PLOS ONE

RESEARCH ARTICLE

SMCis: An Effective Algorithm for Discovery of Cis-Regulatory Modules

Haitao Guo, Hongwei Huo*, Qiang Yu

School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China

* hwhuo@mail.xidian.edu.cn

OPEN ACCESS

Citation: Guo H, Huo H, Yu Q (2016) SMCis: An Effective Algorithm for Discovery of Cis-Regulatory Modules. PLoS ONE 11(9): e0162968. doi:10.1371/journal.pone.0162968

Editor: Martina Stromvik, McGill University, CANADA

Received: March 17, 2016

Accepted: August 31, 2016

Published: September 16, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 61173025, 61373044, and 61502366, the China Postdoctoral Science Foundation under Grant 2015M582621, and the Fundamental Research Funds for the Central Universities under Grant JB150306. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The discovery of *cis*-regulatory modules (CRMs) is a challenging problem in computational biology. Limited by the difficulty of using an HMM to model dependent features in transcriptional regulatory sequences (TRSs), the probabilistic modeling methods based on HMMs cannot accurately represent the distance between regulatory elements in TRSs and are cumbersome to model the prevailing dependencies between motifs within CRMs. We propose a probabilistic modeling algorithm called SMCis, which builds a more powerful CRM discovery model based on a hidden semi-Markov model. Our model characterizes the regulatory structure of CRMs and effectively models dependencies between motifs at a higher level of abstraction based on segments rather than nucleotides. Experimental results on three benchmark datasets indicate that our method performs better than the compared algorithms.

Introduction

The regulation of gene expression involves the binding of transcription factors (TFs) to transcription factor binding sites (TFBSs) [1, 2]. The TFBSs bound by the same transcription factor usually share a conserved DNA sequence pattern called a DNA motif. In higher eukaryotes, gene expression is cooperatively regulated by a number of transcription factors binding to various TFBSs. These TFBSs are tightly clustered and form *cis*-regulatory modules (CRMs) to recruit bound transcription factors and perform more elaborate and accurate regulation. These CRMs are usually scattered across large genomic regions and have lengths ranging from several tens of base pairs (bp) to several thousands of base pairs. We refer to the functional regions harboring CRMs as transcriptional regulatory sequences (TRSs); TRSs include promoter regions, distal DNA regions such as enhancers located in introns, and even other intergenic regions that are far from transcription start sites (TSSs) but still perform implicit regulatory functions. Playing pivotal roles in the regulation of gene expression, CRMs are believed to have a specific regulatory structure, as shown in Fig 1. The computational discovery of CRMs is a key step for constructing a regulatory network.

Experimental identification of the biochemical features [3] closely associated with CRMs, such as occupancy by transcription factors and histone modifications, is an effective method for the discovery of CRMs. However, the experimental determination of these features is costly

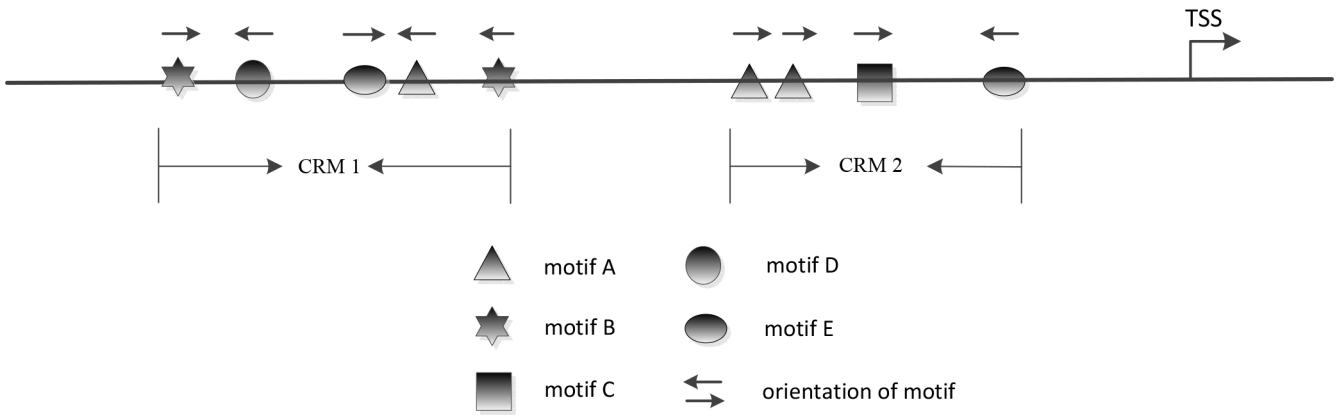

Fig 1. Regulatory structure of a CRM. A CRM is a sequence segment that contains instances of multiple motifs. The orientations of the motifs, the distance between motifs and the relationships between motifs may be key properties of CRMs.

doi:10.1371/journal.pone.0162968.g001

and time consuming, and this approach can be limited by the number of antibodies and cell types available. Therefore, it is necessary to discover CRMs with the aid of computational methods.

The computational methods used to predict CRMs face the following challenges. The CRMs have flexible structural organization; in a CRM, partial motifs show order preferences, and the distances between them are not fixed. It is difficult to accurately describe such a CRM structure. Eukaryotic regulatory regions are usually large, and the motifs constituting CRMs are often short and degenerate, typically 4–20 bp long. It is difficult to identify motifs in such a large potential search space [4, 5]. This challenge makes it difficult to look for CRMs by identifying their component motifs directly from sequences.

Most CRM discovery methods take advantage of the following general features of CRMs: i) Clustering of motifs: multiple cooperating transcription factors binding to a CRM may lead to the clustering of motifs in a small sequence region. ii) Evolutionary conservation: functional sequences exhibit a lower frequency of mutations than non-functional sequences over evolutionary time. iii) Available motif profiles (from existing motif databases such as TRANSFAC [6] and JASPAR [7]): it is simpler to search for motif instances by using profile matrices from motif libraries than to perform *de novo* motif prediction using computational methods.

A variety of models and methods have been proposed to predict CRMs [3, 8–15] in eukaryotic genes. Different methods take advantage of different features of CRMs and use diverse search strategies. These methods can be classified into the following three categories according to the search strategies.

One group of methods searches for CRMs based on window clustering and makes use of the clustering of motifs. Some methods, such as MSCAN [16] and MCAST [17], use a simple means of representing a CRM as a region with a high density of motifs within a window. These methods infer CRMs by counting the number of occurrences of given motifs within a sequence window. Other methods use combinatorial search approaches to look for clusters of motifs that co-occur significantly within a given size window; for example, CisMiner [13] detects CRMs by the fuzzy clustering of closely located motifs and CPModule [18] identifies CRMs based on itemset mining. In essence, the methods in this category assume that the motifs within each sequence window are independent and identically distributed. Moreover, it is not a trivial task to determine a reasonable window size and score thresholds.

