Comprehensive evaluation of gene sequence encoding methods in deep learning

Authors: Li Han, HU Jiming, Sun Xiaoyong, Sun Xiaoyong

Date: 2023-02-09T00:00:00+00:00

Abstract

Background: The prediction of genomic structure has become a hot spot in genome research. At present, the prediction method based on deep learning is more effective and accurate than other machine learning algorithms. Since gene sequence data cannot directly enter the deep learning model, the original data need to be encoded and converted into numerical features before model prediction. As a result, different encoding methods may affect final accuracy.

Methods: In order to explore the performance of different encoding methods, we compared ten strategies in six deep learning models. We also compared the performance of all methods on independent datasets and models from our laboratory. For all models, we used their original parameters.

Results: Dummy encoding, hash encoding, and one-hot encoding perform best in various models. In addition, dummy encoding and one-hot encoding are the best for processing RNA data, while hash encoding is superior to other methods for processing promoter data. Also, when processing part- or full-sequence data, the performance of dummy encoding, hash encoding, and one-hot encoding is similar. Besides that, in sisRNA datasets and prediction models of Arabidopsis and rice, dummy encoding and one-hot encoding achieve higher prediction accuracy.

Conclusions: We conclude that the best encoding method varies when the data set changes. One-hot encoding, dummy encoding, and hash encoding are the three best methods for six models. This study fills the gap on sequence encoding methods in deep learning and can provide a valuable reference for the community.

Full Text

Preamble

Comprehensive Evaluation of Gene Sequence Encoding Methods in Deep Learning

Han LI[1], Jiming HU[1], and Xiaoyong SUN[1],*
[1]Agricultural Big-Data Research Center, College of Information Science and Engineering, Shandong Agricultural University, Tai' an 271018, China

*Corresponding author

Author Contributions: Han LI collected and analyzed the data, performed the experiments, and wrote and revised the manuscript; Jiming HU performed the experiments; Xiaoyong SUN proposed the research problems and designed the research program.

Abstract

Background: The prediction of genomic structure has become a focal point in genome research. Currently, deep learning-based prediction methods demonstrate superior effectiveness and accuracy compared to other machine learning algorithms. Since gene sequence data cannot be directly fed into deep learning models, the raw data must be encoded and converted into numerical features before prediction. Consequently, different encoding methods may affect final accuracy.

Methods: To explore the performance of various encoding strategies, we compared ten methods across six deep learning models. We also evaluated all methods on independent datasets and models from our laboratory. For all models, we used their original parameters.

Results: Dummy encoding, hash encoding, and one-hot encoding performed best across various models. Additionally, dummy encoding and one-hot encoding were optimal for processing RNA data, while hash encoding was superior for promoter data. When processing partial or full-sequence data, the performance of these three encoding methods was similar. Moreover, in sisRNA datasets and prediction models for Arabidopsis and rice, dummy encoding and one-hot encoding achieved higher prediction accuracy.

Conclusions: We conclude that the optimal encoding method varies with the dataset. One-hot encoding, dummy encoding, and hash encoding are the three best methods for the six models evaluated. This study fills a gap in research on sequence encoding methods for deep learning and provides a valuable reference for the community.

Keywords: deep learning; RNA; promoter; encoding methods

Introduction

Genomics serves as a "microscope" for humanity to explore the mysteries of life, with various genomic features helping researchers understand the mechanisms underlying biological processes. The development of high-throughput sequencing technology has generated vast amounts of gene data, enabling comprehensive studies of genomic structure [1]. Existing research demonstrates that important genomic structures such as long non-coding RNAs (lncRNAs), circular RNAs (circRNAs), extrachromosomal circular DNA (eccDNA), and promoters play key roles in regulating biological and life activities [2-6]. Therefore, the recognition and prediction of these important genomic structures can address major problems in biology, medicine, and agriculture [7].

