Published online 9 June 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 13* ***6055–6069***
doi: 10.1093/nar/gkw521

Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences

Matthias Siebert[1,2] and Johannes Söding[1,*]

[1]Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany and [2]Gene Center, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany

Received April 11, 2016; Revised May 23, 2016; Accepted May 29, 2016

ABSTRACT

Position weight matrices (PWMs) are the standard model for DNA and RNA regulatory motifs. In PWMs nucleotide probabilities are independent of nucleotides at other positions. Models that account for dependencies need many parameters and are prone to overfitting. We have developed a Bayesian approach for motif discovery using Markov models in which conditional probabilities of order *k* − 1 act as priors for those of order *k*. This Bayesian Markov model (BaMM) training automatically adapts model complexity to the amount of available data. We also derive an EM algorithm for *de-novo* discovery of enriched motifs. For transcription factor binding, BaMMs achieve significantly (*P* = 1/16) higher cross-validated partial AUC than PWMs in 97% of 446 ChIP-seq ENCODE datasets and improve performance by 36% on average. BaMMs also learn complex multipartite motifs, improving predictions of transcription start sites, polyadenylation sites, bacterial pause sites, and RNA binding sites by 26–101%. BaMMs never performed worse than PWMs. These robust improvements argue in favour of generally replacing PWMs by BaMMs.

INTRODUCTION

The control of gene expression allows the cell to adapt its protein and RNA inventory in response to developmental and environmental cues. At its center lies the binding of proteins to specific motifs in promoters and enhancers to control RNA synthesis rates and to RNAs to regulate their splicing, localization, translation and degradation. The accurate prediction of protein binding affinities to DNA and RNA sequences is therefore of central importance for a quantitative understanding of cellular regulation and of life in general.

Most known models that describe the sequence specificity of transcription factors were deduced from in-vivo binding sites measured by ChIP-seq (1), from *in-vitro* measurements of binding strengths using either protein binding microarrays (PBMs) (2) or *in-vitro* selection coupled to high-throughput sequencing (HT-SELEX) (3), or from bacterial one-hybrid assays (4). To obtain a statistical model of binding specificity from such measurements, motif discovery algorithms learn the model parameters that agree best with the measurements. At present, the standard model for this purpose is the position weight matrix (PWM), and thousands of PWMs for transcription factors are available in motif databases such as JASPAR, HOCOMOCO, SwissRegulon or TRANSFAC (5–8).

PWMs rank binding sites according to the log-odds score $S(x_{1:W}) = \sum_{j=1}^{W} s_j(x_j)$ with contributions $s_j(x_j) = \log_2 [p_j(x_j)/p_{\mathrm{bg}}(x_j)]$ that depend only on single nucleotides x_j in the binding site sequence $x_{1:W} := (x_1 \ldots x_W)$. Here, $p_j(x_j)$ is the probability of nucleotide $x_j \in \{\mathrm{A, C, G, T}\}$ to occur at position j of the binding site, and $p_{\mathrm{bg}}(x_j)$ is the background probability for nucleotide x_j in a representative sequence set.

PWMs cannot model correlations between nucleotides. For example, if 50% of binding site sequences are GATC and the other 50% are GTAC, a PWM will give the same high score to GTTC and GAAC as to the true binding sequences. It cannot learn that at position 2 an A must be followed by a T and T by A.

Nucleotide correlations can originate from (i) stacking interactions that determine binding through DNA ‘shape readout’ (9), (ii) amino acids that contact multiple bases simultaneously (10), (iii) multiple sequence-dependent binding modes of a factor (11–14) and (iv) complex multi-submotif architectures with varying submotif spacings, which are typically bound cooperatively by multiple factors (15–17).

For these reasons one might expect more complex models that do not assume nucleotides to contribute independently to the binding strength to perform better than PWMs. However, the usefulness of such more complex models has been controversially discussed for long (18–20). Zhao *et al.* found PWMs to be as accurate as mixtures of PWMs to describe the binding strengths of transcription factors measured by

[*]To whom correspondence should be addressed. Tel: +49 551 201 2890; Fax: +49 551 201 2803; Email: soeding@mpibpc.mpg.de

PBMs (2,21). Weirauch *et al.* concluded that for >90% of tested transcription factors PWMs performed as well as more complex models to predict PBM binding strengths, and they were just as good in predicting *in-vivo* ChIP-seq binding sites for all factors (22). It is still not clear whether these results are really due to the absence of correlations in binding sites of most transcription factors or to what extent they are explained by the difficulty to train the many model parameters reliably and robustly.

Numerous models incorporating nucleotide dependencies have been developed to improve the modelling of binding site motifs and complex, multipartite motifs. Some learn mixtures of PWMs (23–25) or Markov models (26), or profile hidden Markov models (HMMs) (27). But dependencies generally decrease with increasing distance (3), and therefore most models are based on inhomogeneous Markov models (iMMs), in which the probability of x_j depends on the previous k nucleotides $x_{j-k:j-1}$ (28–30). First-order iMMs, sometimes called dinucleotide PWMs, have been added to the HOCOMOCO and JASPAR databases. HOCOMOCO's dinucleotide PWMs performed better on average than simple PWMs (6) and JASPAR's first-order iMMs yielded significantly better results than PWMs for 21% of 96 tested datasets (30).

The drawback of iMMs is that the number of parameters $W \times 3 \times 4^k$ grows exponentially with k. Already for a second-order model we need 48 parameters per position. To estimate them with 10% accuracy requires ∼100 counts per 3-mer, or 4800 sequences. When fewer sequences are available, more complex models risk being overtrained: they may perform significantly worse than a simple PWM model due to the noisy parameter estimates while showing overly optimistic performance on the training data.

To prevent overtraining, various heuristic methods were suggested that reduce the number of parameters in a data-driven fashion, by pruning the dependency graph describing which positions each motif position j depends on (23,31–33). These methods have several technical drawbacks: (i) they take yes/no decisions, which necessarily lead to a loss of information near the decision boundary. (ii) Optimising a discrete dependency graph is cumbersome: to decide between two alternative graphs one needs to find the optimum model parameters for each graph. Also, since two graphs usually induce models with different numbers of parameters, a likelihood-based optimization is not possible. (iii) The discreteness of the graph topology precludes efficient gradient-based optimization techniques. (iv) Finally, no algorithms have been put forward to train these models on unaligned motifs. These models can therefore not be applied for *de-novo* motif discovery.

Here, we present a Bayesian approach to learn inhomogeneous Markov models for sequence motifs that makes optimal use of the available information while avoiding overtraining. The key idea is that we use the conditional probabilities of order $k-1$ as priors for the conditional probabilities of order k.

Our Bayesian approach is similar to interpolated Markov models (34,35) in that the probabilities of order k are obtained as linear interpolation of the maximum likelihood (ML) estimate for order k and the lower-order probabilities. Various rather ad-hoc methods have been used to set the interpolation weights (34–37) (e.g. by making them depend on the P-value with which the hypothesis can be rejected that the conditional probabilities of order $k-1$ and of order k are noisy estimates of the same underlying distributions (36)). In contrast, the interpolation weights of BaMMs emerge naturally from our probabilistic approach without further assumptions except for the choice of priors.

We analyse how much can be gained by using higher-order inhomogeneous BaMMs over two baseline methods: zeroth-order BaMMs, which are simply PWMs trained with the standard EM-type algorithm as implemented in MEME (38), and our tool XXmotif, which performed favourably in comparison to state-of-the-art motif discovery tools (39). We assessed these different methods by the quality of the models they produced starting from the motif occurrences discovered by XXmotif. We demonstrate consistent improvements by higher-order BaMMs as compared to PWMs on each of a large and heterogeneous collection of datasets with simple and complex motif architectures. Likewise, correlation of predicted binding affinities with quantitative EMSA measurements was substantially improved.

MATERIALS AND METHODS

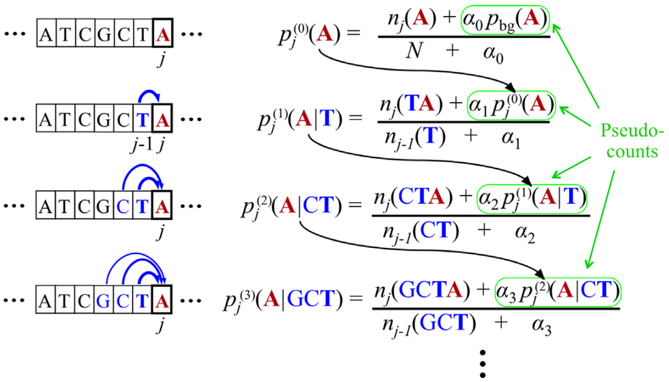
Bayesian Markov model learning

We would like to solve the following task: We have a set of sequences that are enriched in a sought motif, for example a binding motif for a transcription factor or a multi-submotif region with complex architecture. The sequences might have been produced by ChIP-seq or SELEX-seq experiments, they might be promoters of coregulated genes, regions of differential DNase I accessibility, or specific genomic sites such as splice sites. Our goal is to train a model that discovers the location and strength of enriched motifs in the training sequences and that can predict motifs and their strengths in arbitrary sequences.

In the Supplemental Methods (Eqs. (S.1)–(S.7)) we show that, to learn a model for the Gibbs binding energy $\Delta G(\mathbf{x})$ to sequence $\mathbf{x} = x_{1:W}$, we can solve an equivalent statistical learning problem: we learn the probability distribution of motif sequences, $p_{\text{motif}}(\mathbf{x})$, and of background sequences, $p_{\text{bg}}(\mathbf{x})$. Then the binding energy in units of $k_{\text{B}}T$ (where k_{B} is Boltzmann's constant) is, up to a constant, given by the log-odds score, $\Delta G(\mathbf{x})/k_{\text{B}}T = \log[p_{\text{motif}}(\mathbf{x})/p_{\text{bg}}(\mathbf{x})] +$ const. Here, we rely on the common approximation of unsaturated binding (low factor concentration), whereby the probability to observe sequence $\mathbf{x}$ in the training set is proportional to $\exp(\Delta G(\mathbf{x})/k_{\text{B}}T)$.

