

# BayesMD: Flexible Biological Modeling for Motif Discovery

MAN-HUNG ERIC TANG,<sup>1,\*</sup> ANDERS KROGH,<sup>1</sup> and OLE WINTHER<sup>1,2,\*</sup>

## ABSTRACT

We present BayesMD, a Bayesian Motif Discovery model with several new features. Three different types of biological a priori knowledge are built into the framework in a modular fashion. A mixture of Dirichlets is used as prior over nucleotide probabilities in binding sites. It is trained on transcription factor (TF) databases in order to extract the typical properties of TF binding sites. In a similar fashion we train organism-specific priors for the background sequences. Lastly, we use a prior over the position of binding sites. This prior represents information complementary to the motif and background priors coming from conservation, local sequence complexity, nucleosome occupancy, etc. and assumptions about the number of occurrences. The Bayesian inference is carried out using a combination of exact marginalization (multinomial parameters) and sampling (over the position of sites). Robust sampling results are achieved using the advanced sampling method parallel tempering. In a post-analysis step candidate motifs with high marginal probability are found by searching among those motifs that contain sites that occur frequently. Thereby, maximum a posteriori inference for the motifs is avoided and the marginal probabilities can be used directly to assess the significance of the findings. The framework is benchmarked against other methods on a number of real and artificial data sets. The accompanying prediction server, documentation, software, models and data are available from <http://bayesmd.binf.ku.dk/>.

**Key words:** computational molecular biology, gene expression, machine learning, Markov chains, Monte Carlo likelihood, recognition of genes and regulatory elements, sequence analysis, stochastic processes.

## 1. INTRODUCTION

THE IDENTIFICATION OF REGULATORY PATTERNS in DNA is an important step in the process towards understanding the transcriptional program in cells. One of the most important regulation events are

---

<sup>1</sup>Bioinformatics Centre, Department of Molecular Biology, and Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen, Denmark.

<sup>2</sup>DTU Informatics, Technical University of Denmark, Lyngby, Denmark.

\*These authors contributed equally.

DNA-binding proteins (transcription factors, TFs) binding specific sites in DNA (transcription factor binding sites, TFBSs). There are typically an aggregation of TFBSs for certain TFs at the start of genes (transcription start sites, TSS). The region around the transcription start sites is referred to as the promoter (Smale and Kadonaga, 2003).

From a computational perspective, this problem has been approached by informed and ab-initio approaches. In the informed approach, position weight matrices, (containing column-wise probabilities of each nucleic acid to occur) summarize the statistical properties of observed TFBSs. These are scanned over DNA sequences to predict potential TFBSs, as reviewed in Wasserman and Krivan (2003). This method, called motif finding, can be regarded as a fully informed approach for which substantial experimental data to build the models is needed. This data is lacking for most transcription factors. In the ab-initio approach, one makes an unbiased search for over-represented sub-sequences close to transcription start sites for genes that are expected to share some regulatory elements. This motif discovery approach is the topic of this paper. It is equivalent to finding an optimal alignment of sub-sequences which is NP complete (Wang and Jiang, 1994). While many ab-initio approaches have been helpful in experimental studies, lack of sensitivity is often a problem when challenged with DNA sequences that have relevant sizes for biologically motivated problems (often 1000 nucleotides or more). These problems are reviewed in detail in Sandve and Drablos (2006), Pavesi et al. (2004), and Wasserman and Krivan (2003). In the following, the collection of potential binding sites, which can be summarized in a position weight matrix, will be referred to as the motif.

### *1.1. A short literature survey of motif discovery*

Briefly, approaches for computational inference of conserved regulatory motifs in DNA sequences follow several lines of research including the design of sophisticated probabilistic models and enumerative strategies. Enumerative (word-based) motif discovery aims to list all possible n-mers that satisfy an objective function, such as a conservation or significance score, given a maximum allowed number of mismatches. This kind of systematic exploration of the search space for identifying significantly conserved elements produces robust and reproducible predictions. In the recent assessments by Tompa et al. (2005) and Li and Tompa (2006), it was shown that methods like Weeder (Pavesi et al., 2001), which follows this approach, can achieve very good results in predicting known motifs and outperformed the other participants in this survey.

Probabilistic approaches on the other hand, use position weight matrices to represent the overall properties of sites that are assumed to exist in noisy background sequences. Therefore the probabilistic model has at least these two elements: the motif and the background models. The motif model, the position weight matrix, is a multinomial distribution for each position. The distribution of the background sequence can also be taken to be multinomial but more recently Markov models have become de facto standard. Additional parameters such as the width and number of motifs and the order of the background model also have to be specified. In the following we first discuss methods for learning the parameters of the model (plus some extensions) and thereafter how to incorporate biological knowledge.

Parameters of the motif model are inferred either by Bayesian or maximum likelihood methods. Other objective functions closely related to the likelihood have been proposed like the posterior (leading to maximum a posteriori, MAP), sequence specificity (Workman and Stormo, 2000; Pavesi et al., 2001), Z-scores (Sinha and Tompa, 2003), or information content (Kechris et al., 2004). The two most frequently used methods for maximizing the likelihood are expectation maximization (EM) and Gibbs sampling. For example, the popular program MEME (Bailey and Elkan, 1994) uses the EM algorithm to attain the MAP solution. Since the likelihood has several local maxima, good heuristics for seeding and restarts are needed in order to obtain reasonable results. In their landmark article, Lawrence et al. (1993) proposed to use stochastic search based upon Gibbs sampling. This work was followed by improvements and a number of related studies based on the same model which added new features to the model. For example, MotifSampler (Thijs et al., 2002) extended the Gibbs Sampler with higher order background and inference for the number of occurrences. Other contributions like the Gibbs Recursive Sampler (Thompson et al., 2003) or BioProspector (Liu et al., 2001) extended the original approach to model gapped and palindromic motifs. In the latter, a higher order background model was implemented and thresholds on the observed alignments were set to handle the non-occurrence of a motif in a sequence.

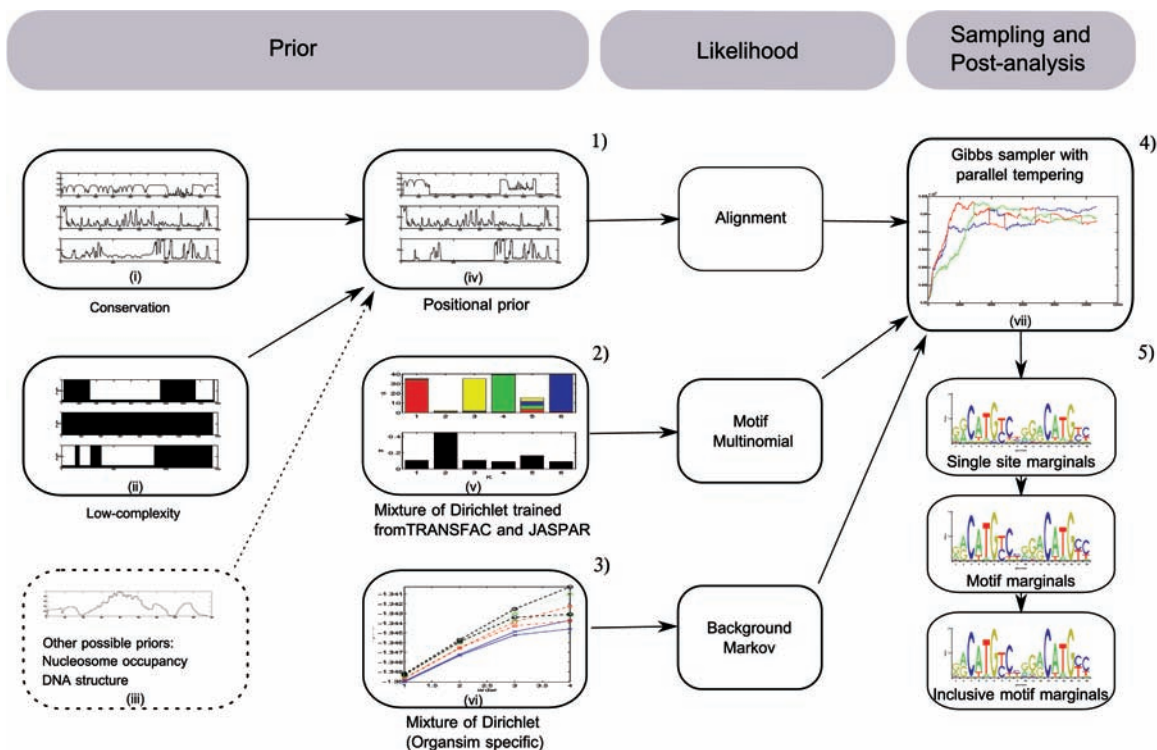
To apply motif discovery (or finding) to a set of co-regulated genes one has to extract the relevant promoters. After having solved this non-trivial task, the central question that arises is what kind of biological prior knowledge is available to improve the performance of the model. This information can be divided into three partly overlapping categories: motif, background and other positional information. For the two first categories, where multinomial type models are used, one may use pseudo-counts and Markov models trained on sets of organism-specific promoter sequences. It is more open-ended how to use, for example, phylogenetic footprinting (Lenhard et al., 2003), which relies on the observation that evolutionary conserved regions are more likely to harbor functional regulatory elements. Wasserman and Fickett (1998) use conservation scores for the post filtering TFBS predictions. More recently, PhyloGibbs (Siddharthan et al., 2005) uses phylogenetic information directly as a part of Gibbs sampling, and Xie et al. (2005) used a conservation index to make an unbiased genome-wide search for motifs. The performance of the Gibbs sampling inference engine can also be improved by incorporating information on nucleosome positioning preferences (Narlikar et al., 2007a, 2007b) or DNA duplex stability (Gordan and Hartemink, 2008) or cis-regulatory modules (Zhou and Wong, 2004). These developments suggest that appropriate identification and incorporation of informative prior knowledge on motif, background and positional biases are crucial for success in probabilistic motif discovery. The Bayesian formulation used in our present work offers a natural and elegant way of incorporating such biological knowledge.

