Nucleic Acids Research Advance Access published January 6, 2010

Nucleic Acids Research, 2009, ***1–14***
doi:10.1093/nar/gkp1180

On the detection and refinement of transcription factor binding sites using ChIP-Seq data

Ming Hu[1,2], Jindan Yu[3,4,5,6], Jeremy M. G. Taylor[2,5], Arul M. Chinnaiyan[3,4,5,7,8] and Zhaohui S. Qin[1,2,*]

[1]Center for Statistical Genetics, [2]Department of Biostatistics, [3]Michigan Center of Translational Pathology, [4]Department of Pathology, [5]University of Michigan Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan 48109, [6]Division of Hematology/Oncology, Department of Medicine, Northwestern University, Chicago, Illinois 60660, [7]Department of Urology, University of Michigan Medical School and [8]Howard Hughes Medical Institute, Ann Arbor, Michigan 48109, USA

Received October 26, 2009; Revised and Accepted December 2, 2009

ABSTRACT

Coupling chromatin immunoprecipitation (ChIP) with recently developed massively parallel sequencing technologies has enabled genome-wide detection of protein–DNA interactions with unprecedented sensitivity and specificity. This new technology, ChIP-Seq, presents opportunities for in-depth analysis of transcription regulation. In this study, we explore the value of using ChIP-Seq data to better detect and refine transcription factor binding sites (TFBS). We introduce a novel computational algorithm named Hybrid Motif Sampler (HMS), specifically designed for TFBS motif discovery in ChIP-Seq data. We propose a Bayesian model that incorporates sequencing depth information to aid motif identification. Our model also allows intra-motif dependency to describe more accurately the underlying motif pattern. Our algorithm combines stochastic sampling and deterministic 'greedy' search steps into a novel hybrid iterative scheme. This combination accelerates the computation process. Simulation studies demonstrate favorable performance of HMS compared to other existing methods. When applying HMS to real ChIP-Seq datasets, we find that (i) the accuracy of existing TFBS motif patterns can be significantly improved; and (ii) there is significant intra-motif dependency inside all the TFBS motifs we tested; modeling these dependencies further improves the accuracy of these TFBS motif patterns. These findings may offer new biological insights into the mechanisms of transcription factor regulation.

INTRODUCTION

Accurately locating the transcription factor (TF)–DNA interaction sites provides key insights into the delineation of the underlying mechanisms of transcriptional regulation. By exploiting the fact that binding sites for a specific TF often show sequence specificity, computational prediction of TF binding sites, or motif finding, has become an indispensible tool for functional genomics research. A variety of different software programs have been developed for motif-finding (1–7) [see Tompa *et al.* (8) for a review of this topic].

The input data for computational motif-finding algorithms are DNA sequences believed to be enriched by the TF binding sites, or motifs. Typical sources of the input data are known co-regulated genes (7), phylogenetic conservation (9) or results from functional genomics experimental assays (1,10–12). For the latter, continually evolving high-throughput technologies, from DNA microarray (13,14) to ChIP-chip (15,16) and now ChIP-Seq (17–20), offer rapidly improving opportunities for motif finding.

ChIP-Seq, or chromatin immunoprecipitation (ChIP) (21,22) followed by ultra-high-throughput sequencing, has emerged as a powerful new technology for genome-wide mapping of protein–DNA interactions and histone modifications (17–20). Through direct sequencing of all DNA fragments from ChIP assays, ChIP-Seq can reveal protein–DNA interaction sites across the entire

---

*To whom correspondence should be addressed. Tel: +1 734 763 5965; Fax: +734 615 8322; Email: qin@umich.edu genome, thus building a comprehensive and high-resolution interactome map for DNA-binding proteins of interest.

From past experience, exploiting the quantitative information provided by high-throughput genomic assays allows scientists to develop more effective motif-finding algorithms. Improvements in motif detection have been reported in studies using microarray (10,11) and ChIP-chip (1,12) data. The newly emerged ChIP-Seq technology has demonstrated remarkable sensitivity and specificity in identifying protein–DNA binding loci across the entire genome with high resolution and few constraints. In excess of 10 000 DNA sequences are routinely being identified as candidates that potentially harbor protein–DNA interaction sites of interest. Such information provides an exciting new venue for motif discovery and refinement.

A *de novo* motif search is a natural follow-up to the identification of ChIP-enriched regions. Not only it is required when the TF binding motif pattern is unknown, but it is also important in cases where TF and its canonical binding motif pattern have been established. After all, it is reassuring to be able to rediscover the known TFBS motif pattern from the input sequences. More importantly, most of the known TF binding motif patterns stored in the various TF binding motif databases or reported in the literature are defined based on limited numbers of experimentally verified TF–DNA interaction sites. Many of these motif patterns could be inaccurate due to limited experimental data. Performing a *de novo* motif search on a large number of ChIP-Seq binding sites has the potential to refine the motif patterns of the TFBS.

While a variety of methods that attempt to identify ChIP-enriched genomic regions from ChIP-Seq experiments (also called 'peak calling') have been described (23–31), little has been developed utilizing ChIP-Seq data for motif finding.

Probability model-based *de novo* motif finding algorithms such as MEME have demonstrated a high level of sensitivity and specificity (2–5,32–36). However, since these methods were developed when only a handful of motif-enriched sequences were available, they do not work well when analyzing large sets of sequences identified by ChIP-Seq. There are at least two limitations that affect their performance: (i) the requirement for going through all bases in all sequences using time-consuming iterative procedures means that these methods do not scale well for the analysis of large sets of sequences generated from ChIP-Seq; (ii) existing methods, which only consider sequence data, are unable to fully utilize the rich information produced from ChIP-Seq. Overlooked information includes the sequencing depth along the ChIP-enriched regions and the overall significance of ChIP-enrichment for each sequence. 'Sequencing depth' refers to the number of ChIP DNA fragments that cover each base. Currently, a common practice for performing motif finding on ChIP-Seq data is to use existing motif-finding tools on a subset of all sequences (e.g. the top 500 sequences or top 10% of all such sequences) (25,26). This is sub-optimal because the small sample size may lead to an inaccurate motif pattern and the selection of top sequences tends to result in motif patterns with inflated information content.

We believe that a more desirable approach is to develop algorithms that can utilize all of the sequence information generated from ChIP-Seq. Not only will this strategy result in the identification of more accurate motif patterns, but also the dramatically increased number of *in vivo* binding sites revealed by ChIP-Seq permits the use of probability models that are more sophisticated than the commonly used product multinomial models (34) for characterizing the motif pattern.

To address these limitations and fully exploit the information provided by ChIP-Seq experiments, we develop a novel model-based motif-finding algorithm named the Hybrid Motif Sampler (HMS). It is specifically designed for ChIP-Seq data and utilizes *all* ChIP-enriched regions identified from ChIP-Seq experiments. In this algorithm, we propose a new probability model that considers both DNA sequence and sequencing depth information that is available from ChIP-Seq experiment. It also allows inter-dependent positions within a motif to be identified. In addition, we propose a novel hybrid searching scheme to significantly expedite the iterative procedure. Our algorithm is capable of processing tens of thousands of sequences and is much faster than the established *de novo* motif-finding tools such as MEME.

MATERIAL AND METHODS

The statistical model

Let $\boldsymbol{R} = (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_J)$ denote a set of J sequences (e.g., DNA sequences in ChIP-enriched regions identified by ChIP-Seq) of length $L_1, \ldots, L_J$. We initially assume that every sequence $\boldsymbol{R}_j$ contains exactly one binding site. In addition, the vector that is formed by the start locations is referred to as the alignment variable, denoted as $\boldsymbol{A} = (a_1, \ldots, a_J)$ where $1 \leq a_j \leq L_j - w + 1, j = 1, 2, \ldots, J$. Here, w is the motif width and is assumed to be known. Given $\boldsymbol{A}$ and w, the aligned sequence motif can be represented by a four by w matrix. Each column of the matrix stores the frequency counts of the four types of nucleotides. Liu *et al.* (34) proposed the *product-multinomial model* to model the nucleotide preferences shown in such matrices. The product-multinomial model has been widely used in EM-based (4,32) and Gibbs sampler-based (3,33,35) motif finding algorithms. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w)$, $\boldsymbol{\theta}_i$ represent the nucleotide preference at the i-th position of the motif and let the probability vector $\boldsymbol{\theta}_0$ represent the nucleotide preference for non-motif positions in these sequences. Each of the $\boldsymbol{\theta}_i, i = 0, 1, 2, \ldots, w$ is a probability vector of length four. For notational simplicity, we use integers 1, 2, 3 and 4 to represent the four types of nucleotides A, C, G and T.