We model the background sequences with a homogeneous Markov model (MM) of order K', $p_{\text{bg}}(x_{1:W}) = \prod_{j=1}^{W} p_{\text{bg}}^{(K')}(x_j|x_{j-K':j-1})$ and the motif sites with an inhomogeneous Markov model (iMM) of order K, $p_{\text{motif}}(x_{1:W}) = \prod_{j=1}^{W} p_j^{(K)}(x_j|x_{j-K:j-1})$. This results in a log-odds score

$$S(\mathbf{x}) = \log_2 \frac{p_{\text{motif}}(\mathbf{x})}{p_{\text{bg}}(\mathbf{x})} = \sum_{j=1}^{W} \log_2 \frac{p_j^{(K)}(x_j|x_{j-K:j-1})}{p_{\text{bg}}^{(K')}(x_j|x_{j-K':j-1})}.$$



··· A T C G C T A ···
j
$p_j^{(0)}(\mathbf{A}) = \frac{n_j(\mathbf{A}) + \alpha_0 p_{bg}(\mathbf{A})}{N + \alpha_0}$
··· A T C G C T A ···
j-1 j
$p_j^{(1)}(\mathbf{A}|\mathbf{T}) = \frac{n_j(\mathbf{TA}) + \alpha_1 p_j^{(0)}(\mathbf{A})}{n_{j-1}(\mathbf{T}) + \alpha_1}$
··· A T C G C T A ···
j
$p_j^{(2)}(\mathbf{A}|\mathbf{CT}) = \frac{n_j(\mathbf{CTA}) + \alpha_2 p_j^{(1)}(\mathbf{A}|\mathbf{T})}{n_{j-1}(\mathbf{CT}) + \alpha_2}$
··· A T C G C T A ···
j
$p_j^{(3)}(\mathbf{A}|\mathbf{GCT}) = \frac{n_j(\mathbf{GCTA}) + \alpha_3 p_j^{(2)}(\mathbf{A}|\mathbf{CT})}{n_{j-1}(\mathbf{GCT}) + \alpha_3}$
Pseudo-counts

Figure 1. Bayesian Markov model training automatically adapts the effective number of parameters to the amount of data. In the last line, if the context GCT is so frequent at position j in the motif that its number of occurrences outweighs the pseudocount strength, $n_j(GCT) \gg \alpha_3$, the third-order probabilities for this context will be roughly the maximum likelihood estimate, e.g. $p_j^{(3)}(\mathrm{A}|\mathrm{GCT}) \approx n_j(GCTA)/n_{j-1}(GCT)$. However, if few GCT were observed in comparison with the pseudocounts, $n_j(GCT) \ll \alpha_3$, the third-order probabilities will fall back on the second-order estimate, $p_j^{(3)}(A|GCT) \approx p_j^{(2)}(A|CT)$. If also $n_j(CT) \ll \alpha_2$, then likewise the second-order estimate will fall back on the first-order estimate, and hence $p_j^{(3)}(A|GCT) \approx p_j^{(1)}(A|T)$. In this way, higher-order dependencies are only learned for the fraction of k-mer contexts that occur sufficiently often at one position j in the motif's training instances to trump the pseudocounts. Throughout this work we set $\alpha_0 = 1$ and $\alpha_k = 20 \times 3^{k-1}$.

Since the binding energy $\Delta G(\mathbf{x})$ is a linear function of the log-odds score, the score is ideal for ranking potential binding sites by their predicted strength.

The central idea for Bayesian Markov Model (BaMM) training is that, to learn the kth-order probability $p_j^{(k)}(x_j|x_{j-k:j-1})$, we can use the order-($k$−1) probability $p_j^{(k-1)}(x_j|x_{j-k+1:j-1})$ as prior information. The latter is an excellent approximation of the former because dependencies between positions generally decrease quickly with increasing distance (3). And since the shorter context $x_{j-k+1:j-1}$ is on average four times more frequent than $x_{j-k:j-1}$, the lower-order probabilities will also be more robustly estimated.

We learn the parameters of the inhomogeneous Markov model by maximising the posterior probability, the product of the likelihood and the prior probability (Eq. (S.12) in Supplementary Methods). A natural prior is a product of Dirichlet distributions with pseudocount parameters proportional to the lower-order model probabilities, with proportionality constants α_k for $k = 1, \ldots, K$ whose size determines the strength of the prior. Maximizing the posterior probability yields

$$p_j^{(k)}(x_j|x_{j-k:j-1}) = \frac{n_j(x_{j-k:j}) + \alpha_k\, p_j^{(k-1)}(x_j|x_{j-k+1:j-1})}{n_{j-1}(x_{j-k:j-1}) + \alpha_k},$$

which is illustrated in Figure 1. For frequently occurring k-mers $x_{j-k:j-1}$ the counts dominate over the pseudocounts and we can accurately estimate the conditional probabilities from the counts. For k-mers with few counts the pseudocounts dominate and the probability reverts to the estimate at order $k-1$, which in turn may be dominated by the estimate at order $k-2$, and so forth, down to an order where the number of counts dominates the pseudocounts. In this way, conditional probabilities are learned only for those k-mers for which they can be robustly estimated, while other conditional probabilities are approximated by robustly estimated lower-order probabilities.

We fixed the prior strengths $\alpha_0 = 1$ and $\alpha_k = \beta \times \gamma^{k-1}$ for $k \geq 1$, with hyperparameters $\beta = 20$ and $\gamma = 3$. The increasing strength of pseudocounts with increasing k reflects the prior belief that dependencies should quickly decline with distance (3). Owing to this rather strong regularization, we prevent overtraining on all datasets (see, e.g. Supplementary Figure S18). As background model, we always train a second-order homogeneous BaMM on the set of positive training sequences, with the default α setting from our tool XXmotif (39) ($\alpha_k = 10$ for all $k \geq 0$). This choice gives good performance as trimers capture the properties of background sequences in sufficient detail without learning the motifs themselves.

Motif discovery using Bayesian models

When the motif sites in the training sequences are not known a priori, we need to learn a good model and at the same time find motif instances in sequences that are often hundreds or thousands of nucleotides long. Most motif discovery algorithms train PWMs. We derieve here an expectation maximization (EM) algorithm to train BaMMs. A formal derivation is given in the Supplementary Methods.

The goal is to estimate the model parameters $\mathbf{p}_{\mathrm{motif}}^{(K)}$, which is a vector containing the $W \times 4^{K+1}$ conditional probabilities $p_j^{(K)}(x_{K+1}|x_{1:K})$ for any $K+1$-mer $x_{1:K+1}$. The EM algorithm cycles between E- and M-step. In the E-step, we re-estimate the probabilities for a motif to be present at position i of sequence n,

$$r_{ni} := p(z_n = i|\mathbf{x}_n, \mathbf{p}_{\mathrm{motif}}^{(K)}) = \frac{p(\mathbf{x}_n|z_n = i, \mathbf{p}_{\mathrm{motif}}^{(K)})}{\sum_{i'} p(\mathbf{x}_n|z_n = i', \mathbf{p}_{\mathrm{motif}}^{(K)})}.$$

We use the zero-or-one-occurrence-per-sequence (ZOOPS) model (38) (for the motivation see Supplementary Methods), and the hidden variable z_n indicates at which position the motif is present in sequence n. In the M-step we use the new r_{ni} to update the model parameters $\mathbf{p}_{\mathrm{motif}}^{(k)}$ for all orders $k = 0, \ldots, K$. This update equation looks exactly the same as the previous equation for known motifs locations, except that now the counts $n_j(x_{1:k})$ are interpreted as fractional counts computed according to

$$n_j(x_{1:k}) := \sum_n r_{ni} \mathbb{I}(x^n_{i+j-k:i-j-1} = x_{1:k}).$$

The indicator function returns 1 if the logical expression is true and 0 otherwise. The update of model parameters in the M-step runs through all orders from $k = 0$ to $k = K$, each time using the just updated model parameters from the order below. We iterate the EM algorithm until convergence.

RESULTS

Nucleotide dependencies in transcription factor binding sites

We show how BaMMs can improve by 29% the prediction of transcription factor binding sites learned from ChIP-seq-enriched sequences by modelling the correlations between nucleotides in their binding sites.

We evaluated our approach on 446 human ChIP-seq datasets for 94 sequence-specific transcription factors associated with RNA polymerase (RNAP) II from The ENCODE Project Consortium (1). Positive sequences were compiled from up to 5000 peak regions with highest confidence by extracting ±102 bp around peak summits. We initialized our BaMM learning with the motif instances that XXmotif uses to build its top PWM model but added two positions to both 5′- and 3′-ends of the models. We sampled synthetic background sequences with the same length as the positive sequences but 100 times as many, using the trimer frequencies from the positive sequences. To compare PWMs with higher-order BaMMs without any influence from other source except model order, we treated PWMs as zeroth-order BaMMs.

The performance in discriminating binding from background sequences was assessed using four-fold cross-validation, by sorting in descending order the sequences by their maximum log-odds score over all possible motif positions and recording the cumulated number of correct predictions (TP) and false predictions (FP) above a score threshold. Lowering the score threshold from maximum to minimum we trace out a curve of TP versus FP. The normalized version in which one plots the true positive rate (TPR), the fraction of TP out of all positive sequences, versus the false positive rate (FPR), the fraction of FP out of all background sequences, is called receiver operating characteristic (ROC) curve. The partial area under the ROC curve (pAUC) up to the fifth percentile of FPR (insets in Figure 2A and B) is a good measure of performance, because at $\text{FPR} > 0.05$ and $\text{TPR} = 1$ the precision, i.e. the fraction of predictions that are correct (i.e. truly bound), has already fallen to below $1/(1 + 0.05 \times 100) = 0.167$. The pAUC therefore summarises the part of the ROC curve most relevant in practice for predicting factor binding sites and is preferable over the AUC.

