Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4150106, 8 pages
https://doi.org/10.1155/2022/4150106

Research Article

DNA Sequence Specificity Prediction Algorithm Based on Artificial Intelligence

Xiandun Zhai and Adilai Tuerxun

School of Forensic Medicine, Henan University of Science and Technology, Luoyang 471000, China

Correspondence should be addressed to Xiandun Zhai; dna@haust.edu.cn

Received 12 August 2022; Revised 3 September 2022; Accepted 21 September 2022; Published 3 October 2022

Academic Editor: Lianhui Li

DNA sequence specificity refers to the ability of DNA sequences to bind specific proteins. These proteins play a central role in gene regulation such as transcription and alternative splicing. Obtaining DNA sequence specificity is very important for establishing the regulatory model of the biological system and identifying pathogenic variants. Motifs are sequence patterns shared by fragments of DNA sequences that bind to specific proteins. At present, some motif mining algorithms have been proposed, which perform well under the condition of given motif length. This research is based on deep learning. As for the description of motif level, this paper constructs an AI based method to predict the length of the motif. The experimental results show that the prediction accuracy on the test set is more than 90%.

1. Introduction

DNA sequence specificity refers to the ability of DNA sequences to bind specific proteins [1]. These specific proteins, also known as a transcription factor (TF), play a central role in gene regulation such as transcription and alternative splicing [2]. The segment in the DNA sequence that binds to transcription factors is called the transcription factor binding site. The same transcription factor often binds to multiple sites in the sequence. These binding sites are conservative to some extent. It is generally recognized that they share the same sequence pattern, which is called a motif [3]. Finding this sequence pattern in a given DNA sequence set is called motif discovery [4]. Obtaining DNA sequence specificity is very important for establishing the regulatory model of the biological system and identifying pathogenic variants [5]. DNA sequence specificity is closely related to transcription factor binding sites. Studying DNA sequence specificity can help understand the process of DNA transcription and then establish a more complete regulatory model of biological systems. In addition, researchers can explain and identify the variation of disease by studying the specificity of DNA sequence and the relationship between specific diseases or determine the discovery of disease or disease trend in the early stage of disease through sequence level research, which is also of far-reaching significance to the medical industry.

With the development of chromatin immunoprediction sequencing (chip SEQ) and other technologies, it is becoming more and more important to use computational methods to study the specificity of DNA sequences. It takes a lot of manpower, material, and financial resources to study the specificity of DNA sequence by using biochemical experimental methods. Nowadays, the sample size of data related to biological information.

DNA sequence specificity is usually characterized at the motif level, and position weight matrix (PWM) is a representation of motifs The elements in the matrix describe the frequency of the four bases at each position. The mode of motif level characterization is easy to explain and supports the rapid scanning of binding sites on the genome scale. Based on this description, existing studies mainly detect DNA sequence specificity through motif discovery. At present, many motif mining algorithms have been proposed.

Given the length of the motif, these methods can effectively mine the motif. With the application of deep learning in some sequence problems [6–9], this study uses deep learning to study the specificity of DNA sequence.

2. Related Work

2.1. Phantom Representation. At present, the two most commonly used modes of phantom representation are consensus [10] and PWM. For highly conserved sequence sites, a consensus has good expression ability. For low conservative sequence sites, more nonpreferred character information is ignored, resulting in poor expression of consistent sequences. PWM has more flexible expression ability than consensus and can well express low conserved sequence sites. In addition, there are other ways to represent the motif. For example, IUPAC code [11], which is an improved motif representation method based on consistent sequence, contains 15 coding characters in total. IUPAC representation is more accurate than consistent sequence. In addition to expressing the most frequent characters in a column of aligned motif instances, it can also express the proportion relationship of multiple characters in some cases. In addition, the motif representation methods of consensus and PWM assume that each base position of the motif is independent of each other. In view of this situation, researchers have modeled on the basis of PWM, and the relevant models include the hidden Markov model, Bayesian network, and generalized PWM. A sequence logo [12] is a graphical representation of DNA motifs, which is composed of several columns. Each column is a stack of one or several bases. Weblog provides an interface to generate sequence logos. Users can enter a set of phantom instances to get the corresponding sequence logos.

