HOKKAIDO UNIVERSITY

Title	Predicting PRDM9 Binding Sites by a Convolutional Neural Network and Verification Using Genetic Recombination Map
Author(s)	Nakamura, Takahiro; Endo, Toshinori; Osada, Naoki
Citation	IPSJ Transactions on Bioinformatics, 15, 9-16 https://doi.org/10.2197/ipsjtbio.15.9
Issue Date	2022
Doc URL	https://hdl.handle.net/2115/96102
Rights	Notice for the use of this material The copyright of this material is retained by the Information Processing Society of Japan (IPSJ). This material is published on this web site with the agreement of the author (s) and the IPSJ. Please be complied with Copyright Law of Japan and the Code of Ethics of the IPSJ if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof. All Rights Reserved, Copyright (C) Information Processing Society of Japan. Comments are welcome. Mail to address editj@ipsj.or.jp, please.
Type	journal article
File Information	IPSJ Trans. Bioinf. 15_9-16.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

IPSJ Transactions on Bioinformatics Vol.15 9–16 (Mar. 2022)

[DOI: 10.2197/ipsjtbio.15.9]

Original Paper

Predicting PRDM9 Binding Sites by a Convolutional Neural Network and Verification Using Genetic Recombination Map

Takahiro Nakamura[1] Toshinori Endo[1,2] Naoki Osada[1,2,a)]

Received: October 26, 2021, Accepted: December 17, 2021

Abstract: PR domain-containing 9 (PRDM9) is a zinc-finger protein that binds to specific DNA motifs and induces the crossing-over between chromosomes, resulting in a high recombination rate around binding sites. Currently, the binding sites of PRDM9 are predicted with methods based on motif matching and Position-specific Weight Matrix (PWM). Meanwhile, the Convolutional Neural Network (CNN) has shown superior performance in recent studies to identify protein-binding regions in general, and it is expected to perform well in PRDM9 binding site prediction. In this study, we compared the performance of PWM and CNN for predicting PRDM9 binding sites with not only test data but also the correlation between the prediction score for a fragment and the local recombination rate to evaluate the performance without overfitting effects. Approximately 170,000 genomic DNA fragments of the human genome containing the Chromatin Immuno-Precipitation with high-throughput sequencing (ChIP-seq) peak of PRDM9 were used for constructing PWM and CNN. We found that CNN outperformed PWM in terms of area under the ROC curve and other metrics. Furthermore, the prediction scores of CNN correlated more strongly with the local recombination rate than PWM. We discuss that the superior performance of CNN would be in part due to the ability of CNN to capture the feature of surrounding sequences of actual PRDM9-binding sites.

Keywords: binding site prediction, CNN, PRDM9, genetic recombination

1. Introduction

Genetic recombination is a fundamental and important process in eukaryotes, and studies are underway to elucidate the mechanism of recombination hotspots, where recombination frequently occurs in early meiosis. PR domain-containing 9 (PRDM9) is a zinc finger protein that induces chromosomal recombination in mammalian germ cells. It binds to the DNA sequence in a sequence-specific manner with its zinc finger domain and localizes in a narrow (1–2 kb) hotspot [1], [2], [3], [4]. PRDM9 trimethylates the surrounding histones at H3K4 and H3K36 with its PR/SET domain [5], [6], [7], [8]. Topoisomerase sporulation-specific 11 (SPO11) is followingly recruited, which forms a double-strand break, and crossing-over occurs at the site [8], [9]. *PRDM9* is highly polymorphic [10]; there are three major alleles, A, B, and C alleles, which consist of 90%, 5%, and 1.5% of alleles in the non-African population, respectively [2], [4].

The binding sites of PRDM9 can be identified by Chromatin Immuno-Precipitation with high-throughput sequencing (ChIP-seq) experiments [8]. However, it is impossible to determine all the binding sites using a limited number of experimental conditions [11], which hampers the comprehensive understanding of the DNA-protein binding mechanisms. Previous studies showed that the canonical 13-mer binding motif of PRDM9 (CCNCC-NTNNCCNC), which is enriched in hotspots, is insufficient to explain all the hotspots [1], [2], [12], [13].

There are two approaches for exploring the factors that determine hotspots. One is to investigate features of hotspots without using PRDM9 binding information. In recent years, important features of recombination hotspots were captured by neural networks [14], [15]. However, these have not yet been able to fully explain the features. Another approach is to investigate the role of PRDM9 binding for producing recombination hotspots. The relationship among PRDM9 binding, chromatin state, and recombination was explored in many studies using mice and primates [16], [17], [18]. [19], [20], [21].

In the latter approach, to predict the binding sites of PRDM9, simple motif string searches and Position-specific Weight Matrix (PWM) have been used in previous studies. PWM represents the DNA motifs to which a protein binds [22]. It is created on the basis of the frequency of nucleotides at each position or the binding properties predicted from the amino acid sequence of the zinc finger [23], [24]. Meanwhile, Convolutional Neural Network (CNN), which can learn complex features from large amounts of DNA sequence data, has been reported to predict protein-binding sites more accurately than other methods including PWM [25]. Although CNN could be a superior prediction method than PWM for PRDM9 binding site prediction, there has been no comparison of PRDM9 binding prediction between PWM and CNN.

[1] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[2] Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[a)] nosada@ist.hokudai.ac.jp

© 2022 Information Processing Society of Japan

In general, when evaluating a model, the prediction accuracy is verified using test data obtained in advance, apart from the training data used to tune the model parameters [26]. However, as overfitting to noise is expected [25], the evaluation would not be independent of the original data. Therefore, it is difficult to use only one experimental data to evaluate the generalized performance of models. To amend the effect of overfitting to one type of training data, combining a validation method using test data and a method that does not require test data is desirable.

