Briefings in Bioinformatics, 2017, 1–13

doi: 10.1093/bib/bbx026
Software Review

An algorithmic perspective of *de novo cis*-regulatory motif finding based on ChIP-seq data

Bingqiang Liu, Jinyu Yang, Yang Li, Adam McDermaid and Qin Ma

Corresponding author: Qin Ma, Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, Brookings, SD, 57007, USA. Tel.: 1-706-254-4293; E-mail: qin.ma@sdstate.edu

Abstract

Transcription factors are proteins that bind to specific DNA sequences and play important roles in controlling the expression levels of their target genes. Hence, prediction of transcription factor binding sites (TFBSs) provides a solid foundation for inferring gene regulatory mechanisms and building regulatory networks for a genome. Chromatin immunoprecipitation sequencing (ChIP-seq) technology can generate large-scale experimental data for such protein–DNA interactions, providing an unprecedented opportunity to identify TFBSs (a.k.a. *cis*-regulatory motifs). The bottleneck, however, is the lack of robust mathematical models, as well as efficient computational methods for TFBS prediction to make effective use of massive ChIP-seq data sets in the public domain. The purpose of this study is to review existing motif-finding methods for ChIP-seq data from an algorithmic perspective and provide new computational insight into this field. The state-of-the-art methods were shown through summarizing eight representative motif-finding algorithms along with corresponding challenges, and introducing some important relative functions according to specific biological demands, including discriminative motif finding and cofactor motifs analysis. Finally, potential directions and plans for ChIP-seq-based motif-finding tools were showcased in support of future algorithm development.

Key words: cis-regulatory elements; ChIP-seq; motif finding; algorithm

Introduction

Cis-regulatory motifs are usually conserved short DNA sequences, which tend to be 8–20 bp long [1]. Typically, they are transcription factor binding sites (TFBSs) and play significant roles in regulating transcription rates of nearby genes and further control their expression levels. Hence, *de novo* motif prediction and related analyses (e.g. motif scan and comparison) provide a solid foundation for the inference of gene transcriptional regulatory mechanisms in both prokaryotic and eukaryotic organisms [2, 3]. Moreover, these techniques also substantially contribute to some system-level studies, such as regulon modeling and regulatory network construction [2, 4, 5]. With the rapidly growing availability of sequenced genomes and advanced biotechnologies, substantial computational techniques have been carried out to identify motifs from query DNA sequences. Nevertheless, the variations among motifs and their short length make their discovery a challenging problem.

Substantial efforts have been devoted in seeking a reliable and efficient way for motif identification over the past few decades. Since the 1980s, identifying motifs in provided promoters has been one of the most prevalent approaches, and numerous tools have been developed [6–13], such as AlignACE, BioProspector, CONSENSUS, MDscan, MEME, CUBIC, MDscan and BOBRO [10, 11, 13–24]. Some of these tools have been successfully applied to various organisms for regulatory network construction [2, 5]. The underlying mechanism is that the co-regulated genes should exhibit overrepresented common motifs in their promoter regions. Although considerable efforts have been made, one non-negligible limitation is the high false-positive rates in predictions [9, 25–27]. Under the assumption that

Bingqiang Liu is an associate professor in School of Mathematics at Shandong University, Jinan Shandong, P. R. China.
Jinyu Yang is a PhD student in Department of Mathematics and Statistics at South Dakota State University, Brookings, SD, USA.
Yang Li is a PhD student in School of Mathematics at Shandong University, Jinan Shandong, P. R. China.
Adam McDermaid is a PhD student in Department of Mathematics and Statistics at South Dakota State University, Brookings, SD, USA.
Qin Ma is the director of Bioinformatics and Mathematical Biosciences Laboratory and an assistant professor in Department of Agronomy, Horticulture and Plant Science at South Dakota State University, SD, USA. He is also an adjunct assistant professor in BioSNTR.
Submitted: 20 December 2016; **Received (in revised form):** 22 February 2017

the motifs in promoters tend to evolve at a lower rate and therefore be more conserved than nonfunctional surrounding sequences, some phylogenetic footprinting-based algorithms have been developed to reduce the false-positive rate, such as PhyloGibbs, Footprinter, PhyloCon and MicroFootprinter [14, 28–32]. A phylogenetic footprinting strategy was firstly proposed in 1988 [33, 34] and has significantly improved the state-of-the-art performance in this field. However, the majority of programs based on phylogenetic footprinting did not make full use of the phylogenetic relationship of query promoter sequences from various genomes [21]. Owing to this limitation, some promoters from highly divergent species could be included and the motif instances are not conserved enough to carry out motif prediction [35–37]. Most recently, Liu *et al.* [4, 38] developed two computational pipelines aiming to break this bottleneck. Specifically, they extracted phylogenetic relationships from regulatory sequences using a combinatorial framework based on 216 selected representative genomes to refine the orthologous promoter set. It is noteworthy that all the methods mentioned above could be potentially improved by integrating additional experimental data.

The rapid development of high-throughput biotechnologies [39–48] has provided new insight and powerful support for regulatory mechanism analyses and genome-scale regulatory network elucidation. In particular, chromatin immunoprecipitation sequencing(i.e. ChIP-seq followed by high-throughput DNA sequencing) provides massive protein–DNA interactive information and has been successfully applied to genome-wide analyses of transcription factor (TF) binding, histone modification markers and polymerase binding [39, 49]. This technology uses one chosen protein (i.e. TF, histone or DNA polymerase of interest) to bind the whole-genome sequences [50, 51]; then some physical techniques, such as ultrasonic, are used to shear the cross-linked DNA sequences into segments after protein–DNA interaction [39]; finally, these segments will be sequenced into short reads [52, 53]. These reads could be mapped onto their reference genome, if available, using Bowtie [54], BWA [55], etc. Based on the mapping results, the motif-enriched genomic regions could be identified by peak-calling tools [56], such as SPP [57], MACS [58], CisGenome [59], FindPeaks [60], QuEST [61] and PeakRanger [62]. These regions will be the potential binding sites of the chosen protein.

Recent studies suggest that ChIP-seq can be effectively integrated into, and benefit, TFBS discovery tools [57, 61, 63–73]. It provides high-throughput motif signals and allows genome-scale discovery in a cell. More accurate binding regions (peaks) can be derived from ChIP-seq experiments, thus leading to more reliable prediction performance [9]. However, the peaks detected from ChIP-seq data can be up to a few hundred base pairs, while the documented *cis*-regulatory motifs are usually only as long as 8–20 bp [74]. Therefore, an *ab initio* motif discovery method is still indispensable to (i) identify the accurate binding sites from these ChIP-seq peaks, and (ii) build conserved motif profiles for further study in transcriptional regulation. Unfortunately, some widely used motif discovery tools, e.g. MEME and WEEDER [75], cannot be directly used on ChIP-seq peaks, as they are designed for co-regulated promoter sequences with limited size. Recently, some efforts have been made to rectify this problem by modifying traditional motif-finding tools to adapt to the ChIP-seq data [68, 74, 76] or designing specific strategies for ChIP-seq-based motif finding [73, 77]. The computational challenges of these tools include, but not limited to, (i) huge amounts of sequenced ChIP-seq reads can make motif finding a computationally infeasible problem [9]; (ii) failure to identify the motifs associated with cofactors of the ChIP-ed TF [73] or *cis*-regulatory modules [78]; (iii) lack of insight in integration of ChIP-seq data sets from multiple TFs [79]; (iv) the traditional false-positive issue in motif prediction, caused by the noise in ChIP-seq technology [74]; (v) lack of an efficient way to determine the correct lengths of motifs except exhaustively enumerating each length within an interval [18, 80, 81]; and (vi) weak support in elucidation of the mutual interactions among multiple motifs from larger ChIP-ed data sets [82–84], which is important in disease diagnosis through gene regulatory network construction.

