BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 4 2012, pages 487–494
doi:10.1093/bioinformatics/btr695

Sequence analysis

MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences

Phil Arnold[1], Ionas Erb[2], Mikhail Pachkov[1], Nacho Molina[3] and Erik van Nimwegen[1,*]

[1]Biozentrum, University of Basel, Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland, [2]Bioinformatics and Genomics program, Center for Genomic Regulation (CRG) and Pompeu Fabra University (UPF), Barcelona, Spain and [3]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Probabilistic approaches for inferring transcription factor binding sites (TFBSs) and regulatory motifs from DNA sequences have been developed for over two decades. Previous work has shown that prediction accuracy can be significantly improved by incorporating features such as the competition of multiple transcription factors (TFs) for binding to nearby sites, the tendency of TFBSs for co-regulated TFs to cluster and form *cis*-regulatory modules and explicit evolutionary modeling of conservation of TFBSs across orthologous sequences. However, currently available tools only incorporate some of these features, and significant methodological hurdles hampered their synthesis into a single consistent probabilistic framework.

Results: We present MotEvo, a integrated suite of Bayesian probabilistic methods for the prediction of TFBSs and inference of regulatory motifs from multiple alignments of phylogenetically related DNA sequences, which incorporates all features just mentioned. In addition, MotEvo incorporates a novel model for detecting *unknown functional elements* that are under evolutionary constraint, and a new robust model for treating gain and loss of TFBSs along a phylogeny. Rigorous benchmarking tests on ChIP-seq datasets show that MotEvo's novel features significantly improve the accuracy of TFBS prediction, motif inference and enhancer prediction.

Availability: Source code, a user manual and files with several example applications are available at www.swissregulon.unibas.ch.

Contact: erik.vannimwegen@unibas.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 7, 2011; revised on November 23, 2011; accepted on December 15, 2011

1 INTRODUCTION

What the sequence specificities of different transcription factors (TFs) are and where in the genome their transcription factor binding sites (TFBSs) occur remain central questions in gene regulation. For over two decades, a large number of computational methods has been developed that aim to support answering such questions, see e.g. Bulyk (2003); Hannenhalli (2008) for reviews. Although much progress has been made, it remains highly challenging to obtain accurate computational TFBS predictions, especially on a genome-wide scale. For example, although from a biophysical point of view identical sequence segments should have equal affinity for the TF, one typically finds that only a small fraction of the sequences with high binding affinity act as functional TFBSs. That is, TFBS functionality is context dependent and thus researchers have searched for additional features that are predictive for the functionality of putative sites.

One approach that has proven particularly fruitful is comparative genomic analysis of the conservation of putative TFBSs across related species, i.e. putative TFBSs that are highly conserved are generally more likely to be functional. A large number of approaches for incorporating conservation information has been proposed including several simple *ad hoc* methods, e.g. Kellis *et al.* (2003), but it has become clear that highest performance is obtained by methods that use explicit evolutionary models for the evolution of TFBSs along a phylogeny (Hawkins *et al.*, 2009; Moses *et al.*, 2004; Siddharthan *et al.*, 2005). As a consequence, there has been considerable interest in extending methods for regulatory motif finding and TFBS prediction to include such explicit phylogenetic models. For example, the well-known Gibbs sampling (Lawrence *et al.*, 1993) and expectation–maximization strategies (Bailey and Elkan, 1994) for *ab initio* motif finding have been extended to methods that work on multiple alignments of orthologous sequences and use explicit evolutionary models (Siddharthan *et al.*, 2005; Sinha *et al.*, 2004).

Beyond conservation information, other features have also proven highly useful in improving the accuracy of TFBS prediction. For example, especially in higher eukaryotes, functional TFBSs often come in clusters where multiple binding sites for a small subset of TFs co-occur in close proximity to each other (Davidson, 2001). Several methods were developed that, instead of looking for TFBSs for one TF at a time, explicitly look for clusters of sites for a collection of TFs. These methods have been especially successful in identifying *cis*-regulatory modules that are distal to their target gene (Frith *et al.*, 2001; Rajewsky *et al.*, 2002). It has also proven useful to take into account the contribution of many weak binding sites and the competitive binding of multiple TFs to a DNA sequence by explicitly considering all possible configurations of non-overlapping binding sites using a dynamic programming procedure, e.g. Rajewsky *et al.* (2002); Roider *et al.* (2007); Wasson and Hartemink (2009).

*To whom correspondence should be addressed.

However, currently these methodologies are spread over multiple computational tools each of which only implements some of these methods. For example, the methods for finding *cis*-regulatory modules have been extended to analyze pairs of aligned species (Sinha *et al.*, 2003) but not to general multiple alignments and phylogenetic relationships. Other methods can only make predictions for one TF at a time, ignoring the competitive binding of multiple TFs [e.g. Moses *et al.* (2004); Sinha *et al.* (2004)] and the methods that incorporate sophisticated models for explicitly considering all possible binding configurations of multiple TFs cannot incorporate conservation information [e.g. Wasson and Hartemink (2009)]. Beyond this, as we show below, current methods that incorporate explicit evolutionary models make several implicit assumptions that cause 'pathologies' that significantly affect their performance.

Here we present a computational tool, MotEvo, that integrates a suite of Bayesian probabilistic methods for the prediction of regulatory sites and motifs on multiple alignments of phylogenetically related sequences. MotEvo not only implements and extends the functionality of many of the tools and methods mentioned in the introduction into a single integrated method, it also incorporates a number of new features that address the 'pathologies' that current methods suffer from, as we explicitly demonstrate below.

2 METHODS

MotEvo takes as input either sequences from a single species or multiple alignments of orthologous sequences from several species, a collection of one or more position-specific weight matrices (WMs), and a phylogenetic tree relating the species. The user designates one of the species as the 'reference species' and MotEvo can then be asked to provide the following.