### *1.2. Bayesian formulation of the motif discovery problem*

In this section, we will highlight the underlying line of thinking that motivates our framework. A potential problem with maximum likelihood approaches is that they ignore uncertainty of parameter estimates, i.e., two identical position weight matrices estimated from counts from say 10 and 100 sites have the same inferential status. The most elegant solution to this counter-intuitive situation is the Bayesian formulation of statistical inference. For a tutorial on biological sequence analysis, see Liu and Lawrence (1999). Bayesian statistics takes into account all sources of uncertainties by first describing all the parameters of the system with probability distributions and then performing the inference by averaging over the parameters. Probabilistic motif discovery has three elements accounting for the distributions of sites that build up the motif, the background and the position of the sites (alignments). These three elements will be discussed in turn below. Figure 1 illustrates this modularity in the context of our framework.

Most probabilistic models assume that the nucleotide positions within the motif are independent from each other and that the multinomial parameters in the position probability matrices follow the conjugate Dirichlet prior distribution (in MAP this simply leads to the use of pseudo-counts). However, since positions in motifs tend to be either very conserved (e.g. almost entirely As) or non-specific (no preference for any base compared to the background), the prior distribution is better described by a mixture of Dirichlet distributions. Because the distributions are in a conjugate family, the multinomial parameters can be averaged out (marginalized) analytically to form a mixture of compound multinomials (or mixture of Dirichlet multinomial, MoDM). These can be assessed by training on known motifs—matrix models—constructed by actual biological experiments. In our work, the parameters of the motifs prior are learned from all matrices in the TRANSFAC (Matys et al., 2006) or JASPAR (Sandelin et al., 2004; Bryne et al., 2008) databases using an EM algorithm that optimizes the MoDM likelihood function. Xing et al. (2002, 2004) and Xing and Karp (2004) use an even more advanced model, a Hidden Markov Dirichlet-Multinomial (HMDM), which can also model site-clustering—i.e., that conserved positions tend to be adjacent (Kechris et al., 2004). Importantly, the known motifs are based on finite data, which gives a constraint on how many parameters that can be conveniently trained without overfitting. Thus, the prior is less complex than the real binding preference of the transcription factors. More specific knowledge about the sought motif can of course also be built into the prior (Thompson et al., 2003) leading to a continuous transition between motif discovery and motif finding (where a sequence is scanned with pre-defined models from TRANSFAC or JASPAR).

The background model is typically chosen as a stationary higher order Markov model (Thijs et al., 2002; Thompson et al., 2003), but as above, this model is simpler than real promoter sequences. Real promoter sequences tend to be non-stationary, e.g., a mixture of AT- and CG-rich sequences, the local nucleotide content can vary considerably (Thompson et al., 2003) and will also depend upon the distance to the transcription start site (TSS). We use a Markov generalization of the MoDM to model the background,



**FIG. 1.** Summary diagram of the BayesMD framework. The high-level organization in steps 1–5 is described first and details given thereafter. Full-scale versions of sub-figures are given in the Supplementary Material. Steps 1–3 show the three elements of the probabilistic model: the positional prior, the likelihood for the motif, and the likelihood for the background, respectively. Step 4 represents the Gibbs sampling inference engine, and step 5 shows the three different ways of obtaining candidate motifs from the sample. Detailed description: Subfigures (i)–(iii) show three different ways of obtaining positional information, e.g., (i) is conservation along three sequences. In (iv), this information is combined with a prior over the number of occurrences of sites to produce a positional prior. Step (v) shows the parameters of Dirichlet mixture prior trained on the TRANSFAC database. The upper curve shows the pseudo count like parameters for each nucleotide for each of the six mixture components. There are four nucleotide specific and two homogeneous components. The lower plot shows the weight of each component. Step (vi) shows the training and test log likelihood per nucleotide for the mixture background prior trained on yeast promoters sequences as a function of the order of the Markov model for different number of mixture components. The test performance is increasing with both the order and the number of components. A leveling off tendency for the test performance is seen for the highest values. In step 4, inference is performed by exact marginalization of multinomial parameters of the motif and background models and by sampling for the site alignments. Sub-figure (vii) shows the log likelihood of three Gibbs sampling chains with parallel tempering. Parallel tempering interpolates between the posterior and more smooth distributions with less local maxima. The “temperatures” of two chains are periodically swapped with the Metropolis acceptance probability. The blue line is the posterior sample. Note also the burn-in regime. In step 5, motifs candidates with high marginal probability are found among sites with high probability (which are combined in single marginal motifs). Inclusive motifs follow a less strict definition including sites that occur in the majority of the samples together with the core sites of the motif. This gives a higher number of sites than the marginal motif (88 versus 75 in the p53 example) without affecting the information content of the motif (the logs).

i.e., the multinomial is replaced by a Markov model which is trained on a large set of promoter regions. This is similar to NestedMICA (Down and Hubbard, 2005), which uses a pure maximum likelihood mixture of Markov chains. Even these advanced models cannot always distinguish random recurring patterns from true motifs, and it is therefore of major importance to seek a better understanding of higher level structural and regulatory “codes.” However, in some instances, complementary experimental information, which can be encoded in the alignment prior, can compensate for the inadequate background model.

The alignment prior encodes information about the preferential positions of the sites in the promoter sequence (Thompson et al., 2003) and the distribution of the number of occurrences of motifs. In this

work we derive a prior which makes it possible for the user to specify both these quantities and infer the number of occurrences by sampling. In previous approaches, the number of occurrences had been found in a recursive manner (Thompson et al., 2003) or estimated by maximum likelihood approaches (Thijs et al., 2002). The genomic context in which the regulatory elements are present provides valuable information that can be modeled into the alignment prior, like for example, chromatin structure (accessibility) (Narlikar et al., 2007a, 2007b), DNA structural properties (Gordan and Hartemink, 2008). In our work, we propose an intuitive way to combine several sources of information, for example, evolutionary conservation or low complexity into the positional prior (Fig. 1).