A second group of methods builds probabilistic models for CRMs and identifies the sequence regions matching a statistical model of a motif cluster better than a background model. Except

for a small number of methods based on discriminative models, such as HexDiff [19], regulatory Potential [20] and CRFEM [21], these methods use generative models. The most commonly used generative model is the HMM [22]. The HMM can provide a statistically reliable measure of the occurrences of CRMs and motifs, and it can characterize the regulatory structure of CRMs. Additionally, the expectation-maximization (EM) algorithm used in model learning can automatically estimate a large number of model parameters. The methods based on HMM models often represent TRSs that contains motifs and CRMs as observations generated by a hidden Markov stochastic process. Compared with the window clustering methods, the methods in this category do not require the consideration of window sizes and score thresholds. Early methods, such as CisModule [23] and Cluster-Buster [24], implement simple HMM models with the states describing motifs and intra-module and inter-module backgrounds to infer CRMs. Geometric distributions for state durations in the HMM putatively specify the inter-motif and inter-module distances. However, these methods only model combinations of motifs; they do not consider any preferential ordering of motifs within a CRM. Later methods, such as Stubb [25] and BayCis [26], further extend this model by introducing transitions between motif states.

A third group of methods searches for CRMs in evolutionarily conserved regions. Some methods, such as MorphMS [27] and StubbMS [25] (the multi-species version of Stubb), first identify conserved regions by using pairwise or multiple sequence alignments in the regulatory regions of related genes, then model the motif clusters within those regions by using a TFBS evolution stochastic model to identify conserved CRMs. However, since the regulatory regions of most genes suffer from a large number of events such as shuffling, deletion and duplication, these methods are difficult to align them. To get around this problem, other methods, such as EEL [28] and ReLA [29], have been proposed. These methods align the pre-identified motif instances instead of raw sequences to detect conserved motif cluster regions. Although the methods in this category have shown promising prediction performance for CRMs, they are limited to related species and thus do not always work.

Of all these methods, the probabilistic modeling methods based on HMMs are the most common and most effective. However, the traditional HMM has two drawbacks that limit its prediction performance. First, HMM state durations are implicitly assumed to be geometric distributions. This assumption is unrealistic because the distances between motifs within a CRM may not be well described by a geometric distribution. Second, the HMM is an unwieldy way to model large numbers of dependences. Current methods based on HMMs usually assume that motifs within a CRM are generated independently by corresponding HMMs. However, transcription factors bound to a CRM cooperate with each other to regulate gene expression. This behavior implies that the motifs corresponding to these transcription factors may be correlated. Thus, the independence assumption may cause predictions to be inaccurate. Although the HMM may be extended to model these correlations by adding extra states and parameters, the extended model may require excessive computational work.

To address these problems, this paper presents a probabilistic modeling method called SMCis. The method builds a CRM discovery model based on a hidden semi-Markov model (HSMM) [30]. We use this sophisticated HMM at a higher level of abstraction (i.e., segments rather than nucleotides) to characterize the regulatory structure of CRMs. Unlike general CRM discovery methods, we consider the distances and ordering of motifs within a CRM instead of simply regarding a CRM as a cluster of motifs. Specifically, we infer the CRM structure from the frequencies of motif occurrences and the dependences and distance specificities between motifs within a CRM. The dependences and distance specificities between motifs within a CRM encode gene regulation information. Modeling these features helps to improve the accuracy of CRM discovery. We test our method on three annotated real biological datasets and

compare it with current published methods. Experimental results suggest that our method performs better than the compared algorithms.

Materials and Methods

The HSMM has more modeling power than the HMM and can explicitly model the state durations and the long-range dependencies between observations and states. Thus, the HSMM is a natural and effective approach for modeling TRSs and CRMs. We used an HSMM to characterize the organization of TRSs and the putative transcriptional regulatory structure of CRMs. The structure models the dependencies and distance specificities between motifs within a CRM and describes the internal organization of a CRM.

In this section, we first introduce the details of the model. Then, the algorithms for learning and inference are given. Finally, we describe an algorithm to reduce the search space of the model.

Construction of the HSMM

The HSMM [30] is an extension of the classical HMM. In contrast to an HMM, each state of which emits one observation, each state of the HSMM can emit strings of observations. In our model, an observation denotes an observed nucleotide. The observations emitted by an HSMM state are governed by a segment model. The segment model gives a joint model for random-length strings of observations. Formally, given an observation sequence $o_{1:t} = o_1 o_2 \ldots o_t$ generated by state s, the segment model [31] can be expressed as follows:

$$P(o_{1:t}, t|s) = P(t|s)P(o_{1:t}|t, s) = d_s(t)e_s(o_{1:t}) \quad (1)$$

The segment model consists of a duration distribution $d_s(t)$, describing the probability that observations generated by state s have length t, and an emission model $e_s(o_{1:t})$, giving the emission probability that state s generates the particular observations $o_{1:t}$.

We use an HSMM to describe the regulatory structure of TRSs. Fig 2 shows the HSMM state diagram that describes the regulatory structure of a TRS. In the HSMM, TRSs containing CRMs are organized in a two-level hierarchy. At the top level, each TRS is viewed as a concatenation of CRMs and inter-module (global) backgrounds. At the bottom level, each CRM is considered a combination of motifs and intra-module (local) backgrounds. The motifs and intra-module backgrounds at the bottom level are viewed as nucleotide segments but are not further divided. Formally, in the model, a CRM is denoted by two dummy states, c_s and c_e, where c_s initializes a CRM instance and c_e correctly terminates the CRM instance. Let b_g and b_c denote the global and local backgrounds, respectively. We let $M = \{m_1, m_2, \ldots, m_K\}$ represent the set of motif states and define $M' = \{m_1', m_2', \ldots, m_K'\}$ to be its reverse complement. In addition, to make the HSMM well defined, we add the initial state **S** and the termination state **E**. Therefore, the state space of the whole model is denoted by $H = \{\mathbf{S}, \mathbf{E}\} \cup \{c_s, c_e\} \cup M \cup M' \cup \{b_g\} \cup \{b_c\}$.

To capture the dependencies between adjacent motifs within a CRM, we define the direct transitions between motif states, as shown in Fig 2. Dependencies between motifs within a CRM implicitly specify the ordering in the spatial arrangement of these motifs. The ordering of motifs within a CRM may affect combinatorial transcriptional regulation [32]. Thus, modeling the dependencies between motifs helps to uncover the mechanism of TFBS regulation within a CRM.

Motifs within a CRM may comply with specific spacing requirements to allow corresponding transcription factors to bind to them. Several previous studies [33] have suggested that a large number of motif pairs suffer from distance constraints under selection forces and exhibit significant distance specificity in human promoters. To model the motif distance specificity,

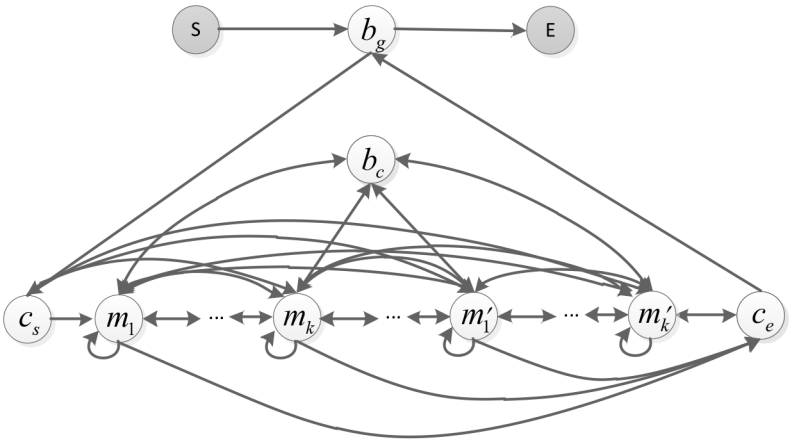

Fig 2. The SMCis HSMM state transition diagram. Nodes represent emission states, including motifs, inter-module backgrounds and intra-module backgrounds. Shadow nodes represent the initial and termination states of the model and the initial and termination states of the CRMs. Arrows indicate permissible transitions between states.

doi:10.1371/journal.pone.0162968.g002

we approximate the state durations of intra-module backgrounds with a more flexible distribution instead of the geometric distribution implicitly assumed by an HMM.

