Computational Molecular Biology 2025, Vol.15, No.3, 112-121
http://bioscipublisher.com/index.php/cmb

Feature Review Open Access

Deep Learning for Predicting Gene Expression from Genomic Sequences

Shiying Yu
Biotechnology Research Center, Cuixi Academy of Biotechnology, Zhuji, 311800, China
Corresponding author: shiying.yu@cuixi.org

Computational Molecular Biology, 2025, Vol.15, No.3 doi: 10.5376/cmb.2025.15.0011
Received: 03 Mar., 2025
Accepted: 14 Apr., 2025
Published: 02 May, 2025

Preferred citation for this article:
Yu S.Y., 2025, Deep learning for predicting gene expression from genomic sequences, Computational Molecular Biology, 15(3): 112-121 (doi: 10.5376/cmb.2025.15.0011)

Abstract Different cell types of higher organisms share the same genomic sequence but have distinct gene expressions, which is attributed to complex gene regulatory mechanisms. Cracking the regulatory rules of gene expression is of vital importance for understanding diseases and life processes. This review examines the research progress on predicting gene expression from genomic sequences using deep learning, including data sources and processing, model architecture design, prediction methods, performance evaluation and interpretability analysis, current challenges and the latest advancements, and illustrates them through case studies of specific species. Finally, the prospects of the integration of deep learning and multi-omics in the future and its potential impact in precision medicine and functional genomics were prospected.

Keywords Deep learning; Genomic sequence; Gene expression; Gene regulation; Multiomics integration

1 Introduction

The genome may seem uniform, but that doesn't mean all cells work mechanically. For instance, different cells in the same person may have exactly the same genome, but their gene expression patterns can be vastly different. This is not merely a simple cause-and-effect relationship, but rather a covert manipulation by various complex regulations. What's more interesting is that only about 2% of the human genome is responsible for directly encoding proteins, while the remaining large non-coding sequences - accounting for 98% - are often overlooked but contain crucial information that determines when and under what conditions genes are expressed (Zhang et al., 2019). To truly understand the occurrence of diseases or the subtle changes in the life process, it is necessary to clarify the role these "silent" fragments play in it. Some people have suggested that directly predicting the expression pattern of genes from sequences might be a key step in cracking this "regulatory code", and it could also bring new breakthroughs to medical and biological research (Beer and Tavazoie, 2004).

In gene regulation, distance is not absolute. The three-dimensional folding of chromatin enables enhancers that were originally separated by tens of thousands of bases to "come close" to promoters and remotely participate in regulation (Robson et al., 2019). Don't think that only the area close to the starting point is important - promoters are usually right next to the transcription starting point, but the positions of enhancers can be as far as the corner of the world. The cis-regulatory elements in the genome, such as promoters and enhancers, essentially provide binding sites for transcription factors to control transcriptional activity. Once the DNA sequence changes and the functions of these components are disrupted, the gene expression level may be rewritten, showing different traits and even causing diseases. Therefore, clarifying the correspondence between sequences and expressions has always been an unavoidable challenge in the study of gene regulation (Li et al., 2018).

The data of genomics is getting larger and larger, and high-throughput technologies are outputting information wave after wave. Faced with such a scale, traditional analytical methods often find themselves struggling, while deep learning is gradually making its way into researchers' toolboxes. It can automatically extract features from complex data, especially excelling at those nonlinear laws that traditional methods fail to capture. In fact, experiments have already been conducted in gene expression prediction: the accuracy of deep neural networks is often higher than that of the old methods (Chen et al., 2016). Just think about it. By integrating deep learning with

genomics, decoding the mechanisms of gene regulation and promoting precision medicine in the future might be a key approach (Drusinsky et al., 2024).

2 Genomic Sequence Data and Gene Expression Characteristics

2.1 Sources and processing methods of genomic sequences

For those engaged in genomic research, the first step is often to "find data". Most of these sequences come from reference genomes in public databases, or they may be the results of sequencing projects that have been made public. After obtaining the data, it's not as simple as simply throwing it into the model and calling it a day - you need to first perform preprocessing such as format conversion and quality checks. Which specific segments to take depends on the research objective: for instance, if the aim is to analyze the promoter, a sequence upstream of the gene would be extracted. If remote control is to be considered, the more distant areas should also be taken into account. Next comes the encoding required by the model, converting the four bases A, C, G, and T into digital form. A common practice is single-hot encoding, where each base is represented by a sparse vector of length 4. Only after all these steps are completed can the sequence be sent into the deep learning model for training (El-Tohamy et al., 2024).