For *de novo* motif finding, the parameter of main interest in our model is the alignment variable $\boldsymbol{A}$.

Lawrence *et al.* (3) proposed a Gibbs sampler-based approach in which the posterior distribution for alignment a_j can be expressed as:

$$\begin{aligned} p(a_j = l|\boldsymbol{\theta_0}, \boldsymbol{\Theta}, \boldsymbol{R_j}, \boldsymbol{A_{-j}}) &\propto \prod_{k=1}^{4} \theta_{0k}^{h_k(\boldsymbol{R_j})} \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1})} \\ &\propto \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1})} \end{aligned} \qquad \mathbf{1}$$

where $\boldsymbol{A_{-j}} = (a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_J)$ and the functions $h_k(), k = 1, 2, 3, 4$, returns the number of nucleotides of type k.

For $\boldsymbol{\theta_0}$ and $\boldsymbol{\Theta}$, as an alternative to sampling them from posterior conditional distributions as in a standard Gibbs sampler, one can use the predictive updating technique (34) to integrate them out. Alternatively, the posterior means can be used to approximate the updated parameters during iteration. More details of these strategies can be found in Liu *et al.* (34).

Allowance for some sequences that do not contain the motif

In the model above, we assume that every sequence $\boldsymbol{R_j}$ contains exactly one motif. However, this is not the case in real data. To increase specificity, as most motif-finding algorithms have done, it is highly desirable that we generalize the method to allow some sequences to be motif-free. We introduce a binary indicator variable I_j, where $I_j = 1$ indicates that $\boldsymbol{R_j}$ contains at least one motif, and $I_j = 0$ otherwise. In the algorithm, I_j is set to 1 if the average of likelihood ratios observing the motif in the sequence $\boldsymbol{R_j}$, denoted as z_j, is greater than 1. i.e.

$$z_j = \frac{1}{L_j - w + 1} \sum_{l=1}^{L_j-w+1} \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1})}, I_j = I_{\{z_j \geq 1\}} \qquad \mathbf{2}$$

After updating I_j, we only conduct motif search on the sequences with $I_j = 1$.

Modeling sequencing depth

The model described in equation (1) assumes that binding motifs are equally likely to occur at all positions in each sequence. This is reasonable when no information beyond the input DNA sequences is considered. However, such a model is no longer sufficient for analyzing ChIP-Seq data since additional information beyond the DNA sequences is available and should be incorporated. In particular, it has been shown that the sequencing depth in each ChIP-enriched region is indicative of the motif location (25,28). Figure S1 in the Supplementary Data shows that the majority of motifs are tightly packed near the peak summit (the location inside each peak with the highest sequence coverage depth), especially for the highly significant peaks.

To capitalize on the extra information provided by ChIP-Seq, we propose adding to the method an informative prior distribution of the motif location based on the sequencing depth. There are multiple ways to assign such priors. The simplest strategy is to make the prior probabilities directly proportional to the sequencing depth in each sequence. However, since sequencing depth is affected by many factors, such as local GC content, using a prior distribution like this may result in 'over fit'. Alternatively, a parametric distribution that approximates the sequencing depth can be used to obtain the prior probabilities. In this study, we set the prior probabilities to be proportional to a discretized Student's *t*-distribution with three degrees of freedom and rescaled such that the prior probabilities form a step function with a fixed step-size (25 bp in this study). The prior probabilities are symmetric and centered at the peak summit (most peak-calling software provides the exact location of the summit). Specifically, the prior probabilities that a motif starts at position l can be expressed as:

$$p(a_j = l) \propto t_3\left(\text{int}\left[\frac{|l + w/2 - s_j| + u/2}{u}\right]\right) \qquad \mathbf{3}$$

Where t_3 is the probability density function of the Student's *t*-distribution with three degrees of freedom, s_j is the location of the peak summit, w is the motif width, u is the step size (25 bp in this study) in the step function and int $[\cdot]$ returns the integer part of a real number. Please see Supplementary Figure S2 for an illustration of the prior probabilities. The reason that we choose Student's *t*-distribution instead of a normal distribution is because it better allows for some motif locations to be far from the peak (the standard deviation of Student's *t*-distribution with three degrees of freedom is 1.73, compared to one for standard normal distribution).

Modeling intra-motif dependency

The classical product-multinomial model assumes that the positions within the motif are independent of each other (37). However, recent studies indicate that some positions of TF binding motifs exert an inter-dependent effect on the binding affinities of TF's (38–41). These findings imply that the commonly used product-multinomial model may be too simplistic in characterizing the binding sites. Models that allow for dependent positions likely will provide a better fit of the data. The significantly increased quantity of motifs identified by ChIP-Seq enables us to consider a more sophisticated model that can take into account the intra-motif dependency.

There have been numerous attempts to incorporate into models the inter-dependency among positions within a motif. King and Roth (42) introduced a non-parametric representation of motifs that allows arbitrary dependencies among positions. Barash *et al.* (43) suggested multiple Bayesian network models to represent dependencies among motif positions. Zhou and Liu (44) proposed a generalized weight matrix model in which a 16-component multinomial model is used to model two dependent positions jointly.

Here, we extend the generalized weight matrix model of Zhou and Liu. To take greater advantage of the abundant sequence information made available by the ChIP-Seq technology, our model allows up to three positions to be inter-dependent.

Detection of dependent positions

Given a set of aligned putative binding motifs, our goal is to identify positions that show inter-dependency. Here, 'inter-dependency' implies that the frequency of certain nucleotide combinations spanning multiple positions deviates from the expected frequency when assuming an independent motif model. As an example, for a pair of positions, if the frequency of a particular dinucleotide, say AC, is much higher or lower than the product of frequency of nucleotide A in the first position and frequency of nucleotide C in the second position, we conclude that the two positions are dependent.

A variety of methods have been proposed in the literature to search for such inter-dependent positions. Barash *et al.* (43) applied machine learning approaches to infer the structure of a Bayesian network that best represents the underlying motif. Zhou and Liu (44) proposed a Metropolis-type iterative procedure to identify pairs of inter-dependent positions. Given the abundant motif data from ChIP-Seq, we implement a comprehensive search strategy to go through all pairs of positions within the motif to determine whether there is evidence of dependency. To be specific, for any two positions i and j among $w(w-1)/2$ possible pairs, we first obtain probability estimates of the 16 dinucleotides assuming either a 16-component multinomial model (dependent) or the product of two four-component multinomial models (independent). Let the number of motifs be represented by M. The term $g_x(r_i)$ represents the number of motifs whose i-th position is occupied by nucleotide x and the term $g_{xy}(r_i, r_j)$ represents the number of motifs whose i-th and j-th positions are occupied by nucleotides x and y, respectively. The probability estimates under the two competing models are $\hat{\eta}_x(r_i) = g_x(r_i)/M$ and $\hat{\eta}_{xy}(r_i, r_j) = g_{xy}(r_i, r_j)/M$, respectively. We then calculate the Hamming distance between the two sets of estimates as

$$d_{ij} = \sum_{x=1}^{4}\sum_{y=1}^{4}\left|\hat{\eta}_{xy}(r_i, r_j) - \hat{\eta}_x(r_i)\hat{\eta}_y(r_j)\right| \qquad \mathbf{4}$$

Under the hypothesis that the two positions are independent, we expect that distance $d_{ij} = 0$, excluding sampling variability; larger d_{ij} indicates stronger inter-dependency between positions i and j. In this study, we designate positions i and j to be dependent if $d_{ij} > 0.2$. The threshold is determined from the empirical null distribution of d_{ij} infer through simulations. More details can be found in the Supplementary Data.

