Multi-LSTM: A Multiscale Long Short-Term Memory-Based Framework for Predicting Time-Series Transcriptomic Gene Expression

Ying Zhou
Southeast University
Erteng Jia
Southeast University
Huajuan Shi
Southeast University
Zhiyu Liu
Southeast University
Yuqi Sheng
Southeast University
Min Pan
Southeast University
Jing Tu
Southeast University
Qinyu Ge (geqinyu@seu.edu.cn)
Southeast University
Zuhong Lu
Southeast University

Article

Keywords: Long short-term memory, Time-Series, Gene Expression, Empirical mode decomposition, Intrinsic mode functions

Posted Date: May 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1586379/v1

Abstract

RNA degradation can significantly affect the results of gene expression profiling, with subsequent analysis failing to faithfully represent the initial gene expression level. It is urgent to have an artificial intelligence approach to better utilize the limited data to obtain meaningful and reliable analysis results in the case of data with missing destination time. In this study, we propose a method based on signal decomposition technique and deep learning, named Multi-LSTM. It is divided into two main modules, one decomposes the collected gene expression sequences by empirical mode decomposition (EMD) algorithm to obtain a series of subsequences with different frequencies to improve data stability and reduce modeling complexity. The other is based on long short-term memory (LSTM) as the core predictor, aiming to deeply explore the temporal nonlinear relationships embedded in the subsequences. Finally, the prediction results of subsequences are reconstructed to obtain the final prediction results of time-series transcriptomic gene expression. The results show that EMD can efficiently reduce the nonlinearity of the original sequences, which provides a reliable theoretical support to reduce the complexity and improve the robustness of LSTM models. Overall, the decomposition-combination prediction framework can effectively predict gene expression levels at unknown time points.

1. Introduction

The fate of RNA transcripts in dead tissues and the decay of isolated RNA are not subject to strict regulation unlike in vivo normal cells.[1] Whether most transcripts decay at a similar rate remains unknown at present.[2] Understanding RNA degradation patterns is critical for studies of samples collected in the field or samples collected in clinical settings, as these tissue samples are often not immediately stored in conditions that can prevent RNA degradation.[3] For example, obtaining donated samples in the clinic may require several hours of transport for RNA extraction and sequencing analysis. It is crucial to investigate the decay pattern of RNA degradation over time, because neglecting the initial RNA quality changes in some clinical samples may even lead to incorrect judgments of disease. Besides, some RNA extraction methods needs to take several hours. However, samples collected during certain low-quality fieldwork remain the only means of solving specific problems.[4] In these cases, the extracted RNA is usually partially degraded and may not faithfully represent the initial gene expression level.[5] Moreover, ignoring the initial RNA quality changes in clinical samples may lead to misjudgment of the disease. Analyzing the effect of RNA degradation on gene expression is essential to obtain meaningful and reliable gene expression data and to ensure reproducible results. Therefore, an artificial intelligence method is needed to predict the decay pattern of RNA degradation over time. Then, the irreversible change in gene expression level of RNA that occurred during this time can be understood clearly. What's more, it is possible to predict the true gene expression level of RNA when degradation has not occurred by the time of degradation and the existing gene expression level.

Time-series based gene expression data has become one of the most fundamental methods for studying biological processes. Most biological processes are dynamic, and the use of time series data can characterize the function of specific genes, discover linkages and regulatory relationships between genes,

and find their clinical applications.[6–8] Time series data can not only capture transient changes in gene expression, but also demonstrate the sequence of events that occur during biological processes and study the temporal patterns of gene expression.[9] These transcriptome time series experiments require observational sampling of biological systems, preparation of transcriptome libraries, and sequencing at each time point. The current time series experiments have a high overhead. Their sampling time points are generally not numerous and they are only used with some specific study subjects. The data it can obtain is limited. Artificial intelligence computational methods are particularly important for humans, who can only perform limited experiments directly. Therefore, there is an urgent need for tools that can study limited time-series-based gene expression data or frameworks and methods that can use time-series-based gene expression data for analytical modeling.[10] So far, this field has been developed and studied by lots of scientists. For example, Lakizadeh A[11] proposed the use of time-series-based gene expression data to complement protein-protein interaction (PPI) networks to detect protein complexes. Wise A[12] developed SMARTS, a method that combines static and time-series-based data from multiple individuals to reconstruct condition-specific unsupervised feedback networks. Qian B[13] used deep neural networks for time-series classification tasks. However, there is no method that combines long short-term memory (LSTM) [14] deep learning models for gene expression prediction of time series.

It is the first prediction framework constructed by deep learning methods for time series data of isolated RNA from tissue samples. The time-series-based prediction is able to extract causality because the information on gene expression changes over time flows in one direction. Overall, we developed a new LSTM-based deep learning model Multi-LSTM to predict gene expression at target time points in transcriptome time series data. (Fig. 1) This model is capable of predicting gene expression at initial time points from gene expression at subsequent time points. It can also predict gene expression at future times from gene expression at previous time points. This study has important implications for those studies that lack initial time samples. Moreover, applying the LSTM deep learning model to the analysis of transcriptome time series data can obtain the missing RNA-seq gene expression data, which can facilitate the functional prediction of transcriptome gene expression profiles and the discovery of disease-related genes. With the increase in the number of time-series-based biological experimental data, it will be possible to make full use of the time-series information based on this study to gain a deeper understanding of the gene functions and molecular mechanisms in biological processes.