So far, there have been a few reviews on motif finding focused on the analysis of ChIP-seq data [77, 85, 86]. Tran *et al.* [77] introduced nine Web tools with numerous details in their usages and applications. For each of them, the authors showcased the requirement of the input format, default parameters, output format, basic characteristics of the tool and a brief introduction of the algorithm. Rather than focusing on individual algorithms or tools, Lihu *et al.* [85] were dedicated to reviewing seven ensemble methods, with input and output requirements, and a brief introduction of the included algorithms. Kulakovskiy *et al.* provided a systematic history of tool development in motif finding before the ChIP-seq technology, and advantages as well as computational challenges of using ChIP-seq in motif finding. Several ChIP-seq-based methods and their applications were reviewed, focusing on their overall stories or workflows [86]. However, it is still under investigation in a detailed algorithmic aspect, as the performance of existing methods is far from satisfactory, based on both efficiency and performance. From the algorithmic point of view, this article presents the main challenges and characteristics of *de novo* motif finding in ChIP-seq peaks, through systematically reviewing eight representative tools (Table 1). Specifically, the review mainly focuses on the motif-finding techniques adopted by these methods, as well as such additional specific functions as discriminative motif finding and cofactor motif identification. Through summarizing the existing limitations and revealing algorithmic potentials, several promising directions for further improvements of ChIP-seq-based motif finding have been proposed, both in methodological aspects and concrete applications.

Motif representation and identification

A motif represents a set of DNA segments with the same length, which are binding sites for the same TF. Each segment of the motif is called an instance, and different instances of the same motif tend to be similar with each other on sequence level (Figure 1A). A representation model of a motif, to demonstrate the similarity of its instances, is expected to accurately capture the characteristics of protein–DNA binding activity of its corresponding TF [90].

The most straightforward model to denote the binding preference of a TF on each position along a motif is the 'consensus' sequence (e.g. AGTCA or AGTCG for the motif in Figure 1A), which is composed of the concatenation of the most frequent nucleotide on each position. It can be seen as the ancestor of the binding sites of the same TF, with an assumption that these sites evolved from it. Although the consensus presents the characteristics of a motif in each position in a simple and clear way, the variations in this motif are absent in this model. The 'degenerate consensus' was proposed to fill this gap, using IUPAC wild cards to replace the exact nucleotides (A, G, C and T). For example, W means both A and T in this position could be recognized by the TF of this motif (Figure 1B) [9].

Table 1. Features of the tools for ChIP-seq motif finding

Programs	MICSA	ChIPMunk	DREME	RSAT peak-motifs	FMotif	SIOMICS	Discrover	RPMCMC
Word-based			Yes	Yes	Yes	Yes	Yes	
Profile-based	Yes	Yes						Yes
Techniques and criteria	MEME+PSSM score	IC-like functions +sequence weight	Word count +Fisher test	RSAT: oligo-analysis, dyad-analysis, local word analysis, MEME and ChIPMunk	Word enumeration +suffix tree+IC -based z-score	IC-like functions +tree methods	HMM +multiple testing corrected P-value	MCMC
Discriminative motif finding			Yes	Yes			Yes	
Cofactors	Multiple motifs		Multiple diverse motifs	Multiple diverse motifs		Motif modules	Multiple diverse motifs	Multiple diverse motifs
Motif comparison				RSAT compare-motifs		STAMP		TOMTOM
Year of release and references	2010 [79]	2010 [69]	2011 [73]	2012 [87]	2014 [74]	2014 [78]	2014 [88]	2015 [89]

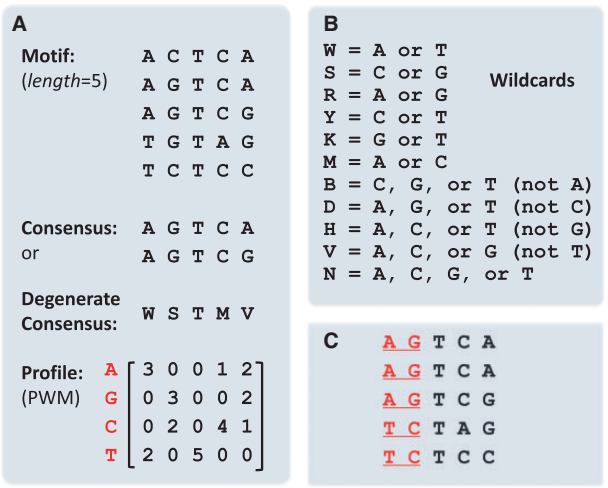

Figure 1. Representation of a motif: (**A**) an example of motif consensus, degenerate consensus and profile; (**B**) a full list of wild cards in the degenerate consensus; (**C**) a different motif but has same profile with motif in (**A**) .

A more accurate and most commonly used model is the 'motif profile'. A profile is built by aligning the available instances of a motif M and counting the frequency of each nucleotide at each position ($f_{i,j}$). These frequencies give rise to a typical matrix representation of a motif profile ($M_f = \{f_{i,j}\}_{4\times l}$, Figure 1A), called a 'position weight matrix' (PWM). An alternative way of constructing the PWM is using the probability distribution to replace frequencies ($M_p = \{p_{i,j}\}_{4\times l}$). Specifically, these frequencies will be divided by the number of binding sites of this motif, and such a representation of the PWM in Figure 1A is shown in Equation (1):

$$M_p = \{p_{i,j}\} = \begin{bmatrix} 0.6 & 0 & 0 & 0.2 & 0.4 \\ 0 & 0.6 & 0 & 0 & 0.4 \\ 0 & 0.4 & 0 & 0.8 & 0.2 \\ 0.4 & 0 & 1 & 0 & 0 \end{bmatrix} \quad (1)$$

Taking the background frequencies of each nucleotides into consideration, the PWM in Equation (1) can be further modified as $M_g = \{g_{i,j}\}$, where $g_{i,j} = \log(f_{i,j}/b_i)$ and $b_i$is the probability of the ith nucleotide of (A, G, C and T) appearing in the background sequences. Based on this version of PWM, we can calculate the 'information content' (IC) to evaluate how conserved this motif is, i.e. Equation (2):

$$I(M) = \sum_{j=1}^{1}\sum_{i=1}^{4} p_{i,j} \log \frac{p_{i,j}}{b_i} \quad (2)$$

In addition, the matrix can also be used to evaluate how a given DNA segment s, with length l, is consistent with motif M by calculating the score:

$$score(s) = \sum_{j=1}^{l}\sum_{i=1}^{4} p_{i,j} \log \frac{p_{i,j}}{b_i} \times \delta_{i,j} \quad (3)$$

where $\delta_{i,j} = 1$, if the jth nucleotide of s is the ith nucleotide of (A, G, C and T), $\delta_{i,j} = 0$ otherwise. A problem of this model is that the probability $f_{i,j}$ could be zero for a small set of binding sites, giving rise to negative infinity in Equations (2) and (3). A common method to avoid this bias is adding a certain value

(pseudocount) for each position of the motif [91]. Through system simulation and analysis, Nishida *et al.* found that the optimal pseudocount value is correlated with the entropy of a motif profile. Specifically, the less conserved motif profiles prefer larger pseudocount value, and 0.8 is suggested in general.

As shown above, multiple approaches for modeling TF's DNA-binding specificity have been developed. A systematic comparison of these approaches can provide substantially valuable information for further motif-finding algorithm design. DREAM5 Consortium organized a competition on motif representation models by applying 26 approaches to *in vitro* protein-binding microarray data [92]. These approaches adopt various strategies, including, but not limited to, *k*-mers model, PWM, hidden Markov models (HMMs) and dinucleotides. The *k*-mers and PWM are the two main strategies, which show similar average performance on multiple data sets. However, they have substantial differences in terms of individual performance on a few data sets, indicating that motif finding is sensitive for model selection. Another interesting observation is that the IC of a motif may not fully represent its accuracy. It is obviously contrary with the basic principle of general motif-finding tools; thus, deeper investigation into this area is still needed to improve the motif representation models.

These motif representation models are still not perfect with a common disadvantage that they ignore the correlation among different positions in a motif. For example, the motif in Figure 1C has the same PWM as the motif in Figure 1a, but apparently, the first two positions in this motif are correlated and dependent to each other. Hence, a high order Markov model is suggested to be integrated into PWM matrix [24]. Meanwhile, it is unsure whether a known PWM can fairly represent the whole population scenario, as the frequencies of nucleotides in each position are calculated only from the known binding sites of a TF. A motif profile built on partially identified bindings sites of a TF may induce bias when it is used to interpret the global binding preference, especially when this profile is used to model the orthologous binding sites from various species. A more fundamental debate is: Do the nucleotides with lower frequencies imply lower binding ability? At the time of this writing, there are still no clear answers to this question, and deeper thought about above concerns will bring potential ways to improve existing representation models of *cis*-regulatory motifs.