- Posterior probabilities for a TFBS for each possible WM to occur at each position in the input sequences of the reference species.
- The probability, at each position, for an *unknown functional element* (UFE) to occur, i.e. a TFBS for an unknown motif not contained in the input set.
- The estimated site densities for each WM and for UFEs (possibly fitted to the data).
- Updated versions of the WMs (fitted to the data).
- Log-likelihood ratio scores, at each position, for the occurrence of *cis*-regulatory modules containing TFBSs for the input motifs.

We now discuss the methods that MotEvo uses to calculate these quantities.

2.1 Binding site configurations

We first introduce some notation. We denote by $\{S\}$ a collection of multiple alignments of sequences, by S an individual multiple alignment (or a segment from such an alignment) and by s an individual sequence or sequence segment. To indicate the segment of length l from sequence s, starting at position $(i+1)$ we use the notation $s_{[i,l]}$, and similarly $S_{[i,l]}$ indicates columns $(i+1)$ through $(i+l)$ of the multiple alignment S. Column numbers are always counted with respect to the position in the reference sequence. As in most approaches, we assume that nucleotides at different positions in TFBSs are statistically independent and use position-specific WMs to represent TF binding specificities. We denote a collection of WMs by $\{w\}$ and a single WM from the set by w. The weight matrix entry w^i_α denotes the probability that nucleotide α occurs at position i of a binding site for WM w.

MotEvo considers all ways in which configurations of TFBSs (Fig. 1) for the WMs $\{w\}$ can be assigned to the sequences of the *reference species*. To explain how MotEvo calculates probabilities of possible configurations, we first explain how MotEvo scores a single hypothesized TFBS.

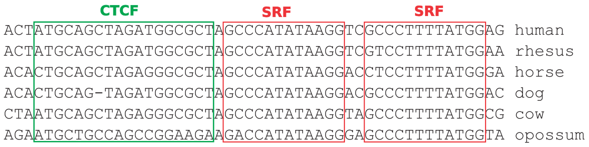

CTCF
SRF
SRF
ACTATGCAGCTAGATGGCGCTAGCCCATATAAGGTCGCCCTTTTATGGAG human
ACTATGCAGCTAGATGGCGCTAGCCCATATAAGGTCGTCCTTTTATGGAA rhesus
ACACTGCAGCTAGAGGGCGCTAGCCCATATAAGGACCTCCTTTTATGGGA horse
ACACTGCAG-TAGATGGCGCTAGCCCATATAAGGACGCCCTTTTATGGAC dog
CTAATGCAGCTAGAGGGCGCTAGCCCATATAAGGTAGCCCTTTTATGGCG cow
AGAATGCTGCCAGCCGGAAGAAGACCATATAAGGGAGCCCTTTTATGGTA opossum

Fig. 1. A segment from a multiple alignment of orthologous mammalian sequences, together with an example configuration of three binding sites: one for the motif of the CCCTC-binding factor TF (CTCF) and two for the motif of the serum response factor TF (SRF). Human is considered the reference species. Note that hypothesized TFBSs are not allowed to overlap within one configuration.

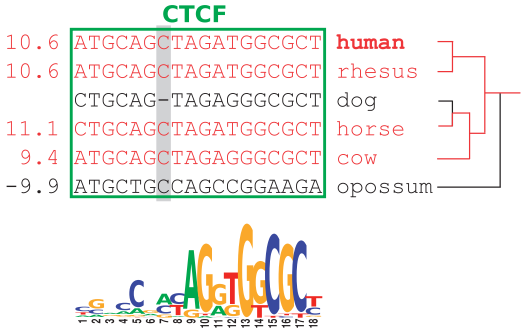

CTCF
10.6 ATGCAGCTAGATGGCGCT human
10.6 ATGCAGCTAGATGGCGCT rhesus
CTGCAG-TAGAGGGCGCT dog
11.1 CTGCAGCTAGATGGCGCT horse
9.4 ATGCAGCTAGAGGGCGCT cow
-9.9 ATGCTGCCAGCCGGAAGA opossum
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Fig. 2. A single hypothesized TFBS for CTCF on a segment of the multiple alignment. For each sequence s, the species from which it derives is shown on the right, and its WM score is shown on the left. The species with WM score larger than zero are selected (red sequences), the subtree involving these species of the full phylogenetic tree (red subtree of the black tree on the right) is obtained and the probability of the selected sequences is calculated under an evolutionary model that incorporates selective constraints set by the WM. A sequence logo (Schneider and Stephens, 1990) of the CTCF motif is shown below the alignment.

2.2 Probabilities under the evolutionary model

Figure 2 shows a single hypothesized site for the CTCF motif from the TFBS configuration of Figure 1. MotEvo calculates a probability ratio $P(S|w,T)/P(S|b,T)$ for observing this multiple alignment segment assuming that the sequences are evolving under constraints set by the WM w and assuming that the sequences are evolving 'neutrally' under a background model b, given the phylogenetic tree T. In contrast to most algorithms that implement explicit phylogenetic models, e.g. Moses *et al.* (2004), MotEvo takes into account that functional TFBSs may only occur in a subset of the species. Sites may either have been truly lost or gained in some species during evolution, or sites may appear to have been lost as a consequence of errors in the multiple alignments. After experimenting with several procedures for treating these possibilities, including explicit models that incorporate rates of gain and loss of sites along different branches of the tree, we found that the most robust results are obtained using the following *species selection* procedure.

We follow the generally made assumption that TFs bind DNA in a fixed configuration, so that TFBSs for a single TF have a fixed length. Consequently, only species that are gaplessly aligned with respect to the reference can have an orthologous TFBS at the same location in the alignment. For example, in Figure 2 the alignment implies that no orthologous site appears in dog. For every species that is gaplessly aligned relative to the reference, MotEvo calculates the probability of its sequence

s under the WM and under a background model b. The probability $P(s|w)$ of a sequence segment s under the WM w is simply given by the product of WM-components, i.e.

$$P(s|w)=\prod_{i=1}^{l} w_{s_i}^{i}, \tag{1}$$

where l is the length of the WM and s_i is the nucleotide occurring at position i in sequence s.

