A Sequential Monte Carlo Method for Motif Discovery

Kuo-ching Liang, Xiaodong Wang, *Fellow, IEEE*, and Dimitris Anastassiou, *Fellow, IEEE*

***Abstract*—We propose a sequential Monte Carlo (SMC)-based motif discovery algorithm that can efficiently detect motifs in datasets containing a large number of sequences. The statistical distribution of the motifs is modeled by an underlying position weight matrix (PWM), and both the PWM and the positions of the motifs within the sequences are estimated by the SMC algorithm. The proposed SMC motif discovery technique can locate motifs under a number of scenarios, including the single-block model, two-block model with unknown gap length, motifs of unknown lengths, motifs with unknown abundance, and sequences with multiple unique motifs. The accuracy of the SMC motif discovery algorithm is shown to be superior to that of the existing methods based on MCMC or EM algorithms. Furthermore, it is shown that the proposed method can be used to improve the results of existing motif discovery algorithms by using their results as the priors for the SMC algorithm.**

***Index Terms*—Genomic sequence, motif discovery, resampling, sequential Monte Carlo (SMC).**

I. Introduction

EFFORTS by various genomic projects have steadily expanded the pool of sequenced DNA data. Motifs, or DNA patterns found in different locations within the genome, are often of interest to biologists. By seeking out these similarities exhibited in sequences, we can further our knowledge on the functions and evolutions of these sequences.

Motif discovery algorithms can be broadly divided into three major categories: consensus sequence-based algorithms, projection-based algorithms, and profile-based algorithms. Consensus sequence-based algorithms include *WINNOWER* [1], and examples of projection-based algorithms include *Projection* in [2] and uniform projection motif finder (UPMF) in [3]. The third category of motif discovery algorithms, the profile-based algorithms, attempts to describe the instances of a motif collectively, by modeling their statistical behavior, namely, the distribution of the four nucleotides at the different locations within a motif. In profile-based algorithms, a position weight matrix (PWM) is used to model such statistical behavior. For motif of length w, the PWM is $4 \times w$ matrix where each column of the matrix is a vector of length 4, corresponding to the probability of observing each of the four nucleotide at the position. In general, the PWM is assumed to be an unknown parameter which is to be estimated by the algorithm together with the locations of the different instances of the motif in the sequences.

In [4], by treating the locations of the motifs in each sequence as missing information, an expectation-maximization (EM) algorithm is proposed to estimate and locate the motifs. In [5] and [6], MEME, an algorithm based on EM, is introduced with support for finding unknown number of motifs and unknown number of occurrences in the sequences. In [7]–[9], *Gibbs Motif Sampler* and *AlignACE* are proposed based on the Gibbs sampler, a Markov chain Monte Carlo (MCMC) algorithm, to estimate the PWM and the locations of the motifs in the sequences. Moreover, in [10], the Gibbs sampler-based *BioProspector* is proposed to treat the two-block motif model and palindromic patterns.

In this work, we take the profile-based approach and propose a solution based on the sequential Monte Carlo (SMC) algorithm to treat cases of two-block motif models, unknown number of motif instances, multiple motifs, and using the SMC algorithm to refine the result of other algorithms. In a follow-up work [11], we have proposed another profile-based deterministic tree-search method, the so-called deterministic sequential Monte Carlo (DSMC) algorithm, to discover motifs of unknown length, and motifs with insertion/deletion mutations.

The sequential Monte Carlo methodology is a family of statistical inference methods that are more powerful than the traditional MCMC techniques. It has been shown in [12] that the SMC methods provide an efficient alternative to Gibbs sampling in Dirichlet mixture models, which is the distribution of choice for the priors in our work. The SMC algorithm sequentially explores the data. However, the underlying problem does not necessarily have to be sequential in nature. Examples of SMC algorithms proposed for problems of this kind where Gibbs sampling solutions already exist can be found in the nonparametric Bayesian matrix factorization; in electropherogram basecalling algorithms; and in feature-based object recognition. In our work, we propose SMC algorithms that can handle single-block model, two-block model with unknown gap length, motifs of unknown length, motifs with unknown abundance, and sequences with multiple unique motifs. Furthermore, the SMC algorithm can also be used as a second-pass algorithm, using other algorithm's results as inputs, and further improve those estimates.

The remainder of the paper is organized as follows. In Section II, we present the system model for the motif discovery problem for the single block model. In Section III, we derive

Manuscript received February 14, 2007; revised May 19, 2008. First published May 28, 2008; last published August 13, 2008 (projected). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Elias S. Manolakos. This work was supported in part by the U.S. Office of Naval Research (ONR) under Grant N00014-03-1-0039.