The HSMM contains three types of emission states: motif state $m \in M \cup M'$, global background state b_g and local background state b_c. Once entering an emission state, the HSMM first resorts to a duration model to determine the duration for the visit. Then, it uses a state emission model to emit this number of nucleotides until it transitions to the next state. The specific state duration and state emission models are defined in the following sections.

State duration model

In the model, state duration models are addressed by the length distributions that represent the length characteristic of nucleotides generated by a particular state.

Position weight matrices (PWMs) [34] of all the motifs are obtained directly from databases (such as TRANSFAC and JASPAR). Given a motif state m, let w_m be its length. The following formula defines the probability distribution that a sequence generated by the motif state m has length u_m:

$$d_m(u_m) = \begin{cases} 1, & if\ u_m = w_m \\ 0, & if\ u_m \neq w_m \end{cases}. \tag{2}$$

In our model, we do not make any assumption regarding the distance distribution between CRMs and we still use a geometric distribution. The probability that a background sequence generated by the global background state b_g has length u_g is defined as follows:

$$d_{b_g}(u_g) = (1 - p_g)^{u-1} p_g, \tag{3}$$

where p_g is the parameter of the distribution.

The sequence modeled as an intra-module background defines the distance between adjacent motif instances. The geometric distribution under an HMM is not a good approximation of the distance between motifs. Based on previous research [26], we use a negative binomial distribution to approximate the distribution of the distances between motifs. This distribution

is given as follows:

$$d_{b_c}(u_c) = \binom{u_c - 1}{r - 1} \pi^{u_c - r}(1 - \pi)^r, \tag{4}$$

where b_c is the local background state, u_c is the length of a local background sequence generated by the state b_c, and r and π are the parameters of the distribution function.

State emission model

The HSMM state emission models define joint probability distributions of nucleotides at particular sites generated by a particular state.

For the global background and local background states, we use the mth order and m'th order local Markov models, respectively. In a kth order local Markov model with $k = m$ or m', the probability of generating a background sequence $o_{1:v}$ is given as follows:

$$e_b(o_{1:v}) = \prod_{i=1}^{v} P(o_i | o_{i-k:i-1}), \tag{5}$$

where b denotes the global or local background state.

In Eq (5), $p(o_i|o_{i\text{-}k:i\text{-}1})$ is the conditional probability of the nucleotide o_i occurring at position i given k preceding nucleotides $o_{i\text{-}k:i\text{-}1}$. This probability can be computed as follows:

$$P(o_i | o_{i-k:i-1}) = \frac{P(o_{i-k:i})}{\sum_{o_i=\mathrm{A}}^{\mathrm{T}} P(o_{i-k:i-1}, o_i)}. \tag{6}$$

In Eq (6), $p(o_{i\text{-}k:i})$ and $p(o_{i\text{-}k:i\text{-}1})$ are estimated from the frequencies of all (k+1)-mers and all k-mers in an i-centered window with length 2D, respectively. Here, D is a predefined parameter. To optimize the computational efficiency, a sliding window approach is used. The sliding window approach scans the whole sequence in one pass, and the conditional probabilities of nucleotides at all positions are calculated. The detailed steps are as follows:

1. Count each (k+1)-mer within the current window, and calculate and store the conditional probability of the nucleotide at each position within the current window.
2. Move the window in a fixed step size, update counts of the (k+1)-mers within the current window, and calculate and store the conditional probability of the nucleotide at each position within the current window.
3. Repeat steps i) and ii) until the conditional probabilities of the nucleotides at all positions are calculated.

For the motif state, we use the standard product multinomial (PM) model [35]. The PM model is a simple motif model based on PWMs, which assumes that nucleotides at all positions in the motifs are independent. Given a motif m, let its PWM $\Theta = [\theta_1, \theta_2, \ldots, \theta_L]$, where θ_i ($1 \leq i \leq L$) is a column vector of the frequencies of nucleotides A, T, G and C. The probability of generating a motif instance $o_{1:L}$ by the motif state m is given as follows:

$$e_m(o_{1:L}) = \prod_{i=1}^{L} \theta_{o_i,i}, \tag{7}$$

where o_i is the nucleotide at position i of the motif instance $o_{1:L}$.

Inference and learning

In an HMM, model parameters can be estimated by using the Baum-Welch algorithm [36]. The Baum-Welch algorithm uses the EM algorithm [36] to find the maximum likelihood estimate of model parameters. We extend this algorithm to obtain an algorithm called the modified Baum-Welch to estimate the model parameters of our HSMM.

Given a TRS o, let s denote one of its state paths and let u denote the corresponding duration sequence. Let θ denote all of the model parameters. The likelihood of a TRS is defined as follows:

$$L(\theta) = \sum_{s,u} P(s, u, o|\theta). \tag{8}$$

Let θ^t represent the value of parameter θ at t-th iteration, and the corresponding Q-function [36] is defined as the likelihood of conditional expectation:

$$Q(\theta|\theta^t) = \sum_{s,u} P(s, u|o, \theta^t) \log P(o, s, u|\theta). \tag{9}$$

We compute Eq (9) by using the modified Baum-Welch algorithm based on the EM algorithm to iteratively converge to a locally optimal θ.

Once the model is trained, we use the Posterior Viterbi algorithm [37] to discover CRMs in the TRSs. Combining ideas of the Viterbi algorithm [36] and the posterior decoding algorithm [36], the Posterior Viterbi algorithm finds the legal path with the maximum joint posterior probability in the posterior probability space. Formally, given an un-annotated TRS $o_{1:T}$, the Posterior Viterbi algorithm is used to find a state path $s = s_{1:N}$ and a corresponding state duration sequence $u = u_{1:N}$ according to the following equation:

$$(s, u) = \underset{N, s=s_{i:N} \in A_p}{\arg\max} \prod_{i=1}^{N} P(s_i, u_i|o_{1:T}, \theta), \tag{10}$$

where N is the possible number of states and A_p is the set of the allowed posterior paths through the HSMM model.

Reducing the search space

The HSMM can provide better expressive power than the HMM, but it adds an additional dimension to infer state durations and needs to explicitly evaluate different segmentations in the learning and inferring of the model [31]. Here, each segmentation determines a mapping from a TRS to a set of state labels corresponding to a state path. To reduce the search space, we locate all putative motif instances of given PWMs that are significant matches and may form a CRM before parsing a TRS. In parsing the TRS, the HSMM only considers the state paths through the positions of the motif matches and combines these pre-identified motif instances to identify the best motif clusters as candidate CRMs.