Finally we discuss the computational task of carrying out the averaging over the model parameters and how to infer candidate motifs from this statistical ensemble. The stochastic variables in the motif discovery model are multinomial parameters of the motif and of the background and the alignments (position relative to the transcription start site). Given the alignments, the multinomial parameters can be integrated out analytically (Liu and Lawrence, 1999; Siddharthan et al., 2005). The summation over the possible alignments are usually deemed too computationally complex to be carried out exhaustively (although this is exactly what enumeration approaches like Weeder aim at). Instead sampling based methods or variational Bayes (Xing et al., 2002, 2004, Xing and Karp, 2004) are used. A very important question that arises is how to infer motifs from sites that have been sampled from the input promoter sequences. The fraction of samples containing a specific site is an estimate of this so-called marginal probability. However, a motif is made up of several sites that should occur simultaneously in the sample, i.e., we are looking for motifs and not sites with high marginal probabilities. Finding out which sites constitutes a motif is thus a combinatorial problem which is complicated further by the fact that motifs with fewer sites are bound to have higher probability than those with many sites. That is why most approaches partly abandon the Bayesian philosophy and instead look for the MAP estimate. Some programs (Thompson et al., 2003; Siddharthan et al., 2005) have an option to continue sampling after the MAP estimate is found in order to address significance as expressed through the MAP marginal probability. In our work, we solve the combinatorial problem by looking for motifs that are made up of sites coming from a hit list of the most frequently occurring sites. The output is a list of possible (likely similar) motifs sorted according to their probability. Figure 1 summarizes the proposed framework and its features: biological priors for site positions, number of sites, motif and background; advanced sampling scheme; and marginal probabilities for motifs in the post-analysis step.

## 2. RESULTS

In this section, we describe the main ingredients in the probabilistic approach (details are given in Methods and Supplementary Materials; see [www.liebertonline.com](http://www.liebertonline.com) for Supplementary Materials) and present three evaluations of our BayesMD tool. These highlight the effect of including a priori biological knowledge in the motif and background modules (Tompa assessment benchmark data sets), the robustness to background noise and use of a priori occurrence information (synthetic data sets) and the use of positional priors based upon sequence complexity and conservation (ChIP-PET p53 data set).

### 2.1. Probabilistic sequence model

In this section, we describe all the ingredients in the probabilistic model used for motif discovery. The basic building block in the Bayesian approach is the Dirichlet-Multinomial (DM) in which letters (nucleotides) are drawn from a multinomial and the parameters of the multinomial are drawn from the Dirichlet distribution. As observed by Xing et al. (2002, 2004) and Xing and Karp (2004), this distribution is too simple to describe the actual distribution of nucleotides in binding sites which tend to be a mixture of very conserved positions (strongly favoring exactly one of the nucleotides) and positions with little bias towards any of the positions. The obvious generalization which can account for this biological phenomena is a mixture of DM (MoDM): letters are drawn from a multinomial and the parameters of the multinomial are drawn from a mixture of Dirichlet distributions (Xing et al., 2002). An advantage of the DM and its mixture extension is that the parameters can be marginalized out analytically leaving the alignments (the positions of the sites building the motif) as the only remaining quantities to be inferred. In general we can

thus write the joint probability of the sequences  $\mathbf{S} = \{s_1, \dots, s_N\}$  and the alignment tensor  $\mathbf{A}$  (element  $a_{mnr}$  is the starting position of the  $r$ th occurrence of the  $m$ th motif in the  $n$ th sequence) as

$$P(\mathbf{A}, \mathbf{S}|\mathbf{B}) = \prod_m P_m(\mathbf{S}(\mathbf{A}_m)|\mathbf{A}_m, \mathbf{B}_m) P_{\text{bg}}(\mathbf{S}_{\text{bg}}|\mathbf{A}, \mathbf{B}_{\text{bg}}) P(\mathbf{A}|\mathbf{B}_{\text{align}}), \quad (1)$$

where  $P_m$  is the distribution for motif  $m$  with parameters  $\mathbf{B}_m$ ,  $\mathbf{S}(\mathbf{A}_m)$  is shorthand for the sequences contained in motif  $m$  and  $P_{\text{bg}}$  is the background distribution for sequences not in motifs  $\mathbf{S}_{\text{bg}} = \mathbf{S} \setminus \{\mathbf{S}(\mathbf{A}_m)\}$ . Figure 1 gives a graphical representation of Equation (1) and the explicit expressions for the different terms are given in Methods.

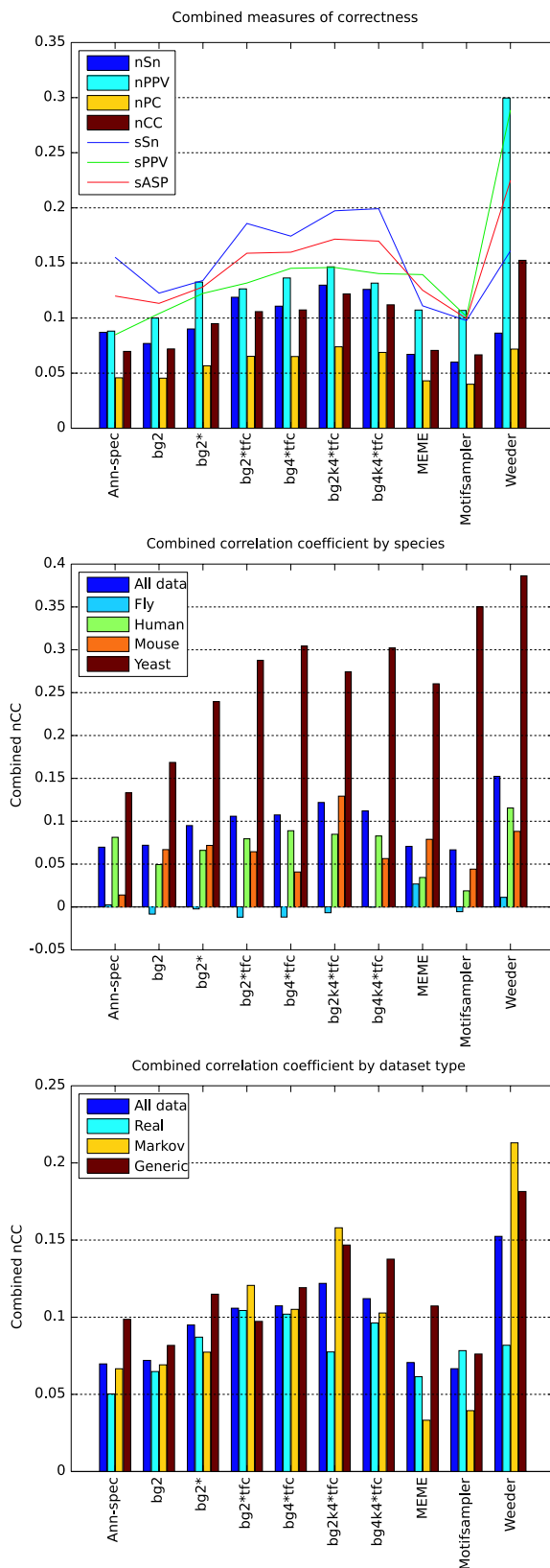
Bayesian inferences about motifs are obtained from (marginals of) the posterior probability  $P(\mathbf{A}|\mathbf{S}, \mathbf{B}) = P(\mathbf{A}, \mathbf{S}|\mathbf{B})/P(\mathbf{S}|\mathbf{B})$ , see Methods. We define the prior over alignments  $P(\mathbf{A}|\mathbf{B}_{\text{align}})$  such that we can specify the distribution over the number of occurrences and positional biases of sites in each of the sequences. The only fixed quantity is the width of the motifs which has to be provided by the user. Since we in most cases cannot make an exhaustive enumeration of all possible alignments, we estimate the posterior probabilities using a Monte Carlo approach similar to the Gibbs sampling approach of Lawrence et al. (1993; see Methods for details).

## 2.2. Evaluation on Tompa data set

The many data sets provided in the Tompa assessment (Tompa et al., 2005) offered an excellent opportunity to make extensive tests on how the different components of the motif discovery model contribute to the performance and to compare different methods. We focused on the various refinements of the motif and background models while keeping the alignment prior and MCMC parameters fixed. The length of the sought motif was set to 12 as it is a good compromise that enables our method to capture most motifs. We searched for a single motif, with no a-priori positional bias, occurring from 0 to 3 times in each sequence. The occurrence prior was set such that it was twice as likely to observe 0 or 1 motif than 2 or 3 in a sequence. The MCMC parameters (number of samples, phase-shift interval, number of parallel chains) were set to their default values in all tests.

