Supporting Text

1 Notations and the Problem Setting

Our strategy is to first use existing *de novo* motif finding algorithms and TF databases to compose a list of putative binding motifs, $\mathcal{D} = \{W_1, \dots, W_D\}$, where D is in the range of 50 to 100, and then simultaneously modify these motifs and estimate the posterior probability for each of them to be included in the CRM through a Monte Carlo method. Let $\boldsymbol{\mathcal{S}}$ denote the set of n sequences, with lengths $L_1, L_2, \dots, L_n$, respectively, corresponding to the upstream regions of n coregulated genes. We assume that the CRM consists of K TFs with the corresponding PSWMs. Both the PSWMs and K are unknown and need to be inferred from the data. The jth TFBS in the ith sequence is denoted as $A_{i,j}$, and $\boldsymbol{A}$ is the collection of these set locations. Associated with each site is its *type* indicator $T_{i,j}$, with $T_{i,j}$ taking one of the K values. We model the dependence between $T_{i,j}$ and $T_{i,j+1}$ by a $K \times K$ probability transition matrix V. The distance between neighboring TFBSs in a CRM, $d_{ij} = A_{i,j+1} - A_{i,j}$, is assumed to follow $Q(\ ;\lambda)$, a truncated geometric distribution, and the distribution of nucleotides in the *background* sequence a multinomial distribution with unknown parameter $\boldsymbol{\rho} = (\rho_A, \dots, \rho_T)$. (In our applications we have taken l=0 for the l^{th} order Markov chain, though a higher order Markov chain could be used.)

Next, we let $\boldsymbol{u}$ be a binary vector indicating which motifs are included in the module, i.e. $\boldsymbol{u} = (u_1, \dots u_D)^T$, where

$$u_j = \begin{cases} 1, & \text{if the } j^{\text{th}} \text{ motif type is present in the module,} \\ 0, & \text{otherwise.} \end{cases}$$

By construction, $|\boldsymbol{u}| = K$. Thus, the information regarding K is completely encoded by $\boldsymbol{u}$. In light of these new notations, the set of PSWMs for the CRM is redefined as $\boldsymbol{\Theta} = \{W_j : \ u_j = 1\}$. Since now we restrict our inference of CRM to a subset of $\mathcal{D}$, the probability model for the observed sequence data, Eq. 1 in the main text, needs to be rewritten as:

$$P(\boldsymbol{\mathcal{S}} \mid \mathcal{D}, V, \boldsymbol{u}, \lambda, \boldsymbol{\rho}) = \sum_{\boldsymbol{A}} \sum_{\boldsymbol{T}} P(\boldsymbol{\mathcal{S}} \mid \boldsymbol{A}, \boldsymbol{T}, \mathcal{D}, V, \boldsymbol{u}, \lambda, \boldsymbol{\rho}) P(\boldsymbol{A} \mid \lambda) P(\boldsymbol{T} \mid \boldsymbol{A}, V), \tag{1.1}$$

From the above likelihood formulation, we need to simultaneously estimate the optimal $\boldsymbol{u}$ and the parameters $\mathcal{D}$, V, λ, and $\boldsymbol{\rho}$. To achieve this end, we adopt the Bayesian method by first giving a prior distribution on the parameters:

$$P(\mathcal{D}, V, \boldsymbol{u}, \lambda, \boldsymbol{\rho}) = f_1(\mathcal{D} \mid \boldsymbol{u}) f_2(V \mid \boldsymbol{u}) f_3(\boldsymbol{\rho}) g_1(\boldsymbol{u}) g_2(\lambda),$$

where the $f_i(\cdot)$'s are (product) Dirichlet distributions; $g_1(\boldsymbol{u})$ represents a product of D Bernoulli(p_0) distributions (p_0 is the prior probability of including any motif in the CRM); and $g_2(\lambda)$ a generally flat Beta distribution. The posterior distribution of the parameters then has the form

$$P(\mathcal{D}, V, \boldsymbol{u}, \lambda, \boldsymbol{\rho} \mid \boldsymbol{\mathcal{S}}) \propto P(\boldsymbol{\mathcal{S}} \mid \mathcal{D}, \boldsymbol{u}, V, \lambda, \boldsymbol{\rho}) f_1(\mathcal{D} \mid \boldsymbol{u}) f_2(V \mid \boldsymbol{u}) f_3(\boldsymbol{\rho}) g_1(\boldsymbol{u}) g_2(\lambda). \tag{1.2}$$