Results

We tested our method on three real biological datasets: the muscle-specific expression system, the liver-specific expression system and the *Drosophila* early embryonic development system. We refer to these as the muscle dataset, the liver dataset and the *Drosophila* early development dataset, respectively. The sequences in the muscle and liver datasets are from co-regulated genes, and the sequences in the *Drosophila* early development dataset are from orthologous genes.

We chose six methods and compared their prediction performances with that of SMCis on these datasets. These methods encompass a wide spectrum of extant models: BayCis [26], Stubb [25], MSCAN [16], MotEvo [38], Cluster-Buster [24] and ReLA [29].

Evaluation

To avoid bias toward any particular measure, we used the correlation coefficient (CC) [39] and the F1-score [40] of precision and recall to evaluate the overall prediction performance of our method on all datasets.

Not all of the evaluated methods provide information about motifs within the predicted CRMs; thus, we compared the results at the nucleotide level. Given the prediction results of a method, the CC and F1 scores are defined as follows:

$$\mathrm{CC} = \frac{\mathrm{TP} \times \mathrm{TN\text{-}FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{11}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}}. \tag{12}$$

In Eq (11), TP is the number of nucleotides correctly predicted as CRMs, TN is the number of nucleotides correctly predicted as background, FP is the number of actual background nucleotides mistakenly predicted as CRMs, and FN is the number of actual CRM nucleotides mistakenly predicted as background. In Eq (12), the precision and recall are defined as follows:

- Precision (Pr), Pr = TP / (TP + FP), measures the ratio of correctly predicted CRMs to the total number of predicted CRMs.
- Recall (Re), Re = TP / (TP + FN), measures the ratio of correctly predicted CRMs to the total number of actual CRMs.

Because it uses all four values, TP, TN, FP and FN, the CC is a balanced measure of the overall performance of a method. The CC is interpreted statistically as the correlation between positions in predicted CRMs and positions in actual CRMs. The value of CC ranges from -1 to +1; a value of +1 indicates that the prediction results are fully coincident with the actual results, and a value of -1 means the opposite; the value tends toward 0 when the predictions are close to random.

Precision and recall are antagonistic measures, which means that a higher recall usually comes at the cost of lower precision. The F1-score is defined as the harmonic mean of the antagonistic pairs to balance them. The value of the F1-score ranges from 0 to +1, with +1 indicating perfect predictions.

Results on the muscle and liver datasets

The datasets. The muscle and liver datasets were originally compiled by Wasserman et al. [41, 42], and these datasets have been widely used to evaluate the prediction performance of CRM discovery methods. Klepper et al. [8] expanded the datasets and used them as benchmarks. The two datasets used here were from Klepper et al. [8].

The muscle dataset consists of five motifs and 24 sequences. The five motifs are Mef2, Myf, Sp1, SRF and Tef, which play important roles in the transcriptional regulation of vertebrate muscle gene expression. The average length of the 24 sequences is 850 bp, and these sequences come from rat, human, chicken and cow. The 24 sequences contain 84 instances of the five motifs in total. Each sequence contains one CRM. The average length of these CRMs is 120 bp, ranging from 14 to 294 bp.

The liver dataset includes four motifs and 12 sequences. The four motifs are HNF-1, HNF-3, HNF-4 and C/EBP, which regulate liver-specific gene expression. Each of the 12 sequences, except for the sequence *M*19524 (943 bp long), has a length of 1 Kbp, and the sequences come from human, mouse, rat and chicken. These sequences contain 14 CRMs in total, and each sequence has one or two CRMs. The lengths of these CRMs range from 22 to 176 bp, and the average length is 112 bp.

Experimental setup. For the two datasets, we used all sequences as the training and test sets. For the methods Stubb, MSCAN and MotEvo, which depend on a window size, we set the window size to 200 bp and kept the remaining parameters at their default values. The default settings were kept in the other methods.

CRM prediction performance. For all the methods, we calculated the CC, F1-score, precision and recall by summing up all four values TP, TN, FP and FN over all sequences on each of the two datasets, as shown in Figs 3 and 4.

Fig 3 shows that SMCis has the highest CC and F1-score on the muscle dataset; its prediction precision is higher than that of other methods, and its recall is more than half of these methods. As is apparent from Fig 3, different methods have different trends and achieve different balances on this dataset. MotEvo, Stubb, Cluster-Buster and ReLA tend to make predictions with very high recall, but they make a lot of false predictions and do not achieve reasonable balances between precision and recall. BayCis tends to make conservative predictions to ensure high precision, and this is reflected in the fact that its recall is significantly lower than that of the other methods. MSCAN emphasizes neither precision nor recall but provides very high precision while ensuring high recall.

As shown in Fig 4, all methods except ReLA produced better predictions on the liver dataset than the muscle dataset. SMCis had the highest CC score and F1-score on this dataset. Although SMCis had a mid-level recall score, it had high precision that was only slightly lower than that of MSCAN and significantly higher than other methods.

Results on the *Drosophila* early development dataset

The datasets. For the dataset, we used seven motifs—Gt, Hb, Tll, Cad, Kni, Bcd and Kr—which drive the transcriptional regulation of *Drosophila* early embryonic development, and we downloaded PWMs for these motifs from the iDMMPMM database [43]. We selected a subset containing nine genes: *kni*, *Kr*, *hb*, *tll*, *btd*, *eve*, *h*, *ftz* and *prd*, from the genes that orchestrate the anterior-posterior axis patterning in the *Drosophila* early embryo. For each gene, all available orthologous sequences were collected for the learning and inference of the model. We defined each gene search region as a sequence region 40 Kbp long, which was obtained by extracting 20 Kbp downstream and upstream of the TSS. Ortholog information and chromosome coordinates were acquired from the FlyBase database [44]. The seven motifs and nine genes constitute the *Drosophila* early development dataset used in the experiment. We collected CRMs for these genes from the REDfly database [45], merged the overlapping CRMs, and then took them as a benchmark.

Experimental setup. SMCis was further compared with the other methods on the *Drosophila* early developmental dataset. For Stubb, we chose its multi-species version StubbMS (a Stubb module, which is referred to as Stubb in the following description). StubbMS extends Stubb by applying phylogenetic comparisons between related organisms. These methods were tested on the dataset as follows:

1. For SMCis and BayCis, on for each gene dataset, we took *D. melanogaster* genes as the test set (the REDfly database only collected the annotated CRMs of *D. melanogaster* genes and

