On the value of intra-motif dependencies of human insulator protein CTCF

Text S3: Mathematical details

1 Basic definitions

The input data for de novo motif discovery consists of a set of DNA sequences of variable length. We denote the $j$-th nucleotide in the $i$-th input sequence by $X_{i,j} \in \mathcal{A} = \{A, C, G, T\}$, the $i$-th sequence by $\vec{X}_i = (X_1, \ldots, X_{L_i})$, and the complete data set by $\mathbf{X} = (\vec{X}_1, \ldots, \vec{X}_N)$. We denote the reverse complement of sequence $\vec{X}$ by $\text{rc}(\vec{X})$.

2 Model

2.1 ZOOPS model

The modularity of the ZOOPS model allows for combining an arbitrary motif model with parameters $\Theta_m$ with an arbitrary flanking model with parameters $\Theta_f$. The binary latent variable $u_i$ handles the situation that the $i$-th sequence contains ($u_i = 1$) or contains not ($u_i = 0$) a binding site. We model the position of the binding site of width $W$ by the latent variable $v_i \in \{1, \ldots, L_i - W + 1\}$. The binding site may occur on both strands, so we introduce a third latent variable $s_i \in \{F, R\}$, indicating whether the binding site occurs on the forward strand ($s_i = F$) or on the reverse complement strand ($s_i = R$). We denote the latent variables of the complete data set by $\vec{u} = (u_1, \ldots, u_N)$, $\vec{v} = (v_1, \ldots, v_N)$, and $\vec{s} = (s_1, \ldots, s_N)$.

The conditional likelihood of $\mathbf{X}$ given the latent variables $(\vec{u}, \vec{v}, \vec{s})$ and parameters $\Theta_m, \Theta_f$ is given by

\[
P(\mathbf{X}|\vec{u}, \vec{v}, \vec{s}, \Theta_m, \Theta_f) = \prod_{i=1}^{N} P_f(\vec{X}_i|\Theta_f)^{\delta_{u_i,0}} \cdot P_f(X_{i,v_i}, \ldots, X_{i,v_i+W-1}|\Theta_f)^{\delta_{u_i,1}} \cdot P_m(X_{i,v_i}, \ldots, X_{i,v_i+W-1}|\Theta_m)^{\delta_{s_i,F}} \cdot P_m(\text{rc}(X_{i,v_i}, \ldots, X_{i,v_i+W-1})|\Theta_m)^{\delta_{s_i,R}},
\]
where $\delta_{a,b}$ denotes the Kronecker delta, which is 1 if $a = b$ and 0 otherwise. We assume $v_i$ to be uniformly distributed, that is, $P(v_i) = \frac{1}{L_i}$. We denote the probability that a sequence contains a binding site by $\nu$, that is, $P(u_i|\nu) = \nu \delta_{u_i,1} (1-\nu) \delta_{u_i,0}$ for $i \in (1, \ldots, N)$. We further denote the probability a binding site is on the forward strand as $\sigma$, that is, $P(s_i|\sigma) = \sigma \delta_{s_i,F} (1-\sigma) \delta_{s_i,R}$ for $i \in (1, \ldots, N)$. We define the component parameter set $\Theta_c = (\nu, \sigma)$. Assuming statistical independence among $u_i, v_i, s_i$, the probability over the space of latent variables is given by

$$P(\vec{u}, \vec{v}, \vec{s}|\Theta_c) = \prod_{i=1}^{N} P(u_i|\nu) P(v_i) P(s_i|\sigma).$$

(2)

For computing the (marginal) likelihood of data $X$ given parameters $\Theta = (\Theta_m, \Theta_f, \Theta_c)$, we sum over all latent variables, yielding

$$P(X|\Theta) = \sum_{\vec{u}} \sum_{\vec{v}} \sum_{\vec{s}} P(X|\vec{u}, \vec{v}, \vec{s}, \Theta_m, \Theta_f) P(\vec{u}, \vec{v}, \vec{s}|\Theta_c).$$

(3)

For the case studies performed in this paper, we utilize the modular implementation of the ZOOPS model provided by the open-source software library Jstacs [1].

