doi: 10.1093/nargab/lqab026

Bayesian Markov models improve the prediction of binding motifs beyond first order

Wanwan Ge, Markus Meier, Christian Roth and Johannes Söding*

Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Received November 2, 2020; Revised March 11, 2021; Editorial Decision March 22, 2021; Accepted March 30, 2021

ABSTRACT

Transcription factors (TFs) regulate gene expression by binding to specific DNA motifs. Accurate models for predicting binding affinities are crucial for quantitatively understanding of transcriptional regulation. Motifs are commonly described by position weight matrices, which assume that each position contributes independently to the binding energy. Models that can learn dependencies between positions, for instance, induced by DNA structure preferences, have yielded markedly improved predictions for most TFs on *in vivo* data. However, they are more prone to overfit the data and to learn patterns merely correlated with rather than directly involved in TF binding. We present an improved, faster version of our Bayesian Markov model software, BaMMmotif2. We tested it with state-of-the-art motif discovery tools on a large collection of ChIP-seq and HT-SELEX datasets. BaMMmotif2 models of fifth-order achieved a median false-discovery-rate-averaged recall 13.6% and 12.2% higher than the next best tool on 427 ChIP-seq datasets and 164 HT-SELEX datasets, respectively, while being 8 to 1000 times faster. BaMMmotif2 models showed no signs of overtraining in cross-cell line and cross-platform tests, with similar improvements on the next-best tool. These results demonstrate that dependencies beyond first order clearly improve binding models for most TFs.

INTRODUCTION

Gene expression is regulated through the binding of transcription factors (TFs) to specific recognition motifs within promoter and enhancer DNA sequences. These binding motifs typically contain 6 to 12 only partially conserved bases (1–3). Learning quantitative models from experimental data that allow us to accurately predict the binding affinities of TFs to any given sequence is important for quantitatively predicting transcription rates from regulatory sequences.

The task of *de novo* motif discovery is to infer from experimental data a statistical or thermodynamic model that can then predict the binding affinity of a TF of interest for any sequence up to a constant (see Supplementary Methods subsection S1.2). Motif models can be inferred from numerous types of experiments (4). Common *in vivo* techniques are ChIP-seq (5) and bacterial-one-hybrid (6), while most modern *in vitro* approaches are SELEX-based (7–9). These measurements result in sets of hundreds to millions of bound sequences from which the binding motif model is deduced based on the statistical enrichment of binding sites compared to a background set of unbound sequences or a background model for random sequences.

The dominant model for describing the binding affinity of transcription factors to DNA target sequences has been the position weight matrix (PWM). This model assumes that the binding energy can be decomposed into a sum of contributions from each of the nucleotides in the binding site. By Boltzmann's law, this is equivalent to assuming statistical independence between nucleotides at different positions of the binding site. The PWM model has been enormously successful because for the vast majority of transcription factors it achieves quite high accuracy for predicting the binding affinity of high-affinity binding sites with only $3W$ parameters for a binding site of W nucleotides. However, modeling the nucleotide inter-dependency often yields better motif predictions than PWMs (10–12). One reason is that the stacked, neighboring bases largely determine the physical properties of DNA, such as their equilibrium bending angle, minor groove width, propeller twist or helical twist. The information on the geometric orientation of the bases propagates within the DNA for several positions before fading out, creating a dependence of the DNA physical properties on nucleotide pairs, triplets and longer k-mers. Since TFs recognize their target sites not only using hydrogen bonds but also using their structural fit, TF-binding motifs show preferences depending on k-mer words (13), particularly in the flanking regions outside the hydrogen bonding core region (14). Furthermore, alternative bind-

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

ing modes of TFs (15,16) can lead to poor performance of PWMs.

During the past decade, it has become increasingly evident that weak binding sites in enhancers and promoters play an important role in determining transcriptional activity (17–21), and PWMs have limitations to describe the affinities for weak binding sites accurately. Therefore, various more refined models have been developed that depart from the simplifying assumption of independence of motif positions (22–24). Prime among them are inhomogeneous Markov models of order k, in which the probability to observe a certain nucleotide at position i depends on the previous k nucleotides at $i - k$ to $i - 1$. A zeroth-order Markov model is therefore equivalent to a PWM. Dinucleotide weight matrices (DWMs) are equivalent to first-order models, in which the probability of a nucleotide depends on its direct predecessor, and they have shown improved accuracy over PWMs (25–27).

For Markov models of higher order k, the large number of $W \times (4^{k+1} - 1)$ parameters can lead to overfitting on the training data and hence bad predictive performance. To address this limitation, our group had proposed a special type of Markov model, the Bayesian Markov model (BaMM) (28), in which the probability for a nucleotide at position i of the motif, for example the last nucleotide in `ACTCG`, is estimated by adding to the actual counts of `ACTCG` pseudo counts based on how often the shorter $(k - 1)$-mer `CTCG` has been observed in the binding sites. The probability for `CTCG` in turn is estimated by adding its counts to pseudo counts based on how often the word of length $k - 1$, `TCG`, has been observed, and so forth. This procedure can be derived formally in a Bayesian framework with Dirichlet priors. Our software BaMMmotif indeed improved on previous PWM-based methods for *de novo* motif discovery and binding site prediction on *in vivo* data (28).

Here we present BaMMmotif2, an open-source software written entirely from scratch in C++. It contains a novel algorithm for its seed finding stage, which gives it greatly improved speed and slightly improved sensitivity in comparison to BaMMmotif. We improved the robustness of the BaMM-based motif refinement stage using sequence masking. BaMMmotif2 can also learn positional preference profiles for binding site locations from the training data.

Higher-order models have the ability to learn several low-order motifs overlaid on top of each other (29). It was therefore surmised that at least a part of the improvements of higher-order models on cross-validation benchmarks using ChIP-seq sequences could stem from learning not only the main binding motif of the ChIPped factor but also, overlaid, the binding motifs of cooperating factors whose binding sites tended to co-occur with it (18). This would of course defeat the purpose of learning the binding affinity of the ChIPped factor. In a different cell type, for instance, in which different co-binding factors are expressed, such a mixed motif might perform badly. It has also been suggested that more complex models could learn complex, non-specific sequence biases characteristic of the measurement technique, which would allow them to be distinguished from the background sequences. These platform-dependent biases could result from the library preparation, amplification, and ligation biases (30).

We therefore designed a set of benchmark experiments with a focus on detecting such overfitting (Figure 1): (i) 5-fold cross-validation on ChIP-seq and HT-SELEX data; (ii) cross-cell-line validation on ChIP-seq data for the same TFs; (iii) model training on ChIP-seq data and testing on HT-SELEX data for the same TFs and (iv) vice versa. Scheme (I) examines how the models generalize to unseen data, especially when data are limited.

Our results demonstrate that BaMMmotif2 does not show signs of overfitting but rather learns the binding affinity of only the factor of interest, and that BaMMmotif2 is the most sensitive and fastest tool among the ones tested here. Furthermore, BaMMmotif2 keeps improving the performance with increasing model orders and scales better with larger datasets.

MATERIALS AND METHODS

The BaMMmotif2 algorithm

BaMMmotif2 consists of a seeding stage and a motif refinement stage. The purpose of the seeding stage is to exhaustively identify motifs enriched in the input sequences in comparison to a second-order Markov background model trained also on the input sequences. Each of the motifs below a P-value cut-off is refined by the BaMM-based refinement stage.