Figure 2A shows the ratios of pAUC for first-order BaMMs to the pAUC for PWMs (order 0) on each of the 446 datasets. On almost all sets the pAUC increases and the average relative increase is 16%. Strikingly, fifth-order BaMMs perform considerably better than first-order BaMMs, yielding an average fold pAUC increase over PWMs of 29% (Figure 2B). Also, on none of the 446 dataset they are clearly worse than PWMs, showing that overtraining is effectively prevented. Higher-orders are particularly beneficial for the more challenging datasets with low pAUC values.

Figure 2C–E illustrates the improvements for specific datasets. The precision-recall curves summarise predictive performance, showing the precision TP/(TP + FP) versus the recall (= sensitivity), the fraction of all bound sequences that are predicted at this precision.

We developed *sequence logos for higher orders* to visualise the BaMMs. We split the relative entropy $H\left(\mathbf{p}_{\text{motif}} \,\middle|\, \mathbf{p}_{\text{bg}}\right) = \sum_{\mathbf{x}} p_{\text{motif}}(\mathbf{x}) \log_2[p_{\text{motif}}(\mathbf{x})/p_{\text{bg}}(\mathbf{x})]$ into a sum of terms, one for each order. The logos show the amount of information contributed by each order *over and above what is provided by lower orders*, for each oligonucleotide and position.

The well-studied CCCTC-binding factor (CTCF) has been implicated in the establishment of topologically associating domains and the formation of regulatory chromosome interactions (40). A fifth-order BaMM for CTCF achieves 14% higher pAUC than a PWM (orange triangle in Figure 2B and C, left). The first-order sequence logo identifies the added information (Figure 2C, right). For example, at position 16 an A is preferentially followed by a G and a G by a C, relative to the zeroth-order model. The first-order dependencies may reflect the intricate interplay of a subset of CTCF's 11 zinc-finger (ZnF) domains.

Transcription factor MafK of the AP-1 family of basic-region leucine zippers (bZIP) can bind DNA as homodimer or heterodimer. Depending on its multimeric state, MafK targets the 13 bp T-MARE or the 14 bp C-MARE motif. These are composed of a 7 bp and 8 bp core sequence, respectively, flanked by GC elements on both sides (41). A fifth-order BaMM for MafK achieves a 19% higher pAUC than a PWM (green triangle in Figure 2B). Most of this improvement is already present in first order (Figure 2D, left). The first-order logo shows that two alternative DNA recognition modes are represented by the BaMM. While the T-MARE is primarily modelled in zeroth order (Figure 2D, middle), the C-MARE is modelled via first-order dependencies (Figure 2D, right). The upstream AT-rich region seen in zeroth order is revealed by the first-order logo to be a poly(dA:dT) tract. This indicates that MafK reads out the narrowed DNA minor groove width known to be induced by poly(dA:dT) tracts (42).

The bZIP transcription factor JunD binds two half-site motifs separated by one or two base pairs. The preferred spacing is cell-type-specific and depends on the availability of oligomerization partners (43). A fifth-order BaMM for JunD achieves a 96% higher pAUC than a PWM (blue triangle in Figure 2B and E, left). The model represents a mixture of two binding sites, ATGA-S-TCAT (S = G or C) and the slightly less frequent ATGA-CG-TCAT. The right half-sites TCAT are aligned in the model, whereas the left half-sites are displaced from each other by one position. In positions 3 to 6 of the zeroth-order logo, this results in mixtures of A + T, T + G, G + A and A + C (Figure 2E, middle), while in the first-order logo it results in contributions AT + TG, TG + GA and GA+AC at positions 4–6 (Figure 2E, right).

Supplementary Figures S1–S9 contain analyses for further transcription factors: BATF, c-Jun, c-Fos, Hnf4a, IRF4, NF-YB, NRSF, PU.1 and ZnF143. Remarkably, for some datasets, e.g. ZnF143, we still observe substantial improvements at order three or higher.

Improvements from flanking nucleotides

We show here that, by including the nucleotides flanking the core binding sites of transcription factors, we can substantially increase the predictive performance of BaMMs

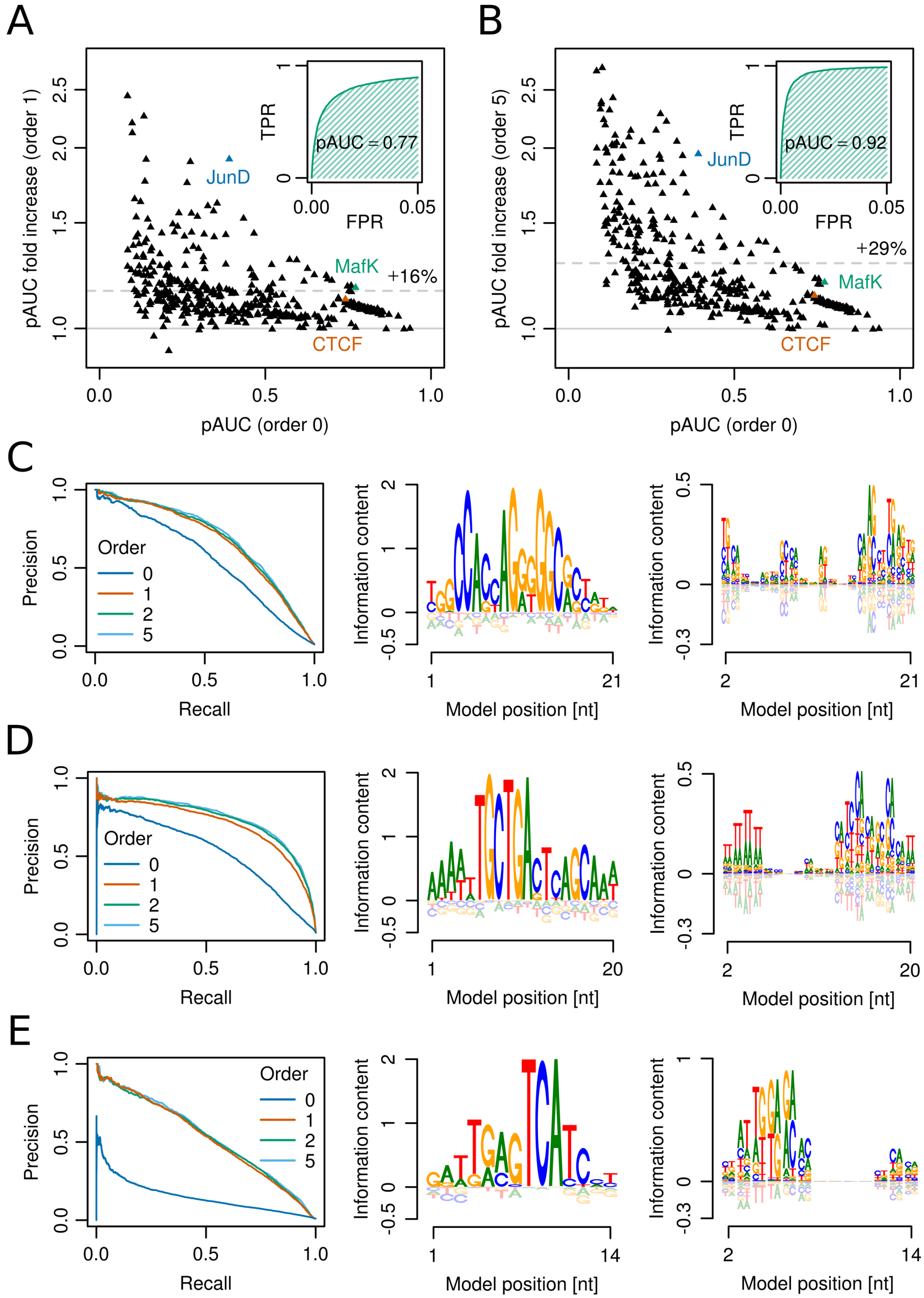


Figure 2. Modelling nucleotide dependencies in transcription factor binding motifs improves motif discovery and prediction. (**A**) Factor of increase in partial area under the ROC curve (pAUC) of first-order BaMMs versus zeroth-order BaMMs (PWMs) on 446 ChIP-seq datasets for transcription factors from ENCODE. The average performance increase is 16% (dashed line). Y-scale is logarithmic. Inset: partial ROC curve for PWM of MafK binding. (**B**) Same as (**A**) but showing the increase in pAUC of fifth-order BaMMs versus PWMs. Inset: fifth-order BaMM of MafK binding. (**C**) CTCF models learned from ChIP-seq sites in Mcf-7 cells. Predictive performance (left) for BaMMs of increasing order. Zeroth-order (middle) and first-order (right) sequence logos of second-order BaMM. (**D**) Same as (**C**) for MafK binding measured in HepG2 cells. (**E**) Same as (**C**) but for JunD binding in HepG2 cells. Positions 2–6 of the first-order model (right) encode the two half-sites ATGAC and XATGA shifted from each other by one nucleotide.

but less so for PWMs, widening the performance gain of BaMMs over PWMs to 36%.

Two recent studies pointed out that for some transcription factors the nucleotides flanking the core binding site make sizeable contributions to the binding specificity (44,45). These contributions are probably owed to shape readout around to the core site, for example the minor groove width (42). Since DNA shape is largely determined by base-stacking interactions, we expect shape readout to lead to strong next-neighbour nucleotide dependencies. If this is true, higher-order models should profit particularly from the inclusion of flanking nucleotides.

