Reconstruction of Novel Transcription Factor Regulons Through Inference of Their Binding Sites

Supplementary Material

Abdulkadir Elmas¹, Xiaodong Wang¹* and Michael S Samoilov²

Email: Abdulkadir Elmas – elmas@ee.columbia.edu; Xiaodong Wang – wangx@ee.columbia.edu; Michael S Samoilov – mssamoilov@lbl.gov

¹Correspondence: wangx@ee.columbia.edu
¹Department of Electrical Engineering, Columbia University, 500W 120th Street, 10027 New York, USA
Full list of author information is available at the end of the article
²Department of Bioengineering, QB3 California Institute for Quantitative Biosciences UC Berkeley, USA.

1 System model

1.1 High-confidence co-regulated gene set estimation via sparse biclustering

Given the genome-wide expression profiles we want to estimate a group of genes that are putatively co-regulated with the coding gene of a TF of interest. Co-regulation of a pair of genes can be quantified by the expression coherency across the data points (conditions).

Let $M \in \mathbb{R}^{n \times p}$ be the gene expression data consisting of $n$ genes and $p$ conditions (features). Let $m_i$ denote the $i$-th row of $M$. For a given subset of features ($\subseteq \{1, ..., p\}$), the observation pair $(m_i, m_j)$ can be represented as a beam around the linear regression line with slope $a_{ij}$ and intercept $b_{ij}$ (e.g., Fig. 1). Then the feature selection problem is to select a subset of features that minimizes the sum of residuals from the best regression.

![Figure 1](image-url) Scatter plot of the features corresponding to expression rates of genes lexA and dinI. The data is obtained from [1].
For the bicluster generated for the $i$-th observation (e.g., a TF’s coding gene), feature selection can be represented by a vector, $s_i \in \{0, 1\}^p$, such that $s_{ik} = 1$ if the feature $k$ is selected, and 0 otherwise. To avoid trivial solutions –not selecting any feature– a sparse penalty term is introduced to the problem, i.e., $\beta_1 \| 1 - s_i \|_1$, where $\beta_1$ is the regularizer coefficient and $1$ is the vector of 1s. Therefore, the feature selection problem for the $i$-th gene can be represented as,

$$
\min_{s_i, a_{ij}, b_{ij}} \sum_{k=1}^{p} s_{ik} (m_{ik} - a_{ij} m_{jk} - b_{ij})^2 + \beta_1 \| 1 - s_i \|_1 ,
$$

(1)

where $s_{ik} \in \{0, 1\}$.

To find coherent observations under the conditions given by $s_i$ a gene selection vector $w_i \in \{0, 1\}^n$ is introduced, such that $w_{ij} = 1$ if there is a strong linear coherence between the observation pair $(m_i, m_j)$, and $w_{ij} = 0$ otherwise. Similarly, another sparse regularizer is added to penalize trivial biclusters, i.e., $\beta_2 \| 1 - w_i \|_1$.

To favor co-up-regulated and co-down-regulated observation samples, an importance matrix $D \in \mathbb{R}^{n \times n \times p}$ is introduced, where $d_{ijk} \in [0, 1]$ indicates the importance of the $k$-th feature for the observation pair $(m_i, m_j)$ which is computed as a function of the euclidean distance of the $k$-th feature to the central point $(\bar{m}_i, \bar{m}_j)$ in the scatter plot. Relaxing $w_{ij} \in \{0, 1\}$ to $w_{ij} \in [0, 1]$ and $s_{ik} \in \{0, 1\}$ to $s_{ik} \in [0, 1]$, the gene expression biclustering problem for the $i$-th gene can then be cast as,

$$
\min_{w_i, s_i, a_{ij}, b_{ij}} \sum_j w_{ij} \sum_k s_{ik} \frac{1}{d_{ijk}} (m_{ik} - a_{ij} m_{jk} - b_{ij})^2 + \beta_1 \| 1 - s_i \|_1 + \beta_2 \| 1 - w_i \|_1 ,
$$

(2)

where $w_{ij} \in [0, 1]$, $s_{ik} \in [0, 1]$ [2].