2.2 Measurement and standardization of gene expression data

To measure gene expression, RNA sequencing (RNA-SEq) is now widely used, while in the past, the chip method was more common. RNA-seq will count the number of transcripts of each gene, but these raw readings are not reliable for direct comparison with different samples and need to be standardized first. Common metrics include RPKM and FPKM. Some people prefer TPM because it is more convenient for cross-sample comparisons (Zhao et al., 2020). The purpose of standardization is to balance out the differences caused by sequencing depth and gene length, so that the expression levels of samples can be compared on the same table. If necessary, logarithmic transformation can also be performed on the expression matrix or batch effects can be processed to make the data cleaner. The processed expression data can eventually be paired with genomic sequences to train the model (Zhao et al., 2021).

2.3 Dataset integration and feature extraction strategies

When conducting deep learning modeling, data must not be scattered here and there. Sequences and expression results from different sources often have to be pieced together into a unified dataset in the end to ensure sufficient sample size and diversity. However, integration is not simply splicing. Data from different platforms need to be batch corrected and normalized first; otherwise, they simply cannot be compared. There are also many considerations when preparing the features for model input. Although deep learning can directly learn from the original sequence end-to-end, if there is not enough data, adding some artificial features appropriately can be helpful, such as the k-mer frequency of the sequence, the GC content, or the known binding sites of transcription factors. When it comes to integrating multi-omics data, it is also necessary to align information and sequences such as chromatin accessibility and histone modifications one by one according to the positions of genes or genomes. Only when these integrations and feature extractions are done well can the model better capture biological signals.

3 Fundamentals and Model Architecture of Deep Learning

3.1 Overview of common deep learning models (CNN, RNN, transformer)

For genomic sequence analysis, the common models are actually just a few types: convolutional neural networks, recurrent neural networks, and the Transformer, which has been very popular in recent years. Let's start with the Transformer. It relies on its self-attention mechanism to connect the relationships at various positions in a sequence regardless of the distance, and can handle regulatory information of hundreds of kilobytes (Lan, 2024). CNN, on the other hand, is more adept at mining local features, such as short motifs in DNA, which are frequently used in such tasks. RNNS (like LSTM) are originally strong at handling sequential data and can capture certain long-term dependencies. However, when it comes to extremely long genomic sequences, they struggle a bit, so their application is not as frequent. Overall, CNN is efficient, RNN can retain sequential information, and

Transformer covers the entire world. Researchers usually choose or mix and match according to the task requirements (Figure 1) (Almotairi et al., 2024).

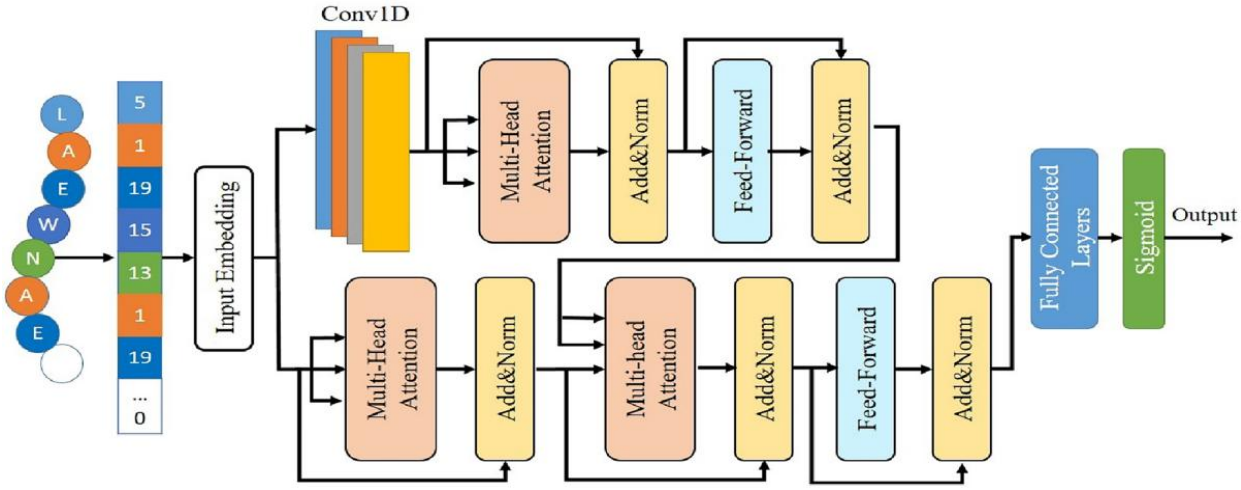

Figure 1 Hybrid transformer-CNN architecture for predicting hemolytic activity of peptides (Adopted from Almotairi et al., 2024)

3.2 Model selection and architecture design principles

When developing a gene expression prediction model, there is no "one-step" formula, but some basic ideas still cannot be bypassed. First, let's talk about the matching between the model and the input sequence: If the main regulatory information is at the proximal end of the gene, a model like CNN that is good at capturing local features is sufficient. However, once the remote enhancers come into play, it is necessary to consider architectures that can handle long-distance dependencies, such as Transformers (Tu et al., 2024). When considering the complexity of the model, don't forget to look at the amount of data - if the data is small, don't build a too deep network. Appropriate use of regularization is more stable. Only with a large number of samples can there be space to stack deeper layers. Some people also use multi-task learning, allowing the model to predict multiple related outputs at once, and the generalization ability is often better (Zeng et al., 2015). Finally, do not overlook biological knowledge. Incorporating these prior information appropriately can also enhance the interpretability of the model. By adhering to these practices, the performance and biological rationality of the model will be more guaranteed.