2. Results And Discussions

2.1. Multi-scale analysis based on EMD decomposition

The raw gene expression data are shown in Fig. 2a, and the data to be predicted are highlighted with blue dots. Specifically, gene expression data related to apoptosis and degradation of RNA samples at different time points were utilized, and the gene names and expressions are provided. These prediction points are the gene expression data at 0 hours when the RNA has not yet undergone degradation. It can be seen that this data exhibits complex nonlinearity and non-stationarity, which makes the prediction of gene expression at specific locations extremely difficult. How to solve the gene expression sequence

nonlinearity problem and make the deep learning neural network operate on a more regular time is the key to improve the prediction accuracy and model robustness. In this regard, the empirical mode decomposition (EMD) [15] signal decomposition technique was utilized, which aims to reduce the nonlinearity of the original gene expression sequence and decompose it into a series of stable components with different frequencies. This study is based on each component and combined with a deep learning predictor, which is expected to improve the accuracy of model analysis.

The decomposition results are shown in Fig. 2b. After EMD component, the original gene expression signal is divided into nine different frequency components of intrinsic mode functions (IMF). [16] As a new data analysis method, empirical modal decomposition has obvious advantages in dealing with non-smooth and non-linear data. EMD is especially suitable for handling terahertz time-domain signals. Compared with the simple harmonic function, IMF does not have the constant amplitude and frequency in the simple harmonic function and has amplitude and frequency that vary as a function of time. The system can achieve remarkably less reconstruction errors at extremely low oversampling rates. The different components contain different gene expression information. The correlation between each component and the original data is quantified by Spearman [17] rank-order relationship, and the correlation heat map is shown in Fig. 2c. It is obvious that different components contain different gene expression information, and aggregating all components will obtain the total gene expression information. It indicates the high reliability of the multi-scale analysis method proposed in this study. Meanwhile, compared with the original sequence in Fig. 2a, each component shows the trend of recent harmonic function and has a strong change pattern. Finally, the trends of such data can be effectively captured using LSTM, which provides accurate prediction of unknown gene expression.

2.2. Construction of core predictor model using LSTM

First, the LSTM model is built for each IMF component to obtain the corresponding prediction results. The prediction results of the nine components are shown in Fig. 3a-i. It can be seen that the LSTM can accurately capture the change trend of each component after decomposition, and the prediction accuracy is satisfactory. Figure 4a-i shows the comparison of the predicted results of each component with the original results. It can be seen that the prediction results of each component are closer to the original data. There is no difference between the prediction results of each component and the original data, and all the p-values > 0.05.

In order to quantify the prediction accuracy of each component more directly, root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2) were selected as the prediction performance evaluation indexes. Among them, RMSE can be used to assess the degree of inconsistency between the predicted and true values of the model. Theoretically, the smaller its value is, the better the performance of the model is. In addition, the MAE can be used to complement the evaluation of the performance of this model. Meanwhile, the induced effects of the model can be evaluated by the determinants statistic R^2. The R^2 is an excellent statistical indicator of the closeness between the predicted and true values. The results of R^2 of each component are shown in Fig. 5a-i. In general, the

performance of this model can be fully evaluated by the above three items. The results of the evaluation of each component are shown in Table 1. These results illustrate the effectiveness of the multi-scale analysis based on EMD decomposition in various aspects.

Table 1
RMSE, MAE and R^2 values for each IMF component.

	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8	IMF9
RMSE	0.1332	0.0871	0.0959	0.0806	0.0817	0.0688	0.0202	0.0081	0.0172
MAE	0.1059	0.0661	0.0812	0.0647	0.0639	0.0492	0.0160	0.0060	0.0151
R^2	0.9016	0.8332	0.8619	0.8765	0.9442	0.9550	0.9617	0.9777	0.9914

2.3. Target time point prediction based on LSTM

The final gene expression prediction results can be obtained by recombining the prediction results of each component. We made predictions for the two time points of 0 h (Fig. 6a, b) and 6 h (Fig. 6c, d) in the original data, respectively. Both predictions were excellent either by predicting the initial unknown time point from the known later time point or by predicting the later unknown time point with the known earlier time point. In addition, we calculated the determinant statistics R^2 between the predicted results and the true values. It can be seen that the fit between the predicted and true values is satisfactory, and more than 90% of the data points are located in the elliptical confidence band of the fitted curve. (Fig. 6b, d) The forward time prediction results have an RMSE of 0.2558, MAE of 0.2105, and R^2 of 0.4771. The overall prediction error is low and there is a high agreement between the prediction results and the true values. The RMSE value for the reverse time prediction was 0.4002, the MAE value was 0.3169, and the R^2 was 0.6145. It also shows low error and excellent agreement between the prediction results and the true values. The prediction results in both directions indicate that this prediction model can well obtain the gene expression results at unknown destination time. In summary, it can be seen that our proposed Multi-LSTM model can dig deeper into the change pattern of gene expression. Moreover, the prediction results can excellently reflect the expression level of genes at unknown time.