2.2. DNA Sequence Specificity Prediction. The prediction of DNA sequence specificity at the motif level is motif discovery. Phantom discovery is a complex biological computing problem, which is often simplified when solving. A typical simplified version is the implanted motif search (PMS) problem [13].

In reality, not every sequence in a given sequence set must contain motif instances. Therefore, the definition of the PMS problem is not applicable to the actual situation. Based on this consideration, in addition to the existing parameters of PMS, researchers often introduce a parameter quorum, whose value range is any real number from 0 to 1, which represents the ratio of the number of sequences containing motif instances to the number of all sequences. This kind of PMS problem with parameter q is called qpms problem [14].

According to whether all the motifs that meet the input parameter constraints can be found, the qpms motif discovery algorithm can be divided into accurate qpms algorithm and approximate qpms algorithm. The basic idea is to search all candidate sets of motifs in the solution space of the motif, and use the first mock examination scoring function to obtain the highest scoring motif. Motif discovery is a NP hard problem. An exhaustive search of all possible situations will bring considerable time and space overhead. Time performance is an important index to evaluate the quality of accurate algorithms. In the past decade, many precise qpms algorithms have been proposed [15, 16]. Accurate algorithms based on sample pattern driving, such as pmsprune [17], stemfinder [18], travstrr [19], and qpms9, include two stages: sample driving and pattern driving. Precise algorithms based on suffix trees, such as weed [20] and fmotif [21], establish suffix tree indexes of input sequences to accelerate the verification of candidate motifs. Approximate algorithms, such as pairmotif+ [22] and qpms10 [23]. For example, pairmotif+ selects some l-mer pairs in the input sequence so that at least half of the motif instance pairs are included. Then, filter out the l-mer pairs with low weight, so that the remaining l-mer pairs contain at least one phantom instance pair. Finally, an approximation strategy is used to reduce the computational verification of candidate motifs, so that the algorithm can complete the motif discovery task in one hour.

Although there are many effective qpms algorithms, they are not effective in the discovery of motifs in real DNA sequence data. In order to deal with these problems, the algorithm does not rely on the calculation of some parameters. For example, meme [24] uses the strategy of expectation maximization to refine the motif and makes the initial motif achieve local optimization through the iterative update of step e and step M. Gibbs sampling [25] first selects part of the initial sequence to generate the initial motif, then update the initial sequence in the iterative update process through Gibbs sampling and select the updated final motif according to the score. In addition to refinement, the quality of the initial motif is also important.

The above-given algorithm is time-consuming when dealing with large data sets. Among them, the search space of qpms algorithm increases rapidly with the increase of data scale. Based on the iterative optimization algorithm, the number of initial motifs and the amount of calculation of refinement of initial motifs increase with the increase of data size. For example, samselect [26], meme chip [27], and micsa [28], select a part of the sequence from the entire data set for motif discovery, which will greatly reduce the running time, but it is difficult to identify the motif with a weak signal. Pairmotifchip [29] mining and merging similar substring pairs from the input sequence to get the motif. Its running time is mainly spent on mining similar substring pairs, which increases in square order with the increase of data size. In addition, there are algorithms based on word frequency statistics [30], such as prosampler, apms [31], and mces [32].

2.3. DNA Sequence Specificity Prediction Method at Sequence Level. In addition to the motif level, DNA sequence specificity is also characterized at the sequence level, and the most effective method is based on deep learning. In 2015, alipanahi et al. First proposed a deep learning model called deepbind for sequence level DNA sequence specificity prediction. In the same year, Zhou et al. Proposed more convolution layers and pooling layers, further proving the feasibility of applying deep learning to sequence level DNA

sequence specificity prediction. On this basis, the subsequent research has done a series of improvements, mainly including model-based improvement, data-based improvement, and improvement based on adding auxiliary features.

