Computational and Structural Biotechnology Journal 20 (2022) 296–308

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

regCNN: identifying *Drosophila* genome-wide *cis*-regulatory modules via integrating the local patterns in epigenetic marks and transcription factor binding motifs

Tzu-Hsien Yang *, Ya-Chiao Yang [1], Kai-Chi Tu [1]

Department of Information Management, National University of Kaohsiung, Kaohsiung University Rd, 811 Kaohsiung, Taiwan

ARTICLE INFO

Article history:
Received 15 September 2021
Received in revised form 10 December 2021
Accepted 10 December 2021
Available online 18 December 2021

Keywords:
cis-regulatory modules
Transcriptional regulation
Epigenetic regulation
Transcriptional factor binding sites

ABSTRACT

Transcription regulation in metazoa is controlled by the binding events of transcription factors (TFs) or regulatory proteins on specific modular DNA regulatory sequences called *cis*-regulatory modules (CRMs). Understanding the distributions of CRMs on a genomic scale is essential for constructing the metazoan transcriptional regulatory networks that help diagnose genetic disorders. While traditional reporter-assay CRM identification approaches can provide an in-depth understanding of functions of some CRM, these methods are usually cost-inefficient and low-throughput. It is generally believed that by integrating diverse genomic data, reliable CRM predictions can be made. Hence, researchers often first resort to computational algorithms for genome-wide CRM screening before specific experiments. However, current existing *in silico* methods for searching potential CRMs were restricted by low sensitivity, poor prediction accuracy, or high computation time from TFBS composition combinatorial complexity. To overcome these obstacles, we designed a novel CRM identification pipeline called regCNN by considering the base-by-base local patterns in TF binding motifs and epigenetic profiles. On the test set, regCNN shows an accuracy/auROC of 84.5%/92.5% in CRM identification. And by further considering local patterns in epigenetic profiles and TF binding motifs, it can accomplish 4.7% (92.5%–87.8%) improvement in the auROC value over the average value-based pure multi-layer perceptron model. We also demonstrated that regCNN outperforms all currently available tools by at least 11.3% in auROC values. Finally, regCNN is verified to be robust against its resizing window hyperparameter in dealing with the variable lengths of CRMs. The model of regCNN can be downloaded athttp://cobisHSS0.im.nuk.edu.tw/regCNN/.

1. Introduction

Transcription regulation in metazoan genomes is controlled by the binding events of transcription factors or regulatory proteins on specific DNA segments in a modular manner [1,2]. These modular DNA regulatory sequences are called *cis*-regulatory modules (CRMs). CRMs regulate the correct spatial–temporal differential gene expression in developmental stages and determine the specific cell types of the developing cells in cell differentiation [3]. Malfunction of certain CRMs can lead to genetic diseases or cancers [4,5]. Hence discovering and understanding genome-wide distributions of CRMs in metazoa species are essential in constructing the transcriptional regulatory network that helps identify medical diagnoses to genetic disorders [6,7].

While traditional reporter-assay approaches can provide an in-depth understanding of CRMs, these methods require careful and innovative experiment designs, leading to the low-throughput and cost-inefficient nature [8]. It is generally believed that by integrating diverse genomic data types, reliable CRM predictions can be generated [1,9,10]. And great efforts have been made by various research groups for obtaining genomic datasets in different aspects. Hence, to perform genome-scale CRM scanning, researchers often first resort to computational algorithms before specific experiments. Based on the understanding of CRM functions from comparative and molecular biology, different *in silico* CRM identification methods were developed. Depending on the modeling techniques, these algorithms can be divided into three different

* Corresponding author.
E-mail addresses: thyangza1025@nuk.edu.tw (T.-H. Yang), a1073314@mail.nuk.edu.tw (Y.-C. Yang), a1073348@mail.nuk.edu.tw (K.-C. Tu).
[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2021.12.015

categories. The first type of CRM prediction algorithms is constructed from the concepts of functional sequence conservation [11–13]. Algorithms categorized in this type identify conserved non-coding sequences as CRM candidates from the alignment between related species. The second type of CRM scanning methods is designed based on mining transcription factor binding site (TFBS) combinations that may lead to CRM functions [13–18]. Such tools use probability modeling to consider all possible combinations of TFBS positioning in a given DNA segment for putative CRMs. The third type of CRM prediction algorithms jointly considers different genome-wide chromatin-binding protein and histone sequencing data to annotate potential functional regulatory sequences [19–22]. CRMs are now known to recruit regulatory DNA/chromatin binding proteins through their epigenetic profiles. And various efforts have been made to probe the comprehensive genome-wide epigenetic profiles using the chromatin immunoprecipitation (ChIP) sequencing techniques [23,24]. The epigenetic profiles of CRM sequences, such as H3K4me and nucleosome depletion signals, can be recognized by different functional transcription factors and regulatory proteins, thus determining CRM regulatory functions [25]. Based on identifying different combinations of average signal values of epigenetic marks over the given chromosomal range, many chromatin type landscapes were created [26,27].

However, all previous methods suffer from certain drawbacks. First, methods developed from sequence conservation assume that functional genomic elements are usually conserved and undergo purifying selection in evolution. However, researchers have discovered that many non-coding sequences containing functional regulatory elements are not conserved among species [28,29]. Some CRMs even belong to specific species and therefore lack phylogeny information [30]. Hence, methods based on the conservation assumption can only discover a fractional number of CRMs, causing low prediction sensitivity. Second, strategies constructed based on TF binding event combinations or k-mer information often require trying out the TFBS motif positioning combinations that usually grow exponentially [16]. The combinatorial complexity causes these algorithms to convey prohibitive high computation time, hindering researchers from carrying out genomic CRM studies. Further, with no prior knowledge, it is hard to accurately discover CRMs that consist of more than two different TF binding targets for these TFBS motif enrichment models [31]. Third, models trained only on average values of epigenetic signals on the given segments overlook the detailed local summarizing information in TF binding motifs, chromatin-binding protein target sites, and epigenetic marks. Using limited selected simple co-occurrence of certain epigenetic marks to infer CRMs can result in many false positives in genome-wide investigation [32,33,22]. In summary, current existing CRM identification algorithms are restricted by the issues of low sensitivity, combinatorial complexity, or unsatisfactory prediction accuracy.

In this research, we devised an integrative deep learning model called regCNN to overcome the aforementioned problems. regCNN encompasses the base-by-base distributions and local summarizing patterns in sequence conservation, TF binding motifs, and epigenetic marks (nucleosome-free and nucleosome-variant regions, chromatin-binding protein target sites, and histone modifications) to improve CRM identification. These local summarizing regulatory patterns were considered by regCNN through consecutive hierarchical convolution operations. We trained, cross-validated, and tested regCNN on a literature-curated CRM ground truth dataset gathered from the REDfly repository [34]. We adopted the model organism *Drosophila melanogaster* in this research since there are comprehensive curated experimentally verified CRMs in this species. By considering the local TFBS preference and epigenetic profiling summarizing feature patterns, regCNN can obtain higher CRM identification accuracy (Accuracy = 84.5%, auROC = 92.5%) than the pure multi-layer perceptron model that uses only the average values of these same epigenetic signals and TFBS preference scores (Accuracy = 79.2%, auROC = 87.8%) on the set-aside test set. Further, we demonstrated that regCNN outperforms all currently available *Drosophila* CRM prediction tools by at least 11.3% auROC values on the test set. And integrating both the TFBS motif datasets and epigenetic profiling datasets can contribute to better CRM identification (auROC = 92.5%) than considering only either TFBS datasets (auROC = 88.9%) or epigenetic datasets (auROC = 88.5%). In addition, by applying different test set partition schemes, it was verified that regCNN generalizes well to genomic sequences. In the last, regCNN is confirmed to be robust against its resizing window size hyperparameter adopted in the information retrieval stage. The regCNN deep network model is freely available at http://cobisHSS0.im.nuk.edu.tw/regCNN/.

2. Methods and Datasets

2.1. Data preprocessing

regCNN integrates five types of transcriptional regulation-related datasets that consider thorough aspects of modular transcriptional regulation. We adopted the model organism *Drosophila melanogaster* in this research since there are comprehensive human-curated verified CRMs in this species. The *Drosophila melanogaster* genome (dmel_r6.03_FB2014_06) was downloaded from Flybase [35], and the literature curated CRMs were retrieved from REDfly [34]. Further, the following five different transcriptional regulation-related datasets are preprocessed and used in regCNN: conservation score data, TF binding motifs, nucleosome-free and nucleosome-variant sites, chromatin-binding protein target sites, and histone modification information. Preprocessing of the five genres of datasets is depicted in the following subsections.

2.1.1. Conservation score data

Some regulatory sequences in genomes are conserved among different related species [36]. In this research, we adopted the base-by-base phastCons conservation scores [37] that utilized the hidden Markov model (HMM) to summarize the conservation alignment between *D. melanogaster* and other 26 insects (*D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. biarmipes*, *D. suzukii*, *D. ananassae*, *D. bipectinata*, *D. eugracilis*, *D. elegans*, *D. kikkawai*, *D. takahashii*, *D. rhopaloa*, *D. ficusphila*, *D. pseudoobscura*, *D. persimilis*, *D. miranda*, *D. willistoni*, *D. virilis*, *D. mojavensis*, *D. albomicans*, *D. grimshawi*, *Musca domestica*, *Anopheles gambiae*, *Apis mellifera*, and *Tribolium castaneum*). The multiple alignment phastCons scores among these species were downloaded from UCSC Genome Browser [38].

2.1.2. TF binding motif data

Combinatorial binding events of different TFs on regulatory sequences can trigger CRM functions [2]. The binding affinity and motifs of different TFs can be represented in the form of position weighted matrices (PWMs). We downloaded the PWMs of 158 *D. melanogaster* TFs from the JASPAR database [39] (JASPAR 2020 version). Details of the 158 TF binding motifs can be found through the provided link in the "Data Availability" subsection. The motif odds ratio OR_{ij} measuring the occurrence of TF motif j at position i is calculated by the following formula:

$$OR_{ij} = \prod_{k=0}^{n-1} \frac{P_{TF_j}(s_k)}{P_{bg}(s_k)},$$

where s_k is the base pair of the kth position after i, n is the length of the motif of TF_j, $P_{TF_j}(s_k)$is the binding probability of s_k on the kth

position of the TF_j binding motif, and $P_{bg}(s_k)$is the background nucleotide distribution for the nucleotide type of s_k. The motif PWMs were calibrated by adding a small ϵ and re-normalizing to 1 to avoid zero probability. In this research, ϵ was selected to be 10^{-6}. After computing the odds ratios of the 158 TFs, we obtained the 158-way TFBS motif data results for the given genomic region.