Currently, machine learning and deep learning have been widely applied to genomic structure prediction. As long as biological sequence data are converted into numerical features, models can automatically predict structures [8-10]. Many algorithms—including the Markov model [11], second-order Markov model [12], hidden Markov model [13,14], pseudo dinucleotide composition (PseDNC) [15-17], k-spectrum nucleotide pair frequency (KSNPF) [18], and k-space nucleotide composition (KSNC) [19,20]—have been widely used for genomic structure prediction based on machine learning. However, machine learning' s limitation lies in its need for manual feature extraction. Deep learning solves this problem by automatically learning features, thereby enabling genomic structure prediction [21-23]. Presently, one-hot encoding [24] is prevalent in nearly all research, and many mature sequence encoding methods based on deep learning have been developed, such as k-mer [25] and word vector [26]. Despite the application of many encoding methods to deep learning models, most existing studies typically adopt a single method directly. Different encoding methods may cause models to learn different features, leading to variations in prediction accuracy. Hence, exploring the effects of different sequence encoding methods on deep learning model predictions is essential.

In this study, we compared ten encoding methods: baseN encoding, binary encoding, complementary encoding, dummy encoding, effect encoding, hash encoding, index encoding, one-bit encoding, one-hot encoding, and ternary encoding. We evaluated these methods based on model prediction accuracy and conducted experiments on independent datasets and models. Compared to other methods, dummy encoding, hash encoding, and one-hot encoding performed best. Dummy encoding and one-hot encoding were more suitable for RNA data, while hash encoding was superior for promoter data. The three encoding methods performed equally well on partial and full-sequence data. Furthermore, dummy encoding and one-hot encoding also achieved the best performance on stable intronic sequence RNA (sisRNA) [27] datasets and prediction models for Arabidopsis and rice. These results demonstrate that besides one-hot encoding, dummy encoding and hash encoding can also enable models to learn more features and achieve higher prediction accuracy (Fig. 1 [Figure 1: see original paper]).

2.1. Datasets

The five sequence datasets used in this study were derived from existing studies and are described in detail below (Table 1): (1) circRNAs [28], downloaded from circBase (http://www.circbase.org/), with a final sequence length of 80 bp after preprocessing; (2) Linear RNAs, downloaded from EMBL-EBI (https://www.ebi.ac.uk/), including Arabidopsis thaliana, maize, rice, and their mixed data, with a final sequence length of 200 bp after preprocessing; (3) Zea mays lncRNAs from Meng et al. (2021) [29], using the longest sequence length as the standard and performing zero-padding on shorter sequences; (4) E. coli promoter from Shujaat et al. (2020) [30], with a sequence length of 81 bp; and (5) Saccharomyces cerevisiae promoter from Vaishnav et al. (2022) [31], with a sequence length of 110 bp. Due to hardware limitations, 100,000 sequences were randomly selected from this dataset for analysis.

2.2. Encoding Methods

This study employed ten sequence encoding methods, with all methods and experimental procedures illustrated in Fig. 2 [Figure 2: see original paper].

2.2.1. One-bit encoding: Proposed by Church et al. in 2012 [32], this method encodes each base in the DNA sequence with a single bit: A as 0, T as 1, G as 1, and C as 0.

2.2.2. Index encoding: This method converts discrete features into continuous numerical variables, normalizing discrete data so that each category has a separate index. Using index encoding, A is encoded as 1, T as 2, G as 3, and C as 4.

2.2.3. Complementary encoding: Since DNA bases pair complementarily, we propose complementary encoding to mimic this principle by encoding complementary bases with opposite numbers: A as 1, T as -1, G as 2, and C as -2.

2.2.4. BaseN encoding: This method represents base N using N unique numbers to represent all elements in the sequence, reducing the number of features while efficiently representing data and improving memory usage. Using baseN encoding, A is encoded as (0, 1), T as (0, 2), G as (0, 3), and C as (0, 4).

2.2.5. Dummy encoding: This method assigns categorical variables by converting DNA sequence data into binary variables. Unlike one-hot encoding, dummy encoding selects one category as a reference for variables with n classification attributes, generating n-1 categories (using n-1 features to represent n categories). Using dummy encoding, A is encoded as (0, 0, 1), T as (0, 1, 0), G as (1, 0, 0), and C as (0, 0, 0).