MotEvo allows for a variety of background models. In the simplest model, there are four parameters b_α representing the probabilities for a nucleotide α to occur at a background position. MotEvo also allows k-th order background models in which the probability of a nucleotide depends on the k preceding nucleotides, i.e. 4^{k+1} conditional probabilities $P(s_i|s_{i-1}s_{i-2}\dots s_{i-k})$ that are estimated from the input sequences by default. For the simple single-nucleotide background model, $P(s|b)$ is given by replacing the WM entries $w_{s_i}^{i}$ with the corresponding background probabilities b_{s_i} in Equation (1), and analogously for the higher order models.

We refer to the log-ratio $\log[P(s|w)/P(s|b)]$ as the *WM score* of sequence s. For the species selection procedure, MotEvo selects all species for which the WM score is larger than zero (the red sequences in Fig. 2). The key assumption that MotEvo now makes is that, whatever the reason is for the apparent loss of the TFBS from certain species, they should not contribute to the evolutionary evidence for a TFBS to occur at this position in the reference species. Specifically, we obtain the subtree T' that is defined by the subset of 'red' species (indicated in Fig. 2) and *replace* the probability ratio $P(S|w,T)/P(S|b,T)$ by the one obtained using only this subtree, i.e. by $P(S|w,T')/P(S|b,T')$. This ensures that the 'black' sequences in Fig. 2 do not contribute to the ratio $P(S|w,T)/P(S|b,T)$.

The probability ratio $P(S|w,T)/P(S|b,T)$ is the product of independent contributions from the individual alignment columns, i.e.

$$\frac{P(S|w,T)}{P(S|b,T)}=\prod_{i=1}^{l}\frac{P(S_i|w^i,T)}{P(S_i|b,T)}, \tag{2}$$

where w^i denotes the i-th WM column.

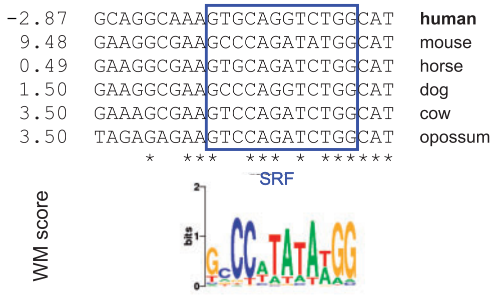
Imagine an alignment column that is evolving under the constraints set by WM column w^i and consider one branch of the phylogenetic tree T. Distances d along the branches of T are measured by the number of expected substitutions per neutrally evolving site. The key evolutionary quantities are the probabilities $P(\alpha|\beta,w^i,d)$ that, when evolving under WM column w^i along a branch of length d, a base β evolves into a base α. MotEvo uses a F81 model (Felsenstein, 1981). In this model, the transition probabilities are given by

$$P(\alpha|\beta,w^i,d)=\delta_{\alpha\beta}e^{-d}+w_\alpha^i\left(1-e^{-d}\right). \tag{3}$$

Although it is straightforward to implement more sophisticated evolutionary models, such as the model of Halpern and Bruno (1998) in practice the F81 model behaves very similarly. We chose the F81 model for consistency with the UFE model calculations, which require the F81 model to be computationally tractable (see below).

The probability $P(S_i|w^i,T)$ is given by multiplying the probabilities $P(\alpha|\beta,w^i,T)$ for the transitions at each of the branches of the tree, setting α to the corresponding nucleotide for branches leading to the leafs of the tree, and summing over the unknown nucleotides at all internal nodes of the tree. To calculate the sum over the nucleotides at the internal nodes, we use recursion relations introduced by Felsenstein (1981) (see Supplementary Materials for details).

For the single nucleotide background model, the probability $P(S|b,T)$ is calculated entirely analogously, i.e. simply replacing the WM column w^i with the column of background frequencies b in the above equations. One novel feature of MotEvo is that it allows the use of higher order background models in a phylogenetic setting by estimating the sequence context at internal nodes by averaging over their descendants in the tree (see Supplementary Materials for details).



```
-2.87  GCAGGCAAAGTGCAGGTCTGGCAT  human
 9.48  GAAGGCGAAGCCCAGATATGGCAT  mouse
 0.49  GAAGGCGAAGTGCAGATCTGGCAT  horse
 1.50  GAAGGCGAAGCCCAGGTCTGGCAT  dog
 3.50  GAAAGCGAAGTCCAGATCTGGCAT  cow
 3.50  TAGAGAGAAGTCCAGATCTGGCAT  opossum
             *  ***  *** * ******
```

Fig. 3. A small segment of a multiple alignment of orthologous mammalian sequences. The stars at the bottom of the alignment indicate columns that are perfectly conserved across all six species. A hypothetical binding site for the SRF motif is indicated (blue box) and the WM scores are shown for each of the sequences s. Without the UFE model, MotEvo assigns a posterior probability of 0.97 for a site for SRF to occur at this position, whereas with the UFE this probability drops to 0.01.

2.3 Unknown functional elements

Even though the number of TFs for which WM models are available is increasing steadily, the sequence specificity of the large majority of TFs is still unknown for most model organisms. Consequently, within the input alignments, there are likely many binding sites for TFs that are not represented by the WMs in our set $\{w\}$. Moreover, these TFBSs will often show significant evolutionary conservation, i.e. much more than can be expected under the background model and this can have undesirable consequences, as illustrated in Figure 3.

In this example, a TFBS for the SRF motif is predicted with high probability, despite the fact that the sequences show very poor matches to the WM (with the exception of mouse). The reason for this pathological behavior is that algorithms that do not explicitly take into account that UFEs may occur in the input sequences, are forced to choose at each position of the alignment between assuming that the alignment segment contains a TFBS for one of the WMs in the input set, or that the alignment segment contains neutrally evolving sequences. Given a segment that is much more conserved than can be expected under neutral evolution, such algorithms may thus assign a high posterior to a site for WM w occurring, even when the sequences in the segment poorly match the WM.