K. Liang and X. Wang are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: kcliang@ee.columbia.edu; wangx@ee.columbia.edu; anastas@ee.columbia.edu).

D. Anastassiou is with the Center for Computational Biology and Bioinformatics (C2B2), Columbia University, New York, NY 10027 USA.

Digital Object Identifier 10.1109/TSP.2008.926194

1053-587X/$25.00 © 2008 IEEE

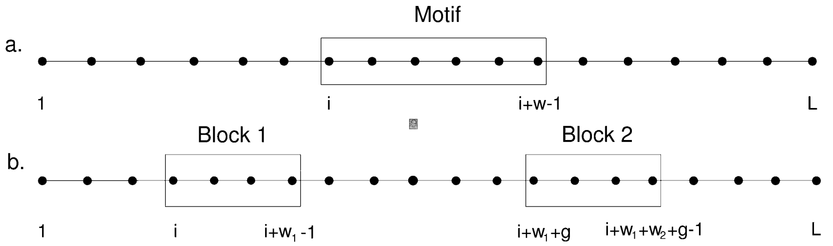


Fig. 1. Position weight matrix models. (a) Model for a single-block motif with motif length w. (b) Two-block motif of lengths w_1 and w_2, and gap length g.

the SMC motif discovery algorithm for the single block model. In Section IV, we introduce modifications to the single-block model algorithm to support other motif models. In Section V, we provide experimental results on both real and synthetic datasets. Section VI concludes the paper.

II. System Model

Let $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_T\}$, with $\boldsymbol{s}_t = [s_{t1}, \ldots, s_{tL}]$, be the set of DNA sequences of length L where we wish to find a common motif. Let us assume that a motif of length w is present in each one of the sequences. A single block motif model is shown in Fig. 1(a). The distribution of the motif is described by the $4 \times w$ position weight matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_w]$, where the column vector $\boldsymbol{\theta}_j = [\theta_{j1}, \ldots, \theta_{j4}]^T, j = 1, \ldots, w$, is the probability distribution of the nucleotides $\{A, C, G, T\}$ at the jth position of the PWM. The remaining nonmotif nucleotides are assumed to follow a Markovian distribution with probabilities given by $\boldsymbol{\Theta}_0$.

In our state-space model, the states represent the locations of the first nucleotides of the different occurrences of the motif in the sequence, whereas the observation for the state at step t is the entire nucleotide sequence, $\boldsymbol{s}_t$. Since the ending $w-1$ nucleotides in a sequence are not valid locations for the beginning of a motif with length w, at step $t, t = 1, \ldots, T$, the state, denoted as x_t, takes value from the set $\mathcal{X} = \{1, 2, \ldots, L_m\}$, where $L_m = L - w + 1$.

Let $\boldsymbol{a}_{t,x_t}$ be a sequence fragment of length w from $\boldsymbol{s}_t$ starting from position x_t in $\boldsymbol{s}_t$, and denote $\boldsymbol{a}^c_{t,x_t}$ as the remaining fragment from $\boldsymbol{s}_t$ with $\boldsymbol{a}_{t,x_t}$ removed. For example, for $\boldsymbol{s}_t = [AAAAGGGGAAAA]$ and $x_t = 5$ with $w = 4$, $\boldsymbol{a}_{t,x_t} = [GGGG]$ and $\boldsymbol{a}^c_{t,x_t} = [AAAAAAAA]$. Let us further define a vector $\boldsymbol{n}(\boldsymbol{a}) = [n_1, n_2, n_3, n_4]$ where $n_i, i = 1, \ldots, 4$, denotes the number of different nucleotides in the sequence fragment $\boldsymbol{a}$. Given the vectors $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_4]$ and $\boldsymbol{n} = [n_1, \ldots, n_4]$, we define

$$\boldsymbol{\theta}^{\boldsymbol{n}} \triangleq \prod_{j=1}^{4} \theta_j^{n_j}. \tag{1}$$