2.2.6. Ternary encoding: Similar to binary encoding but using ternary logic (0, 1, 2) that more closely resembles human thinking patterns. Using ternary

encoding, A is encoded as (0, 0, 1), T as (0, 0, 2), G as (0, 1, 0), and C as (0, 1, 1).

2.2.7. Effect encoding: Similar to dummy encoding in using n-1 features to represent n categories, but differs by replacing lines containing only 0 in dummy encoding with -1. Using effect encoding, A is encoded as (0, 0, 1), T as (0, 1, 0), G as (1, 0, 0), and C as (-1, -1, -1).

2.2.8. Binary encoding: This method converts categorical features into numerical values via an ordinal encoder, then converts these values into binary numbers. Using binary encoding, A is encoded as (0, 0, 0, 1), T as (0, 0, 1, 0), G as (0, 0, 1, 1), and C as (0, 1, 0, 0).

2.2.9. One-hot encoding: Also known as one-bit efficient encoding, this method uses n features to represent n categories and is one of the most commonly used sequence encoding methods in bioinformatics. Using one-hot encoding, A is encoded as (0, 0, 0, 1), T as (0, 0, 1, 0), G as (0, 1, 0, 0), and C as (1, 0, 0, 0).

2.2.10. Hash encoding: This algorithm converts arbitrary-length input to fixed-length output, using new dimensions to represent categorical features like one-hot encoding. However, hash encoding can represent any number of categories using n new features. Using hash encoding, A is encoded as (0, 0, 0, 0, 1), T as (0, 0, 0, 1, 0), G as (0, 0, 1, 0, 0), and C as (0, 1, 0, 0, 0).

2.3. Models

This study utilized six models, including two previously proposed in our laboratory: DeepCircRNA (http://deepbiology.cn/DeepCircRNA) and DeepAS (http://deepbiology.cn/DeepAS). DeepCircRNA is based on a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), while DeepAS combines CNN, LSTM, and Gated Recurrent Unit (GRU). We also compared the ten encoding methods across four published models from other laboratories: SpliceFinder [33], PlncRNA-HDeep [29], pcPromoter-CNN [30], and Evolution-gpu [31].

2.4. Performance Evaluation

To measure how well different encoding methods preserve sequence information, we used accuracy as the primary metric to evaluate each method' s advantages and disadvantages. Additionally, we employed each model' s original evaluation metrics, including precision, recall, and F1-score. The calculation formulas are:

Accuracy = (TP + TN) / (TP + FP + TN + FN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1-score = 2TP / (2TP + FP + FN)

where TP represents true positive samples, TN represents true negative samples, FP represents false positive samples, and FN represents false negative samples.

3. Results

In the following sections, we discuss only the influence of different encoding methods on model prediction accuracy, using all original model parameters.

3.1. Performance of Ten Encoding Methods in Six Models

3.1.1. Model Training: Using ten encoding methods to process all sequence data, we trained the corresponding models. For DeepCircRNA, baseN encoding, index encoding, and one-bit encoding yielded the lowest initial accuracy, while hash encoding and one-hot encoding achieved initial accuracy above 0.7. As training progressed, accuracy grew slowly and eventually stabilized. After training, baseN encoding, index encoding, and one-bit encoding showed significantly lower accuracy than other methods, with one-hot encoding achieving the highest accuracy (Fig. 3A [Figure 3: see original paper]).

For DeepAS, baseN encoding, index encoding, and one-bit encoding again produced the lowest initial accuracy across all datasets. Dummy encoding, hash encoding, and one-hot encoding achieved the highest accuracy in the fewest epochs, while baseN encoding and index encoding exhibited unstable performance (Fig. 3B [Figure 3: see original paper]).

