, corresponding respectively to the learning rate, the number of hidden layer nodes, and the regularization coefficient. In terms of model structure, the number of input channels is the feature dimension, the maximum position encoding is 512, the number of self-attention heads is 4, the total number of key channels is 128, the number of hidden nodes in the BiLSTM layer is obtained through optimization, the dropout ratio is 0.01, and the output dimension of the fully connected layer is 1. The training parameters adopt the adam optimizer, with a maximum of 1000 training rounds and a batch size of 256. The dataset is shuffled in each round. The initial learning rate is obtained through optimization, with a learning rate decline factor of 0.1 and a decline period of 850. The L2 regularization coefficient is also obtained through optimization, and the gradient clipping threshold is 10. The execution environment is automatically selected.

Experiments were conducted using multiple machine learning algorithms, and the results are shown in Table 1. The bar comparison results of each indicator are shown in Figure 5.

DOI: 10.54254/2755-2721/2026.31276

Table 1. The results of the comparative experiment

Model	MSE	RMSE	MAE	MAPE	R^2
AdaBoost	110.327	10.504	8.374	8.438	0.329
GBDT	118.431	10.883	8.623	8.677	0.369
Decision tree	244.69	15.643	12.1	12.933	-0.363
ExtraTrees	112.626	10.613	8.719	8.706	0.297
Random Forest	115.93	10.767	8.8	8.752	0.356
XGBoost	107.467	10.367	8.736	8.796	0.371
CatBoost	112.256	10.595	8.885	8.767	0.339
DE-BiLSTM-Transformer	95.028	9.748	8.013	8.017	0.488

The DE-BiLSTM-Transformer algorithm proposed in this paper demonstrates significant advantages in all evaluation indicators. Its MSE 95.028, RMSE 9.748, MAE 8.013 and MAPE 8.017 are all the lowest values among all the comparison algorithms, while R^2 0.488 is the highest value. In comparison Although classic machine learning algorithms such as XGBoost, GBDT and Random Forest have demonstrated certain predictive performance, all their error metrics are higher than those of DE-BiLSTM-Transformer, and the R^2 values are 0.371, 0.369 and 0.356 respectively, which are significantly lower than those of the proposed algorithms The performance of AdaBoost, ExtraTrees and CatBoost is similar to that of the above algorithms, with error indicators ranging from 110 to 118 and R^2 values within the range of 0.297 to 0.339. However, the Decision tree algorithm performs the worst, with all error indicators being the highest and the MSE reaching 244.69 The RMSE is 15.643, the MAE and MAPE are 12.1 and 12.933 respectively, and the R^2 value is -0.363, indicating that it cannot effectively fit the data pattern.

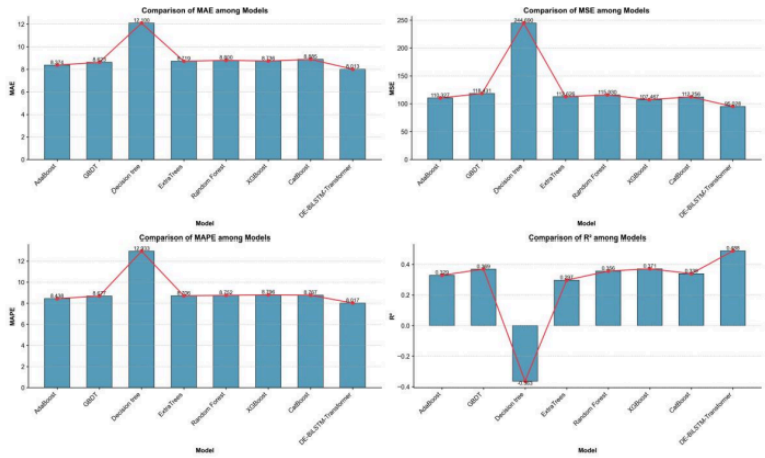

Figure 5. The bar comparison charts of each indicator

5. Conclusion

Housekeeping genes are the core genes that support the basic life activities of cells. The stable regulation of their expression levels is crucial for key physiological processes such as cell metabolism and signal transduction. Once their expression is abnormal, it is easy to cause cell functional disorders. Transcription factors, as the core molecules for regulating gene expression, can

DOI: 10.54254/2755-2721/2026.31276

form specific binding patterns in the promoter regions of housekeeping genes, build complex regulatory networks, and directly determine the transcriptional efficiency and expression level of genes. In view of the deficiencies of the existing models, this paper proposes the DE-BiLSTM-Transformer regression algorithm, which integrates the parameter optimization ability of the differential evolution algorithm, the advantage of BiLSTM in extracting local temporal features, and the ability of Transformer to capture long-distance dependencies, in order to adapt to the prediction of household gene expression. After correlation analysis, violin graph analysis and comparison with multiple machine learning algorithms, this algorithm has advantages in all evaluation indicators. MSE 95.028, RMSE 9.748, MAE 8.013, MAPE 8.017 are all the lowest, and R^2 0.488 is the highest. Although classic algorithms such as XGBoost and GBDT have certain predictive performance, they have higher errors and lower R^2 (0.356-0.371). AdaBoost and others have similar performances, and Decision tree has the worst performance (MSE 244.69, R^2 -0.363). In conclusion, the prediction accuracy and data fitting effect of this algorithm are superior to those of traditional algorithms. It can more accurately capture the complex relationships of data, providing a reliable solution for the prediction of gene expression in household management. It also lays a methodological foundation for the research on gene regulatory mechanisms and the analysis of related physiological processes.

References

[1] Guan, Jiantao, et al. "A near-complete cucumber reference genome assembly and Cucumber-DB, a multi-omics database." Molecular Plant 17.8 (2024): 1178-1182.
[2] Chung, Youngmin, et al. "Accurate spatial gene expression prediction by integrating multi-resolution features." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[3] Jia, Huiqiong, et al. "Neural network-based predictions of antimicrobial resistance phenotypes in multidrug-resistant Acinetobacter baumannii from whole genome sequencing and gene expression." Antimicrobial Agents and Chemotherapy 68.12 (2024): e01446-24.
[4] Redmond, Ethan J., et al. "Stable and dynamic gene expression patterns over diurnal and developmental timescales in Arabidopsis thaliana." New Phytologist 246.3 (2025): 1147-1162.
[5] Shimazaki, Shun, et al. "Building a machine-learning model to predict optimal mevalonate pathway gene expression levels for efficient production of a carotenoid in yeast." Biotechnology journal 19.1 (2024): 2300285.
[6] Balakittnen, Jaikrishna, et al. "A novel saliva-based miRNA profile to diagnose and predict oral cancer." International Journal of Oral Science 16.1 (2024): 14.
[7] Nataren, Nathalie, et al. "Advancing Remote and Low-Resource Healthcare Delivery with Microbiopsy Skin Sample mRNA Housekeeping Gene Stability Analysis Across Various Temperatures." bioRxiv (2024): 2024-11.
[8] Zhang, Dongdong, et al. "Genome-wide analysis of the ABC gene family in almond and functional predictions during flower development, freezing stress, and salt stress." BMC Plant Biology 24.1 (2024): 12.
[9] Liu, Jinxiang, et al. "Chromosome-level genome assembly of humpback grouper using PacBio HiFi reads and Hi-C technologies." Scientific data 11.1 (2024): 51.
[10] Li, Qionghou, et al. "Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits." Plant communications 5.10 (2024).