2.1.3. Nucleosome-free and nucleosome-variant sites

Regulatory sequences wrapped within nucleosomes may be prohibited from carrying out their functions [40]. Hence identifying the depletion or variants of nucleosomal structures can help identify active CRMs. We adopted the nucleosomal depletion signals probed via DNaseI hypersensitive sites (five embryo samples under five different developmental stages (stage 5, 9, 10, 11, and 14) and one replicate of Kc167 cell line) from the works of Thomas et al. [41]. And we also gathered the nucleosome-variant H2Av data in *Drosophila* embryos from the work of Mavrich *et al.* [42]. Detailed accession numbers for these sequence probing experiments can be found in the "Data Availability" subsection. The sequencing `.fastq` files were downloaded from Sequence Read Archive (SRA). We used the bowtie [43] short read alignment tool with default parameters to map the reads back to the Drosophila melanogaster genome (dmel_r6.03_FB2014_06). And the nucleosomal score (NS) of a given condition at position i is computed by taking the logarithm of the RPM (reads per million) [44] value at position i:

$$\mathrm{NS_i} = \log \frac{10^6 \ * \sum_{\mathrm{read} \in R} \mathbb{1}[\mathrm{read} \cap g_i]}{\#R},$$

where R is the collection of all mapped reads under the given experimental condition, $\mathbb{1}[.]$ is the indicator function, $\#R$ is the cardinality of R, g_i is the chromosomal coordinate of position i, and $\mathrm{read} \cap g_i$ refers to the condition that the read covers g_i.

2.1.4. Chromatin-binding protein target sites and histone modification information

Core histone modifications are now known to recruit different chromatin-binding proteins that interact with TFs or remodel the chromatin structures [25]. The "histone codes" resulting from the combinations of histone modifications and chromatin-binding protein target sites can have critical effects on CRM functionalities [1,25]. In total, 57 ChIP-seq experiments for probing the binding sites of 31 chromatin-binding proteins and 36 ChIP-seq experiments for investigating 15 histone modifications were downloaded from the modENCODE project [23] and reprocessed. The detailed list of the ChIP-seq data can be found in the "Data Availability" subsection. Since the 57 chromatin-binding protein ChIP-seq experiments and the 36 histone modification ChIP-seq experiments were performed using similar protocols, we applied the same procedure to analyze these ChIP-seq datasets. More specifically, each of the ChIP-seq data was mapped to the *Drosophila melanogaster* genome (dmel_r6.03_FB2014_06) by bowtie [43] using default parameters. And for each of the ChIP-seq experiments, the base-by-base sequencing scores (SC) for this ChIP-seq experiment are defined by the following formula based on the concept of RPM [44]:

$$\mathrm{SC_i} = \log \left(\frac{\sum_{\mathrm{read} \in C} \mathbb{1}[\mathrm{read} \cap g_i]}{\#C} \ * \left(\frac{\sum_{\mathrm{read} \in I} \mathbb{1}[\mathrm{read} \cap g_i]}{\#I} \right)^{-1} \right),$$

where C is the collection of pooled mapped reads in ChIP replicates under the given condition, I is the collection of pooled reads in input wild-type replicates of the corresponding same experimental condition as $C, \mathbb{1}[.]$ is the indicator function, g_i is the chromosomal coordinate of position i, and $\mathrm{read} \cap g_i$ refers to the condition that the read covers g_i.

2.1.5. Dataset score normalization

Since the data magnitude scale can dominate the performance of deep learning models in a biased way, we performed data normalization as the last step of data pre-processing to reduce the scale effect. The data normalization formula is defined as follows:

$$x_{k\_i} = \frac{x_{k\_i} - \min(\mathbf{x_k})}{\max(\mathbf{x_k}) - \min(\mathbf{x_k})},$$

where $\mathbf{x_k}$ is the collection of scores derived from the kth specific experimental dataset, x_{k_i} is the ith data point in $\mathbf{x_k}$, and $\min(\mathbf{x_k})/\max(\mathbf{x_k})$ is the minimum/maximum of $\mathbf{x_k}$. By applying the transformation, every individual experimental dataset $\mathbf{x_k}$ categorized in the above five genres was normalized into $[0, 1]$, making them free of range biases.

2.2. The regCNN deep network

We designed the regCNN model in this research to identify genome-wide CRMs with high accuracy. The aforementioned five types of transcription regulation-related datasets were integrated in regCNN to boost model performance. The overall regCNN CRM identification process can be divided into three different stages (See Fig. 1): Stage I (Information retrieval), Stage II (Local pattern extraction and summarization), and Stage III (CRM identification). Details of each stage are elucidated in the following subsections. And the architecture hyperparameters are also marked in Fig. 1.

2.2.1. Stage I: information retrieval

Since CRMs are of variable-length, we designed an information retrieval and dimension transformation procedure to allow the model to accept variable-length sequences (See Fig. 1-Stage I). First, experimental data in the aforementioned five categories of datasets observed in the given region are extracted. In total, the numbers of experiments in the aspects of conservation, TFBS motifs, nucleosome-free and nucleosome-variant sites, chromatin-binding protein target sites, and histone modifications are 1, 158, 7, 57, and 36, respectively. Each of the extracted results is regarded as an l x 1 array, where l is the length of the given region. Every array is zero-padded to 512 x 1 if l is smaller than 512 bps or is down-sampled to 512 x 1 using the `openCV INTER_LINEAR` interpolation package [45] if l is longer than 512 bps. Finally, these data arrays are stacked together to form a 512 x 259 multi-channeled tensor.

2.2.2. Stage II: local summarizing pattern extraction

After the information retrieval steps of Stage I, a base-by-base 512 x 259 data tensor F is formed. Convolutional neural networks are now known to have excellent performance in extracting detailed spatial and temporal patterns [46]. Hence, we incorporated hierarchical convolution operations in regCNN to mine out local regulatory patterns that can help identify cellular CRMs (See Fig. 1-Stage II). Hierarchical convolution operations are designed by applying a consecutive series of convolutions and activation functions to the data tensor. The convolution operations can be formulated as

$$C_n = K_n \circledast D,$$

where D is the data tensor applied to the convolution, $\circledast$ denotes for the 1D convolution operator, K_n is the nth kernel filter with the size of $(2k+1)$ and the same depth of D, and C_n is the nth channel of the output convoluted vector C. The element-wise

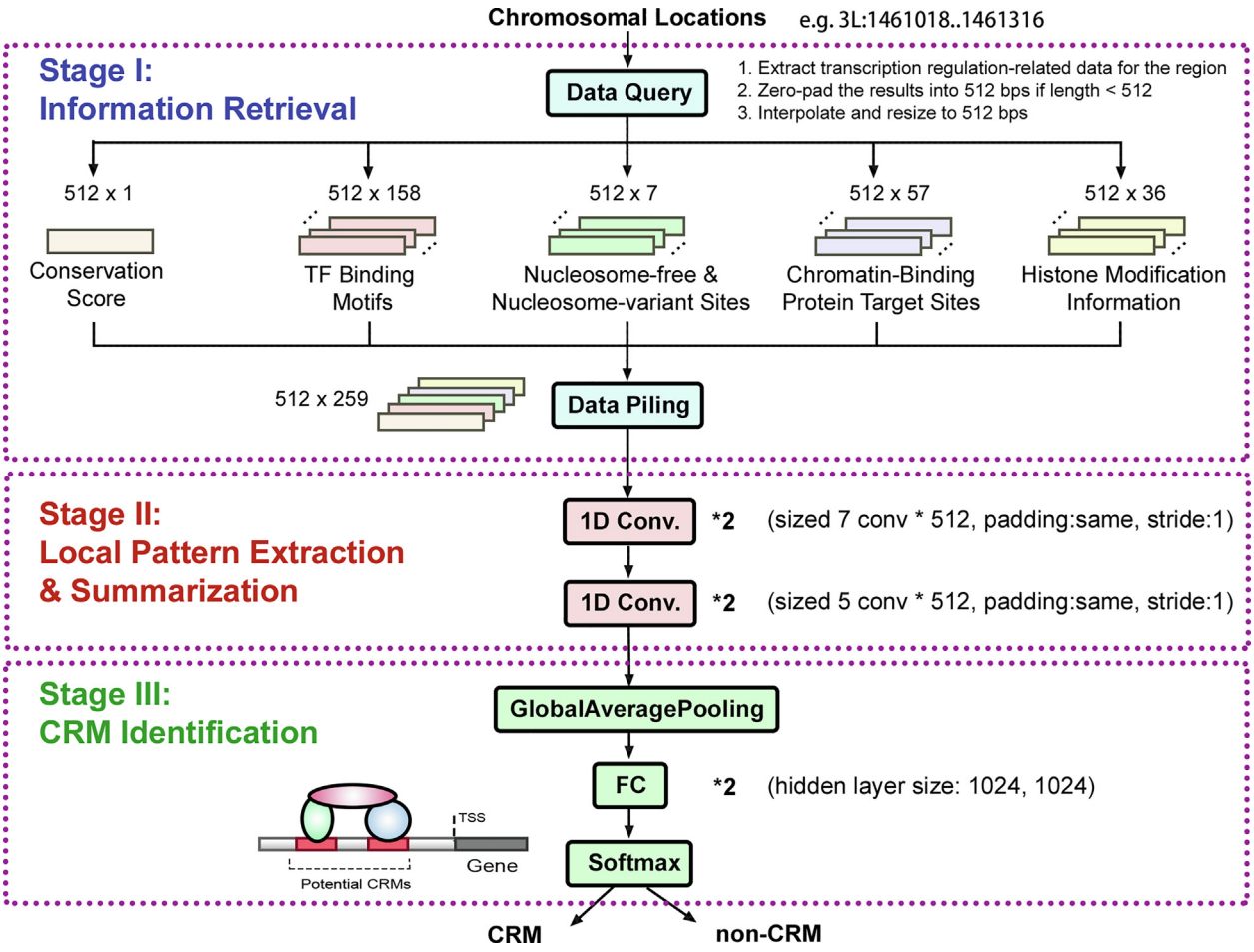

Fig. 1. The overview the of regCNN model. regCNN can be divided into three stages. In Stage I, experimental results of datasets from the five transcription regulation-related genres are retrieved. In Stage II, the local summarizing patterns in different types of data are extracted by hierarchical consecutive convolutions. In the last stage, CRMs are discriminated from non-CRM sequences based on the aggregated local summarizing patterns. The details of the regCNN deep network architecture are also marked for each of the network operations.

results of the $\circledast$ operation in the above equation are defined by the following formula:

$$C_n(i) = \sum_{p=-k}^{k}\sum_{q=1}^{259} K_n(p+k+1,q) * D(i+p,q),$$

where $C_n(i)$ is the ith element of the output vector C_n, and the calculation is zero-padded to make the output C_n of the same length as D. In our settings, the kernel moving stride is all set to 1 in the formula. In the designed regCNN, parametric rectified linear units (PReLU) are used as the activation functions. The final four (kernel size, kernel number) pairs picked from cross-validation are (7, 512), (7, 512), (5, 512), and (5, 512), respectively. The hierarchical convolution operations in regCNN help obtain the local base-by-base summarizing patterns in epigenetic profiles, TF binding information, and conservation scores.