For SpliceFinder, since we could not obtain the original datasets and its work was similar to our DeepAS study, we used all DeepAS datasets for training and testing. BaseN encoding, index encoding, and one-bit encoding showed the lowest initial accuracy, and their final accuracy remained significantly lower than other methods (Fig. 3C [Figure 3: see original paper]).

For PlncRNA-HDeep, one-bit encoding consistently showed lower accuracy than other methods throughout training (Fig. 3D [Figure 3: see original paper]). For Evolution-gpu, one-bit encoding produced the lowest Pearson Correlation Coefficient (PCC) throughout training (Fig. 3E [Figure 3: see original paper]). Overall, baseN encoding, index encoding, and one-bit encoding were the most unstable and least accurate across all model training processes.

3.1.2. Model Testing: For DeepCircRNA, one-hot encoding achieved the highest accuracy, while index encoding performed worst (Table 2). For DeepAS, dummy encoding achieved the highest accuracy in Arabidopsis and mixed datasets, while one-hot encoding excelled in maize and rice datasets (Table 3). Notably, baseN encoding and index encoding showed the most unstable performance, with one-bit encoding also underperforming.

For SpliceFinder, dummy encoding achieved the highest accuracy in Arabidopsis, rice, and mixed datasets, while one-hot encoding performed best in maize datasets (Supplementary Table 1). For PlncRNA-HDeep, one-hot encoding achieved the highest accuracy, F1-score, and recall, while dummy encoding

achieved the highest precision (Supplementary Table 2). For pcPromoter-CNN, after 5-fold cross-validation, hash encoding achieved the highest accuracy and Area Under the ROC Curve (AUC) (Supplementary Table 3). For Evolution-gpu, hash encoding produced the highest PCC (Supplementary Table 4).

Based on these results, we conclude that dummy encoding, hash encoding, and one-hot encoding perform best, enabling models to learn more features and achieve better prediction results.

3.2. Performance of Ten Encoding Methods with Different Data Types

Our work utilized two dataset types: RNA and promoter datasets. For RNA datasets, one-hot encoding performed best in human, maize, rice (DeepAS), and Zea mays datasets, while dummy encoding excelled in Arabidopsis, rice (SpliceFinder), and mixed datasets (Fig. 4A [Figure 4: see original paper]). For promoter datasets, hash encoding achieved the best prediction performance (Fig. 4B [Figure 4: see original paper]). Therefore, we consider dummy encoding and one-hot encoding more suitable for RNA data, while hash encoding is better suited for promoter data.

3.3. Comparison of Ten Encoding Methods with Partial- and Full-Sequence Data

Since original sequences vary in length, preprocessing is required before encoding. In this study, all data were preprocessed through sequence truncation or padding to ensure consistent length. We categorized the preprocessed data as partial sequences (intercepted from whole sequences according to specific rules) or full sequences (using a standard length with zero-padding for shorter sequences). In our datasets, human, Arabidopsis, maize, rice, mixed, and E. coli data are partial sequences; Zea mays is full-sequence data; and Saccharomyces cerevisiae promoter data include both types. For partial-sequence data, dummy encoding, hash encoding, and one-hot encoding performed best (Supplementary Fig. 1A). For full-sequence data, one-hot encoding performed best, with hash encoding and dummy encoding following closely (Supplementary Fig. 1B). For data containing both types, hash encoding performed best, followed by one-hot encoding and dummy encoding (Supplementary Fig. 1C). Thus, we conclude that these three encoding methods perform similarly on partial- and full-sequence data.

3.4. Testing of Ten Encoding Methods on Independent Datasets and Models

To verify the performance of the ten encoding methods on other genomic structures, we conducted experiments on independent datasets and models from our laboratory. Using sisRNA datasets and a newly constructed sisRNA prediction model, we tested Arabidopsis and rice with training sets of 5,000 samples and

test sets of 1,000 samples. Twenty percent of the training set was used for validation, with a 1:1 positive-to-negative sample ratio. Sequence lengths were 200 bp for Arabidopsis and 600 bp for rice, using a CNN-GRU model. Dummy encoding achieved the highest prediction accuracy for Arabidopsis, while one-hot encoding performed best for rice (Supplementary Table 5).