We therefore analysed the contribution of flanking nucleotides on BaMMs of various orders by comparing models of the same length as found by XXmotif with models extended by four base pairs on either side. For the 8-bp-extended models, we also extended all sequences by 8 bp to keep the same search space size. We then trained original and 8-bp-extended BaMMs.

PWMs increase their pAUC by 5% on average (Figure 3A), with a few datasets showing considerably better and others considerably worse performance with 8-bp-extended models. Nicely, fifth-order BaMMs indeed increased their pAUC much more, by 19% on average, and only a single dataset shows >5% worse performance with an extended model (Figure 3B).

When comparing 8-bp-extended PWMs with 8-bp-extended BaMMs of fifth order (Figure 3C), the average increase in pAUC was 36%. Most strikingly, extended fifth-order BaMMs significantly (at 6.25% level) outperformed the zeroth-order BaMMs and also the XXmotif models (Figure 3D) in the vast majority of datasets (97% and 99%, respectively), even though XXmotif compared favourably to state-of-the-art motif discovery tools (39).

Figure 3E shows the results for the second-order BaMM of basic helix–loop–helix (bHLH) family transcription factor USF1. The additional information in the flanking regions in first order (right logo) leads to an increase in pAUC of 20% (blue triangle in Figure 3B) and an increase of precision from 25% to 90% for low recall (Figure 3E left). The strong influence of flanking nucleotides has also been demonstrated for other bHLH transcription factors (44,46), including CBF1, a homolog of USF1 in *Saccharomyces cerevisiae*. Similar analyses for the transcription factors GR, IRF1 and c-Fos can be found in Supplementary Figure S10.

Piqued by this success, we asked how much can be gained by including a still larger sequence context around core sites. We chose CTCF for its importance in chromatin organization and extended the core model by 25 bp on either side. Again, only higher-order models profit markedly (Supplementary Figure S11). The predictive performance for the second-order model reaches an impressive recall of 52% at 95% precision, whereas the zeroth-order BaMM predicts only 14% true sites at that precision.

Quantitative prediction of binding affinities

To attain a quantitative understanding of transcriptional regulation, we need to predict accurately the factor occupancies on regulatory sequences. We demonstrate here that BaMMs trained on ChIP-seq data can predict binding affinities directly measured by biophysical methods considerably more accurately than PWMs and a number of competing methods.

Sun *et al.* (47) measured dissociation constants (K_d) of the pioneer ZnF transcription factor Klf4 for binding to various sequences by competitive EMSA experiments. Thirty three sequences had a single mutation and 25 sequences had multiple mutations to the 10 bp Klf4 consensus motif. As in Sun *et al.* (47), we computed the logarithms of each K_d divided by a K_d^{ref}. For the sequences with single mutations K_d^{ref} was their median K_d and for the sequences with multiple mutations the K_d closest to their mean.

We trained BaMMs of increasing complexity using 101 bp sequences extracted around the 5000 strongest ChIP-seq peaks from (48). We plotted the log ratios of EMSA K_d's versus the corresponding predictions from our models. We compared the performance of BaMMs of increasing order by means of the Pearson correlation between measured and predicted log K_d ratios.

Overall, the Pearson correlation improves with increasing BaMM order (solid lines in Figure 4A, left). While the zeroth-order BaMM successfully predicts Klf4 affinities to singly mutated binding sites, it fails for the multiply mutated binding sites (Figure 4A, middle). In contrast, the fifth-order BaMM succeeds on both sets (Figure 4A, right).

To confirm these results, we performed a similar analysis on a dataset of competitive EMSA measurements for 64 double-stranded oligonucleotide probes containing potential FoxA2 binding sites (49). These were correlated with predictions from the deep learning method DeepBind (50) and various other methods, which were trained on a FoxA2 ENCODE ChIP-seq dataset, and with prediction from a number of published PWMs for FoxA2. We repeated the analysis for our BaMM predictions and obtained a Spearman correlation of $r = 0.831$, better than the best competing method, DeepBind ($r = 0.814$ and 0.784) (Figure 4B). This is remarkable, as no parameters were adjusted and our BaMMs were not developed with the aim of quantitative prediction of binding affinities.

These results indicate that, at least for some factors, BaMMs of order ≥3 are required to satisfactorily predict binding affinities to low-affinity binding sites. This is reflected in the information content of the higher-order sequence logos (Supplementary Figure S13).

Predicting RNAP II transcription start sites

Whereas our previous analyses were based on simple motifs composed of a single binding site, we now assess higher-order Markov models for modelling complex motifs, regulatory regions that are targeted by multiple, cooperatively binding factors. Such complex motifs can be composed of multiple, non-obligatory submotifs with variable spacings and strengths.

The core promoter is the region of approximately ±50 bp around the transcription start site (TSS) that is required to initiate transcription. RNAP II core promoters can be classified into two classes (51), exhibiting a TSS distribution with a narrow peak (NP) or a broad peak (BP). In animals the former tend to be correlated with highly regulated genes

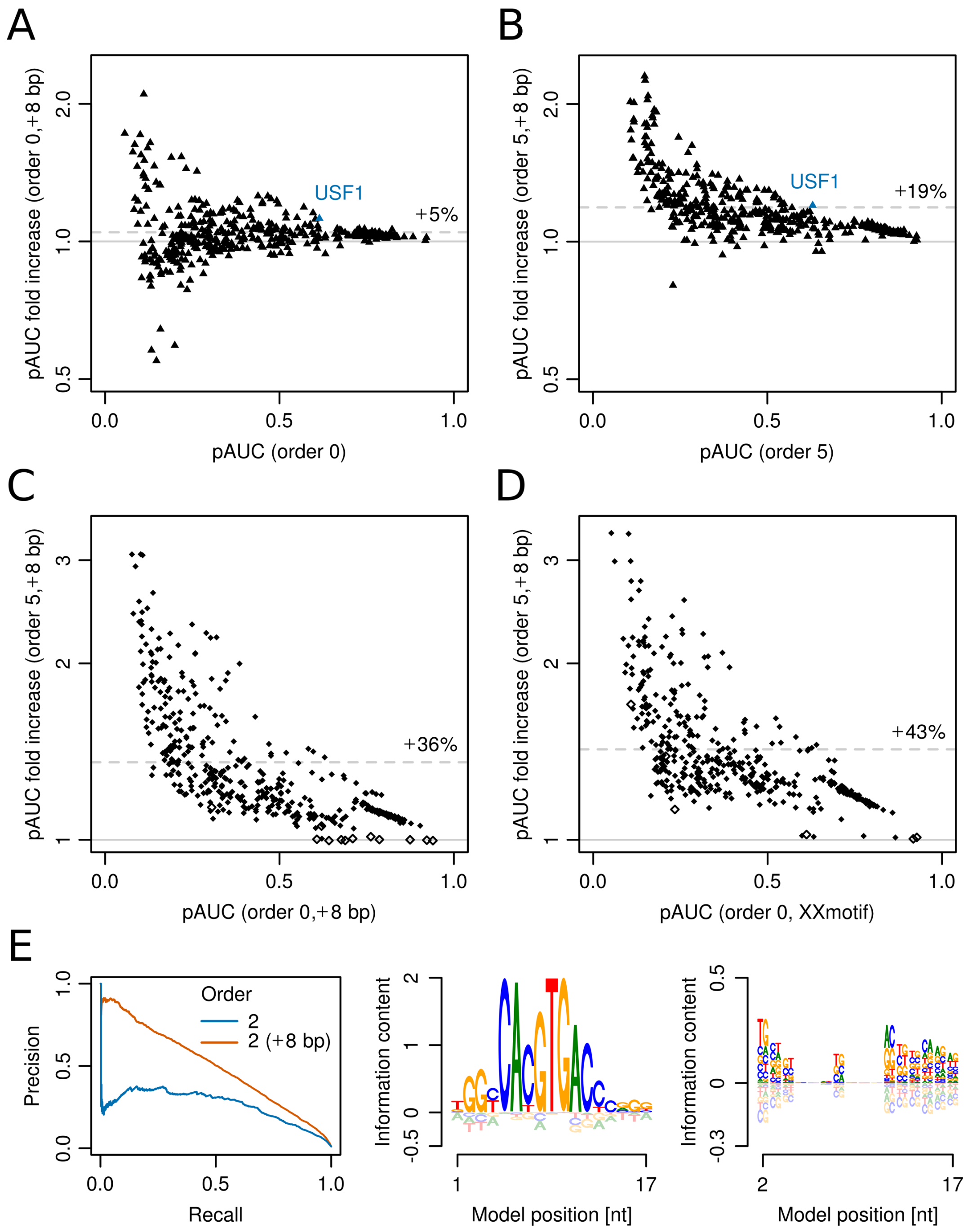


Figure 3. Nucleotides flanking the core binding sites of transcription factors may contribute greatly to the specificity of higher-order models. (**A**) Factor of increase in performance (on log scale) of 8-bp-extended versus unextended zeroth-order BaMMs (PWMs) on 446 ChIP-seq datasets for transcription factors from ENCODE. The mean increase is 5% (dashed line). (**B**) Performance increase of fifth-order 8-bp-extended versus unextended BaMMs. (**C**) Performance increase of fifth-order 8-bp-extended BaMMs versus zeroth-order 8-bp-extended BaMMs (PWMs). Significant improvements ($P = 1/16 = 6.25\%$) are obtained on 97% of all datasets (filled diamonds). The remaining 12 datasets show insignificant differences (open diamonds). (**D**) Performance increase of fifth-order 8-bp-extended BaMMs versus PWMs refined by XXmotif. (**E**) Results for 8-bp-extended USF1 model learned from ChIP-seq sites in the H1-hESC line.