Posterior distribution

We take a Bayesian approach and consider two different models to describe the motif pattern. In the first one, we assume all positions within the motif are independent. There are two sets of parameters in this model: alignment variable $\boldsymbol{A}$ and multinomial distribution probability vector $\boldsymbol{\theta_i}, i = 0, 1, \ldots, w$. The prior distributions for $\boldsymbol{A}$ are multinomial with probabilities defined as in equation (3). Adopting a conjugate prior distribution for each $\boldsymbol{\theta_i}$, which is $Dirichlet(\alpha_{0,1}, \ldots, \alpha_{0,4})$, the posterior probabilities that a motif starts at position l can be expressed as:

$$\begin{aligned} &p(a_j = l|\boldsymbol{\theta_0}, \boldsymbol{\Theta}, \boldsymbol{R_j}, \boldsymbol{A_{-j}}) \\ &\propto I_{\{z_j>1\}}\prod_{i=1}^{w}\prod_{k=1}^{4}\left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1})+\alpha_{0,k}} \qquad p(a_j = l) \end{aligned} \qquad \mathbf{5}$$

As suggested in Liu *et al.* (34), the above conditional distribution can be closely approximated by replacing θ_{ik} by its posterior mean given the current alignment vector $\boldsymbol{A_{-j}}$:

$$\hat{\theta}_{ik} = \frac{h_k(\boldsymbol{r}_{-j,\boldsymbol{A}_{-j}+i-1}) + \alpha_{0,k}}{N_{0[-j]} + \alpha_{0,k}}, N_{0[-j]} = \sum_{s\neq j} I_{\{z_s\geq 1\}} \qquad \mathbf{6}$$

For background (non-motif) regions, it has been shown that employing a Markov model to capture weak dependency in background DNA sequences improves the sensitivity and specificity of motif finding compared to an independent model in equation (1). In this study, we use a third-order Markov model as in Liu *et al.* (2) to characterize the background sequences. Under such a model, the probability of observing DNA sequence fragment $\{r_{s,t}, r_{s,t+1}, \ldots, r_{s,t+w-1}\}$ in the background can be represented by

$$\begin{aligned} P(\text{Background}_{s,t}) &= P(r_{s,t})P(r_{s,t+1}|r_{s,t})P(r_{s,t+2}|r_{s,t+1}, r_{s,t}) \\ &\quad P(r_{s,t+3}|r_{s,t+2}, r_{s,t+1}, r_{s,t})P(r_{s,t+4}|r_{s,t+3}, r_{s,t+2}, r_{s,t+1})\ldots \end{aligned} \qquad \mathbf{7}$$

In this background model, the $3 \times 4^3 = 192$ conditional probabilities are estimated from human promoter sequences downloaded from UCSC genome browser website. The dataset contains 5 kb upstream sequences of annotated transcription starts for all RefSeq genes with annotated 5′-UTRs.

After incorporating these modifications, the complete posterior distribution for $a_j = l$ becomes

$$p(a_j = l|\boldsymbol{\theta_0}, \boldsymbol{\Theta}, \boldsymbol{R_j}, \boldsymbol{A_{-j}}) \propto I_{\{z_j>1\}}\frac{\prod_{i=1}^{w}\prod_{k=1}^{4}\hat{\theta}_{ik}^{h_k(r_{j,l+i-1})+\alpha_{0,k}}}{P(Background_{j,l})}p(a_j = l) \qquad \mathbf{8}$$

In the second model, we consider intra-motif dependency. Within the motif, we assign positions into two disjoint groups: groups of independent positions S and groups of dependent position pairs P where $P = \{(i,j) : d_{ij} > 0.2\}$. By modeling dependent positions jointly, the probability 'matrix' $\boldsymbol{\Theta}$ becomes an amalgamation of vectors of length four (modeling single positions) and vectors of length 16 (modeling pairs of dependent positions). The prior distributions for the two types of $\boldsymbol{\theta_j}$'s are $Dirichlet(\alpha_{0,1}, \ldots, \alpha_{0,4})$ and $Dirichlet(\beta_{0,1,1}, \ldots, \beta_{0,1,4}, \beta_{0,2,1}, \ldots, \beta_{0,4,4})$ respectively. The complete posterior distribution for $a_j = l$ in the dependent model is

$$p(a_j = l|\boldsymbol{\theta_0}, \boldsymbol{\Theta}, \boldsymbol{R_j}, \boldsymbol{A_{-j}}) \propto \frac{I_{\{z_j>1\}} \cdot U \cdot V \cdot p(a_j = l)}{P\ (\mathrm{Background}_{j,l})}$$

$$U = \prod_{i \in S} \prod_{k=1}^{4} \hat{\theta}_{ik}^{h_k(r_{j,l+i-1})+\alpha_{0,k}} \qquad \mathbf{9}$$

$$V = \prod_{i_1, i_2 \in P} \prod_{k_1=1}^{4} \prod_{k_2=1}^{4} \hat{\theta}_{i_1,i_2}^{h_{k_1 k_2}(r_{j,l+i_1-1}, r_{j,l+i_2-1})+\beta_{0,k_1,k_2}}$$

Here the counting function $h_{k_1 k_2}()$, whose argument is a set of positions, counts the frequency of the 16 dinucleotides for a pair of positions within the motif. The above model can be extended easily to allow three-way inter-dependent positions.

Acceleration via prioritized hybrid Monte Carlo

To streamline this motif-finding algorithm in order to handle a large number of input sequences, we develop a prioritized hybrid strategy to increase computation speed with only minimal if any sacrifice in accuracy. Unlike a standard Gibbs sampler where motif alignment variables are sampled stochastically from *all* sequences, only a small proportion, π, of all sequences are subjected to stochastic sampling. For the remaining sequences, we select the alignment variable deterministically by identifying the position that corresponds to the highest probability as given by equation (8) or (9). Since the deterministic approach is much faster than the stochastic one and the proportion π we use is often quite small ($\pi \sim 10\%$), this hybrid strategy is much faster than the standard Gibbs motif sampler (3).

For each iteration, the proportion of sequences undergoing stochastic search is constant, but a different set of sequences is selected each time. We have automated the process of selecting a subset of sequences for stochastic search. All the sequences identified from the ChIP-Seq experiment are rank-ordered according to their ChIP-enrichment. Assume we run N iterations in each Gibbs sampler. In the i-th iteration, we sample a fixed number of $\pi \times J$ sequences from a multinomial distribution $mult(J, p_{i1}, \ldots, p_{iJ})$. At the beginning of the iteration, we use a monotonically decreasing triangle probability distribution, which assigns higher probability to sequences with higher ChIP-enrichment. As the iteration proceeds, the slope of the triangle gradually becomes flatter so that the oversampling of higher ChIP-enriched sequences diminishes. In the last iteration, the distribution becomes uniform. For the i-th iteration, we have

$$p_{ij} \propto c_{ij} = J - j + 1 - \frac{J/2 - j + 1}{N - 1} \times (i - 1), \qquad \mathbf{10}$$
$$i = 1, \ldots, N; j = 1, \ldots, J.$$

Implementation

We have developed a software program that implements the algorithms described in this manuscript. The HMS program is a Gibbs sampler type iterative procedure. To reduce the possibility that the Markov chain converged to a local mode, we run multiple Markov chains and choose the motif pattern that corresponds to the highest likelihood as the final motif pattern. The number of parallel chains and the number of complete iterative cycles within each chain are specified by users. Within each chain, the iterative procedure can be broken down into three steps. In the first step, we use a traditional product multinomial model in which all positions are assumed independent of each other. We further assume every sequence contain one motif. In the second step, we again assume all positions are independent, but we allow some sequences to be motif-free. In the final step, we adopt the generalized motif model that allows intra-motif dependency. The HMS program, including the source code is freely available at http://www.sph.umich.edu/csg/qin/HMS/.

Performance evaluation using simulated data

In the simulation study, we are interested in evaluating the performance of various *de novo* motif finding algorithms from two perspectives: first, the number of times a program successfully detects the motif inserted into each of the 100 simulated datasets; second, the accuracy of the inferred motif pattern given that the motif has been found.