To avoid such spurious predictions, MotEvo explicitly takes into account that the input alignments will contain well-conserved segments for motifs other than those in our input set $\{w\}$, which we call *Unknown Functional Elements* (UFEs). To calculate the probability $P_{\text{ufe}}(S_k|T)$ of a single alignment column under the UFE, we integrate the probability $P(S_k|w^k,T)$ over all possible WM columns w^k, i.e.

$$P_{\text{ufe}}(S_k|T)=\int P(S_k|w^k,T)P(w^k)dw^k, \tag{4}$$

where $P(w^k)$ is a prior distribution over possible alignment columns for which MotEvo uses a Dirichlet prior (see Supplementary Material). To be able to calculate such integrals analytically, MotEvo uses the F81 model for the evolution along each branch of the tree. The Supplementary Materials provide details of the calculation of this integral.

Another parameter used by MotEvo is the length l_u of the UFE model, which is generally set to the typical length of TFBSs. The UFE model is then treated as any other WM, and the probability ratio $P_{\text{ufe}}(S|T)/P(S|b,T)$ is calculated for the l_u consecutive alignment columns of an hypothesized site of the UFE.

For the example shown in Figure 3, when the UFE model of length 10 is used, the posterior of the SRF motif drops to 0.01 and sites for the UFE are predicted in this area with moderate posteriors. Note that MotEvo can also be run using *only* the UFE model. In this way, a conservation profile that quantifies the evidence for purifying selection across the alignments can be obtained without the need of providing specific motifs (Molina and van Nimwegen, 2008), i.e. providing a functionality similar to algorithms such as phastcons (Siepel *et al.*, 2005).

2.4 Forward/backward algorithm

In contrast to MONKEY (Moses *et al.*, 2004) and other algorithms that scan with a single WM at a time, MotEvo predicts TFBSs for an arbitrary number of WMs and considers all possible configurations of non-overlapping TFBSs. Above we calculated the WM/background probability ratio $P(S|w,T)/P(S|b,T)$ for alignment segment S assuming a single hypothesized site for w. The probability ratio for an entire alignment given a configuration containing multiple binding sites is simply the *product* of the ratios for each of the binding sites. Note that all parts of the alignment where no binding sites occur do not contribute to this ratio, i.e. their contributions cancel between numerator and denominator.

To assign prior probabilities to configurations, MotEvo assumes that scanning the reference species sequence from left to right, there is at each position a probability π_w for a site for WM w to start. For notational simplicity, we consider both the UFE model and the background model b as members of our set $\{w\}$ of WMs. The prior probability for a binding site configuration in which there are n_w sites for WM w is then proportional to

$$\prod_w (\pi_w)^{n_w}. \tag{5}$$

Note that we have the normalization condition $\sum_w \pi_w = 1$.

To calculate posterior probabilities, we will need to sum over the probability ratios of all possible binding site configurations, i.e. calculate a partition sum. To this end, we use recursion relations similar to those of the forward/backward algorithm used in the theory of hidden Markov models (Durbin *et al.*, 1998). Let the sum of the probability ratios of all possible configurations of TFBSs up to position n in the reference species be denoted by F_n. Noting that any configuration ending at position n in the reference species has to end with a site for one of our WMs (which now include the background and UFE models), we have the recursion relation

$$F_n = \sum_{w \in \{w\}} \pi_w \frac{P(S_{[n-l_w,l_w]}|w,T)}{P(S_{[n-l_w,l_w]}|b,T)} F_{n-l_w}, \tag{6}$$

where l_w is the length of WM w.

Instead of moving from left to right over the multiple alignment, we can also move from right to left and define R_n as the sum over the probability ratios of all possible binding site configurations from position n until the end of the alignment. Further details are provided in the Supplementary Materials.

2.5 TFBS predictions

Once the forward and backward sums F_n and R_n have been obtained, we can calculate the posterior probabilities $P(w,n|S,\{w\},T)$ that a binding site for WM w occurs at positions $n+1$ through $n+l_w$:

$$P(w,n|S,\{w\},T) = \frac{F_n \frac{P(S_{[n,l_w]}|w,T)}{P(S_{[n,l_w]}|b,T)} \pi_w R_{n+l_w+1}}{F_L}, \tag{7}$$

where F_L is the sum over probability ratios of all configurations for the entire alignment of length L. Note that the sum over all configurations in which a site for w occurs at n is equal to the sum over all possible configurations up to position n and all configurations from $(n+l_w+1)$ onwards. Using the procedures described above, posterior probabilities $P(w,n|S,\{w\},T)$ at every position n for every WM w can be calculated in a time that is proportional to product of the length of the multiple alignment L, the number of WMs and the number of species in the alignment. This linear scaling of the run-time allows MotEvo to make comprehensive site predictions for very large sequences using a large number of WMs in relatively short computational times. For example, when a TF is known to have different types of binding sites, e.g. half-sites separated by spacers of different lengths, MotEvo can easily run with WMs for each site type in parallel.

Note that, when running on a single sequence, MotEvo is equivalent to a statistical mechanical model that calculates the binding frequencies along the genome of multiple factors competing for binding along the genome, i.e. on a single sequence MotEvo is equivalent to the approach presented in Wasson and Hartemink (2009). In this biophysical interpretation, the WM scores correspond to binding energies, the WM priors correspond to the concentrations of the different factors and the posterior probabilities correspond to the fractions of time a given TF is bound at a given site.

2.6 Prior updating

The prior probability distribution over binding site configurations is parametrized by the vector π that gives the expected binding site density π_w for each WM w, including the background and UFE models. This prior π can be specified by the user, but MotEvo can also use an expectation–maximization algorithm to find the vector π that maximizes the probability of the observed alignments $\{S\}$.