In this study, we compared the performance of the PWM and CNN in terms of area under the ROC curve (AUC) and other metrics using test data. We also examined the correlation between the predicted binding score for a genomic fragment and the recombination rate of the region to test whether CNN outperforms PWM, which can avoid overfitting effects. We observed a stronger correlation between the prediction scores and local recombination rates in CNN than in PWM. In addition, we revealed that the DNA fragments with high prediction scores in CNN had higher local recombination rates, even without ChIP-seq peaks. Further, we examined the feature that CNN recognizes and discussed the reason for the different performance from PWM.

2. Materials and Methods

2.1 Materials

We retrieved data of ChIP-seq experiments performed by Altemose et al. (2017) to obtain DNA-PRDM9 binding sequence data [8] (https://www.ncbi.nlm.nih.gov/geo/, GSE99407). The experiments obtained 170,198 nonoverlapping fragments by transfecting HEK293T cell line with N-terminal YFP-tagged human PRDM9 of B-allele. Each ChIP-seq peak was placed at the center of a 301-bp region. We assigned around 10% of the peak regions (17,009) to positive test data and the remaining 90% (153,189) to positive training data for CNN. For negative data, non-overlapping 100-bp-length sequences that are at least 1,000 bp away from any ChIP-seq peaks were extracted from the human reference genomes (hg19).

We used a recombination map estimated by pyrho [21] considering the population demography using genomic polymorphism data. In this study, we used the estimations based on the demography of the CEU (Utah residents with northern and western European ancestry) population.

Overview of the workflow for validation with test data is represented in **Fig. 1**.

2.2 Preparation for Test Data

To verify the prediction accuracy of CNN and PWM, we created PRDM9-binding test data. Positive test data were generated by cutting out 100-bp-length DNA fragments containing 31 bp around the peaks (ChIP-seq core region) at random positions within the fragments. Negative test data were randomly sampled so that the positive data constituted 2% of total test data, based on the ratio of the ChIP-seq-positive to -negative regions, yielding 833,441 negative test regions.

2.3 PWM Used in the Study and its Scoring

We used PWMs obtained from the same ChIP-seq data and

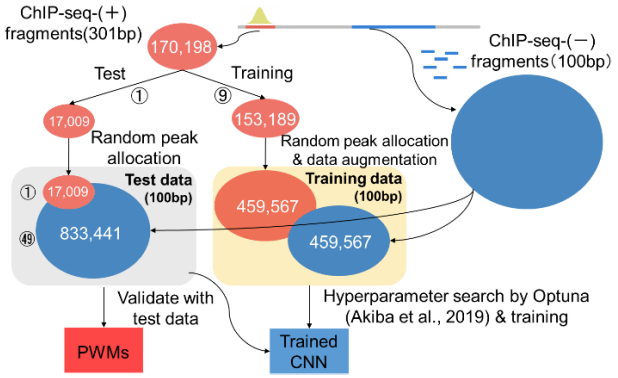

Fig. 1 Overview of data preparation and prediction. The ChIP-seq-positive and -negative DNA fragments were assigned to test and training datasets and used for evaluation. The numbers in the red and blue circles represent the number of fragments.

constructed using the Bayesian de novo motif-finding algorithm [27] (Additional file of Altemose et al. [8]). In total, 17 PWMs were obtained.

For scoring, the log-likelihood ratio score S, which is obtained based on the occurrence probability of a motif considering the nucleotide frequency in the background sequence, was calculated. The nucleotide frequencies of the background sequences were calculated for each chromosome and strand. The score $S_{i,j,k}$ is calculated by summing up the log-likelihood values for all sites of the j-th PWM, starting from the k-th site of the i-th fragment for both strands [28].

Next, $S_{i,j,k}$ was summarized to the score for the i-th fragment and the j-th PWM, $S_{i,j}$, expressed as follows:

$$S_{i,j} = \max_{k} S_{i,j,k} \tag{1}$$

There are two ways to summarize the score for the i-th fragment. One is to accept the maximum score among N PWMs, $S_{\max,i}$, expressed as follows:

$$S_{\max,i} = \max_{1 \le j \le N} S_{i,j} \tag{2}$$

The other method is to sum the scores of all N PWMs, $S_{\mathrm{sum},i}$, expressed as follows:

$$S_{\mathrm{sum},i} = \sum_{j=1}^{N} 2^{S_{i,j}} \tag{3}$$

We determined how many PWMs should be used for the prediction. All PWMs were ranked by the probability of hitting within 100 bp of ChIP-seq peaks. We added PWMs one by one from the one with the maximum score and evaluated the performance in terms of AUC using $S_{\max,i}$ and $S_{\mathrm{sum},i}$. As a result, AUC was maximized using up to the 15th PWM with $S_{\mathrm{sum},i}$, and the condition was adopted as the final PWM scoring (**Table 1**).

2.4 CNN and its Scoring

We modified the equivariant Bayesian convolutional network (EBCN) [14] written in Python using deep learning libraries: TensorFlow-GPU (Version 1.13.1) [29] and Keras-GPU (Version 2.3.1). The model employs Monte Carlo (MC) dropout [30], in which the dropout is performed during not only training but also

© 2022 Information Processing Society of Japan

Table 1 AUC values using different cutoff of the PWM number.

# of PWM used	Max Method	Sum Method
1	0.8330	0.8330
2	0.8363	0.8383
3	0.8329	0.8358
4	0.8250	0.8279
5	0.8301	0.8339
6	0.8364	0.8403
7	0.8414	0.8451
8	0.8449	0.8488
9	0.8469	0.8510
10	0.8503	0.8557
11	0.8528	0.8584
12	0.8512	0.8565
13	0.8568	0.8615
14	0.8577	0.8623
15	0.8582	0.8629
16	0.8203	0.8294
17	0.7941	0.8040

Table 2 One-hot encoding for nucleotides.