For the former, since we know the true location of all inserted motifs in the simulated datasets, we are able to directly verify whether each motif site predicted by the testing software is correct. Within each simulated dataset, we declare that the inserted motif is found if the proportion of sequences in which the program correctly identifies the true motif location is greater than 20%.

For the latter, we measure the accuracy of an inferred motif pattern by calculating the average Hamming distance between the true probability matrix $\boldsymbol{\Theta}$ and its prediction denoted as $\hat{\boldsymbol{\Theta}}$:

$$h = \frac{1}{w} \sum_{i=1}^{4} \sum_{j=1}^{w} \left| \theta_{ij} - \hat{\theta}_{ij} \right| \qquad \mathbf{11}$$

Small h indicates close resemblance of the predicted motif pattern to the truth.

Performance evaluation using real data

Given a set of sequences identified by ChIP-Seq, we want to discern which *de novo* motif-finding algorithm produces a more accurate motif pattern. Since the exact true motif pattern is unknown, we use motif enrichment as the criterion. We assume that among multiple motif patterns, the one that is most enriched in the ChIP-Seq-identified regions relative to random controls is closest to the true motif pattern.

We use a cross-validation scheme to assess motif enrichment. The original dataset is equally divided into halves: a training set and a testing set. The input sequences are restricted to within 200 bp in length and centered at the peak summit (≤ 100 bp toward each side of the peak summit). For the testing set, we create a control set composed of randomly selected DNA promoter sequences (within 5 kb upstream of the transcription start site) as in Zhou and Liu (44) matched by number of sequences and length of each sequence. We run each motif-finding program on the training set to identify the motif pattern, and then utilize this pattern to scan both the testing and the corresponding control sets to assess how many sequences contain the motif. We employ a set of significance thresholds and calculate the corresponding empirical false discovery rate (FDR) (45) and motif enrichment, as measured by chi-squared test statistics for a 2 × 2 contingency table. The empirical FDR is estimated by dividing the number of control sequences that contain the motif by the number of testing sequences that contain the motif. We repeat the scheme five times for each dataset and report the average test statistics corresponding to each FDR level.

We plot the curves of the empirical FDR versus the chi-squared test statistics when the empirical FDR is between 0 and 0.2. To accomplish this, we equally divide the empirical FDR into ten consecutive windows and calculate the mean of the chi-squared test statistics from five cross validations (when the corresponding empirical FDRs fall into the same window). Since the curve representing the most enriched motif pattern will be the highest, we use area under the curve (AUC) as a quantitative assessment of the overall motif enrichment. Higher AUC indicates further motif enrichment.

Estrogen receptor ChIP-Seq experiment on MCF7 cells

To test the algorithms in a real setting, we have conducted a ChIP-Seq experiment to survey genome-wide binding of estrogen receptor (ER) on the MCF-7 breast cancer cell line. ER is a hormonal TF that, when liganded by estrogen, binds specially to estrogen response elements (ERE) and plays a critical role in breast cancer development. Identifying ER target genes and refining the ERE motifs are thus of significant interest. A brief description of the experimental protocol is shown in the next paragraph. More details can be found in the Supplementary Data.

Briefly, MCF-7 cells were grown in RPMI media supplemented with 10% FBS to 50% confluence. The cells were then hormone-starved for three days prior to treatment of the vehicle control or 10 nM β–estradiol for 45 min. The cells were then harvested for ChIP analysis using an antibody against estrogen receptor (ER)-α (sc-543x, Santa Cruz) or against IgG. The ChIP-enriched DNA was evaluated for significant enrichment of positive control genes and then subjected to ChIP-Seq sample preparation and short-read sequencing using Illumina Genome Analyzer (Illumina Inc., San Diego, CA, USA) following the manufacturer's protocols. The raw sequencing images were analyzed using the Illumina analysis pipeline, and the sequencing reads were subsequently aligned to the human reference genome (NCBI v36, hg18) using ELAND software (Illumina Inc.), producing sequencing reads of 35 bp. Only sequencing reads that are uniquely mapped to the human reference genome with up to two mismatches were included for further analysis as delineated in this study. We have submitted ER ChIP-Seq data (raw and processed) into the GEO database; the accession number of this dataset is GSE19013. We used the HPeak software program, a HMM-based peak calling program developed by our group, to define the ChIP-enriched regions. Details of the HPeak software program can be found in the Supplementary Data.

RESULTS

Simulation study

Independent motif models. The goal of this simulation study was to evaluate the ability of HMS to identify the correct motif patterns. We use the default setting for HMS which adopts the informative prior and allows intra-motif dependency. For comparison, we also tested a simpler version of HMS that assumes all positions are independent. In addition, we applied two established motif-finding software tools, MDscan (1) and MEME (4) on the same sets of simulated data. Following the simulation scheme of Liu *et al*. (1), four motif models were manually created (Supplementary Table S1A), representing two different motif widths (8 bp and 16 bp), and two different degrees of conservation measured by information content (1.42 and 0.93). The information content is defined as:

$$\frac{1}{w}\sum_{i=1}^{w}\sum_{j=1}^{4}p_{ij}\log_2(4p_{ij}) \qquad \mathbf{12}$$

where p_{ij} is the proportion of base j at the motif position i. Information content ranges from 0 to 2, reflecting the weakest to the strongest motifs. Finally, two different motif abundance schemes (Supplementary Table S1B) were considered for a total of eight combinations in the simulation study. The eight simulation settings covered a wide range of scenarios. The combination of short motif width, weak motif information content and low motif abundance was the most challenging.

For each setting, we simulated 100 test datasets. Each dataset contains 3000 sequences of 200 bp in length. To mimic real human data, all the sequences were generated from a third-order Markov model with parameters estimated from the collection of 5 kb promoter sequences of annotated genes in the human genome. Hypothetical motifs were generated from product multinomial models with specified length and information content. The proportion of sequences that contained a motif followed one of the two abundance schemes mentioned in the previous paragraph. We assumed that each sequence contained at most one motif.

We next derived the empirical distribution from real ChIP-Seq data of CTCF and NRSF of the motif start locations in a 200 bp window centered at the peak summit. We strategically inserted the motifs in these sequences following this empirical distribution. As a consequence, the motif locations were biased toward the