The fast seeding stage. This method is described in detail in Supplementary Section S1.1. Briefly, we first count the number of occurrences of each non-degenerate W-mer word in $\{A, C, G, T\}^W$ ($W = 8$ in this study) in the input sequences. From here on, we only inspect the count array and not the sequences anymore, making the runtime of the seeding stage almost independent of the input set size. By default, reverse complements are mapped to the alphabetically lower of the two W-mers.

For each W-mer, an enrichment z-value is calculated, which is the number of standard deviations with which the observed W-mer count surpasses its expected count. The expected count is calculated using a second-order homogeneous Markov model as a background sequence model, trained on the input sequences. Following the idea of (31), we determine all locally optimal W-mers. These are the W-mers with a better enrichment z-value than any of its direct neighbors one substitution away. We use each of the locally optimal W-mers to initialize a search for locally optimal W-mer patterns in the 10-letter IUPAC alphabet {A, C, G, T, S, W, R, Y, M, K, N}, where the last six letters stand for C or G, A or T, A or G, C or T, A or C, and G or T, respectively. For each such locally optimal IUPAC pattern, a PWM is derived from all matches in the input sequences to the degenerate pattern. The PWMs are then refined by applying the expectation-maximization (EM) algorithm in the multiple-occurrences-per-sequence (MOPS) model. We merge PWMs together that overlap by at least $W - 2$ highly similar matrix columns. Finally, the PWMs are reranked by their AvRec scores (explained in the next section) and written into an output file in MEME format (32), which is passed to the refinement stage.

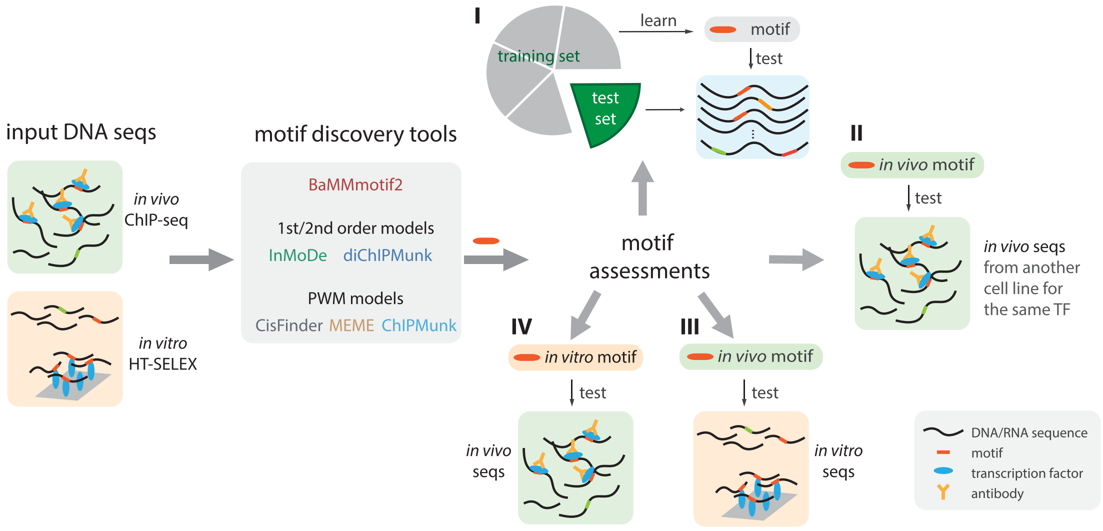

Figure 1. Benchmark pipeline for *de novo* motif discovery. Five state-of-the-art motif discovery tools and BaMMmotif2 learned motif models on *in vivo* and *in vitro* transcription factor binding datasets. The learned models were then assessed (I) by 5-fold cross-validation on the same type of data, (II) by cross-cell-line validation, and (III, IV) by cross-platform validations.

The refinement stage. The refinement stage is initialized with the motif occurrences found by the PWMs passed to it from the seeding stage. The length of the motif is extended by 2 bp on both ends by default to ensure that we do not miss information in the flanking regions. Each seed model is refined into an inhomogeneous Bayesian Markov model (BaMM) of order K using the EM algorithm (Supplementary Section S1.5). Each such refinement is independent of the refinements of the other seed motifs. Motifs can overlap with motifs already discovered in a previous refinement stage. A BaMM is an interpolated Markov model in which the conditional probability of base $x_i \in \{A, C, G, T\}$ at position i is calculated by combining the counts $n_i(x_{i-k:i})$ of k-mer $x_{i-k:i}$ with pseudo counts estimated from lower-order probabilities $p_i^{\mathrm{BaMM}}(x_i|x_{i-k+1:i-1})$:

$$p_i^{\mathrm{BaMM}}(x_i|x_{i-k:i-1}) = \frac{n_i(x_{i-k:i}) + \alpha_k\, p_i^{\mathrm{BaMM}}(x_i|x_{i-k+1:i-1})}{n_{i-1}(x_{i-k:i-1}) + \alpha_k}.$$

Here, the hyper-parameter α_k determines how much weight to give to the lower-order. The probabilities of order $k-1$ are again obtained by adding to the k-mer count the pseudo counts from order $k-2$, and so on down to order 0. In this way, when the number of occurrences observed for $(k+1)$-mer $x_{i-k:i}$ is much smaller than the number of pseudo counts $\alpha_k \times p_i^{\mathrm{BaMM}}(x_i|x_{i-k+1:i-1})$, the higher order falls back to the lower order: $p_i^{\mathrm{BaMM}}(x_i|x_{i-k:i-1}) \approx p_i^{\mathrm{BaMM}}(x_i|x_{i-k+1:i-1})$. In this way, BaMMs adapt the order that is learned in a data- and motif position-specific fashion to the amount of data (k-mer counts) available. We assume that the correlation between nearby bases declines with their distance. This is reflected in the pseudo-parameters α_k increasing with order k. For BaMMmotif2, we kept the same setting as in BaMMmotif, $\alpha_k = 7 \times 3^k$.

The motif model is optimized with the EM algorithm by maximizing the likelihood of the input sequences assuming zero or one motif occurrence per sequence (the ZOOPS model). It models the bound sequence using a Kth-order inhomogeneous BaMM $p_{\mathrm{motif}}^{K}(\mathbf{x})$ (where $\mathbf{x} = x_{1:W}$ is the binding site), and models the other unbound sequence regions using a K'th-order homogeneous BaMM $p_{\mathrm{bg}}^{K'}(\mathbf{x})$ (K' is 2 by default). This background sequence model is trained by default on the input sequences. Potential binding site sequences $\mathbf{x}$ are ranked by their score $S(\mathbf{x}) = \log(p_{\mathrm{motif}}^{K}(\mathbf{x})/p_{\mathrm{bg}}^{K'}(\mathbf{x}))$. In the weak binding limit, this score is proportional to the Gibbs free energy ΔG of binding (Supplementary Section S1.5).

The ZOOPS model is used for its computational convenience. Since actually more than one protein can bind to a sequence, the many-motif-occurrences-per-sequence model would be more appropriate. If the protein can bind in more than one conformation and thereby with more than one distinct motif, ideally all distinct motifs should be modeled and learned at once, using dynamic programming to sum over all possible binding configurations (33).