2.2 Flanking model

We model all nucleotides of a sequence that are not part of a binding site by a homogeneous Markov model of order $M$, abbreviated hMM($M$) and specified by the probability distribution $P_f$. Due to the homogeneity of the Markov model, the ratio between the number of parameters and the number of nucleotides in the data set is large enough to support even higher orders without observing overfitting effects [2]. The parameter set $\Theta_f$ consists of parameters $\vec{\lambda}^z$ (with $0 \leq z \leq M$), which encode the logarithms of the conditional probabilities of observing a symbol given its $z$ predecessors. More precisely, $\vec{\lambda}^z$ consists of $|A|^z$ conditional probability distributions of dimension $|A|$, and $\lambda_{a_1,...,a_z,b}^z$ is the conditional probability of finding symbol $b$ when the last $z$ observations are $a_1, \ldots, a_z$. We write the log-likelihood of the hMM($M$) for an arbitrary sequence $\vec{X}$ as

$$\log P_f(\vec{X}|\Theta_f) = \sum_{z=0}^{M-1} \lambda_{X_1,...,X_z,X_{z+1}}^z + \sum_{\ell=M+1}^{L} \lambda_{X_{\ell-M},...,X_{\ell}}^M.$$

(4)

2.3 Motif model

After having specified the flanking model, we next focus on the motif model. In analogy to the flanking model, we intend to model statistical dependencies among adjacent nucleotides.

To this end, we use an inhomogeneous parsimonious Markov model (PMM) [3], which is a variant of parsimonious Markov models [4, 5]. PMMs generalize variable order Markov models (VOMMs) [6, 3], as they are probabilistic models based on a generalization of
context trees [7]. They are capable of combining different context sequences into sets, represented by a **parsimonious context tree** (PCT), which is defined as follows.

A PCT is a rooted, labeled, and fully balanced tree, which we subsequently denote by \( \tau \). With exception of the root, each node is labeled with a subset of \( \mathcal{A} \). The labels of all children of an arbitrary inner node form a partition of \( \mathcal{A} \), i.e., each symbol \( a \in \mathcal{A} \) occurs exactly once in the labels of the sibling nodes.

We obtain a set of strings by building the cross product of the labels of the nodes on the path from an arbitrary leaf up to the root. We call this set **context**.

An inhomogeneous parsimonious Markov model of order \( D \) for a sequence \( \vec{Y} \) of fixed width \( W \) utilizes exactly \( W \) PCTs, denoted by \( \vec{\tau} = (\tau_1, \ldots, \tau_W) \), one for each position in the motif. For the ease of presentation, we exclude from the following discussion the first \( D \) PCTs, which have an increasing maximal depth of 0, \ldots, \( D - 1 \). We denote a single context as \( w \), and the set of all contexts represented by a specific parsimonious context tree \( \tau \) by \( \mathcal{C}_\tau \). The PMM contains conditional probability parameters for each position \( \ell \in \{1, \ldots, W\} \) and for each context \( w \in \mathcal{C}_\tau \). Hence, we denote the logarithm of the probability of observing symbol \( a \in \mathcal{A} \) at position \( \ell \) under the condition that \((Y_{\ell-D}, \ldots, Y_{\ell-1}) \in w\) by parameter \( \lambda_{\ell w a} \).

Each leaf of a PCT contains a conditional probability distribution for observing a specific symbol given the context of that leaf. Let \( \lambda_{\ell w} \) denote the parameters of one leaf of the tree at position \( \ell \), we denote the parameters corresponding to a complete tree at a given position by \( \Lambda_{\ell \tau} \). We combine all parameters of the entire motif model \( \Theta_m = (\vec{\Lambda}_\tau, \vec{\tau}) \) accordingly. Using this notation, we define the log-likelihood of a parsimonious Markov model for multiple sequences by

\[
\log P_m(\vec{Y}|\Theta_m) = \sum_{\ell=1}^{W} \sum_{w \in \mathcal{C}_\tau} \sum_{a \in \mathcal{A}} N_{\ell wa} \lambda_{\ell wa},
\]

where \( \vec{Y} \) is a set of aligned sequences of length \( W \) and \( N_{\ell wa} \) is the number of occurrences of symbol \( a \) at position \( \ell \) in all sequences of \( \vec{Y} \) whose subsequence from position \( \ell - D \) to \( \ell - 1 \) is in \( w \). The only difference to fixed-order Markov models is \( w \). Instead of allowing \( w \) to be an arbitrary set of strings, a Markov model requires \( w \) to be one single string.