We assessed the following combinations of background and motif models: The most basic model denoted “bg2” makes no use of a-priori biological knowledge. We set all parameters (pseudo counts) of the Dirichlets to one and let the background be second order in order for the model to be able to pick non-trivial correlations in the investigated sequences. The next step, “bg2\*,” consisted in replacing the uniform background by a second order background model with Dirichlet parameters obtained by training on all available promoter sequences for each of the different organisms in the study (Methods). In “bg2\*tfc,” we set the motif model to a six component mixture of Dirichlet multinomials trained on all TRANSFAC entries (see Methods). In “bg4\*tfc” we used a fourth order background model, and in “bg2K4\*tfc” and “bg4K4\*tfc,” we used a four component mixture of Markov for the background of order of two and four, respectively. The results for these models together with those of the other methods from the Tompa assessment are shown in Figure 2. The inferred motif is in each case chosen as the top inclusive marginals as defined in Methods. The inclusive marginals predicts more sites than the actual motif marginals, since only a certain fraction (set to two-thirds here) of all sites in the inclusive motif have to appear simultaneously in the sample. This less strict definition is found to be in a bit better accordance with the performance measures of the assessment; we obtained a high number of positive predictions paying a relatively small cost of more false positives.

The evaluation of our model compared with some selected methods: MEME, Motif-Sampler, Weeder and Ann-Spec shows that our basic DM model with a second order background (bg2) achieves similar results as the stochastic methods (like MEME, Motif-Sampler and Ann-Spec), showing an overall nucleotide Positive Predictive Value (nPPV) of 0.100 against 0.107, 0.106, and 0.088, respectively (Fig. 2, top). We evaluated the benefit of implementing a more complex model in the subsequent simulations. We see a significant improvement by replacing the basic second order background by a mixture model of the same order trained from promoter data. Indeed, the overall nPPV of bg2\* climbs to 0.132, which is significantly higher than bg2 and the three other stochastic methods. The use of a mixture motif model to bg2\*tfc enabled BayesMD to capture more real sites, as for example the nCC increases from 0.064 to 0.104 from bg2 to bg2\*tfc (Fig. 2, bottom). More conserved sites that could have been previously considered as noise



**FIG. 2.** Evaluation of the Tompa Benchmark for various configurations of our program and some selected methods. The top panel summarizes the combined measures of correctness over all 56 datasets for each method at the nucleotide and site level. The middle panel shows the combined correlation (nCC) coefficient by species, and the bottom panel shows the nCC coefficient by dataset type.

(yeast, Markov data sets) were also captured by using the mixture model. For example, we significantly improved the nCC from 0.168 in bg2 to 0.287 in bg2\*tfc for yeast and from 0.069 to 0.120 for Markov datasets (Fig. 2, center and bottom).

The flexibility of our background model can improve our results for some genomes. In our example, using a four-component mixture of Markov instead of just a single Markov component was more accurate on mouse data sets, as the nCC increased from 0.071 to 0.129 between bg2\*tfc and bg2K4\*tfc. The overall best performance of BayesMD was achieved with bg2K4\*tfc which uses a mixture model for both motif and background (2nd order Markov). This configuration outperformed the other stochastic methods in terms of sensitivity (nSn), positive predicted value (nPPV), and correlation coefficient (nCC). For example, we get a nCC of 0.121 against 0.070, 0.066, and 0.069, respectively (Fig. 2, top). Compared to Weeder, bg2K4\*tfc, shows a better sensitivity (sSn of 0.197 against 0.160 for Weeder) but has lower positive predictive values (sPPV of 0.145 against 0.288). This can be explained by the fact that our method tends to return more sites than what Weeder does, since our model does not set constraints on the consensus (number of mismatches) but only on the posterior. However, the correlation coefficients are of the same range (nCC of 0.121 for bg2K4\*tfc and 0.152 for Weeder), which indicates that the performances of our probabilistic method can perform as well as an enumerative method. Note that in this evaluation the use of higher order backgrounds did not significantly change our average performance but contributed more specifically to refine the model for very conserved motifs such as sites in yeast data sets. We see that adding more flexibility to our model improves performance. With the amount of promoter data available for especially the higher organisms, it is very likely that increasing the number of mixture components and higher order background models order could improve the model especially for non-synthetic data sets.

### 2.3. Evaluation on synthetic data

It is usually difficult to evaluate the performance of a motif discovery tool because very few annotated and curated datasets are available for this task. In order to assess the robustness of our method without introducing sequence biases, we used the set of synthetic data from Down and Hubbard (2005), consisting of human intergenic regions in which known transcription factor binding sites from the JASPAR collection have been inserted in sequences of increasing length. The expected outcome in such a study is that most motif discovery methods will find the correct sites if the sequence length is small, but eventually fail when the sequence length increases. The first four of these data sets consist of 100 sequences for which 50 contains one instance of the motif (in other words, one planted site from the model in every two input sequence). We used this information in the prior, setting the number of sought motifs to one and the occurrence prior was set to have the probability 0.5 for having zero or one occurrence. The last set consisted of 100 sequences spiked with 50 CREB instances and 50 Tal1beta instances. In Down and Hubbard (2005), the latter were considered as decoys, i.e. distracters and the test was designed to assess the ability of each program to find the CREB motifs in presence of Tal1beta sites. However, we sought both and consequently set the sought number of motifs to two.

In Table 1 we show the longest sequence for which the tool (MEME, NestedMICA, and BayesMD) recovers the sought motif (as in Down and Hubbard [2005] evaluated by visual inspection of the inferred

TABLE 1. DISCOVERY OF THE DIFFERENT MOTIFS FROM 50 INSTANCES INSERTED IN 100 SEQUENCES OF VARIOUS LENGTHS

<i>Motif/tool</i>	<i>MEME</i>	<i>N'MICA</i>	BayesMD
HLF	150	1000*	1000*
c-FOS	300	500	800
HFH-1	1200	1200	2000*
CREB	400	600	800*
CREBdecoy	200	600	800*

The number in the table indicates that the motifs are recovered up to that sequence length.

\*This is the longest sequence data set provided.

matrix as a sequence logo). In all cases, BayesMD showed better abilities than MEME and at least as good as NestedMICA to capture the single motifs that were inserted in the sets of sequences of variable length. In the second test, where a decoy Tal1beta motif was inserted in the CREB datasets, BayesMD was able to recover the instances of both inserted motifs each time by using the two motif search mode. This series of tests on synthetic data showed that prior knowledge can readily be used and that BayesMD is robust and sensitive as it was able to discover motifs regardless the background even in presence of instances of other binding sites.

It is not surprising that both BayesMD and NestedMICA outperform MEME, because it is well known that maximization of the likelihood with EM can be caught in local minima. There are some similarities and differences between BayesMD and NestedMICA that are worth discussing in detail. They both use advanced sampling methods (parallel tempering and nested sampling) that increase the likelihood of finding configurations with high posterior probability. Which one is best will depend upon the problem at hand and parameter settings. BayesMD uses a more advanced motif model: a mixture versus a multinomial and the background model is trained on a much larger set of promoter sequences. The latter is important as Down and Hubbard find that a first order background is best whereas the results of the Tompa test indicate that at least a fourth order mixture model give the best performance.