The model combines signal processing technology and deep learning with biomedicine, and can be widely applied to gene function prediction, disease-related gene discovery and transcriptomic analysis. In the present study, only the expression data of RNA degradation-related genes were utilized, and the data volume can be expanded in the future to apply Multi-LSTM to a wider range of fields.

3. Materials And Methods

3.1. Data acquisition and pre-processing

The data used in this study were partly obtained from previously published articles.[18] Time-specific clean data of the build sequencing results have been uploaded to the NCBI database. The alignment of

genes from time-series transcriptome sequencing results was performed by Hisat2.[19] The conversion between data was performed using Samtools.[20] The number of reads mapped to each gene for each sample was calculated using FeatureCounts. A catalog of RNA degradation-related genes was downloaded from NCBI, and the corresponding expressions of the genes associated with each sample were extracted. Data pre-processing includes data cleaning, data integration, data transformation and reduction of experimental data before model construction. These processes can improve the training speed of subsequent models and the accuracy and credibility of experimental results. In this study, the raw gene expression data were preprocessed with censoring and normalization. The RNA-seq data were normalized by heat map using ggplots in the R package.

3.2. EMD-based decomposition signal processing

The purpose of EMD algorithm is to decompose the original signal into some series of IMFs, and then obtain the time-frequency relationship of the signal by Hilbert transform. Taking the terahertz time-domain spectral signal as an example, the basic calculation process of the EMD algorithm is as follows.

Step1: The terahertz time-domain spectral signal $\mathrm{x(t)}$ is calculate for all the maxima and minima points. All the curves form the upper envelope$u(t)$ were made by connecting all the maxima points with the cubic spline function. The resulting curves form the lower envelope $v(t)$ were made by connecting all the minima points with the cubic spline function.

Step2: The mean value of the upper and lower envelopes is calculated by the formula:

$$m(t) = \frac{u(t) + v(t)}{2}$$

1

The components of the EMD algorithm are defined as:

$$\mathrm{h(t) = x(t) - m(t)}$$

2

Step3: Determine whether the component $\mathrm{h(t)}$ satisfies the definition of IMF. If $\mathrm{h(t)}$ satisfies the definition of IMF, then $\mathrm{h(t)}$ is the first IMF component filtered out, and it is noted as $\mathrm{c_1(t)}$. If $\mathrm{h(t)}$ does not satisfy the definition of IMF, $\mathrm{h(t)}$ is taken as the original data and the above two steps are repeated until the first IMF component is calculated, which is also labeled as $\mathrm{c_1(t)}$.

Step4: The first IMF component is raised from the original signal to obtain the residual signal, which is calculated as follows.

$$r_1\mathrm{(t) = h(t) - c_1(t)}$$

3

Repeat the above three steps with $r_1(\mathrm{t})$ as the new signal to be analyzed, so that the second IMF component $\mathrm{c}_2(\mathrm{t})$ is obtained. Remove the second IMF component from $r_1(\mathrm{t})$ to obtain the new residual term.

$$r_2(\mathrm{t}) = r_1(\mathrm{t}) - \mathrm{c}_2(\mathrm{t})$$

4

The above steps are repeated continuously until the stopping criterion of the EMD algorithm is satisfied. The stopping criterion of the EMD algorithm uses the Cauchy convergence criterion, a test that requires the normalized squared difference between two adjacent extraction operations to be sufficiently small, as defined by:

$$SD_k = \frac{\sum_{t=0}^{T} \left| h_{k-1}(t) - \quad h_k(t) \right|^2}{\sum_{t=0}^{T} h_{k-1}(t)^2}$$

5

In the formula: T denotes the signal length, and the decomposition ends when SD_k is less than the set threshold value.

Root mean square error (RMSE), also known as root-mean-square deviation, is a commonly used measure of the difference between the values. Its formula is:

$$RMSE = \sqrt{\frac{\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)}{(n-1)}}$$

6

The mean absolute error (MAE) is the average of the absolute values of the deviations of all individual observations from the arithmetic mean. Mean absolute error avoids the problem of errors canceling each other out and thus accurately reflects the magnitude of the actual prediction error. Its formula is:

$$MAE = \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|}{n}$$

7

The R^2 coefficient of determination (R^2) is a statistical measure of how close the regression prediction is to the true data point. Its value ranges from 0 to 1 and represents the percentage correlation between the predicted value and the actual value of the target variable. If a model has an R^2 value of 0, the model does not predict the target variable at all. Conversely, if the R^2 value is 1, the model is considered to be

able to predict the target variable accurately. A value between 0 and 1 indicates that the corresponding percentage of the target variable in the model can be explained by the input characteristics. The formula for this is:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

8

In the equations of RMSE, MAE, and R^2, n is the number of samples, $\hat{y}_i$ is the expression value of the target gene predicted by the ith sample. y_i is the true expression value of the target gene in the ith sample, and $\bar{y}$ is the sample mean.