3.3 Model training, validation and hyperparameter optimization

When training gene expression prediction models, it is generally impossible to avoid the set of supervised learning: as long as there are known gene sequences and corresponding expression values, they can be used as samples. Don't think that just throwing it in and it's done. There are still many details in the process. For instance, the data must first be divided into a training set, a validation set and a test set. The validation set is used to adjust the model structure and hyperparameters, and also to prevent overfitting (Makarova et al., 2021). Sometimes an early stop strategy is added to prevent the model from over-learning (Dorka et al., 2023). The loss function also depends on the type of task. For regression, mean square error is commonly used, while for classification, cross-entropy is often employed. As for hyperparameters - such as learning rate, network depth, and regularization coefficient - they are usually repeatedly adjusted on the validation set through grid or random search. Finally, it is the turn of the independent test set to verify the effect. Only by completing such a round can a model with stable performance and strong generalization ability be obtained.

4 Prediction Methods Based on Deep Learning

4.1 Identification and coding strategies of gene regulatory elements

There is not just one approach to handling DNA sequences. The most common one is, of course, single-heat coding, which converts bases into sparse vectors and allows the model to learn the characteristics of the regulatory elements on its own. Sometimes, however, researchers also carry a bit of prior knowledge and directly label potential regulatory regions or specific motif positions in the input to ensure that the model does not miss key

segments. Deep CNNS are very good at finding short sequence motifs from these encoded sequences, and the convolutional kernels seem to be secretly capturing the binding sites of transcription factors (Choong and Lee, 2017). The gameplay of Transformer is different. Usually, position encoding is added to enable the model to perceive the relative relationships at various points in the sequence. In recent years, some people have even treated DNA as a kind of "language", using word embeddings or pre-trained models to extract features (Chen et al., 2020). Only when the coding is well selected can the model's efficiency in utilizing regulatory information be high.

4.2 End-to-end sequence input gene expression prediction model

Some studies simply throw the original sequence directly into the model and let it calculate the gene expression value by itself. This end-to-end approach has been proven to work. For instance, some people have used CNN to directly predict the expression levels in different tissues from the whole genome sequence, and the accuracy is quite high. The Enformer that emerged later was even more powerful. It introduced the Transformer and could process sequence information up to 100kb at a time. The correlation between the prediction results and the actual expression was significantly better than that of traditional CNNS, and it could also infer the impact of non-coding variations on the expression (Stefanini et al., 2023). Many such end-to-end models conduct multi-task learning to simultaneously predict expressions under multiple organizations or conditions in order to enhance generalization ability. They can automatically interpret the regulatory information in the sequence without the need for manual feature extraction, thus becoming a powerful tool for studying the regulation of gene expression (Ramprasad et al., 2024).

4.3 Comprehensive prediction method combining multi-omics data

In fact, DNA sequences alone cannot explain all the differences in gene expression. This is something everyone has long realized. As a result, multi-omics integrated prediction methods have gradually gained popularity in recent years. Researchers will incorporate epigenetic information, three-dimensional genomic structure and other data into sequence models to make the regulatory background more complete. For instance, if chromatin accessibility and histone modification data are input as additional features along with the sequence, it can help the model determine which fragments are active (Dong et al., 2024b). Some people have also incorporated the three-dimensional interaction frequency of chromatin, making it easier for the model to identify the effect of distal enhancers on genes (Merelli et al., 2015). The actual results are quite astonishing: After introducing remote interaction information, the relevant performance of gene expression prediction soared from 0.46 to 0.93. Overall, this type of multi-omics integration model not only makes more accurate predictions but also brings more biological details.

5 Model Performance Evaluation and Interpretive Analysis

5.1 Performance evaluation metrics and benchmark testing

There is no single "universal indicator" for evaluating gene expression prediction models. A common practice in research is to conduct a Pearson correlation between the predicted values and the measured values. This linear correlation coefficient has almost become a mandatory data in many papers (Mikhaylova and Thornton, 2019). Meanwhile, indicators that measure deviation, such as mean square error, are often taken into account incidentally (Ji et al., 2023). If the problem is changed to a classification of high and low expression, the evaluation method will change again, and indicators such as accuracy rate and recall rate will have to come in handy. Generally, the generalization ability of the model is tested on completely independent test data. The same dataset is also used to compare this method with traditional machine learning models or previous deep models to see how much improvement there is. Some teams even directly cross-validate the prediction results with large expression databases, such as GTEx's multi-organization expression profiles, to see if the model can also work for real data. Only by combining multiple indicators and strict benchmark tests can the predictive level of the model be seen more clearly.