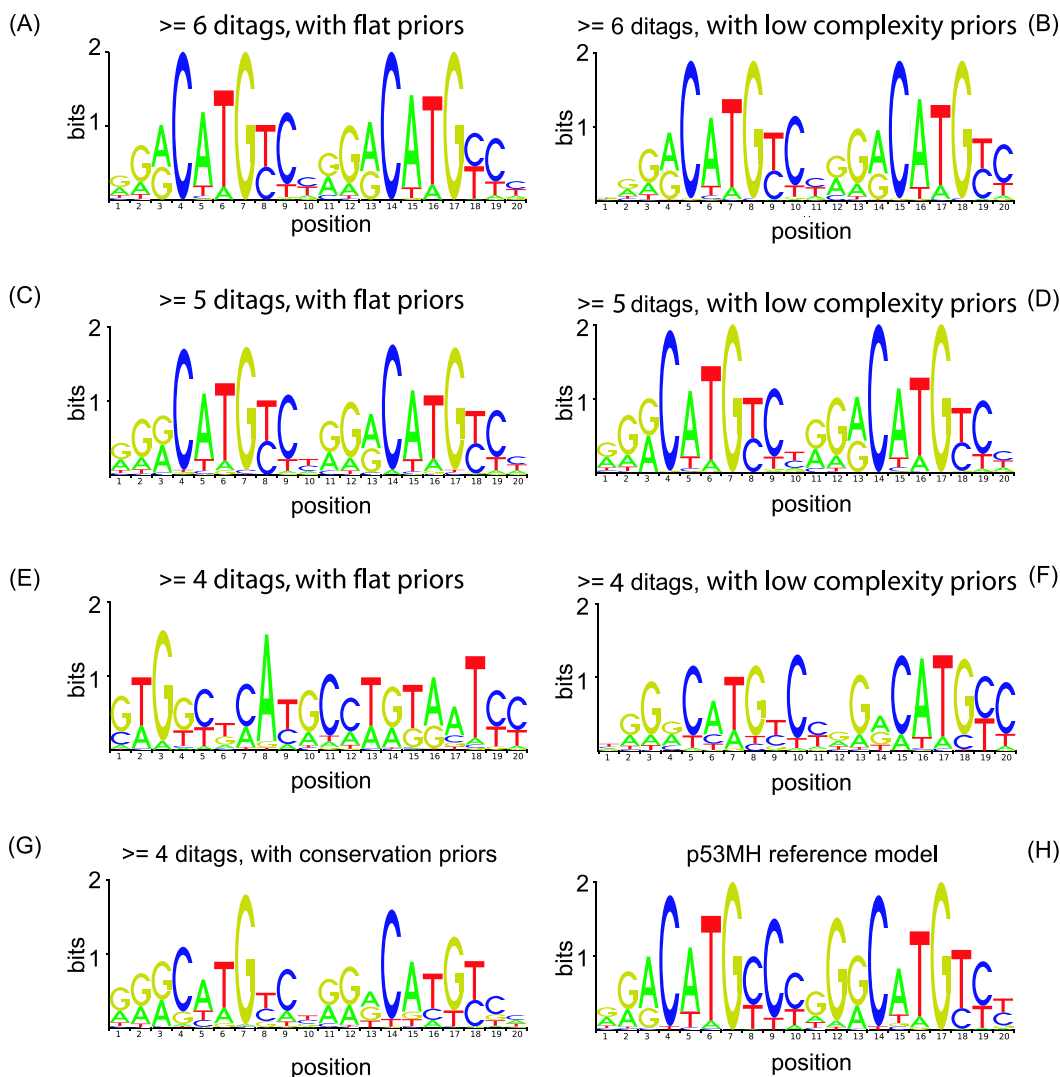
#### 2.4. Evaluation on ChIP-PET defined p53 binding sites

ChIP-PET methodology (Ng et al., 2005) enables high-resolution genome-wide survey of transcription factor binding sites. It consists of cloning ChIP DNA fragments that are then converted into pair-ended ditags (PETs) and mapped to the genome. This approach was used by Wei et al. (2006) to perform a global mapping of p53 TFBS in the human genome. 323 enriched PET tag clusters with three or more counts have been identified with this method. Previously, the characterization of the p53 binding loci in this data set was performed on 39 hand-selected PET 6+ cluster sequences using the de-novo motif discovery tool GLAM (Frith et al., 2004). A p53 motif model was derived from 30 predicted p53 binding sites after a motif refinement step.

It was empirically found in Wei et al. (2006) that PET clusters that effectively contain p53 binding sites should contain six ditags or more. This data set thus provides an excellent opportunity to test robustness of our approach by progressively including sequences from clusters of decreasing p53 binding site enrichment quality. We can furthermore investigate the effect of using soft-masking positional priors based either upon complexity of sequence content or conservation as extracted by the phastcons track of the UCSC genome browser. For the sequence complexity prior we assigned a relative probability of one to ten of low/high complexity nucleotides (as represented by small and capital letter in the sequence). For phastcons values, we set the relative prior to a baseline value of 0.1 and add 0.9 times the phastcons score giving a relative positional prior between 0.1 and 1. This approach relies on the assumption that true binding sites should be more likely to be located in high sequence complexity and/or evolutionary conserved regions. As it will be shown below this gives a very simple yet versatile method for integrating evolutionary information.

Since p53 binding sites were found in only 30 out of the 39 selected sequences by Wei et al. (2006), it is reasonable to assume that not all sequences contain the binding site. We therefore conservatively set the occurrence prior to 50–50 for zero and one occurrence. Because the reference model described in Hoh et al. (2002) extends 20 nucleotides, we searched motifs with a width of 20.

Results of our predictions are summarized by the sequence logos in Figure 3 and the ROC curves in Figure 4. Similarly to the previous section, successful recovery of the sought motifs is assessed by visual inspection of the matrix of the best predicted inclusive motif as a sequence logo. First we used a flat positional prior on all 68 6+ ditag sequences. With this configuration, the p53 binding motif was successfully recovered from 68 predicted sites, with and without soft-masking. The sequence logo that we obtained using low complexity priors (Fig. 3B) was very similar to the reference models described by Hoh et al. (2002), (Fig. 3H) and that of Wei et al. (2006) (not shown). The reason we get the number of sites equal to the number of sequences is because this is the default maximum number of sites (see Supplementary Materials). When we raise the upper limit slightly, more sites that contribute to the inclusive motif log likelihood are found with little change in the logo. However, to keep only the strongest signals, we report only for the default setting in the following. The result for 6+ set shows that the former results

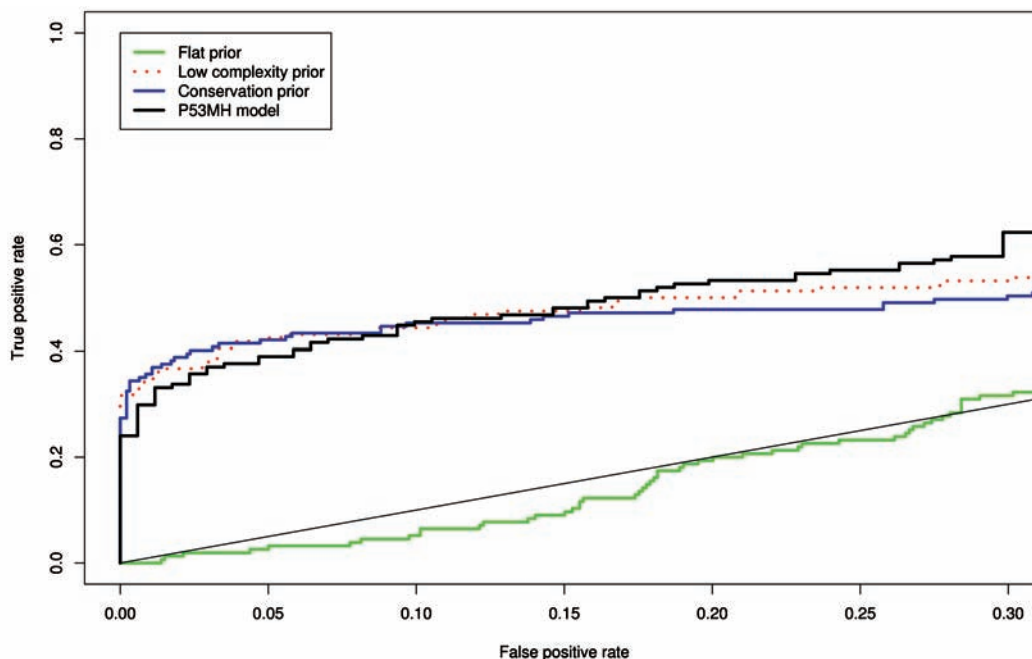


**FIG. 3.** Motif predictions on sequences from ChIP-PET clusters of decreasing p53 binding site enrichment quality. The p53 motif was successfully recovered in the original 68 PET 6+ sequences using flat priors (A) or low complexity positional priors (B). We obtained similar results by including 35 PET-5 sequences (C,D). We observed a strong noise affecting the results while adding 63 PET-4 sequences to this set while using only flat priors (E), but we successfully recovered the motif for PET 4+ sequences using low complexity priors and conservation (F,G). The p53MH model of Hoh et al. (2002) is shown as a reference.

clearly underestimate the number of sites and that our method is capable of recovering the motif from an unrestricted search over all sequences in the set. The small differences in nucleotide intensities in Figure 3A are due to the presence of less conserved p53 binding sites, affecting parts of the motif with lower information content like in positions 8–13.

Encouraged by this result, we added 35 sequences from PET clusters of size 5 to the original test set of 68 sequences, and performed a new evaluation on this 103-sequence data set. We again recovered the p53 motif with minor changes in the logo; see Figure 3C for flat positional prior and Figure 3D for low sequence complexity positional prior.