3.3. Construction of core predictors using LSTM

LSTM is a special recurrent neural network (RNN) proposed to solve the deficiency of RNN in the long-range dependency problem. [21] LSTM achieves the selection of new information addition and the control of information accumulation rate by adding a gating mechanism. The addition of gating control makes up for the shortcomings of RNN network structure. Compared with the classical RNN network in which there is only a single tanh cyclic body, LSTM adds three gate structures and one memory unit. The flow chart is shown in Fig. 1b. Let W be the gate weight and b be the bias. The gate structure can be described as:

$$g(x) = \sigma\left(W_x + b\right)$$

9

The three gates in the LSTM gate structure are the input gate, the output gate and the forget gate. The input gate mainly reflects the number of information stored in the cell state c_t at the current moment of the input sequence x_t at this time. The output state of the forget gate mainly determines the amount of information from the cell state c_t to the output value of the output gate at this moment. Each gate state update of the LSTM is calculated as:

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right)$$

10

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$

11

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right)$$

12

Memory unit update:

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right)$$

13

Status update:

$$c_t = f_t \bullet c_t + i_t \bullet \tilde{c}_t$$

14

$$h_t = o_t \bullet \tanh\left(c_t\right)$$

15

The formula σ is the sigmoid activation function. W_i, W_f, W_o, b_i, b_f, b_o are the weight matrices and biases of the input gate, forgetting gate and output gate. W_c, b_c are the weight matrices and biases of the memory cell after updating. U_i, U_f, U_o, U_c are the state quantity weight matrices. h_{t-1} is the previous moment state quantity.

4. Conclusion

In this study, we propose a method based on signal decomposition techniques and deep learning, called Multi-LSTM. It is a prediction framework to explore the expression changes of RNA degradation-related genes in the initial hours of the sample, which is important for studying the effect of RNA degradation on gene expression levels. The first module decomposes the collected gene expression sequences by empirical mode decomposition algorithm to obtain a series of subsequences with different frequencies in order to improve data stability and reduce modeling complexity. The second module uses a LSTM neural network as the core predictor, aiming to deeply explore the temporal nonlinear relationships embedded in the subsequences. Finally, the prediction results of subsequences are reconstructed to obtain the final transcriptome time series gene expression level prediction results. The results show that EMD can efficiently reduce the nonlinearity of the original sequences, which provides a reliable theoretical support to reduce the complexity and improve the robustness of LSTM models. Meanwhile, the decomposition-combination prediction framework can effectively predict gene expression levels at unknown time points by combining the robust temporal nonlinearity analysis capability of LSTM. The combination of LSTM and effective EMD decomposition algorithm not only improves the accuracy of prediction results, but

also has low prediction errors. The results show that both forward and reverse time series can be predicted accurately. The combination of deep learning neural networks with biomedicine will lead to great breakthroughs in gene function prediction, disease related gene discovery and transcriptome time series analysis.

Abbreviations

empirical mode decomposition (EMD); long short-term memory (LSTM); intrinsic mode functions (IMF); root mean square error (RMSE);mean absolute error (MAE); coefficient of determination (R^2); recurrent neural network (RNN).

Declarations

Authors' contributions:

Author contributions: Y. Z. did the investigation, formal analysis, writing original draft, data analysis. E.J. did the investigation, methodology. H.S. did the investigation, methodology. Z.L. did the investigation, methodology. Y.S. did the data analysis. M.P. did the investigation, methodology. J.T. did the supervision, resources. Q.G. did the methodology, conceptualization, writing review and editing. Z.L. did the supervision.

Data Availability Statement:

The data of this study are freely available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA826142. The source code used to support the findings of this study are freely available to the academic community at https://github.com/habbyzy/Multi-LSTM.

Acknowledgments:

Not applicable.

Conflicts of Interest:

The authors declare no conflict of interest.

Sample Availability:

Samples of the compounds are not available from the authors.

Funding:

This research was supported by the National Natural Science Foundation of China [61801108] and Natural Science Foundation of Jiangsu Province [BK20201148, BK20211166].