Sequence specificity prediction requires local continuous features learned by CNN, but it also requires global features of the sequence. The results showed that danq was more accurate than deepsea in predicting DNA sequence specificity. Lanchantin et al. Proposed the continental/highway MLP framework in 2016. The model has three convolution layers, each layer has 128 convolution cores with a length of 5, and finally connects five layers of highway MLP, of which each layer has 32 neurons. The model replaces the full connection layer with highway MLP. The results show that the prediction accuracy of the network model using the highway is higher. Chen et al. Added the kernel method on the basis of CNN to make the training speed of the model faster and more accurate. This method uses the homogeneous Gaussian dot product kernel to replace the traditional convolution kernel. Although the time to calculate the homogeneous Gaussian dot product kernel is increased, the training time of the convolution kernel is also reduced. Salekin et al. Proposed deepsnr in 2018. The model adds deconvnet layer after the convolution layer and pooling layer, which is an effective means to realize feature visualization in deep learning. This enables deepsnr not only to predict the specificity of DNA sequence, but also to output the probability that each base in the sequence belongs to the motif, and then to more accurately locate the position of the motif in the sequence. With the development of natural language processing, improved models based on attention mechanism have also been proposed. In 2019, deepgrn was proposed. This model established a model for DNA sequence specificity prediction and achieved better performance. On this basis, in 2020, researchers proposed network models based on attention mechanisms such as tbinet and deepett.

The improvement method based on data is described as follows: Zhang et al. Proposed the model hocnn based on high-order coding in 2018. The model encodes adjacent bases as one element. One hot coding based on a single base will make all data particularly independent, but in the DNA sequence, adjacent bases have an interactive effect, which is not completely independent. Zhang et al. Changed the previous coding method and adopted the data coding method of high-order coding considering the interaction and interaction of adjacent bases. It was proved that the high-order coding method was superior to the traditional method. However, because the learnable parameters of hocnn increased exponentially, its performance deteriorated with the increase of high-order degree. The results show that using high-order coding within a reasonable range will indeed improve the accuracy of the model. Zhang et al. Proposed to increase the original training data through some strategic means on the basis of the original training data unchanged in 2018. The first strategy is to introduce DNA antistrand data to participate in training, and its label is consistent with the original data. Through this means, the amount of data can be doubled; the second strategy is to add the surrounding information of the DNA sequence. For example, the original 101 long DNA sequence is expanded left and right in the genome to obtain a 151 long sequence, and then it is divided into three 101 long DNA sequences from the left, middle and right. In this way, the amount of data is expanded to three times the original. The results show that using these two strategies will make the model achieve better prediction accuracy than the original data.

The improved method based on additional information features is described as follows: Jing et al. In reality, the expression of some genes is related to their cellular characteristics, so it is of practical significance to add these characteristics. In addition, Quang et al. Accelerated the motif discovery algorithm based on expectation maximization-yamda by combining the deep learning library and GPU in 2018. Zhang et al. Proposed wscnn.

Existing methods do not utilize cofactor information of specific TF. In biological systems, specific TFs often participate in the regulation together with nonspecific TFs. These nonspecific TFs are called cofactors, and their binding sites with DNA sequences are called cofactor binding sites. When the specific TF expression signal is weak, the introduction of the information of these cofactor binding sites is expected to further improve the accuracy of the model. In addition, the existing methods do not consider that the implicit expression of the same base in different DNA sequences may be different. Usually, fixed coding is used to encode DNA sequences. The use of dynamic coding is expected to make the model more realistic, so as to further improve the accuracy of the model.

2.4. AI Technology. Today, AI is used in almost all industries, providing technical advantages for all companies that integrate AI on a large scale. According to McKinsey, compared with other analysis technologies, AI is likely to create $600billion in retail value and bring 50% incremental value to the banking industry. In transportation and logistics, potential revenue increased by 89%.

2.4.1. Application of AI in Medical Industry. The latest trend in utilizing AI healthcare systems is for every recent scenario. In recent years, it is necessary to develop such a system, which can transform classic diagnostic tools through AI based diagnosis, so as to have a better future. In addition, AI based assistive systems are also very useful for doctors who have a shortage of doctors in emergencies (such as the covid-19 pandemic). The AI based system has been trained on millions of data points collected from the electronic health records of millions of patients, and this AI based system can better diagnose the long history of patients than doctors. For example, mfine CO is one of them. It provides services for medical staff, laboratory technicians, and doctors of systems with AI functions. The platform can easily connect you with hundreds of doctors nearby. More than 700000 users have trusted mfine to become a family doctor on the road, especially in nonemergency situations. The image or visual content contains a lot of information to identify the application of instance = AI, which involves the algorithm training of these data. In the healthcare sector,