Nucleotide	One-hot vector
A	$[1,0,0,0]^T$
C	$[0,1,0,0]^T$
G	$[0,0,1,0]^T$
T	$[0,0,0,1]^T$

prediction, and produces the average of the repeated output predictions. In addition, the model is internally adjusted so that the output for an input forward sequence is equal to the output of its reverse complement sequence. These features make the accuracy less dependent on the amount of training data, the internal representations simple [31], and inconsistency in the model accuracy from training to training small. It also has an advantage for understanding features CNN detects (i.e., finding motifs), and we adopt this model. The input is a DNA fragment of constant length, which is converted into a matrix of 4 × input length by one-hot vectorizing the nucleotides (**Table 2**). Then, the matrix undergoes multiple convolutional and pooling layers, and scores (0–1) for PRDM9 binding through a dense layer with softmax function are output.

2.5 Creating Training Data for CNN and the Network Optimization

Positive training data were created using 90% of the ChIP-seq-positive region using the same method for generating the positive test data (Fig. 1). To construct a robust CNN, we performed data augmentation for ChIP-seq-positive data. We trained the model using three-fold augmented positive fragments with randomly shifting the peak three times for each fragment. ChIP-seq-negative data were prepared to match the number of positive augmented data. The prediction without the augmentation showed smaller AUC than the prediction with the augmentation (data not shown).

We employed the EBCN structure (recombination topology) developed by Brown and Lunter (2019) [14] and further optimized the network structure and hyperparameters. We tuned the number of filters in each convolutional layer, kernel size of the first convolutional layer, number of internal convolutional layers, kernel size of inner convolutional layers, pool size, learning rate, coefficient of the L2 normalization term, dropout rate, batch size, optimizer, and activation function of the convolutional layer (**Table 3**). In particular, batch size was optimized from among 32, 64, 128, 256, 512, 1024, and 2048, optimizer was from among SGD [32], Momentum [33], and Adam [34], and activation function was from among ELU [35], ReLU [36], SELU [37], and LReLU [38]. The hyperparameter auto-optimization library, Optuna, was employed [39].

Table 3 List of hyperparameters and searching range and step.

Hyperparameters	Minimum	Maximum	Step
The number of filters	2	100	2
The kernel size of first filter	2	40	2
The number of internal conv. layers	1	3	1
The kernel size of internal filter	2	4	1
Pool size	2	6	2
L2 coefficient	0	0.01	0.0001
Learning rate	0.0001	0.1	None
Dropout rate	0	1	None

Table 4 Optimized network structure of final CNN.

Layers	Parameters
Equivariant Conv1D Layer 1	Filters = 88; kernel size = 14; activation function = LReLU
Equivariant MC Dropout Layer 1	Dropout rate = 0.3980
Spatial MaxPooling1D Layer	Pool size = 2
Equivariant Conv1D Layer 2	Filters = 88; kernel size = 3; activation function = LReLU
Equivariant MC Dropout Layer 2	Dropout rate = 0.3980
Reverse Complement Sum Pooling Layer	None
Global Spatial MaxPooling1D Layer	None
Dense Layer	Activation function = Softmax

Table 5 Hyperparameters of backpropagation.

Hyperparameter	Optimal Value
Learning rate	1.737×10^{-3}
L2 coefficient	0.0
Batch size	128
Optimizer	Adam

Approximately 10% of the training data were used for the hyperparameter optimization. The epoch was set to 20, and the trials were conducted 100 times. After the hyperparameter search, if the network structure conflicts with equivariance, we trimmed at most 1 bp of the input sequence (**Tables 4**, **5**).

3. Results

3.1 CNN Outperformed PWM in Prediction Accuracy Using Test Data

We trained CNN after the hyperparameter optimization. The training was repeated five times, 50 epochs for each, and the model with the highest accuracy was adopted as the final model. We compared the AUC of CNN and PWM, using the test data. The AUC of PWM and CNN were 0.8629 and 0.9167, respectively (**Fig. 2**). To evaluate using other statistics, the optimal threshold was decided using the ROC curve. The point whose

© 2022 Information Processing Society of Japan

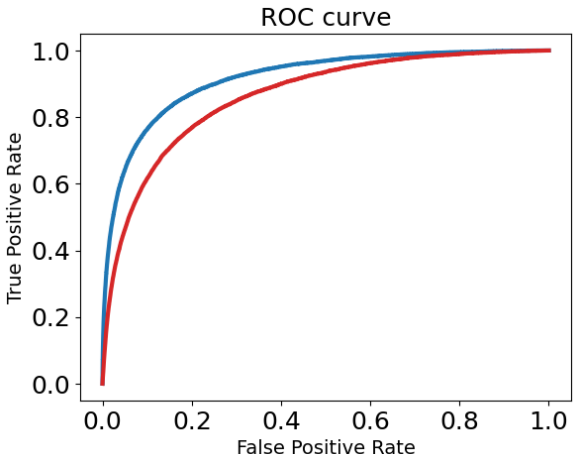

Fig. 2 Receiver operating characteristic (ROC) curve with test data. The red and blue lines represent the ROC curves of PWM and CNN, respectively.

Table 6 Metrics of each prediction method.

Metrics	CNN	PWM
Accuracy	0.8399	0.7911
Sensitivity (True positive rate)	0.8395	0.7778
Specificity	0.8399	0.7914

coordinates were closest to (0, 1) was used for the threshold [40]. Accuracies of CNN and PWM were 0.8399 and 0.7911, respectively. In **Table 6**, we also show sensitivity and specificity.