References

1. Ioannidis, J. P. A. Is Molecular Profiling Ready for Use in Clinical Decision Making? The Oncologist. 12, 301–311 (2007).
2. Gallego, R. I., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: Impact of RNA Degradation On Transcript Quantification. BMC BIOL. 12, 42 (2014).
3. Fouda, M. A. et al. Effect of Seasonal Variation in Ambient Temperature On RNA Quality of Breast Cancer Tissue in a Remote Biobank Setting. EXP MOL PATHOL. 112, 104334 (2020).
4. Opitz, L. et al. Impact of RNA Degradation On Gene Expression Profiling. BMC MED GENOMICS. 3, 36 (2010).
5. Shen, Y. et al. Impact of RNA Integrity and Blood Sample Storage Conditions On the Gene Expression Analysis. Volume 11, 3573–3581 (2018).
6. Jin, C. & Cukier, R. I. Machine Learning Can be Used to Distinguish Protein Families and Generate New Proteins Belonging to those Families. The Journal of Chemical Physics. 151, 175102 (2019).
7. E. El-Attar, N., M. Moustafa, B. & A. Awad, W. Deep Learning Model to Detect Diabetes Mellitus Based on DNA Sequence. Intelligent Automation & Soft Computing. 31, 325–338 (2022).
8. Liang, J., Huang, X., Li, W. & Hu, Y. Identification and External Validation of the Hub Genes Associated with Cardiorenal Syndrome through Time-Series and Network Analyses. Aging (Albany NY). 14, 1351–1373 (2022).
9. Zhou, J. et al. Deep Learning Sequence-Based Ab Initio Prediction of Variant Effects On Expression and Disease Risk. NAT GENET. 50, 1171–1179 (2018).
10. Karim, F., Majumdar, S., Darabi, H. & Chen, S. LSTM Fully Convolutional Networks for Time Series Classification. IEEE ACCESS. 6, 1662–1669 (2018).
11. Lakizadeh, A., Jalili, S. & Marashi, S. PCD-GED: Protein Complex Detection Considering PPI Dynamics Based On Time Series Gene Expression Data. *J THEOR BIOL*. 378, 31–38 (2015).
12. Wise, A. & Bar-Joseph, Z. SMARTS: Reconstructing Disease Response Networks From Multiple Individuals Using Time Series Gene Expression Data. BIOINFORMATICS. 31, 1250–1257 (2015).
13. Qian, B. et al. Dynamic Multi-Scale Convolutional Neural Network for Time Series Classification. IEEE ACCESS. 8, 109732–109746 (2020).
14. Cheng, X., Wang, J., Li, Q. & Liu, T. BiLSTM-5mC: A Bidirectional Long Short-Term Memory-Based Approach for Predicting 5-Methylcytosine Sites in Genome-Wide DNA Promoters. *MOLECULES*. 26, 7414 (2021).
15. Zhang, Q. & Zheng, X. Y. Walsh Transform and Empirical Mode Decomposition Applied to Reconstruction of Velocity and Displacement from Seismic Acceleration Measurement. Applied Sciences. 10, 3509 (2020).
16. Yang, C., Ling, B. W., Kuang, W. & Gu, J. Decimations of Intrinsic Mode Functions Via Semi-Infinite Programming Based Optimal Adaptive Nonuniform Filter Bank Design Approach. SIGNAL PROCESS. 159, 53–71 (2019).

17. Zhou, Y. & Li, S. BP Neural Network Modeling with Sensitivity Analysis On Monotonicity Based Spearman Coefficient. CHEMOMETR INTELL LAB. 200, 103977 (2020).
18. Jia, E. et al. Effects of Brain Tissue Section Processing and Storage Time On Gene Expression. ANAL CHIM ACTA. 1142, 38–47 (2021).
19. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A Fast Spliced Aligner with Low Memory Requirements. NAT METHODS. 12, 357–360 (2015).
20. Li, H. et al. The Sequence Alignment/Map Format and SAMtools. BIOINFORMATICS. 25, 2078–2079 (2009).
21. Xie, W., Luo, J., Pan, C. & Liu, Y. SG-LSTM-FRAME: A Computational Frame Using Sequence and Geometrical Information Via LSTM to Predict miRNA–Gene Associations. BRIEF BIOINFORM. 22, 2032–2042 (2021).

Figures

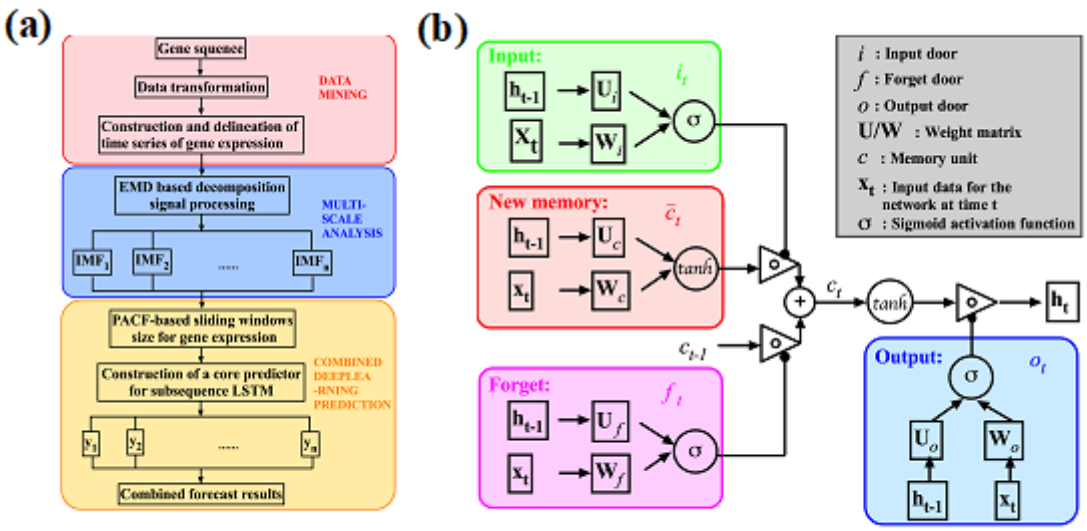

Figure 1

Multi-LSTM flowchart. **(a)**: Flowchart of data processing. Step 1: Data mining. The data used in this study were obtained from previously published articles.[18] RNA degradation-related genes and the expression corresponding to each time were extracted and normalized. Step 2: Multi-scale analysis. The original signal is decomposed into IMF_1-IMF_9 by EMD algorithm, and then the time-frequency relationship of the signal is obtained by Hilbert transform. Step 3: Deep learning combined prediction. The core predictor is constructed by LSTM and the prediction results of each subseries are integrated. **(b)**: LSTM flow chart. The three gates in the LSTM gate structure are the input gate, the output gate and the forget gate. The input gate mainly reflects the number of information stored in the cell state at the current moment of the input sequence at this time. The output state of the forget gate mainly determines the amount of