One can use the general Markov chain Monte Carlo (MCMC) strategy [1] to make inference from (1.2). But, given the flexibility of the model and the size of the parameter space, it is unlikely that a standard MCMC approach can converge to a good solution in a reasonable amount of time. In fact, both the Gibbs module sampler [2] and CISMODULE [3] used a direct MCMC method to infer the CRM from a posterior distribution simpler than (1.2) with fixed $\boldsymbol{u}$.

2 The EMCMODULE Procedure

With a starting set of putative binding motifs $\mathcal{D}$, we simultaneously modify these motifs and estimate the posterior probability for each of them to be included in the CRM through iterations of the following Monte Carlo sampling steps: (i) Given the current collection of motif PSWMs (or sites), sample motifs into the CRM by evolutionary Monte Carlo (EMC); (ii) Given the CRM configuration and the PSWMs, update the motif site locations; and (iii) Given motif site locations, update the corresponding PSWMs and other parameters. Each of these steps is described in detail in the following subsections.

2.1 Evolutionary Monte Carlo for module selection

It has been demonstrated that the EMC is effective for sampling and optimization with functions of binary variables [4]. Conceptually, we should be able to apply the EMC method directly to select motifs comprising the CRM, but a complication here is that there are many continuous parameters such as the W_j's, λ, and V. We cannot just fix these continuous parameters (as in the usual Gibbs sampler) and update the CRM composition because some of them vary in dimensionality when a putative motif in $\mathcal{D}$ is included or excluded from the CRM. We therefore have to integrate out the continuous parameters $\boldsymbol{\Theta}$ and V analytically and condition on variables $\boldsymbol{A}$ and $\boldsymbol{T}$ when updating the CRM composition.

Let $\boldsymbol{u}$ be defined as in the previous section. Since each u_i has a prior probability of p_0 to be 1, we can compute analytically the marginalized conditional posterior probability for a module configuration $\boldsymbol{u}$:

$$P(\boldsymbol{u} \mid \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{\mathcal{S}}) \propto p_0^{|\boldsymbol{u}|}(1-p_0)^{D-|\boldsymbol{u}|} \int P(S \mid \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\rho}, V, \lambda) P(\boldsymbol{\Theta}, V \mid \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{u}) P(\boldsymbol{\rho}) P(\lambda) d\boldsymbol{\Theta} d\boldsymbol{\rho} dV d\lambda, \tag{2.3}$$

where both $\boldsymbol{\Theta}$ and V are dependent on $\boldsymbol{u}$; and $\boldsymbol{A}$ and $\boldsymbol{T}$ are the sets of locations and types, respectively, of all putative motif sites (for all the D motifs in $\mathcal{D}$). Thus, only when the indicator u_i for the weight matrix W_i is 1, do its site locations and types contribute to the computation of (2.3). When we modify the current $\boldsymbol{u}$ by excluding a motif type, its site locations and corresponding motif type indicators are removed from the computation of (2.3).

To conduct EMC, we need to prescribe a set of temperatures, $t_1 > t_2 > \cdots > t_M = 1$, one for each member in the population. Then, we define

$$\pi_i(\boldsymbol{u}_i) \propto \exp[\log P(\boldsymbol{u}_i \mid \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{\mathcal{S}})/t_i],$$

and let $\pi(\boldsymbol{U}) \propto \prod_{i=1}^{M} \pi_i(u_i)$. The population $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$ is then updated iteratively using two types of moves: *mutation* and *crossover*.