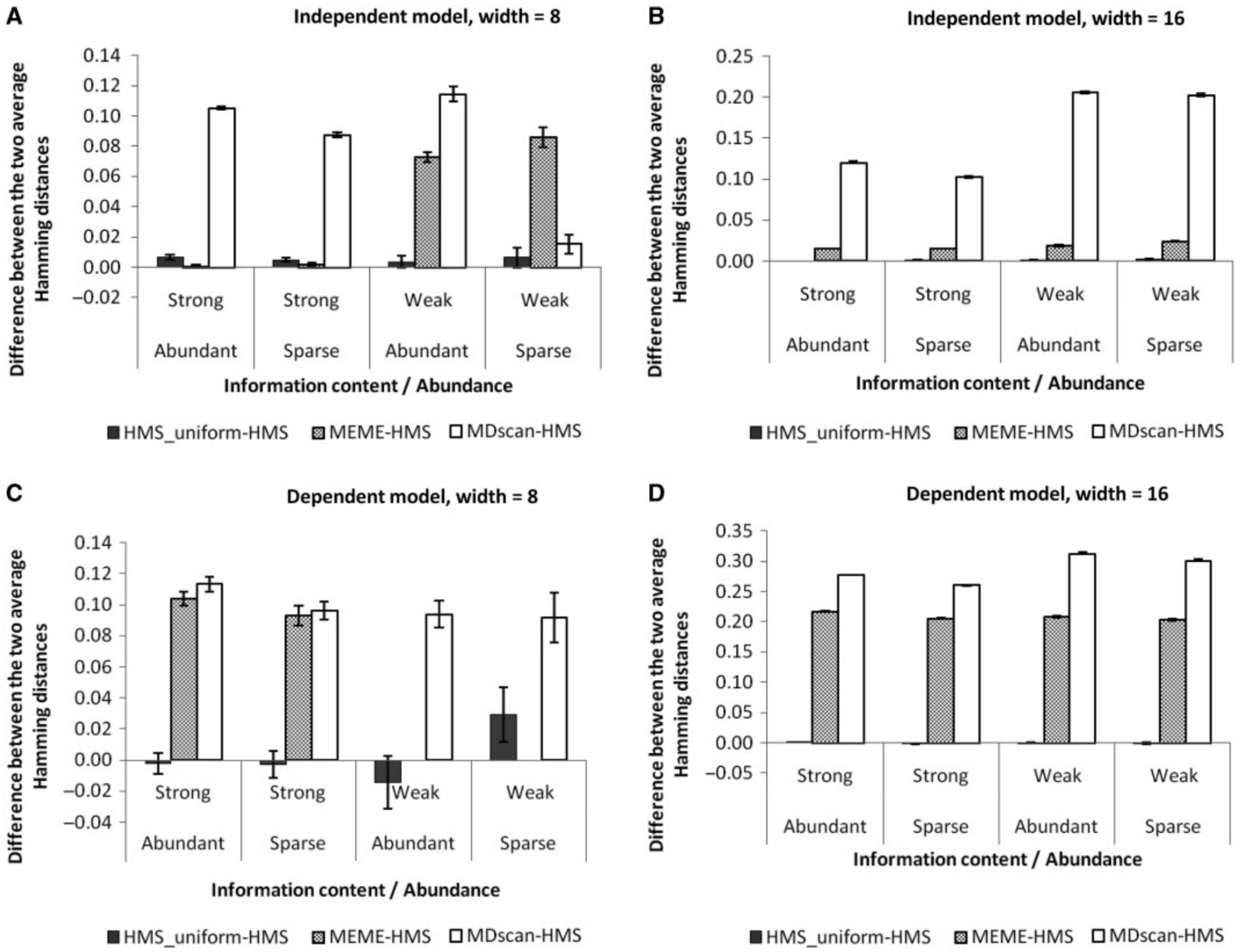

Figure 1. Performance comparison on simulated data with independent and dependent motif model. The y-axis represents the difference between two sets of average Hamming distances resulted from two different motif finding methods. The error bars represent the standard deviation of the difference between two sets of average Hamming distances across 100 simulated datasets. (**A**) Independent, motif width = 8 bp. (**B**) Independent, motif width = 16 bp. (**C**) Dependent, motif width = 8 bp. (**D**) Dependent, motif width = 16 bp.

center of the sequence, which was assumed to be the location of the peak summit.

We applied MDscan, MEME and HMS to every dataset. Two versions of HMS were used in the comparison. One assumed an informative prior (proportional to a discretized and rescaled Student's t-distribution with three degrees of freedom) that favored motif start locations near the peak. The other, denoted as HMS_uniform, assumed a uniform prior for the motif start location throughout the genome. As described in the 'Materials and Methods' section, we used the successful motif detection rate and the accuracy of predicted motif pattern as measurements of performance.

For the motif detection rate, both versions of HMS achieved perfect results in all eight simulation settings. MEME and MDscan achieved perfect results in six and four settings, respectively. MEME achieves equal or higher detection rate than MDscan in all but one setting (Supplementary Table S2A).

We next compared performance on motif pattern prediction accuracy. The prediction accuracy is defined as the average Hamming distance between predicted and true $\boldsymbol{\Theta}$ for each method and each dataset. See equation (11) in the 'Material and Methods' section for the expression for the average Hamming distance. To compare methods, within each simulation setting, we performed a paired t-test between the average Hamming distances obtained using HMS and that of a competing method (HMS_uniform, MEME and MDscan). Among the 100 datasets, we only considered the ones in which all methods successfully detected the right motif. Significantly smaller average Hamming distance (P-value < 0.01) was observed in six out of eight simulation settings when comparing HMS to MEME, and in seven out of eight settings when comparing HMS to MDscan (Figure 1A, 1B and Supplementary Table S2A). In addition, we found that adopting the informative prior for the proposed HMS method results in more accurate motif pattern prediction

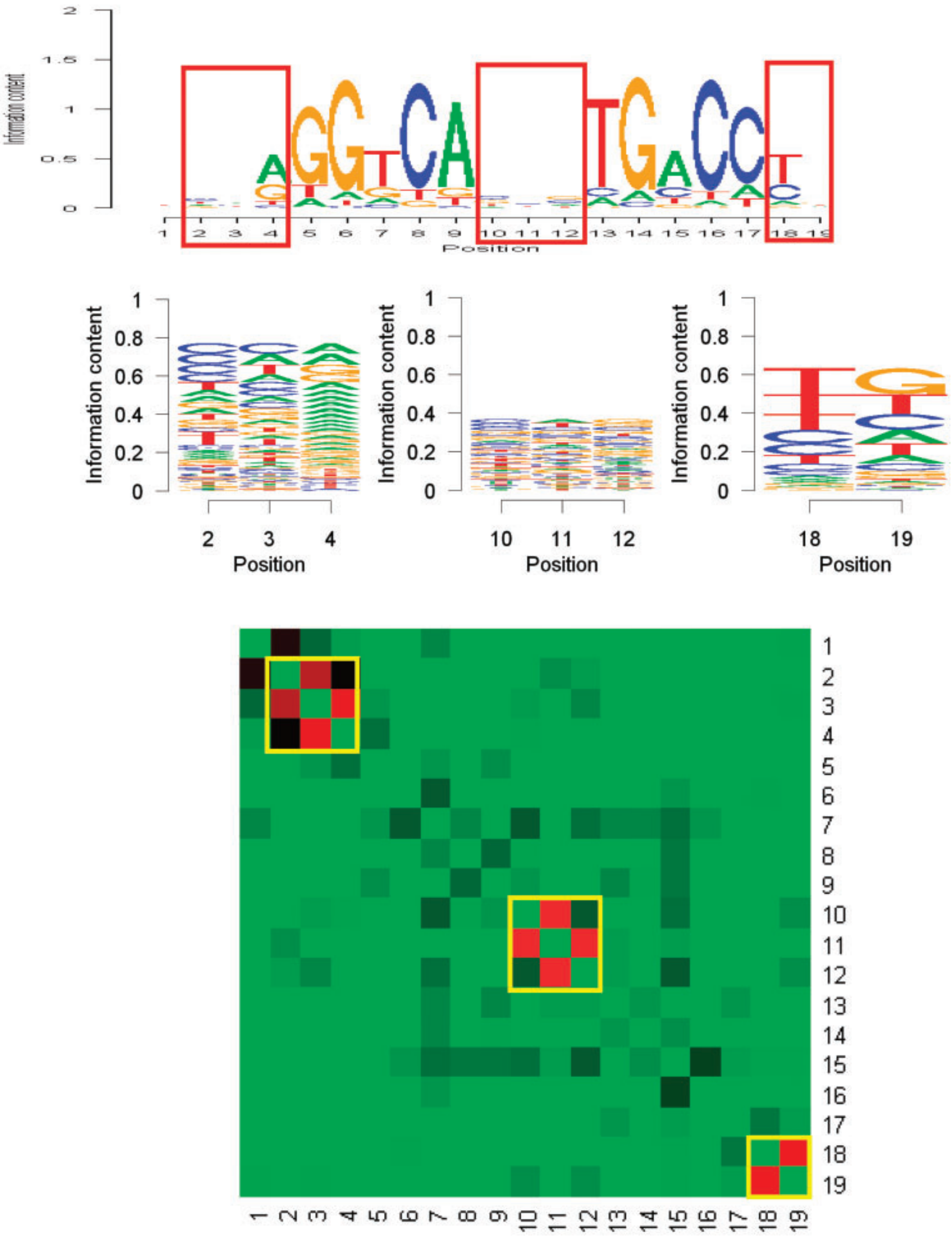

Figure 2. Illustration of the unbiased exhaustive survey of all pairs of positions within the ER motif to assess the strength of their dependency. The differences in Hamming distance between the independent and dependent models are plotted in a heatmap. Larger differences (in dark red color) indicate higher dependency. Dependent triples: position 2, 3 and 4, position 10, 11 and 12. Dependent pairs: position 18 and 19. Dependent positions are illustrated in the box on the logo plot and the heatmap. The logo plots are generated using R package 'seqLogo'. The subgraphs of multi-nucleotide logo plots were generated using a program that we modified from SeqLogo (please see Section 5 in the Supplementary Data for more details). To make the logo plots more readable, we changed the range for y-axis from 0–2 to 0–1 in the subfigures for multi-nucleotide logo plot.

in all eight simulation settings than when using the uniform prior (Supplementary Table S2A).