By optimizing (2) for a given regulatory gene $i$ and expression matrix $M$, one can obtain a bicluster where the genes within the bicluster exhibit high linear-coherency with each other. One may expect that the high coherency in gene expression could indicate a common transcriptional activity. However, this correlation may occur either by chance (e.g. resulting from experimental noise) or because of indirect effects underlying the experiments, thereby integration of additional biological evidence into the problem can be crucial. In particular, we can employ the fitness data to check whether the knockout strains of the bicluster genes can show similar responses (fitness) to a particular stress, i.e., we may assume that the co-regulated genes – when knocked out– are more likely to cause similar growth rates for the organism. Following this intuition, the algorithm looks to find the knock outs that have similar growth patterns under a subset of stress conditions.

Let $F \in \mathbb{R}^{n \times q}$ be the fitness data consisting of $n$ genes and $q$ conditions. For the observation pair $(f_i, f_j)$ in $F$ we determine the feature selection by

$$
\min_{\tilde{s}_i, a_{ij}, b_{ij}} \sum_{\ell=1}^{q} \tilde{s}_{i\ell} (f_{i\ell} - a_{ij} f_{j\ell} - b_{ij})^2 + \tilde{\beta}_1 \| 1 - \tilde{s}_i \|_1 ,
$$
where $\hat{s}_{i\ell} \in \{0, 1\}$, and $\hat{\beta}_1$ control the feature selection in $F$. We optimize the observation selection such that

$$\sum_{\ell} \frac{\hat{s}_{i\ell}}{d_{i\ell}} (f_{i\ell} - a_{ij}f_{j\ell} - b_{ij})^2 + \hat{\beta}_1 \|1 - \hat{s}_i\|_1 + \hat{\beta}_2 \|1 - w_i\|_1,$$

where $\hat{\beta}_2$ controls the gene selection and $d_{i\ell}$ is the importance of the $\ell$-th feature for the observation $(f_i, f_j)$.

For a given regulatory gene $i$, gene expression matrix $M$, and fitness data matrix $F$, the biclustering problem can then be written as

$$\min_{w_{ij}, s_{ik}, a_{ij}, b_{ij}} \sum_j w_{ij} \left( \sum_k \frac{s_{ik}}{d_{ijk}} (m_{ik} - a_{ij}m_{jk} - b_{ij})^2 + \sum_{\ell} \frac{\hat{s}_{i\ell}}{d_{i\ell}} (f_{i\ell} - a_{ij}f_{j\ell} - b_{ij})^2 + \beta_1 \|1 - s_i\|_1 + \hat{\beta}_1 \|1 - \hat{s}_i\|_1 + (\beta_2 + \hat{\beta}_2) \|1 - w_i\|_1 \right), \quad (3)$$

where $w_{ij} \in [0, 1]$, $s_{ik} \in [0, 1]$, and $\hat{s}_{i\ell} \in [0, 1]$. Since (3) is not jointly convex in $W, S, \hat{S}, A$, and $B$, we propose an iterative procedure for solving it as follows.

Initialization: $A$ is initialized to $1 \cdot 1^T$ since the genes’ expression levels (rows) are normalized to $[0, 1]$. $B$ is normalized as the linear regression line for each observation $(m_i, m_j)$ passing through the central point $(\bar{m}_i, \bar{m}_j)$ with slope $a_{ij}$. $S$ and $\hat{S}$ are initialized as selecting the feature points whose importance distance to regression line remain within a certain threshold, i.e., $s_{ik} = 1$ if the $k$-th data point’s distance to the central point $(\bar{m}_i, \bar{m}_j)$ is larger than some predefined threshold. $W$ is computed from initial values of $S, \hat{S}, A$, and $B$. Each variable is then updated through optimizing its cost function when the other variables are fixed.

Updating $W$: When $S, \hat{S}, A$, and $B$ are fixed (3) becomes a convex linear objective of $W$, i.e.,

$$\min_{w_{ij}} w_{ij} \left( \sum_{k=1}^p \frac{s_{ik}}{d_{ijk}} (m_{ik} - a_{ij}m_{jk} - b_{ij})^2 + \sum_{\ell=1}^q \frac{\hat{s}_{i\ell}}{d_{i\ell}} (f_{i\ell} - a_{ij}f_{j\ell} - b_{ij})^2 - (\beta_2 + \hat{\beta}_2) \right) \quad \text{s.t.} \quad w_{ij} \in [0, 1], \quad (4)$$

which can be solved as $w_{ij} = 1$ if $\left( \sum_{k=1}^p \frac{s_{ik}}{d_{ijk}} (m_{ik} - a_{ij}m_{jk} - b_{ij})^2 + \sum_{\ell=1}^q \frac{\hat{s}_{i\ell}}{d_{i\ell}} (f_{i\ell} - a_{ij}f_{j\ell} - b_{ij})^2 \right) < (\beta_2 + \hat{\beta}_2)$, and $w_{ij} = 0$ otherwise.