### 3 Prior

As structure prior, we assume a distribution of the following form

\[
\log P(\vec{\tau}|\kappa) = \kappa \cdot \sum_{\ell=1}^{W} |C_{\tau_\ell}| - c,
\]

where \( \sum_{\ell=1}^{W} |C_{\tau_\ell}| \) is the number of leaves of all PCTs \( \vec{\tau} \) used by the motif model, and \( c \) is the normalization constant.

To adapt the number of leaves of the model, that is, the total number of leaves of all context trees, we vary the hyperparameter \( \kappa \). Since each leaf represents a set of
conditional probability parameters, we are capable of influencing the model complexity by varying $\kappa$. However, $\kappa$ only encodes a-priori assumptions about the model complexity. The same structure prior, i.e. the same value of $\kappa$, might yield models of very different complexity for two different data sets. As parameter prior for the motif model we use a transformed product of Dirichlet distributions

$$\log P(\vec{\Lambda}|\vec{\tau}) = \sum_{\ell=1}^{W} \sum_{w \in C_{\ell}} \log \text{Dir}(\vec{\lambda}_{\ell w} | \vec{\alpha}_{\ell w}),$$

(7)

with

$$\log \text{Dir}(\vec{\lambda}_{\ell w} | \vec{\alpha}_{\ell w}) = \log \Gamma \left( \sum_{a \in A} \alpha_{\ell w a} \right) - \sum_{a \in A} \log \Gamma(\alpha_{\ell w a}) + \sum_{a \in A} \alpha_{\ell w a} \lambda_{\ell w a}$$

(8)

We require the hyperparameters $\vec{\alpha}_{\ell w}$ to satisfy the equivalent sample size constraints, which yields a natural computation of the hyperparameters $\vec{\alpha}_{\ell w}$ from the equivalent sample size $\eta$ that is inspired by Bayesian networks [8], namely $\alpha_{\ell w} = \frac{\eta |w|}{|A|}$, with $d = \min(\ell - 1, D)$. In this study, we further restrict the parameter prior to a symmetric Dirichlet, leaving $\eta$ as sole free hyperparameter of the parameter prior. For the component parameters $\Theta_c$ we assume $P(\Theta_c) = P(\nu)P(\sigma)$ with $P(\nu) = \text{Beta}(a_1, a_2)$ and $P(\sigma) = \text{Beta}(b_1, b_2)$, where $\vec{a}$ and $\vec{b}$ are hyperparameters of a Beta distribution. We set – in the spirit of the equivalent sample size concept – the hyperparameters $\vec{a} = (\eta, \eta)$ and $\vec{b} = (\frac{\eta}{2}, \frac{\eta}{2})$. In all case studies, we use $\eta = 64$.

4 Learning

In order to learn a parsimonious Markov model from a set of promoter sequences, we apply the maximum a posteriori principle to the ZOOPS model. We intend to solve the following problem

$$\hat{\Theta} = \arg\max_{\Theta} P(\Theta | X)$$

$$= \arg\max_{\Theta} \frac{P(X | \Theta) P(\Theta)}{P(X)}$$

$$= \arg\max_{\Theta} \log P(X | \Theta) + \log P(\Theta).$$

This cannot be solved analytically due to the presence of latent variables $\vec{v}$ (existence a binding sites), $\vec{u}$ (positions of binding sites) and $\vec{s}$ (strand orientation of binding sites). However, the maximum of the a posteriori density of the model can be approximated by a modified EM algorithm [9].

In this work, we do not update the parameters $\Theta_f$ of the flanking model during the modified EM algorithm, but estimate them in advance from the entire data set and keep them fixed during the iteration procedure. As a consequence, we consider $\Theta_f$ to be given
for the rest of the supplementary material, and denote from now on the tupel \((\Theta_m, \Theta_c)\) as \(\Theta\).