these types of data were collected to analyze radiologic reports for retinal scans. Computer vision - an AI based tool for diagnosing diseases such as cancer (India finds more than 1 million cases of cancer every year, but the country has only about 2000 pathologists with oncology experience) can get results faster and cheaper. For example, qure AI is one of the healthcare start-ups that can provide medical imaging services based on AI algorithms. It is based on the deep learning technology that uses millions of images (such as X-ray, MRI, and CT scanning) for training. AI cannot replace doctors, but it can certainly provide important help for doctors and medical staff.

2.4.2. Application of AI in Aviation Industry. Recently, many start-ups have put forward dynamic ideas for the aviation industry. AI based software can extract thousands of procurement management information from the aviation manufacturing industry. The software simplifies procurement and establishes a plan for assembling the final product. Another AI based startup will extract PDF documents and all unreadable old text files (such as aircraft manuals) and display them in readable format for new digital technologies-Stelae technologies stelae's technology can be inserted into any on premises and cloud solution to extract documents. These also include defense aircraft modules stored 20 years ago. Asia aerospace is a robotics and AI company dedicated to developing UAV based solutions to provide viable intelligence from aviation data. Their capabilities and intellectual property rights in the entire UAV technology stack (including hardware, software, and analysis) enable them to build in-depth customized aerial remote sensing tools.

2.4.3. Application of AI in Business. AI has changed the assembly line of the manufacturing industry through automation. In the early days, these tasks were performed by humans, mainly repetitive tasks. Analysis tools and CRM (Customer Relationship Management) platforms are now supported by AI algorithms such as machine learning and deep learning algorithms to provide faster, more reliable, and better services. Nowadays, it is common for chat robots to provide customer service for e-commerce companies on websites. The automation of these service industries has led to the work problems of academic and it consulting institutions. AI plays an important role in it, which promotes the development of the robot industry. Usually, robots are preprogrammed to perform some repetitive tasks, but now, intelligent robots are created by combining AI algorithms, which can perform tasks based on previous experience without explicit programming.

2.4.4. Application of AI in Travel and Transportation. AI has higher and higher requirements for the tourism industry. AI can complete all kinds of work related to travel, such as arranging travel to recommending hotels, flights, and the best routes to customers. The tourism industry is using AI driven chat robots, which can interact with customers like people to achieve a better and faster response.

2.4.5. Application of AI in Autonomous Vehicle. Recently, the automotive industry has changed the way of using AI autonomous vehicle. For example, Tesla launched the car status supported by the virtual assistant teslabot, which is a Facebook Messenger chat robot for Tesla owners. It provides services such as unlocking, positioning your car in the parking space, and keyless driving, which makes it one of the best cars ever. At present, all industries are committed to developing an autonomous vehicle, which can make your journey safer. These cars collect a large amount of data through the sensors used in the car, such as the visual data of the surrounding environment, which are processed and served in real time. An autonomous vehicle is equipped with advanced tools to collect information, including long-range radar, camera, and lidar.

2.4.6. Application of Artificial Intelligence in Education. Through artificial intelligence technology, significant changes have taken place in the education sector. For example, great changes have taken place in the interaction between teachers and students. Teaching methods and automatic scoring system can give educators more time to carry out other research work. On the other hand, students can learn in their own way and have rich resources. Chat robots can act as teaching assistants to communicate with students so that instructors have more time. AI in the future can be used as students' personal virtual tutors, which can be easily accessed anytime, anywhere.

2.4.7. Application of Artificial Intelligence in Agriculture. Since the beginning of the industrial revolution in the 19th century, the agricultural sector has been using technology. In the 21st century, the agricultural sector has taken a step forward through digitization, automation, predictive analysis, and crop monitoring. For example, the Tamil Nadu e-government Department has developed a mobile application that can scan crop images and identify diseases and provide solutions at farmers' fingertips (including disease names, pests for specific diseases).