5.2 Model interpretability techniques (feature visualization, attention mechanism, etc.)

The predictions made by deep learning are often very accurate, but researchers have never been able to figure out exactly what the model "thinks". To explain these black boxes, people have come up with many solutions. By

visualizing features, the sequence fragments that the model cares about the most can be identified. A common practice is to calculate the contribution score of the input sequence to the output, such as using gradient methods or DeepLIFT, directly marking the bases that have the greatest impact on the prediction, and incidentally infer the signals that the model values (Xiao et al., 2025). The Transformer with an attention mechanism is more interesting. It can reveal the regions that the model focuses on from the attention weights and even infer the regulatory relationship between remote enhancers and promoters (Figure 2) (Liu et al., 2024). In actual analysis, the key motifs picked out by the model are often concentrated in open chromatin and overlap with the sites expressing quantitative traits, with quite clear biological significance. Through these interpretable tools, the predictive basis of the model becomes visualized, and the results are more convincing.

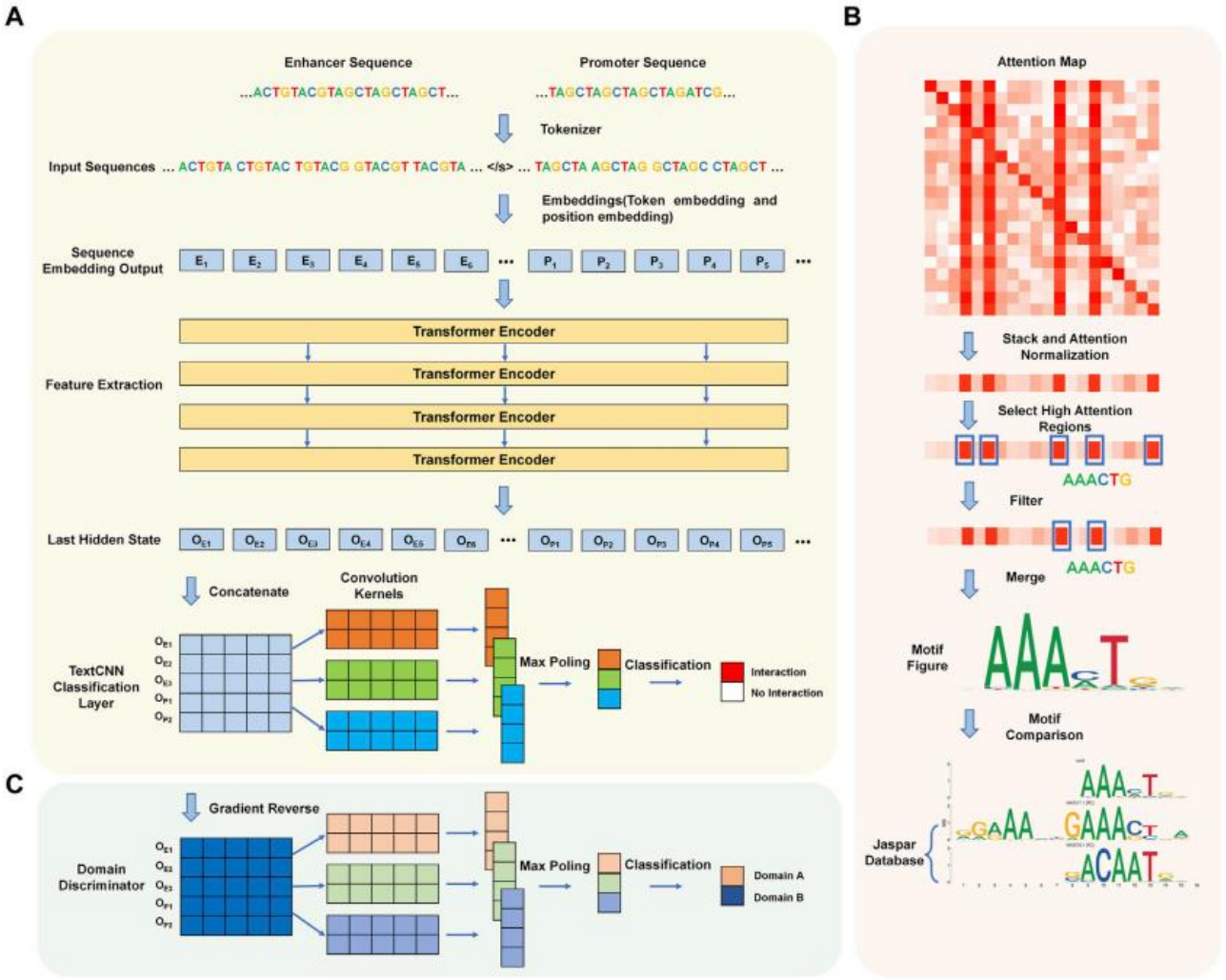

Figure 2 Workflow of TF-EPI (Adopted from Liu et al., 2024)

