SCIENTIFIC REPORTS

OPEN

Received: 15 December 2016
Accepted: 2 May 2017
Published online: 12 June 2017

WSMD: weakly-supervised motif discovery in transcription factor ChIP-seq data

Hongbo Zhang, Lin Zhu & De-Shuang Huang

Although discriminative motif discovery (DMD) methods are promising for eliciting motifs from high-throughput experimental data, due to consideration of computational expense, most of existing DMD methods have to choose approximate schemes that greatly restrict the search space, leading to significant loss of predictive accuracy. In this paper, we propose Weakly-Supervised Motif Discovery (WSMD) to discover motifs from ChIP-seq datasets. In contrast to the learning strategies adopted by previous DMD methods, WSMD allows a "global" optimization scheme of the motif parameters in continuous space, thereby reducing the information loss of model representation and improving the quality of resultant motifs. Meanwhile, by exploiting the connection between DMD framework and existing weakly supervised learning (WSL) technologies, we also present highly scalable learning strategies for the proposed method. The experimental results on both real ChIP-seq datasets and synthetic datasets show that WSMD substantially outperforms former DMD methods (including DREME, HOMER, XXmotif, motifRG and DECOD) in terms of predictive accuracy, while also achieving a competitive computational speed.

As the main regulators of transcription process, transcription factors (TFs) can modulate gene expression by binding to special DNA regions, which are known as TF binding sites (TFBS). Previous researches have concluded that TFs are relatively conserved in the long-term evolution, and are inclined to bind to DNA sequences that follow specific patterns, which are commonly called TFBS motifs[1–3]. Recognition of these motifs is fundamental for further understanding of the regulatory mechanisms, and is still a challenging task in computational biology[4, 5].

In the past few decades, due to the rapid development of high-throughput sequencing technology, a variety of experimental methods have been developed to extract TF-DNA binding regions. In particular, ChIP-Seq, which combines chromatin immunoprecipitation with high-throughput sequencing, greatly improves the amount and spatial resolution of generated data, which are conducive to the studies of modeling TF binding *in vivo*[6, 7]. However, ChIP-Seq also brings two challenges for motif discovery methods: (i) The enormous amount of potential TF binding regions yielded from a single experiment requires highly scalable motif discovery (MD) tools[8, 9]; (ii) The large quantities of datasets also increase the possibility of finding multiple enriched sequence features, and most of them may either be false positives or not directly related to the problem of interest, which make it necessary for MD tools to be capable of understanding the nature of motif signals and determining the relevant ones[10–12].

Currently, many algorithms tailored for high-throughput datasets have been proposed[13–15]. Among existing approaches, the discriminative motif discovery (DMD) methods offer a promising strategy for simultaneously addressing the aforementioned two challenges[16–18]. Similar to traditional MD methods, DMDs treat the peak regions of ChIP-seq dataset as foreground sequences, where the motif instances are assumed to be statistically enriched; unlike traditional methods, however, they represent non-binding regions in the foreground using some carefully selected background sequences, then reformulate MD as the extraction of sequence features which could discriminate the foreground sequences from background sequences.

Computationally, the learning of DMDs is more difficult than general discriminant tasks encountered in machine learning: on one hand, the learner not only needs to correctly classify the sequences as foreground or background, but also has to locate the binding sites in foreground examples; on the other hand, the learning

Institute of Machine Learning and Systems Biology, College of Electronics and Information Engineering, Tongji University, Shanghai, 201804, P.R. China. Hongbo Zhang and Lin Zhu contributed equally to this work. Correspondence and requests for materials should be addressed to D.-S.H. (email: dshuang@tongji.edu.cn)

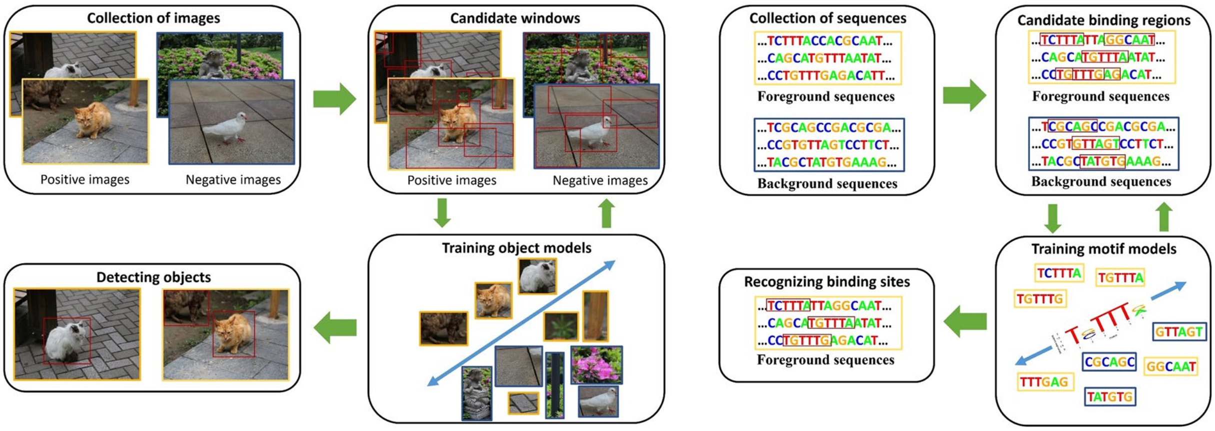

Figure 1. An overview of object detection and discriminative motif discovery. (**a**) Object detection. (**b**) Discriminative motif discovery.

objectives of DMDs are generally nonconvex, nondifferentiable, and even discontinuous, and are thus difficult to optimize. To circumvent such difficulties and improve scalability, current DMD methods typically do not search for motif directly over the complete parameter space, but instead adopt approximate schemes that could sacrifice both accuracy and expressive power. For example, the motifs learned by DREME[19] and motifRG[18] are limited to the discrete IUPAC space, while HOMER[20] and XXmotif[21] choose to refine motifs by only tuning some external parameters.

Meanwhile, in the computer vision community, object detection (OD) is an important and quite challenging application: Given a set of positive images that contain the object of interest, and another set of negative images that don't contain the object, OD aims to classify the test images accurately as positive or negative, and locate the objects of interest in positive images simultaneously. Such kind of problems is also called weakly supervised learning (WSL) in the machine learning community, and various successful techniques have been proposed therein[22–26].

The DMD task shares many features with OD: we know that TF binds to specific sites in ChIP-seq enriched regions, but we don't know exactly where, just as we don't know the exact location of the object of interest in a positive image. In addition, the framework of OD generally consists of four steps (Fig. 1(a)): (1) Collection of training images; (2) Generation of candidate windows that are likely to include the considered object; (3) Learning the object models and refining the candidate windows iteratively with some WSL technologies; (4) Detecting the windows which contain the object with the optimal object models. Similar to OD, DMD also generally consists of four corresponding steps (Fig. 1(b)) which take collection of foreground and background DNA regions as input sequences, then identify the relevant motif by alternating between extracting candidate binding regions and training motif model, finally recognize the real binding sites with the reported motif. Due to these apparent similarities between DMD and OD, and the excellent performance of WSL in OD, it seems natural to integrate the WSL technologies into DMD framework to address the challenges brought by high-throughput ChIP-seq datasets.