4. Discussion

One-hot encoding is among the most widely used methods in previous studies, and our study confirms its excellent performance. This may be because one-hot encoding uses n binary characteristics to represent n categories, allowing models to learn regular patterns. However, one-hot encoding is not irreplaceable, and other methods may achieve similar effects. Therefore, exploring alternative simple and effective encoding methods can provide more possibilities for sequence encoding.

Our study demonstrates that dummy encoding and one-hot encoding perform best on RNA data, while hash encoding excels on promoter data. Moreover, dummy encoding, hash encoding, and one-hot encoding perform well on both partial- and full-sequence data. These findings indicate that one-hot encoding is not the only suitable method for sequence data; dummy encoding and hash encoding may also help models learn more features to achieve excellent prediction accuracy.

We evaluated ten encoding methods across six models, but our research has limitations. First, we only conducted experiments on RNA and promoters. Future work will test our findings on more genomic features, including enhancers, tandem repeats, transposons, DNA methylation, and proteins. Second, we focused on five species; we plan to include more species to enrich our understanding of these encoding algorithms.

5. Conclusion

This paper compares the performance of ten encoding methods across six deep learning models based on prediction accuracy. We also evaluated these methods on independent datasets and models. The results indicate that dummy encoding, hash encoding, and one-hot encoding enable models to achieve higher prediction accuracy. Specifically, dummy encoding and one-hot encoding are most accurate for RNA data, while hash encoding performs best for promoter data. Furthermore, in Arabidopsis and rice sisRNA datasets, dummy encoding and one-hot encoding respectively achieve the highest prediction accuracy. Our study fills a gap in sequence encoding method research and provides a valuable reference for bioinformatics experts predicting important genomic structures.

ACKNOWLEDGEMENTS: This work was supported by the National Natural Science Foundation of China (grant number 32070684 to X.S.). We thank the Supercomputing Center at Shandong Agricultural University for technical

support.

References

[1] K.D. Christensen, D. Dukhovny, U. Siebert, R.C. Green, Assessing the costs and cost-effectiveness of genomic sequencing, J. Pers. Med. 5 (4) (2015) 470-486.
[2] B.L. Gudenas, L. Wang, Prediction of lncRNA subcellular localization with deep learning from sequence features, Sci Rep 8 (1) (2018) 16385.
[3] S. Wang, X. Cheng, Y. Li, M. Wu, Y. Zhao, Image-based promoter prediction: a promoter prediction method based on evolutionarily generated patterns, Sci Rep 8 (1) (2018) 17695.
[4] J. Ye, L. Wang, S. Li, Q. Zhang, Q. Zhang, W. Tang, K. Wang, K. Song, G. Sablok, X. Sun, H. Zhao, AtCircDB: a tissue-specific database for Arabidopsis circular RNAs, Brief Bioinform 20 (1) (2019) 58-65.
[5] K. Wang, H. Tian, L. Wang, L. Wang, Y. Tan, Z. Zhang, K. Sun, M. Yin, Q. Wei, B. Guo, J. Han, P. Zhang, H. Li, Y. Liu, H. Zhao, X. Sun, Deciphering extrachromosomal circular DNA in Arabidopsis, Comput Struct Biotechnol J. 19 (2021) 1176-1183.
[6] A.M. Oudelaar, D.R. Higgs, The relationship between genome structure and function, Nat Rev Genet 22 (3) (2021) 154-168.
[7] W.N. Moss, J.A. Steitz, Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA, BMC Genom 14 (2013) 543.
[8] W.Y. He, C.Z. Jia, Q. Zou, 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction, Bioinformatics 35 (4) (2019) 593-601.
[9] M.P. Niu, J. Wu, Q. Zou, Z.D. Liu, L. Xu, rBPDL: Predicting RNA-binding proteins using deep learning, IEEE J Biomed Health Inform 25 (9) (2021) 3668-3676.
[10] Y.P. Lei, S.Y. Li, Z.Y. Liu, F.P. Wan, T.Z. Tian, S. Li, D. Zhao, J.Y. Zeng, A deep-learning framework for multi-level peptide-protein interaction prediction, Nat Commun 12 (1) (2021).
[11] C. Pian, G. Zhang, F. Li, X.D. Fan, MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model, Bioinformatics 36 (2) (2020) 388-392.
[12] J. Yang, K. Lang, G. Zhang, X. Fan, Y. Chen, C. Pian, SOMM4mC: a second-order Markov model for DNA N4-methylcytosine site prediction in six species, Bioinformatics 36 (14) (2021).
[13] T.M. Ji, A Bayesian hidden Markov model for detecting differentially methylated regions, Biometrics 75 (2) (2019) 663-673.
[14] A. Dhar, D.K. Ralph, V.N. Minin, F.A. Matsen, A Bayesian phylogenetic hidden Markov model for B cell receptor sequence analysis, PLoS Comput Biol 16 (8) (2020) e1008030.
[15] W. Chen, H. Ding, X. Zhou, H. Lin, K.C. Chou, iRNA(m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition, Anal Biochem 561-562 (2018) 59.