2.2.3. Stage III: CRM identification

In Stage III of regCNN, the local summarizing patterns are used to identify potential CRMs. The global average pooling operation first mean-aggregates each pattern tensor. Then two layers of fully connected operations are applied to map the aggregated patterns into a CRM-separable high dimensional space. Finally, potential CRMs are identified by a classification softmax layer (See Fig. 1-Stage III). The operations can be written as

$$p = \mathrm{softmax}(a(a(SW_1)W_2)),$$

where S is the mean-aggregated pattern tensor, p is the probability that the given chromosomal region contains CRMs, W_1 and W_2 are the trainable network parameter weight matrices, $a(.)$ is the nonlinear activation function, and `softmax` is the operation that transforms the computed scores into probability using exponential normalization. In this research, we selected PReLU as the activation function and 1,024 as the hidden layer size. Dropout layers were added between the fully connected layers to equip regCNN with good generalization performance for unknown sequences.

2.3. regCNN training hyper-parameter selection

Several hyper-parameters of regCNN were selected in the training process using the fivefold cross-validation technique. Based on random search over a range of values for each hyperparameter, the followings were the final chosen hyper-parameters: (1) learning rate schedule: cosine warm-up to 1e-4 in four epochs, then exponential decay with decay rate 0.95 before epoch 25 and 0.9 after epoch 25; (2) Optimization method: Adam; (3) number of epochs: 60; (4) neuron initialization: Xavier initialization; (5) Batch training size: 256; (6) The non-linear activation function: PReLU. Dropout layers (dropout rate = 0.1) were added to regularize the

training process and prevent over-fitting that may kill model generalization.

3. Results

3.1. Overview of regCNN

regCNN integrates the conservation data, the TFBS motif data, the nucleosome-free and nucleosome-variant sites, the chromatin-binding protein ChIP data, and the histone modification ChIP information to discriminate CRMs from non-functional sequences in the *Drosophila melanogaster* genome (See Fig. 1). A given chromosomal range is provided as the input information for regCNN. Then the overall CRM identification process of regCNN can be divided into three stages: (1) Stage I: Information retrieval. regCNN first queries the calculated base-by-base score arrays for the specified chromosomal region from each of the above-mentioned datasets. For sequence ranges shorter than 512, the scores are zero-padded to 512 bps. If the given range is longer than 512, each array of the calculated experimental results is individually down-sampled to 512 bps using interpolation. Then the 259 experimental scores from datasets in the five categories are stacked together to form a 512 x 259 data tensor. (2) Stage II: Local summarizing pattern extraction. A sequence of hierarchical convolutions is applied to the data tensor to summarize the local combinatorial patterns of different regulatory features for the given chromosomal region. The pattern tensors summarize the local regulatory information in different regulation aspects that may potentially recruit and form special regulatory protein complexes. (3) Stage III: CRM identification. Based on the feature patterns, regCNN calculates the likelihood of the sequence to be a putative CRM. The overall network architecture and the layer hyper-parameters of regCNN are depicted in Fig. 1. More detailed descriptions of the architectures of regCNN can be found in the "The regCNN deep network" subsection.

3.2. Ground-truth CRM dataset

We gathered the experimentally verified CRM sequences from the literature as the positive set and randomly picked non-functional sequence segments as the negative set in this research. The positive set and the negative set together form the CRM ground-truth dataset. We further divided the ground truth dataset into the training/validation set and the test set. Since comprehensive literature CRM repositories are currently only available in *Drosophila*, we hence selected *Drosophila melanogaster* as the demo model organism in this research. The experimentally verified CRM positive set was downloaded and processed from the REDfly database [34] (v9.5.1, database updated at 09/01/2021, under dmel_r6.03_FB2014_06). In total, 28,119 experimentally verified CRMs from the literature were included in the positive CRM set. The basic statistic distributions (distances of CRMs to their closest genes, the numbers of TF binding sites per CRM, and the CRM lengths) of these experimentally verified CRMs are summarized in Fig. 2. The median CRM length is 501 bps, and the median of non-zero CRM distances to their closest genes is 2,715 bps. 26,402 CRMs overlap with at least one gene and have zero distance to their closest genes. Notice that Fig. 2-b plots the numbers of TF binding sites per CRM (average TFBSs per CRM = 7) for only the 1,694 CRMs with known experimentally verified TFBSs (downloaded from REDfly v9.5.1) on them.

To obtain the negative set of non-CRM segments, we followed the procedure proposed by Su et al. [9] to extract negative random sequences. We sampled the non-CRM sequences from the introns, exons, and intergenic regions based on the *Drosophila* genome

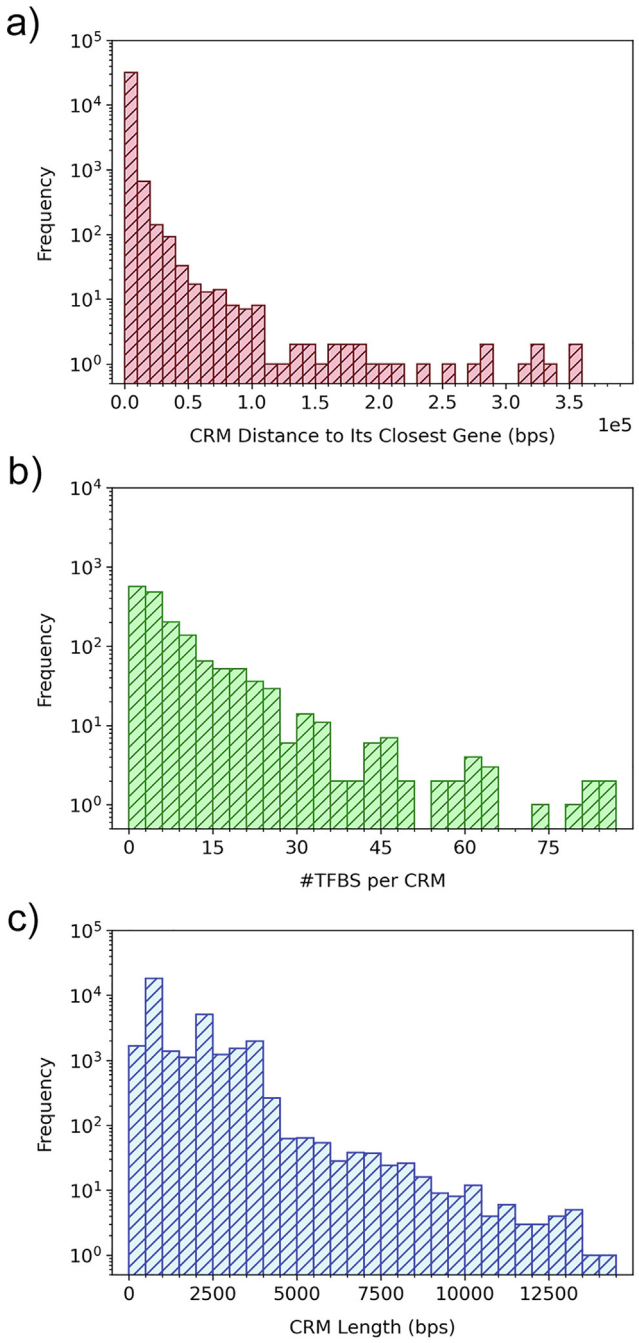

Fig. 2. The basic statistic distributions of the CRMs gathered in this research. (a) The distribution of distances of CRMs to their closest genes. (b) The distribution of the numbers of TF binding sites per CRM. (c) The distribution of CRM lengths.

deposits (FlyBase [35] dmel_r6.03_FB2014_06) using the following criteria listed by Su *et al.*: (1) the sampled sequences should not overlap with known CRMs in the positive set; (2) the overall length distributions of the positive set and the negative set are similar; (3) 6 bps from 5' and 3' ends are avoided in introns/exons that are shorter than 83 bps to eliminate splice donor/acceptor sites; (4) 150 bps from 5' and 3' ends are excluded in introns or exons that are longer than 300 bps to reduce splice regulatory sequences; (5) 1000 bps from 5' and 3' ends are not used in intergenic regions to be free of promoter sites and post-transcriptional modification sites. These five criteria minimize the possibility of mistakenly selected functional sequences that may reside in the exons,

introns, or intergenic regions into the negative set. We obtained 25,849 non-CRM sequences to form the negative set. The ground-truth CRM and non-CRM set can be downloaded using the link provided in the "Data Availability" subsection. From the ground-truth CRM dataset, we first took one-tenth of the dataset (2,847 CRMs and 2,590 non-regulatory sequences) as the set-aside test set. To avoid potential data contamination issues caused by sequence overlapping, we further enforced a restriction that sequences in the test set have no overlapping with any sequences in the training/validation set. Therefore, those sequences that overlap with some sequence in the test set had been removed from the training/validation set. The rest of the CRM and non-CRM sequences (25,272 CRMs and 23,259 non-regulatory sequences) were used as the training/validation set for regCNN model cross-validation training.

