Supplement

This is a supplement of the article "SDMEA：DNA Motif Enrichment Analysis Based on Specificity Detection", containing three parts: 1) analysis of limitations faced by traditional frameworks of DNA motif enrichment analysis (MEA); 2) the training and evaluation metrics of specificity detection model Deepbind; 3) the information of the 17 motifs used in the experiments.

1. Limitations of Traditional Motif Enrichment Analysis framework

The traditional motif enrichment analysis framework, as shown in Figure 1, includes data preparation, motif instance search, and multiple statistical tests, but there are significant defects in each link.

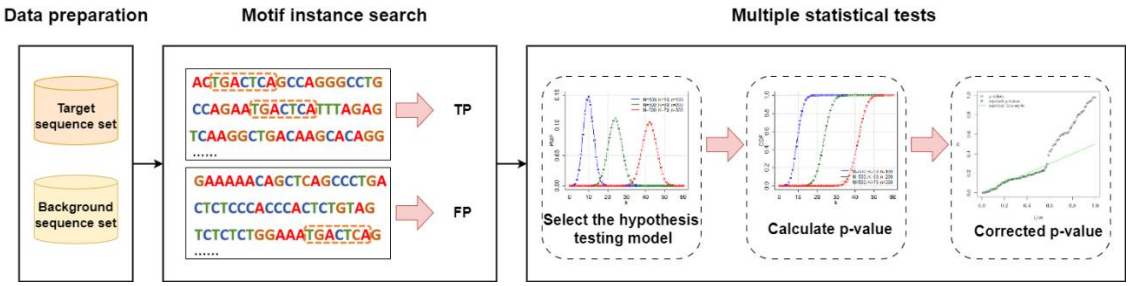

Fig.1 Traditional Motif Enrichment Analysis Framework

First, at the level of motif instance search, fixed threshold division is not conducive to the accurate identification of motifs. The experimental results of the RUNX3 motif (as shown in Figure 2) show that the accuracy of motif search fluctuates significantly with the change of p-value threshold. Increasing the threshold can improve the recognition rate of positive samples, but it will introduce a large number of false positives; reducing the threshold may miss key motifs. More complicatedly, the optimal division thresholds of different motifs are inherently different (as shown in Figure 3), and they will dynamically change with the characteristics of background sequences. A unified threshold will inevitably lead to a decrease in the recognition accuracy of some motifs. In addition, the site-level search mode only focuses on base frequency and ignores contextual information, which may generate false positive sites that meet the Position Weight Matrix (PWM) characteristics but cannot specifically bind to transcription factors.

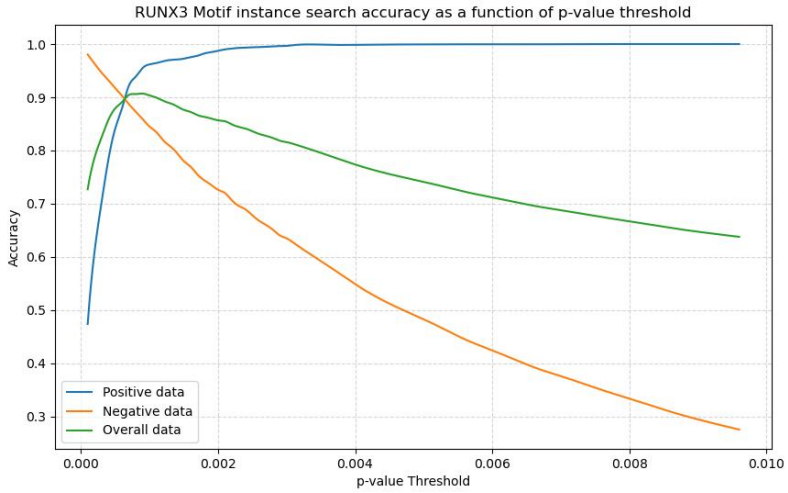

Fig. 2 Changes in the Accuracy of RUNX3 Motif Instance Search with p-value Threshold