Updating $S$: When $W, S, \hat{S}, A$, and $B$ are fixed, the objective function is a convex (linear) function of $S$, i.e.,

$$\min_{s_{ik}} s_{ik} \left( \sum_{j} \frac{w_{ij}}{d_{ijk}} (m_{ik} - a_{ij}m_{jk} - b_{ij})^2 - \beta_1 \right) \quad \text{s.t.} \quad s_{ik} \in [0, 1], \quad (5)$$
then $S$ can be updated in closed form such that $s_{ik} = 1$ if $\sum_j \frac{w_{ij}}{d_{ijk}} (m_{ik} - a_{ij} m_{jk} - b_{ij})^2 < \beta_1$, and $s_{ik} = 0$ otherwise.

**Updating $\hat{S}$:** Similarly, when $W, S, A$, and $B$ are fixed, the objective function is convex in $\hat{S}$, i.e.,

\[
\min_{\hat{s}_{i\ell}} \left( \sum_j \frac{w_{ij}}{d_{ij\ell}} (f_{i\ell} - a_{ij} f_{j\ell} - b_{ij})^2 - \hat{\beta}_1 \right)
\]

s.t. $\hat{s}_{i\ell} \in [0, 1],$

and $\hat{S}$ is updated in closed from such that $\hat{s}_{i\ell} = 1$ if $\sum_j \frac{w_{ij}}{d_{ij\ell}} (f_{i\ell} - a_{ij} f_{j\ell} - b_{ij})^2 < \hat{\beta}_1$, and $\hat{s}_{i\ell} = 0$ otherwise.

**Updating $A$ and $B$:** When $W, S$, and $\hat{S}$ are fixed, $A$ and $B$ can be solved as a least-squares linear regression problem.

The procedure iterates by updating $W, S, \hat{S}, A$, and $B$ until the objective function converges. After convergence, we check if the bicluster have sufficient members (intuitively between 5-50 genes) to discover a regulatory motif from their upstream sequences. Before motif discovery, a pre-processing may be necessary for refining those inputs. As one can expect, the obtained high-confidence set may contain false positive genes and the corresponding upstream sequences can thereby deteriorate the motif estimation. Even some true positive genes may lack the corresponding regulatory sites in their upstream sequences which makes the motif finding problem ill-defined. Next, we will demonstrate a filtering procedure to eliminate those potentially uninformative genes from the obtained set.

### 1.2 Filtering high-confidence gene set

We aim to eliminate genes that presumably lack the regulatory patterns in their upstream, which is generally observed in gene groups that belong to the same operon. If the reference knowledge about the TF is available (e.g., consensus sequence, motif, etc.) we could easily detect such genes by scanning their upstream regions. However, for novel genomes such prior information is usually not available. On the other hand, for most sequenced genomes the operon predictions data are available. In such data, the adjacently co-regulated genes are grouped into operons whereby the common regulatory sites are located in the upstream of the first-in-operon genes. Using this information, an intuitive elimination procedure could be first including only the first-in-operon genes and then sequentially filtering in the downstream genes as necessary. Our filtering procedure is outlined as follows.

Assume that the high-confidence set $\mathcal{H}$ consist of the genes appearing in three different operons named as $a, b, c$, e.g., $\mathcal{H} = \{a_1, a_2, a_3, b_1, b_2, c_1, c_2, c_3\}$, where $a_1$ is the first-in-operon gene in the operon $a$ and $a_2$ is the second-in-operon gene, and so forth for the other operons $b$ and $c$.

- Determine the first-in-operon genes $\mathcal{F} = \{a_1, b_1, c_1\}$ and the downstream genes $\mathcal{D} = \{a_2, a_3, b_2, c_2, c_3\}$ in the high-confidence set by using the operon predictions database [3]. Start the new high-confidence set with only the first-in-operon genes, e.g., $\mathcal{H} = \mathcal{F} = \{a_1, b_1, c_1\}$.

- For each operon in $\mathcal{D}$, calculate the difference $S$ in co-expression score between a pair of adjacent genes, where the co-expression score is defined as \("the
coherence of gene expressions (correlation) with the given TF gene”, e.g.,
\[ S(a_2) = \text{score}(a_1) - \text{score}(a_2), \]
where \( \text{score}(x) = \text{Corr}(x, \text{TFgene}) \).