Learning positional binding preferences. BaMMmotif2 can learn the positional binding preferences for enriched motifs with respect to the center of the input sequences. By aligning the sequences around some anchor feature, such as a transcriptional start site, a 3' splice site, or a binding site of some other transcription factor, the distance preference between enriched motifs and the reference feature can be learned. We parameterize the positional probability distribution with one parameter per position and ensure smoothness by adding L_2 penalties for the differences between successive sequence positions (Supplementary Section S1.5).

Masking sequences during the motif refinement stage. Sequences from *in vivo* experiments such as ChIP-seq commonly contain several distinct motifs from other TFs that together co-regulate their target genes. This can create two types of problems during the refinement stage. First, instead of refining the motif from the seeding stage, the model in some cases tends to learn two or even more motifs in the same higher-order model, as this often improves the likelihood on the training data. Second, if the seed motif is less enriched or less informative than other motifs in the positive sequence set, the model can switch from the seed motif to these other motifs. In this way, the weaker motif is not discovered at all. To avoid these two problems, we introduced a masking step in the EM optimization. We score all possible motif start positions in the input sequences using the PWM passed from the seeding stage to the refinement stage. We mask out all but the top X% of positions ($X = 5$ in this study) and ignore these positions in the EM iterations of the refinement stage.

Motif assessment using average recall (AvRec)

To assess the performance of a classifier such as a motif model, one often plots the true positive predictions (TP) versus the false positive predictions (FP) over all score thresholds. Normalizing FPs and TPs to a maximum of 1 by plotting the true positive rate TPR = TP/Positives versus the false positive rate FPR = FP/Negatives yields the receiver operating curve (ROC). The often-used area under the ROC curve (AUC) is not a good quality measure for a motif model because in many applications the fraction of positive sequences (those carrying the motif) is much smaller than the number of negative sequences. When scanning the human genome for CTCF binding motifs in windows of 100 bp, for example, the ratio is about 1:30. At this ratio, a false discovery rate FDR = TP/(TP + FP) below 50% requires a ratio FPR/TPR < 1/30. So 29/30 = 97% of the ROC plot, the part with FDR > 50%, would be irrelevant. A predictor could have an AUC of 95% and never reach an FDR below 50%.

We therefore previously developed the Average Recall (AvRec) score (34), which averages the recall (the same as true positive rate and sensitivity) over a range of TP:FP ratios from 1:1, corresponding to FDR = 0.5, to 1:100, corresponding to FDR = 1/101 (Figure 2A). The AvRec score therefore considers the range of FDR most relevant in practice and has the additional benefit that a different positive-to-negative ratio than 1 simply results in a vertical shift of the AvRec curve on the logarithmic y axis.

To calculate the AvRec score, we first simulated 10-fold more negative than positive sequences using a second-order Markov background model learned on the positive set. We computed the motif scores $S(x_{i:i+W-1}) = \log_2\left(p^{K}_{\text{motif}}(x_{i:i+W-1})/p^{K'}_{\text{bg}}(x_{i:i+W-1})\right)$ for all possible binding positions i (excluding the masked positions) and took the best score for each sequence. All sequences are sorted by descending score. The false positive count FP is the cumulative number of sequences from the negative set above the score cut-off, and TP is the cumulative number of positive sequences above the score cut-off.

Benchmark design

The performance of BaMMmotif2 was evaluated together with five state-of-the-art motif discovery tools, MEME (32) as the most cited tool, CisFinder (35) for its speed and ability to run on large datasets, ChIPMunk (36) and diChIPMunk (27), which are used for generating the PWMs and dinucleotide PWMs in the HOCOMOCO database (37), and InMoDe (24), which can learn inhomogeneous Markov models of order 2 and beyond.

The processing of the ChIP-seq and HT-SELEX data is described in detail in Supplemental Material II. The motif discovery tools were run on the input sequence sets with default parameters, and four CPU cores were used for tools that could be parallelized (CisFinder, MEME and BaMMmotif2). For tools that learn multiple motif models from one dataset, the motif models ranked top by the tools were benchmarked.

To assess the model performance over the given sequences, we first performed the benchmark on both human ChIP-seq (38) and HT-SELEX datasets (39) using 5-fold cross-validation (Figure 1I). The cross-cell-line validation was applied to ChIP-seq data (Figure 1II) and the cross-platform validations were applied to both ChIP-seq data HT-SELEX data (Figure 1III and IV). A more detailed description including tool settings can be found in Supplementary Section SII.

RESULTS

Model performance on *in vivo* and *in vitro* data

We learned *de novo* motifs with each of the six tools on 427 ChIP-seq datasets for 93 transcription factors from the ENCODE project (38) and evaluated their performance using 5-fold cross-validation (Figure 1I).

As an example, we compare in Figure 2A and B the AvRec plot of a fifth-order BaMM with a second-order InMoDe model for the Elf2 motif, trained and tested on 5000 sequences of length 208 bp via 5-fold cross-validation. At a positives-to-negatives ratio of 1:1 (bold blue line) and a TP:FP-ratio of 10:1 (see y axis, corresponding to an FDR of 1/11), the BaMMmotif2 model achieves a recall of 0.81 and the InMoDe model achieves 0.69. At a positives-to-negatives ratio of 1:10 and a TP:FP-ratio of 10:1 (broken blue line), or, equivalently, at a positives-to-negatives ratio of 1:1 and a TP:FP-ratio of 100:1, the models achieve recalls of 0.12 and 0.13, respectively. When comparing AvRec scores between fifth-order BaMMs with second-order models from InMoDe across all 427 ChIP-seq datasets, BaMMs attain higher AvRec scores for 415 (97%) of the datasets, and the median AvRec of BaMMs is 13.6% higher than the one of InMoDe models (Figure 2C). This improvement is universal across TF domain families (40) (Figure 2C).

Overall, the PWM-based tools, CisFinder, MEME and ChIPMunk, are outperformed by the tools using higher-order models. BaMMmotif2 with first-order models performs on par with InMoDe and better than the first-order tools such as diChIPMunk. Fifth-order BaMMs achieve even better AvRec scores, as seen in the box plots and AvRec cumulative distributions of Figures 2D and E, and in one-on-one comparisons in Supplementary Figure S2A. We also compared BaMMmotif2 with our previous tool BaMMmotif (28). BaMMmotif2 is 10 times faster while being slightly more sensitive (Supplementary Figure S3).