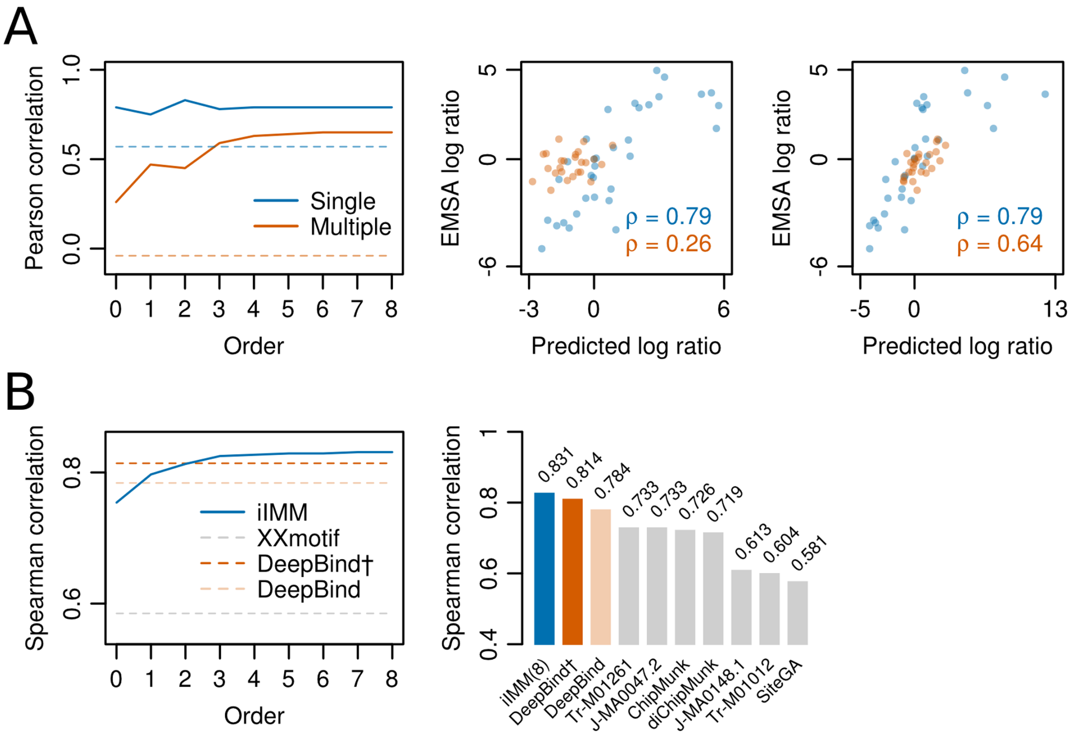


Figure 4. Higher-order BaMMs boost accuracy of binding affinity predictions for weak sites. (**A**) Left: Pearson correlation between $\log(K_d/K_d^{\mathrm{ref}})$ values for Klf4 binding to singly and multiply mutated consensus sites and $\log(K_d/K_d^{\mathrm{ref}})$ values predicted with models trained on Klf4-bound sequences from ChIP-seq. Solid lines: BaMM models; dashed lines: PWMs from XXmotif. Middle: $\log(K_d/K_d^{\mathrm{ref}})$ values measured by competitive EMSA assay versus values predicted by zeroth-order BaMM trained on ChIP-seq data. Right: same but predictions from fifth-order BaMM. Affinities to weak, multiply mutated sites (orange) are very badly predicted using a PWM (correlation 0.26) but decently using a fifth-order BaMM (correlation 0.64). (**B**) Left: same as (**A**), but showing Spearman rank correlations for FoxA2 binding affinities measured for 64 putative binding sites. DeepBind† and DeepBind differ only in model length (16 versus 24 bp). Right: Spearman correlations of our eighth-order BaMM and various other methods, adopted from Alipanahi *et al.* (50).

and more frequently carry TATA-box and Initiator motifs, while the latter are correlated with housekeeping genes and have fewer, more poorly defined motifs.

We clustered and filtered TSSs measured in *D. melanogaster* by cap analysis of gene expression (CAGE) (52), resulting in 15 971 TSS clusters, assigned to 11 536 unique genes, which we classified into 7262 NP and 8709 BP core promoters. Furthermore, we modelled ribosomal protein (RP) gene core promoters using 92 core promoter sequences, corresponding to 86 unique RP genes listed in the RPG database (53).

We again used four-fold cross-validation, training on 75% of the TSSs, testing on 25% of TSSs and pooling results of the four test sets. Because of the different peak widths for NP, BP and RP core promoters (90% of CAGE tags are contained in regions of 9, 47 and 23 bp around the peak mode, respectively), we took training sequences of lengths 109, 147 and 123 bp around the mode of each TSS peak and trained models of length 101 bp on them. We trained the second order background model on 501 bp sequences centered around TSSs. For each test sequence of 501 bp around a TSS, the position with the largest log-odds score was taken to be the predicted TSS. When the prediction was within 4, 23 or 11 bp in the case of NP, BP and RP promoters, respectively, it counted as a true positive prediction, otherwise as false positive. The precision was the fraction of sequences with true positive predictions.

Figure 5A shows the positional distribution of predictions around NP TSSs for BaMMs of order 0, 1 and 2, normalized to a random predictor with uniform density. While NP core promoters are relatively well predicted within ±4 bp by all three models, the second-order model achieves a precision of 76%, 26% higher than the precision of the PWM model (60%). These improvements are reflected by a concomitant increase in total information content of the models (middle top). For even higher model orders, the precision saturates but, crucially, does not show any signs of overfitting (Supplementary Figure S16).

The zeroth-order sequence logo of the second-order NP core promoter BaMM (Figure 5A, right) reveals the Initiator motif at the TSS, the TATA box near −32 bp, and the motif 10 and downstream promoter elements (MTE, DPE). The left inset shows first-order dependencies in the region around the TATA box, which partly arise from the variable positioning of the TATA box with respect to the TSS. The right inset covers a region of overlapping DPE and E-box motifs and gives an idea how such overlapping, alternative motifs can be represented by a first-order BaMM. (See Supplementary Figure S14 for complete sequence logos.)

The BP TSSs are evidently much harder to predict than the NP TSSs (Figure 5B), owing to the scarcity and poor information content of their motifs. For difficult cases, however, BaMMs show particularly clearcut gains: the precision achieved by a second-order BaMM of 44% is twice

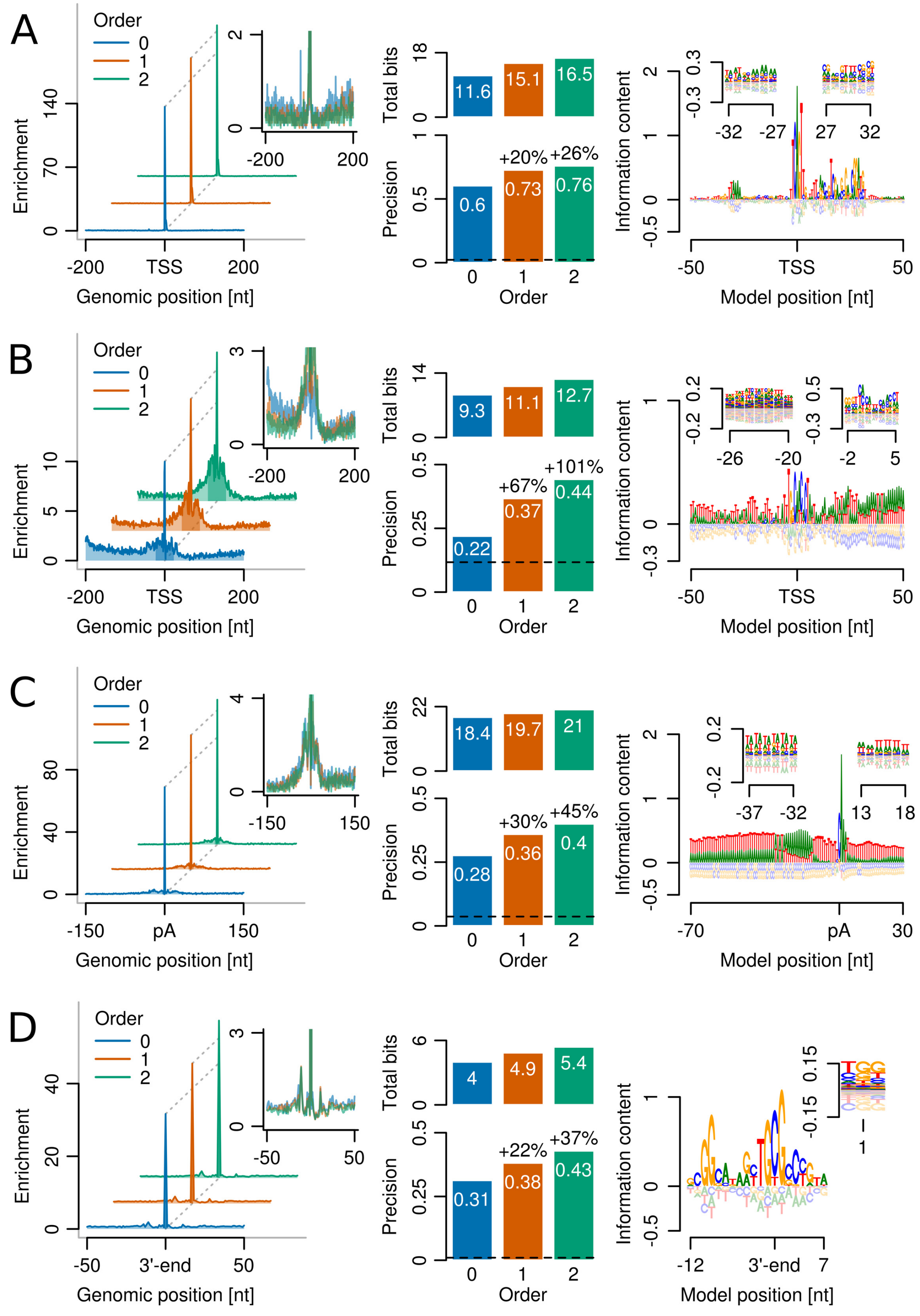


Figure 5. Higher-order BaMMs excel at predicting complex, multipartite motifs. (**A**) Left: positional distribution of TSS predictions around measured TSSs of 7262 narrow peak (NP) core promoters in *D. melanogaster*, normalized to random prediction with uniform density. Inset: same with expanded y-axis to show false predictions. Middle bottom: Fraction of sequences with correct predictions, defined to lie within 4 bp of measured TSS peak mode. Dashed line: precision of random predictor. Middle top: Total information content in BaMMs. Right: zeroth-order sequence logo of secon-order BaMM. Insets: first-order sequence logos in region covering the TATA box (left) and the DPE and E-box motifs (right). (**B**) Same as (**A**) but for TSSs of 8709 broad peak (BP) promoters. Correct predictions are defined to lie within 23 bp of measured TSSs. Logo insets show second- and first-order contributions. (**C**) Same as (**A**) but for polyadenylation (pA) sites from *S. cerevisiae*. Correct predictions are within 5 bp of measured pA sites. Logo insets show first-order contributions at efficiency and U-rich elements. (**D**) Same as (**A**) but for RNAP pause sites from *E. coli*. Correct predictions are within 0 bp of measured pause sites.