- Add the downstream genes with score \( S < \text{threshold} \), and check if (i) there are \( \geq 10 \) genes in \( \mathcal{H} \) or (ii) 50% of the downstream genes are filtered in.
- If (i) or (ii) satisfies, return \( \mathcal{H} \). Otherwise, continue adding more downstream genes by incrementing the threshold within a given range then comparing the next set of gene pairs in the downstream (see Algorithm 1).

**Algorithm 1 Downstream gene filter**

```plaintext
Require: set \( \mathcal{H} = \emptyset \), and \( \mathcal{D}_0 = \mathcal{D} 
Require: set \( i = 2 \), and \( \text{threshold} = \min(\text{thresholdRange}) 
1: for every operon \( x \in \mathcal{D} \), e.g., \( x \in \{a, b, c\} \), do 
2: \quad \text{Calculate} \( S(x_i) = \text{score}(x_{i-1}) - \text{score}(x_i) \) 
3: \quad if \( S(x_i) < \text{threshold} \) then 
4: \quad \quad \mathcal{H} = \mathcal{H} \cup x_i 
5: \quad \mathcal{D} = \mathcal{D} \setminus x_i 
6: \quad end if 
7: end for 
8: if \( |\mathcal{H}| \geq 10 \) then 
9: \quad return \( \mathcal{H} \) 
10: else if \( |\mathcal{D}| \leq \frac{|\mathcal{D}_0|}{2} \) then 
11: \quad return \( \mathcal{H} \) 
12: else 
13: \quad \text{threshold} = \text{threshold} + \frac{0.1}{1\text{+|H_new|}} 
14: \quad if \text{threshold} \geq \max(\text{thresholdRange}) \text{ then} 
15: \quad \quad i = i + 1 
16: \quad \quad \text{threshold} = \min(\text{thresholdRange}) 
17: \quad end if 
18: \quad jump to 1 
19: end if
```

Setting a sensible \( \text{thresholdRange} \) is important, e.g., starting with a negative value may be intuitive. A negative \( S \) occurs when the gene’s correlation score is higher than that of a preceding gene, which indicates that the gene is likely directly-regulated by a TF and the regulatory pattern could be present in the corresponding gene’s upstream region.

### 1.3 Motif discovery with BAMBI2b

Genes regulated by the same TF usually share a common nucleotide pattern in their promoter regions where the TF specifically recognizes this pattern and binds. The locations of such TF binding sites and the properties of the corresponding motif can be computationally derived by motif discovery algorithms. Here, we employ the Bayesian Algorithm for Multiple Biological Instances of motif discovery –BAMBI [4]– to estimate the unknown motif for the putatively co-regulated genes obtained in the previous section.

BAMBI is designed to discover nucleotide motifs via estimating the number, length and locations of the binding instances in a collection of sequences, where each sequence may contain one, several or no instances of the motif. The algorithm is based on the sequential Monte Carlo (SMC) technique which processes one input sequence at a time efficiently in a sequential manner. The system is modeled as an HMM, where the hidden state at time \( t \) corresponds to the current input’s state vector \( \mathbf{x}_t \) composed by the number of motif instances \( n_t \) and their locations in the current sequence (observation) \( s_t \). As the iteration proceeds, given \( t \) sequences \( \mathcal{S}_t = \{s_1, \ldots, s_t\} \), BAMBI aims to identify the number of motif instances and their
locations for each sequence, i.e., $X_t = \{x_1, \ldots, x_t\}$. Since the distribution of the state vectors given observations $p(X_t|S_t)$ is not known and the samples are usually not available for this posterior, an approximation is implemented by taking weighted samples (particles), $X_t^k$, $k=1, \ldots, K$, from a trial density $q(X_t|S_t)$ within the sequential importance sampling (SIS) framework [5]. The sequential algorithm is obtained by setting $q(X_t^k|S_t) = q(X_t^k|S_{t-1})q(x_t^k|X_{t-1}^k, S_t)$, and computing the sample weights

$$w_t^k \propto w_{t-1}^k \frac{p(s_t|X_t^k, S_{t-1})p(x_t^k|X_{t-1}^k, S_{t-1})}{q(x_t^k|X_{t-1}^k, S_t)}.$$ (7)

A resampling procedure [6] is employed to reduce the increasing variance of ineffective samples with very small weights resulting from the phenomenon known as the degeneracy problem [7].