As a first attempt to combine the DMD framework with the WSL technologies, in this paper we propose a novel MD approach named WSMD (Weakly-Supervised Motif Discovery) to identify motifs from ChIP-seq datasets. Firstly, we propose to learn the optimal motifs by maximizing classification accuracy (CA), which is one of the most popular metrics in pattern recognition. Similar to other widely used metrics in DMDs, CA is also based on the contingency table; meanwhile, CA has the additional advantage that it can be easily reformulated as a continuous function using convex surrogates. Then, we show that the resulting optimization problem is essentially equivalent to latent support vector machine (LSVM), a widely studied formulation in OD. With this inherent similarity, many WSL learning strategies, which have excellent performances in OD, can be utilized to solve this optimization task. In contrast to the learning strategies adopted by previous DMD methods, WSMD allows "global" optimization of PWMs in continuous space, thereby reducing the information loss of model representation and improving the quality of resultant motifs. Finally, we compare the performance of WSMD with five well-known DMD methods (DREME, XXmotif, HOMER, motifRG and DECOD). The experimental results on 134 real ChIP-seq datasets show that the motifs found by WSMD have better statistical significance, as measured by three commonly used evaluation criteria (AUC under ROC, Fisher's exact test score and the Minimal Hyper-Geometric score). Further in-depth experiments on two groups of synthetic datasets also show that the individual steps of our method have advantages over the five benchmarked methods. The R package of our algorithm is available at https://github.com/hongbozhang0808/WSMD.

Methods

Overview of motif discovery.

Motif discovery is one of the most studied branches in bioinformatics, and the existing literature is vast. Here we give a brief overview of the related works and recommend the interested authors to[13, 17, 27] for detailed reviews. Based on the model representation, motif discovery algorithms can be generally classified as 'word-based' or 'probabilistic-based'[28]. Word-based algorithms model TF binding affinities

	Motif present	Motif absent
Foreground	*TP*	*FN*
Background	*FP*	*TN*

Table 1. Contingency table. Here *TP* and *FP* stand for true and false positives, *TN* and *FN* for true and false negatives, respectively.

with consensus sequence, which represents the predominant nucleotide at each site with IUPAC ambiguity codes[19,29,30]. On the other hand, probabilistic-based algorithms generally perform local searches for the most represented segments in input sequences and represent them as probabilistic models. One of the most commonly used probabilistic models is Position Weight Matrix (PWM)[8,31,32]. Compared with consensus sequence, PWM is a more nuanced representation of motif. It models TF binding affinities with a 4 × *l* matrix which describes binding affinities as probability distributions over DNA alphabet.

Computationally, DMDs typically search an extremely large space of candidate motifs to look for the motif that maximizes an "objective function" which quantifies the degree of discriminability. Naturally, the choice of objective functions would significantly affect the quality of elicited motifs. In practice, the objective functions of DMDs are generally built upon the statistical features of input sequences, and one of frequently considered features is the statistic that describes whether a sequence contains a target motif. Based on this statistic, we can construct a contingency table (Table 1) which tabulates the number of foreground/background sequences that contain or don't contain a motif instance. Many objective functions of DMDs are defined using the contingency table[33], such as the Fisher's exact test score adopted by DREME and SeAMotE[16] and the relative frequency difference used by DECOD[8] and DIPS[34].

Aside from being able to accurately measure the motif discriminability, the learning objectives of DMDs should also allow for efficient and effective optimization, which are especially important when dealing with high-throughput datasets. While the aforementioned contingency-table-based metrics all could reasonably quantify the "discrimination score" of motifs, they are generally nonconvex, nondifferentiable, and even discontinuous, and are thus difficult to optimize numerically. To alleviate such difficulties, a variety of heuristic searching procedures were applied in the previously mentioned DMD methods. For example, DREME and motifRG greedily refine the initial motifs in site-by-site manner, and restrict the search for motifs to the discrete IUPAC space. On the other hand, despite differences in implementation details, HOMER and XXmotif essentially adopt the same strategy: they iteratively tweak the site score threshold so as to maximize the motif enrichment in foreground versus background sequences, then update the PWMs using *k*-mers scored above the selected detection threshold. Although such strategies could provide PMW-based motif representations to avoid the drawbacks of DREME and motifRG, they are still limited since only one indirect motif parameter (i.e., the detection threshold) is optimized.

Here, we model the transcription factor binding specificities with PWMs and adopt the classification accuracy (CA) as the learning objective, which measures the proportion of true predicted results (both *TP* and *TN*) among the total number of sequences. Similar to the aforementioned metrics such as Fisher exact test score, CA is also based on contingency table. Besides, as a widely used statistical measure in pattern classification, CA has the additional advantage that it can be easily reformulated as a continuous function using convex surrogates. Meanwhile, as will be detailed later, the resulting optimization problem is essentially equivalent to latent support vector machine (LSVM), which is widely studied in the OD literatures. Furthermore, by exploiting this connection, efficient learning scheme can be designed to solve the proposed optimization task.

CA-based discriminative object function.

We used unified notations in this paper. Lower case italic letters such as x denote scalars, bold lower case letters represent vectors, such as $\mathbf{x} = (x_i) \in \mathrm{R}_m$. Bold upper case letters denote matrices, such as $\mathbf{X} = (x_{ij}) \in \mathrm{R}_{m \times n}$, and bold upper case italic letters represent vector sets, such as $\boldsymbol{X} = (\mathbf{x}_i)$, $\mathbf{x}_i \in \mathrm{R}_m$. $|\mathbf{x}|$ and $|\boldsymbol{X}|$ return the size of vector $\mathbf{x}$ and $\boldsymbol{X}$ respectively.

We would like to solve the following problem: Given $\boldsymbol{F}$ and $\boldsymbol{B}$ as foreground and background sequence set, respectively, our task is to learn a motif, represented by PWM $\mathbf{P} \in \mathrm{R}_{4 \times l}$, which can distinguish $\boldsymbol{F}$ from $\boldsymbol{B}$ with the maximal CA. Formally, the object function can be written as follows

$$\max_{\mathbf{P}} \left(\frac{TP + TN}{TP + FN + TN + FP} \right), \tag{1}$$

which is equivalent to the minimization of classification error:

$$\max_{\mathbf{P}} \left(\frac{TP + TN}{TP + FN + TN + FP} \right) \Leftrightarrow \min_{\mathbf{P}} \frac{1}{|\boldsymbol{F}| + |\boldsymbol{B}|} (FN + FP). \tag{2}$$