[16] J. Song, J.J. Zhai, E. Bian, Y.J. Song, J.T. Yu, C. Ma, Transcriptome-Wide annotation of m5C RNA modifications using machine learning, Front Plant Sci 9 (2018) 519.
[17] T. Fang, Z.Z. Zhang, R. Sun, L. Zhu, J.J. He, B. Huang, Y. Xiong, X.L. Zhu, RNAm5CPred: Prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition, Mol Ther Nucleic Acids 18 (2019) 739-747.
[18] X. Chen, Y. Xiong, Y.B. Liu, Y.Q. Chen, S.D. Bi, X.L. Zhu, m5CPred-SVM: a novel method for predicting m5C sites of RNA, BMC Bioinform 21 (1) (2020) 489.
[19] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes, Comput Struct Biotechnol J 18 (2020) 906-912.
[20] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation, Plant Mol Biol 103 (1-2) (2020) 225-234.
[21] Z. Abbas, H. Tayara, K.T. Chong, 4mCPred-CNN-Prediction of DNA N4-methylcytosine in the mouse genome using a convolutional neural network, Genes (Basel). 12 (2) (2021) 296.
[22] T.H. Yang, S.C. Shiue, K.Y. Chen, Y.Y. Tseng, W.S. Wu, Identifying piRNA targets on mRNAs in C. elegans using a deep multi-head attention network, BMC Bioinform 22(1) (2021).
[23] Y. Li, F.K. Kong, H. Cui, F. Wang, C.Q. Li, J.Q. Ma, SENIES: DNA shape enhanced two-layer deep learning predictor for the identification of enhancers and their strength, IEEE/ACM Trans Comput Biol Bioinform (2022) PP.
[24] X.M. Zheng, S.G. Xu, Y. Zhang, X.X. Huang, Nucleotide-level convolutional neural networks for pre-miRNA classification, Sci Rep 9 (1) (2019) 628.
[25] M. Tahir, M. Hayat, K.T. Chong, Prediction of N6-methyladenosine sites using convolution neural network model based on distributed feature representations, Neural Netw 129 (2020) 385.
[26] J.W. Hong, R.T. Gao, Y. Yang, CrepHAN: Cross-species prediction of enhancers by using hierarchical attention networks, Bioinformatics 37 (20) (2021) 3436-3443.
[27] S.N. Chan, J.W. Pek, Stable intronic sequence RNAs (sisRNAs): an expanding universe, Trends Biochem Sci. 44 (3) (2018) 258-272.
[28] Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014. 20(11):1666-1670.
[29] J. Meng, Q. Kang, Z. Chang, Y.S. Luan, PlncRNA-HDeep: plant long noncoding RNA prediction using hybrid deep learning based on two encoding styles, BMC Bioinform 22 (Suppl 3) (2021) 242.
[30] M. Shujaat, A. Wahab, H. Tayara, K.T. Chong, pcPromoter-CNN: a CNN-based prediction and classification of promoters, Genes (Basel) 11 (12) (2020) 1529.
[31] E.D. Vaishnav, C.G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis,

