OXFORD

Briefings in Bioinformatics, 16(6), 2015, 964–973

doi: 10.1093/bib/bbv022
Advance Access Publication Date: 17 April 2015
Paper

A review of ensemble methods for *de novo* motif discovery in ChIP-Seq data

Andrei Lihu and Ştefan Holban

Corresponding author. Andrei Lihu, Str. Take Ionescu 45 ap. 9, 300043 Timişoara, Romania. Tel.: +40 742545392; E-mail: andrei.lihu@gmail.com

Abstract

De novo motif discovery is a difficult computational task. Historically, dedicated algorithms always reported a high percentage of false positives. Their performance did not improve considerably even after they adapted to handle large amounts of chromatin immunoprecipitation sequencing (ChIP-Seq) data. Several studies have advocated aggregating complementary algorithms, combining their predictions to increase the accuracy of the results. This led to the development of ensemble methods. To form a better view on modern ensembles, we review all compound tools designed for ChIP-Seq. After a brief introduction to basic algorithms and early ensembles, we describe the most recent tools. We highlight their limitations and strengths by presenting their architecture, the input options and their output. To provide guidance for next-generation sequencing practitioners, we observe the differences and similarities between them. Last but not least, we identify and recommend several features to be implemented by any novel ensemble algorithm.

Key words: next-generation sequencing; motif discovery; ensemble methods; ChIP-Seq; transcription factors

Introduction

De novo computational DNA motif discovery is central to understanding and controlling gene expression. Motifs are typically short nucleotide sequences [5–20 base pairs (bp) in length] that are overrepresented statistically. They may appear several times across or within genes and it is conjectured that they possess biological significance, as they often represent the sequence-specific binding sites for 'transcription factors' (TFs) and other classes of regulatory proteins [1]. Motif finding is a computationally daunting task: given a collection of sequences, one must find an unknown but frequent pattern of unknown length, while taking into account possible mutations, deletions or insertions. A motif is generally contiguous, found on both strands of the DNA, and it can also be palindromic or gapped [2].

Before the 'next-generation sequencing' (NGS) era, algorithms for motif finding were designed for promoter analysis, receiving as input a set of a few hundred sequences of co-regulated genes [3]. The advent of 'chromatin immunoprecipitation' (ChIP [4]) combined with high-throughput NGS increased the accuracy of locating *in vivo* the 'transcription factors' binding sites' (TFBS), from several thousands of bp down to 300 bp [5]. Also known as 'ChIP sequencing' (ChIP-Seq) [6], this novel method can produce a large amount of data with increased precision and lower noise (e.g. Illumina HiSeq X Ten® System can produce over 1 TB of data per run). Nowadays, ChIP-Seq is the protocol of choice for most genome-wide investigations related to protein–DNA interactions and holds a paramount role in epigenetics research (e.g. mapping histone modifications) [7].

Historically, motif finding algorithms suffered from low-performance issues. Early studies, e.g. [8], showed that even the best-performing algorithm did not surpass levels of 13% in sensitivity and 35% in precision. A major problem was also a high rate of false positives, as discussed in [9]. After the introduction of NGS, with the aid of peak finding preprocessing methods, the input sequences have become shorter and more likely to be centered on the actual TFBS, and so the search could be circumscribed around 50–200 bp around the peaks [10]. Unfortunately, because algorithms for *de novo* motif discovery had to deal with massive amounts of data, their performance was affected by too many false positives and long execution times. To overcome

Andrei Lihu is a postdoc at the Politehnica University of Timişoara. His research interests are bioinformatics, motif prediction in particular, machine learning and evolutionary computation.
Ştefan Holban is a full professor of Computer Science at the Politehnica University of Timişoara. His research interests include data mining, pattern recognition and their applications in computational biology.
Submitted: 15 January 2015; **Received (in revised form):** 17 March 2015

these obstacles, some algorithms achieved small gains by taking into account phylogenetic information from closely related species [11] or by using different motif representations [12–14]. Nevertheless, since 2000s, even before the emergence of NGS technologies, based on the belief that authentic motifs could be identified by more than one method, studies have increasingly encouraged researchers to combine the results of various algorithms [15–18]. 'Ensemble methods', implemented as 'ensemble tools', delivered most improvements in motif proficiency.

Despite the widespread usage of ensemble methods in the NGS community, reviews on *de novo* motif prediction methods have focused only on individual algorithms [2] and on related web applications [19]. We believe that NGS practitioners and other interested parties may find useful an up-to-date review of ensembles for *de novo* motif discovery. To provide guidance in choosing the right tool, this article analyzes all the ensembles from scientific literature that, according to our best knowledge, were specifically designed to operate with data from ChIP-Seq experiments. First, we outline a set of basic notions regarding motif prediction along with fundamental algorithms that are found in most composite systems. Then, we briefly enumerate several compound methods from the period before the wide-scale dissemination of NGS technologies. Afterward, we center our attention on modern ChIP-Seq ensembles. We describe their structure, the input parameters and how the results are presented to the user. We show their advantages and disadvantages from a practical point of view and we highlight some desired characteristics that can be incorporated in future ensemble methods.

Fundamentals of *de novo* motif discovery

Motif finding is an NP-complete problem. In a formulation, known as 'planted motif search', it is stated that given n sequences, one must find an implanted pattern of length l (a l-*mer*) with at most d mutations [20]. If no mutations are considered, for any sequence with length m that contains an l-mer, there are $(m-l+1)^n$ possible solutions, thus rendering any brute force search computationally impracticable.

Motifs are commonly represented using consensuses or profile matrices. Given a set of aligned sequences for a TF, a consensus sequence is built choosing the predominant nucleotide from each position, while a 'degenerate consensus' is built taking into account the most frequent nucleotides per position and represented with IUPAC ambiguity codes [21]. The prevalent representation is based on profiles [22], which are $4 \times l$ 'positional weight matrices' (PWM) with four rows for nucleotides and l columns for motif sites. To construct a PWM, a collection of aligned sequences and a random background model are needed. Entries in the matrix represent log-likelihoods of the site-specific frequency of nucleotides versus the background model. The 'PWM score' (PWMS) indicates how far a sequence is from a random one and how well it conforms to a given motif profile. PWMS is computed by summing the elements of PWM that positionally match a given oligo-sequence [23]. PWMs can be illustrated with sequence logos [24].