For word-based MD methods, the calculation of (2) is easy since the occurrence of the motif is well defined. For probabilistic-based methods such as ours, an additional site score threshold b is required, and whether a sequence contains a motif is defined based on whether it contains a site scored above the threshold[35]. Formally, *FN* and *FP* in (2) can be defined as follows

$$\begin{cases} FN = \sum \mathrm{sgn}(\mathrm{E}(\mathbf{P}, \mathbf{s}_i) - b < 0), \ \mathbf{s}_i \in \boldsymbol{F} \\ FP = \sum \mathrm{sgn}(\mathrm{E}(\mathbf{P}, \mathbf{s}_i) - b > 0), \ \mathbf{s}_i \in \boldsymbol{B} \end{cases}, \tag{3}$$

where sgn (·) is an indicator function which returns 1 if the argument is true and 0 otherwise. Let $\boldsymbol{S}$ be the set of all possible l-mers in sequences **s**, E (**P**,**s**) returns the maximal binding energy of all elements in $\boldsymbol{S}$ with motif **P**:

$$\mathrm{E}(\mathbf{P},\ \mathbf{s})\ =\ \max_{(}\mathrm{E}_{(}\mathbf{P},\ \mathbf{s}^{\mathrm{sub}}{}_{)},\ \mathbf{s}^{\mathrm{sub}}\ \in\ \boldsymbol{S}_{)}, \tag{4}$$

where the site-level binding energy is defined as

$$\mathrm{E}(\mathbf{P},\ \mathbf{s}^{\mathrm{sub}})\ =\ \sum_{i=1}^{l}\log\left(\mathbf{P}_{(i,(\mathbf{s}^{\mathrm{sub}})_i)}\right). \tag{5}$$

The formulation in (5) can be further simplified: for a given l-length DNA sequence $\mathbf{s}^{\mathrm{sub}}$, we can encode it as a $4l$-length binary feature vector $\mathbf{x}^{\mathrm{sub}}$ by transforming each nucleotide into a 4-dimensional vector using "one-hot encoding":

$$\begin{cases} \text{"A "}\ =\ \{1,\ 0,\ 0,\ 0\},\ \text{"C"}\ =\ \{0,\ 1,\ 0,\ 0\},\\ \text{"G"}\ =\ \{0,\ 0,\ 1,\ 0\},\ \text{"T"}\ =\ \{0,\ 0,\ 0,\ 1\}. \end{cases} \tag{6}$$

We can also convert the logarithm of **P** to a $4l$-length vector **w** by concatenating its column vectors as a single vector. Then the binding energy (5) can be evaluated as the inner production of **w** and $\mathbf{x}^{\mathrm{sub}}$:

$$E\ (\mathbf{P},\ \mathbf{s}^{\mathrm{sub}})\ =\ \mathbf{w}^{T}\mathbf{x}^{\mathrm{sub}}. \tag{7}$$

By applying Equations (3), (4) and (7) to(2), the objective function is rewritten as

$$\min_{\mathbf{w},b}\frac{1}{|\boldsymbol{F}|\ +\ |\boldsymbol{B}|}\sum_{\mathbf{s}\ \in\ \boldsymbol{F}\bigcup\boldsymbol{B}}\mathrm{sgn}\left(y_{\mathbf{s}}\left(\max_{\mathbf{s}^{\mathrm{sub}}\ \in\ \boldsymbol{S}}(\mathbf{w}^{T}\mathbf{x}^{\mathrm{sub}})\ -\ b\right)\ <\ 0\right), \tag{8}$$

where $y_{\mathrm{s}} = 1$ if $\mathbf{s} \in \boldsymbol{F}$ and -1 otherwise.

Numerically, the indicator function sgn(x) is still non-convex and its optimization is NP-hard in general, thus most works in the machine learning literatures replace the indicator function with a convex upper bound that has better computational guarantees[36]. Here we specifically choose the hinge function[37], which can be defined as follows

$$\mathrm{hinge}\ (x)\ =\ \max(0,\ 1\ -\ x). \tag{9}$$

Additionally, we add an L2 norm penalty term $\|\mathbf{w}\|_2^2$ to the objective function to avoid overfitting, and rewrite Equation (8) as

$$\min_{\mathbf{w},b}\|\mathbf{w}\|_2^2\ +\ \frac{c}{|\boldsymbol{F}|\ +\ |\boldsymbol{B}|}\sum_{\mathbf{s}\ \in\ \boldsymbol{F}\ \bigcup\ \boldsymbol{B}}\max\left(1\ -\ y_{\mathbf{s}}\left(\max_{\mathbf{s}^{\mathrm{sub}}\ \in\ \boldsymbol{S}}(\mathbf{w}^{T}\mathbf{x}^{\mathrm{sub}})\ -\ b\right),\ 0\right), \tag{10}$$

where c controls the tradeoff between the classification error rate and the complexity of training model. By introducing a slack variable ξ_i for each sequence $s_i \in \boldsymbol{F} \cup \boldsymbol{B}$, we can transform (10) into a less convoluted form as

$$\begin{aligned} &\min_{\mathbf{w},\xi,b}\|\mathbf{w}\|_2^2\ +\ \frac{c}{|\boldsymbol{F}|\ +\ |\boldsymbol{B}|}\sum\xi_i,\\ &\text{s.t. } y_{\mathbf{s}}\left(\max_{\mathbf{s}^{\mathrm{sub}}\in\boldsymbol{S}}(\mathbf{w}^{T}\mathbf{x}^{\mathrm{sub}})\ -\ b\right)\ +\ \xi_i\ \geq\ 1,\ \xi_i\ \geq\ 0,\ \forall\ \mathbf{s}\ \in\ \boldsymbol{F}\ \bigcup\ \boldsymbol{B}. \end{aligned} \tag{11}$$

The above-mentioned objective function is analogous to the classical latent SVM (LSVM), which is a widely used formulation of WSL in OD[22]. There, instead of DNA sequences, the input foreground and background datasets are labeled images, and **w** could be regarded as a vectorized "template" which describes the object of interest; on the other hand, while in motif learning, the latent variable $\mathbf{s}^{\mathrm{sub}}$ represents the most potential binding site in sequence, in OD it instead represents the sub-region in a picture that most resembles the object to be detected.

As will be detailed later, this LSVM-based formulation of DMD simultaneously provides three advantages over the existing DMD methods reviewed in the previous sections:

- Unlike the discrete searching space adopted in DREME and motifRG, formulation (11) directly learns PWMs in a continuous space, thus it is anticipated that the information loss of model representation is reduced and the quality of resultant motifs is improved.
- Compared with the greedy approach used by DREME and motifRG, and the indirect refinement strategy applied by HOMER and XXmotif, this formulation allows "global" optimization of PWM by taking all the positions of PWM into account at the same time.
- Additionally, there are a wide spectrum of existing literatures in the WSL domain that focus on developing efficient solver for LSVM, which can be adapted to obtain high-quality solutions efficiently for (11).

Weakly-supervised motif discovery (WSMD). Based on the formulation (11), in this section we outline the basic framework of WSMD, which can be divided into 5 stages:

Preprocessing. We split each input sequence and its reverse-complement with an l-length sliding window to obtain a bag of l-mers, where l is the desired motif length. Then we encode each l-mer as a feature vector with (6) to formulate each foreground/background sequence as a positive/negative set of feature vectors.