The length-$M$ motif is described as a $4 \times M$ PWM, i.e., $\theta = [\theta_1, \ldots, \theta_M]$, where $\theta_j$, $j = 1, \ldots, M$, represents the unknown nucleotide distributions (i.e., A, C, G, T/U) in each position of the motif that will be estimated. Since each promoter sequence can have one, several or no binding site (instance) of the motif their abundance also needs to be estimated. Assuming $N$ instances per sequence as the upper bound, the distribution of the number of instances is described by an unknown vector $\lambda = [\lambda_0 \ldots \lambda_N]$, where $\lambda_j$ is the proportion of sequences with $j$ instances of the motif.

TF binding sites have intrinsic properties that should be taken into account in the model, such as the location of conserved bases in the half-sites and their symmetry. We modified BAMBI to incorporate such properties of the TFBS. In particular, we can use the general two-block model [8] to describe the palindromic or direct/inverted repeat patterns in the half-sites. Estimating the location and size of the core segments will be introduced later under the “Two-block model” section. To model the motif symmetry we consider three types, i.e., (i) palindromic, (ii) inverted-repeat and (iii) direct-repeat symmetry. We represent the distribution of the symmetry of instances as an unknown vector, i.e., $\sigma = [\sigma_0, \sigma_1, \sigma_2]$, where $\sigma_0$ represents the proportion of instances with palindromic symmetry, and $\sigma_1$ ($\sigma_2$) represents the proportion of instances with inverted(direct)-repeat symmetry.

At each step $t$, given the first $t$ sequences $S_t$ the sequential algorithm estimates the number of instances in each sequence and their locations, $X_t$, within the Bayesian framework. The motif matrix $\theta$ is modeled as Dirichlet random vectors. The distribution of the number of instances $\lambda$ and the distribution of the symmetry of instances $\sigma$ are also modeled as Dirichlet random vectors. A sufficient statistic $\alpha^t$ is used to characterize the distribution of $\theta$, and it can be updated sequentially, i.e., $\alpha^t = \{\alpha^t_{m,j}\}_{m=1, \ldots, M}^{j=1, \ldots, 4}$, where $\alpha^t_{m,j} = [\alpha^t_{m,1}, \ldots, \alpha^t_{m,4}]^T$ is the parameters for the $m$-th Dirichlet distribution corresponding to the $m$-th column in PWM $\theta_m$. Then the parameters of $p(\theta|X_t, S_t)$ can be sequentially updated from $p(\theta|X_{t-1}, S_{t-1})$ as

$$\alpha^t_{m,j} = \alpha^{t-1}_{m,j} + n(a_{t,x_t}(m)),$$

where $a_{t,x_t}$ is the subsequence of $M$ letters in the sequence $s_t$ whose location is given by $x_t$. Similarly, a sufficient statistic is selected for the distribution of the number
of instances $\lambda$ which is Dirichlet distributed with parameters $\gamma^t = [\gamma_0^t \ldots \gamma_K^t]$. The sequential update is obtained by $\gamma^t = \gamma^{t-1} + j(x_t)$, where $j(x_t)$ is a vector of zeros except for a 1 indicating the number of instances of the motif in the $t$-th sequence. A sufficient statistic for the distribution of the symmetry of instances $\sigma$ is also defined. $\sigma$ is a Dirichlet distribution with parameters $\delta^t = [\delta_0^t \delta_1^t \delta_2^t]$ which can be updated sequentially by $\delta^t = \delta^{t-1} + u(a_{t,x_t})$, where

$$u(a_{t,x_t}) = [u_0 \ u_1 \ u_2] = \left[ \begin{array}{c} \sum_{m=1}^{M/2} I_{\{a_{t,x_t}(m)=\bar{a}_{t,x_t}(M-m)\}} \\ \sum_{m=1}^{M/2} I_{\{a_{t,x_t}(m)=a_{t,x_t}(M-m)\}} \\ \sum_{m=1}^{M/2} I_{\{a_{t,x_t}(m)=a_{t,x_t}(m+\frac{M}{2})\}} \end{array} \right].$$

$I_{\{\}}$ is the indicator function which takes the value 1 if a nucleotide pair observed in $a_{t,x_t}$ obeys the given symmetry type, palindromic, inverted-repeat, or direct-repeat, and the operator $\bar{\cdot}$ takes the Watson-Crick complement of the given letter.