D.A. Thompson, J.Z. Levin, F.A. Cubillos, A. Regev, The evolution, evolvability and engineering of gene regulatory DNA, Nature 603 (7901) (2022) 455-463.
[32] G.M. Church, Y. Gao, S. Kosuri, Next-generation digital information storage in DNA, Science 337 (6102) (2012) 1628.
[33] R.H. Wang, Z.S. Wang, J.P. Wang, S.C. Li, SpliceFinder: ab initio prediction of splice sites using convolutional neural network, BMC Bioinform 20 (Suppl 23) (2019) 652.

Figure Legends

Fig. 1: The best encoding methods across all data. We assign a value of 1 to the encoding method with the best performance and 0 to other methods. The heatmap indicates that hash encoding, dummy encoding, and one-hot encoding are the optimal methods.

Fig. 2: Ten encoding methods and complete experimental process. M1: one-bit encoding; M2: index encoding; M3: complementary encoding; M4: baseN encoding; M5: dummy encoding; M6: ternary encoding; M7: index encoding; M8: binary encoding; M9: one-hot encoding; M10: hash encoding.

Fig. 3: Training process of models. In the figure legend, the method at the top maximizes prediction accuracy (or PCC). (A) Training process of human circRNAs using ten encoding methods in DeepCircRNA. (B) Training process of Arabidopsis, maize, rice, and mixed datasets using ten encoding methods in DeepAS. (C) Training process of Arabidopsis, maize, rice, and mixed datasets using ten encoding methods in SpliceFinder. (D) Training process of Zea mays lncRNAs using ten encoding methods in PlncRNA-HDeep. (E) Training process of Saccharomyces cerevisiae promoter using ten encoding methods in Evolution-gpu.

Fig. 4: Performance of ten encoding methods in different data types. In the figure legend, the method at the top has the highest prediction accuracy (or PCC). (A) Performance in RNA datasets. Radial bar charts show that dummy encoding and one-hot encoding achieve the highest prediction accuracy. (B) Performance in promoter datasets. Radial bar charts indicate that hash encoding achieves the highest accuracy (or PCC).

Supplementary Fig. 1: Performance of ten encoding methods in full-sequence and part-sequence data. In this figure, the method with the highest prediction accuracy (or PCC) appears at the top. (A) Performance in part-sequence data. Funnel plots show that dummy encoding, hash encoding, and one-hot encoding perform best. (B) Performance in full-sequence data. Funnel plots show that one-hot encoding performs best, with hash encoding and dummy encoding following closely. (C) Performance in data containing both part and full sequences. Funnel plots show that hash encoding performs best, followed by one-hot encoding and dummy encoding.

Figures

Fig. 1: The best encoding methods of all data. We assign a value of 1 to the encoding method with the best performance, and 0 to other methods. The heatmap indicates that hash encoding, dummy encoding and one-hot encoding are the best methods.

Fig. 2: Ten encoding methods and complete experimental process. M1: one bit encoding; M2: index encoding; M3: complementary encoding; M4: baseN encoding; M5: dummy encoding; M6: ternary encoding; M7: index encoding; M8: binary encoding; M9: one-hot encoding; M10: hash encoding.

Fig. 3: Training process of models. In figure legend, At the top is the method to maximize the prediction accuracy (or PCC) of the model. A. Training process of human circRNAs using ten encoding methods in DeepCircRNA. B. The training process of Arabidopsis, maize, rice and their mixed datasets using ten encoding methods in DeepAS. C. The training process of Arabidopsis, maize, rice and their mixed datasets using ten encoding methods in SpliceFinder. D. Training process of Zea mays lncRNAs using ten encoding methods in PlncRNA-HDeep. E. Training process of Saccharomyces cerevisiae promoter using ten encoding methods in Evolution-gpu.