information from the cell state to the output value of the output gate at this moment. The output gates of gate control often use the *sigmoid* function as the activation function, while the activation functions of the input gates and memory cells usually use *tanh*.

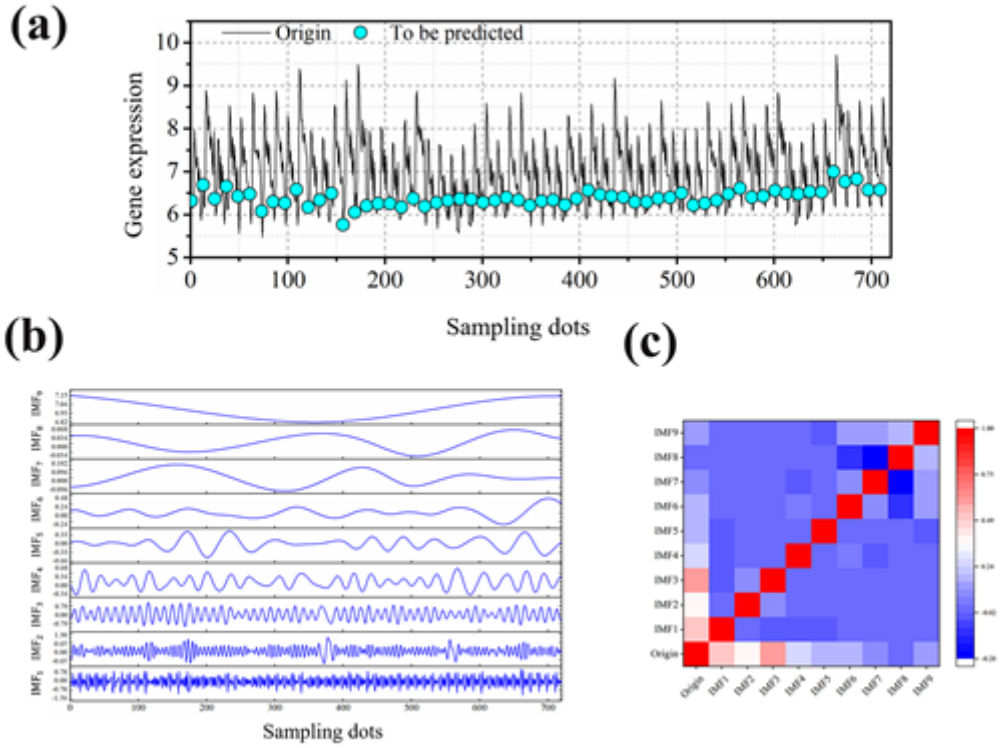

Figure 2

EMD-based decomposition of the original data. **(a)** Expression of the genes to be analyzed. The data to be predicted are highlighted with blue dots. **(b)** The results obtained from the original data based on EMD decomposition for each component of IMF_1-IMF_9. **(c)** Spearman correlation of each component of IMF_1-IMF_9 after EMD decomposition with the original sequence.