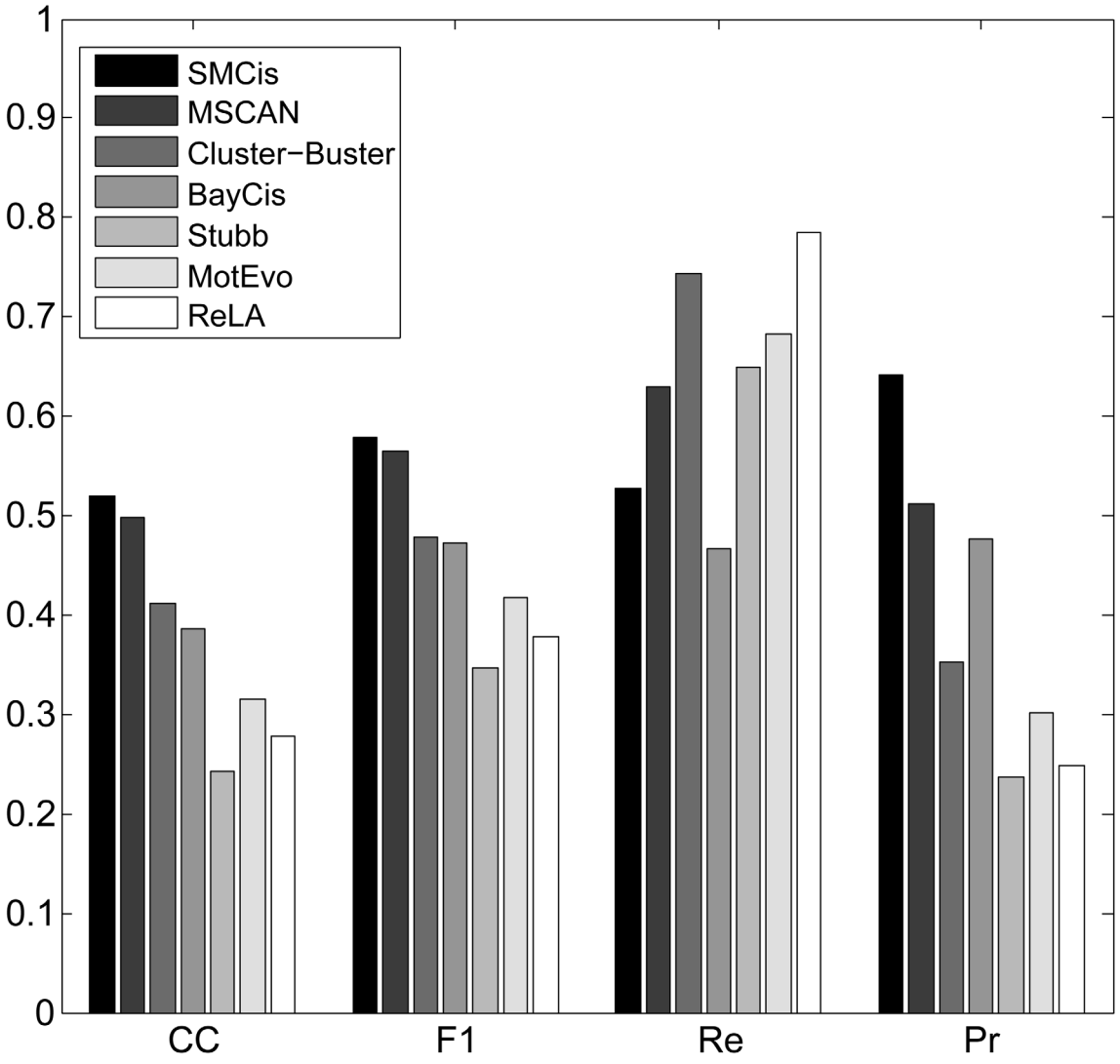

Fig 3. Performance of all methods on the muscle dataset.

doi:10.1371/journal.pone.0162968.g003

had not yet collected CRMs of other *Drosophila* species genes) and orthologous genes from other species of *Drosophila* as the training set.

2. For MotEvo and Cluster-Buster, we tested them on the whole orthologous group but only focused on the predictions on *D. melanogaster* genes.
3. For Stubb, which requires a pairwise sequence alignment, we selected the species used in the original paper (*D. melanogaster* and *D. virilis*) and tested Stubb on the corresponding genes.
4. For MSCAN, which predicts CRMs on a single sequence, we tested it only on the corresponding *D. melanogaster* genes.
5. For ReLA, on each gene dataset, we designated the corresponding *D. melanogaster* gene as the reference sequence, and the other *Drosophila* species genes were compared with the reference sequence.

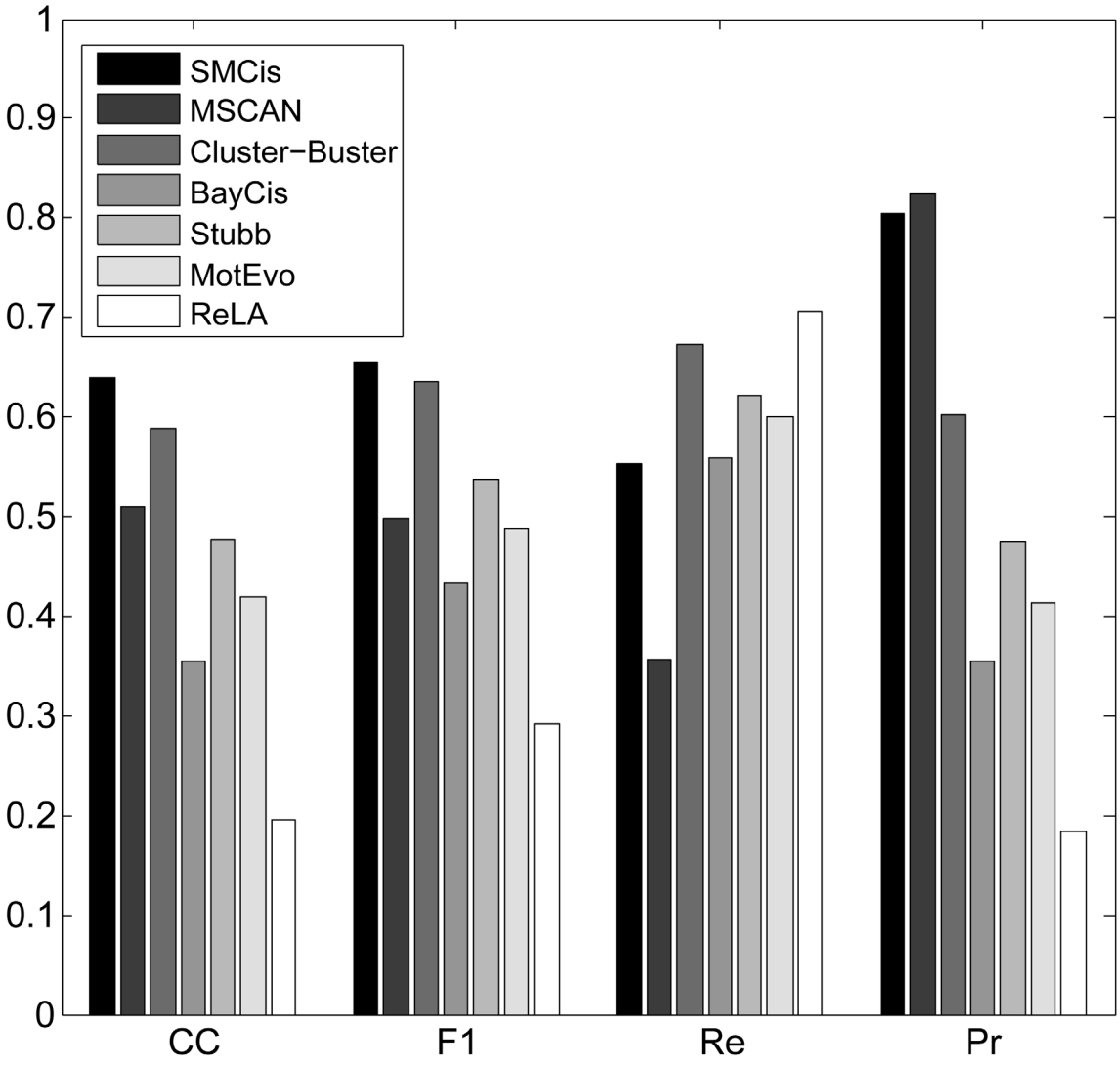

Fig 4. Performance of all methods on the liver dataset.

doi:10.1371/journal.pone.0162968.g004

For the methods that depend on a window size, we set the window size to 800 bp (approximately the average length of CRMs on this dataset) and kept the remaining parameters at their default values. The default settings were maintained for the other methods.