The influence of these variables can be averaged out in the SIS procedure, whereby the measurement model depends on $\theta$ and $\sigma$, and the state transition depends on $\lambda$. The importance distribution can be chosen with $q(x_t|X_{t-1}^k,S_t) = p(x_t|X_{t-1}^k,S_t)$, where

$$p(x_t|X_{t-1}^k,S_t) \propto p(s_t|x_t,X_{t-1}^k,S_{t-1})p(x_t|X_{t-1}^k,S_{t-1})$$

$$= \int p(s_t|x_t,X_{t-1}^k,S_{t-1},\theta)p(\theta|x_t,X_{t-1}^k,S_{t-1})d\theta$$

$$\int p(x_t|X_{t-1}^k,S_{t-1},\lambda)p(\lambda|X_{t-1}^k,S_{t-1})d\lambda. \quad (8)$$

Then, for each particle, $k=1,\ldots,K$, the importance weight can be updated as

$$w_t^k \propto w_{t-1}^k \sum_{x_t} p(s_t|x_t,X_{t-1}^k,S_{t-1})p(x_t|X_{t-1}^k,S_{t-1}). \quad (9)$$

Given the PWM $\theta$, the background distribution $\theta_0$, the state $x_t$ at time $t$, and the symmetry distribution $\sigma$, we can express in closed-form the likelihood of the sequence $s_t$ in (8) as

$$\int p(s_t|x_t,X_{t-1}^k,S_{t-1},\theta)p(\theta|x_t,X_{t-1}^k,S_{t-1})d\theta$$

$$= \theta_0^n(a_{t,x_t}) \prod_{m=1}^{M} E_m^n(a_{t,x_t}(m)) p(a_{t,x_t}|\theta), \quad (10)$$

where $\theta^n = \prod_{j=1}^{J} \theta_j^{n_j}$, $n(a) = [n_1, \ldots, n_4]$ with $n_j, j=1,\ldots,4$, being the number of times the $j$-th letter appears in the instance $a$, and $a_{t,x_t}^c$ is the sequence in $s_t$ resulting from removing the instance $a_{t,x_t}$.
The last term in (10) is the likelihood of the instance $a_{t,x}$ given the PWM $\theta$, and the motif’s underlying symmetry $\sigma$. The influence of the symmetry can be averaged out by computing the integral $p(a_{t,x}|\theta) = \int p(a_{t,x}|\theta,\sigma)p(\sigma|\theta)d\sigma$. For a given instance $a_{t,i}$, as a measure of dissimilarity (asymmetry) we use the Kullback-Leibler (KL) divergence between the PWM’s base variations indicated by the symmetry type observed in $a_{t,i}$. Since the KL divergence is a non-symmetric distance measure we define a symmetrized similarity metric and write the closed-from expression for $p(a_{t,x}|\theta)$ as follows.

\[
p(a_{t,x}|\theta) = \int p(a_{t,x}|\theta,\sigma)p(\sigma|\theta)d\sigma = p(a_{t,x},\theta,s)\mathbb{E}[\sigma] \]

\[
= \frac{1}{1 + H(\theta^0_\alpha, \theta^0_\pi) + H(\theta^1_\alpha, \theta^1_\pi)} \sum_{s=0}^{2} \delta_s , \tag{11}
\]

where

\[
H(\theta^0_\alpha, \theta^0_\pi) = \frac{M/2}{4} \sum_{m=1}^{M/2} \sum_{j=1}^{4} -I\{a_{t,x}(m)=a_{t,x}(M-m)\}
\]

\[
\theta_m(j) \log(\theta_{M-m}(4-j)),
\]

\[
H(\theta^1_\alpha, \theta^1_\pi) = \frac{M/2}{4} \sum_{m=1}^{M/2} \sum_{j=1}^{4} -I\{a_{t,x}(m)=a_{t,x}(M-m)\}
\]

\[
\theta_m(j) \log(\theta_{M-m}(j)),
\]

\[
H(\theta^2_\alpha, \theta^2_\pi) = \frac{M/2}{4} \sum_{m=1}^{M/2} \sum_{j=1}^{4} -I\{a_{t,x}(m)=a_{t,x}(m+M/2)\}
\]

\[
\theta_m(j) \log(\theta_{m+M/2}(j)),
\]