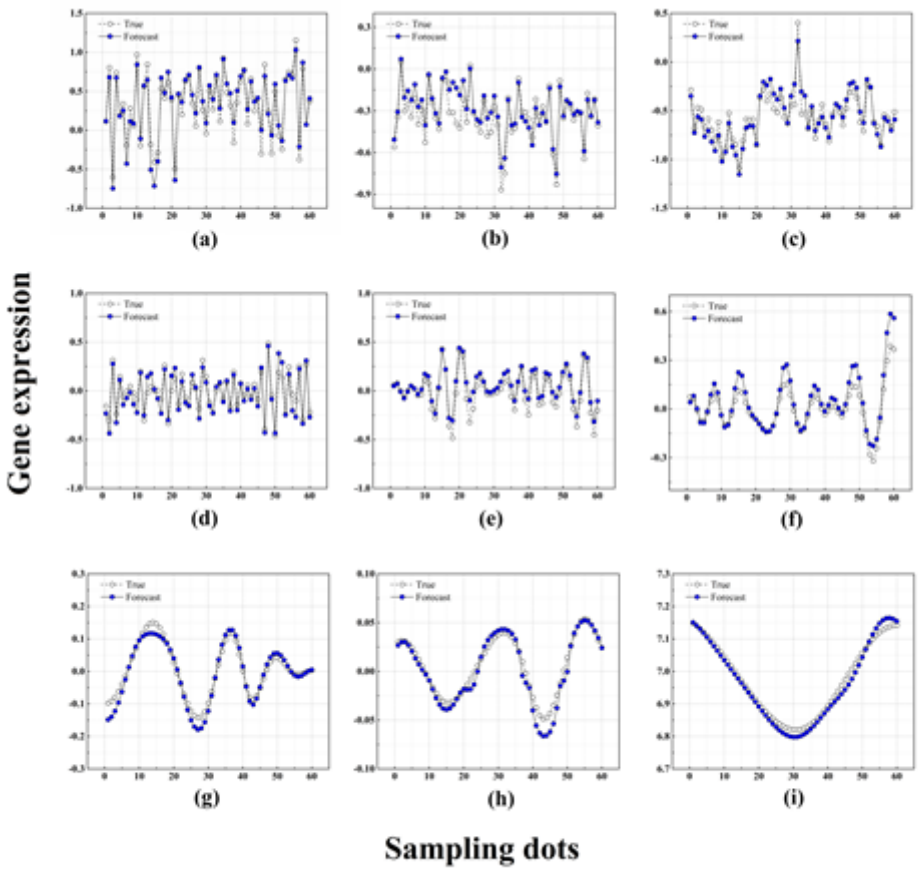

Figure 3

The prediction results of each component using LSTM. **(a)-(i)** correspond to the results of components IMF_1-IMF_9, respectively.

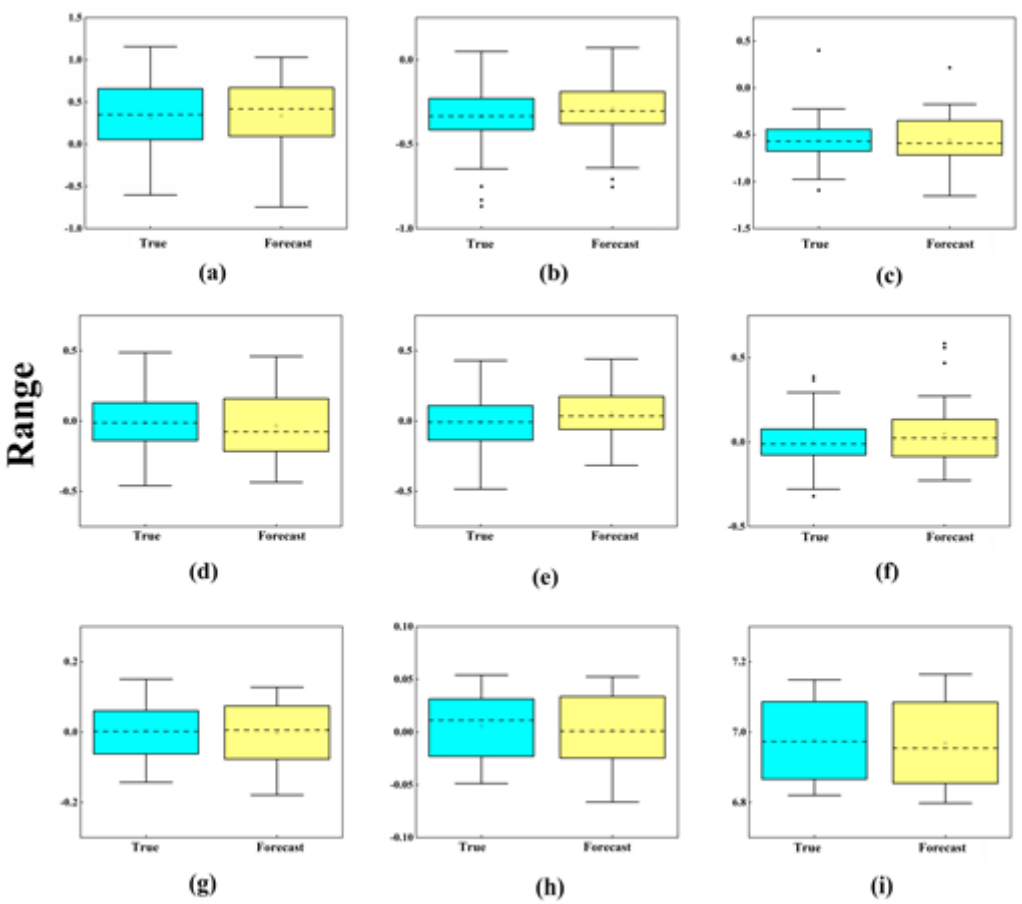

Figure 4

Proximity between the predicted values and the original data. **(a)-(i)** correspond to the results of components IMF_1-IMF_9, respectively. The results show that there is no difference between the predicted and true values of each component, and all p-values > 0.05.