We start with an initial prior vector π and calculate the posterior probabilities $P(w,n|S,\{w\},T)$ for all alignments in the set $\{S\}$. We then calculate, for each WM, the sum n_w of the posterior probabilities $P(w,n|S,\{w\},T)$ over all positions n in all alignments S. That is, n_w represents to the total expected number of binding sites for WM w. Using these MotEvo calculates a new prior vector

$$\pi_w = \frac{n_w}{\sum_{w' \in \{w\}} n_{w'}}, \tag{8}$$

and calculates new posterior probabilities $P(w,n|S,\{w\},T)$ using this new prior vector. This procedure is iterated until the prior vector converges. It is easy to show, see e.g. van Nimwegen (2007), that this expectation–maximization procedure maximizes the probability of the input alignments with respect to the prior vector π.

2.7 Enhancer prediction

Enhancers are *cis*-regulatory elements on the genome that are distal to the promoter of the gene whose expression they regulate. They are characterized by a high density of TFBSs for a particular subset of TFs and they are typically a few hundred base pairs in length. They can occur both upstream, downstream or in an intron of their target gene (Arnosti and Kulkarni, 2005).

To find enhancers, MotEvo extends previously developed algorithms (Rajewsky *et al.*, 2002; Sinha *et al.*, 2003) to multiple alignments with an arbitrary number of species and phylogenetic relationships. A window of a given length (typically a few hundred base pairs) is slid over the input alignments and for each window MotEvo predicts posterior probabilities of binding site occurrence for the set of input WMs. Importantly, MotEvo then updates the priors π_w *separately* for each window, allowing it to adapt the binding site densities to each window. Note that, because TFBSs cannot overlap within a single configuration, the prior updating roughly tries to maximize the number of TFBSs for the input WMs that can be bound at the same time to a given region. To assign a final *enhancer score* to a window, MotEvo calculates the log-ratio of the sum of the probabilities of all possible binding site configurations and the probability of the configuration with only background columns.

2.8 Weight matrix refinement

MotEvo also implements an expectation–maximization procedure for refining WMs based on its TFBS predictions. Formally, the idea is to maximize the probability of the entire input data $\{S\}$ with respect to the WMs $\{w\}$, starting from the WMs that were provided as input. As shown in the Supplementary Materials, this maximization can be obtained to a good approximation, using the following procedure.

Starting from the input WMs $\{w\}$, MotEvo first predicts posterior probabilities $P(w,n|S,\{w\},T)$ for TFBS occurrence of each WM w at each position n in each input alignment S. It then calculates, for each WM w, each position i in the WM and each nucleotide α, the sum $n^i_\alpha(w)$ of the posterior probabilities of all putative TFBSs that have a nucleotide α occurring at position i of the site in the reference species. That is, the statistics $n^i_\alpha(w)$ calculate the expected total number of TFBSs for WM w that, in the reference species, have nucleotide α at position i. Note that MotEvo can also be instructed to ignore sites with posteriors below a cut-off in calculating these sums. MotEvo then updates the WM as follows:

$$w^i_\alpha = \frac{n_{i\alpha}(w)}{\sum_\beta n_{i\beta}(w)}. \qquad (9)$$

Site predictions are then performed with these updated WMs and this procedure is repeated until the WMs converge. Using this procedure, MotEvo thus extends the functionality of the PhyME algorithm (Sinha *et al.*, 2004), allowing for the simultaneous expectation–maximization of multiple motifs in parallel.

3 RESULTS

Although a key benefit of MotEvo as a computational tool is that it integrates cutting-edge methods within a single executable, we here focus on evaluating the performance benefits of the novel features that MotEvo implements. For benchmarking, we collected five datasets (Jothi *et al.*, 2008; Valouev *et al.*, 2008) in which chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) was performed for the human TFs CTCF, GABP, NRSF, SRF and STAT1. Using the peak finder MACS (Zhang *et al.*, 2008), we determined the regions bound by the TF in question for each ChIP-seq dataset. Binding regions that occurred in more than one dataset were removed to ensure that (at least in the conditions tested) only one TF is bound to each region. Finally, we selected the 900 regions with highest enrichment from each dataset (see Supplementary Materials for details).

Using pairwise genome alignments from the UCSC database (Karolchik *et al.*, 2008), we extracted orthologous regions from six other mammals (mouse, dog, cow, monkey, horse and opossum) and obtained seven-way multiple alignments using T-coffee (Notredame *et al.*, 2000). For NRSF, GABP, SRF and STAT1 known WM motifs were taken from the literature (Vlieghe *et al.*, 2006) and a CTCF WM was inferred from ChIP-chip data in fly (Holohan *et al.*, 2007).

To test the new features, we predict TFBSs with MotEvo on the benchmarking dataset both using the feature and with the feature turned off. We then compare the TFBS predictions and evaluate to what extent the TFBS predictions are able to infer which region was bound by which TF. In addition, we compared MotEvo's performance on these datasets with those of MONKEY (Moses *et al.*, 2004) and PhyloScan (Carmack *et al.*, 2007; Palumbo and Newberg, 2010). Finally, we test the performance of motifs obtained using MotEvo's motif refinement and compare it with the performance of motifs inferred by MEME (Bailey and Elkan, 1994).

3.1 The UFE model strongly reduces spurious predictions

We argued above that, without the UFE, highly conserved regions are often mistakenly predicted as TFBSs for WMs in our set, even when the sequences poorly match the corresponding motif. To test this, we predicted TFBSs on all regions, once including the UFE and once without it. For each TF, we determined the WM score of the sequence occurring in the reference species at each predicted TFBS, and constructed histograms of the distributions of WM scores (Fig. 4 and Supplementary Fig. S1).