as high as for the zeroth-order model. Similar to NP core promoters, the BaMM of BP core promoters represents sequence motifs that slightly vary in positioning and also distinct, overlapping motifs (Figure 5B, right). For instance, the Ohler6 element (54) and the DNA recognition element (DRE) are admixed but distinguishable in the second-order sequence logo (left inset). Similarly, the core promoter elements Ohler1 and Ohler7 (54) are overlapping each other but are distinguishable in higher orders (right inset).

The information density in higher orders of the NP and BP models seems rather low, but it sums up to considerable sizes across the modelled region (Supplementary Figures S14 and S15). We speculate that the nucleotide dependencies in our models reflect, at least in part, DNA structural properties of core promoter regions that contribute to TSS recognition (55).

Models of RP gene core promoter sequences do not profit from higher orders (Supplementary Figure S17A). However, despite the low number of sequence instances, even a fifth-order BaMM is not overfit (Supplementary Figure S18A).

Prediction of polyadenylation sites

Sequence elements around the RNAP II polyadenylation (pA) site induce transcription termination by recruiting the cleavage and polyadenylation machinery to the pA site. In *S. cerevisiae*, 3′-end processing sequence signals were detected in the range from roughly 70 bp upstream to 30 bp downstream of the pA site (56) including the UA-rich efficiency element (EE), the A-rich positioning element (PE) and U-rich elements.

We extracted sequences of length 401 bp around 4228 pA sites in *S. cerevisiae* from 4173 unique genes using major transcript isoform annotations (57). The training and analysis was performed analagous to the core promoter prediction, using four-fold cross-validation. Training was done on the regions from −70 to +30 bp around the pA sites and training of the BaMM background model on the full 401 bp sequences. Testing was done on full-length 401 bp sequences, with true positive predictions defined to lie within 5 bp of the annotated pA site.

Again, higher-order BaMMs outperform PWM models by a wide margin, improving the 28% precision of the PWM model by 45% up to 40% for the second-order BaMM (Figure 5C, left and middle bottom). The second-order BaMM comprises all known 3′-end processing elements (Figure 5C, right). The first-order correlations are necessary to model the EE (left inset) and the downstream T-rich region (right inset), which is revealed to be a poly(dA:dT) tract in higher orders (inset and Supplementary Figure S16). Again, even models of very high order did not suffer from overtraining (Supplementary Figure S18B).

Prediction of bacterial RNAP pause sites

Pausing of RNAP during transcription has regulatory functions in RNA folding, recruitment of factors to mRNAs, and transcription termination. Larson *et al.* (58) measured RNAP pause sites in *E. coli* and *B. subtilis* using nascent elongating transcript sequencing (NET-seq) and identified 16 and 12 bp RNAP pause sequence signatures, respectively.

We extracted sequences of length 121 bp centered at 11 648 *E. coli* and 6809 *B. subtilis* pause sites. 20 bp models were trained on the regions from −12 to +7 bp around pause sites, which adds 2 bp in *E. coli* and 4 bp in *B. subtilis* to either end of the pause site motifs defined by Larson *et al.* The second-order background model was trained on the entire genome. The assessment was analogous to the TSS and pA site predictions, using 4-fold cross-validation and defining correct predictions as being precise within 0 bp.

The zeroth-order BaMM predicts 31% of *E. coli* pause sites correctly, the first-order BaMM increases the precision to 38%, and the second-order BaMMs to 43% (Figure 5D). This suggests that pause sites might have a specific signature of DNA structural properties reflected in higher-order nucleotide dependencies. Off-site predictions up- and downstream of the pause index, e.g. at −11 bp, are presumably caused by local similarities in the sequence features (Figure 5D, right).

Beside the GpG dinucleotide at the 5′-end of the RNA-DNA hybrid, 10 bp upstream of the 3′-end, another distinctive feature of the consensus sequence described by Larson *et al.* is the occurrence of TpG or CpG at the location of the 3′-end of the nascent transcript and incoming nucleoside triphospate. The CpG dinucleotide of the template strand was recently shown to inhibit elongation and induce G-to-A errors when spanning the active site of RNAP (59). Our second-order BaMM refines this signature by revealing that after a TpG a G is favoured, whereas CpG is more likely to be followed by a T or C (Figure 5D, right, inset).

In *B. subtilis*, the precision is only about half as high as in *E. coli*, but improvements through higher-order BaMMs are more marked. The third-order BaMM reaches 21% precision, an increase of 55% over zeroth-order (Supplementary Figures S17B and S18C).

Pause site models differ substantially in all orders between the gram-negative *E. coli* and the gram-positive *B. subtilis*, except for a GpG dinucleotide at the upstream edge of the RNA–DNA hybrid and a pyrimidine at the downstream edge.

Prediction of protein–RNA binding sites

In cells,mRNAs are actively kept in a largely unfolded state by energy-dependent processes (60). In contrast to DNA, which forms a relatively stiff double helix, mRNAs are therefore mostly single-stranded and extremely flexible. This leads to profound differences in the sequence specificity of RNA- versus DNA-binding. DNA sequences similar to the consensus sequence will usually be bound in a very similar overall protein–DNA conformation, whereas a single mutation from high-affinity RNA motif will usually cause the highly flexible mRNA to change its structure quite dramatically in order to minimise the binding energy. Such behaviour strongly violates the assumption that the binding energy can be approximated by independent energy contributions from each nucleotide, and it comes as no surprise that PWMs are poor models for RNA binding factors (61). We were therefore wondering whether BaMMs would be more appropriate.

We used a dataset of binding sites of 25 mRNP biogenesis factors from *S. cerevisiae* measured *in vivo* using PAR-CLIP

(62,63). We extracted 25 nt sequences centered around the crosslinked uridines of the strongest 2000 binding sites. We randomly sampled 20 000 25 nt background sequences centered around uridines in the transcriptome of *S. cerevisiae*. Our analysis of the performance to discriminate between bound and background sequences proceeds in a way analogous to the benchmark in Figures 2 and 3.

Figure 6A shows the discriminative power of second-order BaMMs compared to zeroth-order, PWM models. The performance of all models is low in comparison to DNA binders in Figure 2 when we consider that here we used a ratio of background to positive sequences of 10:1 instead of 100:1. BaMMs outperformed PWM models for all RNA-binding proteins ($P = 3 \times 10^{-8}$, Wilcoxon one-sided signed-rank test, $n = 25$), in two cases doubling the pAUC values.

The most drastic improvement was seen for the SR-like factor Hrb1 (Figure 6B). The crosslinked U at position 0 (not shown) is frequently flanked by A and G upstream and downstream, respectively (Figure 6, middle), which are also enriched around other RNA-binding protein crosslink sites (Supplementary Figure S19). In the first- and second-order logos we can discern a CUG-rich region upstream of the crosslink site (Figure 6B, right), representing up to five successive CUG repeats, which cannot be learned by the PWM due to their variable positioning. Hrb1 was not known to bind CUG-rich sequences, but our observation makes sense in light of the fact that it contains three RRM domains just like CELF1, which is known to bind CUG-rich ssRNAs (64).

Nab3 profits greatly from higher orders (Figure 6C) because the Nab3 motifs UCUU and CUUG are learned together with the Nrd1 motifs UGUA and GUAG, which are enriched near Nab3 crosslink sites (65) (Supplementary Figure S19A). A typical, very moderate gain is seen for the Mex67 adaptor protein Yra1 (Supplementary Figure S6D and Figure S19B).

DISCUSSION

In this study, we developed a Bayesian approach to train inhomogeneous Markov models that uses the conditional probabilities from lower order $k - 1$ as prior for order k. The BaMMs trained with this scheme can be regarded as a variation of interpolated Markov models. Unlike the various heuristic schemes that have been proposed for choosing the interpolation weights (34–37), in our Bayesian approach we do not need to make any ad hoc choices and merely require two hyperparameters to set the strengths of the Bayesian priors for all model orders.

Our scheme also sidesteps the common approach of pruning the discrete dependency graph of the Markov model in order to limit the model's complexity (23,31–33). Instead, we use continuous, soft, data-driven cut-offs which are effectively realized using Bayesian priors. We thereby avoid the cumbersome discrete optimization of a dependency graph and can make use of simple and effective optimization methods. This also allowed us to develop an EM-based algorithm for motif discovery using BaMM training.