3.2 CNN Outperformed PWM in Prediction Accuracy Validated Using Recombination Map

To evaluate prediction power using recombination rate, the entire autosomal genome was divided into fragments of 100-bp length from the start position, and fragments without ambiguous nucleotide (N) were selected. For each fragment, the average recombination rate (cM/bp) was calculated from the recombination map. We predicted the PRDM9 binding score for each fragment using CNN and PWM. The correlation coefficient (ρ) between the prediction score and recombination rate was statistically significant for CNN ($\rho = 0.180$, p-value $< 1.0 \times 10^{-47}$) and for PWM ($\rho = 0.138$, p-value $< 1.0 \times 10^{-47}$), but the correlation coefficient was higher in CNN than in PWM.

For the visualization shown in **Fig. 3**, the scored fragments were sorted in descending order by prediction score, and fragments were grouped into 10 bins (10% of data for each bin). The mean and standard error of recombination rate in each bin was calculated and shown in Fig. 3. For CNN, the average recombination rate was high in bins with high prediction scores and low in bins with low prediction scores (Fig. 3 B). To observe the correlation within the group of high scores, we grouped the top 10% of fragments into 1% bins and examined the average recombination rate. Again, CNN showed a stronger correlation than PWM (Fig. 3 A).

3.3 Can PWM and CNN Detect Potential Recombination Hotspots without ChIP-seq Peaks?

To examine whether PWM and CNN detect potential recombination hotspots missed by the ChIP-seq experiments, we plot the prediction score and average recombination rate in each bin for

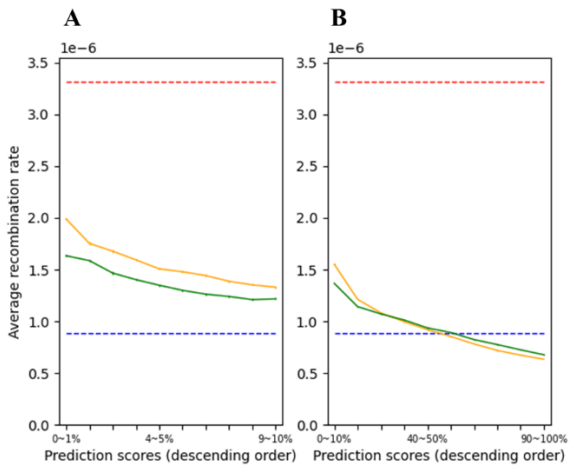

Fig. 3 Correlation between recombination rate and prediction score for all fragments. The yellow and green lines represent the mean recombination rate in bins for CNN score and PWM score, respectively. The error bars are drawn in the figure but are nearly invisible because of the large sample sizes. The average recombination rate of all ChIP-seq-negative and -positive fragments are shown in the blue and red dashed lines, respectively. A) The top 10% of data were divided into 10 bins (1% of all data for each bin). B) All data were divided into 10 bins (10% of all data for each bin).

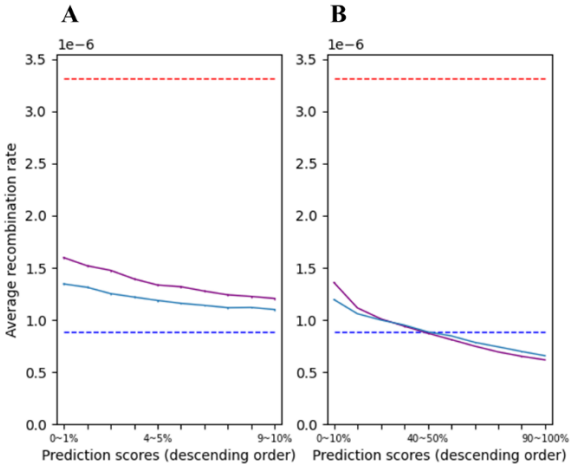

Fig. 4 Correlation between recombination rate and prediction score for ChIP-seq-negative fragments. The purple and light blue lines represent the mean recombination rate in bins for CNN score and PWM score to ChIP-seq-negative fragments, respectively. The error bars were drawn in the figure but invisible because of the large sample size. The average recombination rate of all ChIP-seq-negative and -positive fragments are shown in blue and red dashed lines, respectively. A) The top 10% of data were divided into 10 bins (1% of all data for each bin). B) All data were divided into 10 bins (10% of all data for each bin).

ChIP-seq-negative fragments (**Fig. 4**). Although the correlation became somewhat weaker, we observed a statistically significant correlation between prediction scores and recombination rates. The correlation coefficient between the prediction score and the recombination rate (ρ) was 0.129 (p-value $< 1.0 \times 10^{-47}$) in PWM and 0.169 (p-value $< 1.0 \times 10^{-47}$) in CNN. The results indicate that PWM and CNN properly capture sequence features of recombination hotspots.

3.4 CNN Captures Sequence Features Surrounding PRDM9-Binding Site

The results shown in Sections 3.1 and 3.2 indicated that CNN

© 2022 Information Processing Society of Japan

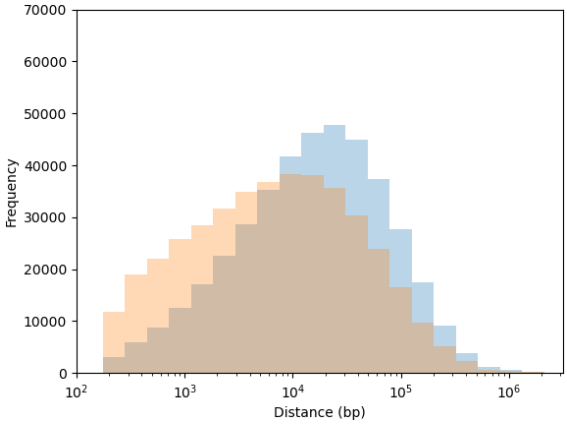

Fig. 5 Distance to the nearest ChIP-seq peak from ChIP-seq-negative/CNN-positive fragments. The orange histogram represents the frequency of distance between the center of the CNN-positive fragment and the ChIP-seq peak. The blue histogram represents the frequency of distance between the center of the CNN-negative fragment and the ChIP-seq peak.