Algorithms for motif detection are classified as 'word-based' or 'profile-based', depending on the motif representation [3].

The first category comprises consensus algorithms. For each possible l-mer, they gather from the input sequences its approximate occurrences with at most e mutations and rank them based on their overrepresentation. To prune the search space (4^l possible patterns), a typical method like Weeder [25] uses suffix-trees to hold data and enforces some constraints on locations where mismatches are allowed.

Profile-based algorithms perform heuristic searches by iteratively optimizing an initial PWM. Although the search space spans across all possible solutions, they avoid an exhaustive enumeration. Profile-based methods can cater to longer input sequences. Iteratively, these methods select some positions from the input set, align their associated sequences, build a PWM and score the obtained model. An example of such an algorithm is 'Multiple Expectation Maximization for Motif Elicitation' (MEME) [26], a multi-start local method that begins with separate profiles for each input l-mer, then selects the current best profile to be optimized deterministically in further 'expectation maximization' (EM) steps. MEME cannot detect spaced dyads *per se*, only as separate motifs. The Gibbs sampler [27] is another profile-based algorithm that can be seen as MEME's stochastic counterpart. Unlike MEME, it overcomes the generation of too many initial profiles by building only one random initial profile that is subsequently improved [23]. Both algorithms have inherent drawbacks: they assume the presence of a motif in each input sequence or they may prematurely end in local optima.

Older algorithms for promoter analysis had to adapt to cope with the massive amount of short reads from ChIP-Seq experiments. Newer Gibbs and MEME variants process only a subset of the input, while the rest is ignored [28]. For example, ChIPMunk [29] adopted a greedy optimization strategy combined with EM. Novel word-based algorithms were also designed for speed, like 'discriminative regular expression motif elicitation' (DREME) [30], that limits its search to motifs of maximum 8 bp. Nonetheless, MEME and Weeder are still used in predicting binding sites from ChIP-Seq data, but they require a higher computational effort.

Motif predictions can be checked against experimentally validated TFBS from dedicated databases (e.g. TRANSFAC [31], JASPAR [32], UniPROBE [33]) by using tools like TOMTOM [34] or STAMP [35]. Peak calling tools [36] can preselect ChIP-Seq regions as input for motif finding, but they can also work in parallel to verify the authenticity of predicted motifs.

Assessing the performance of motif discovery algorithms has always been a challenge. Regardless of the dramatic increase in the amount of data produced by the sequencing machines, a widely accepted benchmark did not materialize. Before the emergence of NGS, first attempts to compare the performance of methods for promoter analysis consisted in measuring the ability to find implanted motifs in randomly generated sequences [37]. Soon, data sets procured from laboratory replaced the implantation approach [8, 38]. In the ChIP-Seq period, algorithms are assessed using hybrid data collections that still contain some promoters (see [15] or [39]) or using sequences derived from ChIP-Seq assays like Chen *et al.*'s data set [40].

The performance of *de novo* motif detection algorithms has been always far from satisfactory [8], being afflicted by false positives [9]. A 'receiver operating characteristic' (ROC) curve, drawn to examine the relation between sensitivity and specificity [41], can illustrate the problem. In this regard, ensemble methods emerged as a viable solution.

Ensemble methods before the ChIP-Seq era

Ensemble methods combine different and complementary algorithms to improve the accuracy of prediction, in the same way any real-life important decision is made relying on advice from