In the mutation operation, a unit $\boldsymbol{u}_k$ is randomly selected from the current population and mutated to a new vector $\boldsymbol{v}_k$ by changing the values of some of its bits chosen at random. The new member $\boldsymbol{v}_k$ is accepted into the population with probability $\min(1, r_m)$, where

$$r_m = \pi_k(\boldsymbol{v}_k)/\pi_k(\boldsymbol{u}_k). \tag{2.4}$$

In the crossover step, two individuals, $\boldsymbol{u}_j$ and $\boldsymbol{u}_k$, say, are chosen at random from the population. Then, a crossover point x is chosen randomly over the positions 1 to D, and two new units $\boldsymbol{v}_j$ and $\boldsymbol{v}_k$ are formed by switching between the two individuals the segments on the right side of the crossover point. The two "children" are accepted into the population to replace their parents $\boldsymbol{u}_j$ and $\boldsymbol{u}_k$ with probability $\min(1, r_c)$, where

$$r_c = \frac{\pi_j(\boldsymbol{v}_j)\pi_k(\boldsymbol{v}_k)}{\pi_j(\boldsymbol{u}_j)\pi_k(\boldsymbol{u}_k)}. \tag{2.5}$$

If rejected, the two parents are kept unchanged. After convergence, the samples of $\boldsymbol{u}_M$ (corresponding to temperature $t_M = 1$) follow the target distribution (2.3).

2.2 Dynamic programming method for sampling motif sites

The second part of the algorithm consists of updating the motif sites conditional on a CRM configuration (i.e., with $\boldsymbol{u}$ fixed). For simplicity, we describe the method for a single sequence $S = (s_1, \ldots, s_L)$– the same procedure is repeated for all sequences in the data set. Let $\Omega = (\boldsymbol{\Theta}, \boldsymbol{\rho}, V, \lambda)$ denote the set of all parameters in the model, for a fixed $\boldsymbol{u}$. For the simplicity of notation, we assume that all motifs are of width w. Let $g(i, j, k, \boldsymbol{u}) = P(s_{[i,j,k]} \mid \Omega, \boldsymbol{u})$ denote the probability of observing the part of the sequence S from position i to j, with a motif of type k $\{k \in \mathcal{D} : u_k = 1\}$ occupying positions from $j - w + 1$ to j ($k = 0$ denotes the background). Let $Q(\ ;\lambda)$ denote the geometric(λ) distribution truncated at w, i.e. $Q(d; \lambda) = (1-\lambda)^{d-w}\lambda \quad (d = w, w+1, \ldots)$. Let $K = \sum_{k=1}^{D} u_k$ denote the number of motif types in the module. For notational simplicity, let us assume that $\boldsymbol{u}$ represents the set of the first K motifs, indexed 1 through K. Also, since the motif site updating step is *conditional* given $\boldsymbol{u}$, here we drop the subscript $\boldsymbol{u}$ from $g(i, j, k, \boldsymbol{u})$ in the remaining part of the section.

In the *forward summation* step, we recursively calculate the probability of different motif types ending at a position j of the sequence:

$$g(1, j, k) = \left[\sum_{i<j} \sum_{l=1}^{K} g(1, i, l)\ V_{l,k}\ Q(j - i - w; \lambda)\ + P(s_{[1,j-w,0]}|\boldsymbol{\rho})\right] g(j - w + 1, j, k). \tag{2.6}$$

By convention, the initial conditions are: $g(0,0,k) = 1$ for $k = 0, 1, \ldots, K$, and $g(i,j,k) = 0$ for $j < i$ and $k > 0$.

In the *backward sampling* step, we use the Bayes theorem to calculate the probability of motif occurrence at each position of a sequence, starting from the end of the sequence. Given that a motif of type k ends at position i in the sequence, the probability that the next motif further ahead in the sequence spans position $(i' - w + 1)$ to i', (where $i' \leq i - w$), and is of type k', is given by:

$$\frac{g(1,i',k')\ P(s_{[i'+1,i-w,0]}|\boldsymbol{\rho})\ g(i-w+1,i,k)\ Q(i-i'-w;\lambda)\ V_{k',k}}{g(1,i,k)}. \tag{2.7}$$

All the required expressions in (2.7) have already been calculated during the calculation (2.6).