Finally, we added 63 more sequences from PET 4 clusters to our test set. We repeated same the procedure as for the PET 5+ and 6+ datasets and then compared the performances of each configuration to the reference model. We made ROC curves of each predicted profiles and for the P53MH model (Fig. 4). Only the 0–30% false positive range is shown on Figure 4, because our work aims to predict transcription factor binding sites with good accuracy. The negative set was made of 1000 random human promoters, providing



**FIG. 4.** ROC curves for the p53 profiles derived from the PET 4+ sequences. The curves are drawn for all positive cut-offs within the 0–30% false positive range, and the performance of the p53MH model was added for reference. Results agree with the observations of Figure 3. Loss of specificity due to noise is observed for the profile using flat priors, while the profiles derived with low complexity and conservation priors show similar or better performance than that of the p53MH model.

us with a realistic background that could contain p53 or other TFBSs. Using informative positional priors (blue and red curves), our method was able to recover the p53 motif with similar accuracy (10–30% interval) or higher accuracy (0–10% interval) than that of the P53MH model (black curve), which was also shown by the logos of Figures 3F and 3G. Despite the similarities in the sequence logos of Figure 3E and 3H, we can see that the matrix derived from the flat prior model lacks of accuracy, since the green ROC curve follows the random diagonal line. Unlike the two informative prior models, the flat prior model was unable to compensate the noise caused by the poor quality of some sequences in the dataset which possibly did not contain any P53 binding sites (false positives). In the ROC analysis, we used only the positive scores to build the curves and chose a realistic negative set to emphasize the differences between each method by “pulling down the curves.” This explains the relatively low ROC scores for the reference model.

This assessment on P53 ChIP-PET data demonstrates that our method is very sensitive, and that low sequence complexity and conservation information can be crucial for successful motif discovery. Encouragingly, it also shows that the ChIP-PET technology is more reliable in recovering true binding sites than previous motif discovery approaches have suggested Wei et al. (2006).

### 3. DISCUSSION

We have proposed a flexible, fully Bayesian model for motif discovery consisting of motif, background and alignment modules, summarized by Figure 1. The motif model is a multinomial at each position in the motif with a mixture of Dirichlet as prior for the parameters. This mixture prior is learned from empirical motif data, for example, all TRANSFAC matrices. The background counts are modeled with a higher order Markov model, where the prior again are a mixture of Dirichlet learned from all available organism-specific promoter sequences. Finally, in the alignment module, the user can specify both positional preferences of the location of motif instances (based upon for example tiling array information, low complexity, nucleosome

positioning, DNA stability or conservation) and the a priori probabilities for the number of occurrences of a motif in a sequence. Inference in the model, i.e. integration of multinomial parameters and alignments is carried out with a combination of analytical integration and sampling, and the stability of the Markov Chain Monte Carlo is enhanced by the use of the advanced sampling method called parallel tempering. Candidate motifs are not simply obtained as those sites most frequent in the sample, i.e. have the largest marginal probability. Rather, the marginal probability of motifs is obtained by requiring that the sites in motif must occur simultaneously in the sample. In order to make the search for possible motifs computationally feasible we limit the candidate sites to those with substantial marginal likelihood.

We now turn to a discussion of our empirical findings. The success of Weeder in the recent Tompa assessment (Tompa et al., 2005) of motif discovery tools seems to indicate that enumerative approaches are superior to probabilistic approaches. It has been the ambition of this work to investigate how well flexible probabilistic models tuned on all available a priori motif and promoter data can perform. As far as the Tompa assessment goes, we are not there yet. It is therefore interesting to investigate what are the differences and strengths/weaknesses of the two approaches.

One difference is enumeration versus sampling. We found that it was necessary to introduce a sampling method (parallel tempering) with better convergence properties than standard (Gibbs) sampling. So in some cases the cause of failure, if we don't opt for full enumeration, is not the model but rather the inference engine. The results in Section 2.3 illustrates well the clear advantage of using parallel tempering (as in BayesMD or NestedMICA) rather than standard sampling (as in MEME).

The Weeder definition of a motif is quite restrictive; only a few mismatches are allowed. This means that there is a bias towards only finding very similar instances of motifs. Probably few real motifs are so alike and the success might in part be owed to the fact that the Tompa sets are biased towards few very alike sub-sequences. On the other hand standard (non mixture) motif parameter priors are surely too non-specific as we observe that using a mixture distribution as prior we picked up more of those motifs found exclusively by Weeder. Moreover, unlike standard probabilistic approaches which infer motifs by merging the most probable sites together, our novel definition for motifs marginalizes a combination of sites. This additional constraint guarantees the similarity of all instances as consistently occurring combinations of sites should be more similar than sites that are taken individually. This introduces indirectly a restriction in the consensus of the most probable motif which improves the capability of our probabilistic method to find highly conserved motifs.

Another difference is that Weeder uses a relatively high order background model to weed out low complexity candidate motifs in the post enumeration step. We observed that going to more advanced background models (higher order and mixture) only increased the performance to a certain point. This is contrary to the test performance of the trained background model on promoter sequences (see Supplementary Materials). These results showed that with respect to modeling the promoter sequence, an even more complex model is to be preferred. This might indicate that either the Tompa sets are not representative of real promoter sequences (which is surely true because not all background sets are real) or that the complex background models in part have learned the sought motifs because they appear frequently enough in the promoter training sequences. The results for the motif model and to a lesser degree for the background indicate that we have enough biological a priori information to use even more advanced and flexible models. So the probabilistic models used so far have been too simplistic.

Explicit knowledge about negative sets, i.e., sequences without instances of the sought motif, can also be exploited in the probabilistic framework to get around the problem of fitting motifs with advanced background models. In the simplest approach we may update the background model with counts from these sequences. More advanced discriminative approaches are also possible (Workman and Stormo, 2000). The motif prior used in this paper can be extended to include dependencies between positions (Xing et al., 2002) or bias towards known motifs (Thompson et al., 2003), but probably the main improvements will come with the advancement in our understanding of regulatory "codes" on different levels—i.e., the information encoded in the promoter and other regulatory parts of the genes. Indeed, our BayesMD framework introduces an alignment prior that allows the user to represent such information in a relative simple way. We demonstrated in our prediction study of p53 PET clusters, the clear advantage of having informative alignment priors based on conservation and low complexity over a flat model. Other studies (Narlikar et al., 2007a, 2007b; Gordan and Hartemink, 2008) have proposed in a similar way to integrate DNA properties and nucleosome position preferences into the alignment component. The conclusions from

all these studies and our own suggest that the main further improvements of probabilistic approaches will come from the integration of higher level regulatory signals (structural, nucleosome code, conservation). This will take motif discovery from the generic towards more the realistic modeling.

## 4. METHODS

In this section, we describe the Bayesian modeling framework, the Markov chain Monte Carlo (MCMC) sampling scheme and the post-analysis used to infer motifs from the MCMC samples.

### 4.1. Dirichlet-Multinomial (DM)

In the simplest Bayesian approach to modeling sequence data, the data is modeled with a multinomial distribution and the parameters of the multinomial are modeled with the conjugate distribution, the Dirichlet. Here we go one step further and let the parameters being drawn from a mixture of Dirichlets to reflect the fact that, for example, background promoter sequences are diverse spanning (A + T)- to (C + G)-rich and purine (A + G)- to pyrimidine (C + T)-rich sequences. A single Dirichlet cannot adequately model this.

**Multinomial.** The multinomial with parameters  $\theta$  (lying in the probability simplex  $\theta_i \geq 0$  and  $\sum_i \theta_i = 1$ ) assigns the following probability to a set (or sequence) of letters  $\mathbf{s}$ , for example  $\mathbf{s} = \{A, A, C, T, G\}$ ,

$$P(\mathbf{s}|\theta) = \prod_i \theta_i^{c_i}, \quad (2)$$

where  $c_i = \sum_{j=1}^L \delta(s, i)$  is the number of counts/occurrences of letter  $i$ , in the example  $c_A = 2$  and  $c_C = c_G = c_T = 1$ .