We tested our BaMMs in a cross-validation setting on hundreds of ENCODE ChIP-seq datasets, RNAP II core promoter sequences and polyadenylation sites, bacterial RNAP pause sites and RNA-bound sites from PAR-CLIP measurements, using the same hyperparameters. On all datasets, BaMMs yielded sizeable improvements, typically around 30–40% increase in precision. BaMMs of order 5 led to a significant improvement (at significance level 0.0625) over PWMs on 97% of the 446 ENCODE ChIP-seq datasets, while the performance was very similar on the remaining 3% (Figure 3C). For comparison, the Markov models recently introduced in the JASPAR database (TFFMs) showed significantly improved performance on only 21% of 96 ENCODE ChIP-seq datasets (66). Fifth-order BaMMs also showed 12% better performance on average than first-order BaMMs on the 446 ENCODE ChIP-seq datasets (Figures 2A and B, Supplementary Figure S21).

A strength of BaMMs is that one does not need to decide on a case-by-case basis which order to choose, since there is no disadvantage in always choosing a relatively high order such as $k = 5$ or $k = 8$: Despite the theoretically high number of parameters at larger orders, BaMMs never deteriorated in performance with increasing order k, in contrast to simple inhomogeneous Markov models, which were prone to overtraining (Supplementary Figures S12, S18 and S20). And in many cases the performance of BaMMs did still improve up to quite high orders (see e.g. Figure 4 and Supplementary Figure S21).

Given this success with modelling nucleotide correlations, why did k-mer-based methods not clearly outperform PWM-based models in the DREAM5 challenge (22)? First, whereas ChIP-seq measurements are done with full-length transcription factors, many of which possess multiple DNA-binding domains and binding partners, PBM measurements are mostly done with single DNA binding domains. Second, we think that, in contrast to ChIP-seq measurements, the amount of information present in a PBM measurement is often not enough to learn a model more detailed than a PWM. To see why this might be the case, we note that each 8-mer occurs 16 times and each 10-mer occurs once on the PBMs used in (22). Due to the measurement noise, usually only the affinities to 8-mers are considered, since these can be averaged over the 16 measurements. However, for many DNA-binding domains the nucleotides flanking the core 8-mer probably have a considerable influence on binding strength. This means that the 16 measurements of each 8-mer are convoluted by the effects of the flanking nucleotides and could therefore be too unreliable to allow for training complex models with many parameters. Another way to look at the question is this: in a PBM measurement typically only 1% of probes, i.e. 400, are significantly bound and will carry most of the information, which might not be enough to estimate the $(4^2 - 1) \times 10 = 150$ parameters of a first-order model reliably enough.

We developed higher-order sequence logos to visualise the information learned in the various orders on top of what is contained in lower orders. As illustrated in several examples, the logos could often explain the origin of the added value in higher orders.

For transcription factors, the added value is owed to variable submotif spacings, variable dimerization partners, and DNA shape readout, neither of which can be adequately

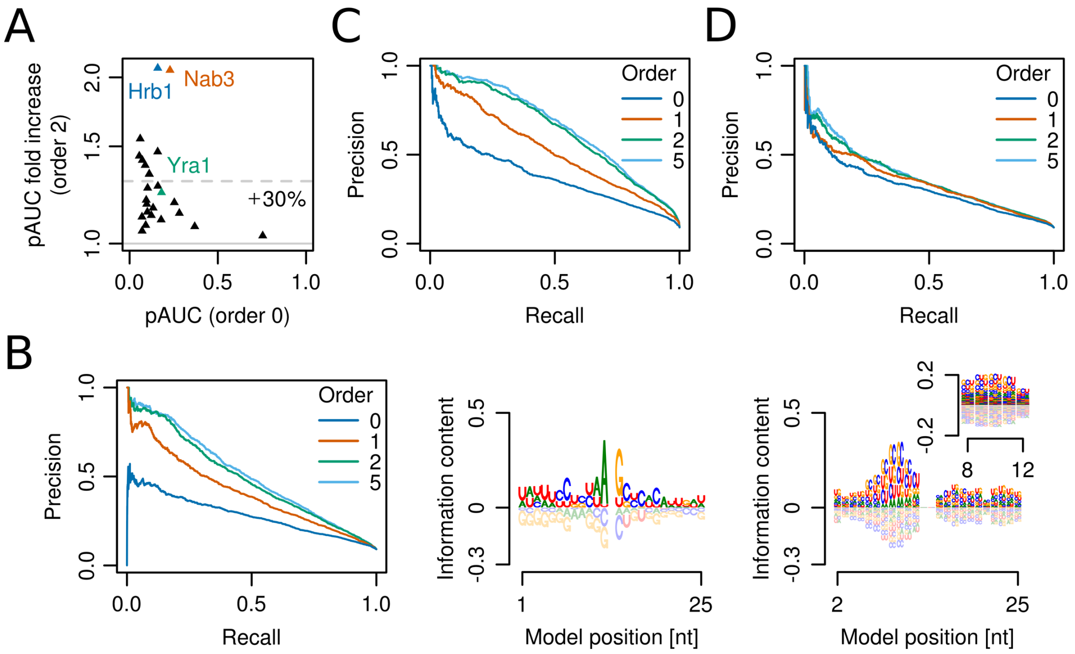


Figure 6. Modelling the sequence specificity of RNA binding factors is challenging but improves with higher orders. (**A**) Increase of prediction performance of second- versus zeroth-order BaMMs for binding sites of 25 mRNP biogenesis factors from *S. cerevisiae* measured by PAR-CLIP. Dashed line: mean fold increase. (**B**) Left: higher-order BaMMs lead to sizeable gains in precision and recall for predicting Hrb1 binding sites. Right: sequence logos for order 0 and 1 of second-order BaMM (central crosslinked U was removed from the zeroth-order logo). (**C** and **D**) Effect of BaMMs model order on prediction performance for Nab3 (C) and Yra1 (D).

learned with PWMs. Variable submotif spacings could also be learned by profile HMMs or mixtures of PWMs, however they lack flexibility to model dinucleotide preferences due to DNA shape readout and to describe more complex architectures with variable presence of motifs. Also, HMM training is slow as it requires running a forward-backward algorithm in each iteration of the EM algorithm.

Extension of the core transcription factor binding motif by 8 bp improved fifth-order BaMMs by 19% but PWMs by only 5%, due to the ability of higher orders to model the DNA shape constraints around the core binding site. Implicitly learning structural properties might work better than explicitly including them in the model (67), since *any* DNA physical property will be reflected in specific, learnable oligonucleotide preferences.

We tested the ability of PWMs and BaMMs trained on ChIP-seq data to predict binding affinities for two transcription factors measured by EMSA on datasets of sequences near their consensus binding sequences. BaMMs improved predictions of PWMs and a number of other methods. The improvements are likely owed to weak sites more than a single substitution away from the consensus. On singly mutated sequences, the PWM predicted binding affinities as well as the BaMM, while on doubly mutated sites it showed a dismal correlation of 0.26 while the BaMM achieved 0.64 (Figure 4). Hence PWMs learn to predict mostly the high-affinity sites correctly, as the energies of all sequences a single mutation away from consensus can still be described with their simple energy model, and the breakdown of the PWM performance for doubly mutated sites reflects the breakdown of the additivity assumption according to which nucleotides contribute individually to the binding energy. As low-affinity sites have been reported to be important for the specificity and robustness of gene expression (68), improvements in predicting binding to weak sites will be important for quantitative modelling of transcriptional regulation.

Complex, multipartite motifs profited from the flexibility of BaMMs to represent multiple submotifs at variable spacings and strengths. This flexibility may be useful to predict binding sites of cooperatively binding transcription factors, which often prefer certain spacings and orientations (15,69). It might prove particularly powerful for predicting binding sites of factors with multiple DNA-binding domains. Zinc fingers (ZnFs) comprise up to a third of human transcription factors (70) and they contain on average 10 DNA-binding domains, which might partly explain why clearly defined motifs could be found for 8% of them (3). Although the 15 ββα-type ZnFs (http://v1.factorbook.org) in our ENCODE ChIP-seq datasets improve about as much as the other tested factors between zeroth and fifth order (Supplementary Figure S22), the striking prediction performance and specificity we observed for the fifth order BaMM of Znf143 (Supplementary Figure S9) and a 67-bp-long model of CTCF (Supplementary Figure S11) indicates that the complex binding sites of some ZnF transcription factors with multiple DNA-binding domains might be well predictable using long BaMMs.

At present, a limitation of BaMMs in comparison to Bayesian network models (e.g. (23,32)) is that nucleotide dependencies are only modelled within consecutive *k*-mers. Overcoming this would be useful for transcription factors that change their binding mode depending on the sequence and to learn complex motif architectures with correlated submotif occurrences, for example. It seems straightforward to generalise the presented approach by making each position j dependent on up to k not necessarily neighbouring upstream positions using a heuristic selection strategy. Because the described EM algorithm is fast—a fifth-order BaMM with 21 positions from 5000 sequences of 205 bp is learned within 83 s on a single core of a 3.4 GHz Intel Core i7-2600 CPU—we may choose a generous value $k = 5$ or even higher in practice.

To further improve the performance of our BaMM-based motif discovery tool BaMM!motif, we will learn automatically from the data (i) the hyperparameters and (ii) the model length, a nontrivial task since the changing number of parameters precludes a simple optimization of the likelihood or posterior probability. We will also (iii) develop an efficient method to estimate the biological significance of discovered motifs, and (iv) introduce positional priors. (v) Since BaMM!motif's speed is at present limited by the seed motif discovery code from XXmotif, we will accelerate this code substantially. One future challenge will also be (vi) to develop a rigorous approach that can deal with quantitative data such as fluorescence intensities from HT-SELEX and protein binding microarrays and peak strengths of ChIP-seq measurements.