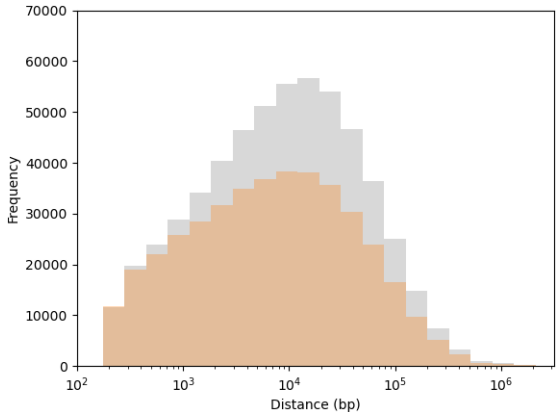

Fig. 6 Distance to the nearest ChIP-seq peak from ChIP-seq-negative/PWM-positive fragments. The orange histogram represents the frequency of distance between the center of the CNN-positive fragment and the ChIP-seq peak (same as Fig. 5). The gray histogram represents the frequency of distance between the center of the PWM-positive fragment and the ChIP-seq peak.

outperformed PWM in predicting PRDM9 binding, and the result presented in Section 3.3 suggested the possibility of detecting PRDM9 binding sites undetected by ChIP-seq experiments. However, the high recombination rate of fragments with high prediction scores did not necessarily mean that the fragments contain actual PRDM9 binding sites. There are two possible reasons. One possibility is that CNN finds the potential binding sites missed in the ChIP-seq experiments due to differences in the conditions of PRDM9 binding between actual germ cells and cultured cells, or it detects binding sites of other alleles rather than B-allele used for the experiments. Another possibility is that CNN detects the recombination hotspots using surrounding features of the PRDM9 binding site rather than the binding motif.

To test the latter possibility, we examined the distances from ChIP-seq-negative/CNN-positive fragments to the nearest ChIP-seq peaks for autosomal regions. To define CNN-positive fragments, we used the optimal threshold used for the evaluation of accuracy. We obtained 4,118,963 CNN-positive/ChIP-seq-negative fragments, and 22,062,475 CNN-negative/ChIP-seq-negative fragments. We randomly extracted equal numbers of CNN-positive and -negative fragments without overlap (411,886 fragments each) from the ChIP-seq-negative fragments and examined the distance to the nearest ChIP-seq peak (**Fig. 5**). The results showed that the CNN-positive DNA fragments were significantly closer to ChIP-seq peaks than the CNN-negative fragments (p-value $< 1.0 \times 10^{-47}$, Mann–Whitney U test).

For comparison, we conducted the same procedure using PWM-positive fragments. We obtained 5,579,323 PWM-positive/ChIP-seq-negative fragments and 20,602,115 PWM-negative/ChIP-seq-negative fragments. We randomly extracted 557,922 PWM-positive fragments without overlap. The CNN-positive fragments were still significantly closer to ChIP-seq peaks than the PWM-positive fragments (p-value $< 1.0 \times 10^{-47}$, Mann–Whitney U test) (**Fig. 6**).

4. Discussion

4.1 Advantages of Validation Using Recombination Map

In this study, we compared the performance of CNN and PWM in terms of the correlation between prediction scores and recombination rates as well as the evaluation using test data. In these two methods, CNN consistently outperformed PWM, and the method using recombination map was confirmed its usefulness for validation. The method allows us to evaluate models without the influence of overfitting to the ChIP-seq data and verify the prediction accuracy even for ChIP-seq-negative data. This validation method is not limited to PWM or CNN but can be applied to a variety of prediction models.

Table 7 Nucleotide composition in test data.

True state of fragment	AT count	GC count	GC content
ChIP-seq-positive	756,326	944,574	55.53%
ChIP-seq-negative	49,076,589	34,267,511	41.12%

4.2 ChIP-seq-positive Fragments Are GC-rich

The results shown in Fig. 5 indicated that CNN potentially captures the surrounding sequence features as well as directly recognizing binding sites of PRDM9. We hypothesized that this could be in part due to the cytosine-rich nature of the PWM motif and the change in GC content caused by the biased gene conversion that occurs after recombination repair [41]. We found that the GC contents of ChIP-seq-positive fragments in test data were significantly skewed toward high GC% (**Table 7**) (p-value $< 1.0 \times 10^{-47}$, Chi-squared test). This high GC content can be one explanation for the features around the motifs that CNN recognized.

4.3 How Much Canonical Binding Motif Has Importance when CNN Predicts?

We further conducted an analysis using the 13-mer motif to examine how much importance the CNN places on the binding site motif of the PRDM9 when scoring a fragment. First, we classified sequences that were erroneously determined as positive by CNN and PWM (i.e., false-positive fragments), after excluding the overlap. We counted the number of occurrences of the 13-mer motif in each group (**Table 8**). The results showed there was a significantly higher proportion of 13-mer motifs in the PWM-false-positive fragments than that in the CNN-false-

© 2022 Information Processing Society of Japan

Table 8 The number of the 13-mer motif occurrence for each class.