3.3. regCNN can distinguish CRMs from random sequences

regCNN was trained, validated, and tested using the ground truth CRM dataset. On the training/validation set (25,272 CRMs and 23,259 non-regulatory sequences), 5-fold cross-validation was applied to fully utilize the data for both model optimization and over-fitting control [47]. Using the 5-fold cross-validation technique, the training/validation set was divided into five different subsets. First, the model was trained on four subsets. Then the remaining one fold was used as the validation set to select model hyperparameters and architectures. This process was repeated five times by treating every one fold of the training/validation set as the validation set at a time. A final average performance was evaluated using the receiver operating characteristic (ROC) curves [44]. The ROC curves plots (1-specificity) against sensitivity when the threshold is adjusted:

$$\text{Sensitivity (recall)} = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{FP + TN},$$

where TP is the number of identified true CRM sequences, FP is the number of wrongly identified false CRM segments, TN is the number of correctly identified non-CRM sequences, and FN is the number of known CRMs to be mistakenly regarded as non-regulatory sequences. A CRM identification tool is said to have good discriminating power if it shows high sensitivity while the threshold is chosen to maintain high specificity (low 1-specificity) in the ROC curve. In this case, the model will have a high auROC (area under the ROC curve) value in the ROC curve plot. The precision-recall curve (PRC) is also generated, and the auPRC (area under the PRC) value is computed for evaluating the trade-off between precision and recall. The larger the auPRC value is, the better trade-off between precision and recall can be achieved. We also evaluated the CRM identification accuracy by using the precision and F1 measure [48]:

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}},$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN},$$

The training process was monitored by the learning curve approach (See Fig. 3-a). In learning curve plots, accuracy values are recorded according to training epochs. It can be used to verify that the model is well optimized and is not over-fitted. As shown in Fig. 3-a, the training and validation learning curves both approach a plateau, showing that the training process converges. Moreover, the two curves also bear a minor gap between each other, verifying that the designed network is not over-fitted. Therefore, the regCNN model was well trained. The ROC curve and PRC results are summarized in Fig. 3-b and 3-c. The fivefold validation ROC curves

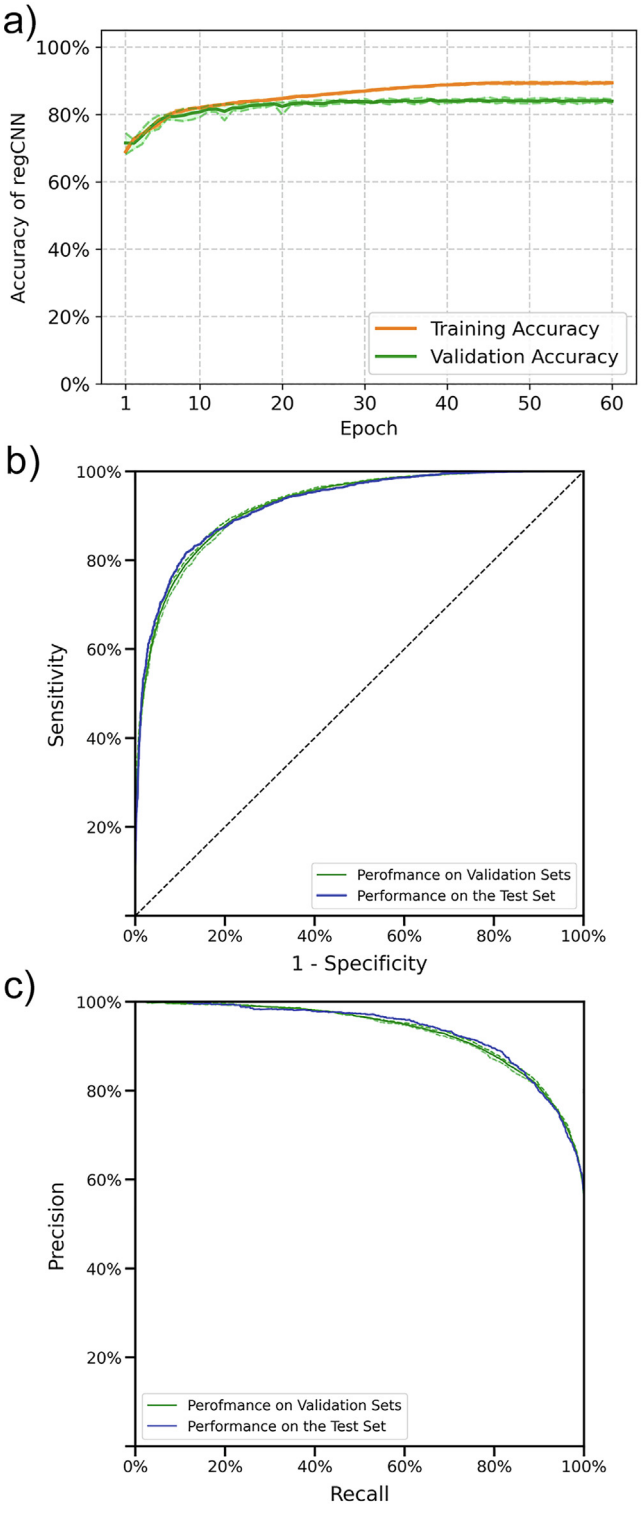

Fig. 3. The model performance of regCNN. (a) The 5-fold cross-validation learning curves of regCNN. In this figure, dashed lines represent the upper and lower bounds of the results on the 5-fold cross-validation sets. And the solid line refers to the average value of the 5-fold results. (b) The ROC curves of the cross-validation results and the test result of regCNN. (c) The PRCs of the cross-validation results and the test result of regCNN.

of regCNN are in the upper-left corner with an average validation auROC value of 92.4%, and the average validation PRC of regCNN is in the upper-right corner with an average validation auPRC value of 93.1%. Therefore, the ROC curves and PRCs reveal the ability of regCNN to discriminate CRMs from those randomly chosen negative sequences. Other average performance metrics of regCNN on the 5-fold validation sets are listed in Table 1. To evaluate the generalization of regCNN in identifying CRMs when applying to new sequences, we performed the generalization evaluation on the set-aside test set. As shown in Fig. 3-b and 3-c, the ROC curve and PRC of regCNN on the test set (test auROC = 92.5%, auPRC = 93.3%) closely track the average 5-fold validation ROC curve and PRC, indicating good general CRM-identification performance of regCNN for new sequences since the distribution of the test set is quite similar to the training/validation set. Other performance metrics calculated on the test set also lead to the same conclusion as the auROC/auPRC values (summarized in Table 1). All in all, regCNN is well-trained and can successfully discriminate CRMs from random genomic sequences with high sensitivity and specificity.

3.4. regCNN improves its performance by considering the base-by-base local regulatory patterns

To assess the improvement from considering the base-by-base local summarizing patterns of transcription regulation-related features, we re-trained the previously reported pure multi-layer perceptron (MLP) model proposed by Li et al. [22] on *Drosophila* and compared the performance of it with regCNN. The pure MLP model is a deep multi-layered neural network based on the average values of the signals within the given genomic range. The architecture of the average-based pure MLP model can be obtained by removing the hierarchical convolutions from regCNN, i.e., MLP has only two dense layers (hidden layer size = 1024) on the average values. Dropout layers were used to boost the performance of pure MLP during the training phase (dropout = 0.2). We also ensured the model training convergence of pure MLP using the learning curve technique. As shown in Fig. 4-a, pure MLP converges in the training epochs and is free of model-overfitting, implying that the comparison with regCNN is fair. We evaluated the performance of pure MLP and regCNN on the reserved test set using ROC curves and PRCs (Fig. 4-b and 4-c). The comparison shows that regCNN obtains a 4.7%/3.8% improvement in test auROC/auPRC values (regCNN 92.5%/93.3% vs. pure MLP 87.8%/89.5% in auROC/auPRC, respectively) by the local pattern summarizing convolutional layers. In addition, regCNN outperforms the pure MLP model by 5.3% in test accuracy values (regCNN 84.5% vs. pure MLP 79.2%, see Table 2 for other metrics) due to the local summarizing patterns. In summary, by considering the local patterns of various transcription regulation-related data, regCNN is confirmed to outperform the average value-based pure MLP model.

3.5. regCNN outperforms existing Drosophila CRM prediction methods on the test set

Various CRM prediction tools have been developed over the decades. We compared the performance of regCNN with other existing CRM prediction tools. The performance of different CRM prediction tools can be fairly estimated using the reserved test set. We compared the performance of regCNN on this test set with

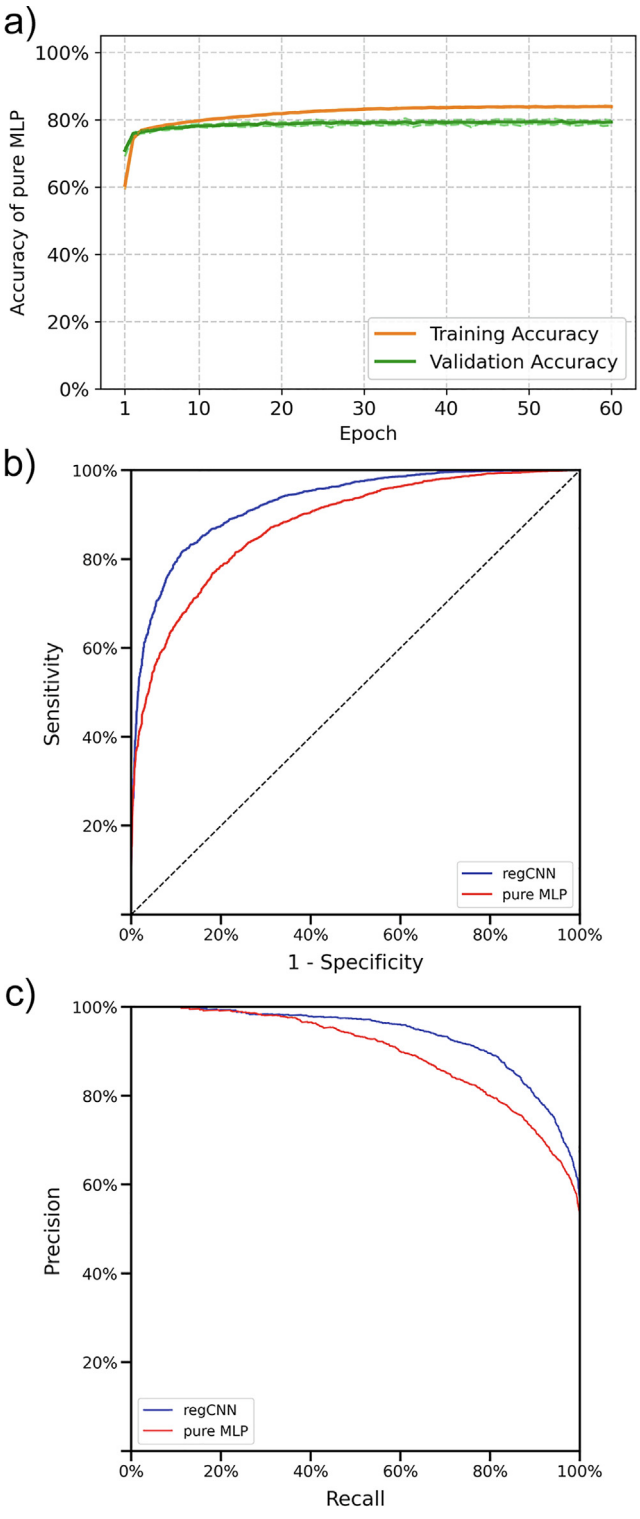

Fig. 4. Comparison between regCNN and the pure multi-layer perceptron (MLP) model. (a) The learning curves of the pure MLP model. (b) The ROC curves for regCNN and pure MLP on the test set. (c) The PRCs for regCNN and pure MLP on the test set.