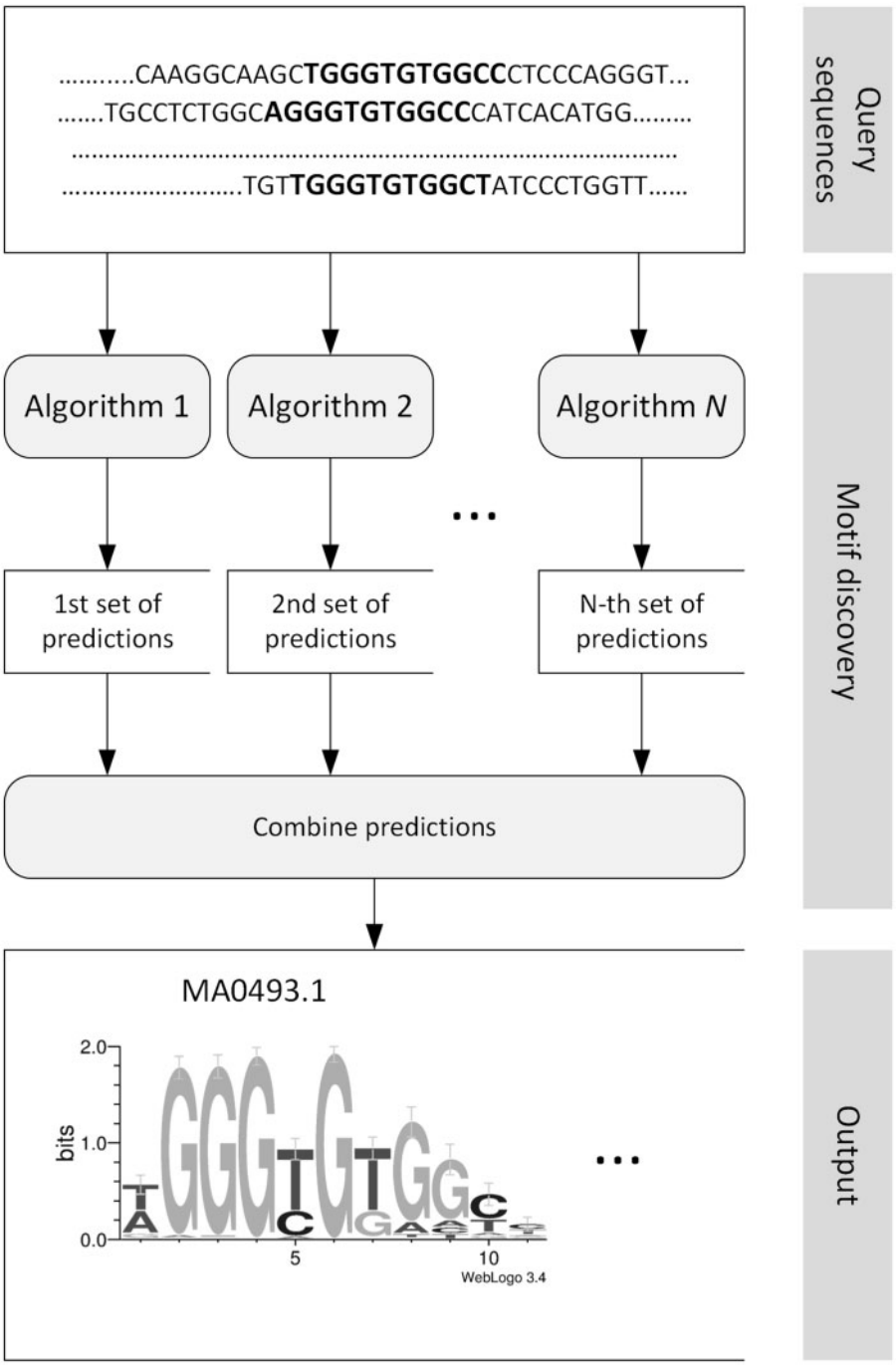

Figure 1. Overview of an ensemble method. In this example, the query sequences contain at least one *a priori* unknown motif. Some of its corresponding oligo-sequences are highlighted in bold characters for illustrative purposes only. *N* different algorithms independently process the same input and each generates its own set of predicted motifs. Afterward, all *N* sets of obtained motifs are pooled together and combined using a particular clustering and/or voting procedure. Finally, only significant top motifs are reported in the ensemble's output. In this illustration, the resulted first most significant motif, corresponding to the binding site of Krueppel-like factor 1, is represented with its sequence logo.

several experts. Such methods have flourished in the field of machine learning [42] and, subsequently, in bioinformatics [43, 44]. Although there are many types of aggregate methods (e.g. bagging [45], boosting [46], mixture of experts [47]), most *de novo* motif discovery ensembles are similar, being composed of several algorithms that process the same input sequences. They pool together and coalesce their predictions, then select top-ranking solutions with a higher confidence, as shown in Figure 1. We follow the same naming convention as in [18] and [48] and we refer to the above-mentioned approach, also called 'meta-server', as 'ensemble method'. These methods are implemented in practice as ensemble tools that can be accessed as stand-alone local applications or via the web.

A timeline for *de novo* motif discovery ensembles, emphasizing the partition between older tools for promoter analysis and newer methods for ChIP-Seq, is provided in Figure 2. As shown

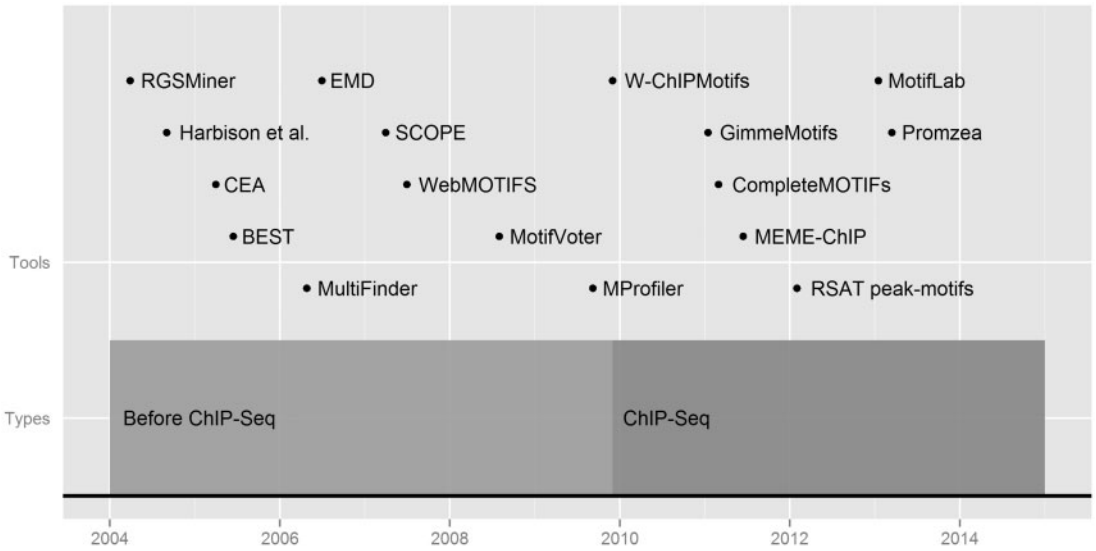

Figure 2. A timeline for ensemble methods. Ensembles are classified into two categories. The first category comprises early ensembles devised initially for promoter analysis before the emergence of ChIP-Seq assays. The second category comprises newer ensembles designed to handle massive amounts of read counts from ChIP-Seq experiments. All methods are plotted based on their publication date.

Table 1. Ensemble methods before ChIP-Seq era[a]

Ensemble	Components	Running time estimations
RgS-Miner	Gibbs sampler, MEME, AlignACE	7.77%–64.74%–27.49%
Harbison *et al.*	AlignACE, MEME, MDScan, method from Kellis *et al.*, MEME_c, CONVERGE	N/A
BEST	AlignACE, BioProspector, CONSENSUS, MEME	0.56%–1.27%–53.04%–45.13%
MultiFinder	MDScan, BioProspector, MEME, AlignACE	0.08%–0.71%–93.56%–5.65%
CEA, EMD	AlignACE, MEME, BioProspector, MDScan, MotifSampler	1.21%–77.50%–0.90%–0.12%–20.27%
SCOPE	BEAM, PRISM, SPACER	N/A
WebMOTIFS	MEME, AlignACE, MDScan, Weeder	54.03%–8.31%–0.08%–37.58%
MotifVoter, MProfiler	MITRA, Weeder, SPACE, AlignACE, ANN-Spec, BioProspector, Improbizer, MDScan, MEME, MotifSampler	7.20%–5.54%–14.13%–6.93%–5.17% –0.09%–5.08%–0.46%–32.50%–22.90%

[a]First appearances in Table 1 with their respective citations: AlignACE [49], MDScan [50], method from Kellis *et al.* [51], CONSENSUS [52], BioProspector [53], MotifSampler [54], BEAM [55], PRISM [56], SPACER [57], MITRA [58], SPACE [59], ANN-Spec [60], Improbizer [61].
Aggregate methods are listed along with their component algorithms and the running time percentage for each component. Comma-separated ensembles contain the same algorithms; however, they are improved versions in terms of the clustering/voting procedure. To show which component has the longest execution time, we measured the mean running time of each component across the 13 data sets from Chen *et al.* To fit the input requirements of all algorithms, we used a random sample of 500 sequences from each of the 13 Chen *et al.*'s mouse embryonic stem cells ChIP-Seq data sets [40] provided in [30]. Ensembles were run with default parameter values.

in the drawing, even before the rise of ChIP-Seq-capable methods, there have been a series of endeavors to harness the power of multiple algorithms (see Table 1). They are briefly described in the next paragraphs.