The modified EM algorithm updates iteratively weighted estimates \(\gamma^{(t)}\) for the latent variables based on parameters \(\Theta^{(t)}\) (E-step) and parameters \(\Theta^{(t+1)}\) based on weighted estimates of latent variables \(\gamma^{(t)}\) (M-step).

### 4.1 E-step

In the E(xpectation)-step, we estimate – for each sequence – the probability of each assignment of latent variables, given the current parameters.

\[
\gamma^{(t)}_{iklm} = \frac{P(u_i = k, v_i = l, s_i = m| \vec{X}_i, \Theta^{(t)})}{\sum_{u_i} \sum_{v_i} \sum_{s_i} P(u_i, v_i, s_i, \vec{X}_i|\Theta^{(t)})} = \frac{P(\vec{X}_i|u_i = k, v_i = l, s_i = m, \Theta^{(t)})P(u_i = k|\nu^{(t)})P(v_i = l)P(s_i = m|\sigma^{(t)})}{\sum_{u_i} \sum_{v_i} \sum_{s_i} P(\vec{X}_i|u_i, v_i, s_i, \Theta^{(t)})P(u_i|\nu^{(t)})P(v_i)P(s_i|\sigma^{(t)})}
\]

### 4.2 M-step

In the M(aximization)-step, we compute the parameters given the weights of the latent variables by utilizing the Q-function, defined by

\[
Q(\Theta, \Theta^{(t)}) = \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \sum_{s_i} \gamma^{(t)}_{iu_i\nu_i\sigma_i} \log P(\vec{X}_i, u_i, v_i, s_i|\Theta) \quad (9)
\]

We obtain the next set of parameters \(\Theta^{(t+1)}\) as set of parameters that maximize the Q-function plus logarithm the parameter prior.

\[
\Theta^{(t+1)} = \arg\max_{\Theta} \left( Q(\Theta, \Theta^{(t)}) + \log P(\Theta) \right)
\]

\[
= \arg\max_{\Theta} \left( \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \sum_{s_i} \gamma^{(t)}_{iu_i\nu_i\sigma_i} \log P(\vec{X}_i, u_i, v_i, s_i|\Theta) + \log P(\Theta) \right)
\]

\[
= \arg\max_{\Theta} F(\Theta)
\]
while $\mathcal{F}(\Theta)$ is fully written as

\[
\mathcal{F}(\Theta) = \log P(\Theta_m) + \log P(\nu) + \log P(\sigma)
\]

\[
+ \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \sum_{s_i} \gamma_{iu_iv_is}^{(t)} \log P_f(X_i|\Theta_f)^{\delta_{u_i,0}}
\]

\[
+ \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \sum_{s_i} \gamma_{iu_iv_is}^{(t)} \log P_f(X_i,1,\ldots,X_{i,v_i-1}|\Theta_f)^{\delta_{u_i,1}}
\]

\[
+ \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \sum_{s_i} \gamma_{iu_iv_is}^{(t)} \log P_m(X_i,v_i,\ldots,X_{i,v_i+W-1}|\Theta_m)^{\delta_{u_i,1}\delta_{s_i,F}}
\]

\[
+ \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \sum_{s_i} \gamma_{iu_iv_is}^{(t)} \log P_m(rc(X_i,v_i,\ldots,X_{i,v_i+W-1})|\Theta_m)^{\delta_{u_i,1}\delta_{s_i,R}}
\]

\[
+ \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \gamma_{iu_iv_is}^{(t)} \log(\nu^{\delta_{u_i,1}}(1-\nu)^{\delta_{u_i,0}})
\]

\[
+ \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \gamma_{iu_iv_is}^{(t)} \log(\sigma^{\delta_{s_i,F}}(1-\sigma)^{\delta_{s_i,R}})
\]

\[
+ \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \gamma_{iu_iv_is}^{(t)} \log \frac{1}{L_i - W + 1},
\]