Image caption: (A) Cell type-specific EPI detection network structure. Generally, it includes four steps: tokenization, sequence embedding, feature extraction and classification. (B) The process of de novo motif discovery. (C) Model expansion for cross-cell type EPI detection. The Domain Discriminator is used during the model training process to determine whether the input data comes from the source cell line or the target cell line (Adopted from Liu et al., 2024)

5.3 Result verification and experimental data comparison

It has become a consensus that no matter how reliable a prediction is, it must be backed by real data and experiments. Usually, the results of the model are compared with the experimental measurement values that did not participate in the training at all to see if they match. Some people have conducted large-scale reporter gene experiments in human studies to test the model's judgment on the impact of mutations on expression (Avsec et al., 2021). The performance of Enformer was quite accurate, almost consistent with the experiments. There are similar examples in the field of plants: Researchers will perform site-directed mutagenesis on the promoter elements that

the model considers key, and then directly measure the expression changes (Lin and Jiang, 2021). The results are very consistent with the predictions. Only such independent verification can prove that the output of the model truly has biological significance and make it more convincing in scientific research and clinical applications.

6 Challenges and Latest Developments

6.1 Data sparsity and sample bias issues

In the field of gene expression prediction, for deep learning to be trained well, there must first be a sufficient amount and diversity of data, but this is precisely the difficulty. High-quality expression profiles are not easy to obtain. The cost of covering a species comprehensively is not low, and the sample size often fails to keep up with the appetite of super-large models. Even if data is collected, there are still other troubles: the number of highly expressed and low-expressed genes is often unbalanced, the model tends to favor the majority class, and unpopular low-expression patterns may be directly ignored (Wang and Hu, 2023). If the training set is overly focused on certain organizations or experimental conditions, it will also cause the model to generalize poorly to situations it has never seen before. Not to mention the noise in the experimental measurement. Once the model is overly fitted with this noise, its generalization ability will be dragged down. To solve these problems, we need to focus on both sides: on the one hand, continuously accumulate larger and more diverse datasets; on the other hand, in training, adopt strategies such as data augmentation and loss weighting to minimize bias as much as possible, making the model more stable and more generalized (Jaichitra et al., 2023).

6.2 The balance between model interpretability and biological significance

Although deep learning has strong predictive capabilities, once the model becomes too complex, researchers find it difficult to figure out exactly how it reaches its conclusions, which has become a headache for many people. Everyone not only wants to know what results the model gives, but also wants to figure out "why" there is such an output. If it is completely like a black box, no matter how high the accuracy rate is, it will not help much in promoting the understanding of biology. Some people have attempted to work on the model structure, such as adding interpretable modules or prior constraints, to link the internal structure of the network with biological processes, making it naturally more transparent (Zhang et al., 2022). Some teams also directly use visualization methods to break down the black box, looking for clues from the motifs or attention distributions learned from the model, and then convert them into biological hypotheses for verification (Hanczar et al., 2020). Just enhancing interpretability should not lower performance. This requires the entire community to come up with solutions together, developing new networks and algorithms that not only maintain prediction accuracy but also make the model more "explainable".

6.3 Frontier research trends and potential breakthrough directions

Research on gene expression prediction has been constantly emerging with new ideas. In recent years, a notable trend has been the integration of pre-trained large models into genomics. Researchers first train general deep models on massive genomic sequences and then fine-tune them for specific tasks, such as predicting gene expression. DNABERT is a case in point. It pre-trains DNA as a "language", providing a strong representation of sequence features for downstream predictions (Dong et al., 2024a). Meanwhile, generative deep learning has also begun to be applied to regulatory sequence design. With the help of these generative models, scientists can synthesize new regulatory elements with specific expression effects, which holds great promise in gene therapy and synthetic biology (Yang et al., 2025). Of course, algorithms, computing power and data are all constantly advancing. New models and methods are expected to follow one after another, helping us gain a deeper understanding of gene regulatory networks and also broadening the possibilities of practical applications.