CRM prediction performance. For each method, we counted TP, FP, TN and FN on each gene dataset and then calculated the CC scores of the predicted results on the gene dataset. To evaluate the overall performance of all the methods on the whole dataset, we also calculated all the measures for each method on the whole *Drosophila* early development dataset.

Fig 5 shows CC scores for all the methods on each gene dataset. Overall, SMCis achieved a better prediction performance than the other methods. SMCis had the highest CC scores on four of the nine genes tested: *hb*, *eve*, *h* and *ftz*. It had the second best CC scores on *btd* and *tll* and the third best on *kni* and *Kr*.

Fig 5 shows that the prediction performance of all methods exhibited a relatively consistent tendency on most gene datasets. We speculate that the properties of the CRMs of different genes, such as CRM length and the number of motifs involved, may have a significant impact

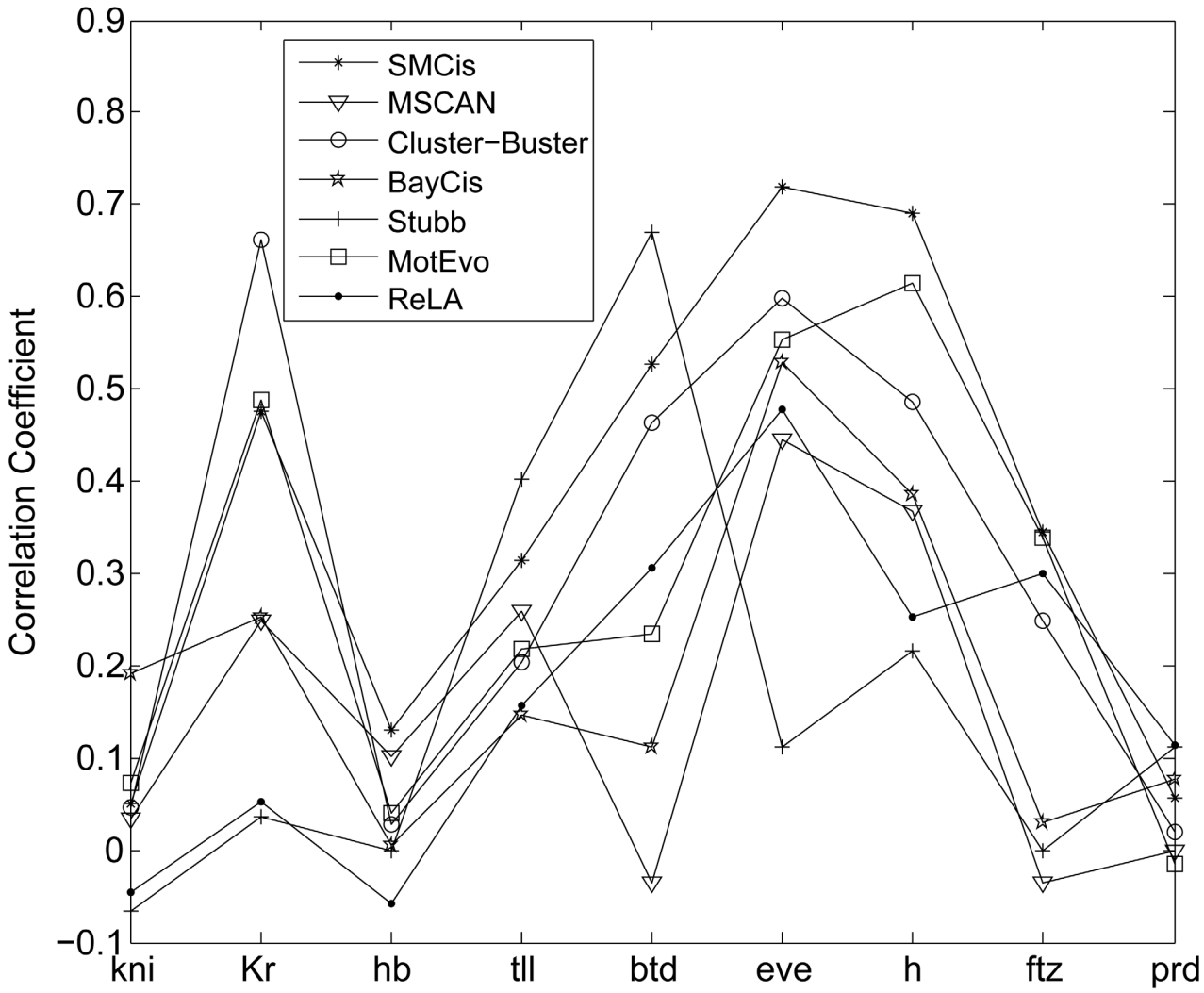

Fig 5. CC scores of all methods, calculated for single genes on the *Drosophila* early development dataset.

doi:10.1371/journal.pone.0162968.g005

on the performance of CRM discovery methods. Generally, the CRMs that are long and contain multiple instances of known motifs are easier to predict. For instance, the *eve* enhancer *eve_stripe*_3+7 contains up to 49 instances of the four motifs, including *kni* and *hb*, and all the methods correctly predicted the CRM. Although it is difficult to design a perfect search strategy that is insensitive to data, most of the probabilistic modeling methods exhibited good adaptability.

Fig 6 shows all of the measures for these methods on the whole dataset. The CC scores and F1-scores of all the methods dropped more sharply on this dataset compared with the results on the muscle and liver datasets. Despite this, the scores were higher than all other compared methods for both CC and F1-score. SMCis still achieved stable prediction performance; it had the highest prediction precision while ensuring a reasonable recall.