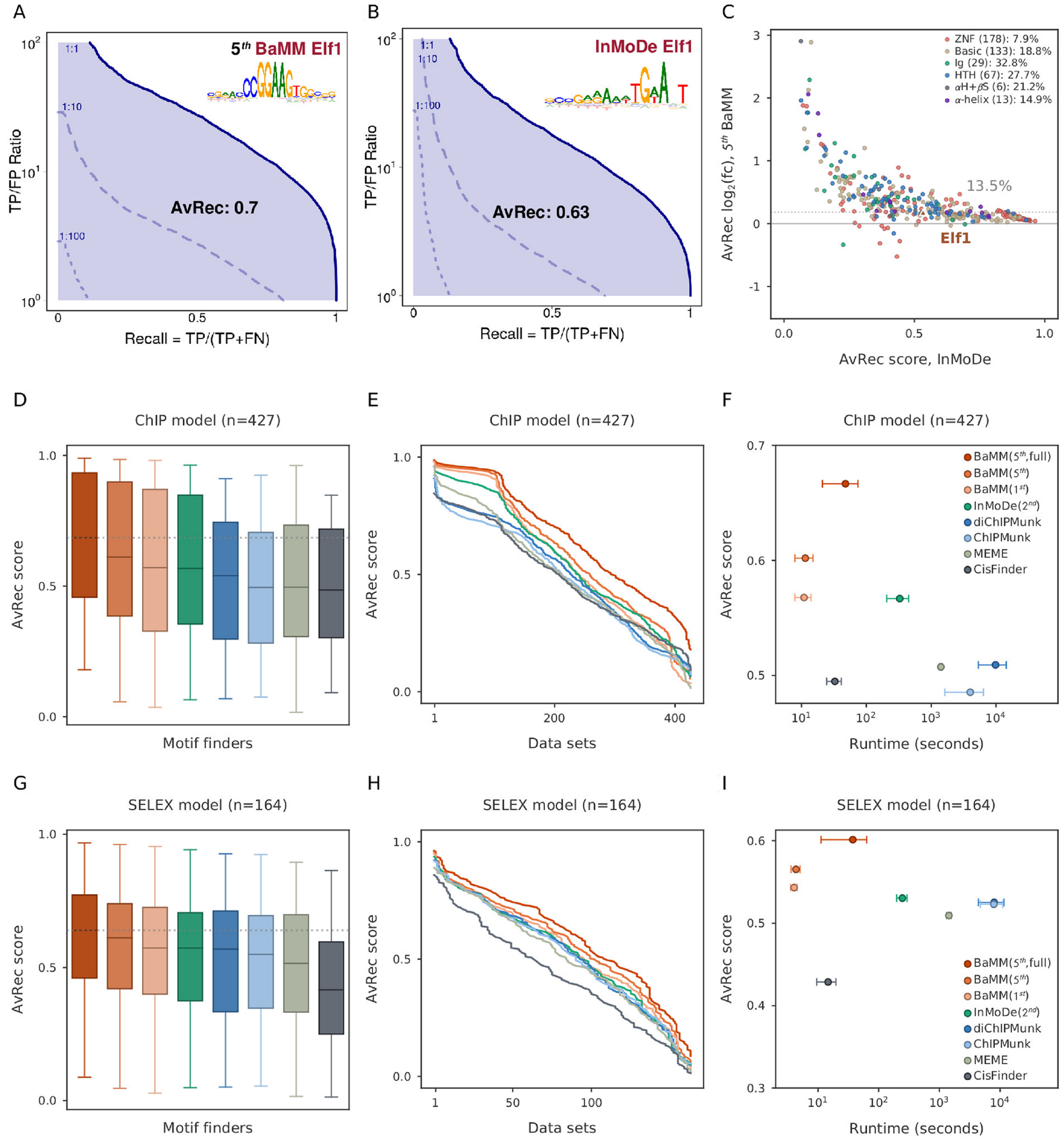

Figure 2. Performance of *de novo* motif discovery tools on *in vivo* and *in vitro* datasets. (**A**) AvRec analysis for fifth-order BaMM on the Elf1 ENCODE dataset. The AvRec is the recall averaged in log space over TP-to-FP ratios between 10^0 and 10^2. This ratio range corresponds to a precision between $1/(1 + 1)$ and $100/(1 + 100) = 0.99$. Bold line: 1:1 ratio of positives to negatives. At 1:10 ratio (dashed) and 1:100 (dotted), the curves are shifted down by a factor of 10 and 100, respectively. Inset: motif logo of Elf1. (**B**) Same as (A) for the InMoDe model of Elf1. (**C**) $\log_2$ of AvRec fold change between fifth-order BaMMmotif2 and InMoDe models versus the AvRec of InMoDe. Each dot represents one dataset. Elf1 is highlighted in a brown triangle. Dot colors represent different TF superfamilies defined by (40). ZNF: Zinc-finger DNA-binding domains, Basic: Basic domains, Ig: Immunoglobulin fold, HTH: Helix-turn-helix domains, αH+βS: alpha-helices exposed by beta-structures, αH: Other all-alpha-helical DNA-binding domains. The median AvRec fold change and the number of motifs are shown in the legend. The overall median $\log_2$ fold change is 13.5%. (**D**) AvRec distributions as box plot, with boxes indicating 25%/75% quantiles and whiskers 95%/5% quantiles. Color code: see the legend in (**F**). (**E**) Cumulative distribution of AvRec scores on the 427 datasets. (F) Average runtime per dataset on four cores versus the median AvRec score. InMoDe and (di)ChIPMunk are not parallelized and ran on a single core. Whiskers: ±1 standard deviation. BaMM (5th, full): no masking step. (**G–I**) Analogous to (D–F) but for 164 HT-SELEX datasets from the Taipale lab (39).

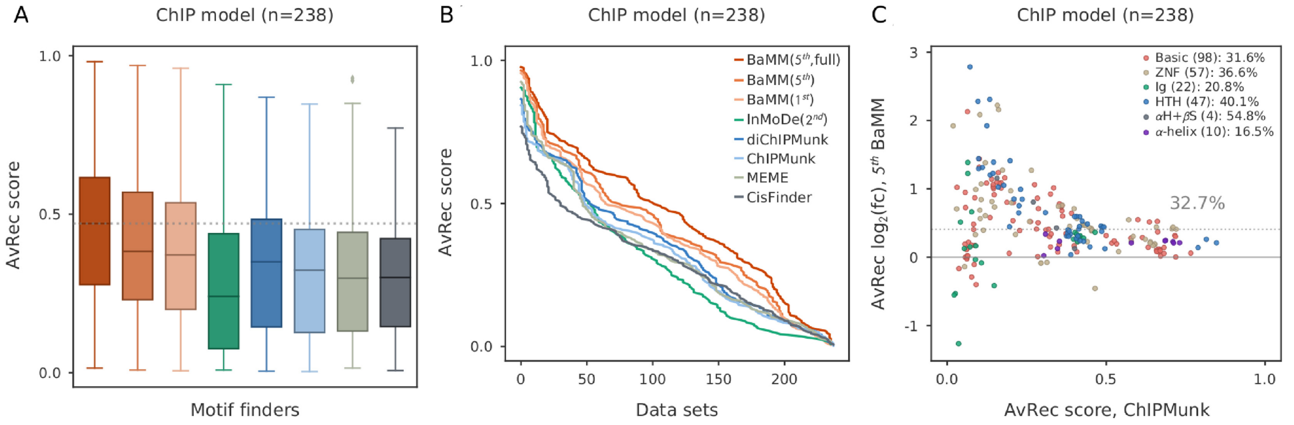

Figure 3. Cross-cell-line validation. 119 pairs of ENCODE datasets were used in this benchmark in which the same TF had been ChIPped in different cell lines. (**A**) AvRec distributions for 2 × 119 models that were tested on a ChIP-seq dataset from a different cell line than they were trained on. (**B**) Cumulative distributions of AvRec scores. (**C**) Log_2 fold change in AvRec between fifth-order BaMMs and ChIPMunk for each of the 238 datasets. The median improvement is 32.7%. Same legend as Figure 2C.