Seeding. We start by enumerating all exact words (without wildcards) of a given length k ($k = 6$ by default), then calculate the substring minimal distance (SMD)[38] between every pair of word and input sequences. Here, SMD is defined as the minimal Hamming Distance (HD) between a word and the subsequences of an input sequence. The "discrimination score" of each word is then calculated as the probability that it has a smaller SMD in a foreground sequence than in a background. Then, the k-mers with the highest discrimination scores are retained for further optimization.

Refinement. In the refinement step, WSMD takes the top-scored k-mers from the Seeding stage as input, and optimizes them using(11). Although at first appearance, Equation (11) is a complex-formed nonconvex optimization problem under large number of constraints, we can still solve it efficiently by using a simple coordinate-descent-style LSVM optimization strategy[22]. The core idea is to exploit the fact that if the latent variables that mark the bound regions of each input sequences are given, the problem (11) reduces to a convex quadratic programming (QP) which can be solved efficiently using off-the-shelf software such as Mosek and CPLEX. In summary, the PWM is optimized iteratively with two alternating steps: **Update-step:** Update the TF binding regions for both foreground and background sequences using the current PWM; **QP-step:** Solve the resultant QP problem to update PWM. This procedure is repeated until the objective function value converges.

Extension. Generally, the seed length k is smaller than the desired motif length l. In order to extend the refined motif to length l, we firstly add uniform weights at x positions upstream and $l-k-x$ positions downstream of the motif respectively, where x varies between 0 and $l-k$. Such a protocol yields $l - k + 1$ initial PWMs of length l, which are then again optimized using Equation (11), and the one that achieves the minimal objective function value is reported as the final motif.

Masking. In practice, the input dataset often contains multiple motifs, and often each motif explains a subset of the data[39], which requires motif finders to be capable of extracting multiple non-redundant motifs from one dataset. To fulfill this requirement, a commonly adopted strategy in existing DMD methods is to mask the "most potential" binding regions in foreground sequences for the reported motifs, and then repeat search procedure to find other motifs. In WSMD, this can be done by simply removing corresponding feature vectors from the positive set.

Additional details about WSMD are presented in Supplementary Section 1.

Results

In this section, the performance of WSMD is systematically evaluated by comparing it with five widely used DMD algorithms, including DREME, HOMER, XXmotif, motifRG and DECOD. We first conducted an experiment on a comprehensive collection of real ChIP-Seq datasets to show that the performance of WSMD is superior or competitive w.r.t. the other methods. Then, with further experiments on synthetic datasets, we performed in-depth analysis of the refinement and extension strategies of WSMD respectively. At last, we compared the running time of four DMD methods.

Performance comparison on real data. To assess WSMD on real data under different conditions, we collected 134 ChIP-seq datasets from ENCODE (see Supplementary Section 2 for the complete list). For each ChIP-seq dataset, 2000 top ranking peaks were chosen as foreground sequences. On the other hand, the choice of background sequences can significantly affect the results of DMDs. It is widely recognized that the background sequences have to be selected to match the statistical properties of the foreground set[29, 40–42], otherwise the elicited motifs could be biased. To achieve this, a commonly adopted strategy in the literature is to generate artificial background sequences based on the statistical features of foreground sequences, however, previous studies have demonstrated that such sequences are relatively "easy" negatives and could result in underestimation of false-positive rates[43]. Following[18, 43], we instead obtained a background sequence for each peak by randomly choosing a sequence of the same length and lies 0–200 nt from the edge on either up or down strand.

Here, 3-fold cross-validation was used as the performance evaluation scheme. In other words, for each ChIP-seq dataset we took the corresponding set of positive/negative sequences and partitioned them into 3 sets ("folds") of roughly equal size, a PWM was trained on two folds and then evaluated on the other fold. One of the most intuitive approach for assessing DMD methods is to evaluate the similarity between predicted motifs and the reference motifs retrieved from dedicated databases[44–46]. However, this evaluation protocol is problematic since existing motif catalogs are still incomplete and may contain errors. In addition, many TFs could bind DNA cooperatively as heterodimers that alter their respective binding specificity[47], yet motifs of these heterodimers still remain underrepresented in current motif repositories[48]. Similar to[9, 21], we instead adopt the following reference-free metrics:

Firstly, the AUC (the area under the receiver operating characteristic curve), a widely used evaluation criterion in both machine learning and motif discovery[18, 27, 49, 50], was evaluated to gauge and compare discriminating power of different motifs reported by four DMD methods. Figure 2(a) summarizes the average test AUC performance of four tools on 134 datasets. It is evident that WSMD almost always achieves the best discriminability on test datasets in comparison with other methods. Additionally, the paired t-test and Wilcoxon signed-rank test between WSMD and the other methods were conducted to quantify the advantages of WSMD in test AUC