Table 1
The summary of the cross-validation (CV) results and the test set result for regCNN.

regCNN	auROC	auPRC	Accuracy	Sensitivity	Precision	F1	Specificity
CV mean	92.4%	93.1%	84.2%	85.0%	84.6%	84.8%	83.2%
Test set	92.5%	93.3%	84.5%	84.3%	85.9%	85.1%	84.7%

Table 2
The summary of the test results of regCNN and the pure MLP model.

Method	auROC	auPRC	Accuracy	Sensitivity	Precision	F1	Specificity
regCNN	92.5%	93.3%	84.5%	84.3%	85.9%	85.1%	84.7%
Pure MLP	87.8%	89.5%	79.2%	78.3%	81.2%	79.8%	80.1%

10 currently available *Drosophila* CRM prediction tools (the pure MLP model, gkm-SVM [49], *cis*Module [13], *cis*ModScan [50], MCAST [14], ClusterBuster [15], MultiModule [13], MorphMS [11], *cis*PlusFinder [12], and ComSPS [51]). We excluded tools that are no longer available to the public in this comparison. For algorithms developed based on the concept of sequence alignment (MultiModule, *cis*PlusFinder, and MorphMS), we gathered the multiple alignment results of *Drosophila melanogaster* and *Drosophila persimilis* from the UCSC genome browser database [38]. MorphMS provides two log-likelihood ratio (LLR) scores against different null models. Both LLR1 and LLR2 were included in this comparison. For tools that require TF binding motif information (*cis*ModScan, Cluster-Buster, MCAST, MorphMS, and ComSPS), we adopted the same 158 TF binding PWMs used in regCNN as the inputs. We further compared regCNN with the local k-mer feature scoring tool named gapped-kmer-SVM classifier (gkm-SVM [49]). gkm-SVM utilizes the noise-resistant gapped k-mer features that convey essential DNA properties of the given sequence to identify functional elements in the genome. We downloaded LS-GKM [52], which is the newly updated gkm-SVM version for large datasets, and evaluated its performance on the reserved test set. In all tools, the fly genome sequences were downloaded from Flybase (dmel_r6.03_FB2014_06). Default parameters were applied in the CRM prediction process for each tool. And for most of the tools, the default sliding window sizes of 200 bps or 500 bps were both tested since most tools were designed with default window sizes of 200 bps and 500 bps [9].

The comparison results of regCNN with other tools are summarized in Fig. 5. Different tools were evaluated on the reserved test set, and the ROC curves and PRCs were generated. On this test set, regCNN reveals an auROC value of 92.5%, an auPRC value of 93.3%, and an accuracy value of 84.5%. Compared with the pure MLP model (auROC = 87.8%, aurPRC = 89.5%, accuracy = 79.2%), a 4.7%/3.8%/5.3% improvement in the auROC/auPRC/accuracy value can be observed for regCNN. And regCNN also outperforms gkm-SVM (auROC = 81.2%, auPRC = 82.2%, accuracy = 63.8%) by 11.3%/11.1%/20.7% in auROC/auPRC/accuracy values. In addition, more than 28.6% auROC value improvements (92.5%–63.9%, for the best existing tool MorphMS using LLR1 with the window size of 200) in CRM discrimination are achieved compared with the rest of the prediction tools (see Fig. 5). Notice that tools based on mere motif search (*cis*ModScan, Cluster-Buster, *cis*Module, MCAST, and ComSPS) only improve the performance slightly over random guesses on this test set. This may result from the TF composition combinatorial issue that hinders these tools from performing well when considering the combinations of more than two TFs. And tools (regCNN and pure MLP model) that utilize the epigenetic profiling information and TF binding preference can achieve much better results on genomic CRM identification. Based on the performance comparison on the test set, we conclude that by integrating more comprehensive genomic datasets using deep learning, regCNN generalizes well to genomic sequences and outperforms current existing tools.

4. Discussions

4.1. Using both TF binding preference and epigenetic profiling data can boost CRM identification accuracy

Current CRM identification tools usually rely on the enrichment of either the TF binding site data or epigenetic profiling signals. regCNN integrates both the TFBS data and epigenetic profiling datasets (nucleosome-free and nucleosome-variant sites, chromatin-binding protein target sites, and histone modification information) along with sequence conservation data to improve the CRM screening performance. Epigenetic profiling has been

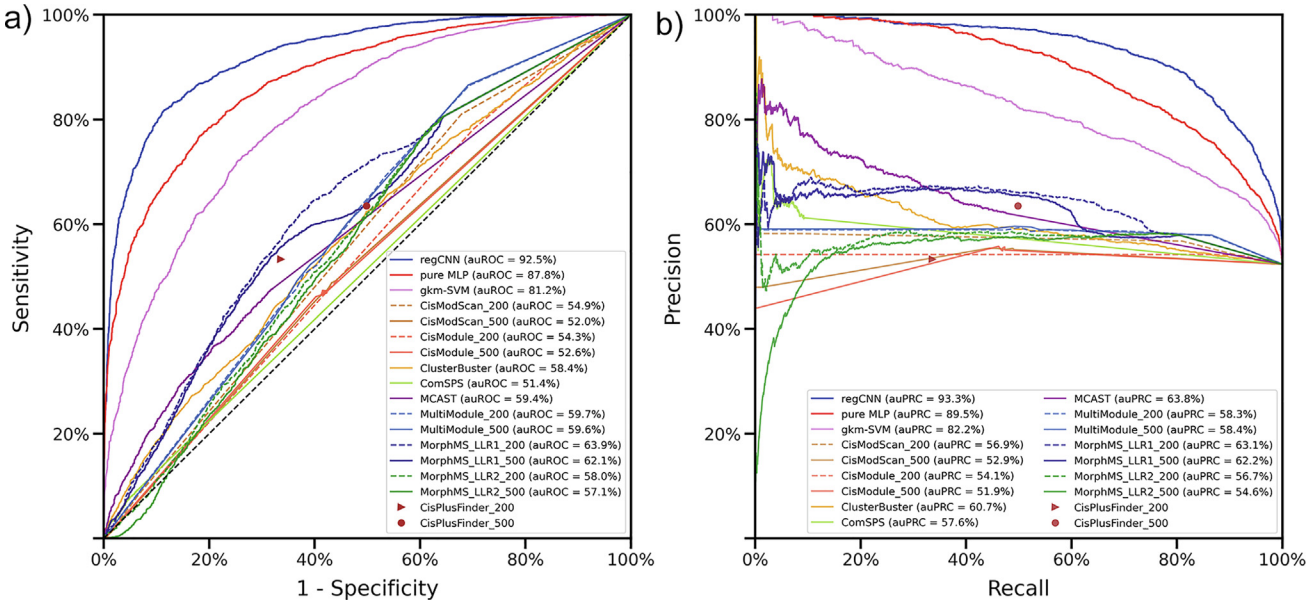

Fig. 5. (a)/(b) ROC curve/PRC comparison between regCNN and other CRM prediction tools on the test set.

shown to be closely related to the transcription mediator complex formation in metazoa species [1,31,53]. Hence, we next show that considering the local patterns of both TFBS and epigenetic profiling can enhance the CRM discrimination power. For this purpose, we trained an epigenetic profiling-depleted regCNN model using only the TFBS data and conservation scores on the training/validation set. And we also trained a TFBS-depleted regCNN model using only the epigenetic profiling data and conservation scores. Similar network architectures were retained in the epigenetic profiling-depleted model and the TFBS-depleted model. The results are plotted in Fig. 6. We used the learning curve technique to affirm the network convergence and well-fitting of the epigenetic profiling-depleted model (Fig. 6-a) and the TFBS-depleted model (Fig. 6-b). We then tested the performance of the epigenetic profiling-depleted model and the TFBS-depleted model on the test set. As shown in Fig. 6-c and 6-d, regCNN (auROC/auPRC = 92.5%/93.3%) obtains around 4% auROC improvement and 3% auPRC enhancement over both the epigenetic profiling-depleted model (auROC/auPRC = 88.9%/90.2%) and the TFBS-depleted model (auROC/auPRC = 88.5%/89.9%). The accuracy value of regCNN (accuracy = 84.5%) is also around 4% better than the epigenetic profiling-depleted model (accuracy = 80.2%) and the TFBS-depleted model (accuracy = 80.3%, see Table 3). Summarizing the performance metrics and the auROC/auPRC results, we conclude that considering the summarizing patterns in both the epigenetic profiling and TFBSs can boost CRM identification since they can reveal more subtleties in the transcription mediator complex formation.