2.4.8. Application of Artificial Intelligence in Game Industry. As early as 1949, artificial intelligence (AI) entered the game industry. It is a chess game in which a chess player plays chess against a computer. AI can be used for game purposes. AI machines can play strategic games such as chess. On machines, machines need to consider a large number of possible places. AI is deployed to build simulation similar to reality, such as human beings such as unpredictable things and human beings such as emotional things.

2.4.9. Application of Artificial Intelligence in Social Media. Technology in the 21st century has changed the human social interaction in the AI era. Various platforms have changed society through the purposes of everyone who uses this platform every day, from individuals to enterprises. But have you ever thought about how this platform works with multiple users at the same time? The general answer is AI or AI. Most of your feeds on this platform are affected by AI.

3. Method

Motif discovery is helpful to find out the sequence fragments with biological significance in DNA sequence and plays an important role in the study of gene expression regulation. At present, a large number of motif discovery algorithms have emerged, which mainly determine the length of the motif through the following three methods: the length of the motif is specified by the user before the discovery of the motif. The fragments with high information content are cut out from the extension matrix found by the motif, and the length of the motif is determined according to this. These methods to determine the length of the phantom have some shortcomings, such as difficult to determine the threshold, time-consuming, and inaccurate.

3.1. Basic Ideas. This section takes the ATF3 motif as an example to show its corresponding training sample. The sites of the ATF3 motif in the DNA sequence were extended to both sides, and the extension site set P of the ATF3 motif was obtained. Take V as a training sample of ATF3, and take the length of ATF3 as a label.

$$M_{ij} = \frac{\mathrm{occ}(P, i, j)}{\sum_{x\in\sigma} \mathrm{occ}(P, x, j)},$$
$$V_j = \sum_{i\in\sigma} M_{ij} \ln \frac{M_{ij}}{b_i}. \tag{1}$$

This paper makes the machine learning model automatically extract and learn the characteristics of high relative entropy fragments in the relative entropy vector.

3.2. Overall Framework of the Model. The overall framework of predicting phantom length is shown in Figure 1, including sample data construction, prediction model construction, and prediction model application.

3.3. Sample Data Construction. In this paper, the PWM of human transcription factor binding sites was obtained in the Jaspar database, and the corresponding chip SEQ data set was obtained in encode database. The training samples need to be constructed from the fragment of the phantom. If the motif discovery algorithm finds a motif fragment, then this fragment is often a fragment with high relative entropy. In addition, when constructing samples, the length k of the high relative entropy fragment should not be too different from the real motif length L, otherwise many sites located by this fragment may not be real motif sites, thus affecting the quality of the constructed samples. Therefore, the samples are constructed by intercepting the segments with high relative entropy and large length in the real phantom PWM. Specifically, given the length L of a real phantom, its range is generally 8 to 21, and the range of interception length k is determined by the following formula:

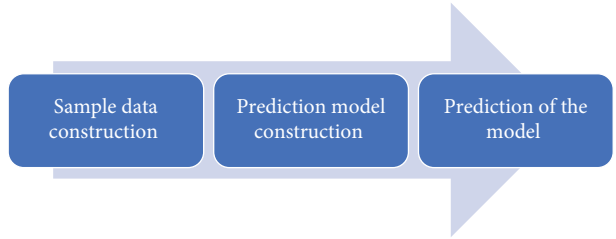

Figure 1: Overall prediction process of the model.

$$k \in \begin{cases} [l-3, l], & l \geq 11, \\ [8, l], & 8 \leq l < 11. \end{cases} \tag{2}$$

Some of the appearance sites of motif fragments were screened instead of all the appearance sites to construct samples. PWM detection tool moods is used to locate the occurrence site of the whole motif or motif fragment in DNA sequence data set D. It is found that many loci in p-do not fall into P. If the samples are directly constructed with all sites in P-, the quality of the samples may be relatively low, that is, the complete phantom information may be missing in the constructed relative entropy vector v.

For the obtained real motifs and chip SEQ data sets, the number of chip SEQ data sets corresponding to motifs of different lengths is not uniform, especially since some motifs do not find the corresponding chip SEQ data sets.