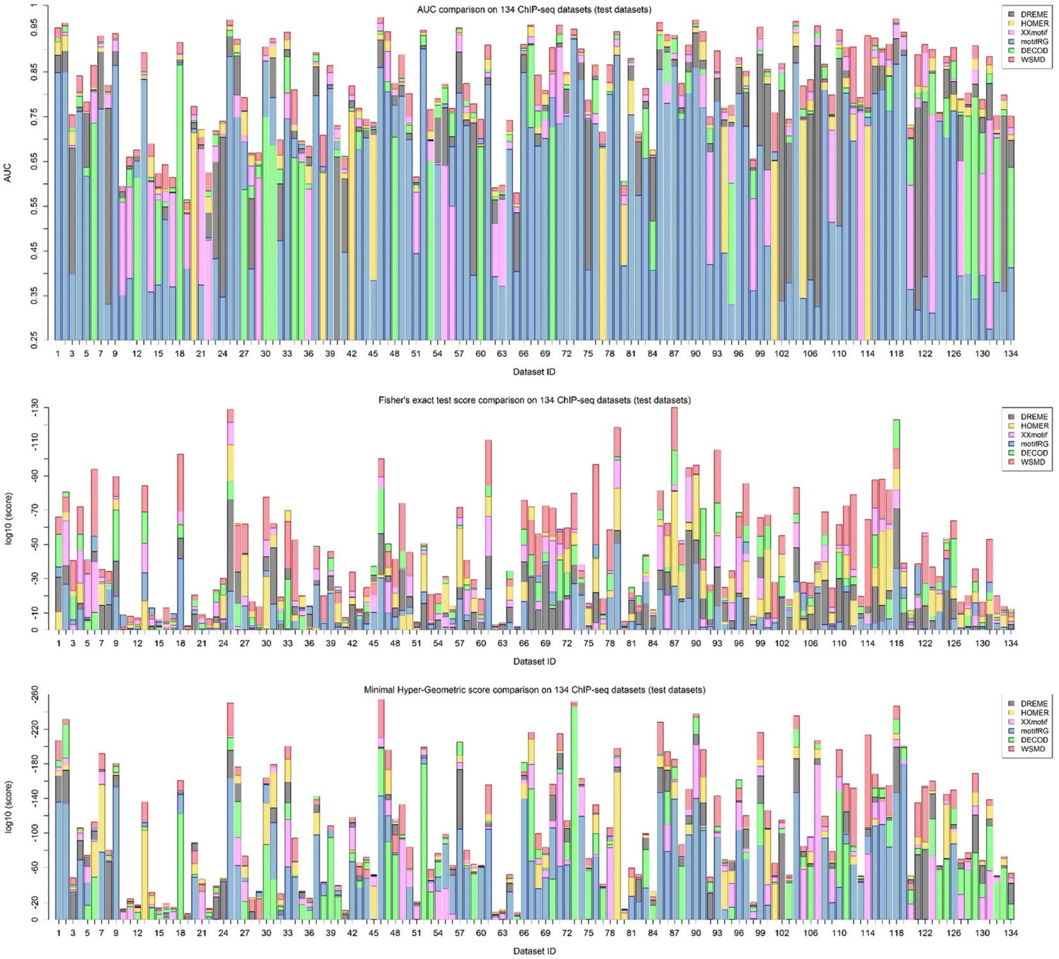

Figure 2. 3-fold cross-validation test performance on three reference-free evaluation criteria over 134 datasets. The performances of six methods on same dataset were plotted on one horizontal bar while differing in colors. In this way, the lines with different colors in one horizontal bar present the performance archived by corresponding tools, and the height of box with different colors can show the performance improvement of corresponding tools compared with the one performing more poorly. (**a**) 3-fold cross-validation test performance on AUC over 134 datasets. (**b**) 3-fold cross-validation test performance on Fisher's Exact Test score over 134 datasets. (**c**) 3-fold cross-validation test performance on Minimal Hyper-Geometric score over 134 datasets.

(Table 2 rows 1–3), and the average training and test AUC on all the 134 datasets for all the algorithms were also reported (Table 3 rows 1–3). As them shows, WSMD have a considerable advantage over other five methods.

Similarly, the Fisher's Exact Test score (Fisher's score) and Minimal Hyper-Geometric score (MHG score) (see Supplementary Section 3.1 & 3.2 for their rigorous mathematical definitions), which are respectively the learning objectives of DREME and HOMER, were used to quantify the relative enrichment of reported motifs in corresponding foreground datasets. Since Fisher's Exact Test needs a pre-defined threshold to count the motif occurrence, following[35], we set the threshold as the top 0.1% quantile of all site-level binding energies in the background sequences. Figure 2(b,c) presents the performance of four methods on Fisher's score and MHG score. As it shows, WSMD performs orders of magnitude better than other methods on most of datasets. For example, the paired t-test and Wilcoxon signed-rank test between WSMD and DREME on Fisher's score both returned very low P-values which highlight the advantages of WSMD in almost all datasets, even though DREME is specifically designed to optimize such a score (Table 2 rows 4–6). The average Fisher's and MHG scores on all the 134 datasets for each algorithms are also reported in Table 3 (rows 4–9).

Performance comparison on synthetic data. Our earlier studies on real ChIP-seq datasets have validated that the performance of WSMD is superior or competitive in comparison with DREME, HOMER, XXmotif, motifRG and DECOD. In this section, we tried to provide further insights on the advantages of our refinement and extension strategies based on artificially created test sequences. In contrast to the real datasets used in the previous sections, the constructive process of synthetic datasets explicitly defines the true binding positions, therefore the quality of elicited motif can be evaluated by directly assessing its accuracy for predicting binding sites on the nucleotide and binding-site level.

On AUC WSMD w.r.t.	DREME	HOMER	XXmotif	motifRG	DECOD
Paired t-test P-value	1.85e-42	3.19e-23	2.61e-24	9.19e-30	1.61e-23
Wilcoxon signed-rank test P-value	5.30e-24	9.08e-23	5.80e-24	4.96e-24	2.07e-23
On Fisher's score WSMD w.r.t.	DREME	HOMER	XXmotif	motifRG	DECOD
Paired t-test P-value	3.79e-34	5.25e-21	7.31e-23	2.26e-29	5.78e-18
Wilcoxon signed-rank test P-value	7.77e-24	9.18e-22	7.14e-23	1.74e-23	3.12e-19
On MHG score WSMD w.r.t.	DREME	HOMER	XXmotif	motifRG	DECOD
Paired t-test P-value	6.97e-18	1.71e-15	6.76e-18	7.81e-28	1.81e-18
Wilcoxon signed-rank test P-value	8.96e-20	1.40e-18	8.60e-20	7.77e-24	4.59e-21

Table 2. The paired t-test and Wilcoxon signed-rank test P-value between WSMD and the other methods on different evaluation criteria.

Average AUC (%)	DREME	HOMER	XXmotif	motifRG	DECOD	WSMD
Training datasets	77.43	79.11	77.13	62.91	77.16	82.80
Test datasets	77.01	78.74	76.85	62.83	76.98	81.86
Average Fisher's score (log10)	DREME	HOMER	XXmotif	motifRG	DECOD	WSMD
Training datasets	−37.42	−54.75	−57.81	−26.61	−59.77	−96.83
Test datasets	−18.50	−27.42	−27.77	−13.06	−29.37	−44.91
Average MHG score (log10)	DREME	HOMER	XXmotif	motifRG	DECOD	WSMD
Training datasets	−185.95	−187.62	−173.93	−112.67	−167.23	−219.87
Test datasets	−91.34	−92.82	−86.58	−55.83	−84.16	−108.08

Table 3. The average performance of each method on 134 ChIP-seq datasets. For each tool, its average performance on both training and test datasets are presented.

	Refinement	Extension
Signal width	8	18
Signal IC	2,4,6,8,10,12,14,16	6,8,10,12,14,16,18,20,22,24,26
Decoy width	8	18
Decoy IC	10	20
Total experiments	80	110

Table 4. Parameters of implanted signal and decoy PWMs.

The setup for the simulation study is generally similar to those from previous works[8,11,33,34]. More specifically, two sets of foreground and background datasets were firstly generated, each set consists of 2000 500 bp-long sequences that were sampled from a uniform distribution on DNA alphabets. Then, a signal PWM and a decoy PWM were respectively generated according to different settings of width and information content (IC). The signal PWM was only inserted into foreground sequences, and the decoy PWM was inserted into foreground and background sequences both. The parameters that varied in two sets of experiments are summarized in Table 4, including the width and IC of signal motifs and decoy motifs, and we generated 10 datasets for each set of parameters (more details about the synthetic datasets construction can be found in Supplementary Section 4). In terms of evaluation metrics, we also follow previous studies[13,51] and adopt the nucleotide-level correlation coefficient (nCC) and the average site-level performance (sASP) (see Supplementary Section 3.3 for their rigorous mathematical definitions). Additionally, in order to keep coherence between real and synthetic datasets, the performances of different methods on AUC, Fisher's score and MHG score are also presented for synthetic datasets (see Supplementary Section 5).