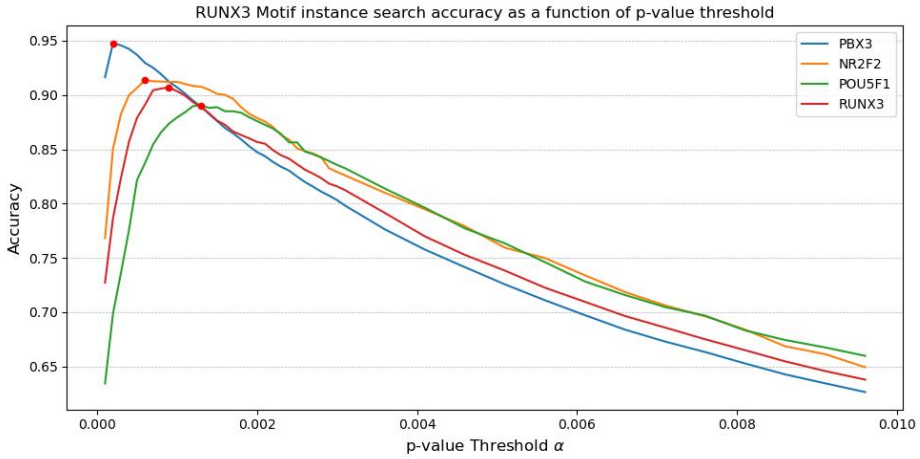

Fig. 3 Differences in Optimal Thresholds Among Different Motifs

Second, the statistical test link faces the risk of double error accumulation. On the one hand, multiple tests lead to error inflation. When analyzing k motifs, the overall false positive rate is $1-(1-\alpha)^k$, where α is the significance threshold. For example, when there are 10 motifs and α=0.05, the overall error rate can reach 40%. While multiple test correction controls the overall false discovery rate with a more stringent threshold, it discards a large amount of potential motif information, thereby affecting the results of motif enrichment analysis. On the other hand, fixed threshold division causes systematic deviation. For example, the 8% false negative rate and 16% false positive rate in the recognition of a certain motif will be amplified through the cascading effect of multiple rounds of searches. The false positive motifs in the first round may be mistakenly used as "real signals" in the co-occurrence test of the next round.

Finally, the description of enrichment degree is biased because the assumptions are

disconnected from reality. Traditional methods judge the enrichment degree by comparing the p-values of different motifs horizontally, but its implicit assumptions such as consistent sample size, independent binding events, and homogeneous background distribution contradict the heterogeneous distribution of motifs in biological scenarios, especially in GC-rich promoter regions, where the deviation is more obvious.

2. Specificity Detection Model's Training

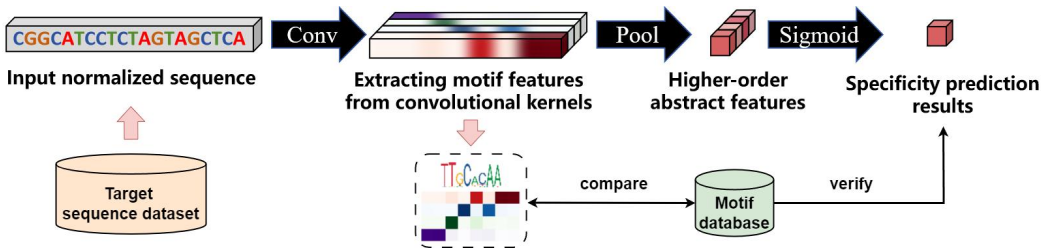

Fig. 4 The Framework of Deepbind Model for Specificity Detection.

The structure of the Deepbind model is shown in Figure 4. This model captures sequence context features through deep learning and dynamically generates negative samples to avoid distribution deviation, effectively improving the specificity and accuracy of motif recognition.

Data processing adopts a strict division strategy: odd peak sequences are used as the training set (positive samples, label 1), and negative samples are shuffled sequences (label 0) dynamically generated during runtime, which maintain dinucleotide frequency. This not only ensures data diversity but also avoids overfitting.