**Dirichlet.** The basic a priori distribution over the probability simplex is the Dirichlet, with parameters  $\mathbf{b}$  (corresponding to pseudo-counts in regularized maximum likelihood), e.g. for four-letter DNA  $\mathbf{b} = (b_A, b_C, b_G, b_T)$ ,

$$P(\theta; \mathbf{b}) = \frac{1}{Z_D(\mathbf{b})} \prod_i \theta_i^{b_i-1} \delta\left(\sum_i \theta_i - 1\right). \quad (3)$$

The normalization constant is written in terms of  $\Gamma$ -functions

$$Z_D(\mathbf{b}) = \frac{\prod_i \Gamma(b_i)}{\Gamma(\sum_i b_i)}.$$

The uncertainty in parameters is handled by integrating them out in the joint probability  $P(\mathbf{s}, \theta; \mathbf{b}) = P(\mathbf{s}|\theta)P(\theta; \mathbf{b})$  to form the probability of the sequence, also called the marginal likelihood  $L(\mathbf{b}; \mathbf{s})$  of  $\mathbf{b}$ ,

$$P(\mathbf{s}; \mathbf{b}) = L(\mathbf{b}; \mathbf{s}) = \int d\theta P(\mathbf{s}, \theta; \mathbf{b}) = \frac{Z_D(\mathbf{b} + \mathbf{c})}{Z_D(\mathbf{b})} \equiv H(\mathbf{b}, \mathbf{c}). \quad (4)$$

This Bayesian distribution for the sequence data is the Dirichlet-multinomial (also known as the compound multinomial and Polya urn).

**Mixture of Dirichlet-multinomials (MoDM).** In the mixture of Dirichlet-multinomials (MoDM) the letters are drawn from a multinomial, Equation (2), and the multinomial parameters are drawn from a mixture of Dirichlets:

$$P(\theta | \{\mathbf{b}_k, \pi_k\}) = \sum_{k=1}^K \pi_k P(\theta; \mathbf{b}_k), \quad (5)$$

where  $\pi_k$  are the mixing proportions,  $\sum_k \pi_k = 1$ , and  $P(\theta; \mathbf{b}_k)$  are Dirichlets having parameters  $\mathbf{b}_k$ . As in Equation (4) above we can integrate out the multinomial parameters to form the marginal likelihood:

$$P(\mathbf{s}; \{\mathbf{b}_k, \pi_k\}) = L(\{\mathbf{b}_k, \pi_k\}; \mathbf{s}) = \int d\theta P(\mathbf{s}, \theta; \{\mathbf{b}_k, \pi_k\}) = \sum_k \pi_k H(\mathbf{b}_k, \mathbf{c}). \quad (6)$$

The set of parameters  $\{(\pi_k, \mathbf{b}_k) | k = 1, \dots, K\}$  should either be learned from data or set according to prior biological knowledge. We use TRANSFAC matrices and promoter sequences to train the motif and background models, respectively (see Supplementary Material). Since we have relatively large amounts of data, overfitting is an important issue so we can train the MoDM with maximum likelihood implemented by an EM algorithm (see Supplementary Material).

#### 4.2. Alignment Prior $P(\mathbf{A}|\mathbf{B}_{\text{align}})$

The motif and background models use information about typical motifs and promoter sequences. However, we might have additional information that are more specific to the data set we are investigating. We can use this information in an alignment prior  $P(\mathbf{A}|\mathbf{B}_{\text{align}})$  that can encode both the a priori knowledge about the specific position of a site (to reflect knowledge about conservation, missing data, information from tilling arrays, nucleosome positioning, DNA structural properties, etc.) and at the same time encode a priori assumptions about the distribution of the number of occurrences of the motif. The latter implies that we need a model where the number of occurrences is not fixed but can adapt to data. In the following, we specify  $P(\mathbf{A}|\mathbf{B}_{\text{align}})$  in terms of the positional prior for each site  $P_{mn}^{\text{pos}}(a_{mnr})$ ,  $\sum_{a=1}^{L_n - W_m + 1} P_{mn}^{\text{pos}}(a) = 1$  and the prior  $P_m^{\text{nocc}}(\cdot)$  over the number of occurrences. One example of how to use conservation information, implemented in the accompanying software, is to assign relatively higher probability for positions which are conserved according UCSC genome browser conservation track (Fig. 1).

We can let the number of occurrences vary without changing the notation used so far by letting the alignment  $a_{mnr}$  have an additional ‘‘absent’’ state  $a_{mnr} \in \{1, \dots, L_n - W_m + 1, \text{absent}\}$ . The variable  $a_{mnr}$  thus now indicates the  $r$ th possible occurrence. We let  $\Omega_{mn}$  denote the set of occurrences and  $|\Omega_{mn}|$  the cardinality of this set. In the Supplementary Material we derive the expression of  $P(\mathbf{A}|\mathbf{B}_{\text{align}})$  in terms of  $P_{mn}^{\text{pos}}(\cdot)$  and  $P_m^{\text{nocc}}(\cdot)$ . The final result for a specific motif and sequence (omitting the  $m$  and  $n$  indices) is

$$P(\mathbf{a}) = \frac{1}{\mathcal{N}(|\Omega|)} P^{\text{nocc}}(|\Omega|) \Theta(\mathbf{a}) \prod_{r \in \Omega} P^{\text{pos}}(a_r)$$

where  $\mathcal{N}(|\Omega|)$  is a normalization constant taking care of combinatorial factors and  $\Theta(\mathbf{a})$  is an indicator function which is one if the occurring sites are non-overlapping and zero otherwise. The normalizer  $\mathcal{N}(|\Omega|)$  only depends upon the number of occurring sites and can be computed effectively before the sampling stage (see Supplementary Material).

#### 4.3. Monte Carlo

Monte Carlo is used to construct a Markov chain of samples drawn from the posterior distribution. When and if the Markov chain has converged, the samples can be used to answer questions about probabilities of motifs and their sites, see next subsection. Typically we let the number of samples increase with the number of sequences and discard the first 30% to allow the chain to converge (burn-in). For the sampling, we adapted the two Monte Carlo moves, sequential and phase-shift, from the pioneering maximum a posteriori (MAP) approach of Lawrence et al. (1993). In order to keep the computational time as small as possible, we fixed the width of the sampled sites while other approaches like MEME (Bailey and Elkan, 1994) or NestedMICA (Down and Hubbard, 2005) allowed the option to set a minimum and a maximum target length.

In the Bayesian sequential update, one element in the alignment tensor  $a_{mnr}$  is considered at a time and the new value in  $\{1, \dots, L_n - W_m + 1, \text{absent}\}$  is chosen according to the relative posterior probability. The sequential update will not be very effective at changing a sub-optimal alignment of a motif, i.e. when the likelihood becomes higher if the location of all sites in a certain motif is changed with the same amount. The phase shift move does exactly this by considering all such shifts in an interval; we use  $-10$  to  $10$ . This move is applied seldom, every  $10^3$  sequential updates in our runs and the new value for the shift is chosen according to relative posterior probability. Trials on artificial data sets has shown that it can be advantageous to allow for overlaps in the phase shift update. These will automatically be removed in the subsequent sequential updates, because it only allows for non-overlapping alignments. We found that  $10^5$  samples (sum of sequential and phase-shift moves) gave stable inferences for motifs for all data sets

investigated (the largest containing  $1.6 \times 10^6$  nucleotides). The actual marginal probabilities for motifs, see below, could vary quite a bit for this sampling size.

To improve convergence properties of the Markov chain, we use parallel tempering (Gregory, 2005). In parallel tempering  $N_T$  chains (we use three as default) are run independently at different inverse temperatures  $\beta_i$ , i.e. we sample from  $P_{\beta_i}(\mathbf{A}_i|\mathbf{S}, \mathbf{B}) \propto [P(\mathbf{S}|\mathbf{A}_i, \mathbf{B})]^{\beta_i} P(\mathbf{A}_i|\mathbf{B})$ .  $\beta = 1$  corresponds to sampling from the posterior and smaller values corresponds to sampling from a smoother distribution. Once every few (we use third) sweep through the sequential moves, a swap of  $\beta$ -values between two adjacent values is performed with a probability given by the Metropolis acceptance probability of the joint distribution of all chains (Gregory, 2005). In order to get a reasonable acceptance probability ( $\sim 50\%$ ), the temperatures are set according to a geometric series with  $T_{\max} = 1/\beta_{\min} = 10^3 / \sum_n L_n$ . The maximum value scales with the total number of nucleotides to counter the scaling of the log likelihood. An illustration of parallel tempering using 3 chains is shown in Figure 1 and Supplementary Figure 5.