3.4. Prediction Model Construction. Figure 2 is a schematic diagram of a common neural network. CNN uses convolution to check different channels for convolution and finally fuses the convoluted results. The matrix is defined as follows:

$$f\left(M_{\mathrm{input}}\right) = \mathrm{net}\left(\mathrm{pool}\left(\left(\mathrm{conv}\left(M_{\mathrm{input}}\right)\right)\right)\right). \tag{3}$$

The reason why ReLu is selected here is that it is simple to realize and has fast convergence speed. Use maximum pooling for sampling, and set the size of the pooling window to 4. The activation function is defined as follows:

$$\mathrm{ReLU}(x) = \max\{0, x\}. \tag{4}$$

The reconstructed vector is fully connected with 14 neurons in the output layer, and the parameter is 1280×14 weights and the same number of paranoid items. The dropout operation is performed during full connection, and some hidden layer neurons are randomly discarded during training, which can effectively prevent the model from overfitting. The value of dropout is set to 0.25. The softmax function is defined as follows:

$$\mathrm{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{14} e^{z_j}}. \tag{5}$$

The cross entropy function is used as follows:

$$H_{y'}(y) = -\sum_{i=0}^{14} y_i \log(y_i). \tag{6}$$

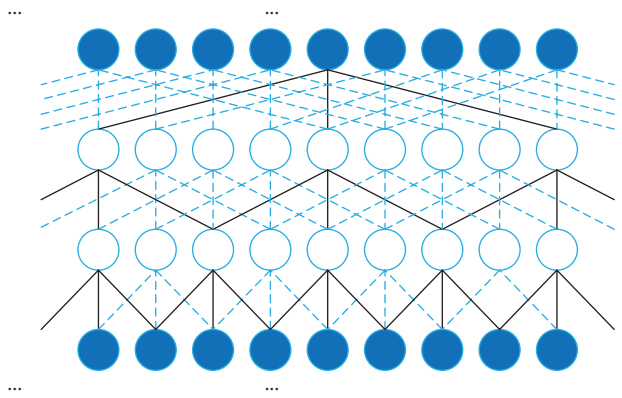

Figure 2: Schematic diagram of neural network.

In the training phase of the model, set 30 epoch iterations (one epoch is the process of training all training samples once), and the batchsize (the number of samples selected for one training) is 500.

4. Experiment

4.1. Experimental Setting. PWM of 280 human models were collected in Jaspar database. The length of these PWM ranges from 8 to 21. The number of PWM corresponding to each PWM length is 20. Among the 280 PWM collected, 74 PWM can obtain their corresponding chip SEQ data in encode database. For the other 206 PWM, simulated DNA sequence data sets are generated for them, respectively. During the experiment, the model was implemented on a Windows environment with a single CPU of 2.4 GHz and 16 GB memory. MPC is defined as follows:

$$mPC = \frac{\text{len}_{\text{overlap}}(m_p, m_k)}{l + l' \text{len}_{\text{overlap}}(m_p, m_k)}. \quad (7)$$

4.2. Model Validation. Cross validation method is used to verify the model. In order to avoid the samples generated by the same PWM appearing in the training set and the verification set at the same time, the sample set is divided according to PWM. For each discount data, train a model according to the other 10% discount data, use this model to predict the discount samples that do not participate in the training, and calculate its prediction accuracy.

Select 20 PWM with tags 18-21 corresponding to different transcription factors (TFs). For each PWM, the fragments with the highest relative entropy whose difference from the original length is no more than 10 are intercepted, and the prediction samples are generated from them.

We analyze the prediction accuracy of motiflen under different length segments of the real phantom. It can be found that the greater the difference between the fragment length and the length of the real phantom, the lower the average prediction accuracy.

4.3. Optimize Existing Motif Discovery Algorithms. Motiflen can not only optimize the motifs found by the existing motif discovery algorithm but also optimize their time performance. Next, take the meme chip as an example to introduce the optimization method and optimization effect of motif.

First, motiflen can be used to optimize the motifs found by existing motif discovery algorithms. Meme chip, a well-known motif discovery algorithm, is taken as an example for discussion. In the experiment, the meme chip mining length interval was set to be 8 to 21, and the first 3000 sequences in the DNA sequence data set were taken for motif discovery. Meme chip was run on 74 sets of real data sets for phantom discovery, and the PWM found under 40 sets of real data sets was selected, which met the overlap with the published PWM. Then, from these found PWM, a prediction sample is generated, a new PWM length is predicted with motiflen, and a new PWM is obtained. This shows that by optimizing the results of the meme chip, motiflen can get a module length closer to the real PWM.