RgS-Miner was an earlier integrated system for promoter analysis [62] that used a k-means algorithm [63] for motif clustering and reported the closest patterns to centroids. When battered by false discoveries obtained with individual methods, Harbison *et al.* in [15] undertook a similar approach by improving the predictions of six algorithms using k-medoids [64]. WebMOTIFS [65] used the same clustering procedure as in Harbison *et al.*'s study to gather the results of four *de novo* algorithms (i.e. MEME, AlignACE, MDScan and Weeder). Another attempt was MultiFinder, a pipeline that merged and ranked motifs using hierarchical clustering with several scoring functions [66]. Ensembles like SCOPE [67] and BEST [68] reported top motifs from individual methods subject to a common scoring function. However, Tmod [69] and Melina II [70] were notable exceptions because, although they could run several algorithms, they did not automatically cluster the results.

Early investigations on the matter of motif discovery argued for a mix of methods [17]. An influential study was Tompa *et al.* assessment of 13 motif finding algorithms on a eukaryotic data set. The survey recommended using a collection of complementary tools to mitigate the low performance of individual algorithms [8]. Accordingly, Hu *et al.*'s complementary study on prokaryotes followed the advice [16] and they sketched the 'consensus ensemble algorithm' (CEA), which collected motifs from multiple runs of the same workflow. After being further improved and renamed 'Ensemble Motif Discovery' (EMD), it outperformed by 22.4% in accuracy the best stand-alone method in a test [18]. EMD clustered and ranked top motifs from several methods using a voting system. The concept was later developed by MotifVoter and MProfiler [48, 71].

Ensemble methods for ChIP-Seq data

Older ensemble methods can still be used to find motifs in ChIP-Seq read counts (e.g. SCOPE in [72]); however, more suitable tools have been developed.

Most recent ensembles share many common features but are composed of methods with different strengths and weaknesses. They seek to provide additional value other than chaining together third-party modules and they offer a better insight into data through an augmented visual presentation. For a user's convenience, most of the tools are available as online portals and many parts of the workflow are automated to save time.

Hereinafter, we describe all the ensembles that were primarily devised to deal with ChIP-Seq data.

'W-ChIPMotifs' (2009) [73] is a web server limited to human or mouse genome analysis. Data can be entered directly in the browser or through file uploading, with an upper limit of 600 KB. Between 10 and 2000 sequences in FASTA format are recommended as input, considering that more could pose a problem to MEME [74], a time and resource consuming algorithm with $O(n^2)$ complexity. The user has the option to conduct differential analysis. If no control data are provided, then one is generated using 5000 random promoter sequences automatically picked from the selected species. W-ChIPMotifs includes MEME, Weeder and a greedy search strategy that relies on indexing—'Mammalian Motif Finder' (MaMF) [75]. The union of their results is assessed against a randomized initial input. The top-scoring motifs are selected in two rounds: the first is based on profile scores, the second on the Fisher significance test. Finally, STAMP matches the resulted motifs in TRANSFAC and JASPAR. The final report contains the predicted patterns along with their sequence logos, PWMs, scores, *P*-values. The web portal automates many steps of the workflow at the expense of less user control and it also lacks the ability to predict alternative binding motifs [19]. Neither the source code nor any installation package is publicly available at this time.

'GimmeMotifs' (2011) [72] is a collection of configurable command-line utilities that can aggregate up to nine motif-finding algorithms. Not being an online portal, it requires some computer expertise to install and configure. All algorithms run in parallel using an 'inter process communication' (IPC) solution—Parallel Python [76]. The only mandatory parameter is the input file. Optional parameters include the reference genome, the algorithms to be run, the maximum running time, the background model, the size of desired motifs and a cutoff *P*-value along with an enrichment level to select significant motifs. Large inputs, given in BED or FASTA format, are trimmed to 20% of their initial size; however, the fraction can be changed. Obtained motifs are filtered using randomized data generated from the remaining 80% of the initial sequences and similar patterns are merged using a clustering procedure based on an information content metric. Statistics for the validation of the captured motifs are the hypergeometric *P*-value, the ROC curve, the 'area under receiver operating characteristic', the 'mean normalized conditional probability' (MNCP) [77] and the absolute enrichment. Two types of background models can be used: a first-order Markov model with frequencies similar to the input dinucleotides' frequencies and another model generated with frequencies of randomly chosen genes around the 'transcription start sites' of the specified genome. Apart from the regular sequence logos, PWMs and motif scores, the output provides a histogram with the motif's position relative to the peak's center [72]. The product is open-source, well documented and installer packages are available for several Linux distributions.

'CompleteMOTIFs' (2011) [78] is a web platform. Alongside motif discovery, it offers a few useful data set operations, peak region annotation and BED-FASTA conversion utilities for mouse (mm9), rat (rn4) and human (hg18, hg19). For motif prediction, the system accepts input data in BED, FASTA and GFF formats, either directly in the browser or through file uploading (maximum 100 MB). CompleteMOTIFs incorporates a motif scanning method Patser [52] and three *de novo* discovery algorithms: CUDA-MEME [79], Weeder and an advanced version of ChIPMunk—ChIPHorde. Weeder is using OpenMP for parallel processing. CUDA-MEME uses the 'Compute Unified Device Architecture' (CUDA) programming model on a 'graphical processing unit' (GPU) to accelerate MEME's execution. There are restrictions depending on the chosen methods to be run: if ChIPMunk is used alone then the maximum number of input bases is limited to 5 million, while for MEME, Weeder or Patser the limit is 500 000. Motif scanning is done by Patser, which uses TRANSFAC and JASPAR as motif compendia, but it also accepts user-defined profiles. After scanning, a background random model is created by shuffling the user's original input or by using pre-compiled upstream sequences from the considered genome. Afterward, this model is involved in calculating a *P*-value corrected for 'false discovery rate', allowing an estimation regarding the significance of results. Finally, top 10 motifs from each of the four methods are collected, ranked and inventoried with STAMP. User options include setting a *P*-value cutoff, choosing a background random sequence type and its nucleotide shuffling parameters, the motif width for MEME, the reference genome and the motif databases. The final report contains information in HTML and text formats that is specific to each algorithm involved in prediction [19]. The source code is not public, but the product can be downloaded on request. The standalone application requires compiling and manually adding the motif databases owing to different licensing policies of algorithm and databases. On the online portal, free accounts are offered to academic users who prefer storing their results on the server.