Inter-dependent motif models. We next conducted simulation studies to evaluate the performance of HMS when some positions within the motif showed inter-dependency. In our simulation, dependency was added to two pairs of positions in the 8 bp motif model and four pairs of positions in the 16 bp motif model. The joint distribution of the pairs was taken from the one predicted for position pair (1,2) in the E2F motif in Zhou and Liu (44) [as shown in Figure 2(b) in the original paper, reproduced in Supplementary Table S3].

In terms of motif detection, both versions of HMS achieved perfect results in five out of the eight simulation settings. MEME and MDscan achieved perfect results in four and two settings respectively. Furthermore, HMS and HMS_uniform reported higher detection rates compared to MDscan and MEME in all simulation settings. Our results also suggest that the HMS method assuming informative prior performed better than the HMS method assuming non-informative prior (Supplementary Table S2B).

When comparing motif pattern prediction accuracy, paired *t*-tests showed that the average Hamming distances between the true and predicted probability matrix $\boldsymbol{\Theta}$ were significantly smaller for HMS than MEME and MDscan in all testable simulation settings (MEME did not identify the correct motif in any dataset under two simulation settings; MDscan only identifies the correct motif in two out of 100 datasets under one simulation settings. Therefore no paired *t*-test is performed for those simulation settings). The performance was similar between the two versions of HMS (Figure 1C, 1D and Supplementary Table S2B).

Real data

To further evaluate the performance of HMS, we tested it along with MDscan and MEME on four real ChIP-Seq datasets. The first three datasets, namely NRSF (neuron-restrictive silencer factor) (18), STAT1 (signal transducer and activator of transcription protein 1) (19), and CTCF (CCCTC-binding factor) (17), are publically available. The ER dataset, however, is newly generated for this study. The details of these four datasets can be found in Table S4A and the Supplementary Data.

Intra-Motif dependency

It is well known that some positions of TF binding motifs exert an inter-dependent effect on the binding affinities of TFs (38–41). However, due to the scarcity of the motifs identified for each TF, it is difficult to detect those dependent positions based solely on the limited motif sequence data. With the introduction of the ChIP-Seq technology, significantly more motif sequences can now be identified, which gives us unprecedented opportunity to identify dependent positions. Using the exhaustive search strategy we outlined in the 'Material and Methods' section, we surveyed the four ChIP-Seq datasets used in this study: NRSF, STAT1, CTCF and ER. The Hamming distance between two probability vectors—$\{\theta_i, i = 1, \ldots, 16\}$ and $\{\theta_i\theta_j, i = 1, \ldots, 4, j = 1, \ldots, 4\}$ were presented in heatmaps (Figure 2 and Supplementary Figure S3). The two sets of probabilities of the 16 dinucleotides were estimated under the independent and dependent models respectively. Larger distance indicated higher dependency. Using the Hamming distance of 0.2 as the threshold, the number of dependent position pairs in the motif ranged from three to five in the four real datasets we studied (Supplementary Table S5). These pairs formed two triplets in NRSF and CTCF motifs, one triplet and one pair in the STAT1 motif and two triplets and one pair in the ER motif. In particular, we found that positions 14 and 15 in the CTCF motif show exceptionally strong dependency. The frequency of dinucleotides AC and GG in these positions were below what would be expected if they were independent. Similarly, the frequency for dinucleotides AG and GC exceeded expectations. The difference in dinucleotide frequencies between independent and dependent motif models exceeded 0.1 in all four relevant cells in the four by four table (Supplementary Table S6E). For other dependent position pairs we identified, their dinucleotide frequencies were summarized in Supplementary Table S6.

An interesting question is that, at position pairs that show significant inter-dependency, whether any particular dinucleotide displays significant enrichment or depletion. To address this, in the 16 dependent position pairs identified from the four motifs, the observed dinucleotide frequencies were compared with the expected frequencies under the assumption that the two positions are independent. We noticed that some dinucleotides, such as TG, CA and AG are over-represented, whereas some dinucleotides, such as CG and TA, are under-represented (Supplementary Figure S4). We found that the overall dinucleotide preference pattern observed is consistent with what has been reported in the literature (46).

Although our search strategy considers all pairs equally, we found that the strongest intra-motif dependency occurred at pairs of adjacent positions (Figure 2 and Supplementary Figure S3). All 16 dependent position pairs we identified in the four motifs were adjacent. This is not surprising given the strong dependency in neighboring positions of DNA sequences. We also found that strong intra-motif dependency often occurred in the so-called 'gap' positions where the motif pattern appeared to be 'weak' according to single-column motif model (e.g. positions 10, 11 and 11 and 12 in the ER motif).

TFBS motif profile comparison

Since both HMS and MDscan were able to rapidly process tens of thousands of DNA sequences without sacrificing much computation time, we fed the entire set of ChIP-enriched regions into these two programs. In this comparison, we only used the top 500 sequences as input for MEME, since this program was not optimized to analyze large numbers of DNA sequences. Next, we applied MAST (47), a motif scanning software that is a companion to MEME, to scan the remaining sequences using the motif pattern identified by MEME. This is a commonly used strategy in motif analysis (26). We also included motif patterns either from the literature [CTCF motif from Kim *et al.* (48)] or from MatBase (Genomatix, Software GmbH, Munich, Germany) for comparison. We used two different versions of HMS in our analysis: the default setting allowing dependency among positions in the motif and HMS_ind assumed all positions are independent. Informative prior for alignment variable $\boldsymbol{A}$ is used in both versions of HMS.

Although the four TFs and their binding motifs were quite diverse, the motif pattern identification results were remarkably consistent. The results from the ER dataset are presented in Figure 3. Results from the three publicly available ChIP-Seq datasets can be found in Figure S5–7 in the Supplementary Data. Inspired by the logo plot (49), we have developed a new plot that can be used to visualize the dinucleotide and trinucleotide motif pattern (Figure 2 and Supplementary Figure S3). This is achieved by modifying the SeqLogo package found in the BioConductor open source software package. More details can be found in the Supplementary Data.

Figure 3A showed that *de novo* motif patterns identified by MEME and HMS from the ER ChIP-Seq dataset. Both patterns were similar to the ER motif stored in MatBase. However, the motif pattern identified by HMS was relatively less conserved (average information content: HMS: 0.64, MEME: 0.71, Genomatix V$ER01: 1.00, Genomatix V$ER02: 1.03, Genomatix V$ER03: 0.89) but more palindromic (reverse compliment) than the other motif patterns (Hamming distance between the two 6-mer half sites after one half site was converted to its reverse complement: HMS: 0.09, MEME: 2.57, Genomatix V$ER01: 4.00, Genomatix V$ER02: 2.18, Genomatix V$ER03: 2.53). The results are encouraging since it is well known that ER binds as a homo-dimer so a palindromic pattern is expected in its TFBS motif.

An intriguing question is if dinucleotides also exhibit the palindromic attribute. Among the five dependent position pairs that HMS identified in the ER motif, position pairs 3–4 and 18–19 were especially well-positioned to serve as a test case for the presence of this palindromic attribute. This is because they are located at the two ends of the ER half sites and do not overlap. We found that the 16 dinucleotide frequencies for positions 3–4 matched almost perfectly with the corresponding dinucleotide frequency at positions 18–19 after reverse compliment transformation (Supplementary Table S7). That is, we did observe dinucleotide dependency at the two ends of the ER motif that exhibited palindromic attribute. This led us to hypothesize that the palindromic property, a hallmark of homer-dimer TF binding motifs, can also be found in the dinucleotide level.