Cancelling all terms that do not pertain $\Theta$, and simplifying the remaining terms, we obtain

\[
\mathcal{F}(\Theta) = \log P(\Theta_m) + \log P(\nu) + \log P(\sigma)
\]

\[
+ \sum_{i=1}^{N} \sum_{v_i} \gamma_{iu_i}^{(t)} \log P_m(X_i,v_i,\ldots,X_{i,v_i+W-1}|\Theta_m)
\]

\[
+ \sum_{i=1}^{N} \sum_{v_i} \gamma_{iu_i}^{(t)} \log P_m(rc(X_i,v_i,\ldots,X_{i,v_i+W-1})|\Theta_m)
\]

\[
+ \sum_{i=1}^{N} \sum_{v_i} \gamma_{iu_i}^{(t)} \log \nu + \sum_{i=1}^{N} \sum_{v_i} \gamma_{iu_i}^{(t)} \log(1-\nu)
\]

\[
+ \sum_{i=1}^{N} \sum_{v_i} \gamma_{iu_i}^{(t)} \log \sigma + \sum_{i=1}^{N} \sum_{v_i} \gamma_{iu_i}^{(t)} \log(1-\sigma).
\]
Maximizing $F(\Theta)$ w.r.t. $\Theta$ yields

$$
\nu^{(t+1)} = \arg\max_{\nu} \log P(\nu) + \sum_{i=1}^{N} \sum_{v_i} \sum_{s_i} \gamma_{i,v_i,s_i}^{(t)} \log \nu + \gamma_{i0v_i,s_i}^{(t)} \log(1 - \nu)
$$

$$
= \arg\max_{\nu} (\gamma_{1.}^{(t)} + a_1) \log \nu + (\gamma_{0.}^{(t)} + a_2) \log(1 - \nu)
$$

$$
= \frac{\gamma_{1.}^{(t)} + a_1}{\gamma_{0.}^{(t)} + a_2},
$$

$$
\sigma^{(t+1)} = \arg\max_{\sigma} \log P(\sigma) + \sum_{i=1}^{N} \sum_{u_i} \sum_{v_i} \gamma_{i,u_i,v_i}^{(t)} F \log \sigma + \gamma_{i,v_i,R}^{(t)} \log(1 - \sigma)
$$

$$
= \arg\max_{\sigma} (\gamma_{..F}^{(t)} + b_1) \log \sigma + (\gamma_{..R}^{(t)} + b_2) \log(1 - \sigma)
$$

$$
= \frac{\gamma_{..F}^{(t)} + b_1}{\gamma_{..R}^{(t)} + b_2},
$$

and

$$
\Theta_m^{(t+1)} = \arg\max_{\Theta_m} \left( \log P(\Theta_m) + \sum_{i=1}^{N} \sum_{v_i=1}^{L_i-W+1} \right.
$$

$$
\left. + \gamma_{i1v_i,F}^{(t)} \log P_m(X_{i,v_i}, \ldots, X_{i,v_i+W-1} | \Theta_m) \right.
$$

$$
\left. + \gamma_{i1v_i,R}^{(t)} \log P_m(rc(X_{i,v_i}, \ldots, X_{i,v_i+W-1}) | \Theta_m) \right)
$$

Let $S = \{(i, v, s) | i \in \{1, \ldots, N\}, v \in \{1, \ldots, L_i - W + 1\}, s \in \{F, R\}\}$ denote the space of all motif occurrences of length $W$ in the data and let $\hat{X}_j$ denote the sequence corresponding to the motif occurrence $j \in S$, we rewrite the previous formula as

$$
\Theta_m^{(t+1)} = \arg\max_{\Theta_m} \left( \log P(\Theta_m) + \sum_{j=1}^{S} \gamma_j \log P_m(\hat{Y}_j | \Theta_m) \right).
$$

We observe this estimator to be identical to that of a single component of a mixture model, permitting to use the optimization algorithm of Gohr et al. [10].

### 4.3 Practical issues

The modified EM algorithm only guarantees to monotonically increase the posterior value in each iteration step, so it can miss the global optimum since the posterior landscape is often not convex. In order to cope with this problem, we start the algorithm multiple times and use the result with the largest posterior. In our studies, we choose ten starts of the EM algorithm and run each EM instance until the log-posterior difference between two subsequent iterations is smaller than $10^{-6}$. 

7
References