Motif signal detection techniques and performance evaluation

The basic computational assumption of motif identification is that they are overrepresented as conserved patterns in given sequences. The scattered instances of a motif are not perfectly identical but similar with each other. Once identified and well aligned, they will show significance in conservation compared with background sequences. Therefore, finding and aligning these instances are the primary issues and remain a challenge in motif finding. ChIP-seq data supply enriched resource for motif finding, yet make this problem more challenging, as the full-size peak sequences from a ChIP-seq experiment are much larger than co-regulated promoters in traditional motif finding.

Motif-finding methods mainly fall into two categories: word-based methods (i.e. consensus-based) and profile-based methods [9, 13]. Word-based methods usually enumerate and compare nucleotides starting from a consensus sequence with a fixed length and a tolerance of mutations. Theoretically, this strategy is able to identify global optimal solutions, but suffers from high computational complexity when it is applied to large-scale input data or when to-be-identified motifs are relatively long and with a large number of mutations [13]. The profile-based methods usually start with some aligned patterns, either randomly chosen [93] or enumerated in a limited subset of input data [93], and refine them based on some criteria on the whole data. These criteria are designed to evaluate the overrepresented significance of aligned profiles from the input sequences. Improvements are mostly conducted in a heuristic way, e.g. neighboring improvement (add or delete patterns to see if the profile goes better, similar to a hill-climbing method) or iterative statistical methods [Gibbs sampling or expectation maximization (EM)]. The profile-based methods usually run faster than word-based methods and have better performance in predicting motifs with complex mutations. However, this kind of methods tend to fail in detection of multiple motifs, especially when the data size is large, as the iterative procedure they adopted often falls into local optimizations, which is difficult to escape [89].

Word-based methods

The dominant model of these methods is the (l, d)-motifs, where l is the width of a motif and d is the maximum number of mutations between a motif instance and the consensus sequence [73, 74, 76, 78, 88, 94]. Intuitively, the goal is to find potential (l, d)-motifs that are significantly presented in the input sequences. These methods can be further divided into two categories: pattern-driven and sample-driven [74]. The former enumerates all 4^l l-mers on character set (A, G, C and T) as consensus sequences to build candidate motifs. The latter only uses real l-mers from input data to generate all possible (l, d)-motifs. Five algorithms specifically designed for motif finding in ChIP-seq peaks, FMotif [74], DREME [73], RSAT peak-motifs [76], SIOMICS [78, 94] and Discrover [88], are reviewed with more details as follows.

DREME [73] is a sample-driven method, which exhaustively searches for exact words and heuristically expands them to words with wild cards. DREME requires a positive data set, i.e. the peaks derived from a ChIP-seq data set, and a negative data set, i.e. different peaks from other ChIP-seq experiments or randomly shuffled version of the positive data. It takes all l-mers in the positive data set and counts the number of their occurrence in the two data sets. The top 100 words that are significantly overrepresented, as determined by Fisher's exact test, are taken as seed motifs. DREME then expands the seed motifs by allowing exactly one additional wild card. The newly generated words are evaluated and only the top 100 most significant words are kept. To save time in the word generation step, DREME estimates the number of sequences matching generated motifs, rather than searching explicitly for them in the data set. However, this program prefers short motifs (with length 4–10 bp), hence is more suitable for monomeric eukaryotic TFs.

The RSAT package [87] provides a pipeline, 'peak-motif', to discover motifs in ChIP-seq peaks [76]. It integrates a set of try-and-test algorithms, using complementary criteria to detect exceptional words, i.e. 'oligo-analysis', 'dyad-analysis', 'position-analysis' and 'local-word-analysis' [95–97]. It is a sample-driven method starting by counting all 6–7 bp long nucleotide occurrences within the input regulatory sequence set and estimating their statistical significances. The expected word frequencies were estimated with a 1st-order or 2nd-order Markov model by default. A lower order can be chosen to improve the sensitivity of frequency estimation in a smaller data set. Finally, the significant words are assembled and converted to PWMs for

further usage. It is worth noting that this algorithm uses calibrated tables, which only takes the uneven nucleotide representation in the noncoding sequences into consideration. This feature, combined with the simplicity of its binomial-based statistical analysis, results in the high efficiency of RSAT 'peak-motifs'. Experimental results also showcased that it is significantly faster than other tools like DREME [73], ChIPMunk [98, 99] and MEME-ChIP [68]. However, the methodological simplicity limits this method to the detection of short and relatively conserved motifs. Longer motifs must be reconstituted by the combination of overlapping short nucleotides.

FMotif [74] is an exhaustive, pattern-driven strategy for finding motifs in ChIP-seq peaks. A similar idea has been used in several traditional motif-finding tools for co-regulated data, e.g. WEEDER [75] and MITRA [100]. One unique advantage of pattern-driven methods is being able to identify planted (l, d)-motifs without prior knowledge of width l. FMotif basically follows the strategy of WEEDER to enumerate all the possible (l, d)-motifs in a depth-first manner and scan all motif occurrences in a suffix tree. Although this strategy has achieved high accuracy in searching for short eukaryotic motifs [25], it still has several drawbacks. For example, it sets a constraint on mutation tolerance d/l, which is not satisfied by many motifs, and the run time will greatly increase in identifying long motifs or motifs with low occurrence frequency. The effect of these drawbacks is getting worse when faced with large-scale ChIP-seq data. To solve these problems, FMotif designed a new suffix tree structure for motif instance searching. It keeps the mismatching information generated when scanning i-mers in suffix tree, and uses them in the searching of $(i+1)$-mers. The new structure dramatically decreases the time complexity of motif instance searching and achieves higher accuracy compared with WEEDER; thus, it makes this enumeration strategy applicable on ChIP-seq data. For example, the run time of FMotif is about 32 min and 4.75 h when identifying planted (16, 5)-motif in the data set with 10 000 and 80 000 nucleotides, correspondingly. However, WEEDER spends >10 h when identifying the same motif in the data set with having 1200 nucleotides [74]. The efficiency can be even improved through a clustering strategy in combining l-mer enumeration (e.g. cisFinder) [101]. One shortcoming of FMotif is higher space complexity because more memory is required to store mismatched information.

The above three methods have made substantial efforts on the motif signal detection procedure. SIOMICS [78, 92] adopts a relatively simpler strategy to build motifs but pays more attention to reassess these motifs based on the significance of a group of motifs (called motif modules). Specifically, SIOMICS takes all l-mers ($l=8$ by default) in given sequences as consensuses, and builds a motif for each of them by searching all segments with at most one mismatch as its motif instances. The motifs are evaluated and ranked by the following Equation (4):

$$\frac{\log(n)}{l}\left[\sum_{j=1}^{l}\sum_{i}^{4} p_{i,j}\log p_{i,j} - \frac{1}{n}\sum_{all-segments}\log(p_0(s))\right], \qquad (4)$$

where n is the number of instances in this motif, p_{ij} is the same as that in Equation (2), and $p_0(s)$ is the probability of generating TFBSs based on background nucleotide frequencies. This score normalizes the effects of motif length and number of instances, thus has no bias in motif interpretation and comparison. It has been successfully used in MDscan, another motif-finding tool for ChIP-chip data [17]. SIOMICS eliminates redundancy in prediction through removing the motif, which has a higher-rank similar motif. The predicted motifs were reassessed as motif modules by a tree strategy, which will be discussed in the 'Find co-factors of TFs from ChIP-seq data' section. The basic assumption is that an individually insignificant motif should be taken as a valuable motif if it is overrepresented when considered as a member of motif module together with the motifs of some other factors.

The word-based methods can be roughly considered as a global optimization strategy, as they enumerate all possible (l, d)-motifs. However, they usually only have a good performance on simulation data but not on real biological data because the (l, d)-motif model is too strict for representing real TFBSs. A high time complexity is another bottleneck of the enumeration strategy. Even though many techniques have been developed to accelerate the searching speed, it is still a tough problem in ChIP-seq motif finding.