Tools that learn higher-order Markov models can learn several motifs in one model, profiting from signals that are merely correlated with the real binding sites (29,41). To find out whether BaMMs are affected or not, we introduced a masking step in the initial iteration of the EM algorithm (see ‘Materials and Methods’ section). We restrict the model refinement with the higher-order BaMM to the 5% potential motif positions with the highest scores scanned by the seeding PWM. In this way, we avoid overfitting and also speed up the refinement by a factor of 10. However, this robustness is paid by a loss in motif model performance (Supplementary Figures S4 and S5). The performance decrease could be caused in part by the limitation of being unable to select better sites during the refinement that were too different from the seeding motif, and in part because sometimes the BaMMs would otherwise have learned more than one distinct motif in a single model. To be on the conservative and robust side, we adopted the masking step in BaMMmotif2 for all our benchmarks in this study, unless explicitly stated otherwise.

Next, we assessed the performances of selected tools on 164 *in vitro* HT-SELEX datasets for 164 TFs (39). Each dataset contains long oligomers of 200 bp. We also sampled 10-fold background sequences using the trimer frequencies from the same input set for estimating true negatives.

For the *in vitro* benchmark we observed overall similar trends as on the ChIP-seq data (Figure 2G–I). CisFinder tends to learn longer motifs than the other tools, which probably helped it on the ChIP-seq data but hurt its performance on the HT-SELEX data. The BaMMs learned without masking (BaMM 5^{th}, full; red) gained only 5% on the masked version (BaMM 5^{th}; orange), whereas the gain had been 12% on the ChIP-seq data. This comparison shows that, on the ChIP-seq data, the fifth-order BaMMs trained without masking indeed tend to learn also motifs of cooccuring TFs that help to distinguish positive from negative sequences. If the goal is to learn the pure binding affinity of the ChIPped TF, masking should therefore be turned on for *in vivo* data.

Assessing consistency of motif models across cell lines

ChIP-seq measurements have cell-type-specific biases associated with difference in chromatin accessibility, in particular of enhancers and promoters, and differences in TF concentrations (42). A motif model that predicts only the binding affinity of the ChIPped TF should also perform well in predicting binding sites of the factor in other cell lines, whereas a motif model that has learned also motifs of co-occurring TFs and other sequence features with no direct effect on the binding affinity of the main TF should generalize badly to other cell lines in which different TFs will often co-occur with the ChIPped TF.

We therefore conducted a cross-cell line benchmark on *in vivo* data. We assessed the performance of models learned on ChIP-seq data from one cell line and tested on ChIP-seq data of the same TF from another cell line. We found 119 pairs of ChIP-seq datasets in the ENCODE database in which the same TF had been ChIPped in two different cell lines. We trained the model on one dataset and tested it on the other, and vice versa, resulting in 238 AvRec scores (Figure 3).

Remarkably, the AvRec scores are around 0.2 lower for all tools than the AvRec scores in Figure 2D obtained when training and testing in the same cell lines, with the PWM-based tools going from AvRec 0.5 to as low as 0.3. This quite dramatic decrease indicates that all models, even the simple PWMs, do not perform well for predicting bound sequences in another cell line. Remarkably, except for InMoDe, the predictive power of the higher-order models does not suffer more than that of the PWMs. This indicates that the higher-order models (except InMoDe) do not tend to overfit to sequence features that are specific to one cell line, such as co-occurring TFs. It is surprising that the fifth-order BaMMs trained without masking maintain or even improve their edge on the other models, despite our expectation that they would be the most prone to overfit on cell type-specific features.

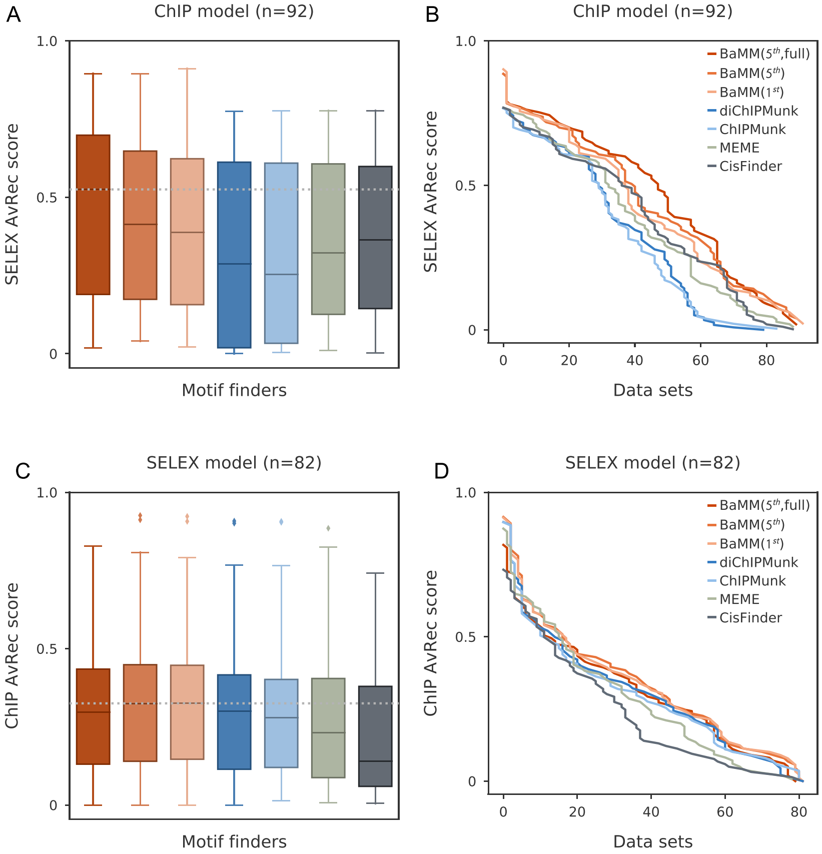

Figure 4. Cross-platform validation. (**A** and **B**) AvRec distributions and cumulative distributions for 92 models trained on ChIP-seq datasets and tested on HT-SELEX datasets for the same TFs using different tools. (**C** and **D**) Same as (A,B) for 82 motif models but trained on HT-SELEX datasets and tested on ChIP-seq datasets.

In vitro models predict *in vivo* binding and vice versa

Each measurement for detecting TF–DNA interactions has its own biases. ChIP-seq has biases from sequence-dependent PCR amplification, cell-type-specific sonication bias, and chromatin structure(43–45), while HT-SELEX has biased nucleotide compositions and depleted palindromes as a result of the library preparation, as well as sequence carry-over bias in selection cycles (46,47). These biases can give optimistic results even in the cross-cell-line benchmark because the model can be overtrained on genomic features that are identical or similar in both cell lines.

To assess how much models base their predictions on technical biases that would improve their performance when tested on the same platform but decrease their performance when tested on a different platform, we performed two cross-platform benchmarks.

First, for each of the tools, we trained a motif model on each of the 140 ChIP-seq datasets for which an HT-SELEX dataset for the same TF, but not necessarily from the same cell line, was available. We discovered that several datasets were of too low quality to give reliable models, and some HT-SELEX datasets showed signs of having had the identity of the TF switched. We therefore selected the 92 ChIP-seq datasets for which at least one of 8 tools achieved an AvRec score of $\geq$0.1. The first- and fifth-order BaMMs achieve better accuracies than the PWM-based models (Figure 4A and B).

Second, for each of the tools we trained a motif model on each of the 82 HT-SELEX datasets for which a ChIP-seq dataset with the same TF was available. We selected the HT-SELEX datasets for which at least one of the 8 tools achieved an AvRec score of $\geq$ 0.1. Again, BaMMs achieved the best AvRec scores. However, we observed no major improvements from first to fifth order (Figure 4C and D). This time, the improvements over PWM-based models are minor. ChIPMunk and diChIPMunk fared badly because they only predict one motif per dataset, while other tools generate several motif candidates and the best one is chosen for comparison.