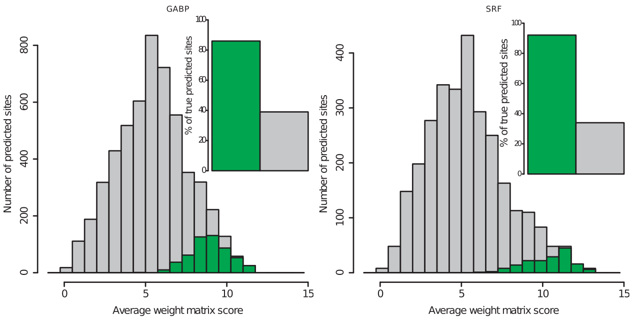


Fig. 4. Comparison of TFBS predictions with and without inclusion of the UFE model. The histograms show the distribution of WM scores (log-ratio $\log[P(s|w)/P(s|b)]$ for the sequence s occurring in the reference species) of the TFBSs predicted by MotEvo with (green) and without (gray) usage of the UFE. The insets show the fractions of predicted sites that fall within the 'correct' regions, i.e that were immunoprecipitated with the corresponding TF. Results are shown for the TFs GABP and SRF. Results for all TFs are shown in Supplementary Figures S1 and S2.

Inclusion of the UFE in general leads to substantial changes in the predicted TFBSs. First of all, the total number of predicted TFBSs is much lower with the UFE. Second, the UFE specifically causes predicted TFBSs that have a weak match to the motif to disappear. Finally, using the UFE the fraction of predicted TFBSs that fall in 'correct' regions, i.e. regions that were immunoprecipitated with the corresponding TF often increases dramatically, i.e. from around 40% to over 90% for three of the five TFs (insets of Fig. 4 and Supplementary Fig. S1). However, from this test it is not clear if this increased specificity comes with a cost in sensitivity of the site predictions, which we address in the following test.

3.2 MotEvo's novel features improve TFBS prediction

To test MotEvo's performance, and the role of its novel features in a realistic setting, we tested how accurately the TFBS predictions can distinguish which region was bound by which of the five TFs. For each motif w and each region r, we assign a score $n(r,w)$ by summing the posterior probabilities of all predicted TFBSs for w in r. We then obtain a sensitivity/positive predictive value (PPV) curve by, as a function of a cut-off on the score $n(r,w)$, calculating the fraction of all regions bound by the corresponding TF that have a score above the cut-off (sensitivity) and the fraction of all regions with score above the cut-off that were indeed bound by the TF (PPV). We gather such sensitivity/PPV curves using MotEvo in its standard form, with the UFE model turned off, and with species selection turned off, i.e. including sequences from all gaplessly aligned species at each putative site. We also obtained binding site predictions for two TFBS prediction algorithms that also incorporate an explicit evolutionary model: MONKEY (Moses *et al.*, 2004) and PhyloScan (Carmack *et al.*, 2007; Palumbo and Newberg, 2010) (see Supplementary Materials for details), and determined their sensitivity/PPV curves.

We first of all see that, without the UFE, MotEvo's performance is dramatically reduced. In particular, because of the large number of spuriously predicted sites, no high specificity can be obtained without the UFE (Fig. 5 and Supplementary Fig. S2). Only at very high sensitivities, i.e. when detecting even the regions with the

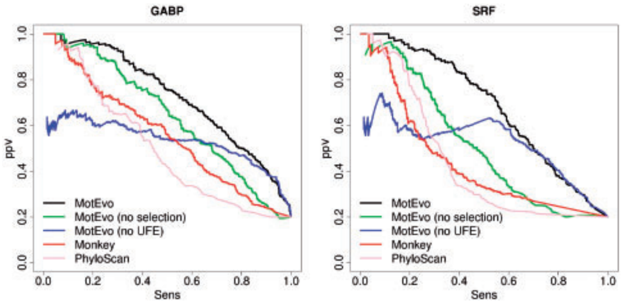


Fig. 5. Comparison of TFBS prediction accuracy for MotEvo (black), MotEvo without the UFE (blue), MotEvo without species selection (green), Monkey (red) and PhyloScan (pink). TFBS predictions were made on the benchmark set of 5 times 900 regions by each of the methods and were then used to predict, for each TF, which of the 4500 regions were bound by the TF. The panels show sensitivity/PPV curves for the performance obtained predicting the regions bound by the TFs SRF and GABP. Results for all TFs are shown in Supplementary Figure S2.

weakest TFBSs, is the performance relatively unaffected. Besides the UFE, MotEvo explicitly considers that TFBSs may have been lost in a subset of the species, either through evolution or simply because of errors in the multiple alignment, by using the 'species selection' scheme described above. We find that species selection leads to an increase in performance for all TFs (Supplementary Fig. S2) ranging from small improvements at some TFs (NSRF, CTCF) to moderate or even very large improvements for others (GABP and SRF, Fig. 5). Thus, MotEvo's method for treating loss and gain of TFBSs within the phylogeny significantly outperforms methods that assume that all sequences in the multiple alignment are evolving under the same selective constraints determined by the WM.

Although we invested some efforts optimizing the performance of MONKEY and PhyloScan on this dataset (see Supplementary Materials), MotEvo significantly outperforms these algorithms (Fig. 5 and Supplementary Fig. S2). Especially striking is the inability of these algorithms to reach high sensitivity for motifs with high information content (NRSF, CTCF). Manual inspection of the differences in the predictions of MotEvo, PhyloScan and MONKEY strongly suggest that MONKEY and PhyloScan's performance is most affected by their inability to deal with alignment segments where a binding site occurs in only some of the species, i.e. where some of the species have either gaps relative to the reference species or low WM scores (see Supplementary Materials for more discussion). These algorithms effectively assume that a functional site must appear in all species of the alignment, and for multiple alignments involving seven species there are many cases where this is too restrictive an assumption.

Finally, we checked whether MotEvo's performance is strongly affected by the alignment algorithm used (Supplementary Fig. S6) and find that sensitivity/PPV curves change only marginally when using different alignment methods. Similarly, use of a higher order background model also only marginally improves performance on these benchmarking datasets (Supplementary Fig. S6).