Profile-based methods

Profile-based motif-finding methods are usually based on an optimization strategy with a motif profile score, e.g. IC for PWM. Specifically, they try to find a collection of DNA segments, giving rise to a motif profile with the highest score among all the combinations of candidates. However, it is impractical to identify the best profile by exhaustive enumeration, as going through all possible combinations will lead to extremely high time complexity. Therefore, heuristic strategies are usually adopted to narrow down the searching space. First, some primary profiles will be built by randomly selecting one or several DNA segments or by an enumerative search in partially selected input sequences (based on some preliminary knowledge). Then they will be improved, aiming to gain higher profile scores during the iteration process. The improvement process is usually conducted iteratively in a heuristic way [102], e.g. EM [80] and Gibbs sampling [15]. It is noteworthy that high-quality motif profiles can be built through combinatorial optimization strategies, e.g. integrate linear programming [103] and graph-based techniques [7, 10, 23]. With ChIP-seq data, a critical challenge to efficiently handle the exponentially increased number of sequences has been posed to the conventional motif discovery methods.

Several traditional motif-finding tools were updated to adapt to ChIP-seq data, and most of them focused on how to improve the time efficiency faced with the larger data size. For example, MEME-ChIP [68] runs MEME as one of its primary motif detection engines; however, it limits the input data as no >600 sequences. Hybrid Motif Sampler (HMS) combines deterministic identification of motif instances and random sampling strategy together to accelerate its iteration process [72]. STEME [104], which can be considered as an updated version of MEME, introduced a suffix tree structure in the EM process to index the sequences set, which decreases the running time to a different magnitude. MICSA uses MEME on a subset of the highest ChIP-seq peaks to identify motifs [79], and then MEME is applied for several rounds on the sequences in which no motifs have been identified. ChIPMotifs is originally designed for ChIP-chip data, and integrates MEME and WEEDER to predict primary motif candidates, followed by a bootstrap resampling approach to determine optimal cutoff and filter against the false-positive candidates [105].

ChIPMunk [69] is a new iterative algorithm for ChIP-seq data that combines greedy optimization with bootstrapping. The evaluation of motif profiles in ChIPMunk is based on the Kullback Discrete Information Content (KDIC), which is calculated by Equation (5):

$$KDIC = \sum_{j=1}^{l}\sum_{i=1}^{4}\frac{\log f_{ij}! - \log N!}{N} - \sum_{j=1}^{l}\sum_{i=1}^{4} f_{ij}\log b_i, \quad (5)$$

where the variables l, N, f_{ij} and b_i are the same as those defined in Equations (3) and (4). Like CONSENSUS, ChIPMunk uses a greedy strategy to find motif profiles that have high KDIC values. The identified motif profiles are ranked based on PWM scores and then improved by an EM-like iterative process. It is worth noting that the rebuilding of motif profiles integrated the weight scores calculated from the ChIP-seq coverage information. Specifically, ChIPMunk set a score to each sequence as its normalized maximal peak value, and these weights are assigned to each position forming the motif profiles. The experiment on several ChIP-seq data set indicated that ChIPMunk obtained a better performance both in running speed and prediction quality compared with MEME, HMS [72] and SeSiMCMC [106].

RPMCMC [89] is a profile-based method, which combines Gibbs motif samplers and a reversible-jump MCMC method. As a parallel version of Gibbs, RPMCMC fully uses parallel mechanisms to accelerate motif finding. This breakthrough means a lot to current motif finding studies, especially in this big data era driven by high-throughput sequencing technology, e.g. ChIP-seq. Although various motif-finding methods have been proposed before, such as DREME, HEGMA, WEEDER and Gibbs motif sampler [15], they have limited power in properly controlling the trade-off between computation time and motif detection accuracy. Fortunately, RPMCMC makes multiple interacting motif samplers in parallel to ensure an acceptable run time compared with other methods. Most importantly, a repulsive force is applied in RPMCMC to separate different motif samplers close to each other, making further contribution in getting rid of local optima. In this way, RPMCMC has achieved extremely promising performances on synthetic promoter sequence and ENCODE ChIP-seq data sets. Meanwhile, it can also be applied to diverse motif identification, i.e. discriminative motifs, and more details regarding this will be discussed in the next section [89].

The profile-based methods stand on an assumption that each nucleotide in a motif instance participates independently in the TF-binding activity. However, more studies have indicated that the neighboring positions have strong dependent effect in some motifs [107]. Taking this issue into consideration, Mathelier and Wasserman [64] designed a new transcription factor flexible model (TFFM) and developed a prediction system based on this model. It uses a 1st-order HMM to elucidate the dinucleotide dependencies. Specifically, the distribution probabilities of nucleotides in position i are dependent on the nucleotide found at position i-1. The performance evaluation on 96 ChIP-seq data sets indicates that the models considering position dependence outperform the other models on 90 data sets of 96, i.e. 94%, in terms of the area under the curves for the corresponding receiver operating characteristic curves [64]. Although this algorithm is not designed to build motif profiles, the TFFM model can be easily modified to integrate the correlation between positions in *de novo* motif finding.

Although several motif representation models have been developed and new methods have been designed for motif finding on ChIP-seq data, the problem is still far from a sound solution because of the complexity of gene regulation, the large-scale of ChIP-seq data and the low specificity of some binding sites. In the future, new models for motif presentation may be created to replace the consensus and profiles model, and improve the accuracy of motif finding.

Assessment criteria and performance evaluation

The main difference between motif finding before and after the next-generation sequencing technology is the dramatically increased data size [9]. Substantial experiments showcased that the most popular traditional motif-finding tools, e.g. MEME and WEEDER, cannot handle the genome-scale ChIP-seq data containing thousands of peaks. Therefore, almost all ChIP-seq-based motif-finding tools focus on improvement in efficiency in addition to accuracy and conduct comprehensive performance evaluations on synthetic data or real biological data sets. The performance evaluation was mainly carried out based on the comparison with other tools, like MEME, WEEDER and the most-cited ChIP-seq-based tool, DREME. The criteria include coverage on known binding sites (e.g. sensitivity or SN), similarity between predicted motif profiles and the documented ones, specificity (SP), positive predicted value (PPV) and the running time on data sets with different scales.

Here, we summarized the performance evaluation of the eight ChIP-seq-based motif-finding tools as a sketch in Table 2. Overall, no method can outperform others in all situations. These eight tools showed their preferences on different data sizes, different motif lengths and advantages on special context of motif finding. Generally, DREME has a good performance on primary motif finding from large-scale data set, along with other two functions, discriminative motif finding and cofactor motif finding, which will be discussed in next section. However, it prefers short motif identification and only outputs profile matrices without information of corresponding motif instances. In this situation, we have to rely on additional motif scanning tools to identify the concrete TF-binding sites. MICSA, the earliest among the eight tools, takes MEME as the search engine; thus, its speed and performance are highly limited by MEME. RSAT peak-motifs and FMotif handle large data sets with a good prediction accuracy and run faster than DREME. Specifically, FMotif shows good tolerance on noise rate of input sequences. SIOMICS works well on finding motif modules and runs faster than DREME but slower than RSAT peak-motifs. Discrover uses DREME to identify motif seeds, and has a better performance on discriminative motif finding than DREME. RPMCMC shows a great potential on finding diverse motifs in real biological data sets. More details can be found in Table 2, from which users may get a clue to appropriate tool selection in specific context of motif-finding applications.