4.2. regCNN is robust against the resizing window sizes

regCNN deals with the variable-length input sequences by using the resizing operation. In Stage I of regCNN, the retrieved experimental data of the input chromosomal region are first zero-padded to 512 bps if the region is shorter than 512 bps. On the other hand, the data are down-sampled to 512 bps using the `openCV` interpolation package [45] if the region is longer than 512 bps. We refer to this resizing length threshold as the resizing window w. In the default implementation of regCNN, w was chosen to be 512. We checked the robustness of regCNN over different resizing window w's in this subsection. In this test, w is chosen to iterate through $2^8, 2^9$ (the default w), and 2^{10}. regCNN with different resizing window w were first trained and validated on the training/validation set. And the ROC curves and PRCs were then used to evaluate the performance of the regCNN models with different resizing window w's on the set-aside test set. The results are shown in Fig. 7. From the learning curves between the training results and the validation results of models with different resizing window w's, these models were ensured to be convergent and well-trained, making the comparison fair and complete (Fig. 7-a). And on the test set, regCNN with w = 256 (accuracy = 83.4%, auROC = 91.5%, auPRC = 92.6%) shows around 1% accuracy and auROC/auPRC value decrease while the metrics are quite similar in models with w = 512 and 1024 (accuracy = 84.5%/84.1%, auROC = 92.5%/92.5%, and auPRC = 93.3%/93.5% for w = 512/1024, respectively. See Table 4, Fig. 7-b, and 7-c). Therefore,

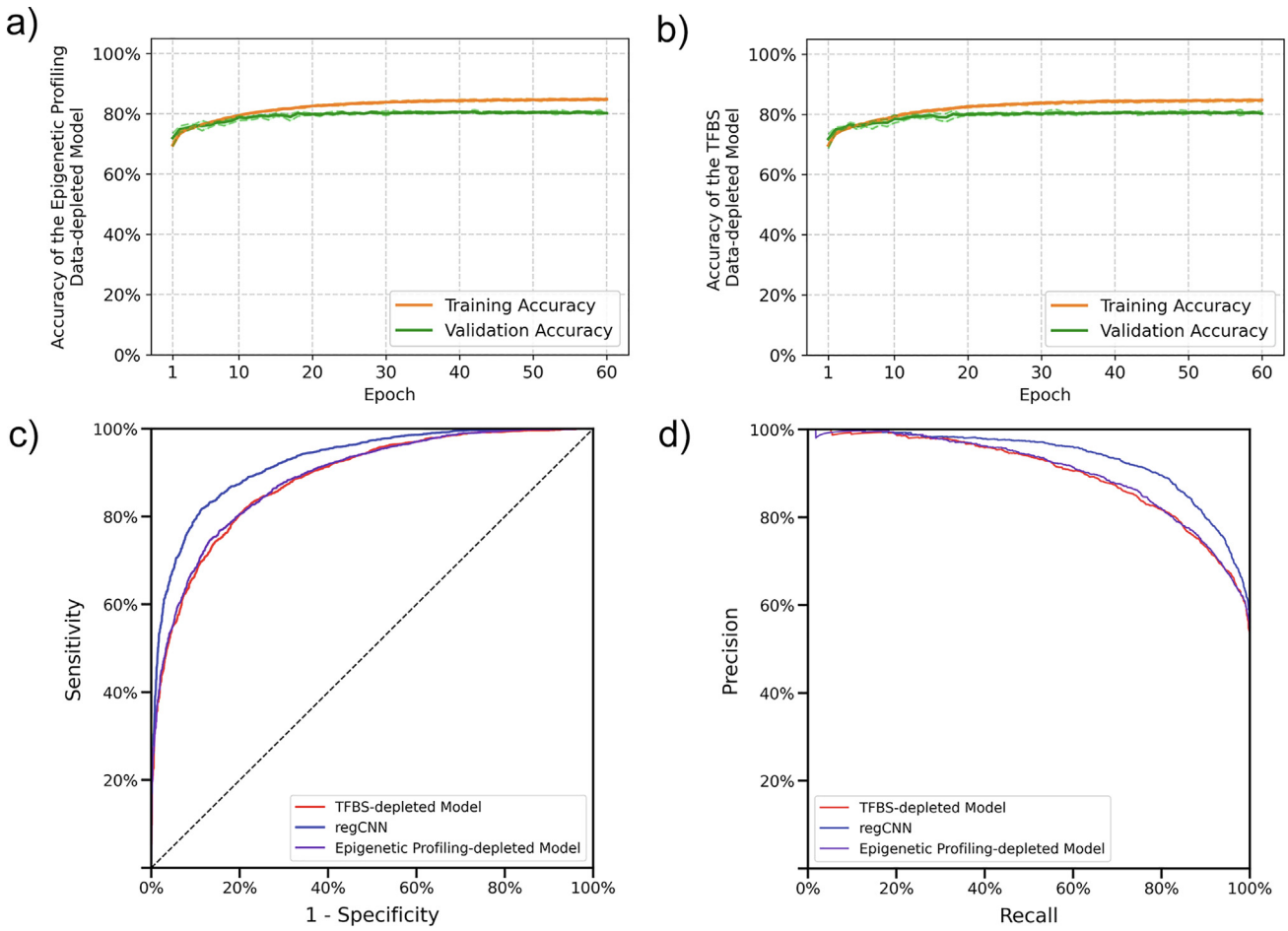

Fig. 6. The comparison of regCNN, the epigenetic profile-depleted model and the TFBS-depleted model. (a) The 5-fold cross-validation learning curves of the epigenetic profiling-depleted model. (b) The 5-fold cross-validation learning curves of the TFBS-depleted model. (c)/(d) The ROC curves/PRCs of regCNN, the TFBS-depleted model, and the epigenetic profiling-depleted model on the test set.

Table 3
The performance of regCNN versus the epigenetic profiling-depleted model and TFBS-depleted model on the test set.

Models	auROC	auPRC	Accuracy	Sensitivity	Precision	F1	Specificity
regCNN	92.5%	93.3%	84.5%	84.3%	85.9%	85.1%	84.7%
TFBS-depleted model	88.5%	89.9%	80.3%	81.6%	81.0%	81.3%	79.0%
Epigenetic profiling-depleted model	88.9%	90.2%	80.2%	79.2%	82.3%	80.8%	81.3%

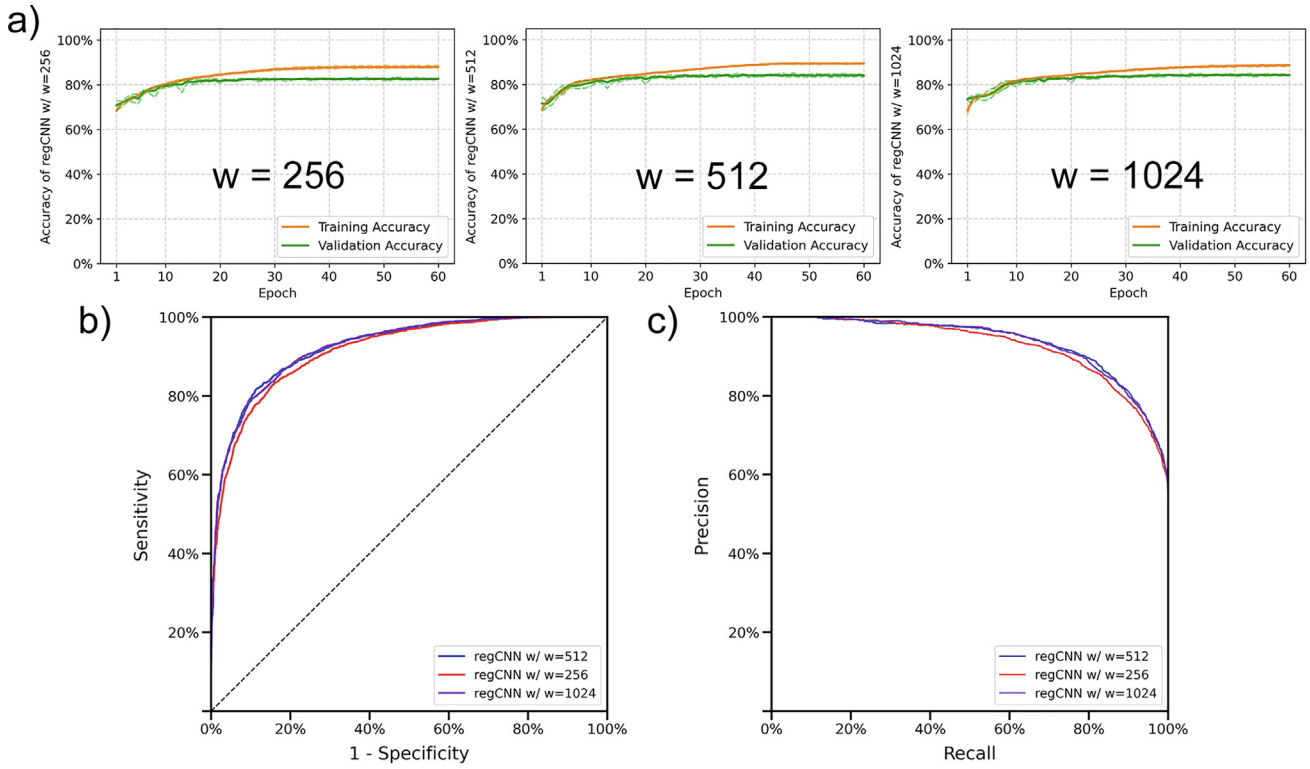

Fig. 7. regCNN is robust against the resizing window *w* in the data retrieval stage. (a) The learning curve of regCNN with resizing window *w* = 256, 512, 1024. (b) The ROC curves of the regCNN models with *w* = 256, 512, 1024 on the test set. (c) The PRCs of the regCNN models with *w* = 256, 512, 1024 on the test set.

Table 4
The performance of regCNN with different resizing window *w*'s on the test set.

w	auROC	auPRC	Accuracy	Sensitivity	Precision	F1	Specificity
256	91.5%	92.6%	83.4%	83.6%	84.6%	84.1%	83.3%
512	92.5%	93.3%	84.5%	84.3%	85.9%	85.1%	84.7%
1024	92.5%	93.5%	84.1%	83.3%	85.9%	84.6%	85.0%

regCNN is robust against different resizing window sizes as long as a sufficient length for base-by-base transcription regulation-related feature information is preserved.

4.3. The impact of variable CRM lengths on CRM prediction

CRMs in metazoa are of variable length. Hence, how the input CRM lengths impact the performance of a CRM identification model is of concern. We divided the test set into three different groups to evaluate the length-induced performance variation of regCNN: short CRMs (lengths $<$ 300 bps), medium-length CRMs (lengths between 300 and 600 bps), and long CRMs (lengths $>$ 600 bps). The performance metrics of regCNN calculated on the short CRMs, medium-length CRMs, and long CRMs are summarized in Table 5. regCNN showed 8.2%/11.0% lower auROC values in the short CRM group when compared with median-length/long test CRMs (auROC = 83.0%, 91.2%, 94.0% in the short CRMs, medium-length CRMs, and long CRMs, respectively). The ill CRM prediction performance on short CRMs may partly be due to the sparsity of short CRMs in the genome. Only 454 short CRMs were curated in the REDfly database, making it hard for models to learn the identification patterns in the short CRM group. It is also noted that longer sequences tend to be better classified by regCNN since longer CRMs usually provide richer local patterns that can help identify functional CRMs. We further computed the length-impact performance deterioration of the pure MLP model. Much worse performance decrease was observed (15.5%/14.7% auROC decrease compared with medium-length/long CRMs) in the short CRMs for the pure MLP model (auROC = 72.7%, 88.2%, 87.4% in the short CRMs, medium-length CRMs, and long CRMs, respectively). Similar trends can be observed in PRCs (see Table 5). In summary, regCNN shows lower length-induced performance deterioration than the pure MLP model and can provide better CRM identification in all CRM groups of different lengths.