BaMMs learned similar information content in the first-order on ChIP-seq and on HT-SELEX data while showing no tendency to learn systematic biases of these platforms (Supplementary Figure S7). This demonstrates how the information in the first-order can help to improve cross-platform predictions.

Extended flanking regions increase motif prediction accuracy

Various studies have shown that the flanking regions outside of the core binding sites affect TF binding, by affecting

DNA shape preferences or by harboring binding sites of co-cooperatively binding TFs at variable spacings (14,48–50). Therefore, we investigated the impact of extending the core motifs, by adding two or four nucleotides on each side in the seeding motifs and refining the extended motifs with BaMMmotif2.

We find that for BaMMs trained on ChIP-seq datasets, extending the models by 2 × 2 or 2 × 4 positions indeed improves the motif model performance across all orders, and more so with increasing order (Figure 5A). The improvement from no added positions to 2 × 4 bp added is by 3% for zeroth order BaMMs (PWMs) and by 11% for fifth-order BaMMs (Figure 5B and Supplementary Figure S8A). This indicates that flanking regions carry information mostly in the higher orders and not much in preferences for specific nucleotides.

It is not clear, however, if these improvements are due to DNA shape preferences that are reflected by preferences for certain di- and tri-nucleotides or by other sequence features of the genomic sequences such as motifs of co-occurring TFs. We therefore repeated the same analysis on HT-SELEX data. We restricted ourselves to long oligonucleotides of 200 bp because short oligonucleotides of 20 to 40 bp might not reflect well enough the physical properties of genomic DNA.

The results on the HT-SELEX data are very similar to those on ChIP-seq data (Figure 5B and D). Again, PWMs gain much less AvRec score through 2 × 4 bp extensions than fifth-order BaMMs (1.3% versus 8%, shown in Supplementary Figure S8B and Figure 5D). This result confirms that the features picked up by the higher orders are not chiefly ones that are specific to genomic sequences but are also learned on *in vitro*-selected sequences and are therefore likely to be associated with DNA structural preferences.

Learning positional binding preferences

Motifs often have certain positional preferences with regard to other motifs or genomic landmarks such as transcription start sites. Therefore, we introduced the possibility to learn the probability distribution of motif positions from the input data (Supplementary Figure S1A). Learning the positional distribution of motifs around ChIP-seq peak positions did not improve the median motif performances (Supplementary Figure S1B and S1C), probably because the information content of the positional distribution is very low when the the distribution is not much narrower than the window size (the information content can be calculated as the difference between the entropies of the two positional distributions). The positional preference is likely to have a positive impact when positioning effects are stronger, such as for splicing motifs around splice sites, core promoter motifs around transcription start sites, or TF binding sites of cooperatively binding TFs.

DISCUSSION

We presented BaMMmotif2, a fast and accurate *de novo* motif discovery algorithm for large-scale transcriptomic data. BaMMmotif2 builds on our earlier theory of Bayesian Markov models (BaMMs) implemented in BaMMmotif. BaMMs employ pseudocounts from model order $k-1$ to stabilize the estimation of the conditional probabilities for order k, for all orders k from 1 to the maximum order (five in this study). In this way, they can learn higher orders if a sufficient number of k-mer counts was observed to estimate them but otherwise fall back to a lower order that can still be estimated safely.

BaMMmotif2 was written from scratch in C++ using explicit AVX2 vectorization and multi-core parallelization. We developed a novel, fast seeding method to find enriched patterns that scales almost independently of the input set size. We also added a masking step to force the refinement stage to only refine the seed motifs and prevent it from learning in addition other predictive features such as co-occurring motifs of other TFs or experimental sequence biases. We also developed a Bayesian approach to learn position binding preferences from the input data.

By their sheer number, ChIP-seq datasets are the dominant source of information for TF binding affinities. Therefore, most benchmark comparisons of *de novo* motif discovery tools have been performed exclusively or predominantly on ChIP-seq data. However, for assessing the quality of models more complex and informative than PWMs, such as higher-order Markov models and mixture models, ChIP-seq data are problematic for several reasons. First, they often have complex sequence biases (42), which higher-order models can learn to distinguish from negative sequences generated with random background models. To alleviate this problem, second order background models should be used, but even this might be insufficient to eliminate learning generic sequence biases of the ChIPped versus random sequences. Second, sequences in ChIP-seq peaks usually contain in addition to the motif of the ChIPped TF the binding motifs of co-binding factors (41). Complex models can improve their predictive performance by scoring sequences highly that contain any of these co-occuring motifs. This is possible even within a short motif length by learning the motifs superposed with each other, with the higher orders preventing mixing and blurring of motifs (29). Although improving the apparent model performance, such models do not describe faithfully the binding affinity of the ChIPped factor.

Our goal was to compare PWM-based motif discovery tools with tools employing more complex models: dinucleotide weight matrices, parsimonious context trees and BaMMs. We therefore set up a cross-cell line benchmark to assess how well the motif models learned in one cell line can predict binding in another cell line. Furthermore, we conducted a cross-platform benchmark, in which we trained the models on ChIP-seq data and tested them on HT-SELEX data, and vice versa. The results show that among the tested tools, those with more complex models still tend to perform better in these benchmarks, albeit with smaller improvements over the PWM-based tools. The improvements from higher orders were particularly marked for the BaMMs. So, most of the information in higher orders seems to be transferable between cell lines and measurement platforms.

Even though we did not see clear signs of overfitting in our BaMMs, we introduced sequence masking as a precaution against overfitting to other motifs and technology- or cell line-dependent sequence biases. We use the seed PWM