7 Case Study: Prediction of Gene Expression in Specific Species

7.1 Case background and research objectives

For this case, we chose corn (*Zea mays*), which is a commonly used model crop. Its genome is large and complex, and gene expression is also influenced by multiple layers of regulation, making prediction even more challenging. Especially for those distant enhancers, they influence gene expression through chromatin spatial structure. However, traditional methods often only look at the proximal sequences of genes, and three-dimensional genomic

information is basically ignored, thus reducing the prediction accuracy. Our idea is to create a deep learning model that incorporates the genomic sequence and chromatin interaction information of corn to more accurately predict the gene expression levels of different tissues and identify key regulatory elements at the same time. Incidentally, I also want to see if multi-omics integration can indeed enhance predictive performance and evaluate the potential value of these regulatory elements in corn growth and breeding.

7.2 The design and implementation process of deep learning models

In this case, we developed a corn gene expression prediction model called DeepCBA. It uses a dual-pathway structure: one pathway processes sequences near the gene promoter, and the other pathway receives distal fragments that have chromatin interactions with the gene. The two streams of data each extract features through a convolutional neural network and then converge at a higher level to jointly output the expression value of the target gene. This design enables the model to simultaneously capture the regulatory effects of the near-end cis element and the long-range enhancer (Wang et al., 2024). During training, we took the gene expression data of multiple corn tissues as the supervisory signal, adopted the mean square error as the loss, and combined cross-validation and regularization to prevent overfitting. After the model was trained, we evaluated it on the test set and performed feature visualization to see exactly which key sequence motifs the model focused on (Zeng et al., 2018).

7.3 Result analysis and practical application value

DeepCBA has performed quite impressively in predicting gene expression in corn. Compared with the model that only uses promoter sequences, when the information of remote chromatin interactions is also added, the predicted correlation increases from approximately 0.47 to 0.93 at once, and the remote regulatory factors are clearly better captured. The model also identified many sequence motifs related to high expression. Most of these motifs are concentrated in the open chromatin regions of corn and highly overlap with the sites expressing quantitative traits, indicating that the features it has learned are in good agreement with the real regulatory elements (Jiang et al., 2020). We also performed site-directed mutagenesis on the promoters of two corn genes, and the results showed that their expression changes were almost consistent with the model predictions. Overall, DeepCBA not only makes predictions more accurate but also identifies key regulatory elements, providing a new tool for functional genomics research and molecular breeding. Researchers can use this tool to screen elements that affect yield or stress resistance and make targeted improvements through gene editing.

8 Future Outlook and Conclusions

The combination of deep learning and multi-omics has opened up a very broad path for gene expression prediction. With the continuous emergence of new data such as single-cell epigenomics and spatial transcriptomics, models can simultaneously absorb multi-level information including genomics, transcriptomics, and epigenomics, and the regulatory networks they construct are also more complete. Future research is likely to bind different omics data and deep models together in order to more accurately restore the full picture of gene regulation. Take disease research as an example. By feeding the model with gene sequence variations, chromatin states, and transcriptome data together, the impact of pathogenic variations on expression can be better predicted, providing a basis for precision medicine. Crop science also has similar demands. Multi-omics models can help explain the changes in gene expression under environmental stimuli and guide the breeding of better varieties. Of course, data integration standardization, model complexity and computational costs are all problems that must be addressed, but technological progress is likely to make this approach the norm in functional genomics research.

The potential of predicting gene expression based on genomic sequences has long been eyed by the medical and biological communities. Especially in precision medicine, such methods can help annotate the functions of a large number of non-coding variations in humans. Many disease-related mutations do not fall in the coding region but may function by altering gene expression. Deep learning models can directly predict the impact of these variations on expression, providing clues for the search of pathogenic mechanisms and drug targets. The research approach of functional genomics is thus changing - in the past, large-scale experiments were relied on to screen regulatory elements. Now, models can first predict potential regulatory sequences across the entire genome and then select

key points for verification, which is much more efficient and may also uncover regulatory factors that are difficult to discover through traditional experiments. It can be said that sequence function prediction driven by deep learning is tightening the connection between genotypes and phenotypes, bringing more precise and efficient research methods to medicine and biology.

If we string together all the previous discussions, the potential of deep learning to predict gene expression using genomic sequences is already quite obvious and is still moving forward. From data preparation, model design to result interpretation, this method is becoming increasingly mature. It can capture sequence features that are difficult to identify by traditional methods, demonstrating unprecedented precision in the study of gene expression regulation and bringing new ideas to functional genomics. The application prospects are also quite broad: from the screening of disease risk variations to molecular breeding of crops, it can be put to good use.

However, for such technologies to play a greater role, joint efforts from both the academic and industrial sectors are still needed. In research, it is necessary to continuously accumulate multi-dimensional high-quality data and also develop more efficient and easier-to-understand model algorithms. In terms of policy, efforts should be made to promote the cultivation of interdisciplinary talents, better integrate life sciences and artificial intelligence, establish an open and shared genome and expression database, and at the same time formulate relevant AI application norms to ensure the reliability and long-term usability of prediction results. As long as both technology and policy keep pace, deep learning-driven gene expression prediction is expected to truly transform the landscape of life science research and take precision medicine and biotechnology innovation to a new level.