Table 5
The performance summary of regCNN and pure MLP on CRMs with different sequence lengths from the test set. We grouped the CRMs in the test set into three categories to evaluate the length impact on the performance: short CRMs (< 300 bps), medium-length CRMs (300–600 bps), and long CRMs (> 600 bps).

CRM Test Group	Model	auROC	auPRC	Accuracy	Sensitivity	Precision	F1	Specificity
short CRMs	regCNN	83.0%	85.4%	74.4%	69.6%	78.0%	73.6%	79.5%
	pure MLP	72.7%	74.6%	67.8%	58.7%	73.0%	65.1%	77.3%
medium-length CRMs	regCNN	91.2%	92.4%	82.6%	82.0%	82.8%	82.4%	83.1%
	pure MLP	88.2%	90.2%	80.4%	77.9%	81.8%	79.8%	82.8%
long CRMs	regCNN	94.0%	94.4%	87.2%	87.2%	89.3%	88.3%	87.1%
	pure MLP	87.4%	89.2%	78.2%	79.4%	80.9%	80.2%	76.8%

4.4. regCNN is robust against different data partition schemes

In applying deep learning or machine learning methods to biological applications, there is a common pitfall of sample feature sharing that frequently occurs in genomes [54]. We estimated the impact of this issue for regCNN. The inflated test accuracy of many biological deep learning models is mainly caused by the data information leakage of share features between the training set and the test set [55]. Therefore, in the preparation of the training/validation set and the test set used in this research, we have enforced a restriction that sequences in the test set have no overlapping with any sequences in the training/validation set to avoid this data snooping issue. Those sequences that overlap with some sequence in the test set were removed from the training/validation set to avoid data snooping. Therefore, the data snooping issue that causes model generalization deterioration was minimized in the original construction of regCNN.

We have further used the chromosome-based and target gene-based data partition schemes to evaluate the level of reduced model generalization caused by data contamination. In the chromosome-based data partition scheme, positive and negative sequences on the X chromosome were reserved as the test set. Under the chromosome-based data partition scheme, 22,925 positive CRMs and 21,206 negative sequences were included in the training/validation set. 5,194 positive CRMs and 4,643 negative sequences were set aside in the test set. On the other hand, in the target gene-based data partition scheme, we randomly picked 215 genes and separated their regulating CRMs and closest negative non-regulating sequences as the test set. In total, 25,567/23,336 positive/negative sequences were used in the

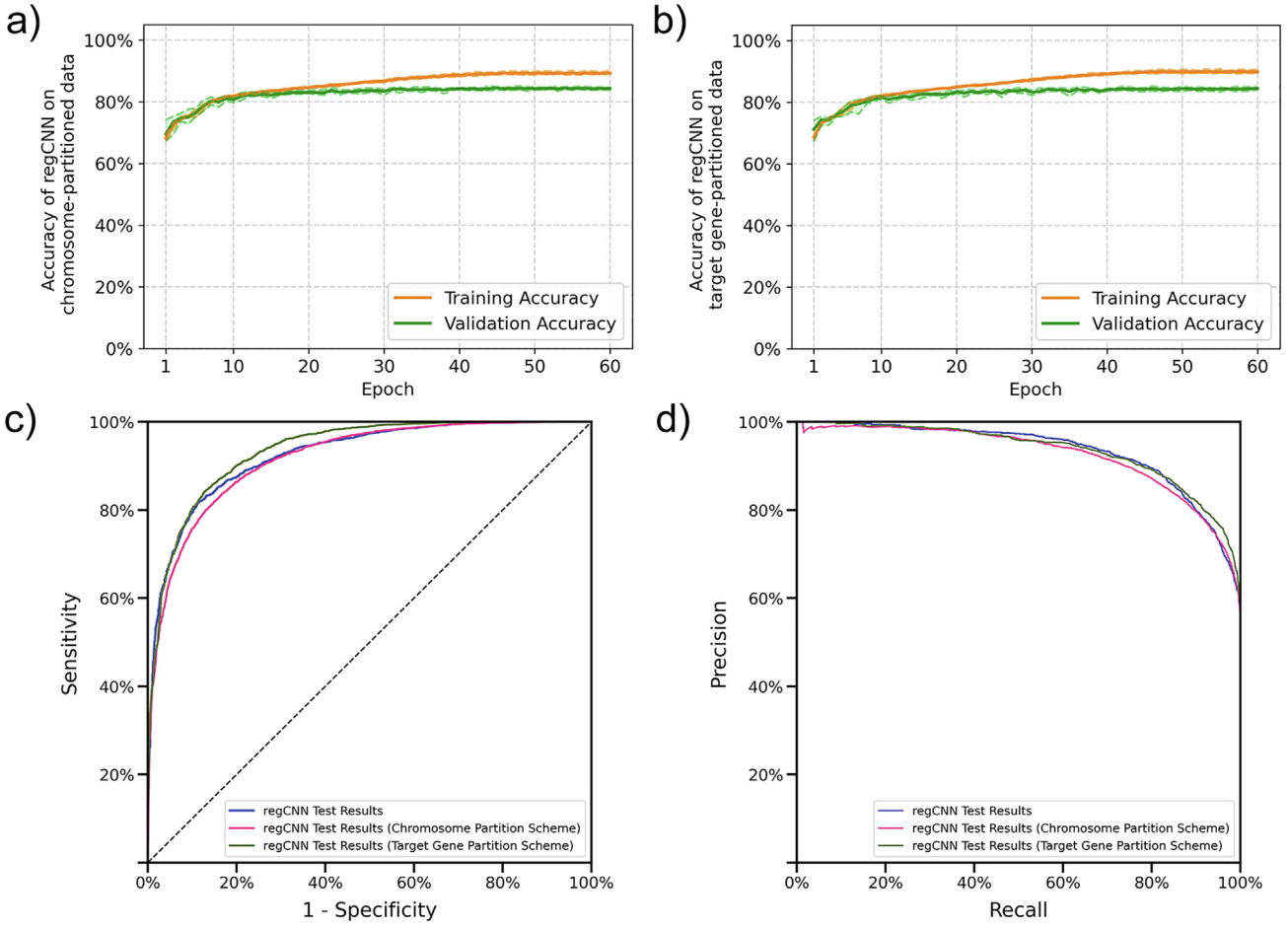

Fig. 8. regCNN is robust against different data partition schemes. (a) The learning curves of regCNN trained with the chromosome-based data partition scheme. (b) The learning curves of regCNN trained with with the target gene-based data partition scheme. (c) The ROC curve comparison of the regCNN models trained under different partition schemes. (d) The PRC comparison of the regCNN models trained under different partition schemes.

training-validation set, and 2,552/2,513 positive/negative sequences were cut out to be the test set under the target gene-based data partition scheme. We also utilized the learning curve and fivefold cross-validation techniques to ensure proper model fitting and convergence under these two data partition schemes (See Fig. 8-a and 8-b). As shown in Fig. 8-c and 8-d, the test auROC/auPRC values of regCNN are 91.8%/92.5% under the chromosome-based data partition scheme, 93.5%/93.5% under the target gene-based data partition scheme, and 92.5%/93.3% under the default random (training-vs.-test) non-overlapping data partition scheme. Under all data partition schemes, regCNN all convey high performance in terms of test auROC/auPRC values, verifying the excellent model generalization of regCNN to newly given sequences. In summary, regCNN is robust and generalizes well to genomic sequences in *Drosophila*.

4.5. The individual dataset importance in the regCNN model

In regCNN, diverse epigenetic datasets, conservation scores, and TFBS motif scores were integrated. Through hierarchical convolution operations, local patterns for each dataset were learned and utilized to identify potential CRMs. We further inferred the feature importance provided by these integrated epigenetic datasets and TFBS scores in regCNN using the SHAP (SHapley Additive exPlanations) tool [56]. SHAP was designed by the Shapley value concept in game theory to explain the importance of feature inputs to the obtained model. We first performed the SHAP analysis on the base-by-base scores of each dataset integrated in regCNN. Then the final SHAP values for individual datasets were summarized by summing the base-by-base SHAP values in each CRM. The final SHAP values of each dataset indicate the feature importance provided by the dataset local patterns in discriminating potential CRMs from random sequences. The top 10 dataset SHAP values in regCNN are plotted in Fig. 9. As shown in Fig. 9, all five genres of transcription-regulation related datasets contribute to these top ten dataset SHAP values. Some of the well-known features and epigenetic marks can be found in the top 10 important datasets [25,31]: nucleosome-free sites (which indicate open chromatin structures), sequence conservation, H3K4me3/H3K9ac ChIP scores (which are enriched in active promoters), and H3K4me1 ChIP scores (which are associated with active enhancers). Therefore, these learned representative CRM features coincide with previous transcriptional regulation studies. From the explainable AI analysis on regCNN, we conclude that regCNN identifies potential CRMs based on biologically interpretable patterns. Notice that regCNN integrates diverse epigenetic experiments conducted in different cell types. For each epigenetic mark, there are multiple experiments conducted under different cellular conditions. Furthermore, in the training process, regCNN was trained on literature-curated *Drosophila* CRMs verified by report-assay experiments under different cellular conditions. The cell type mixtures in both the epigenetic data and the CRM ground-truth dataset guided regCNN to identify all potential CRMs during the training process. Therefore, the devised model summarizes the patterns from diverse cell-type-specific features to identify all potential CRMs in different cellular conditions. The specific functional cellular conditions for each CRM are yet to be determined by further investigation in CRM-gene relations.