Fig. 4: Performance of ten enoding methods in different types of data. In figure legend, the method at the top has the highest prediction accuracy (or PCC). A. Performance of ten encoding methods in RNA datasets. Radial bar charts show that the prediction accuracy of dummy encoding and one-hot encoding is the highest. B. Performance of ten encoding methods in promoter datasets. Radial bar charts indicate that hash encoding can achieve the highest accuracy (or PCC).

Supplementary Fig. 1: Performance of ten encoding methods in full sequence data and part sequence data. In this figure, the method with the highest prediction accuracy (or PCC) is at the top. A. Performance of all encoding methods in part sequence data. Funnel plots show that the best performance is dummy encoding, hash encoding and one-hot encoding. B. Performance of all encoding methods in full sequence data. Funnel plots show that one-hot encoding performs best, and hash encoding, dummy encoding follow closely. C. Performance of all encoding methods in data with both part and full sequence. Funnel plots show that hash encoding performs best, followed by one-hot encoding and dummy encoding.

Tables

Table 1: Datasets and corresponding models.

Species	Data size	Data type	Model	Website
human	65,966	circRNAs	DeepCircRNA	http://deepbiology.cn/DeepCircRNA

Species	Data size	Data type	Model	Website
Arabidopsis	20,000	linear RNAs	DeepAS, SpliceFinder	http://deepbiology.cn/DeepAS
Maize	20,000	linear RNAs	DeepAS, SpliceFinder	https://gitlab.deepomics.org/wangruohan/S
Mixed	60,000	linear RNAs	DeepAS, SpliceFinder	https://github.com/Shujaatmalik/pcPromot CNN
Zea mays	36,000	lncRNAs	PlncRNA-HDeep	https://github.com/kangzhai/PlncRNA-HDeep
E. coli	5,720	promoter	pcPromoter-CNN	https://github.com/1edv/evolution/tree/ma
Saccharomyces cerevisiae	100,000	promoter	Evolution-gpu	

Table 2: Prediction accuracy using ten encoding methods in DeepCircRNA.

Encoding methods	Human
BaseN	
Binary	
Complementary	
Dummy	
Effect	
Index	
One bit	
One-hot	
Ternary	

Table 3: Prediction accuracy using ten encoding methods in DeepAS.

Encoding methods	Arabidopsis	Maize	Mixed
BaseN			
Binary			
Complementary			
Dummy			
Effect			
Index			
One bit			
One-hot			
Ternary			

Supplementary Table 1: Prediction accuracy using ten encoding methods in SpliceFinder.

Encoding methods	Arabidopsis	Maize	Mixed
BaseN			
Binary			
Complementary			
Dummy			
Effect			
Index			
One bit			
One-hot			
Ternary			

Supplementary Table 2: Prediction performance of PlncRNA-HDeep using ten encoding methods.

Encoding methods	Zea mays	Accuracy	F1-score	Precision	Recall
BaseN					
Binary					
Complementary					
Dummy					
Effect					
Index					
One bit					
One-hot					
Ternary					

Supplementary Table 3: Prediction performance of pcPromoter-CNN using ten encoding methods.

Encoding methods	Accuracy
BaseN	
Binary	
Complementary	
Dummy	
Effect	
Index	
One bit	
One-hot	
Ternary	

Supplementary Table 4: PCC of Evolution-gpu using ten encoding methods.

Encoding methods	Saccharomyces cerevisiae
BaseN	
Binary	
Complementary	
Dummy	
Effect	
Index	
One bit	
One-hot	
Ternary	

Supplementary Table 5: Accuracy of ten encoding methods in sisRNAs prediction model.

Encoding methods	Arabidopsis	Rice
BaseN		
Binary		
Complementary		
Dummy		
Effect		
Index		
One bit		
One-hot		
Ternary		

Note: Figure translations are in progress. See original paper for figures.