Secondly, motiflen can be used to improve the time performance of existing motif discovery algorithms. For meme chip, the time consumption of mining mode with fixed motif length is significantly less than that of mining mode with set motif length interval. In this way, motiflen is used to optimize the results of the mining mode of fixed module length in meme chip, and this method is used to replace the mining mode of setting the module length interval, so as to improve the time performance of the algorithm. In the experiment, the length of the meme chip mining module is fixed to 11 for module discovery, and then the PWM with the length of 11 is used to generate prediction samples, and a new PWM is obtained with motif. Because the running time of motiflen can be ignored, the acceleration ratio of the running time is taken as the ratio of the running time of the meme chip with a fixed length of 11 and the running time of the meme chip with a set length range of 8 to 21. We analyzed the results on eight PWM. It can be found that the optimized PWM obtained by motiflen is closer to the published PWM than the fixed length 11 PWM found by meme chip. Moreover, motiflen's correction of the PWM found by the meme chip can not only correct the shorter one to the longer one. When the PWM found by the meme chip is longer than the actual PWM, the model still has the repair function for it. On this basis, motiflen has increased the running time of meme chip by more than 2 times. In addition, the same strategy is used to optimize fmotif algorithm, which is an accurate motif discovery algorithm suitable for large data sets of DNA sequences, and it costs a lot of time. It can be found that the acceleration ratio of motiflen to fmotif can reach more than 100.

5. Conclusion

In reality, the length of the motif is unknown when it is discovered. Based on the shortcomings of the existing module length determination algorithm, this chapter proposes a set of overall solutions to predict the module length with supervised deep learning. Firstly, the motiflen model is verified, and the accuracy of motif length prediction of motiflen on the test set is more than 90%. Experiments show that in the collected DNA sequence data, the longer the length of the real phantom corresponding to the predicted sample, the higher the average prediction accuracy of motiflen. This is because the MPC index has a preference for samples with larger labels under the same absolute error. Select 20 TFs with labels ranging from 18 to 21, intercept fragments with a difference of no more than 10 from the original length according to the relative entropy, generate samples and input them to the motif. The results show that the greater the difference between the length of the motif and the real phantom, the lower the average prediction accuracy of the fragment, but the higher the complementation ability of the motif. Later, we will consider using feature reconstruction and data enhancement methods to iteratively optimize the results to make the results better.

Data Availability

The dataset can be accessed from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Trabelsi, M. Chaabane, and A. Ben-Hur, "Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities," *Bioinformatics*, vol. 35, no. 14, pp. i269–i277, 2019.

[2] M. David, "Transcription factors and DNA play hide and seek," *Trends in Cell Biology*, vol. 30, no. 6, pp. 491–500, 2020.

[3] P. D'haeseleer, "What are DNA sequence motifs," *Nature Biotechnology*, vol. 24, no. 4, pp. 423–425, 2006.

[4] T. L. Bailey, "STREME: accurate and versatile sequence motif discovery," 2020.

[5] S. S. Nishizaki, N. Ng, S. Dong et al., "Predicting the effects of SNPs on transcription factor binding affinity," *Bioinformatics*, vol. 36, no. 2, pp. 364–372, 2020.

[6] Z. Jia, Junyu, X. Zhou, and Y. Zhou, "Hybrid spiking neural network for sleep EEG encoding," *Science China Information Sciences*, vol. 65, 2022.

[7] Z. Jia, Y. Lin, J. Wang et al., "Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1977–1986, 2021.

[8] T. Jia, C. Cai, and Y. Hu, "Forecasting citywide short-term turning traffic flow at intersections using an attention-based spatiotemporal deep learning model," *Transportation Business: Transport Dynamics*, pp. 1–23, 2022.

[9] Z. Jia, X. Cai, and Z. Jiao, "Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging," *IEEE Sensors Journal*, vol. 22, no. 4, 2022.