'MEME-ChIP' (2011) [28, 80] is part of the MEME Suite online platform [81]. Data, entered directly into the browser or by file uploading, must be only in FASTA format and should not exceed 50 MB. It is recommended that the sequences are peak-centered. However, only the middle 100 bp are actually used in the prediction. The users can select a reference database from those available (JASPAR, UniPROBE, etc.) or can provide their own. For vertebrates, 'JASPAR Vertebrates and UniPROBE Mouse' is appropriate for most cases. The ensemble includes two algorithms for *de novo* motif discovery, MEME and DREME, coupled with an algorithm for enrichment analysis—CentriMo ('central motif enrichment analysis') [82]. MEME performs the task of finding long motifs (maximum of 30 bp in length) but, owing to its $O(n^2)$ complexity, only at most 600 randomly sampled input sequences are considered. While default options for MEME work well in most cases, the user can adjust several parameters like the maximum and minimum of sites per motif, motif width or the number of motifs to be returned. The implicit background model is a first-order Markov model built from the input sequences, but the user can upload a custom model. The default expected motif site distribution 'Zero or one occurrence per sequence' suits well the majority of large-scale studies; the others are 'One occurrence per sequence' (fastest) and 'Any number of repetitions' (slowest). To run a parallel version of MEME, a 'Message Passing Interface' (MPI) implementation and

a batch scheduler are needed. Before processing, MEME-ChIP trims the input sequences to 100 bp and centers them, forming the input for DREME, a fast word-based algorithm. Regarding DREME, the user can specify values for search termination conditions, which are the maximum number of reported motifs and the estimated statistical significance, represented by the E-value. DREME and MEME complement each other: 'MEME is highly specific but slower, whereas DREME is less specific but faster'. [80] Because it performs motif enrichment analysis on 'known' motifs, CentriMo is not a *de novo* method. Nonetheless, using CentriMo's output graph of motif probability in sequences, researchers can also identify co-factors and check the quality of the ChIP experiment (see [80]). The final output, obtained by ranking the results of the three algorithms, can be explored in XML, text or interactive HTML forms. Motifs are ordered by their E-value, grouped by similarity and described in detail by their sequence logo, occurrence sites, regular expression, etc. The final report also includes links to other tools, like TOMTOM, for further analysis. MEME-ChIP can analyze large-scale data for any genome. Available as a web server, it is also provided as a web service through the Opal2 platform [83]. It can be downloaded and installed on a local machine and its source code is public.

'RSAT (Regulatory Sequence Analysis Tools) peak-motifs' (2012) [84] is a web-based toolset for detection of cis-regulatory elements, that is accessible in the browser or through SOAP web-services. Alongside motif prediction, it also includes tools for statistics and genome management. This online workbench can be used with any type of genome and it accepts several formats, directly in the browser, through file uploading or pasting a URL of a sequence file from a server. From a BED file, peak-motifs can return sequences for any organism. An optional but distinctive feature of RSAT is represented by the possibility to perform differential analysis: the user can input two sets of sequences, run the motif discovery pipeline and assess the resulted enriched motifs. Before *de novo* discovery, the input peak sequences can be shrunk and filtered in a facultative step. Motifs are predicted with at most four word-based algorithms. Users can select the algorithms to run, the oligomer lengths, the desired Markov order of the random background model and the desired number of motifs returned by each algorithm, among other options. The user can choose from a large list of reference databases, but can also use a custom one. The predicted sites can be exported as custom UCSC tracks that can be viewed in genome browsers. RSAT peak-motifs can be obtained by request and installed on Unix-like operating systems.

'MotifLab' (2013) [85] is a stand-alone Java desktop application for analysis of regulatory regions and motif discovery. It is using Java threads to provide concurrency, but this feature is not fully used in the current version (as stated in the user's manual). An important advantage is the option to include external algorithms, either automatically installed from a preconfigured repository or manually added. The application contains two built-in ensemble methods not meant for ChIP-Seq data. The first one, 'Simple Ensemble', returns the sites and motifs where at least *M* different methods predict at least *N* nucleotides. The second one is Hu *et al.*'s EMD. However, the user can add external ensemble methods for ChIP-Seq using XML configuration files. Depending on the operating system, the workbench currently accepts the following algorithms: ChIPMunk, MEME, AlignACE, Weeder and BioProspector. After running an ensemble, the final report highlights the binding sites with their associated motifs. The consensus representation, the PWM and the sequence logo are displayed for each motif. Besides the results, the user is endowed with a toolset to assess the similarities and differences between distinct patterns. At last, the lack of various and more recent default ensembles in MotifLab and the effort in manual configuration may represent a drawback, but the extensive documentation can overcome this. The tool also runs in command-line mode.

'Promzea' (2013) [86] is an online web-server for detecting motifs in plant species. It is particularly suited for maize (*Zea mays*), rice (*Oryza sativa*) and *Arabidopsis thaliana*. Sequences can be entered directly in the browser or through file uploading (maximum 1 MB). Promoter data can be specified as gene IDs, microarray probe-set IDs (for maize) or FASTA, while ChIP-Seq peaks can be introduced in BED format. Information in cDNA FASTA format is deferred to a BLAST procedure and the system retrieves a corresponding list of promoters to analyze from the database of the selected plant genome, with lengths specified by the user. The ensemble combines three *de novo* algorithms: MEME, BioProspector and Weeder. Their results are mixed, then ranked using the MNCP [77]. MEME contributes with 10 motifs of a maximum size of 10 each, BioProspector with 10 motifs of size 10 each, while Weeder returns motifs with a size between 6 and 10. The results of each method are filtered. For BioProspector and Weeder the filtering is based on a binomial distribution *P*-value test, while for MEME is based on a hypergeometrical distribution *P*-value test. The final report shows each motif along with its score, the algorithm that predicted it, the sequence logo and a graph with the motif's frequency in the input data. Every motif is presented along with the annotated genes that contain it. A claimed advantage of Promzea is that it handles internally the particularities of the three species of interest, like the distribution of the distal cis-acting elements and the high percent of transposable elements in maize and rice genomes.