Acknowledgments

I would like to express my heartfelt thanks to all the teachers who have provided guidance for this study.

Conflict of Interest Disclosure

The author affirms that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Almotairi S., Badr E., Abdelbaky I., Elhakeem M., and Abdul Salam M., 2024, Hybrid transformer-CNN model for accurate prediction of peptide hemolytic potential, Scientific Reports, 14(1): 14263.
https://doi.org/10.1038/s41598-024-63446-5

Avsec Ž., Agarwal V., Visentin D., Ledsam J., Grabska-Barwinska A., Taylor K., Assael Y., Jumper J., Kohli P., and Kelley D., 2021, Effective gene expression prediction from sequence by integrating long-range interactions, Nature Methods, 18(10): 1196-1203.
https://doi.org/10.1038/s41592-021-01252-x

Beer M., and Tavazoie S., 2004, Predicting gene expression from sequence, Cell, 117(2): 185-198.
https://doi.org/10.1016/S0092-8674(04)00304-6

Chen R., Dai R., and Wang M., 2020, Transcription factor bound regions prediction: Word2Vec technique with convolutional neural network, Journal of Intelligent Learning Systems and Applications, 12(1): 1-13.
https://doi.org/10.4236/jilsa.2020.121001

Chen Y., Li Y., Narayan R., Subramanian A., and Xie X., 2016, Gene expression inference with deep learning, Bioinformatics, 32(12): 1832-1839.
https://doi.org/10.1101/034421

Choong A., and Lee N., 2017, Evaluation of convolutional neural networks modeling of DNA sequences using ordinal versus one-hot encoding method, In: 2017 International Conference on Computer and Drone Applications (IConDA), IEEE, pp.60-65.
https://doi.org/10.1101/186965

Dong G., Wu Y., Huang L., Li F., and Zhou F., 2024a, TExCNN: Leveraging pre-trained models to predict gene expression from genomic sequences, Genes, 15(12): 1593.
https://doi.org/10.3390/genes15121593

Dong W., Zhang J., Dai L., Chen J., Wu H., He R., Pang Y., Wang Z., Jian F., Ren J., Liu Y., Tian Y., Liu S., Zhao X., and Xie X., 2024b, Mapping eukaryotic chromatin accessibility and histone modifications with DNA deaminase, bioRxiv, 24: 630236.
https://doi.org/10.1101/2024.12.24.630236

Dorka N., Welschehold T., and Burgard W., 2023, Dynamic update-to-data ratio: minimizing world model overfitting, arXiv, 2303: 10144.
https://doi.org/10.48550/arXiv.2303.10144

Drusinsky S., Whalen S., and Pollard K., 2024, Deep-learning prediction of gene expression from personal genomes, bioRxiv, 27: 605449.
https://doi.org/10.1101/2024.07.27.605449

El-Tohamy A., Amin H., and Badr N., 2024, Integration of deep learning models for enhanced classification of viral DNA sequences across specific viruses and viral families, International Journal of Intelligent Computing and Information Sciences, 24(1): 89-104.
https://doi.org/10.21608/ijicis.2024.279692.1332

Hanczar B., Zehraoui F., Issa T., and Arles M., 2020, Biological interpretation of deep neural network for phenotype prediction based on gene expression, BMC Bioinformatics, 21(1): 501.
https://doi.org/10.1186/s12859-020-03836-4

Jaichitra I., Mohanaprakash T., Poonguzhali C., Janagiraman S., Selvakumaran S., and Maheswari B., 2023, Deep learning for breast cancer prediction in the era of big data: A comparative study of gene expression and DNA methylation, In: 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), IEEE, PP.222-229.
https://doi.org/10.1109/ICSCNA58489.2023.10370563

Ji Y., Green T., Peidli S., Bahrami M., Liu M., Zappia L., Hrovatin K., Sander C., and Theis F., 2023, Optimal distance metrics for single-cell RNA-seq populations, bioRxiv, 26: 572833.
https://doi.org/10.1101/2023.12.26.572833

Jiang J., Xing F., Zeng X., and Zou Q., 2020, Investigating maize yield-related genes in multiple omics interaction network data, IEEE Transactions on NanoBioscience, 19(1): 142-151.
https://doi.org/10.1109/TNB.2019.2920419

Lan B., 2024, Deep learning-based functional prediction model for genome sequence, In: 2024 3rd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI), IEEE, pp.207-212.
https://doi.org/10.1109/ICDACAI65086.2024.00045

Li Y., Hu M., and Shen Y., 2018, Gene regulation in the 3D genome, Human Molecular Genetics, 27(R2): R228-R233.
https://doi.org/10.1093/hmg/ddy164