Compared with the muscle and liver datasets, the sequences on the *Drosophila* early development dataset are much longer and the CRM lengths are more varied, which may affect the prediction performances of the methods. On this dataset, the methods showed significant differences in prediction performance. ReLA achieved better performance than most of the methods,

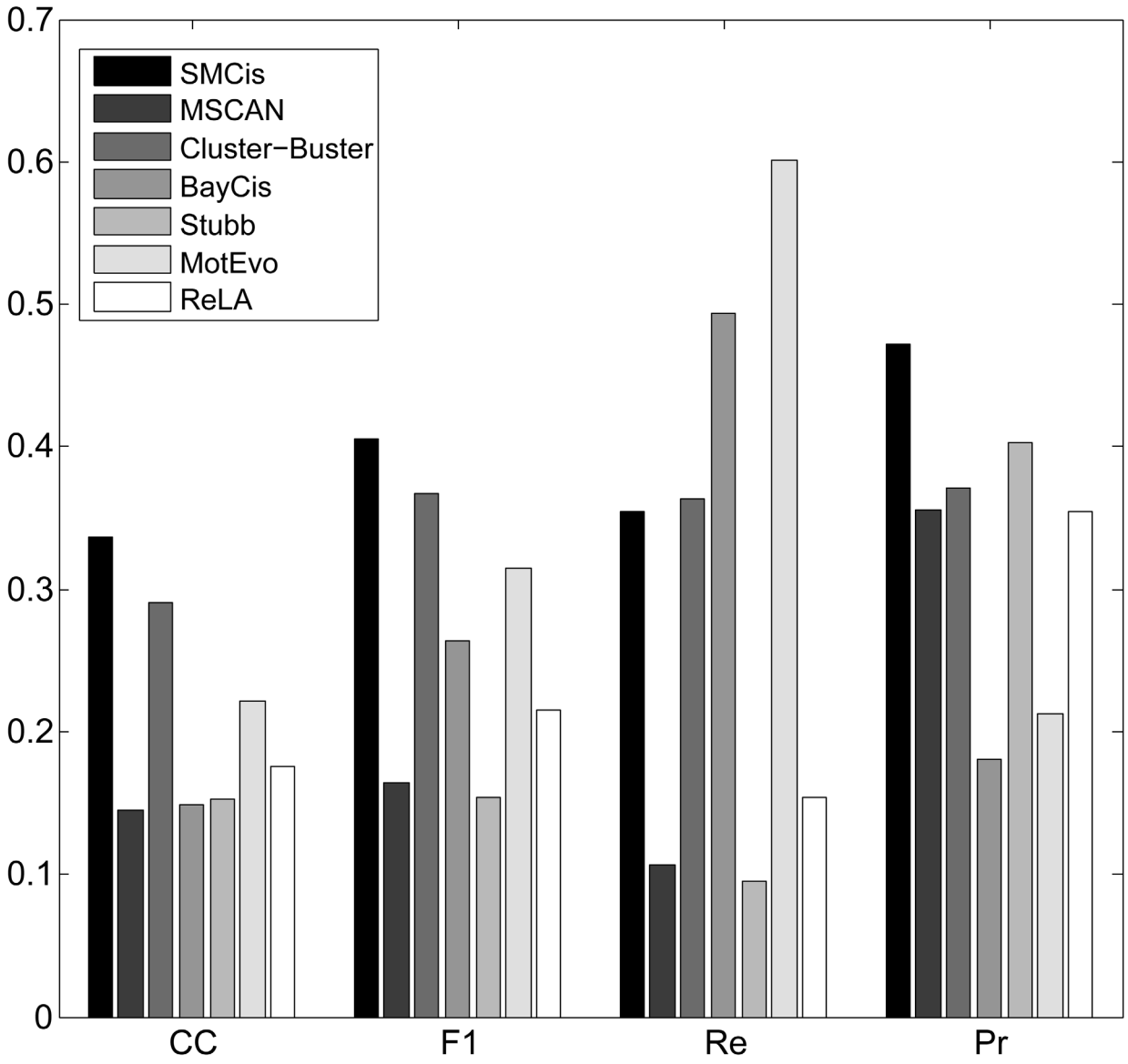

Fig 6. Performance of all methods on the whole *Drosophila* early development dataset.

doi:10.1371/journal.pone.0162968.g006

despite its poor performance on the muscle and liver datasets. In contrast to its good performance on the muscle dataset, MSCAN performed very poorly on this dataset; its CC score was the lowest, and its F1-score was only slightly higher than that of Stubb, which had the lowest CC score. One explanation for this may be that it is difficult for window-based clustering methods to determine reasonable window sizes and score thresholds on this dataset. MotEvo also uses a window-based clustering strategy, but it achieved better performance than MSCAN. MotEvo builds a Bayesian probability model to score motifs clustering within a sliding window, which also illustrates that probabilistic modeling methods may have better adaptability to different types of data. Taking advantage of the evolutionary conservation between species, Stubb had high prediction precision (only second to SMCis), but its recall score was the lowest out of all the methods. SMCis does not directly make use of conservation by aligning sequences, but it implicitly considers the evolutionary conservation between species by characterizing the

conserved regulatory structures of co-regulated or orthologous sequences based on the HSMM. This additional information helps to improve the prediction performance of SMCis.

Discussion and Conclusions

In this paper, we present SMCis, a probabilistic modeling method for predicting CRMs that builds a more powerful CRM discovery model based on an HSMM. In this model, we characterize the regulatory structure of a TRS at a higher level of abstraction (sequence segments rather than nucleotides). Our model views a TRS as a combination of CRMs and inter-module backgrounds and further represents a CRM as a combination of motifs and intra-module backgrounds. Paying more attention to the modeling of the CRM internal structure, we consider not only dependencies between motifs within a CRM but also the distance specificities between the motifs. Compared with other probabilistic modeling methods for CRM discovery, SMCis has the following advantages.

1. The level of abstraction at sequence segments rather than single nucleotides makes the model representations more natural. Representing the regulatory structure of a TRS with a hierarchical organization and explicitly defining CRM states makes the overall architecture of the model clearer.
2. Compared with other methods based on HMMs, our model is more flexible. In the model, we can build an individual model for each type of segment (corresponding to the states of the HSMM model). For example, we can introduce more sophisticated models to capture dependencies between nucleotide sites within a motif. Moreover, our model can use any duration distribution with a specific meaning or a combination of distributions; it is not limited to the implicit geometric distribution in the HMM.

To further improve the prediction performance of CRMs, we will continue working on the following issues in follow-up studies. We will collect more CRM annotations and use more systematic approaches, such as k-fold cross validation, to aid in the selection of model parameters. We will also consider using a Bayesian approach to add *a priori* information and other soft constraints, which will make our method more adaptable to new data because it more easily leads to over-fitting to the training data completely based on the likelihood of the observed data.

Supporting Information

S1 Datasets. The real biological datasets used in our experiments, including the liver and muscle datasets [8] and the *Drosophila* early development dataset from the database REDFly [45].
(ZIP)

S1 File. The test results of all methods.
(ZIP)

S1 Program. The executable program of SMCis.
(ZIP)

S1 Text. Supplement for SMCis and the experiment.
(DOC)

Author Contributions

Conceptualization: HG HH.

Data curation: HG QY.

Formal analysis: HG QY.

Funding acquisition: HH QY.

Investigation: HG.

Methodology: HG HH.

Resources: HH QY.

Software: HG.

Supervision: HH.

Validation: HG QY.

Writing – original draft: HG.

Writing – review & editing: HG HH QY.