Class	The number of the motif occurrence	Total number of fragments	Percentage
CNN-false-positive	232	42,224	0.5495%
PWM-false-positive	1,227	82,618	1.485%
CNN-false-negative	9	745	1.208%
PWM-false-negative	12	1,795	0.6685%

positive fragments (p-value = 4.71×10^{-47}, Chi-squared test). We performed a similar analysis for false-negative fragments; however, there was no significant difference between the CNN-false-negative and PWM-false-negative fragments (p-value = 0.176, Chi-squared test). These results indicate that PWM is prone to output false positives in response to the presence of the motifs more than CNN.

4.4 The Features that CNN Detects and the Reason why CNN Outperformed PWM

To understand the features of the sequences captured by CNN more directly, we performed motif extraction, following the method by Brown and Lunter [14]. We performed a convolution operation to sampled 1% of all non-overlapping fragments without ambiguous nucleotide for each filter in the first layer of the trained EBCN and recorded the sequence (filter length) at the position where the output was maximum for each 100 bp DNA sequence. The DNA sequences that scored up to the top 0.1% were aligned and PWMs corresponding to each filter were created on the basis of the observed frequency of nucleotides at each position. To avoid zero probabilities, pseudocounts of 0.25 were added to correct the nucleotide frequencies at each position [42]. These PWMs should represent motifs that play an important role in prediction, but whether they under- or over-represent motifs in the ChIP-seq-positive fragments is not clear. Therefore, we measured AUC for each PWM with the test data. Because of the nature of EBCN structure, the weights of the first half of filters and the second half of filters are reverse-complementary. The analysis was therefore performed using the first 44 filters. The results are presented in **Table 9**. We found that most of the motifs recorded AUC above 0.5, which indicates they are more likely to appear in the ChIP-seq sequence (overrepresented motifs). On the other hand, eight motifs showed AUC below 0.5, which indicates they are less likely to occur in the ChIP-seq-positive fragments (underrepresented motifs). Some of the overrepresented motifs were similar to the 13-mer motif (examples are shown in **Fig. 7**). The underrepresented motifs were AT-rich (examples are shown in **Fig. 8**), which is consistent with the results shown in Table 7.

While PWM uses a scoring method that adds up the scores of sequences that appear frequently in ChIP-seq-positive fragments, CNN would subtract scores reacting to motifs that are less frequently appear in ChIP-seq-positive fragments, which can be one of the reasons why CNN outperformed PWM.

4.5 Drawbacks of Current Methods for PRDM9-binding Site Prediction

So far, we have compared CNN and PWM, and revealed CNN can be an alternative prediction method for PWM. However, the current prediction ability is restricted for several reasons. In this study, we used ChIP-seq data using PRDM9 B-allele. In the B-allele, there is no change in the zinc fingers recognizing the major 13-mer motif compared to the A allele, but there is a change from serine to threonine in the outside of the motif [2], [4], and the substitution was suggested a slightly altered binding property [2], [8], [10]. Therefore, ideally, a large dataset using other alleles should be included in the training. Sequences that do not bind specifically to the B-allele (or do not bind easily) can be captured using a larger dataset and a CNN with higher prediction performance can be obtained. We also should note that the prediction is based on only DNA sequence information in this study and the method does not take into account chromatin structure or 3D structure information. Such multimodal learnings will allow us to predict the binding site with more accuracy and know further features of PRDM9 binding sites.

Table 9 AUC for extracted PWMs from EBCN.

Filter ID	AUC
1	0.8081
2	0.7741
3	0.7479
4	0.8293
5	0.7601
6	0.4061
7	0.4036
8	0.7263
9	0.8025
10	0.7377
11	0.7834
12	0.5334
13	0.7404
14	0.7524
15	0.7786
16	0.8279
17	0.4074
18	0.6932
19	0.6184
20	0.2483
21	0.7234
22	0.7935
23	0.5156
24	0.8218
25	0.7475
26	0.3441
27	0.7129
28	0.7617
29	0.2307
30	0.6553
31	0.4183
32	0.7969
33	0.7182
34	0.7461
35	0.7163
36	0.5762
37	0.6433
38	0.6587
39	0.6709
40	0.5956
41	0.8106
42	0.8061
43	0.2200
44	0.8057

© 2022 Information Processing Society of Japan

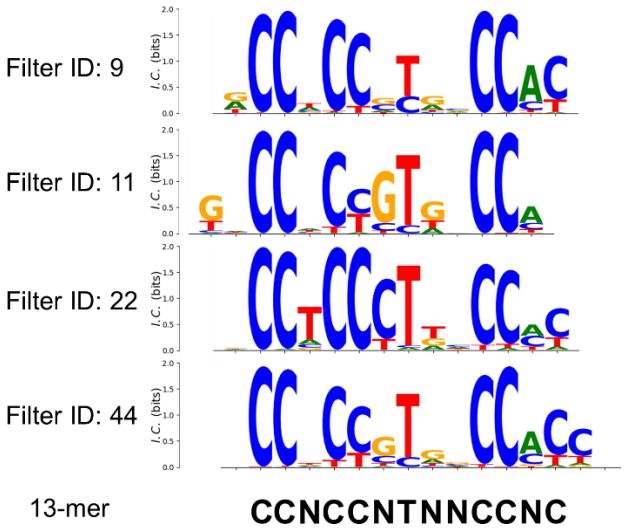

Fig. 7 Overrepresented motifs aligned with the canonical motif. The heights of strings represent Information Content. For ease of viewing, Filter ID 22 was depicted in its reverse-complement.

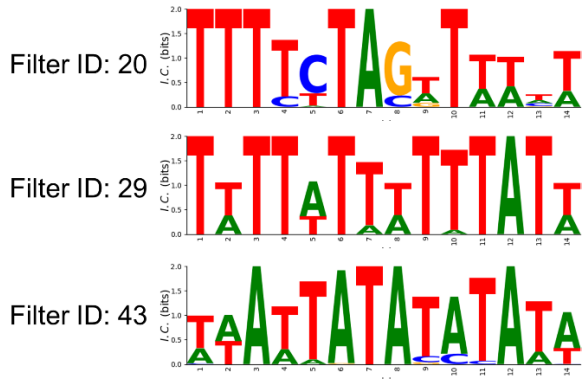

Fig. 8 Underrepresented AT-rich motifs. The heights of strings represent Information Content.