#### 4.4. Marginal probabilities for motifs

In the Bayesian setting, we can answer the question: what is the posterior probability (given our model) that a specific set of sites is part of a motif. As an example, we consider the situation that we are looking for one motif and want to ask what is the probability that whether the sites  $\{a_1, b_1, c_2, d_3\}$ —two sites  $a_1$  and  $b_1$  in sequence one and sites  $c_2$  and  $d_3$  in sequence two and three, respectively—is a part of the motif. This can be calculated as a marginal probability

$$P(\{a_1, b_1, c_2, d_3\}|\mathbf{S}, \mathbf{B}) = \sum_{\mathbf{A}} \delta(a_1; \mathbf{a}_1) \delta(b_1; \mathbf{a}_1) \delta(c_2; \mathbf{a}_2) \delta(d_3; \mathbf{a}_3) P(\mathbf{A}|\mathbf{S}, \mathbf{B}),$$

where  $\mathbf{a}_n$  is shorthand for occurrence in sequence  $n$ ,  $\mathbf{a}_n = (a_{n1}, \dots, a_{nR})$  and  $\delta(a; \mathbf{a})$  is one if  $a$  is equal to any of the components of  $\mathbf{a}$  and zero otherwise. Note that we in this marginal probability average over whatever else might occur together with the sites in question. If we knew what sites to consider, as in the example, the Markov chain can answer this question: we can simply count the fraction of samples where these sites occur together. The problem of course is that we do not know beforehand which sites are relevant so we have to come up with a method for finding candidate motifs. The central idea is to select the sites that we are going to merge into candidate motifs from a hit list of sites with high single site marginal probability. The details of the procedure are described in the Supplementary Material. The program reports the best candidates (given a specified minimum number of sites) for three ways to define motifs that we call single site marginal motifs (simply merging the most probable sites), marginal motifs (combination of sites that co-occur) and inclusive marginal motifs (combination of sites where the majority co-occur). In all three cases, the program reports the rank of motif, the number of sites, the log likelihood of the motif, the marginal probability, the change in accumulated marginal probability (i.e. which proportion is unique to this motif), the accumulated marginal probability and the consensus sequence. Also logos for the highest rank marginal and inclusive marginal motifs are displayed. In standard probabilistic approaches, motifs are inferred in the same manner as in our single-site marginal motif definition, by merging the most probable sites together.

Recently, Thompson et al. (2007) proposed to use the centroid position of sites that only differ by shifts as the best posterior estimator of the site position. In the spirit of the centroid sampler, BayesMD reports when candidate motifs are shifted versions of each other. But even though the centroid in many cases represents an improvement, this method can still break down if the sampler visits distinct motifs with a non-vanishing probability. The joint (or marginal) probability of the set of sites will be low in such a case. Our method proposes two novel definitions of motifs using marginalization for combination of sites, ensuring that sites that contribute to the motif are similar. Finding the sites to consider is a hard combinatorial problem that we solve sub-optimally by only considering combinations of sites that occur frequently in the sample. We find that the consensus sequence of the three motif definitions very often agree on the most predominant motifs, differing mostly in the number of sites and the overall marginal probability. However, when the signal in the investigated sequences is weak, the marginal motifs may pick up more than one motif whereas the signal is smeared out in the consensus of the single site marginals.

## ACKNOWLEDGMENTS

Thanks to Albin Sandelin for his valuable comments on the manuscript and Thomas Down for sharing his training and assessment datasets. This work was supported by a grant from the Novo Nordisk foundation to the Bioinformatics Center.

## DISCLOSURE STATEMENT

No conflicting financial interests exist.

## REFERENCES

- Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Second Int. Conf. Intell. Syst. Mol. Biol.* 28–36.
- Bryne, J., Valen, E., Tang, M., et al. 2008. Jaspas, the open access database of transcription factor binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 36, D102–D106.
- Down, T.A., and Hubbard, T.J. 2005. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res.* 33, 1445–1453.
- Frith, M.C., Hansen, U., Spouge, J.L., et al. 2004. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.* 32, 189–200.
- Gordan, R., and Hartemink, A. 2008. Using DNA duplex stability information for transcription factor binding site discovery. *Pac. Symp. Biocomput.* 453–464.
- Gregory, P.C. 2005. *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press, New York.
- Hoh, J., Jin, S., Parrado, T., et al. 2002. The p53mh algorithm and its application in detecting p53-responsive genes. *Proc. Natl. Acad. Sci. USA* 99, 8467–8472.
- Kechris, K., van Zwet, E., Bickel, P., et al. 2004. Detecting DNA regulatory motifs by incorporating positional trends in information content. *Genome Biol.* 5, R50.
- Lawrence, C.E., Altschul, S.F., Boguski, M.S., et al. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.
- Lenhard, B., Sandelin, A., Mendoza, L., et al. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* 2, 13.
- Li, N., and Tompa, M. 2006. Analysis of computational approaches for motif discovery. *Algorithms Mol. Biol.* 1–8.
- Liu, J.S., and Lawrence, C.E. 1999. Bayesian inference on biopolymer models. *Bioinformatics* 15, 38–52.
- 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.
- Matys, V., Kel-Margoulis, O.V., Fricke, E., et al. 2006. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108–D110.
- Narlikar, L., Gordan, R., and Hartemink, A. 2007a. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.* 3, 2199–2208.
- Narlikar, L., Gordan, R., and Hartemink, A. 2007b. Nucleosome occupancy information improves de novo motif discovery. *RECOMB 2007* 107–121.
- Ng, P., Wei, C.L., Sung, W.K., et al. 2005. Gene identification signature (gis) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2, 105–111.
- Pavesi, G., Mauri, G., and Pesole, G. 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 17, S207–S214.
- Pavesi, G., Mauri, G., and Pesole, G. 2004. In silico representation and discovery of transcription factor binding sites. *Briefings Bioinform.* 5, 217–236.
- Sandelin, A., Alkema, W., Engström, P., et al. 2004. Jaspas: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32, D91–D94.
- Sandve, G.K., and Drablos, F. 2006. A survey of motif discovery methods in an integrated framework. *Biol. Direct* 1, 11.
- Siddharthan, R., Siggia, E.D., and van Nimwegen, E. 2005. Phylogibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* 1, e67.
- Sinha, S., and Tompa, M. 2003. Ymf: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* 31, 3586–3588.

- Smale, S.T., and Kadonaga, J.T. 2003. The RNA polymerase ii core promoter. *Annu. Rev. Biochem.*, 72, 449–479.
- Thijs, G., Marchal, K., Lescot, M., et al. 2002. A Gibbs sampling method to detect over-represented motifs in upstream regions of coexpressed genes. *J. Comput. Biol.* 9, 447–464.
- Thompson, W., Rouchka, E.C., and Lawrence, C.E. 2003. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.* 31, 3580–3585.
- Thompson, W., Newberg, L.A., Conlon, S., McCue, L.A., and Lawrence, C.E. 2007. The Gibbs centroid sampler. *Nucleic Acids Res.* 35 (Web Server issue), W232–W237.
- Tompa, M., Li, N., Bailey, T.L., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23, 137–144.
- Wang, L., and Jiang, T. 1994. On the complexity of multiple sequence alignment. *J. Comput. Biol.* 1, 337–348.
- Wasserman, W.W., and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278, 167–182.
- Wasserman, W.W., and Krivan, W. 2003. In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften* 90, 156–166.
- Wei, C.L., Wu, Q., Vega, V.B., et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124, 207–219.
- Workman, C.T., and Stormo, G.D. 2000. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.* 467–478.
- Xie, X., Lu, J., Kulbokas, E.J., et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature* 434, 338–345.
- Xing, E.P., Jordan, M.I., Karp, R.M., et al. 2002. A hierarchical Bayesian Markovian model for motifs in biopolymer sequences. *Advances in Neural Information Processing Systems 16 (NIPS2002)*. MIT Press.
- Xing, E.P., and Karp, R.M. 2004. Motifprototyper: a profile Bayesian model for motif family. *Proc. Natl. Acad. Sci. USA* 101, 10523–10528.
- Xing, E.P., Wu, W., Jordan, M.I., et al. 2004. Logos: a modular Bayesian model for de novo motif detection. *J. Bioinform. Comput. Biol.* 2, 127–154.
- Zhou, Q., and Wong, W.H. 2004. Cismodule: *De novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *PNAS* 101, 12114–12119.

Address reprint requests to:

*Dr. Ole Winther*

*The Bioinformatics Centre, Department of Molecular Biology*

*Biotech Research and Innovation Centre*

*University of Copenhagen*

*Ole Maaløes Vej 5*

*DK-2200 København Ø, Denmark*

*E-mail: owi@imm.dtu.dk*