2.3 Sampling parameters from posterior distributions

Given the motif type indicator $\boldsymbol{u}$ and the motif position and type vectors $\boldsymbol{A}$ and $\boldsymbol{T}$, we now update the parameters $\Omega = (\boldsymbol{\Theta}, \boldsymbol{\rho}, V, \lambda)$ by a random sample from its conditional distribution. Since conjugate priors (Dirichlet and product Dirichlet distributions) have been assumed for these parameters, their conditional posterior distributions are also of the same form and are straightforward to simulate from.

More precisely, we assume *a priori* that $W_i \sim \prod_{j=1}^{w} \text{Dirichlet}(\boldsymbol{\beta}_{ij})$ (for $i = 1, \ldots, D$); $\boldsymbol{\rho} \sim \text{Dirichlet}(\boldsymbol{\beta}_0)$; $\lambda \sim \text{Beta}(a,b)$, and let the total number of sequences be n. The posterior distributions of these parameters, conditional on $\boldsymbol{u}$, $\boldsymbol{A}$, and $\boldsymbol{T}$, will be also of the same type. For example, the posterior of W_i will be $\prod_{j=1}^{w} \text{Dirichlet}(\boldsymbol{\beta}_{ij} + \boldsymbol{n}_{ij})$, where $\boldsymbol{n}_{ij}$ is a vector containing the counts of the 4 nucleotides at the jth position of all the sites corresponding to motif type i. For those motifs that have not been selected by the module (i.e., motif types with their $u_i = 0$), the corresponding W's still follow their prior distribution. Similarly, the posterior distribution of $\boldsymbol{\rho}$ is $\text{Dirichlet}(\boldsymbol{\beta}_0 + \boldsymbol{n}_0)$, where $\boldsymbol{n}_0$ denotes the frequencies for the 4 nucleotides in the *background* sequence.

Given $\boldsymbol{u}$ (with $|\boldsymbol{u}| = K$), each row of V is assumed to follow an independent Dirichlet. Let the i^{th} row $v_i|\boldsymbol{u} \sim \text{Dirichlet}(\boldsymbol{\alpha}_i)$, where $i = 1, \ldots, K$. For updating V, we note that if $\boldsymbol{m}_{ij}$ $\{i, j \in \mathcal{D} : u_i = u_j = 1\}$ denotes the number of transitions from PSWM types i to j (when i and j have both been included in the module), then the posterior distribution of $\boldsymbol{v}_i$ is $\text{Dirichlet}(\boldsymbol{\alpha}_i + \boldsymbol{m}_i)$. Finally, we denote the distance between consecutive sites on sequence i ($i = 1, \ldots, n$) as $d_{ij} = A_{i,j+1} - A_{ij}$, and assume that each d follows $Q(\ ;\lambda)$, a geometric(λ) distribution truncated at w (as defined in section 2.2). Let $\boldsymbol{d} = \sum_{i=1}^{n} \sum_{j=1}^{|\boldsymbol{A}_i|-1} d_{ij}$ be the total length of sequence covered by the CRMs, where $|\boldsymbol{A}_i|$ is the total number of sites in sequence i, and $|\boldsymbol{A}'| = \sum_{i=1}^{n}(|\boldsymbol{A}_i| - 1)$. Then, the posterior distribution of λ is $\text{Beta}(a + |\boldsymbol{A}'|, b + \boldsymbol{d} - w|\boldsymbol{A}'|)$.

1. Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing.* (Springer-Verlag).

2. Thompson, W, Palumbo, M. J, Wasserman, W. W, Liu, J. S, & Lawrence, C. E. (2004) *Genome Research* **10**, 1967–74.

3. Zhou, Q & Wong, W. H. (2004) *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12114–9.

4. Liang, F & Wong, W. H. (2000) *Statistica Sinica* **10**, 317–342.

5. Berman, B, Nibu, Y, Pfeiffer, B, Tomancak, P, Celniker, S, Levine, M, Rubin, G, & Eisen, M. (2002) *Proc. Natl Acad. Sci. USA* **99**, 757–762.