As an overview, we provide Table 2, which summarizes information about the component algorithms, product type, the availability of their sources, details about installation and the genomes the method is suited for. It is also concerned with input formats and restrictions, parallelization and acceleration, a few relevant user options (number of motifs to be returned and *P*-value) and which databases can be used for motif comparison.

Conclusions

We reviewed seven ensemble tools designed to process ChIP-Seq data and observed their limitations and strengths.

Except RSAT peak-motifs, all tools are a combination of profile-based and word-based algorithms. Three of them, W-ChIPMotifs, Promzea and CompleteMOTIFs, greatly restrict the user input to $<$20 MB, rendering them unfit for large-scale analysis.

While most tools can be used for any genome, the applicability of W-ChIPMotifs and Promzea is limited because the first can be used only for mouse and human, and the second is a better choice for a subset of plant genomes.

MEME-ChIP, followed by RSAT peak-motifs, provides most options for motifs *per se* (width, predicted sites, significance), while W-ChIPMotifs offers the least. RSAT peak-motifs and MEME-ChIP also offer a multitude of reference databases and the ability to use a custom motif compendium, while other tools are either limited to a few (W-ChIPMotifs, GimmeMotifs, CompleteMOTIFs) or do not offer motif comparison at all (Promzea, MotifLab).

Higher order background models may improve the prediction ability of algorithms. Except Promzea, MotifLab and

Table 2. Characteristics of ChIP-Seq ensembles[a]

Ensemble	*De novo* components	P/W	Product type and platform	Sources available	Installation	Parallelization	Accelerator	Input formats	Input restrictions	No. of motifs option	*P*-value option	Motif databases	Targeted genomes
W-ChIPMotifs	MEME, MaMF, Weeder	P, W	On-line portal	No	On request	–	–	FASTA	Maximum upload file 600 KB. Between 10 and 2000 input sequences	No	No	TRANSFAC, JASPAR	Mouse and human
GimmeMotifs	MEME, Improbizer, MDScan, Bioprospector, GADEM, MoAn, MotifSampler, Trawler, Weeder	P, W	Ubuntu, Debian and Fedora Linux command-line tools	Yes	DEB, RPM or install from source	IPC (via Parallel Python)	No	BED, FASTA	–	No	Yes	JASPAR	Any
Complete MOTIFs	CUDA-MEME, Weeder, ChIPHorde/ ChIPMunk	P, W	On-line portal. Fedora, RHEL and Ubuntu Linux application	No	TGZ file or VMWare image (on request)	OpenMP (for Weeder)	CUDA-GPU (for MEME)	FASTA, BED, GFF	Maximum 5 000 000 bp for ChIPMunk. Maximum 500 000 bp for Weeder, MEME	No	No	TRANSFAC, JASPAR, user	Any
MEME-ChIP	MEME, DREME	P, W	On-line portal. SOAP web-services. Linux, OS X, Cygwin com-mand-line tools	Yes	Installs from source	MPI (for MEME)	No	FASTA	Maximum 50 MB input data. Peak-centered sequences, minimum 100 bp each	Yes	Yes	JASPAR, UniProbe etc. and user	Any
RSAT peak-motifs	Oligo-analysis, position-analysis, local-word-analysis, dyad-analysis	W	On-line portal. SOAP web-services. Linux, OS X, UNIX command-line tools	On request	On request. Install script	–	–	FASTA, BED, raw, wconsensus, IG, multi, tab	–	Yes	Yes	JASPAR, UniProbe etc. and user	Any
MotifLab	AlignAce, BioProspector, ChIPMunk, MEME, MotifSampler, Priority, Weeder	P, W	Java desktop application (cross-platform)	No	Java archive (JAR)	Shared memory (Java threads)	No	FASTA, BED, GFF	–	Yes	No	N/A	Any
Promzea	MEME, BioProspector, Weeder	P, W	On-line portal	No	No	–	No	cDNA FASTA, BED, gene ID, micro-array ID	Maximum 1 MB upload	No	No	N/A	Maize, rice and *Arabidopsis thaliana*

[a]First appearances in Table 2 with their respective citations: MoAn [87], Trawler [88], oligo-analysis[89], dyad-analysis [90].

For each ensemble method, its *de novo* component algorithms and several important features are specified. Fields marked with '-' should be interpreted as unavailable or not enough data. The column 'P/W' shows whether an ensemble method implements profile-based (P) and/or word-based (W) discovery methods. 'Parallelization' reports which parallel programming model or technology is used. Information about acceleration technologies is given in 'Accelerator'. 'No. of motifs option' and '*P*-value option' for significance testing are reported for common versions of the tools in default mode. 'Motif databases' are compendia of motifs used by *de novo* ensembles to verify the results. For brevity, not more than two reference databases are displayed, while 'user' represents the possibility to use custom databases. 'Targeted genome' refers to organisms supported by the current tool.

W-ChIPMotifs, all other tools offer the possibility to use a custom higher order random model.

Enrichment analysis complements *de novo* prediction and may help locating additional secondary motifs. MEME-ChIP is the only ensemble method that can perform motif discovery and enrichment analysis altogether. Nonetheless, a differential analysis toward a control set can only be performed in RSAT peak-motifs, CompleteMOTIFs and GimmeMotifs.

Parallelization and acceleration technologies can reduce the running times of motif finding algorithms [79]. CompleteMOTIFs, GimmeMotifs and MEME-ChIP implement different parallel programming models to speed up the computation. CompleteMOTIFs is the only ensemble tool that uses an acceleration technology.

Exposing the functionality through web-services is valuable for programmatic access; however, web portals are better suited for users with less computer expertise. Except GimmeMotifs and MotifLab, all other tools are web applications, out of which only RSAT and MEME-ChIP are available as web-services. The majority of the platforms can be installed on Unix-like environments. As seen in Table 2, not all of them provide the source code. The availability of sources allows verifying the code and adapting it to particular needs.