Development of advanced functions with biological insights

Discriminative motif finding

Traditional motif-finding tools aim to identify a group of conserved motifs in query sequences, which are expected to contain instances of the to-be-detected motifs. Besides these motifs, some other randomly conserved DNA patterns may also exist in these sequences. Distinguishing these false positives from real motifs has always been a big challenge in motif finding. This issue is getting more serious because the high-throughput sequencing techniques bring in more false positives. One solution to this problem is to carry out the discriminative motif finding, which is to find motifs whose occurrence frequencies vary between the query sequence set and several well-defined control sets. The simplest contrast for discriminative motif finding contains a positive set (query sequence set) and a negative set (sequences without binding activity or randomly generated

Table 2. Performance summary of motif-finding tools on ChIP-seq data

Programs	Criteria	Preferences	Speed	Performance and special characteristics
MICSA [79]	Speed; motif coverage	Prefers small data sets	Slow on large data sets	Performance is limited by MEME; performs well on filtering weak peaks
ChIPMunk [69]	Speed; motif coverage	NA	Faster than MEME, slower than DREME; quasi-linear complexity (power 1.27)	More accurate than MEME
DREME [73]	Speed; motif coverage	Handles large ChIP-seq data (thousands of sequences); prefers short motifs	Much faster than WEEDER and MEME; linear time complexity	Performs well at cofactor motif finding and discriminative motif finding; only provides motif profile matrix, no output of motif sites
RSAT peak-motifs [87]	Speed; motif coverage	Handles large ChIP-seq data sets (up to 1 000 000 peaks of 100 bp each)	Faster than DREME, MEME and ChIPMunk	High prediction accuracy
FMotif [74]	Speed; motif coverage; SN and SP on motif sites	Handles large ChIP-seq data; handles long motifs	Fast but requires more memory	High sensitivity and high specificity, predicted accurate motif profiles with high rank (1 or 2) for 12 mESC data; tolerates up to 30% noise sequences
SIOMICS [78]	Speed; motif coverage and specificity on motifs	NA	Slower than RSAT peak-motifs; much faster than DREME, especially on large data sets	Competitive performance with DREME and RSAT in motif finding and cofactor motif finding; competitive performance with DREME and better than RSAT peak-motifs on false-positive control; good motif module prediction
Discrover [88]	Speed; MCC (Matthews correlation coefficient) on nucleotide-level; SN and PPV on binding site level	NA	Comparable with DREME in speed	Better than Bioprospector and MDscan on SN and PPV of prediction; better than DREME on discriminative motif finding
RPMCMC [89]	Speed; SN and PPV	NA	Similar to DREME	Better than WEEDER on SN and PPV of prediction; great potential to mine many reliable diverse motifs

sequences) [88]. The discriminative ideas have been integrated into some traditional motif-finding tools to reduce false positives in prediction. For example, CONSENSUS [102] calculates *P*-values for motif profiles, indicating the probabilities for the same motifs occurring in a hypothetical random sequence set with a similar size. BOBRO [10] calculates such a *P*-value by practically simulating random sequences set in the program as contrast under Poisson distribution. BBR [24] takes coding genomic sequences as background data to reevaluate candidate motifs, and an experiment on *Escherichia coli* genome showed that this strategy reduces the number of false positives. Certainly, the function of discriminative motif finding is not limited to dealing with the false-positive issue. Some biology-driven applications specifically identify motifs overrepresented in a positive set but not in a negative set (called decoy motif problem). Hence, a motif without occurrence differences between two sets is not considerable, no matter how conserved it is [108].

Application of discriminative motif finding is even more important in ChIP-seq data analysis because the data size is usually large and the peaks with or without binding activities naturally compose the positive and negative sets, respectively. In DREME [73], the first step is counting how many positive sequences and negative sequences contain a query word. Then, the *P*-value from Fisher's exact test is calculated for this word. The words with significant *P*-values will be kept as seeds for motif merging. The same evaluation process will be conducted iteratively in generation of motifs by combing similar motif seeds. In application, DREME performed discriminative motif finding by taking Sox2- and Oct4-binding ChIP-seq data in mouse embryonic stem cell (mESC) as positive and negative data set, and found that Oct4 data set has significantly more Oct4-binding sites than the Sox2 data set. The results overturned the previous opinion that Sox2 and Oct4 bind their targets exclusively as a heterodimer [109].

Maaskola and Rajewsky [88] developed an improved HMM-based tool Discrover for discriminative motif finding on large data set. The method consists of three parts: 'seed finding', 'HMM optimization' and 'significance filtering'. The seed finding part is similar to DREME, but uses more objective functions besides the Fisher's exact test *P*-values. Among these objective functions, mutual information of condition and motif occurrence (MICO) and maximum mutual information estimation can be used for discriminative motif evaluation on the contrast with more than two conditions, which extended the application of Discrover. The HMM optimization step starts from a background state trained by the Baum–Welch algorithm. Then, the initial emission probabilities of the motif chains are centered on the seed sequences. The posterior probability of all sequences is evaluated to calculate the expected number of sequences that have motif occurrences. Discrover conducts discriminative

learning by iterative gradient optimization based on the chosen objective functions, which has a linear run time complexity based on the length of the input data. Finally, Discrover refines the final models by a threshold on a corrected MICO-based *P*-value, regardless of if MICO or another objective function was used in previous steps. In addition, Discrover makes extra efforts in detection of multiple motifs, which will be discussed in the next section.

In the application of discriminative motif finding on ChIP-seq data, the various choices of reference data can benefit different biological analysis about gene regulation, and is the main factor affecting the prediction performance [108, 110]. Specifically, the reference data can be (i) randomly generated sequences using a uniform distribution or a Markov process, (ii) the sequences with no binding evidence, (iii) the ChIP-seq peaks for another TF, (iv) the peaks of the same TF but under a different condition, even (v) multiple-level reference data sets based on a detailed grade framework of the signal strength, etc. [94]. This information can reduce false positives and find other special motifs, including decoy motifs, variant motifs for one TF, impoverished motifs and collaborated motifs [108], and thus, can improve the state-of-the-art studies about complex regulation mechanism, e.g. combinatorial regulation and behavior change of TFs among different conditions [110]. It is worth noting that the binding activity of a TF could be affected by epigenetic modifications in a complex fashion, e.g. through local chromatin structure caused by the packaging of DNA, histones and other proteins [111]. Cuellar-Partida *et al.* [111] have used the signal from histone modification ChIP-assay to filter the scanned motifs. The corresponding pitfall raised in discriminative motif finding is that the sequences in a negative data set with no binding evidence may contain potential binding sites. Some algorithms try to mitigate this effect by a special design, which can tolerate a few motif occurrences in negative set [110]. Another phenomenon is that the binding motifs of a TF could be affected by the binding of other TFs. Mason *et al.* [112] proved that Oct4 preferentially binds to different motifs with or without Sox2 bindings within 5000 bp, indicating that the binding motif of Oct4 is context-dependent.

Find cofactors of TFs from ChIP-seq data

In animals and plants, TFs usually regulate gene expression with cooperation of other partner TFs (cofactors) [113]. Growing evidence indicates that cofactors, which interacted directly or indirectly, play an important role in transcription regulation. The binding sites of multiple TFs in a short DNA region often determine the temporal spatial expression pattern of the downstream genes, thus will help to elucidate underlying patterns of combinatorial regulation [114]. The analysis of cofactors has attracted substantial attention in traditional motif finding. For example, BBA [24] in the DMINDA Web server designed a probability model to evaluate the co-occurrences among identified motifs in a given set of regulatory sequences, which can reveal joint regulation relationships by multiple TFs. The cofactor motif finding on ChIP-seq data is usually formulated by detecting multiple, nonredundant motifs in a set of ChIP-seq peak regions [73], or detecting TF patterns whose binding motifs frequently co-occurred on the same set of regulatory regions [113]. Recently, some algorithms have tried to take advantage of the large-scale ChIP-seq data to discover the cofactor motifs.

Researchers want to identify multiple, nonredundant motifs in query ChIP-seq data. The most straightforward way is to scan peak sequences by documented motifs [84, 115]. However, the number of documented motifs is limited compared with the number of TFs, leading to a loss of some real binding sites. Some *de novo* motif-finding methods try to first identify multiple motifs and then build relationships for them with known TFs. One example is the aforementioned DREME [73], which feeds the whole sequences into the corresponding model rather than using only a small portion of the collected ChIP-seq data. It identifies multiple motifs by removing the most statistically significant identified motif derived from Fisher's exact test, and then repeats the search for motifs. DREME achieved remarkable performance in cofactor motifs discovery in mouse cell ChIP-seq data and dramatically outperformed other prevalent algorithms. In particular, DREME discovered >2 times the number of cofactor motifs compared with Amadeus [116] and >10 times compared with Trawler [117]. In addition, the algorithm peak-motifs in the RSAT package also returns additional motifs potentially bound by cofactors, and its application on mouse cell ChIP-seq data successfully detected some cofactor motifs [76].