5. Concluding Remarks

We evaluated PRDM9 binding site prediction methods, PWM and CNN created from the same ChIP-seq data. CNN outperformed PWM in the validation using test data. In addition, to eliminate the effect of overfitting to a particular dataset, we developed a strategy to evaluate the accuracy of prediction by examining the correlation between prediction score and local recombination rate. Consequently, we found that CNN outperformed in both methods. This validation method is applicable for not only PWM and CNN but also a variety of models. Further, we examined the genomic distance between the ChIP-seq peak and CNN-positive fragments, nucleotide composition, and occurrence of the 13-mer motif in fragments and motifs extracted from EBCN. The analysis suggested that CNN recognized not only the binding motif of the zinc finger but also the features of surrounding sequences such as skewed GC content. In addition, a different scoring strategy of CNN from PWM, where CNN detects underrepresented features in ChIP-seq fragments, was confirmed. However, the ability of CNN is currently restricted. With large flexibility, CNN can learn other information such as other alleles' binding sites, chromatin state, and 3D structure of DNA sequence for better prediction, which will allow CNN to make better predictions.

References

[1] Myers, S., Bowden, R., Tumian, A., et al.: Drive Against Hotspot Motifs in Primates Implicates the *PRDM9* Gene in Meiotic Recombination, *Science*, Vol.327, No.5967, pp.876–879 (2010).

[2] Baudat, F., Buard, J., Grey, C., et al.: PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice, *Science*, Vol.327, No.5967, pp.836–840 (2010).

[3] Parvanov, D.E., Petkov, M.P. and Paigen, K.: *Prdm9* Controls Activation of Mammalian Recombination Hotspots, *Science*, Vol.327, No.5967, pp.835–835 (2010).

[4] Paigen, K. and Petkov, M.P.: PRDM9 and Its Role in Genetic Recombination, *Trends Genet.*, Vol.34, No.4, pp.291–300 (2018).

[5] Hayashi, K., Yoshida, K. and Matsui, Y.: A histone H3 methyltransferase controls epigenetic events required for meiotic prophase, *Nature*, Vol.438, No.7066, pp.374–378 (2005).

[6] Grey, C., Clément, A.J.J., Buard, J., et al.: In vivo binding of PRDM9 reveals interactions with noncanonical genomic sites, *Genome Res.*, Vol.27, No.4, pp.580–590 (2017).

[7] Powers, R.N., Parvanov, D.E., Baker, L.C., et al.: The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots *In Vivo*, *PLoS Genet.*, Vol.12, No.6, p.e1006146 (2016).

[8] Altemose, N., Noor, N., Bitoun, E., et al.: A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis, *eLife*, Vol.6, p.e28383 (2017).

[9] Neale, J.M. and Keeney, S.: Clarifying the mechanics of DNA strand exchange in meiotic recombination, *Nature*, Vol.442, No.7099, pp.153–158 (2006).

[10] Alleva, B., Brick, K., Pratto, F., et al.: Cataloging Human *PRDM9* Allelic Variation Using Long-Read Sequencing Reveals *PRDM9* Population Specificity and Two Distinct Groupings of Related Alleles, *Front. Cell Dev. Biol.*, Vol.9, p.3058 (online), DOI: 10.3389/fcell.2021.675286 (2021).

[11] Zeng, Y., Gong, M., Lin, M., et al.: A Review About Transcription Factor Binding Sites Prediction Based on Deep Learning, *IEEE Access*, Vol.8, pp.219256–219274 (2020).

[12] Myers, S., Freeman, C., Auton, A., et al.: A common sequence motif associated with recombination hot spots and genome instability in humans, *Nat. Genet.*, Vol.40, No.9, pp.1124–1129 (2008).

[13] Yang, P., Wu, M., Guo, J., et al.: LDsplit: Screening for *cis*-regulatory motifs stimulating meiotic recombination hotspots by analysis of DNA sequence polymorphisms, *BMC Bioinform.*, Vol.15, No.1, p.48 (2014).

[14] Brown, C.R. and Lunter, G.: An equivariant Bayesian convolutional network predicts recombination hotspots and accurately resolves binding motifs, *Bioinformatics*, Vol.35, No.13, pp.2177–2184 (2019).

[15] Li, Y., Chen, S., Rapakoulia, T., et al.: Deep learning identifies and quantifies recombination hotspot determinants (online), DOI: 10.1101/2021.07.29.454133 (2021).

[16] Wu, M., Kwoh, C.-K., Przytycka, M.T., et al.: Prediction of Trans-regulators of Recombination Hotspots in Mouse Genome, *Proc. 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM '11)*, pp.57–62, IEEE (2011).

[17] Brunschwig, H., Levi, L., Ben-David, E., et al.: Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome, *Genetics*, Vol.191, No.3, pp.757–764 (2012).

[18] Stevison, S.L., Woerner, E.A., Kidd, M.J., et al.: The Time Scale of Recombination Rate Evolution in Great Apes, *Mol. Biol. Evol.*, Vol.33, No.4, pp.928–945 (2016).

[19] Hillmer, M., Wagner, D., Summerer, A., et al.: Fine mapping of meiotic NAHR-associated crossovers causing large *NF1* deletions, *Hum. Mol. Genet.*, Vol.25, No.3, pp.484–496 (2016).