De novo motif discovery algorithms of an ensemble may yield discordant results. Depending on each ensemble's approach [48], discrepant results are not always considered false discoveries. If the ensemble re-ranks all motifs found by individual methods using a scoring function (Promzea, W-ChIPMotifs, MEME-ChIP, etc.), then, depending on their final rank, even the discrepant results may be accepted as solutions (e.g. as in the analysis of the SCL ChIP-Seq data set in [28]). However, if the ensemble relies on the consensus of several motif finders (e.g. the built-in ensembles from MotifLab), then the discordant patterns are considered less likely to be real motifs.

Because ensembles have various strengths, but also weaknesses, it is recommended to perform parallel analyses with several ensemble tools. The results of most ensembles are provided in various text formats that can be further processed or analyzed (TRANSFAC motif format, Weeder format, MEME minimal output format, etc.). To summarize all different outputs from multiple ensemble methods, we recommend using a tool like STAMP (also used to compare results from separate methods in [91]), which accepts formats and mixtures of formats from different methods [35]. If the output is not accepted by STAMP, users can convert it with utilities such as 'convert-matrix' from the RSAT suite or write their own conversion scripts. However, even if there is a strong 'algorithmic' consensus on some motifs, it is still required to inspect the results visually to exclude poor quality motifs.

We conclude that any novel and better ensemble should process large input sequences and be suited for any genome. It should mix profile with word-based methods and allow motif comparison in a multitude of known databases and also in user-provided compendia. The tool should have options to use a higher custom background model and choose motif parameters like width, predicted sites and several significance measures. Other important features to be considered are differential and enrichment analysis. To shorten the running time, it should use both parallelization and acceleration technologies (e.g. mCUDA-MEME [92], that includes MPI, CUDA and OpenMP). It should be available online, directly in the browser and through web-services, but it should also offer the possibility to be installed on a local machine, on any operating system, and should adhere to an open-source policy. Last but not least, the graphical interface should be user friendly and should help the user in exploring and interpreting the results.

Key Points

- Compared with individual algorithms, ensemble methods for *de novo* motif finding may reduce false discoveries.
- Modern ensemble tools that handle ChIP-Seq data include algorithms with complementary strengths and aim to provide a user-friendly interface.
- It is recommended to use at least two ensembles for motif prediction on the same data set because existing tools have limitations.
- By exposing their strengths and weaknesses, this article offers a guide on choosing the right ensembles depending on particular needs.
- We advocate a set of features for a better ensemble tool.

Acknowledgements

We would like to express our gratitude to Sek Won Kong, Harvard Medical School, for providing us with the installation guide and archive for CompleteMOTIFs, as well as to Jacques van Helden, Université d'Aix-Marseille, for the stand-alone version of RSAT peak-motifs. We would like to thank Mihai Udrescu, Politehnica University of Timişoara, for suggesting improvements to earlier versions of this article. We are very grateful to Oana-Andreea Lihu for the valuable comments and suggestions to improve the manuscript.

Funding

This work was partially supported by the strategic grant POSDRU/159/1.5/S/137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007–2013.

References

1. D'haeseleer P. What are DNA sequence motifs? *Nat Biotechnol* 2006;**24**:423–5.
2. Das MK, Dai H-K. A survey of DNA motif finding algorithms. *BMC Bioinformatics* 2007;**8**(Suppl 7):S21.
3. 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.
4. Collas P. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front Biosci* 2008;**13**:929.
5. Sung PBF. Methods, systems and kits for detecting protein-nucleic acid interactions, 2010..
6. Johnson DS, Mortazavi A, Myers RM, *et al*. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 2007;**316**:1497–502.
7. Bailey T, Krajewski P, Ladunga I, *et al*. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* 2013;**9**:e1003326.
8. 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.

9. Zia A, Moses AM. Towards a theoretical understanding of false positives in DNA motif finding. *BMC Bioinformatics* 2012; **13**:151.
10. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008;**36**:5221–31.
11. Hawkins J, Grant C, Noble WS, *et al*. Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics* 2009;**25**:i339–47.
12. Osada R, Zaslavsky E, Singh M. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics* 2004;**20**:3516–25.
13. Bailey TL, Bodén M, Whitington T, *et al*. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* 2010;**11**:179.
14. 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.
15. Harbison CT, Gordon DB, Lee TI, *et al*. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;**431**: 99–104.
16. Hu J, Li B, Kihara D. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res* 2005;**33**: 4899–913.
17. MacIsaac KD, Fraenkel E. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol* 2006;**2**:e36.
18. Hu J, Yang YD, Kihara D. EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC Bioinformatics* 2006;**7**:342.
19. Tran NTL, Huang CH. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biol Direct* 2014;**9**:4.
20. Pevzner PA, Sze SH. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol* 2000;**8**:269–78.
21. Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). *Proc Natl Acad Sci USA* 1986;**83**:4–8.
22. Stormo GD, Schneider TD, Gold L, *et al*. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res* 1982;**10**:2997–3011.
23. Xia X. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica (Cairo)* 2012;**2012**:917540.
24. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990;**18**: 6097–100.
25. Pavesi G, Mauri G, Pesole G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 2001;**17**: S207–14.
26. 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.
27. Lawrence C, Altschul S, Boguski M, *et al*. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993;**262**:208–14.
28. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 2011;**27**:1696–7.
29. Kulakovskiy IV, Boeva VA, Favorov A V, *et al*. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 2010;**26**:2622–3.
30. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 2011;**27**:1653–9.
31. Matys V, Fricke E, Geffers R, *et al*. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;**31**: 374–8.
32. Sandelin A, Alkema W, Engström P, *et al*. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;**32**:D91–4.
33. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 2009;**37**:D77–82.
34. Gupta S, Stamatoyannopoulos JA, Bailey TL, *et al*. Quantifying similarity between motifs. *Genome Biol* 2007;**8**:R24.
35. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 2007;**35**: W253–8.
36. Szalkowski AM, Schmid CD. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief Bioinform* 2011;**12**:626–33.
37. Jia C, Carson MB, Yu J. A fast weak motif-finding algorithm based on community detection in graphs. *BMC Bioinformatics* 2013;**14**:227.
38. Li N, Tompa M. Analysis of computational approaches for motif discovery. *Algorithms Mol Biol* 2006;**1**:8.
39. 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.
40. 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.
41. Henderson AR. Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. *Ann Clin Biochem* 1993;**30**(Pt 6):521–39.
42. Kuncheva LI. *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, New Jersey: Wiley-Interscience; 2004.
43. Larrañaga P, Calvo B, Santana R, *et al*. Machine learning in bioinformatics. *Brief Bioinform* 2006;**7**:86–112.
44. Yang P, Hwa Yang Y, Zhou B, *et al*. A review of ensemble methods in bioinformatics. *Curr Bioinform* 2010;**5**: 296–308.
45. Breiman L. Bagging predictors. *Mach Learn* 1996;**24**:123–40.
46. Schapire RE. The strength of weak learnability. *Mach Learn* 1990;**5**:197–227.
47. Jacobs RA, Jordan MI, Nowlan SJ, *et al*. Adaptive mixtures of local experts. *Neural Comput* 1991;**3**:79–87.
48. Wijaya E, Yiu SM, Son NT, *et al*. MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. *Bioinformatics* 2008;**24**:2288–95.
49. Hughes JD, Estep PW, Tavazoie S, *et al*. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000;**296**:1205–14.
50. 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.
51. Kellis M, Patterson N, Endrizzi M, *et al*. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 2003;**423**:241–54.
52. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999;**15**:563–77.
53. 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.