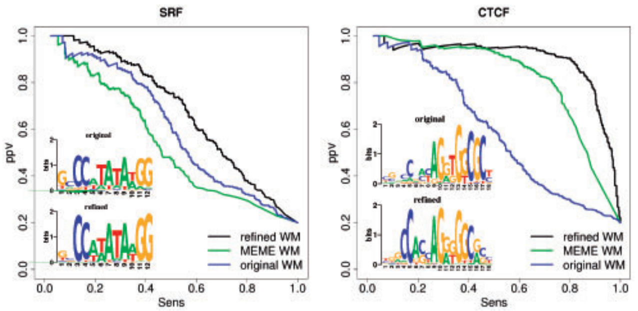


Fig. 6. Comparison of the performance in predicting TF binding of original WMs based on literature, WMs refined by MotEvo and WMs inferred by MEME. Binding sites were predicted by MotEvo on all 5×450 regions in the test-set using the three WMs for each of the five TFs. The predicted TFBSs were used to predict which TF is bound by each region. The panels show sensitivity/PPV curves for the performance obtained using the original (blue), the refined motif (black) and MEME's motif (green) for the TFs SRF and CTCF. Sequence logos of the original and refined WMs are shown in each panel as well. Results for all TFs are shown in Supplementary Figure S3.

3.3 WM refinement improves TFBS predictions

Through the availability of next-generation sequencing technologies, many laboratories have started performing ChIP-seq of TFs of interest, and the number of available datasets is increasing rapidly. As ChIP-seq is able to identify large numbers, i.e. hundreds to thousands, of binding regions for a given TF genome-wide, this offers the possibility to investigate the binding specificity of TFs at much higher levels of resolution than was previously possible.

MotEvo implements an expectation–maximization strategy for refining WMs from an input dataset of binding regions which we here test using the same benchmarking ChIP-seq data for five TFs. For each of the five TFs, we randomly selected 450 of the 900 peaks and pooled them into one large dataset that we used as a *test set*. The remaining 450 regions for each TF were used as a *training set* for WM refinement. Besides MotEvo, we also used MEME (Bailey and Elkan, 1994) to infer a WM motif for each of the five test sets. Sequence logos of the original and inferred motifs are shown in Figure 6 and Supplementary Figures S3 and S4.

To test the performance of these WMs in identifying which TF binds to which target region, we predicted TFBSs for all original and refined WMs on the *test set* using MotEvo. As in the previous section, we assign a score $n(r,w)$ for each WM w to each region r by summing the posterior probabilities of predicted TFBSs, and obtain sensitivity/PPV curves that quantify the performance of the TFBSs in predicting which TF was bound by each region (Fig. 6 and Supplementary Fig. S3).

The improvement that refinement provides over the literature motif ranges from virtually no difference between original and refined WMs (GABP), through moderate improvements (SRF), to very dramatic improvements (CTCF). It is notable that, even when there is a clear difference in performance of the original and refined WMs, the sequence logos appear very similar visually. This illustrates that subtle changes in the WM can have substantial effects on TFBS predictions. Interestingly, in the case of STAT1 the refined motif is closer to a true palindrome than the original motif, suggesting that these sites are bound by the TF in dimer form. This is supported by the literature, i.e. it is known that STAT1 is 'activated' through phosphorylation at a tyrosine, after which STAT1 proteins form homo-dimers that are translocated to the nucleus to regulate transcription (McBride and Reich, 2003). The most dramatic improvement in performance is observed for CTCF.

Table 1. Overlap between the known and predicted blastoderm *cis*-regulatory modules as a function of the number of *Drosophila* species used

Number of species	1	2	3	5	7	9
Performance	0.57	0.70	0.73	0.76	0.82	0.93

This is not surprising given that, in contrast to the other TFs in this set, for CTCF the original WM was based on data from *Drosophila.* That is, it is plausible that the precise sequence specificity of CTCF may differ between *Drosophila* and human.

MotEvo's refined WMs outperform the WMs inferred by MEME for all TFs (Fig. 6 and Supplementary Fig. S3). Although the difference is marginal for NRSF, for some of the TFs MEME's motif performs clearly worse than the literature motif (see Supplementary Materials for further discussion). In summary, our results show that MotEvo's WM refinement consistently improves the ability of the WM to distinguish regions bound by the TF from regions that are bound by other TFs. In addition, this improvement can be very large in some cases.

3.4 Enhancer prediction accuracy increases with the number of species used

Finally, we evaluated MotEvo's ability to predict distal *cis*-regulatory modules (also called enhancers). For testing, we used the set of 76 experimentally validated blastoderm *cis*-regulatory modules (CRMs) that was collected in Ivan *et al.* (2008). The length of these CRMs ranges from ~90 bp to 2 kb. For each CRM, we extracted the flanking region of 5 kb before and after the CRM on the genome. We then multiply-align these regions with other sequenced *Drosophila* species. As it is likely that these regions contain other blastoderm CRMs that are unknown, we shuffle the aligned columns of the flanking region without changing the conservation pattern, that is to say we substitute each column by a column with similar conservation (same gap pattern and subset of species that has the same base as the reference). Exonic regions are also excluded.

To test how prediction accuracy depends on the number of species used in the multiple alignments, we made several pruned versions of the multiple alignments by selecting different subsets of the available species, ranging from only the melanogaster sequence, to sequences from nine available species (see Supplementary Materials and Supplementary Fig. S5 for details). We use MotEvo with seven WMs for Drosophila TFs known to be involved in binding to these enhancers (Bcd, Cad, Dl, Hb, Kni, Kr and Tll), to perform enhancer predictions on all multiple alignments, selecting for each alignment the 900 bp window with the highest enhancer score as the predicted enhancer. To assess the performance, we calculated for each alignment the overlap between the predicted and the known enhancer. As Table 1 shows, the performance increases significantly as the number of species increases, reaching over 90% performance when nine species are used. This illustrates that MotEvo's ability to predict enhancers on multiple alignments of an arbitrary number of species significantly improves accuracy over methods that only allow pairwise analysis.