CONCLUSION

We have developed a Bayesian approach to train higher-order Markov models, which automatically adapts model complexity to the amount of available data position- and k-mer-specifically. The BaMMs learned with this scheme were never overtrained, even at high orders. To our knowledge this is the first method for learning the dependency graph among motif positions of higher-order Markov models that does not require a-priori knowledge of motif locations and that can hence be applied to de-novo motif discovery.

The most remarkable result of this study is the consistency with which higher-order BaMMs yielded solid improvements across various heterogeneous datasets without requiring parameter tuning on each dataset and without a single case of failure. We can therefore answer affirmatively the question of whether nucleotide correlations are significant in transcription factor binding sites and other regulatory regions.

These results argue in favour of making the transition from PWMs to BaMMs as the standard model to describe protein–DNA binding affinities and to offer BaMM models in databases for regulatory and binding site motifs.

AVAILABILITY

The Bayesian Markov Model motif discovery software BaMM!motif is available under GPL at http://github.com/soedinglab/BaMMmotif.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Holger Hartmann for help with XXmotif, Mark Heron and Phillipp Torkler for help with core promoter and PAR-CLIP data, respectively, and all members of the Söding lab for fruitful discussions.

FUNDING

German Research Foundation (DFG) [GRK1721, SFB646]; German Federal Ministry of Education and Research (BMBF) within the frameworks of e:Med and e:Bio [e:AtheroSysMed 01ZX1313D, SysCore 0316176A]; Bavarian Center for Molecular Biosystems (BioSysNet). Funding for open access charge: Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
3. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
4. Najafabadi,H.S., Mnaimneh,S., Schmitges,F.W., Garton,M., Lam,K.N., Yang,A., Albu,M., Weirauch,M.T., Radovani,E., Kim,P.M. *et al.* (2015) C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.*, **33**, 555–562.
5. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
6. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Soboleva,A.V., Kasianov,A.S., Ashoor,H., Ba-Alawi,W., Bajic,V.B., Medvedeva,Y.A., Kolpakov,F.A. *et al.* (2016) HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, **44**, D116–D125.
7. Pachkov,M., Balwierz,P.J., Arnold,P., Ozonov,E. and van Nimwegen,E. (2013) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.*, **41**, D214–D220.
8. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
9. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
10. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
11. Zuo,Z. and Stormo,G.D. (2014) High-resolution specificity from DNA sequencing highlights alternative modes of lac repressor binding. *Genetics*, **198**, 1329–1343.
12. Nakahashi,H., Kwon,K.-R.K., Resch,W., Vian,L., Dose,M., Stavreva,D., Hakim,O., Pruett,N., Nelson,S., Yamane,A. *et al.* (2013) A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.*, **3**, 1678–1689.
13. Fordyce,P.M., Pincus,D., Kimmig,P., Nelson,C.S., El-Samad,H., Walter,P. and DeRisi,J.L. (2012) Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E3084–E3093.
14. Meijsing,S.H., Pufall,M.A., So,A.Y., Bates,D.L., Chen,L. and Yamamoto,K.R. (2009) DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, **324**, 407–410.
15. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015)

DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.
16. FitzGerald,P.C., Sturgill,D., Shyakhtenko,A., Oliver,B. and Vinson,C. (2006) Comparative genomics of Drosophila and human core promoters. *Genome Biol.*, **7**, R53.
17. Ohler,U. (2006) Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res.*, **34**, 5943–5950.
18. 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.
19. Bulyk,M.L., Johnson,P.L.F. 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.
20. 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.
21. Zhao,Y. and Stormo,G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, **29**, 480–483.
22. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
23. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites. In: *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*. ACM Press, NY, pp. 28–37.
24. Hannenhalli,S. and Wang,L.-S. (2005) Enhanced position weight matrices using mixture models. *Bioinformatics*, **21**(Suppl. 1), i204–i212.
25. Georgi,B. and Schliep,A. (2006) Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics*, **22**, e166–e173.
26. Huang,W., Umbach,D.M., Ohler,U. and Li,L. (2006) Optimized mixed Markov models for motif identification. *BMC Bioinformatics*, **7**, 279.
27. Maaskola,J. and Rajewsky,N. (2014) Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res.*, **42**, 12995–13011.
28. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781–790.
29. Kulakovskiy,I., Levitsky,V., Oshchepkov,D., Bryzgalov,L., Vorontsov,I. and Makeev,V. (2013) From binding motifs in ChIP-seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **11**, 1340004.
30. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
31. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
32. Ben-Gal,I., Shani,A., Gohr,A., Grau,J., Arviv,S., Shmilovici,A., Posch,S. and Grosse,I. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, **21**, 2657–2666.
33. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, **43**, e119.
34. Jelinek,F. and Mercer,R.L. (1980) Interpolated estimation of Markov source parameters from sparse data. In: *Proceedings of the Workshop on Pattern Recognition in Practice*. pp. 381–397.
35. Ristad,E. and Thomas,R.G. (1997) Nonuniform Markov models. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 791–794.
36. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.
37. Ohler,U., Harbeck,S., Niemann,H., Nöth,E. and Reese,M.G. (1999) Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics*, **15**, 362–369.
38. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, pp. 28–36.
39. Hartmann,H., Guthöhrlein,E.W., Siebert,M., Luehr,S. and Söding,J. (2013) P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.*, **23**, 181–194.
40. Ong,C.-T. and Corces,V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234–246.
41. Kurokawa,H., Motohashi,H., Sueno,S., Kimura,M., Takagawa,H., Kanno,Y., Yamamoto,M. and Tanaka,T. (2009) Structural basis of alternative DNA recognition by Maf transcription factors. *Mol. Cell. Biol.*, **29**, 6232–6244.
42. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
43. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.
44. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.
45. Levo,M., Zalckvar,E., Sharon,E., Dantas Machado,A.C., Kalma,Y., Lotan-Pompan,M., Weinberger,A., Yakhini,Z., Rohs,R. and Segal,E. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.*, **25**, 1018–1029.
46. Mordelet,F., Horton,J., Hartemink,A.J., Engelhardt,B.E. and Gordân,R. (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics*, **29**, i117–i125.
47. Sun,W., Hu,X., Lim,M.H.K., Ng,C.K.L., Choo,S.H., Castro,D.S., Drechsel,D., Guillemot,F., Kolatkar,P.R., Jauch,R. *et al.* (2013) TherMos: estimating protein-DNA binding energies from in vivo binding profiles. *Nucleic Acids Res.*, **41**, 5555–5568.
48. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
49. Levitsky,V.G., Kulakovskiy,I.V., Ershov,N.I., Oshchepkov,D.Y., Makeev,V.J., Hodgman,T.C. and Merkulova,T.I. (2014) Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-seq data. *BMC Genomics*, **15**, 80.
50. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
51. Rach,E.A., Yuan,H.-Y., Majoros,W.H., Tomancak,P. and Ohler,U. (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol.*, **10**, R73.
52. Brown,J.B., Boley,N., Eisman,R., May,G.E., Stoiber,M.H., Duff,M.O., Booth,B.W., Wen,J., Park,S., Suzuki,A.M. *et al.* (2014) Diversity and dynamics of the Drosophila transcriptome. *Nature*, **512**, 393–399.
53. Nakao,A., Yoshihama,M. and Kenmochi,N. (2004) RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res.*, **32**, D168–D170.
54. Ohler,U., Liao,G., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, RESEARCH0087.
55. Durán,E., Djebali,S., González,S., Flores,O., Mercader,J.M., Guigó,R., Torrents,D., Soler-López,M. and Orozco,M. (2013) Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. *Nucleic Acids Res.*, **41**, 7220–7230.
56. Tian,B. and Graber,J.H. (2012) Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA*, **3**, 385–396.
57. Pelechano,V., Wei,W. and Steinmetz,L.M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.

58. Larson,M.H., Mooney,R.A., Peters,J.M., Windgassen,T., Nayak,D., Gross,C.A., Block,S.M., Greenleaf,W.J., Landick,R. and Weissman,J.S. (2014) A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, **344**, 1042–1047.
59. Imashimizu,M., Takahashi,H., Oshima,T., McIntosh,C., Bubunenko,M., Court,D.L. and Kashlev,M. (2015) Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol.*, **16**, 98.
60. Rouskin,S., Zubradt,M., Washietl,S., Kellis,M. and Weissman,J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.
61. Jankowsky,E. and Harris,M.E. (2015) Specificity and nonspecificity in RNA-protein interactions. *Nat. Rev. Mol. Cell. Biol.*, **16**, 533–544.
62. Baejen,C., Torkler,P., Gressel,S., Essig,K., Söding,J. and Cramer,P. (2014) Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Mol. Cell*, **55**, 745–757.
63. Schulz,D., Schwalb,B., Kiesel,A., Baejen,C., Torkler,P., Gagneur,J., Söding,J. and Cramer,P. (2013) Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell*, **155**, 1075–1087.
64. Edwards,J.M., Long,J., de Moor,C.H., Emsley,J. and Searle,M.S. (2013) Structural insights into the targeting of mRNA GU-rich elements by the three RRMs of CELF1. *Nucleic Acids Res.*, **41**, 7153–7166.
65. Creamer,T.J., Darby,M.M., Jamonnak,N., Schaughency,P., Hao,H., Wheelan,S.J. and Corden,J.L. (2011) Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet.*, **7**, e1002329.
66. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
67. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.
68. Crocker,J., Abe,N., Rinaldi,L., McGregor,A.P., Frankel,N., Wang,S., Alsawadi,A., Valenti,P., Plaza,S., Payre,F. *et al.* (2015) Low affinity binding site clusters confer Hox specificity and regulatory robustness. *Cell*, **160**, 191–203.
69. Spivak,A.T. and Stormo,G.D. (2016) Combinatorial cis-regulation in Saccharomyces species. *G3 (Bethesda)*, **6**, 653–667.
70. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.