and $s=0,1,2$ refers to the given symmetry type, i.e., palindromic, inverted-repeat, and direct-repeat, respectively. We normalize these cross entropy measures by the number of symmetric bases in $a_{t,x}$, i.e., $u_s(a_{t,x})$. By restricting the calculation to only block sites $B$, i.e., $m \in B \subset \{1, \ldots, M/2\}$, the density is then multiplied by $\frac{u_s(a_{t,x})}{1+|B|}$ and log-normalized to average out the influence of the block length $|B|$, which replaces (11) by

\[
p(a_{t,x}|\theta) = \log(\tau + \frac{u_s(a_{t,x})^2}{1+|B|}) \frac{\delta_s}{1 + H(\theta^0_\alpha, \theta^0_\pi) + H(\theta^1_\alpha, \theta^1_\pi)} \sum_{s=0}^{2} \delta_s , \tag{12}
\]

where $\tau$ is a constant to avoid taking log(0) and serves as a tolerance threshold for the asymmetric instances, i.e., the larger is $\tau$ the algorithm is more likely to sample an (asymmetric) instance that possess a larger cross-entropy. By using (12) the expectation term in (10) can be approximated as

\[
\mathbb{E}[\theta^m_{a_{t,x}(m)}] p(a_{t,x}|\theta) \approx \theta^m_{a_{t,x}(m)} \log(\tau + \frac{u_s(a_{t,x})^2}{1+|B|}) \frac{\delta_s}{1 + H(\theta^0_\alpha, \theta^0_\pi) + H(\theta^1_\alpha, \theta^1_\pi)} \sum_{s=0}^{2} \delta_s . \tag{13}
\]
The second integral in (8) is computed analogously.

\[
\int p(\lambda | X_{k,t-1}, S_{t-1}, \lambda) d\lambda = p(i_1 \ldots i_n | n) \pi_n = p(i_1 \ldots i_n | n) \frac{\gamma_{n-1}^{t-1}}{\sum_{i=0}^{N} \gamma_{i-1}^{t-1}},
\]  

(14)

where \(p(i_1 \ldots i_n | n)\) is distributed uniformly.

Given a set of promoter sequences \(S_t = \{s_1, \ldots, s_t\}\), the BAMBI2b algorithm for discovering symmetric two-block motifs is then summarized as follows.

For each sequence, and for each particle,

- Construct the importance distribution by computing (8,10,13,14) and enumerating all possible sample extensions

\[ X_{t}^k(n, i_1, \ldots, i_n) = [X_{t-1}^k(i_1 \ldots i_n)^T]; \]

- Draw sample \(x_{t}^k\) from the importance distribution and set

\[ X_{t}^k = [X_{t-1}^k(x_t^k)]; \]

and update the particle weight using (9);

- Update the sufficient statistics \(\alpha^t, \gamma^t, \delta^t\); then, resample if needed.

Two-block model The bacterial TF binding sites are usually symmetric in sequence where the half-sites obey a palindromic or direct/inverted repeat symmetry. The more informative bases in each half-site are generally conserved in the middle. One can model such motif structure as a pair of informative symmetric sequences (blocks \(B\)) separated by a run of uninformative bases (gap), and a pair of uninformative flanker sequences (\(F\)) located at each end (Fig. 2). To estimate this structure within the SMC framework we used the class-based resampling scheme presented in [6] and augmented the state vector as \(z_t = [x_t^T, m, b, f]^T\), where \(m, b, \) and \(f\) are the parameters for motif length \(M\), block length \(B\), and flanker length \(F\), respectively.

The classes are enumerated as \(\Lambda_{M,B,F} = \{(M, B, F) \mid B \in \{1, \ldots, \frac{M}{2}\}, F \in \{0, \ldots, \frac{M}{2} - B\}\}\), where each class represents a different scenario of the possible motif structures. The Bayesian algorithm for the two-block motif model is similar to the single-block model given in [8] except that the closed-form expressions are modified for the non-block region, i.e., background distribution \(\theta_0\) is used for the uninformative bases \(\{1, \ldots, M\} \setminus B\).

![Figure 2](image)

Figure 2 The two-block structure of a length-\(M\) motif described by the blocks (\(B\)) and the flankers (\(F\)).