References

1. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004; 5: 276–287.
2. Davidson EH, Erwin DH. The regulatory genome: Gene regulatory networks in development and evolution. San Diego (CA): Academic Press; 2006.
3. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet. 2012; 13: 469–483.
4. Zambelli F, Pesole G, Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Brief Bioinform. 2013; 14: 225–237. doi: 10.1093/bib/bbs016 PMID: 22517426
5. Yu Q, Huo H, Zhang Y, Guo H. PairMotif: A new pattern-driven algorithm for planted (l, d) DNA motif search. Plos One. 2012; 7: e48442.
6. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes Nucleic Acids Res. 2006; 34: D108–D110.
7. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010; 38: D105—D110. doi: 10.1093/nar/gkp950 PMID: 19906716
8. Klepper K, Sandve GK, Abul O, Johansen J, Drablos F. Assessment of composite motif discovery methods. BMC Bioinformatics. 2008; 9: 123. doi: 10.1186/1471-2105-9-123 PMID: 18302777
9. Su J, Teichmann SA, Down TA. Assessing computational methods of cis-regulatory module prediction. PLOS Comput Biol. 2010; 6: e1001020. doi: 10.1371/journal.pcbi.1001020 PMID: 21152003
10. Van Loo P, Marynen P. Computational methods for the detection of cis-regulatory modules. Brief Bioinform. 2009; 10: 509–524. doi: 10.1093/bib/bbp025 PMID: 19498042
11. Suryamohan K, Halfon MS. Identifying transcriptional cis-regulatory modules in animal genomes. Wiley Interdiscip Rev Dev Biol. 2015; 4: 59–84. doi: 10.1002/wdev.168 PMID: 25704908
12. Thompson JA, Congdon CB, editors. GAMI-CRM: using de novo motif inference to detect cis-regulatory modules. In: 2014 IEEE Congress on Evolutionary Computation (CEC): Beijing; 2014. pp. 1022–1029.
13. 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: e108065. doi: 10.1371/journal.pone.0108065 PMID: 25268582
14. Nikulova AA, Favorov AV, Sutormin RA, Makeev VJ, Mironov AA. CORECLUST: identification of the conserved CRM grammar together with prediction of gene regulation. Nucleic Acids Res. 2012; 40: e93. doi: 10.1093/nar/gks235 PMID: 22422836
15. Rouault H, Santolini M, Schweisguth F, Hakim V. Imogene: identification of motifs and cis-regulatory modules underlying gene co-regulation. Nucleic Acids Res. 2014; 42: 6128–6145. doi: 10.1093/nar/gku209 PMID: 24682824

16. Alkema WB, Johansson O, Lagergren J, Wasserman WW. MScan: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res. 2004; 32 (Suppl 2): W195–W198. doi: 10.1093/nar/gkh387

17. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. Computational detection of cis -regulatory modules. BioInformatics. 2003; 19 (Suppl 2): ii5–ii14. doi: 10.1093/bioinformatics/btg1052 PMID: 14534164

18. Guns T, Hong S, Marchal K, Nijssen S, editors. Cis-regulatory module detection using constraint programming. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): Hong Kong; 2010.

19. Chan BY, Kibler D. Using hexamers to predict cis-regulatory motifs in drosophila. BMC Bioinformatics. 2005; 6: 262. doi: 10.1186/1471-2105-6-262 PMID: 16253142

20. Kolbe D, Taylor J, Elnitski L, Eswara P, Li J, Miller W, et al. Regulatory potential scores from genome-wide three-Way alignments of human, mouse, and Rat. Genome Res. 2004; 14: 700–707. doi: 10.1101/gr.1976004 PMID: 15060013

21. Gan Y, Guan J, Zhou S, Zhang W. Identifying cis-regulatory elements and modules using conditional random fields. IEEE/ACM Trans Comput Biol Bioinform. 2014; 11: 73–82. doi: 10.1109/TCBB.2013.131 PMID: 26355509

22. Bilmes JA. What HMMs Can Do. IEICE _ Trans Inf Syst. 2006; E89-D: 869–891. doi: 10.1093/ietisy/e89-d.3.869

23. Zhou Q, Wong WH. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci U S A 2004; 101: 12114–12119. doi: 10.1073/pnas.0402858101 PMID: 15297614

24. Frith MC, Li MC, Weng Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res. 2003; 31: 3666–3668. doi: 10.1093/nar/gkg540 PMID: 12824389

25. Sinha S, van Nimwegen E, Siggia ED. A probabilistic method to detect regulatory modules. BioInformatics. 2003; 19 (Suppl 1): i292–i301. doi: 10.1093/bioinformatics/btg1040 PMID: 12855472

26. Lin T-H, Ray P, Sandve GK, Uguroglu S, Xing EP. BayCis: a Bayesian hierarchical HMM for cis-regulatory module decoding in metazoan genomes. In: Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology. Singapore: Springer-Verlag; 2008. p. 66–81.

27. Sinha S, He X. MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLOS Comput Biol. 2007; 3: e216. doi: 10.1371/journal.pcbi.0030216 PMID: 17997594

28. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell. 2006; 124: 47–59.

29. González S, Montserrat-Sentís B, Sánchez F, Puiggròs M, Blanco E, Ramirez A, et al. ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. BioInformatics. 2012; 28: 763–770. doi: 10.1093/bioinformatics/bts024 PMID: 22253291

30. Yu S. Hidden semi-Markov models. Artif Intell. 2010; 174: 215–243. doi: 10.1016/j.artint.2009.11.011

31. Ostendorf M, Digalakis VV, Kimball OA. From HMM's to segment models: a unified view of stochastic modeling for speech recognition. IEEE Transactions on Speech and Audio Proceeding; 1996; 4: 360–378. doi: 10.1109/89.536930

32. Lott SE, Kreitman M, Palsson A, Alekseeva E, Ludwig MZ. Canalization of segmentation and its evolution in drosophila. Proc Natl Acad Sci U S A. 2007; 104: 10926–10931. doi: 10.1073/pnas.0701359104 PMID: 17569783

33. Vardhanabhuti S, Wang J, Hannenhalli S. Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. Nucleic Acids Res. 2007; 35: 3203–3213. doi: 10.1093/nar/gkm201 PMID: 17452354

34. Stormo GD. DNA binding sites: representation and discovery. BioInformatics. 2000; 16: 16–23. doi: 10.1093/bioinformatics/16.1.16 PMID: 10812473

35. Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of Co-expressed genes. Pac Symp Biocomput. 2001: 127–138. PMID: 11262934

36. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press; 1998.

37. Fariselli P, Martelli PL, Casadio R. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. BMC Bioinformatics. 2005; 6 (Suppl 4): S12–S17. doi: 10.1186/1471-2105-6-S4-S12 PMID: 16351738

38. Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. BioInformatics. 2012; 28: 487–494. doi: 10.1093/bioinformatics/btr695 PMID: 22334039

39. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005; 23: 137–144.

40. Shaw WM, Burgin R, Howell P. Performance standards and evaluations in IR test collections: Cluster-based retrieval models Inform Process Manage. 1997; 33: 1–14. doi: 10.1016/S0306-4573(96)00043-X

41. Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol. 1998; 278: 167–181. doi: 10.1006/jmbi.1998.1700 PMID: 9571041

42. Krivan W, Wasserman WW. A predictive model for regulatory sequences directing liver-specific transcription. Genome Res. 2001; 11: 1559–1566. doi: 10.1101/gr.180601 PMID: 11544200

43. Kulakovskiy IV, Makeev VJ. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. Biophysics. 2009; 54: 667–674. doi: 10.1134/S0006350909060013

44. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, et al. FlyBase: enhancing drosophila Gene ontology annotations. Nucleic Acids Res. 2009; 37 (Suppl 1): D555–D559. doi: 10.1093/nar/gkn788

45. Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in drosophila. Nucleic Acids Res. 2011; 39 (Database Issue): D118–D123. doi: 10.1093/nar/gkq999 PMID: 20965965