54. Thijs G, Lescot M, Marchal K, *et al*. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 2001;**17**: 1113–22.
55. Carlson JM, Chakravarty A, Gross RH. BEAM: a beam search algorithm for the identification of cis-regulatory elements in groups of genes. *J Comput Biol* 2006;**13**:686–701.
56. Carlson JM, Chakravarty A, Khetani RS, *et al*. Bounded search for de novo identification of degenerate cis-regulatory elements. *BMC Bioinformatics* 2006;**7**:254.
57. Chakravarty A, Carlson JM, Khetani RS, *et al*. SPACER: identification of cis-regulatory elements with non-contiguous critical residues. *Bioinformatics* 2007;**23**:1029–31.
58. Eskin E, Pevzner PA. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 2002;**18**(Suppl 1):S354–63.
59. Wijaya E, Rajaraman K, Yiu S-M, *et al*. Detection of generic spaced motifs using submotif pattern mining. *Bioinformatics* 2007;**23**:1476–85.
60. Workman CT, Stormo GD. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput* 2000:467–78.
61. Ao W, Gaudet J, Kent WJ, *et al*. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* 2004; **305**:1743–6.
62. Huang HD, Horng JT, Sun YM, *et al*. Identifying transcriptional regulatory sites in the human genome using an integrated system. *Nucleic Acids Res* 2004;**32**:1948–56.
63. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inf Theory* 1982;**28**:129–37.
64. Kaufman L, Rousseeuw PJ. Clustering by means of medoids. *Stat. Data Anal. Based L1-Norm Relat. Methods* 1987;**405**:16.
65. Romer KA, Kayombya G-R, Fraenkel E. WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucleic Acids Res* 2007;**35**:W217–20.
66. Huber BR, Bulyk ML. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics* 2006;**7**:229.
67. Carlson JM, Chakravarty A, DeZiel CE, *et al*. SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Res*. 2007;**35**:W259–64.
68. Che D, Jensen S, Cai L, *et al*. BEST: binding-site estimation suite of tools. *Bioinformatics* 2005;**21**:2909–11.
69. Sun H, Yuan Y, Wu Y, *et al*. Tmod: toolbox of motif discovery. *Bioinformatics* 2010;**26**:405–7.
70. Okumura T, Makiguchi H, Makita Y, *et al*. Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions. *Nucleic Acids Res* 2007;**35**:W227–31.
71. Altarawy D, Ismail MA, Ghanem SM. Pattern Recognition in Bioinformatics. Pattern Recognit. Bioinformatics, 4th {IAPR} International Conference {PRIB} 2009, Sheffield, UK, September 7–9, 2009. Proceedings on 2009; 5780.
72. Van Heeringen SJ, Veenstra GJC. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* 2011;**27**:270–1.
73. Ho JWK, Bishop E, Karchenko P V, *et al*. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* 2011;**12**:134.
74. Kennedy BA, Lan X, Huang TH-M, *et al*. Using ChIPMotifs for de novo motif discovery of OCT4 and ZNF263 based on ChIP-based high-throughput experiments. *Methods Mol Biol* 2012; **802**:323–34.
75. Hon LS, Jain AN. A deterministic motif finding algorithm with application to the human genome. *Bioinformatics* 2006;**22**: 1047–54.
76. Vanovschi V. Parallel Python Software, 2015.
77. Clarke ND, Granek JA. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics* 2003;**19**:212–18.
78. Kuttippurathu L, Hsing M, Liu Y, *et al*. CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. *Bioinformatics* 2011;**27**:715–7.
79. 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**:2170–7.
80. Ma W, Noble WS, Bailey TL. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat Protoc* 2014;**9**: 1428–50.
81. Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**: W202–8.
82. Bailey TL, Machanick P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* 2012;**40**:e128.
83. Ren J, Williams N, Clementi L, *et al*. Opal web services for biomedical applications. *Nucleic Acids Res* 2010;**38**:W724–31.
84. 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.
85. Klepper K, Drabløs F. MotifLab: a tools and data integration workbench for motif discovery and regulatory sequence analysis. *BMC Bioinformatics* 2013;**14**:9.
86. Liseron-Monfils C, Lewis T, Ashlock D, *et al*. Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the Maize Development Atlas. *BMC Plant Biol* 2013;**13**:42.
87. Valen E, Sandelin A, Winther O, *et al*. Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput Biol* 2009;**5**:e1000562.
88. Ettwiller L, Paten B, Ramialison M, *et al*. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat Methods* 2007;**4**:563–5.
89. Van Helden J, André 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.
90. 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.
91. Ding J, Hu H, Li X. SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data. *Nucleic Acids Res* 2014;**42**:e35.
92. Liu Y, Schmidt B, Maskell DL. An Ultrafast Scalable Many-Core Motif Discovery Algorithm for Multiple GPUs. 2011 IEEE International Parallel & Distributed Processing Symposium. Work Phd Forum 2011, 428–34.