Lin Y., and Jiang J., 2021, Rapid validation of transcriptional enhancers using a transient reporter assay, In: Modeling Transcriptional Regulation: Methods and Protocols, Springer US, pp.253-259.
https://doi.org/10.1007/978-1-0716-1534-8_16

Liu B., Zhang W., Zeng X., Loza M., Park S., and Nakai K., 2024, TF-EPI: an interpretable enhancer-promoter interaction detection method based on transformer, Frontiers in Genetics, 15: 1444459.
https://doi.org/10.3389/fgene.2024.1444459

Makarova A., Shen H., Perrone V., Klein A., Faddoul J., Krause A., Seeger M., and Archambeau C., 2021, Overfitting in Bayesian optimization: an empirical study and early-stopping solution, In:2nd Workshop on Neural Architecture Search (NAS 2021) @ ICLR 2021, pp.1-16.

Merelli I., Tordini F., Drocco M., Aldinucci M., Liò P., and Milanesi L., 2015, Integrating multi-omic features exploiting chromosome conformation capture data, Frontiers in Genetics, 6: 40.
https://doi.org/10.3389/fgene.2015.00040

Mikhaylova A., and Thornton T., 2019, Accuracy of gene expression prediction from genotype data with PrediXcan varies across diverse populations, Frontiers in genetics, 10: 261.
https://doi.org/10.1101/524728

Ramprasad P., Pai N., and Pan W., 2024, Enhancing personalized gene expression prediction from DNA sequences using genomic foundation models, Human Genetics and Genomics Advances, 5(4): 100347.
https://doi.org/10.1016/j.xhgg.2024.100347

Robson M., Ringel A., and Mundlos S., 2019, Regulatory landscaping: How enhancer-promoter communication is sculpted in 3D, Molecular Cell, 74(6): 1110-1122.
https://doi.org/10.1016/j.molcel.2019.05.032

Stefanini M., Lovino M., Cucchiara R., and Ficarra E., 2023, Predicting gene and protein expression levels from DNA and protein sequences with Perceiver, Computer Methods and Programs in Biomedicine, 2023, 234: 107504.
https://doi.org/10.1101/2022.09.21.508821

Tu X., and Li Y., 2024, Gene expression pattern recognition algorithm based on deep learning, In: 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA), IEEE, pp.320-325.
https://doi.org/10.1109/ICAICA63239.2024.10823051

Wang A., and Hu Q., 2023, Deep learning models for cancer classification from microarray gene expression profiles, In: 2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI), IEEE, pp.40-44.
https://doi.org/10.1109/CCAI57533.2023.10201310

Wang Z., Peng Y., Li J., Li J., Yuan H., Yang S., Ding X., Xie A., Zhang J., Wang S., Li K., Shi J., Xing G., Shi W., Yan J., and Liu J., 2024, DeepCBA: a deep learning framework for gene expression prediction in maize based on DNA sequences and chromatin interactions, Plant Communications, 5(9): 100985.
https://doi.org/10.1016/j.xplc.2024.100985

Xiao Z., Li Y., Ding Y., and Yu L., 2025, EPIPDLF: a pretrained deep learning framework for predicting enhancer-promoter interactions, Bioinformatics, 41(5): btae716.
https://doi.org/10.1093/bioinformatics/btae716

Yang Z., Su B., Cao C., and Wen, J., 2025, Regulatory DNA sequence design with reinforcement learning, arXiv, 2503: 07981.
https://doi.org/10.48550/arXiv.2503.07981

Zeng T., and Ji S., 2015, Deep convolutional neural networks for multi-instance multi-task learning, In: 2015 IEEE International Conference on Data Mining, IEEE, pp.579-588.
https://doi.org/10.1109/ICDM.2015.92

Zeng W., Wang Y., and Jiang R., 2018, Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network, Bioinformatics, 2020, 36(2): 496-503.
https://doi.org/10.1101/341214

Zhang H., Hung C., Liu M., Hu X., and Lin Y., 2019, NCNet: Deep learning network models for predicting function of non-coding DNA, Frontiers in Genetics, 10: 432.
https://doi.org/10.3389/fgene.2019.00432

Zhang T., Hasib M., Chiu Y., Han Z., Jin Y., Flores M., Chen Y., and Huang, Y., 2022, Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype predictions, Cancers, 14(19): 4763.
https://doi.org/10.3390/cancers14194763

Zhao S., Ye Z., and Stanton R., 2020, Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols, RNA, 26(8): 903-909.
https://doi.org/10.1261/rna.074922.120

Zhao Y., Li M., Konaté M., Chen L., Das B., Karlovich C., Williams P., Evrard Y., Doroshow J., and McShane L., 2021, TPM, FPKM, or normalized counts? A comparative study of quantification measures for RNA-seq data from the NCI patient-derived models repository, Journal of Translational Medicine, 19(1): 269.
https://doi.org/10.1186/s12967-021-02936-w