Performance comparison of refinement strategies. As mentioned before, in WSMD, the refinement of seed motifs is formulated as a unified learning problem(11), which could allow for simultaneous tuning of all motif parameters in continuous space. To confirm the advantage of this novel refinement strategy over the approximate schemes used by DREME, HOMER, XXmotif, motifRG and DECOD, we analyzed the performance of the six tested methods for motif refinement by using the same sets of seeds as input.

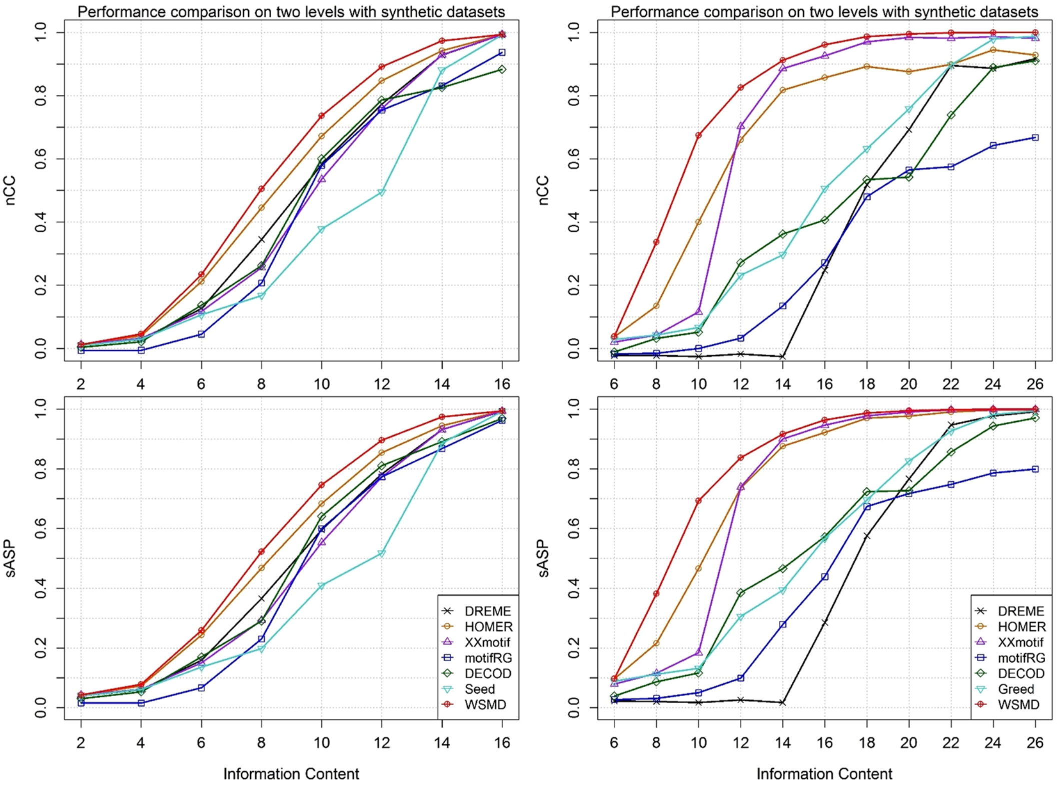

Figure 3. Performance comparison of different refinement and extension strategies. For each IC value, we show the average performance obtained by using each tools over 10 distinct synthetic datasets. (**a**) Performance comparison of different refinement strategies. (**b**) Performance comparison of different extension strategies.

Concretely, to make the comparison fair enough, for each synthetic data, 5 seeds of length 8 were generated using our Seeding procedure, and fed into WSMD, HOMER and XXmotif for further refinement. Then, the best-performing motif reported by each method was used to evaluate its performance. Note that DREME, motifRG and DECOD does not allow optimization of a given motif, thus their performance are measured by running them on corresponding datasets with the width of motifs fixed at 8 and the maximum number of motifs fixed at 5.

Figure 3(a) summarizes the predictive performance of seeds and six DMD tools on datasets of increasing motif IC. The results show that WSMD almost always achieves the best predictive power, followed by HOMER and other tools. This illustrates the particular advantage of our refinement strategy over the ones used by HOMER and XXmotif. The performance comparison on AUC, Fisher's score and MHG score also conform this conclusion (Figures S3–S5).

Performance comparison of extension strategies. Recall that in WSMD, the Extension stage considers all possible scenarios of extending seed motifs to the desired length. In addition, by using the same objective function, this stage is naturally integrated with the Refinement stage. This seems to be a more natural extension scheme compared with the one used by DREME, XXmotif and motifRG, which greedily extends the seed site-by-site until no improvement can be made, and the one used by HOMER and Seeder[38] which simply appends (l−k)/2 positions at both sides symmetrically. In order to validate the benefits of our extension strategy, we conducted extensive experiments on more challenging synthetic datasets described in Table 3. More specifically, WSMD, HOMER and XXmotif were initialized with 5 pre-generated seeds of length 6 and required to recognize a motif of length 18. Besides these three methods, a variant of WSMD, referred to as Greed, was implemented to simulate the greedy extension strategy used by DREME and motifRG. Greed uses the same learning procedure as WSMD, with the exception that the refined motif is extended greedily until the desire length is achieved. Figure 3(b) summarizes the prediction performance of each tool on extension experiments by nCC and sASP. As it shows, WSMD has a significant advantage over any other tested algorithm, which confirms the soundness of formulating extension as a unified optimization task. On the other hand, both DREME and motifRG perform poorly on these datasets, which is expected as they only greedily search for motifs in a discrete space.

Running time comparison. WSMD was implemented with R language. We compared the running time of the six DMD tools on datasets of increasing size. All algorithms were performed on a 3.4 GHz 4-core computer running 64bit-Linux. For each setting of dataset size, the experiment were repeated 10 times and the averaged results were reported. Figure 4 plots the average running time in seconds against the number of sequences. The performance of XXmotif was not presented in the Fig. 4 because its running time is significantly longer, for example it spent 4.413 minutes when dealing with dataset containing 1000 sequences and the running time exceeded half an hour

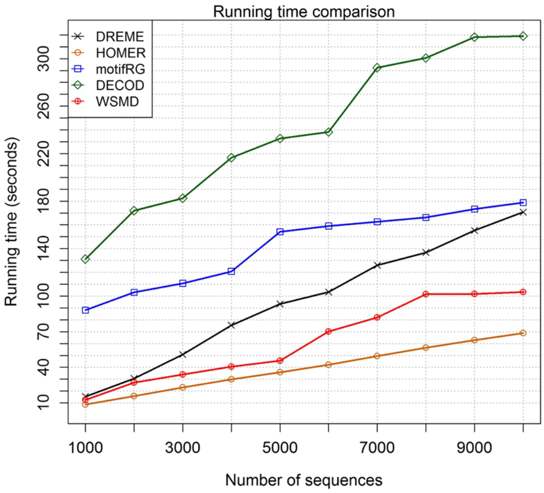

Figure 4. Comparison of running time (seconds) for DREME, HOMER, motifRG, DECOD and WSMD.

when increasing the dataset size to 10000 (Supplementary Table S2). The results show that WSMD is faster than DREME, XXmotif, motifRG and DECOD, while it is still slower than HOMER. However, it has to be emphasized that the search space of HOMER is much smaller and could significantly sacrifice accuracy, as is indicated by previous experiments.