Rather than simply considering individual motifs separately, SIOMICS [78, 94] models the cofactor motifs as motif modules, i.e. a group of TFs with their TFBSs co-occurring in significantly large number of ChIP-seq peaks. Cofactor motifs may occur only in a small set of peaks and thus can be underrepresented individually in all peak regions. Therefore, evaluation of motif modules simultaneously could make the significance of cofactors stand out from the massive background sequences. As described in previous section, SIOMICS identifies conserved motifs by a word-based strategy based on Equation (4). To identify nonredundant motifs, it removes redundant candidates with lower scores, and the remaining ones are used to identify putative motifs. Instead of considering motif candidates individually, SIOMICS represents them as nodes in a tree, in which the more frequent candidates are closer to the root, while the branch of the tree represents a group of co-occurring motifs. By evaluating each branch based on a Poisson clumping heuristic strategy, SIOMICS identifies statistically significant motif modules and takes motif candidates in them as final putative motifs. The performance comparison on the same ChIP-seq data set used by DREME and peak-motifs indicates that more documented cofactor motifs were identified by SIOMICS rather than DREME and RSAT Peak-motifs, with comparable time efficiency.

The above methods identify cofactors from ChIP-seq data for a single TF and have obtained many valuable results. However, synthetic analysis of diverse ChIP-seq data could provide more information about the co-regulation mechanism of TFs. FCOPs [113] is a method for identifying combinatorial occupancy patterns of multiple TFs from diverse ChIP-seq data. FCOPs integrates two kinds of ChIP-seq data sets as input, from TFs occupancy as well as chromatin modification, and it considers two kinds of uncertainty: How possible a peak from ChIP-seq data represents an actual binding activity and how confident the predicted enhancer sequences are. FCOPs generate probabilities from these uncertainties and use them to evaluate the statistical significance of a given set of TFs co-occurring in a given set of enhancers. Obviously, enumerating all possible subset of enhancers has an exponential complexity, which is inefficient. This method adopts a dynamic programming-based method to calculate the probabilities, which leads to linear complexity with respect to the number of the enhancers, both in time and space. Despite this method not providing any DNA motifs, it can be easily integrated into ChIP-seq motif analysis to identify cofactors.

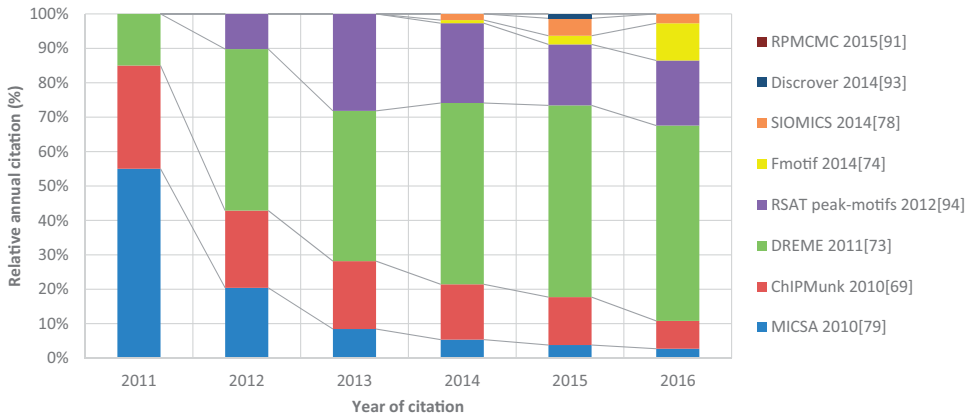

Figure 2. Relative citation of the ChIP-seq motif-finding tools until October 2016. The numbers after a tool name indicate the year of its original release and the corresponding reference. Citation of each tool was extracted from the Bing Academy (http://cn.bing.com/academic/).

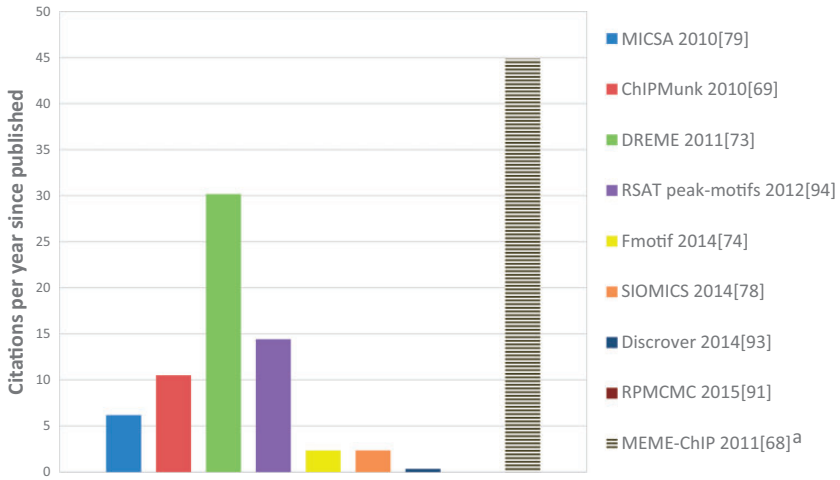

Figure 3. The average annual citation of eight motif-finding tools and MEME-ChIP. [a]MEME-ChIP is a Web server that implants DREME and MEME to detect motifs.

Citation analysis of ChIP-seq motif-finding tools

To analyze the usage of existing ChIP-seq-based motif-finding tools, we collected the citation information of eight tools in Table 1 and a Web server MEME-ChIP, which implants MEME and DREME. The relative citations of the eight tools since 2011 are shown in Figure 2, where we can see that DREME has an overwhelming superiority among these tools, and RSAT peak-motifs and ChIPMunk also have higher relative citation rates. Although the four tools that were developed within the past 3 years (RPMCMC, Discover, SIOMICS and FMotif) introduced some new ideas and obtained better prediction performance, they still fail to get more attention from biologists, and almost all of their citations are from the papers focusing on development of new methods. The main reason is the lack of a user-friendly Web server and a platform with multiple related functions, limiting their usages by people without adequate computational background.

DREME, as one motif-finding engine of MEME-ChIP, has been integrated in the MEME suite, which is the most popular motif-finding and analysis platform. MEME-ChIP obtained more additional citations than DREME (Figure 3), and most of these should be attribute to DREME. RSAT peak-motifs is part of RSAT platform, where a series of modular computer programs is integrated for regulatory signal detection in noncoding sequences. And the relatively high number of citations of ChIPMunk also benefited from its Web service. The methods that only provide source code are difficult to obtain continuous attention. Taking MICSA as an example, it was published in 2010 and was cited frequently within 2 years following publication, but its citation count decreased sharply in recent years. This phenomenon indicated that the user experience of a new algorithm is as important as the quality of algorithm itself.

Discussion

The ChIP-seq peaks provide a foundation for the identification of motifs on a genome scale. In return, the predicted motifs can be used to reevaluate the peaks and even find more potential TF-binding sites that were missed by a ChIP-seq experiment. For example, MICSA [79] takes advantage of *de novo* motif identification to reevaluate the ChIP-seq peaks. In MICSA, only a subset of the highest peaks is fed into MEME for motifs identification, for which the probability of observing a motif by chance (*P*-value) is calculated. Then, the predicted motifs will be used to scan all previously identified peaks, and a score will be assigned to each peak by combining the *P*-values of motifs that occurred on it with its depth. The experimental results

showed that MICSA offered a distinct improvement in TF-binding prediction accuracy with fewer false positives compared with the other 10 approaches. Certainly, searching potential binding sites, using the predicted motifs, is not limited to the peak sequences. It could even be applied on the whole-genome sequences, as some binding activities may not occur under the condition of ChIP experiments, and these binding sites are also valuable for studying the full regulation picture of the interested TFs.

Another issue that should be discussed concerning ChIP-seq motif finding is the difference of its application on various data sources. For example, the binding motifs in eukaryotes and prokaryotes often have their respective characteristics, and the length of the binding regions in eukaryotes is much longer than those in prokaryotes. Therefore, there is a need of special design in motif-finding and analysis techniques.

As with the traditional motif finding, there are some ensemble methods for ChIP-seq-based motif finding [85, 118]. It was always recommended that using several tools in motif finding is a better strategy, as diverse techniques may capture different characteristics of motifs. This idea leads to the development of ensemble motif-finding methods. Most of these methods simply assemble several algorithms together and rank the output of them. For example, W-ChIPMotif [119] includes MEME, WEEDER and MaMF [120], and assesses the output by comparing with a randomized initial input. CompleteMotifs [118] incorporates three *de novo* tools (ChIPMunk, WEEDER and CUDA-MEME) [121], and calculates the *P*-values of predicted motifs by a background random model. However, using more algorithms introduces more false positives, while they have high probability to cover real motifs. Therefore, further analyses are still required for the output of multiple algorithms.