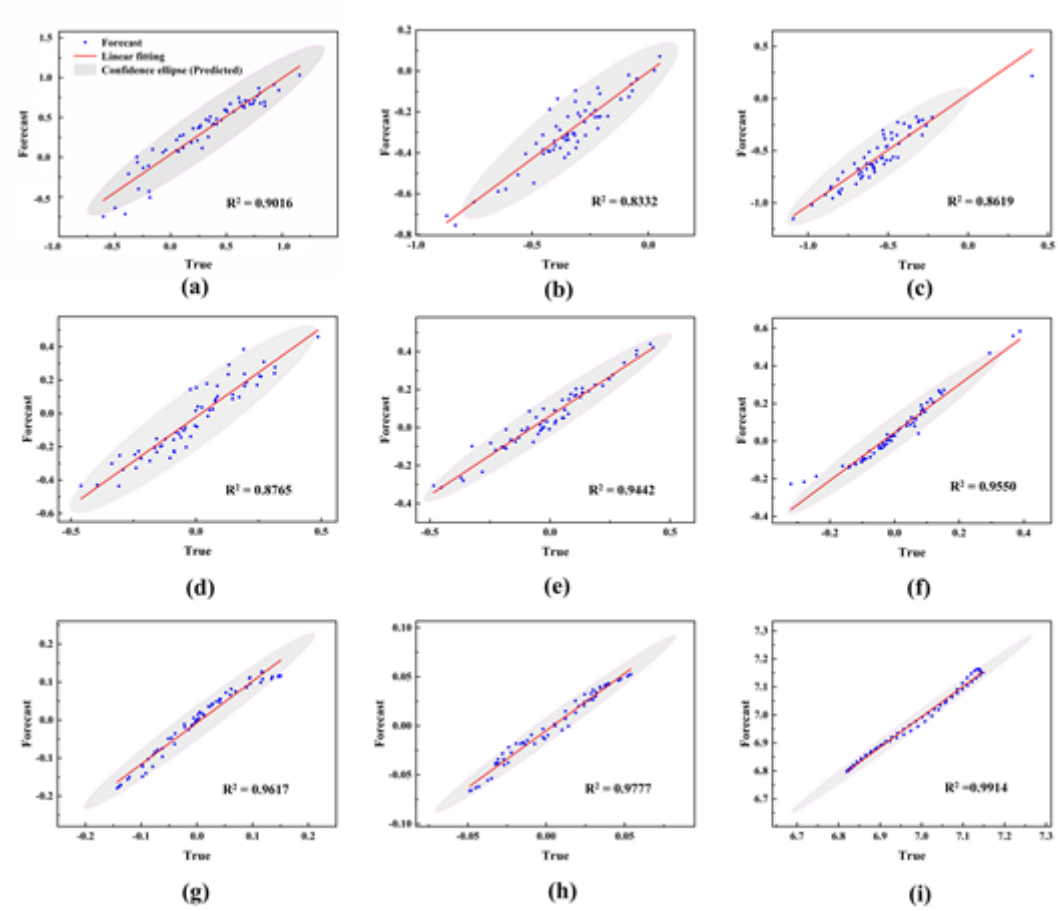

Figure 5

R^2 between predicted and true values of each component. **(a)-(i)** correspond to the results of components IMF_1-IMF_9, respectively.

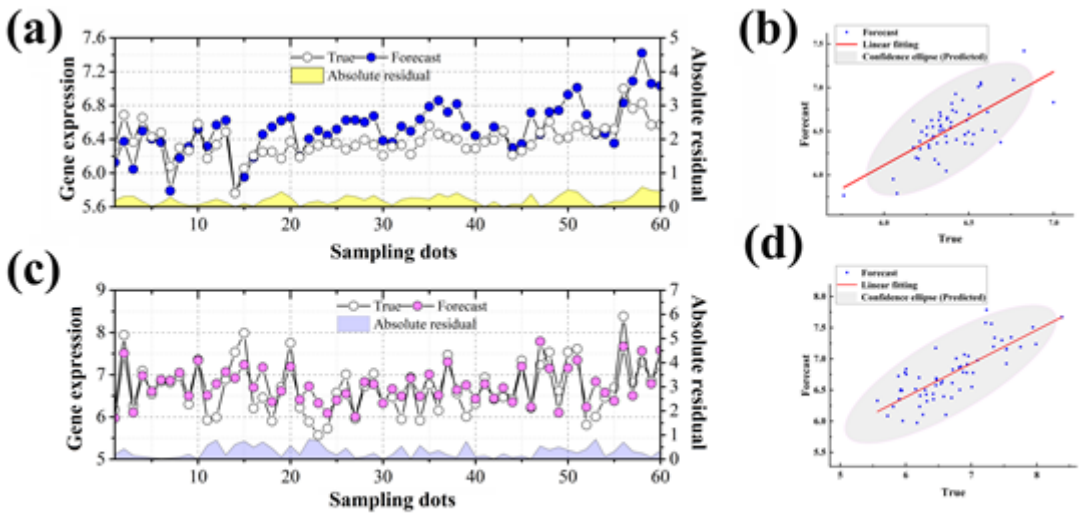

Figure 6

Overall prediction results. **(a, b)**: RMSE value of 0.2558, MAE value of 0.2105 and R^2 of 0.4771 at 0 h. The overall prediction error is low and there is a high agreement between the prediction results and the true values. **(c, d)**: RMSE value of 0.4002, MAE value of 0.3169, and R^2 of 0.6145 for the reverse time prediction of 6 h. It also shows low error and excellent agreement between the prediction results and the true values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInfo.zip