Discussion

In this article, we pointed out the inherent similarities between DMD and OD and thereby proposed a novel method for identifying motifs from ChIP-seq datasets. The core idea of our approach is to learn the optimal motifs by maximizing classification accuracy, which is one of the most popular metrics in pattern classification. Through rigorous mathematical deduction, we proved that the CA-based DMD problem is essentially equivalent to latent support vector machine (LSVM), which is a widely studied formulation of WSL. WSMD outperforms other popular motif finding tools by: (i) Searching for motifs in a continuous space, which could greatly reduce the information loss of model representation; (ii) Formulating DMD problem as an integrated optimization task in which PWM could be refined directly. When tested on real ChIP-seq and synthetic datasets, we showed that the motifs found by WSMD have an excellent performance based on various evaluation criteria. In further experiments on synthetic datasets, where the well-defined 'correct' outcomes are known, WSMD outperforms all benchmarked methods when searching for complicated motifs. Meanwhile, by incorporating ideas from several existing OD literatures, WSMD could also archive competitive speed.

Here, we primarily focus on discussing TFBS motif learning alone. However, it is important to emphasize that the mechanisms by which TFs select their functional binding sites in a cellular environment are highly complex and do not rely purely on recognitions of motifs[48, 52]. In fact, TFs bind only a small fraction of regions that match their corresponding motifs in any given biological condition[48, 53], which means that motif learning alone cannot accurately predict TF binding *in vivo*[43]. For instance, TF binding could be significantly altered by many external factors, such as the DNA shape[54], protein concentration[55], nucleosomes[56, 57], chromatin accessibility[58], cofactors[59, 60], and pioneer TFs[57], etc. Previous researches have also demonstrated that modeling of TF-DNA interactions can benefit greatly from incorporation of these non-sequence features[58, 61–63].

In addition, even if one chooses to only model the sequence specificities of TFs, direct application of motif elicitation tools is still not the ideal choice and does not perform well in practice[43, 64]. This is because the sequence information recognized by a TF is also not limited to the core-binding motif[64]. For example, the lower-order sequence composition (e.g., GC content) of DNA regions that most TFs bind to is often different from that of the rest of the genome[65–67]. In addition, clustered weak binding motifs are often found in the local sequence environment around the core site[64, 68], which is hypothesized to reduce genetic perturbations and help the TFs to reduce the search space of binding sites[64]. Due to these issues, top-performing machine learning methods for sequence-based modeling[43, 69–71] are generally SVMs (e.g., gkmSVM[72] and SeqGL[43]) or neural nets (e.g., DeepBind[69] and Basset[70]) trained using a large set of features that collectively capture the complex properties of bound DNA sequences.

Meanwhile, although motif learning has the limitations mentioned above, it still plays indispensable roles in TF binding study. Firstly, motif information remains to be an integral part of binding models that could incorporate multilayered genomic datasets[61–63]. While dedicated motif databases such as JASPAR[46] and TRANSFAC[44] exist, they are far from complete and a large number of motifs still need to be characterized, such as the motifs of heterodimers. Thus novel computational and experimental technologies are in need to bridge this gap; Secondly, although recent sequence models such as SeqGL and DeepBind show their potential for modeling the overall binding affinity of TFs, they are "black boxes" in nature and difficult to interpret. As a result, *de novo* motif discovery tools and/or motif databases are typically used in the end to analyze outputs of these advanced models[43, 58, 70].

Finally, in spite of its good performance, WSMD still leaves some room for improvement in our previous analysis: (i) The present implementation of WSMD only allows for searching PWMs on ChIP-seq datasets. It is worth studying that whether WSMD could be extended to RNA or protein sequence analysis, as well as to high-order motif models. (ii) Although WSMD can outperform other methods by utilizing a simple coordinate-descent-style learning strategy, its classification accuracy based discriminative object function is still nonconvex and could lead to local minima. There is potential to further improve the performance of WSMD by adapting some more sophisticated strategies discussed in the WSL literatures[23, 73]. Although the above-mentioned future directions are conceptually feasible, they also inevitably lead to more complex learning problems that are computationally expensive in practice. Therefore, we will focus on further exploring the possibilities of applying these ideas without sacrificing scalability.