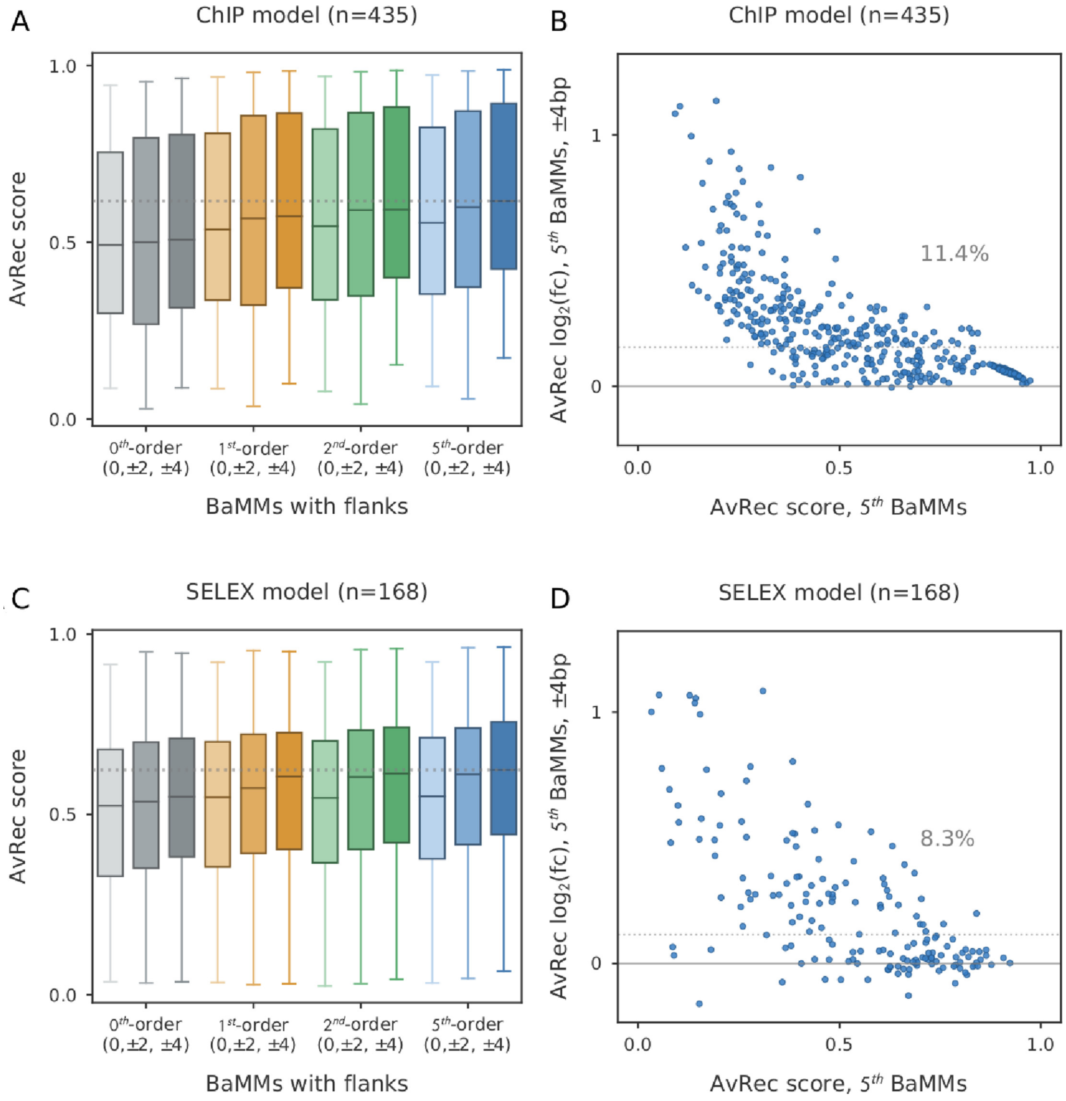

Figure 5. Extending the core motif by flanking positions improves motif performance. AvRec of BaMMs with different numbers of flanking positions added to the core motif, tested by 5-fold cross-validation. (**A**) AvRec distribution on 435 ChIP-seq datasets for models of order 0, 1, 2 and 5, each for three sizes of flanking regions: 0 bp, ± 2 bp and ± 4 bp. (**B**) Log_2 of fold change between fifth-order BaMMmotif2 models with ± 4 bp flanking positions and no added flanking positions. The median AvRec increase is 11.4 %. (**C** and **D**) Same as (A and B) for 168 HT-SELEX datasets.

to mask out all but the top-scoring 5% of positions, and we train the higher-order BaMM only on the remaining 5%. We thereby ensure that only sequence regions that actually carry the seed motif can be learned by the BaMM. The performance drop between training fifth-order BaMMs with and without masking was 8% on HT-SELEX data and 12% on ChIP-seq data (Figure 2D,G; Supplementary Figure S4). This indicates that if higher-order BaMMs profit from learning co-occurring motifs at all, the effect on their performance is quite limited.

Still, if the goal is to learn binding affinities and not just predict motifs from *in vivo* sequence data, we recommend to run BaMMmotif2 with the masking because BaMMs can learn several similar motifs in one single model, such as bipartite motifs with a variable-length spacer or motifs of mono- and dimeric binding modes of a transcription factor. The masking option controls how closely the refined motif has to stay to the seed motif. For instance, masking helps to learn the correct partially related motifs for FoxA2 factor, when training 5th-order BaMMs on a ChIP-seq data (Supplementary Figure S11C). Whether these similar motifs are learned in a single model or are split into two models can vary from case to case. If users want to learn motifs separately, it is therefore recommended to use masking and to experiment with even stricter masking than the default 95%.

On *in vitro* data, masking is not necessary and in order to make use of the 5% improvement we recommend to run BaMMmotif2 without masking. However, even with masking the fifth-order BaMMs still perform competitively with the state-of-the-art tools while being significantly faster.

Transcription factors combine base- with DNA shape readout (13). Instead of studying the TF-DNA binding using only the sequence features, some models utilize DNA shape features predicted from the sequence to enhance motif models (51–53). The shape descriptors these tools use, like minor groove width, helical tilt and bent, or propeller tilt, are predicted from five-mer tables computed using molecular dynamics calculations. Given enough data, it is therefore evident that higher-order models such as BaMMs can learn these DNA structural preferences implicitly, yet are not limited to the pre-defined shape descriptors.

In recent years, deep learning approaches have become popular for learning motif models with very good predictive performance (51,54,55). Such models usually take advantage of contextual information such as co-occuring motifs, which increases their predictive power but serves a different purpose than the models we discuss here: learning a model for the sequence dependence of the binding affinity of a factor. In addition, BaMMs have the advantage or being conceptionally simple and interpretable in terms of *k*-mer dependent energy terms.

In conclusion, we have shown that higher-order models for binding motifs improved binding site predictions on a large collection of ChIP-seq and HT-SELEX datasets, both in cross-validated setting and when training and testing on different experimental platforms and cell lines. Importantly, clear improvements in predictive performance are even seen beyond first order models: BaMMs of fifth order show a solidly improved performance across the bench over the tested state of the art tools, while being significantly faster.

AVAILABILITY

Data

ENCODE database. We evaluated the performance of selected algorithms on human ChIP-seq datasets from the ENCODE portal (38) until March 2020. In total, there are 435 datasets for 93 distinct transcription factors. The top 5000 peak regions sorted by their signal value are selected for each dataset when peaks are >5000, and all peaks are chosen if there are fewer than 5000 peaks. Positive sequences are extracted ±104 bp around the peak summits. Background sequences are sampled by the trimer frequencies from positive sequences, with the same lengths as positive sequences and 10 times the amount of positive sequences. 8 datasets are excluded from all the results because diChIPMunk fails to learn models within 3 h.

HT-SELEX datasets. For HT-SELEX data, we downloaded 164 datasets with 200 bp-long oligomers from Zhu *et al*. (39), which are deposited in the European Nucleotide Archive (ENA) under the accession PRJEB22684. Each dataset represents one non-redundant transcription factor. For each dataset, we selected 5000 sequences from each selection round without any sorting.

The HT-SELEX data contain reads from at least four selection cycles, and the measured binding affinity iteratively increases with the cycles. Thus, we chose the sequences from the fourth selection rounds with detected high affinities for motif training and testing in the main paper. Since ChIPMunk and diChIPMunk took longer than 2 h to run on the full datasets, we selected 5000 sequences out of the millions of reads as training and test sequences. To examine the power of BaMMs in learning the weak binding sites, we also used sequences from the second and third selection rounds. Background sequences are sampled in the same way as described previously.

Software and parameters

The new version of BaMMmotif2 software is implemented in C++ and Python3. The code is licensed under GPLv3 and freely accessible without registration at github.com/soedinglab/PEnG-motif, and github.com/soedinglab/BaMMmotif2, and supported on Linux and MacOS. They are also integrated into our webserver (34).