Hyperparameter tuning is realized through random search, and the optimization objects include key parameters such as learning rate, momentum coefficient, and dropout ratio. After generating parameter combinations through multi-dimensional sampling within the preset search range, a five-fold cross-validation scheme is adopted. Each iteration performs random weight initialization and completes parameter updates within the preset epoch limit. Using the test set AUC value as the indicator, the optimal combination is finally selected to ensure the stable performance of the model.

Model training is based on 690 sets of ChIP-seq data from the ENCODE database accessible at the website https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/, and 23 human motifs matching the JASPAR motif library are screened out. Each motif is trained with an independent model. The experiment uses four evaluation indicators, Sensitivity, Specificity, Accuracy, and AUC (Area Under Curve), where Sensitivity reflects the model's ability to capture specific sequences, Specificity measures the model's ability to identify non-specific sequences, Accuracy comprehensively evaluates the overall classification performance of the model on positive and negative samples and is suitable for scenarios with balanced samples, and AUC evaluates the robustness of the model under different thresholds through the area under the ROC curve, which can ensure stability under different experimental conditions.

We selected 17 specificity prediction models with AUC indicators above 0.9 under different threshold classifications. The results of five-fold cross-validation are shown in Table 1. The average AUC of the motifs to be analyzed reaches 0.943; the average accuracy, specificity, and sensitivity are 0.883, 0.866, and 0.906 respectively. Among them, the AUC of the BATF motif reaches 0.993, and that of the NR2F2 motif reaches 0.976, verifying the high reliability of the model.

Table 1. Indicators of Different Specificity Detection Models on the Test Set

Transcription Factor	AUC	Accuracy	Specificity	Sensitivity
BATF	0.993	0.957	0.94	0.974
CEBPD	0.916	0.842	0.84	0.844
NR2F2	0.976	0.912	0.874	0.95
POU5F1	0.960	0.884	0.828	0.94

3. Motif Information of Experimental Data

Motif ID	Motif Name	Link
MA0462.3	BATF	https://jaspar.elixir.no/download/data/2024/bed/MA0462.3.bed
MA0003.5	TFAP2A	https://jaspar.elixir.no/download/data/2024/bed/MA0003.5.bed
MA0018.5	CREB1	https://jaspar.elixir.no/download/data/2024/bed/MA0018.5.bed
MA0103.4	ZEB1	https://jaspar.elixir.no/download/data/2024/bed/MA0103.4.bed
MA0484.3	HNF4G	https://jaspar.elixir.no/download/data/2024/bed/MA0484.3.bed
MA0490.3	JUNB	https://jaspar.elixir.no/download/data/2024/bed/MA0490.3.bed
MA0508.4	PRDM1	https://jaspar.elixir.no/download/data/2024/bed/MA0508.4.bed
MA0522.4	TCF3	https://jaspar.elixir.no/download/data/2024/bed/MA0522.4.bed
MA0597.3	THAP1	https://jaspar.elixir.no/download/data/2024/bed/MA0597.3.bed
MA0684.3	RUNX3	https://jaspar.elixir.no/download/data/2024/bed/MA0684.3.bed
MA0836.3	CEBPD	https://jaspar.elixir.no/download/data/2024/bed/MA0836.3.bed
MA1102.3	CTCFL	https://jaspar.elixir.no/download/data/2024/bed/MA1102.3.bed
MA1111.2	NR2F2	https://jaspar.elixir.no/download/data/2024/bed/MA1111.2.bed
MA1114.2	PBX3	https://jaspar.elixir.no/download/data/2024/bed/MA1114.2.bed
MA1115.2	POU5F1	https://jaspar.elixir.no/download/data/2024/bed/MA1115.2.bed
MA1508.2	IKZF1	https://jaspar.elixir.no/download/data/2024/bed/MA1508.2.bed
MA1585.2	ZKSCAN1	https://jaspar.elixir.no/download/data/2024/bed/MA1585.2.bed