[20] Summerer, A., Mautner, V.-F., Upadhyaya, M., et al.: Extreme clustering of type-1 *NF1* deletion breakpoints co-locating with G-quadruplex forming sequences, *Hum. Genet.*, Vol.137, No.6, pp.511–520 (2018).

[21] Spence, P.J. and Song, S.Y.: Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations, *Sci. Adv.*, Vol.5, No.10, p.eaaw9206 (online), DOI: 10.1126/sciadv.aaw9206 (2019).

[22] Xia, X.: Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction, *Scientifica*, Vol.2012 (online), DOI: 10.6064/2012/917540 (2012).

[23] Persikov, V.A. and Singh, M.: De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins, *Nucleic Acids Res.*, Vol.42, No.1, pp.97–108 (2014).

[24] Persikov, V.A., Osada, R. and Singh, M.: Predicting DNA recognition by Cys2His2 zinc finger proteins, *Bioinformatics*, Vol.25, No.1, pp.22–29 (2009).

[25] Alipanahi, B., Delong, A., Weirauch, T.M., et al.: Predicting the sequence specificities of DNA- and RNA-binding proteins by deep

© 2022 Information Processing Society of Japan

learning, *Nat. Biotechnol.*, Vol.33, No.8, pp.831–838 (2015).
[26] Eraslan, G., Avsec, Ž., Gagneur, J., et al.: Deep learning: New computational modelling techniques for genomics, *Nat. Rev. Genet.*, Vol.20, No.7, pp.389–403 (2019).
[27] Davies, B., Hatton, E., Altemose, N., et al.: Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice, *Nature*, Vol.530, No.7589, pp.171–176 (2016).
[28] Megraw, M., Baev, V., Rusinov, V., et al.: MicroRNA promoter element discovery in *Arabidopsis*, *RNA*, Vol.12, No.9, pp.1612–1619 (2006).
[29] Abadi, M., Agarwal, A., Barham, P., et al.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv preprint arXiv:1603.04467 (2016).
[30] Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, *Proc. 33rd International Conference on Machine Learning* (*ICML'16*), Balcan, F.M. and Weinberger, Q.K., pp.1050–1059, JMLR.org (2016).
[31] Shrikumar, A., Greenside, P. and Kundaje, A.: Reverse-complement parameter sharing improves deep learning models for genomics (online), DOI: 10.1101/103663 (2017).
[32] Ruder, S.: An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747 (2016).
[33] Qian, N.: On the momentum term in gradient descent learning algorithms, *Neural Netw.*, Vol.12, No.1, pp.145–151 (1999).
[34] Kingma, P.D. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980 (2014).
[35] Clevert, D.-A., Unterthiner, T. and Hochreiter, S.: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), *4th International Conference on Learning Representations, ICLR 2016* (2016).
[36] Glorot, X., Bordes, A. and Bengio, Y.: Deep Sparse Rectifier Neural Networks, *Proc. 14th International Conference on Artificial Intelligence and Statistics* (*AISTATS 2011*), pp.315–323, JMLR Workshop and Conference Proceedings (2011).
[37] Klambauer, G., Unterthiner, T., Mayr, A., et al.: Self-Normalizing Neural Networks, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp.971–980 (2017).
[38] Maas, L.A., Hannun, Y.A. and Ng, Y.A.: Rectifier Nonlinearities Improve Neural Network Acoustic Models, *Proc. ICML* (2013).
[39] Akiba, T., Sano, S., Yanase, T., et al.: Optuna: A Next-Generation Hyperparameter Optimization Framework, *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (*KDD '19*), pp.2623–2631, Association for Computing Machinery (2019).
[40] Akobeng, K.A.: Understanding diagnostic tests 3: Receiver operating characteristic curves, *Acta Paediatr.*, Vol.96, No.5, pp.644–647 (2007).
[41] Duret, L., Eyre-Walker, A. and Galtier, N.: A new perspective on isochore evolution, *Gene*, Vol.385, pp.71–74 (2006).
[42] Nishida, K., Frith, C.M. and Nakai, K.: Pseudocounts for transcription factor binding sites, *Nucleic Acids Res.*, Vol.37, No.3, pp.939–944 (2009).

Appendix

The source codes and supplementary materials which include all PWMs extracted from EBCN are available on the GitHub repository (https://github.com/ibio-nakamura/PRDM9_public).

Takahiro Nakamura is in the second year of the master's course in the Graduate School of Information Science and Technology at Hokkaido University. His research interest is in bioinformatics and machine learning.

Toshinori Endo received his Ph.D. in Genetics from The Graduate University for Advanced Studies, SOKENDAI in 1996. He experienced the post-docs at National Institute of Genetics from 1996, Biolabs, Harvard University from 1997, and Genome Sciences Center, RIKEN from 1999. He was a lecturer at Medical Research Institute, Tokyo Medical and Dental University from 2000 to 2004. He became a full professor of bioinformatics at Graduate School of Information Science Technology, Hokkaido University in 2004. His major interest is molecular evolution, prediction and design of gene functions.

Naoki Osada is an associate professor at Hokkaido University, Japan. He received his Ph.D. in 2002 at the University of Tokyo for his work on genomics and the molecular evolution of primates. He worked as a post-doctoral researcher at the University of Tokyo from 2002 to 2003, and the University of Chicago from 2003 to 2004. During that time he studied bioinformatics and population genetics for speciation. He worked at the National Institute of Biomedical Innovation from 2005 to 2010 as a Scientific Researcher. From 2010 to 2015, he worked at the National Institute of Genetics as an assistant professor. He is currently working on a variety of research topics in molecular evolution and population genetics, including the study of human/primate evolution and gene expression evolution.

(Communicated by *Yoshihiro Taguchi*)

© 2022 Information Processing Society of Japan