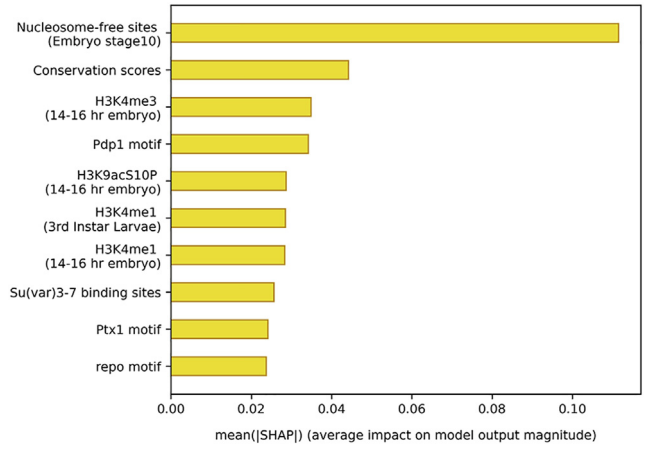

Fig. 9. The inferred dataset feature importance of regCNN represented in SHAP values.

5. Conclusions

Cis-regulatory modules (CRMs), or the modular functional DNA sequences, play essential roles in metazoa transcriptional regulation. In this research, we developed and designed a novel genomic CRM identification method called regCNN. regCNN considers and extracts the base-by-base local summarizing patterns in transcription factor binding and epigenetic profiles. We demonstrated that the designed local pattern extraction convolution architecture (auROC = 92.5%) helps improve at least 4.7% CRM discrimination auROC values over the traditional average-based pure multi-layer perceptron method (auROC = 87.8%). And by considering both the TFBS binding motifs and the epigenetic profiling datasets, 4% auROC improvement can be obtained over the epigenetic profiling-depleted model and the TFBS-depleted model. Moreover, regCNN outperforms all currently available *Drosophila* CRM prediction tools by at least 11.3% auROC values on the collected test set. We also showed that regCNN generalizes well to genomic sequences by applying different test set partition schemes. Finally, regCNN is validated to be robust against its resizing window hyperparameter in dealing with the variable lengths of CRMs. We believe that the designed algorithm can precisely identify genomic modular transcriptional regulatory DNA sequences and thus facilitates future research on metazoa transcriptional regulation.

Data Availability

The accession numbers for the nucleosome-free (DNaseI hypersensitive) and nucleosome-variant sites, the summary of the chromatin-binding protein ChIP experiments, the list of TFs included in the TF binding motif data, and the summary of the histone modification ChIP experiments are deposited at http://cobisHSS0.im.nuk.edu.tw/regCNN/. And the CRM training/validation dataset, the collected CRM test set, and the regCNN model can also be downloaded at http://cobisHSS0.im.nuk.edu.tw/regCNN/.

CRediT authorship contribution statement

Tzu-Hsien Yang: Conceptualization, Methodology, Project administration, Software, Formal analysis, Writing - Original Draft, Writing- Reviewing and Editing. **Ya-Chiao Yang:** Investigation, Software, Writing - Original Draft , Writing- Reviewing and Editing. **Kai-Chi Tu:** Investigation, Software, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Zong-Xiao Yang and Jing-Xi Xu for their help on the initial modENCODE ChIP dataset collection. This study was supported by National University of Kaohsiung and Ministry of Science and Technology of Taiwan (MOST 107-2218-E-390-009-MY3, MOST 110-2222-E-390-001).

References

[1] Yang T-H, Wang C-C, Hung P-C, Wu W-S. cisMEP: an integrated repository of genomic epigenetic profiles and cis-regulatory modules in drosophila. BMC Syst Biol 2014;8(4):S8.
[2] Yang T-H. Transcription factor regulatory modules provide the molecular mechanisms for functional redundancy observed among transcription factors in yeast. BMC Bioinform 2019;20(23):1–16.
[3] Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. Science 2006;311(5762):796–800.
[4] Poulos RC, Sloane MA, Hesson LB, Wong JW. The search for cis-regulatory driver mutations in cancer genomes. Oncotarget 2015;6(32):32509.
[5] Chatterjee S, Kapoor A, Akiyama JA, Auer DR, Lee D, Gabriel S, Berrios C, Pennacchio LA, Chakravarti A. Enhancer variants synergistically drive dysfunction of a gene regulatory network in Hirschsprung disease. Cell 2016;167(2):355–68.
[6] Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U. Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol 2004;2(9):e271.
[7] Mathelier A, Shi W, Wasserman WW. Identification of altered cis-regulatory elements in human disease. Trends Genet 2015;31(2):67–76.
[8] Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. Briefings Functional Genomics Proteomics 2009;8(4):215–30.
[9] Su J, Teichmann SA, Down TA. Assessing computational methods of cis-regulatory module prediction. PLoS Comput Biol 2010;6(12):e1001020.
[10] Niu M, Tabari E, Ni P, Su Z. Towards a map of cis-regulatory sequences in the human genome. Nucleic Acids Res 2018;46(11):5395–409.
[11] Sinha S, He X. MORPH: probabilistic alignment combined with hidden markov models of cis-regulatory modules. PLoS Comput Biol 2007;3(11):e216.
[12] Pierstorff N, Bergman CM, Wiehe T. Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. Bioinformatics 2006;22(23):2858–64.
[13] Zhou Q, Wong WH. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc National Acad Sci USA 2004;101 (33):12114–9.
[14] Bailey TL, Noble WS. Searching for statistically significant regulatory modules. Bioinformatics 2003;19(suppl 2):ii16–25.
[15] Frith MC, Li MC, Weng Z. Cluster-Buster: Finding dense clusters of motifs in dna sequences. Nucleic Acids Res 2003;31(13):3666–8.
[16] Navarro C, Lopez FJ, Cano C, Garcia-Alcalde F, Blanco A. CisMiner: genome-wide in-silico cis-regulatory module prediction by fuzzy itemset mining. PLoS One 2014;9(9):e108065.
[17] Niu M, Tabari ES, Su Z. De novo prediction of cis-regulatory elements and modules through integrative analysis of a large number of ChIP datasets. BMC Genomics 2014;15(1):1047.
[18] Blanchette M, Bataille AR, Chen X, Poitras C, Laganière J, Lefèbvre C, Deblois G, Giguère V, Ferretti V, Bergeron D, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 2006;16(5):656–68.
[19] Kouzarides T. Chromatin modifications and their function. Cell 2007;128 (4):693–705.
[20] Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. A map of the cis-regulatory sequences in the mouse genome. Nature 2012;488(7409):116.
[21] Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. A cis-regulatory map of the Drosophila genome. Nature 2011;471(7339):527–31.
[22] Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinform 2018;19 (1):202.
[23] Washington NL, Stinson E, Perry MD, Ruzanov P, Contrino S, Smith R, Zha Z, Lyne R, Carr A, Lloyd P, et al. The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details. Database: J Biological Databases Curation 2011;2011(23).
[24] Chen A, Chen D, Chen Y. Advances of DNase-seq for mapping active gene regulatory elements across the genome in animals. Gene 2018;667.
[25] Boros IM. Histone modification in Drosophila. Briefings Functional Genomics 2012;11(4):319–31.
[26] Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 2010;471 (7339):480–5.
[27] Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. Systematic protein location mapping reveals five principal chromatin types in drosophila cells. Cell 2010;143(2):212–24.
[28] Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M, et al. Conservation of core gene expression in vertebrate tissues. J Biol 2009;8(3):33.
[29] Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. Functional evolution of a cis-regulatory module. PLoS Biol 2005;3(4):e93.
[30] Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 2010:1186176.
[31] Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet 2012;13(7):469–83.
[32] Yang T-H, Wu W-S. Identifying biologically interpretable transcription factor knockout targets by jointly analyzing the transcription factor knockout microarray and the ChIP-chip data. BMC Syst Biol 2012;6(1):102.
[33] Yang T-H, Wang C-C, Wang Y-C, Wu W-S. YTRP: a repository for yeast transcriptional regulatory pathways. Database: The J Biolog Databases Curation 2014.
[34] Rivera J, Keränen SVE, Gallo SM, Halfon MS. REDfly: the transcriptional regulatory element database for Drosophila. Nucleic Acids Res 2019;47(D1): D828–34.
[35] Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, et al. FlyBase 2.0: the next generation. Nucleic Acids Res 2019;47(D1):D759–65.
[36] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15 (8):1034–50.
[37] A. Siepel, D. Haussler, Phylogenetic hidden Markov models, in: Statistical Methods in Molecular Evolution, Springer, 2005, pp. 325–351..
[38] Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, et al. The UCSC genome browser database: 2021 update. Nucleic Acids Res 2021;49(D1):D1046–57.
[39] Fornes O, Castro-Mondragon JA, Khan A, Van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2020;48(D1):D87–92.
[40] Jansen A, Verstrepen KJ. Nucleosome positioning in Saccharomyces cerevisiae. Microbiol Mol Biol Rev 2011;75(2):301–20.
[41] Thomas S, Li X-Y, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, et al. Dynamic reprogramming of chromatin accessibility during drosophila embryo development. Genome Biol 2011;12(5):R43.
[42] Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, et al. Nucleosome organization in the Drosophila genome. Nature 2008;453(7193):358–62.
[43] Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics 2019;35 (3):421–32.
[44] Yang T-H, Wang C-Y, Tsai H-C, Liu C-T. Human IRES Atlas: an integrative platform for studying ires-driven translational regulation in humans. Database: J Biological Databases Curation 2021.
[45] Bradski G, Kaehler A. Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media Inc; 2008.
[46] Azodi CB, Tang J, Shiu S-H. Opening the black box: Interpretable machine learning for geneticists. Trends Genet 2020;36.
[47] Abu-Mostafa YS, Magdon-Ismail M, Lin H-T. Learning from data, Vol. 4. NY, USA: AMLBook New York; 2012.
[48] Yang T-H. An aggregation method to identify the RNA meta-stable secondary structure and its functionally interpretable structure ensemble. IEEE/ACM Trans Comput Biol Bioinf 2021.
[49] Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol 2014;10(7):e1003711.
[50] Johnson DS, Zhou Q, Yagi K, Satoh N, Wong W, Sidow A. De novo discovery of a tissue-specific gene regulatory module in a chordate. Genome Res 2005;15 (10):1315–24.
[51] Guo H, Huo H. A new algorithm for identifying cis-regulatory modules based on hidden Markov model. BioMed Research Iternational 2017;2017.
[52] Lee D. LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics 2016;32(14):2196–8.
[53] Soutourina J. Transcription regulation by the Mediator complex. Nat Rev Mol Cell Biol 2018;19(4):262–74.
[54] Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. PLoS Comput Biol 2018;14(12):e1006625.
[55] Cao F, Fullwood MJ. Inflated performance measures in enhancer–promoter interaction-prediction methods. Nat Genet 2019;51(8):1196–8.
[56] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions, in. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–77.