We did not include MDscan in our comparison since MDscan was unable to consistently identify the consensus ER motif pattern. In Figure 3B, we plotted the chi-squared test statistics that measured the motif enrichment at different levels of the empirical FDR. Comparing AUC, we found that the motif patterns identified by MEME and HMS showed much higher AUC than the known motif patterns stored in MatBase. We believe that the dramatically increased number of binding sites identified by ChIP-Seq contributed to the refinement of the motif pattern. MEME and a simplified version of HMS (which used an independent mono-nucleotide model, referred as HMS_ind) exhibited a similar result. AUC for HMS, which allowed up to three-way interdependency, was 16.7% higher than MEME (Supplementary Table S8). The improvement is statistically significant when we repeated the cross-validation steps 100 times and compared the AUCs from HMS and MEME using a paired *t*-test (P-value < 1.0*e*-5). We also compared the proportions of ChIP-enriched sequences that contain each of the ER motif patterns shown in Figure 3A. We found that, under the two empirical FDR levels (0.05 and 0.1), the proportion of motif pattern defined by HMS is higher than that from HMS_ind (by 12.95% and 8.07%, respectively). Comparing HMS to MEME under these empirical FDR levels, the proportion of motif pattern defined by HMS again is higher (by 19.52% and 9.20%, respectively). These differences are again significant (P-value < 1.0e-5) when verifying with paired *t*-test comparing results from 100 cross-validations. In addition, we found that proportions of motifs reported by HMS, HMS_ind and MEME are much higher than those found in the MatBase (Supplementary Table S9).

Among the other datasets (NRSF, STAT1 and CTCF), HMS and MEME consistently identified the consensus motif patterns in all trials. MDscan was able to consistently identify only the NRSF motif, but not the ones for the other two datasets. Again, we found that the motif patterns identified by these *de novo* motif-finding tools were more enriched than known motif patterns found in the literature or MatBase. Motif patterns defined by HMS consistently showed higher enrichment and resulted in higher AUC than MEME (Supplementary Figures S4–6, Table S8). Motif patterns defined by HMS are consistently found in more ChIP-enriched sequences than those defined by HMS_ind and MEME at the same empirical FDR levels (Supplementary Table S9). The performance differences are significant except for the STAT1 motif.

Comparison to ChIP-chip data

In order to confirm that the higher enrichment of the motif identified by HMS on ChIP-Seq data was not platform-dependent, we compared an independent set of testing and control sequences using ChIP-chip. Not only the technology is different, but also the cells and antibodies used. Detailed information about these datasets can be found in Table S4B and the Supplementary Data.

Despite all the differences, we found that the ER motif pattern identified by HMS from ChIP-Seq data once again exhibited significantly higher enrichment than those of HMS_ind and MEME (Figure 3C): the improvements of AUC were 17.5%, and 57.4%, respectively (Supplementary Table S8). These differences are statistical significant (P-value < 1.0e-5). Similar plots and AUC comparisons performed on the other three datasets—NRSF, STAT1 and CTCF—showed comparable patterns (Supplementary Figures S4–6, Table S8). These findings support that the motif pattern identified by HMS has a higher accuracy.

Computation time

All computation was performed on Dell PowerEdge 1950 compute nodes with 2.83 GHz CPU processors and 8 GB RAM. To compare the computation time required for each algorithm, we selected the top 500, 1000, 1500, 2000, up to 5000 sequences identified from the NRSF ChIP-Seq data and fed them into the three motif-finding programs—MDscan, MEME and HMS. We found MDscan to be the fastest, with HMS a close second. Computation time increased linearly with the number of sequences for MDscan and HMS; and both were much faster than MEME. The differences are quite dramatic. For real data, computation times for HMS ranged from 0.4 h (NRSF data) to about 2.5 h (CTCF data). However,

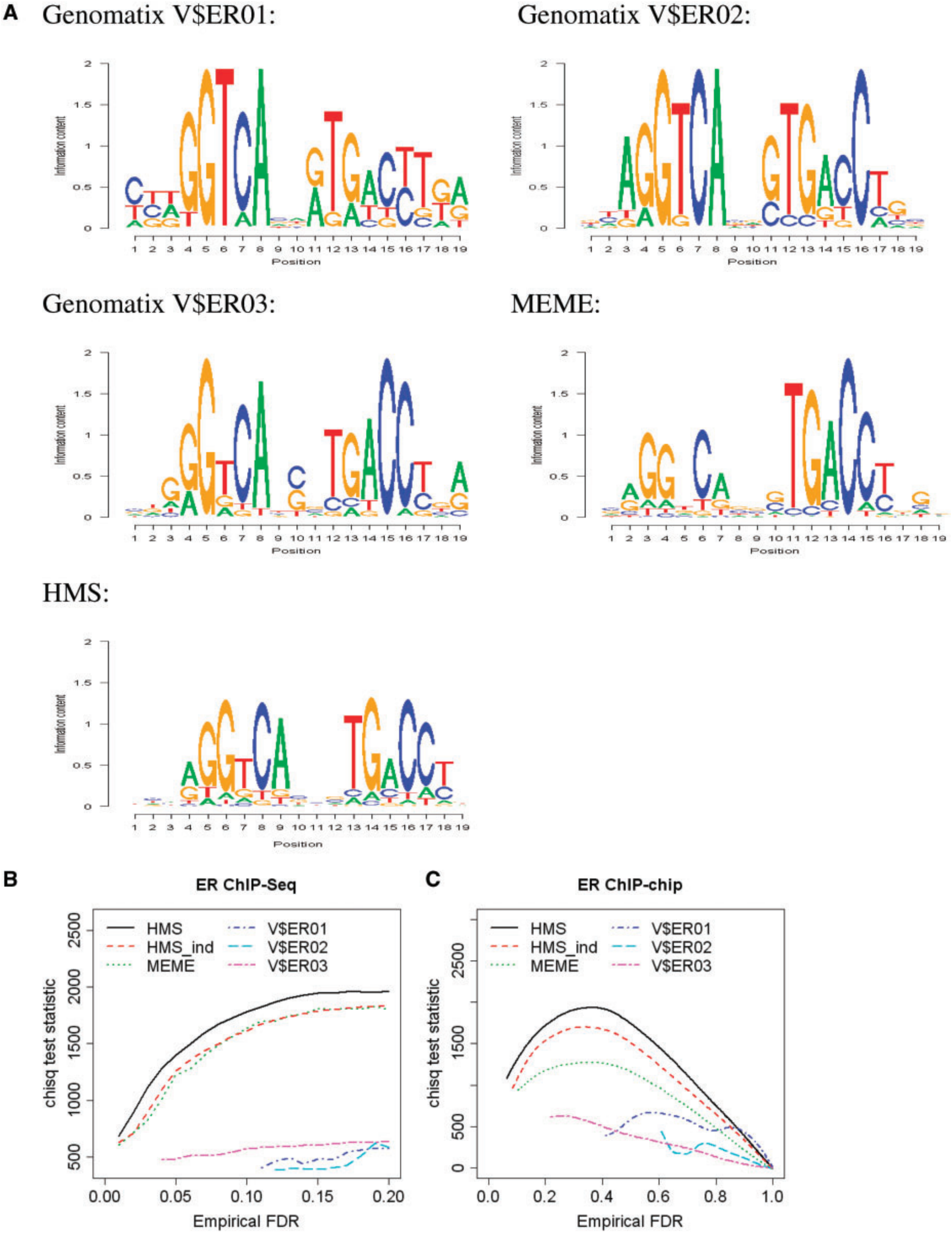

Figure 3. Comparison of ER motif patterns identified by different *de novo* motif-finding tools, as well as known motif patterns stored in the MatBase (Genomatix Software GmBH, Munich, Germany). (**A**) Logo plots (49) of motifs identified by various motif-finding programs as well as the ones stored in the MatBase. The logo plots are generated using R package 'seqLogo'. (**B**) Comparison of motif enrichment in ChIP-Seq for six different motif finding strategies using cross validation. Training sets, testing sets and control sets were generated following the scheme described in the 'Materials and Methods' section (see 'Performance evaluation using real data'). (**C**) Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq. In order to obtain a smooth curve when plotting empirical FDR versus chi-squared test statistics, we applied kernel smoothing using an R function smooth.spline().

since all parallel chains are independent, computation time can be reduced to one tenth if using a multi-processor computing cluster. In contrast, MEME takes much longer; from 13 h (NRSF data) to more than 23 days (CTCF data, job aborted after 23 days of running).