The class-based resampling procedure is summarized as follows.
- Choose the number of particles for each class according to a multinomial distribution with parameters

\[ P_{m,b,f} \sim \{ \hat{P}(m,b,f|S_t) \}_{(m,b,f) \in \Lambda_{M,B,F}}. \]  

(15)

- For every class \( \forall \gamma \in \Lambda_{M,B,F} \), if \( P_{\gamma} < P_{\text{thresh}} \) then set \( P_{\gamma} = P_{\text{thresh}} \).
- Reduce the number particles starting from the class with the largest number particles until \( \sum_{\gamma} P_{\gamma} = P \).
- For every class draw \( P_{\gamma} \) new particles from the set of previous particles with probabilities proportional to their weights.
- Assign equal weights to each new sample within a class, \( w^k_t = \hat{P}(m,b,f|S_t)/P_{m,b,f} \).

**Optimal data feeding**

Because of the sequential nature of BAMBI2b, the order in which the sequences are utilized have a direct impact on the accuracy of estimation. An optimal data feeding should then be defined in order to prioritize the most useful sequences to supply to BAMBI2b. The downstream genes can be supplied in the same order in which they are filtered in by *Algorithm 1*.

To rank-order the first-in-operon genes \( \mathcal{F} \), we can use a similar approach. The genes that are expressed more coherently with the given transcription factor may indicate a proximity to the TF’s binding instances. Following this intuition, we feed the upstream sequences of the first-in-operon genes \( \{x \in \mathcal{F}\} \) in a decreasing order based on the correlation scores, i.e., \( \text{score}(x) = \text{Corr}(x, \text{TFgene}) \), as follows. Let \( \mathbf{M} \in \mathbb{R}^{p \times |\mathcal{F}|} \) be the gene expression matrix corresponding to \( |\mathcal{F}| \) high-confidence genes and \( p \) different conditions. Suppose that \( \mathbf{m}_i \in \mathbb{R}^p \) is the vector of expression rates for the given transcription factor gene located in the \( i \)-th row of \( \mathbf{M} \). The correlation scores for the genes \( x = 1, \ldots, |\mathcal{F}| \), are calculated by Pearson’s correlation

\[
\text{score}(x) = \frac{\sum_{k=1}^{p} (m_{ik} - \bar{m}_x)(m_{xk} - \bar{m}_x)}{\sqrt{\sum_{k=1}^{p} (m_{ik} - \bar{m}_x)^2 \sum_{k=1}^{p} (m_{xk} - \bar{m}_x)^2}},
\]

where \( \bar{m}_x \) and \( \bar{m}_i \) are empirical means of the expression vectors \( \mathbf{m}_x \) and \( \mathbf{m}_i \) of the genes \( i \) and \( x \), respectively. The upstream sequences are then arranged in the order of descending absolute scores \( |\text{score}(x)| \) of the corresponding genes in \( \mathcal{F} \).

**Author details**

1 Department of Electrical Engineering, Columbia University, 500W 120th Street, 10027 New York, USA.
2 Department of Bioengineering, QB3 California Institute for Quantitative Biosciences UC Berkeley, 1700 4th St #214, 94720 California, USA.

**References**

1. Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J., Gardner, J.: Large-scale mapping and validation of Escherichia Coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5 (2007)
2. Shi, Y., Liao, X., Zhang, X., Lin, G., Schuurmans, D.: Sparse learning based linear coherent bi-clustering. Algorithms in Bioinformatics 7534, 346–364 (2012)
3. Price, M., Huang, K., Alm, E., Arkin, A.: A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res 33, 880–892 (2005)
4. Jajamovich, G., Wang, X., Arkin, A., Samoilov, M.: Bayesian multiple-instance motif discovery with BAMBI: inference of recombinase and transcription factor binding sites. Nucleic Acids Res 39, 146 (2011)
5. Dong, B., Wang, X., Doucet, A.: A new class of soft MIMO demodulation algorithms. Signal Processing, IEEE Transactions on 51, 2752–2763 (2003)
6. Vercauteren, T., Guo, D., Wang, X.: Joint multiple target tracking and classification in collaborative sensor networks. Selected Areas in Communications, IEEE Journal on 23, 714–723 (2005)

7. Doucet, A., Wang, X.: Monte Carlo methods for signal processing: a review in the statistical signal processing context. Signal Processing Magazine, IEEE 22, 152–170 (2005)

8. Liang, K., Wang, X., Anastassiou, D.: A sequential Monte Carlo method for motif discovery. Signal Processing, IEEE Transactions on 56, 4496–4507 (2008)