4 DISCUSSION

From a theoretical point of view, the major advantage of the MotEvo method we presented here is that it integrates cutting-edge Bayesian probabilistic methods for the prediction of TFBSs, regulatory motifs and conservation patterns within one consistent theoretical frame work, developing several novel features such as the UFE model and species selection in the process. In addition, our benchmarking tests have demonstrated that these features improve MotEvo's performance, and that MotEvo outperforms currently available methods. Another major advantage of the MotEvo tool is its versatility, i.e. by simple changes to the parameter file the tool can perform a wide array of tasks ranging from motif inference, to enhancer prediction, to conservation profile mapping, site density estimation and of course TFBS prediction. Moreover, essentially all variables used by the algorithm, from phylogeny to background models, to priors, can be controlled by the user, allowing these to be adapted to a wide range of applications. For example, using MotEvo in combination with genome-wide mapping of transcription start sites we have predicted functional TFBSs for hundreds of WMs in proximal promoters in human, mouse (Suzuki *et al.*, 2009), yeast (Chen *et al.*, 2010) and *Escherichia coli*. We have also obtained refined WMs using MotEvo on ChIP-seq datasets for a number of TFs beyond those studied here. All these TFBS and motif predictions are available for download from our SwissRegulon database at www.swissregulon.unibas.ch.

As experimental validation of the functionality of individual TFBSs is extremely labor intensive, it remains highly challenging to estimate the accuracy of large-scale TFBS predictions. In the past, it has sometimes been assumed that ChIP-chip and ChIP-seq datasets can be treated as a gold standard, but our own analysis suggests that computational predictions can, in fact, be considerable more accurate in mapping functional sites than such high-throughput experimental approaches (Chen *et al.*, 2010).

What is clear is that TF binding and function is highly context dependent. For a TF with a short degenerate motif, there may be millions of sites genome wide with motif matches at least as high as known functional sites, but in a typical ChIP-seq experiment only a few thousand of these are found to be actually bound, and even among these only a subset may directly affect gene expression. In the search for variables that provide important context, it is important to distinguish those that are merely *predictive* for the functionality of TFBSs from those that are *explanatory*. Cross-species conservation is an example of an explanatory variable, i.e. highly conserved TFBSs are more likely to be functional, but conservation is not explanatory; sequences in other species cannot explain why a particular sequence is bound or functional in a given species.

In higher eukaryotes, the chromatin state is likely to be an important explanatory variable. In areas where the nucleosomes are densely packaging the DNA, it may be hard for a TF to access an individual site in the DNA. This may explain why functional TFBSs come in clusters of nearby sites for co-expressed TFs: these TFs may passively cooperate in displacing the nucleosomes from the DNA. Such a model can potentially explain the observation that the genomic binding pattern observed for a given TF is dependent on the expression profiles of other TFs (Wilczynski and Furlong, 2010). However, to what extent binding and function of TFBSs is dependent on a high-order 'grammar' of TFBS configurations, i.e. the precise spacing and relative orientation of the sites, is currently unclear. In order to make progress on these important questions, we believe that the largest potential lies in integrating sequence analysis with the analysis of temporal patterns of TF binding, of genome-wide chromatin states and of the expression of potential target genes.

Funding: The work presented here was supported by Swiss National Science Foundation grant (31003A-118318) and by SystemsX.ch through the CellPlasticity project.

Conflict of Interest: none declared.

REFERENCES

Arnosti,D.N. and Kulkarni,M.M. (2005) Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell Biochem.*, **94** 890–898.

Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Sec. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.

Carmack,C.S. *et al.* (2007) PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algor. Mol. Biol.*, **2**, 1.

Chen,K. *et al.* (2010) Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. *Genome Biol. Evol.*, **2**, 697–707.

Davidson,E.H. (2001) *Genomic Regulatory Systems*. Academic Press, San Diego.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge, UK.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Frith,M.C. *et al.* (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.

Halpern,A.L. and Bruno,W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, **5**, 910–917.

Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.

Hawkins,J. *et al.* (2009) Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, **25**, i339–i347.

Holohan,E.E. *et al.* (2007) CTCF genomic binding sites in *Drosophila* and the organisation of the bithorax complex. *PLoS Genet.*, **3**, e112.

Ivan,A. *et al.* (2008) Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol.*, **9**, R22.

Jothi,R. *et al.* (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.

Karolchik,D. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.

Kellis,M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Lawrence,C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

McBride,K.M. and Reich,N.C. (2003) The ins and outs of STAT1 nuclear transport. *Sci. STKE*, **2003**, RE13.

Molina,N. and van Nimwegen,E. (2008) Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res.*, **18**, 148–160.

Moses,A.M. *et al.* (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **5**, R98.

Notredame,C. *et al.* (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol.*, **302**, 205–217.

Palumbo,M.J. and Newberg,L.A. (2010) Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. *Nucleic Acids Res.*, **38**, W268–W274.

Rajewsky,N. *et al.* (2002) Computational detection of genomic cis-regulatory modules, applied to body patterning in the early drosophila embryo. *BMC Bioinformatics*, **3**; doi:10.1186/1471-2105-3-30.

Roider,H.G. *et al.* (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Siddharthan,R. *et al.* (2005) Phylogibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Sinha,S. *et al.* (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19** (Suppl. 1), i292–i301.

Sinha,S. *et al.* (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.

Suzuki,H. *et al.* (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, **41**, 553–562.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.

van Nimwegen,E. (2007) Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, **8** (Suppl. 6), S4.

Vlieghe,D. *et al.* (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.

Wasson,T. and Hartemink,A.J. (2009) An ensemble model of competitive multi-factor binding of the genome. *Genome Res.*, **19**, 2101–2112.

Wilczynski,B. and Furlong,E.E. (2010) Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.*, **6**, 383.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.