Results and analysis scripts

The analysis scripts are available in Jupyter Notebook format at github.com/soedinglab/bamm-benchmark.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Matthias Siebert and Anja Kiesel for help with designing the code structure of BaMMmotif2, the members of the Söding lab, especially Eli Levy Karin, Saikat Banerjee, Milot Mirdita and Ruoshi Zhang for discussions, and the genomics research community for sharing their data.
Author contributions: W.G. developed the BaMMmotif2 software. M.M., C.R. and J.S. developed the seeding stage software (PEnG). W.G. implemented the statistical approach and conducted all the benchmarks. W.G. and J.S. wrote the manuscript. J.S. supervised the research.

FUNDING

DFG [SPP1935 CR 117/6–1]; International Max Planck Research School for Genome Science [IMPRS-GS]. Funding for open access charge: Institutional.
Conflict of interest statement. None declared.

REFERENCES

1. Serfling,E., Jasin,M. and Schaffner,W. (1985) Enhancers and eukaryotic gene transcription. *Trends Genet.*, **1**, 224–230.
2. Argos,P. (1988) A sequence motif in many polymerases. *Nucleic Acids Res.*, **16**, 9909–9916.
3. Mitchell,P.J. and Tjian,R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **245**, 371–378.
4. Jolma,A. and Taipale,J. (2011) Methods for analysis of transcription factor DNA-binding specificity in vitro. In: *A Handbook of Transcription Factors*. Springer, pp. 155–173.
5. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
6. Meng,X., Brodsky,M.H. and Wolfe,S.A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.*, **23**, 988–994.
7. 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.

8. Riley,T.R., Slattery,M., Abe,N., Rastogi,C., Liu,D., Mann,R.S. and Bussemaker,H.J. (2014) SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. In: *Hox Genes*. Springer, pp. 255–278.
9. Isakova,A., Groux,R., Imbeault,M., Rainer,P., Alpern,D., Dainese,R., Ambrosini,G., Trono,D., Bucher,P. and Deplancke,B. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods*, **14**, 316–322.
10. 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.
11. Bulyk,M.L., Johnson,P.L. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
12. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
13. 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.
14. 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.
15. 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. USA*, **109**, E3084–E3093.
16. Zuo,Z. and Stormo,G.D. (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics*, **198**, 1329–1343.
17. Halazonetis,T.D., Georgopoulos,K., Greenberg,M.E. and Leder,P. (1988) c-Jun dimerizes with itself and with c-Fos, forming complexes of different DNA binding affinities. *Cell*, **55**, 917–924.
18. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
19. Crocker,J., Noon,E. P.-B. and Stern,D.L. (2016) The soft touch: low-affinity transcription factor binding sites in development and evolution. In *Curr. Top. Dev. Biol.* Vol. **117**, Elsevier, pp. 455–469.
20. Kribelbauer,J.F., Rastogi,C., Bussemaker,H.J. and Mann,R.S. (2019) Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annu. Rev. Cell Dev. Biol.*, **35**, 357–379.
21. Jiang,J. and Levine,M. (1993) Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, **72**, 741–752.
22. Rastogi,C., Rube,H.T., Kribelbauer,J.F., Crocker,J., Loker,R.E., Martini,G.D., Laptenko,O., Freed-Pastor,W.A., Prives,C., Stern,D.L. *et al.* (2018) Accurate and sensitive quantification of protein-DNA binding affinity. *Proc. Natl. Acad. Sci. USA*, **115**, E3692–E3701.
23. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
24. Eggeling,R., Grosse,I. and Grau,J. (2017) InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics*, **33**, 580–582.
25. Gershenzon,N.I., Stormo,G.D. and Ioshikhes,I.P. (2005) Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res.*, **33**, 2290–2301.
26. Siddharthan,R. (2010) Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS One*, **5**, e9722.
27. 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.
28. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
29. Eggeling,R. (2018) Disentangling transcription factor binding site complexity. *Nucleic Acids Res.*, **46**, e121.
30. Orenstein,Y. and Shamir,R. (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.*, **42**, e63.
31. Nitta,K.R., Jolma,A., Yin,Y., Morgunova,E., Kivioja,T., Akhtar,J., Hens,K., Toivonen,J., Deplancke,B., Furlong,E.E. *et al.* (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *elife*, **4**, e04837.
32. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
33. Sohrabi-Jahromi,S. and Söding,J. (2021) Thermodynamic modeling reveals widespread multivalent binding by RNA-binding proteins. bioRxiv doi: https://doi.org/10.1101/2021.01.30.428941, 01 February 2021, preprint: not peer reviewed.
34. Kiesel,A., Roth,C., Ge,W., Wess,M., Meier,M. and Söding,J. (2018) The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.*, **46**, W215–W220.
35. Sharov,A.A. and Ko,M.S. (2009) Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Res.*, **16**, 261–273.
36. Kulakovskiy,I.V., Boeva,V., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
37. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
38. ENCODE Project Consortium and others (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
39. Zhu,F., Farnung,L., Kaasinen,E., Sahu,B., Yin,Y., Wei,B., Dodonova,S.O., Nitta,K.R., Morgunova,E., Taipale,M. *et al.* (2018) The interaction landscape between transcription factors and the nucleosome. *Nature*, **562**, 76–81.
40. Wingender,E., Schoeps,T., Haubrock,M. and Dönitz,J. (2015) TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.*, **43**, D97–D102.
41. Hunt,R.W. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.*, **15**, 412.
42. Chen,Y., Negre,N., Li,Q., Mieczkowska,J.O., Slattery,M., Liu,T., Zhang,Y., Kim,T.-K., He,H.H., Zieba,J. *et al.* (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods*, **9**, 609–614.
43. Aird,D., Ross,M.G., Chen,W.-S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.
44. Diaz,A., Park,K., Lim,D.A. and Song,J.S. (2012) Normalization, bias correction, and peak calling for ChIP-seq. *Stat. Appl. Genet. Mol. Biol.*, **11**, doi:10.1515/1544-6115.1750.
45. Teytelman,L., Özaydın,B., Zill,O., Lefrançois,P., Snyder,M., Rine,J. and Eisen,M.B. (2009) Impact of chromatin structures on DNA processing for genomic analyses. *PloS One*, **4**, e6700.
46. Zhao,Y., Granas,D. and Stormo,G.D. (2009) Inferring binding energies from selected binding sites. *PLoS Comput. Biol.*, **5**, e1000590.
47. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
48. Levo,M., Zalckvar,E., Sharon,E., Machado,A. C.D., Kalma,Y., Lotam-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.
49. Schöne,S., Jurk,M., Helabad,M.B., Dror,I., Lebars,I., Kieffer,B., Imhof,P., Rohs,R., Vingron,M., Thomas-Chollier,M. *et al.* (2016) Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nat. Commun.*, **7**, 12621.

50. Yella,V.R., Bhimsaria,D., Ghoshdastidar,D., Rodríguez-Martínez,J.A., Ansari,A.Z. and Bansal,M. (2018) Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res.*, **46**, 11883–11897.
51. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.*, **3**, 278–286.
52. Peng,P.-C. and Sinha,S. (2016) Quantitative modeling of gene expression using DNA shape features of binding sites. *Nucleic Acids Res.*, **44**, e120.
53. Samee,M. A.H., Bruneau,B.G. and Pollard,K.S. (2019) A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.*, **8**, 27–42.
54. 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.
55. Kelley,D.R., Snoek,J. and Rinn,J.L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.