Certainly, the accuracy of prediction and the efficiency of methods are still the most important features in ChIP-seq motif finding, and the application of current methods shows that they are still far from completely satisfactory. More deep thinking and advanced techniques are still needed to improve the intrinsic algorithms. For example, the evolution information from phylogenetic footprinting can be useful as prior knowledge in evaluation of candidate binding sites. Besides, additional functions, such as cofactor analysis, evaluating peaks based on motifs and multiple types of discriminative motif finding are attracting more attention in applications. They should be well integrated in motif-finding methods in future studies. Finally, an integrated Web server or platform is essential for the application of new designed methods, as it can provide more functions about upstream data collection and downstream result analysis.

Most recently, deep learning has demonstrated its unprecedented performance in regulatory genomics, especially in ChIP-seq-based motif analysis [122]. This marked performance is mainly achieved by taking advantage of graphics processing units and the capability to extract complex features of motifs from ChIP-seq data. However, it can be improved further if more information (e.g. DNA shape) can be taken into consideration, rather than only the ChIP-seq peaks sequences.

Key Points

- The algorithmic strategies adopted by current motif-finding tools can be generally divided into two categories: word-based and profile-based, which could be implemented with other models like tree, graph and clustering. The combination of these two strategies is a better choice to break the current bottleneck in prediction accuracy and efficiency.
- The new features of current tools, including, but not limited to, discriminative motif identification (DREAM, RSAT 'peak-motifs' and Discrover), cofactor motif detection (DREME, RSAT 'peak-motifs', SIOMICS, RPMCMC, and Discrover), etc., are essential for application of ChIP-seq data analysis and are encouraged to be integrated into newly designed algorithms.
- Existing tools still have a much room for improvement with respect to prediction accuracy and efficiency; the contradictory relationship between the time efficiency and space complexity as well as between prediction accuracy and application universality is generally present in these tools.
- An integrated Web server or platform, providing various analysis functions in data collection and the result analysis, is essential for the spread of new designed methods in their applications.
- In future studies, more mathematical models, e.g. graph model, statistics and combinatorial optimization, could be developed and integrated to design new motif signal detection techniques and new evaluation criteria. In addition, more information, e.g. the dependency between nucleotides and the height of peaks from ChIP-seq data, should be considered to improve the prediction accuracy.

Funding

The State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University; the National Nature Science Foundation of China (grant numbers 61303084 and 61432010 to B.L.); and the Young Scholars Program of Shandong University (grant numbers YSPSDU 2015WLJH19 to B.L.).

References

1. D'Haeseleer P. What are DNA sequence motifs? *Nat Biotechnol* 2006;**24**:423–5.
2. Brohee S, Janky R, Abdel-Sater F, *et al*. Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic Acids Res* 2011;**39**:6340–58.
3. Davidson E, Levin M. Gene regulatory networks. *Natl Acad Sci USA* 2005;**102**:4935.
4. Liu B, Zhou C, Li G, *et al*. Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses. *Sci Rep* 2016;**6**:23030.
5. Baumbach J. On the power and limits of evolutionary conservation–unraveling bacterial gene regulatory networks. *Nucleic Acids Res* 2010;**38**:7877–84.
6. Lawrence CE, Altschul SF, Boguski MS, *et al*. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993;**262**:208–14.
7. Pevzner PA, Sze SH. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol* 2000;**8**:269–78.
8. Nakaki R, Kang J, Tateno M. A novel ab initio identification system of transcriptional regulation motifs in genome DNA sequences based on direct comparison scheme of signal/noise distributions. *Nucleic Acids Res* 2012;**40**:8835–48.

9. 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**:225–37.
10. Li G, Liu B, Ma Q, *et al*. A new framework for identifying cis-regulatory motifs in prokaryotes. *Nucleic Acids Res* 2011;**39**:e42.
11. Chen X, Guo L, Fan Z, *et al*. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics* 2008;**24**:1121–8.
12. Sinha S. PhyME: a software tool for finding motifs in sets of orthologous sequences. *Methods Mol Biol* 2007;**395**:309–18.
13. Das MK, Dai HK. A survey of DNA motif finding algorithms. *BMC Bioinformatics* 2007;**8(Suppl 7)**:S21.
14. Wang T, Stormo GD. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 2003;**19**:2369–80.
15. Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001:127–38.
16. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999;**15**:563–77.
17. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* 2002;**20**:835–9.
18. Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**:W202–8.
19. Olman V, Xu D, Xu Y. CUBIC: identification of regulatory binding sites through data clustering. *J Bioinform Comput Biol* 2003;**1**:21–40.
20. Li X, Wong WH. Sampling motifs on phylogenetic trees. *Proc Natl Acad Sci USA* 2005;**102**:9481–6.
21. Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* 2002;**12**:739–48.
22. Blanchette M, Tompa M. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res* 2003;**31**:3840–2.
23. Li G, Liu B, Xu Y. Accurate recognition of cis-regulatory motifs with the correct lengths in prokaryotic genomes. *Nucleic Acids Res* 2010;**38**:e12.
24. Ma Q, Liu B, Zhou C, *et al*. An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. *Bioinformatics* 2013;**29**:2261–8.
25. Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–44.
26. McCue LA, Thompson W, Carmack CS, *et al*. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* 2002;**12**:1523–32.
27. Simcha D, Price ND, Geman D. The limits of de novo DNA motif discovery. *PLoS One* 2012;**7**:e47836.
28. Siddharthan R, Siggia ED, van Nimwegen E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 2005;**1**:e67.
29. Blanchette M, Schwikowski B, Tompa M. Algorithms for phylogenetic footprinting. *J Comput Biol* 2002;**9**:211–23.
30. Neph S, Tompa M. MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic Acids Res* 2006;**34**:W366–8.
31. Carmack CS, McCue LA, Newberg LA, *et al*. PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol Biol* 2007;**2**:1.
32. Zhang S, Xu M, Li S, *et al*. Genome-wide *de novo* prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res* 2009;**37**:e72.
33. Katara P, Grover A, Sharma V. Phylogenetic footprinting: a boost for microbial regulatory genomics. *Protoplasma* 2012;**249**:901–7.
34. Tagle DA, Koop BF, Goodman M, *et al*. Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* 1988;**203**:439–55.
35. Borneman AR, Gianoulis TA, Zhang ZD, *et al*. Divergence of transcription factor binding sites across related yeast species. *Science* 2007;**317**:815–9.
36. Odom DT, Dowell RD, Jacobsen ES, *et al*. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* 2007;**39**:730–2.
37. Boyle AP, Araya CL, Brdlik C, *et al*. Comparative analysis of regulatory information and circuits across distant species. *Nature* 2014;**512**:453–6.
38. Bingqiang Liu CZ, Zhang H, Guojun L, *et al*. An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes. *BMC Genomics* 2016;**17**:578.
39. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**:669–80.
40. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010;**2**:5384.
41. Wang Z, Gerstein M, Snyder, *et al*. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2008;**10**:57–63.
42. Tsankov AM, Gu H, Akopian V, *et al*. Transcription factor binding dynamics during human ES cell differentiation. *Nature* 2015;**518**:344–9.
43. Wu F, Olson BG, Yao J. DamID-seq: genome-wide mapping of protein-DNA interactions by high throughput sequencing of adenine-methylated DNA fragments. *J Vis Exp* 2016;**107**:e53620.
44. Maragkakis M, Alexiou P, Nakaya T, *et al*. CLIPSeqTools-a novel bioinformatics CLIP-seq analysis suite. *RNA* 2015;**22**:1–9.
45. Hafner M, Landthaler M, Burger L, *et al*. PAR-CliP—a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp* 2010;**41**:e2034.
46. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* 2014;**15**:205–13.
47. Giresi PG, Kim J, Mcdaniell RM, *et al*. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 2007;**17**:877–85.
48. Nutiu R, Friedman RC, Luo S, *et al*. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat Biotechnol* 2011;**29**:659–64.
49. Collas P, Dahl JA. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front Biosci* 2008;**13**:929–43.
50. Kimura H, Sato Y. *DNA Replication and Histone Modification*. Japan: Springer, 2016.
51. Suganuma T, Workman JL. Histone modification as a reflection of metabolism. *Cell Cycle* 2016 **15**:481–2.
52. Qu H, Fang X. A brief review on the human encyclopedia of DNA elements (ENCODE) project. *Genomics Proteomics Bioinformatics* 2013;**11**:135–41.

53. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 2004;**306**:636–40.
54. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
55. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010;**26**:589–95.
56. Szalkowski AM, Schmid CD. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief Bioinform* 2011;**12**:626–33, 628.
57. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 2008;**26**:1351–9.
58. Zhang Y, Liu T, Meyer CA, *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137.
59. Jiang H, Wang F, Dyer NP, *et al*. CisGenome browser: a flexible tool for genomic data visualization. *Bioinformatics* 2010;**26**:1781–2.
60. Fejes AP, Robertson G, Bilenky M, *et al*. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 2008;**24**:1729–30.
61. Valouev A, Johnson DS, Sundquist A, *et al*. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 2008;**5**:829–34.
62. Xin F, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics* 2011;**12**:139.
63. Kuan PF, Chung D, Pan G, *et al*. A statistical framework for the analysis of ChIP-Seq data. *J Am Stat Assoc* 2011;**106**:891–903.
64. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. *PLoS Comput Biol* 2013;**9**:e1003214.
65. Cheng C, Min R, Gerstein M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* 2011;**27**:3221–7.
66. Wu S, Wang J, Zhao W, *et al*. ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data. *Theor Biol Med Model* 2010;**7**(1):18.
67. van Heeringen SJ, Veenstra GJC. GimmeMotifs: a *de novo* motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* 2011;**27**:270–1.
68. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 2011;**27**:1696–7.
69. Kulakovskiy IV, Boeva VA, Favorov AV, *et al*. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 2010;**26**:2622–3.
70. Jothi R, Cuddapah S, Barski A, *et al*. Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008;**36**:5221–31.
71. Mercier E, Droit A, Li L, *et al*. An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. *PLoS One* 2011;**6**:e16432.
72. Hu M, Yu J, Taylor JM, *et al*. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res* 2010;**38**:2154–67.
73. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 2011;**27**:1653–9.
74. Jia C, Carson MB, Wang Y, *et al*. A new exhaustive method and strategy for finding motifs in ChIP-enriched regions. *PLoS One* 2014;**9**:e86044.
75. Pavesi G, Mereghetti P, Mauri G, *et al*. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* 2004;**32**:W199–203.
76. Thomas-Chollier M, Herrmann C, Defrance M, *et al*. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* 2012;**40**:e31.
77. Tran NT, Huang CH. A survey of motif finding web tools for detecting binding site motifs in ChIP-Seq data. *Biol Direct* 2014;**9**:4.
78. Jun Ding HH, Xiaoman L. SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data. *Nucleic Acids Res* 2014;**42**:1645–35.
79. Boeva V, Surdez D, Guillon N, *et al*. *De novo* motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res* 2010;**38**:e126.
80. Bailey TL, Williams N, Misleh C, *et al*. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 2006;**34**:369–73.
81. Hartmann H, Guth?hrlein EW, Siebert M, *et al*. P-value-based regulatory motif discovery using positional weight matrices. *Genome Res* 2013;**23**:181–94.
82. Niu M. *De novo* prediction of cis-regulatory modules in eukaryotic organisms, Dissertations & Theses—Gradworks, The university of North Carolina at Charlotte, 2014.
83. Bolouri H, Ruzzo WL. Integration of 198 ChIP-seq datasets reveals human cis-regulatory regions. *J Comput Biol* 2012;**19**:989–97.
84. Sun H, Guns T, Fierro AC, *et al*. Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection. *Nucleic Acids Res* 2012;**40**:e90.
85. Lihu A, Holban S. A review of ensemble methods for de novo motif discovery in ChIP-Seq data, *Brief Bioinform* 2016;**17**:731.
86. Kulakovskiy IV, Makeev VJ. Motif discovery and motif finding in ChIP-Seq data. *Genome Analysis: Current Procedures and Applications*, Caister Academic Press, Norfolk, UK. 2014, 83.
87. Medina-Rivera A, Defrance M, Sand O, *et al*. RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res* 2015;**43**:W50–6.
88. Maaskola J, Rajewsky N. Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res* 2014;**42**:12995–3011.
89. Ikebata H, Yoshida R. Repulsive parallel MCMC algorithm for discovering diverse motifs from large sequence sets. *Bioinformatics* 2015;**31**:1561–8.
90. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics* 2000;**16**:16–23.
91. Nishida K, Frith MC, Nakai K. Pseudocounts for transcription factor binding sites. *Nucleic Acids Res* 2009;**37**:939–44.
92. Weirauch MT, Cote A, Norel R, *et al*. Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* 2013;**31**:126–34.
93. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994;**2**:28–36.
94. Ding J, Dhillon V, Li X, *et al*. Systematic discovery of cofactor motifs from ChIP-seq data by SIOMICS. *Methods* 2015;**79–80**:47–51.
95. van Helden J, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 1998;**281**:827–42.

96. van Helden J, Rios AF, Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 2000;**28**:1808–18.
97. van Helden J, del Olmo M, Perez-Ortin JE. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res* 2000;**28**:1000–10.
98. Kulakovskiy I, Levitsky V, Oshchepkov D, *et al*. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J Bioinform Comput Biol* 2013;**11**:1340004.
99. Levitsky VG, Kulakovskiy IV, Ershov NI, *et al*. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genomics* 2014;**15**:80.
100. Eskin E, Pevzner PA. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 2002;**18 Suppl 1**:S354–63.
101. Sharov AA, Ko MS. Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Res* 2009;**16**:261–73.
102. Hertz GZ, Hartzell GW III, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* 1990;**6**:81–92.
103. Kingsford C, Zaslavsky E, Singh M. a compact mathematical programming formulation for DNA motif finding. *Lect Notes Comput Sci* 2006;**4009**:13.
104. Reid JE, Wernisch L. STEME: efficient EM to find motifs in large data sets. *Nucleic Acids Res* 2011;**39**:e126.
105. Jin VX, O'Geen H, Iyengar S, *et al*. Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Res* 2007;**17**:807–17.
106. Favorov AV, Gelfand MS, Gerasimova AV, *et al*. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* 2005;**21**:2240–5.
107. Zhao Y, Ruan S, Pandey M, *et al*. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* 2012;**191**:781–90.
108. Redhead E, Bailey TL. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics* 2007;**8**:385.
109. Chen X, Xu H, Yuan P, *et al*. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 2008;**133**:1106–17.
110. Huggins P, Zhong S, Shiff I, *et al*. DECOD: Fast and Accurate Discriminative DNA Motif Finding. *Bioinformatics* 2011;**27**:2361–7.
111. Cuellar-Partida G, Buske FA, McLeay RC, *et al*. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* 2012;**28**:56–62.
112. Mason MJ, Plath K, Zhou Q. Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics* 2010;**26**:2826–32.
113. Teng L, He B, Gao P, *et al*. Discover context-specific combinatorial transcription factor interactions by integrating diverse ChIP-Seq data sets. *Nucleic Acids Res* 2014;**42**:e24.
114. Vaquerizas JM, Kummerfeld SK, Teichmann SA, *et al*. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;**10**:252–63.
115. Ding J, Cai X, Wang Y, *et al*. ChIPModule: systematic discovery of transcription factors and their cofactors from ChIP-seq data. *Pac Symp Biocomput* 2013:320–31.
116. Ettwiller L, Paten B, Ramialison M, *et al*. Trawler: *de novo* regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat Methods* 2007;**4**:563–5.
117. Linhart C, Halperin Y, Shamir R. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res* 2008;**18**:1180–9.
118. Kuttippurathu L, Hsing M, Liu Y, *et al*. CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. *Bioinformatics* 2011;**27**:715–7.
119. Ho JW, Bishop E, Karchenko PV, *et al*. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* 2011;**12**:134.
120. Che D, Li G, Mao F, *et al*. Detecting uber-operons in prokaryotic genomes. *Nucleic Acids Res* 2006;**34**:2418–27.
121. Liu Y, Schmidt B, Liu W, *et al*. CUDA–MEME: accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units. *Pattern Recognit Lett* 2010;**31**:8.
122. Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.