References

1. Elnitski, L., Jin, V. X., Farnham, P. J. & Jones, S. J. M. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Research* **16**, 1455–1464 (2006).
2. Zhao, Y., Granas, D. & Stormo, G. D. Inferring Binding Energies from Selected Binding Sites. *Plos Computational Biology* **5** (2009).
3. Wang, B., Valentine, S., Raghuraman, S., Plasencia, M. & Zhang, X. Prediction of peptide drift time in ion mobility-mass spectrometry. *BMC Bioinformatics* **10**, S9 (2009).
4. Zhang, Z. Z., Chang, C. W., Hugo, W., Cheung, E. & Sung, W. K. Simultaneously Learning DNA Motif Along with Its Position and Sequence Rank Preferences Through Expectation Maximization Algorithm. *Journal Of Computational Biology* **20**, 237–248 (2013).
5. Ji, Z. *et al*. Systemic modeling myeloma-osteoclast interactions under normoxic/hypoxic condition using a novel computational approach. *Scientific Reports* **5**, 13291 (2014).
6. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
7. Robertson, G. *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
8. Huggins, P. *et al*. DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics* **27**, 2361–2367 (2011).
9. Patel, R. Y. & Stormo, G. D. Discriminative motif optimization based on perceptron training. *Bioinformatics* **30**, 941–948 (2014).
10. Mehdi, A. M., Sehgal, M. S. B., Kobe, B., Bailey, T. L. & Boden, M. DLocalMotif: a discriminative approach for discovering local motifs in protein sequences. *Bioinformatics* **29**, 39–46 (2013).
11. Redhead, E. & Bailey, T. L. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *Bmc Bioinformatics* **8** (2007).
12. Ji, Z. *et al*. Predicting the impact of combined therapies on myeloma cell growth using a hybrid multi-scale agent-based model. *Oncotarget* (2016).
13. Tompa, M. *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* **23**, 137–144 (2005).
14. Mason, M. J., Plath, K. & Zhou, Q. Identification of Context-Dependent Motifs by Contrasting ChIP Binding Data. *Bioinformatics* **26**, 2826–2832 (2010).
15. Ichinose, N., Yada, T. & Gotoh, O. Large-scale motif discovery using DNA Gray code and equiprobable oligomers. *Bioinformatics* **28**, 25–31 (2012).
16. Agostini, F., Cirillo, D., Ponti, R. D. & Tartaglia, G. G. SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. *BMC genomics* **15**, 925 (2014).
17. Lihu, A. & Holban, S. A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Briefings In Bioinformatics* **16**, 964–973 (2015).
18. Yao, Z. Z. *et al*. Discriminative motif analysis of high-throughput dataset. *Bioinformatics* **30**, 775–783 (2014).
19. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
20. Heinz, S. *et al*. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).
21. Hartmann, H., Guthohrlein, E. W., Siebert, M., Luehr, S. & Soding, J. P-value-based regulatory motif discovery using positional weight matrices. *Genome Research* **23**, 181–194 (2013).
22. Forsyth, D. Object Detection with Discriminatively Trained Part-Based Models. *Computer* **47**, 6–7 (2014).
23. Ren, W. Q., Huang, K. Q., Tao, D. C. & Tan, T. N. Weakly Supervised Large Scale Object Localization with Multiple Instance Learning and Bag Splitting. *Ieee Transactions on Pattern Analysis And Machine Intelligence* **38**, 405–416 (2016).
24. Crandall, D. J. & Huttenlocher, D. P. In Computer Vision - E*ccv 2006, Pt 1, Proceedings* Vol. 3951 *Lecture Notes in Computer Science* (eds A. Leonardis, H. Bischof & A. Pinz) 16–29 (2006).
25. Wang, X. F., Huang, D. S. & Xu, H. An efficient local Chan–Vese model for image segmentation. *Pattern Recognition* **43**, 603–618 (2010).
26. Li, B., Zheng, C. H. & Huang, D. S. Locally linear discriminant embedding: An efficient method for face recognition. *Pattern Recognition* **41**, 3813–3821 (2008).
27. Weirauch, M. T. *et al*. Evaluation of methods for modeling transcription-factor sequence specificity. *Nature Biotechnology* **31**, 126–134 (2013).
28. Zambelli, F., Pesole, G. & Pavesi, G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings In Bioinformatics* **14**, 225–237 (2013).
29. Lee, D., Karchin, R. & Beer, M. A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Research* **21**, 2167–2180 (2011).
30. Yu, Q., Huo, H. W., Vitter, J. S., Huan, J. & Nekrich, Y. An Efficient Exact Algorithm for the Motif Stem Search Problem over Large Alphabets. *Ieee-Acm Transactions on Computational Biology And Bioinformatics* **12**, 384–397 (2015).
31. Li, L. P., Liang, Y. & Bass, R. L. GAPWM: a genetic algorithm method for optimizing a position weight matrix. *Bioinformatics* **23**, 1188–1194 (2007).
32. Linhart, C., Halperin, Y. & Shamir, R. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Research* **18**, 1180–1189 (2008).
33. Maaskola, J. & Rajewsky, N. Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Research* **42**, 12995–13011 (2014).
34. Sinha, S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics* **22**, E454–E463 (2006).
35. Tanaka, E., Bailey, T. L. & Keich, U. Improving MEME via a two-tiered significance analysis. *Bioinformatics* **30**, 1965–1973 (2014).
36. Ben-David, S., Eiron, N. & Long, P. M. On the difficulty of approximately maximizing agreements. *Journal Of Computer And System Sciences* **66**, 496–514 (2003).
37. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).
38. Fauteux, F., Blanchette, M. & Strömvik, M. V. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics* **24**, 2303–2307 (2008).

39. Ikebata, H. & Yoshida, R. Repulsive parallel MCMC algorithm for discovering diverse motifs from large sequence sets. *Bioinformatics* **31**, 1561–1568 (2015).
40. Fletez-Brant, C., Lee, D., McCallion, A. S. & Beer, M. A. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Research* **41**, W544–W556 (2013).
41. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nature Genet.* **47**, 955 (2015).
42. Orenstein, Y. & Shamir, R. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research* **42**, 10 (2014).
43. Setty, M. & Leslie, C. S. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. *PLoS Comput Biol* **11**, e1004271 (2015).
44. Matys, V. *et al.* TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic acids research* **31**, 374–378 (2003).
45. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic acids research* **37**, D77–D82 (2009).
46. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* **42**, D142–D147 (2014).
47. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
48. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538–554 (2016).
49. Peng, H. *et al.* Prediction of treatment efficacy for prostate cancer using a mathematical model. *Scientific Reports* **6**, 21599 (2016).
50. Zheng, C. H., Zhang, L., Ng, T. Y., Shiu, S. C. K. & Huang, D. S. Metasample-Based Sparse Representation for Tumor Classification. *IEEE/ACM Transactions on Computational Biology & Bioinformatics* **8**, 1273 (2011).
51. Valen, E., Sandelin, A., Winther, O. & Krogh, A. Discovery of Regulatory Elements is Improved by a Discriminatory Approach. *Plos Computational Biology* **5**, e1000562 (2009).
52. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem.Sci.* **39**, 381–399 (2014).
53. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research* **22**, 1798–1812 (2012).
54. Zhou, T. Y. *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 4654–4659 (2015).
55. Wang, J. & Batmanov, K. BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations. *Nucleic acids research* **43**, e147 (2015).
56. Soufi, A. *et al.* Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell* **161**, 555–568 (2015).
57. Barozzi, I. *et al.* Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. *Mol. Cell* **54**, 844–857 (2014).
58. Zeng, H. Y., Hashimoto, T., Kang, D. D. & Gifford, D. K. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* **32**, 490–496 (2016).
59. Slattery, M. *et al.* Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* **147**, 1270–1282 (2011).
60. Siggers, T., Duyzend, M. H., Reddy, J., Khan, S. & Bulyk, M. L. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7**, 14 (2011).
61. Cirillo, D., Botta-Orfila, T. & Tartaglia, G. G. By the company they keep: interaction networks define the binding ability of transcription factors. *Nucleic Acids Research* **43** (2015).
62. Levo, M. *et al.* Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research* **25**, 1018–1029 (2015).
63. Balwierz, P. J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research* **24**, 869–884 (2014).
64. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research* **25**, 1268–1280 (2015).
65. Song, L. Y. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research* **21**, 1757–1767 (2011).
66. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
67. Fenouil, R. *et al.* CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Research* **22**, 2399–2408 (2012).
68. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy *in vivo*. *Nature Genet.* **47**, 1393 (2015).
69. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**, 831–838 (2015).
70. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* **26**, 990–999 (2016).
71. Ghandi, M. *et al.* gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207 (2016).
72. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *Plos Computational Biology* **10**, 15 (2014).
73. Cinbis, R. G., Verbeek, J. & Schmid, C. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1–1 (2015).

Acknowledgements

This work was supported by the grants of the National Science Foundation of China, Nos 61520106006, 31571364, 61532008, 61672382, 61402334, 61472280, 61472173, 61572447, and 61373098.

Author Contributions

H.B.Z. and L.Z. conceived and designed the method, H.B.Z. conducted the experiments and wrote the main manuscript text. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-03554-7

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.