[10] T. D. Schneider, "Consensus sequence Zen," *Applied Bioinformatics*, vol. 1, no. 3, pp. 111–119, 2002.

[11] T. Marschall and S. Rahmann, "Efficient exact motif discovery," *Bioinformatics*, vol. 25, no. 12, pp. i356–364, 2009.

[12] Z. Dong, "An overview of sequence logo technique and potential application direction," *The Frontiers of Society, Science and Technology*, vol. 2, no. 11, pp. 51–57, 2020.

[13] P. A. Pevzner and S. H. Sze, "Combinatorial approaches to finding subtle signals in DNA sequences," *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, vol. 8, pp. 269–278, 2000.

[14] M. Federico, P. Valente, M. Leoncini, M. Manuela, and C. Roberto, "An efficient algorithm for planted structured motif extraction," *Proceedings of the 1st ACM Workshop on Breaking Frontiers of Computational Biology*, pp. 1–6, Ischia, Italy, May 2009.

[15] F. Zambelli, G. Pesole, and G. Pavesi, "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 225–237, 2013.

[16] B. Liu, J. Yang, Y. Li, A. Mcdermaid, and Q. Ma, "An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data," *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 1069–1081, 2018.

[17] J. Davila, S. Balla, and S. Rajasekaran, "Fast and practical algorithms for planted (l,d) motif search," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 4, pp. 544–552, 2007.

[18] Q. Yu, H. Huo, J. S. Vitter, J. Huan, and Y. Nekrich, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 2, pp. 384–397, 2015.

[19] S. Tanaka, "Improved exact enumerative algorithms for the planted (l,d)-motif search problem," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pp. 361–374, 2014.

[20] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucleic Acids Research*, vol. 32, pp. 199–203, 2004.

[21] C. Jia, M. B. Carson, Y. Wang, Y. Lin, and H. Lu, "A new exhaustive method and strategy for finding motifs in ChIP-enriched regions," *PLoS One*, vol. 9, no. 1, Article ID e86044, 2014.

[22] Q. Yu, H. Huo, Y. Zhang, H. Guo, and H. Guo, "PairMotif+: a fast and effective algorithm for de novo motif discovery in DNA sequences," *International Journal of Biological Sciences*, vol. 9, no. 4, pp. 412–424, 2013.

[23] P. Xiao, S. Pal, and S. Rajasekaran, "qPMS10: a randomized algorithm for efficiently solving quorum Planted Motif Search problem," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pp. 670–675, Shenzhen, China, December 2017.

[24] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Research*, vol. 34, no. Web Server, pp. 369–373, 2006.

[25] X. Sheng and K. Wang, "Motif identification method based on Gibbs sampling and genetic algorithm," *Cluster Computing*, vol. 20, no. 1, pp. 33–41, 2017.

[26] Q. Yu, D. Wei, and H. Huo, "SamSelect: a sample sequence selection algorithm for quorum planted motif search on large DNA datasets," *BMC Bioinformatics*, vol. 19, no. 1, pp. 228–243, 2018.

[27] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," *Bioinformatics*, vol. 27, no. 12, pp. 1696-1697, 2011.

[28] V. Boeva, D. Surdez, N. Guillon et al., "De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis," *Nucleic Acids Research*, vol. 38, no. 11, p. e126, 2010.

[29] Q. Yu, H. Huo, and D. Feng, "PairMotifChIP: a fast algorithm for discovery of patterns conserved in large ChIP-Seq data sets," *BioMed Research International*, vol. 2016, Article ID 4986707, 10 pages, 2016.

[30] Y. Li, P. Ni, S. Zhang, Z. Su, and G. Li, "ProSampler: an ultra-fast and accurate motif finder in large ChIP-Seq datasets for combinatory motif discovery," *Bioinformatics*, vol. 35, no. 22, pp. 4632–4639, 2019.

[31] Q. Yu, X. Zhang, Y. Hu, S. Chen, and L. Yang, "A method for predicting DNA motif length based on deep learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

[32] Q. Yu and X. Zhang, "A new efficient algorithm for quorum planted motif search on large DNA datasets," *IEEE Access*, vol. 7, no. 1, Article ID 129626, 2019.