DISCUSSION

The newly emerged ChIP-Seq technology is capable of comprehensively revealing protein–DNA interacting sites across the entire genome with high resolution, which presents both opportunities and challenges for the identification of TFBS motif patterns. Increasing the number of input sequences allowed us to define TFBS motif patterns more accurately. However, most of the existing motif-finding programs such as MEME are not optimized to analyze the large number of input sequences that are generated from ChIP-Seq experiments. In this manuscript, we introduce HMS, a novel computational algorithm, specifically designed for TFBS motif discovery from ChIP-Seq data. It combines stochastic sampling with deterministic optimization in an iterative procedure. The assignment of sequences to these two treatments was dependent on the ranks of the ChIP-enrichment of those regions. This prioritized hybrid Monte Carlo strategy allows us to rapidly analyze tens of thousands of input sequences and produces an accurate estimate of the motif pattern. Our algorithm has the additional advantage of leveraging sequencing depth within each region to aid motif identification. Since the shape of sequencing depth is indicative of likely loci of the motif, using an informative prior gives HMS greater capability to identify weaker motifs than it could otherwise, a clear advancement.

In addition, using HMS we found that there is substantial intra-motif dependency among selected pairs of positions. We identified 16 highly significant position pairs within the NRSF, STAT1, CTCF and ER motifs. All of these position pairs are adjacent to each other, some form triplets. In particular, we noticed a position pair (14 and 15) in the CTCF motif that displays exceptionally strong dependency in which dinucleotides AG and GC are far more frequent than AC and GG at these two positions. Interestingly, we found that dinucleotides at dependent position pairs in the ER motif also exhibit palindromic property, a hallmark for binding motifs of homer-dimer TFs. Using both simulated data and real data, we showed that incorporating dependent positions in a motif model offers further improvement in detecting and characterizing the underlying TF binding motif patterns.

Currently, most *de novo* motif searches on sequences identified by ChIP-Seq are conducted on a subset of all available sequences. This is because searching through the full set of thousands, or even tens of thousands, of input sequences using existing motif-finding tools is extremely time-consuming. Our simulation study showed that this strategy, while convenient, has increased the likelihood of missing the true motif patterns. Further, the probability matrix $\boldsymbol{\Theta}$ inferred with this strategy are often less accurate. In contrast, HMS allows us to analyze the full set of input sequences within only a fraction of the computational time required for existing *de novo* motif-finding tools like MEME. In this study, stochastic search was performed on the top 10% of all sequences. This proportion is adjustable by users. We have experimented increasing or decreasing the 10% cutoff and found that these changes made little difference in the performance of HMS. When applied to multiple real ChIP-Seq datasets, we found that the motif patterns identified by HMS tend to be more enriched than motifs identified by other methods. Remarkably, when comparing the same motif patterns identified from ChIP-Seq data to enriched regions identified from independent ChIP-chip experiments for the same TF, even with different cell types or different antibodies or both, we still found that motif patterns identified by HMS showed higher enrichment in the ChIP-enriched regions relative to random control sequences. This finding suggests that the motif patterns identified by HMS are closer to the underlying motif pattern recognized by the TF.

In this study, we utilized ChIP-enrichment of the peaks to rank order all input sequences, believing that ChIP-enrichment is positively correlated with the motif abundance. However, there are many potential reasons, both biological and technical, that a particular region is sequenced more deeply. These include the availability of the antibody's epitope during the immunoprecipitation step, conformational changes on the TF, abnormality in the cell line such as aneuploidy, bias introduced during the sequencing library construction, nucleotide-induced sequencibility bias (such as GC content) and bias related to alignment (repeat regions, various polymorphisms). These complications will reduce the correlation between ChIP-enrichment and sequencing depth. We believe advanced models that consider these factors will further improve the performance of HMS. Another potential enhancement would be to model the protein–DNA binding affinity indicated by read density using thermodynamic models (50).

In this study, if the motif width is unknown, we run HMS with every possible width within the range specified by the user and report all significant motif patterns. One possible improvement to this step would be to allow motif width w to vary during iterations (51). For example, we may add a Metropolis step, with equal probability of adding or removing one base at one end of the motif, and test whether the new motif pattern provides a better fit with the data. Another possible area for improvement concerns multiple binding sites. Currently, HMS is only designed to search for the primary binding site (i.e. the binding motif of the regulatory protein being ChIP'ed). However, we can also use HMS to identify secondary binding sites by masking the first motif identified and re-running HMS on the masked sequences.

In summary, we showed that ChIP-Seq data can significantly increase our ability to discover and refine TFBS motif patterns. However, new computational tools are needed in order to efficiently and thoroughly handle the ChIP-Seq data, as well as to exploit the various advantages of ChIP-Seq technology. The development of the highly scalable HMS algorithm represents an early attempt. With significant improvement in both accuracy and computation speed, we believe that HMS will be of broad interest to researchers conducting ChIP-Seq experiments and has the potential to accelerate discovery in biomedical research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the executive editor and two anonymous reviewers for constructive comments and suggestions. We would like to thank Christopher Maher in the Chinnaiyan Lab for helpful discussion, Lisa Henn and Jill Granger for proofreading the article.

FUNDING

Burroughs Welcome Foundation; Doris Duke Charitable Foundation; Prostate Cancer Foundation; American Cancer Society; National Institute of Health (K99CA129565, U01CA111275, P30CA46592, P30CA69568, R01GM72007, R01HG005119); Department of Defense (PC080665, BCO75023, W81XWH-09-2-0014) and the Howard Hughes Medical Institute. Funding for open access charge: R01HG005119.

Conflict of interest statement. None declared.

REFERENCES

1. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
2. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138.
3. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
4. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
5. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
6. Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.
7. Stormo,G.D. and Hartzell,G.W. III (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
8. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
9. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.
10. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
11. Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.
12. Shim,H. and Keles,S. (2008) Integrating quantitative information from ChIP-chip experiments into motif finding. *Biostatistics*, **9**, 51–65.
13. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
14. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
15. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
16. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
17. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
18. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
19. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
20. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
21. Solomon,M.J., Larsen,P.L. and Varshavsky,A. (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, **53**, 937–947.
22. Orlando,V. and Paro,R. (1993) Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell*, **75**, 1187–1198.
23. Fejes,A.P., Robertson,G., Bilenky,M., Varhol,R., Bainbridge,M. and Jones,S.J. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 1729–1730.
24. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
25. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
26. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **360**, 5221–5231.
27. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.
28. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
29. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

30. Choi,H., Nesvizhskii,A.I., Ghosh,D. and Qin,Z.S. (2009) Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. *Bioinformatics*, **25**, 1715–1721.
31. Nix,D.A., Courdy,S.J. and Boucher,K.M. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, **9**, 523.
32. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the idenification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
33. Liu,J.S. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene-regulation problem. *J. Am. Stat. Assoc.*, **89**, 958–966.
34. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
35. Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer-membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.
36. Gupta,M. and Liu,J.S. (2003) Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Am. Stat. Assoc.*, **98**, 55–66.
37. Staden,R. (1988) Methods to define and locate patterns of motifs in sequences. *Comput. Appl. Biosci.*, **4**, 53–60.
38. Bulyk,M.L., Johnson,P.L. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
39. Lee,M.L., Bulyk,M.L., Whitmore,G.A. and Church,G.M. (2002) A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics*, **58**, 981–988.
40. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
41. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
42. King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, **31**, e116.
43. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) *RECOMB 2003*. Berlin, Germany.
44. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
45. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
46. Hanai,R. and Wada,A. (1988) The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. *J. Mol. Evol.*, **27**, 321–325.
47. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
48. Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenkov,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.
49. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
50. Leach,A.R. (1996) *Molecular Modelling: Principles and Applications*. NY Longman Pub. Group, White Plains.
51. Jensen,S.T. and Liu,J.S. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, **20**, 1557–1564.