In DNA sequences, a nucleotide is often influenced by the surrounding nucleotides. Thus, we assume for our system model a third-order Markov model for the nonmotif nucleotides in the sequence. Let us denote P^3_{t,x_t} as the probability of $\boldsymbol{a}^c_{t,x_t}$. For example, if $\boldsymbol{a}^c_{t,x_t} = [ATAAG]$, the probability of $\boldsymbol{a}^c_{t,x_t}$ is given by

$$P^3_{t,x_t} = p(A)p(T|A)p(A|A,T)p(A|A,T,A)p(G|T,A,A). \tag{2}$$

In general, the zeroth to third-order Markov chain probabilities for the background nonmotif nucleotides can be averaged over a large genomic region, and are assumed to be known, which we denote as $\boldsymbol{\Theta}_0$. To perform motif discovery using the SMC algorithm, $\boldsymbol{\Theta}_0$ can be given as a known parameter by the user or default values can be used. Since the nucleotides being located in the motif are independent of the other motif nucleotides and nonmotif nucleotides, given the PWM $\boldsymbol{\Theta}$, the background distribution $\boldsymbol{\Theta}_0$, and the state at time t, the distribution of the observed sequence $\boldsymbol{s}_t$ is then given as follows:

$$p(\boldsymbol{s}_t | x_t = i, \boldsymbol{\Theta}) = P^3_{t,x_t} \prod_{k=1}^{w} \boldsymbol{\theta}_k^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))} \triangleq \mathcal{B}(\boldsymbol{s}_t; i, \boldsymbol{\Theta}) \tag{3}$$

where $\boldsymbol{a}_{t,i}(k)$ is the kth element of the sequence fragment $\boldsymbol{a}_{t,i}$, and $\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))$ is a 1×4 vector of zeros except at the position corresponding to the nucleotide $\boldsymbol{a}_{t,i}(k)$, where it is a one.

Inference Problem: From the discussion above, we formulate our inference problem as follows. Let us denote the state realizations up to time T as $\boldsymbol{x} \triangleq [x_1, x_2, \ldots, x_T]$ and similarly the sequences up to time T as $\boldsymbol{S} \triangleq [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_T]$, with the unknown parameter $\boldsymbol{\Theta}$, the position weight matrix. Given the sequences $\boldsymbol{S}$ and the Markovian nonmotif nucleotide distribution $\boldsymbol{\Theta}_0$, we wish to estimate the state realizations $\boldsymbol{x}$, which are the starting locations of the motif in each sequence, and the position weight matrix $\boldsymbol{\Theta}$, which describes the statistics of the motif. In the next section, we derive the SMC algorithm to solve this inference problem.

III. SMC Motif Discovery Algorithm

In this section, we first give a brief overview of the SMC methods. We then derive an SMC motif discovery algorithm for the case where each sequence in the dataset contains exactly one instance of the same motif. In Section IV, we will extend this algorithm to treat more general motif models, including two-block motifs with unknown gap length, motifs of unknown length, motifs with unknown abundance, and sequences with multiple unique motifs.

A. Sequential Monte Carlo Methods

Let us consider the following dynamic model

$$\text{initial state model: } p_{\boldsymbol{\theta}}(x_0) \tag{4}$$

$$\text{state transitions model: } p_{\boldsymbol{\theta}}(x_t | x_{t-1}) \quad \forall t \geq 1 \tag{5}$$

$$\text{measurement model: } p_{\boldsymbol{\theta}}(y_t | x_t) \quad \forall t \geq 1 \tag{6}$$

where x_t and y_t are the state and the observation at time t, respectively, and $p_{\boldsymbol{\theta}}(\cdot)$ are probability density functions depending on some known parameters $\boldsymbol{\theta}$. At time t, we want to make an online inference of the states $\boldsymbol{x}_t = (x_0, \ldots, x_t)$ based on the observation $\boldsymbol{y}_t = (y_0, \ldots, y_t)$. The optimal solution in terms of any common criterion depends only on the conditional pdf $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{y}_t)$. Often, direct computation of this conditional pdf is infeasible due to the complexity of the system; therefore, Monte Carlo methods are employed to estimate it. In most cases, however, drawing random samples directly from the conditional pdf $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{y}_t)$ is also infeasible. Hence, we employ the importance sampling technique to sample from some trial sampling density $q_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{y}_t)$ and properly weigh the samples according to the target distribution. Suppose K random samples $\{\boldsymbol{x}_t^{(k)}, k = 1, \ldots, K\}$ are drawn from $q_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{y}_t)$. The target conditional pdf can then be approximated by

$$\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{y}_t) = \frac{1}{W_t}\sum_{k=1}^{K} w_t^{(k)} \mathbb{I}\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(k)}\right),$$

$$\text{with } w_t^{(k)} = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(k)}|\boldsymbol{y}_t)}{q_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(k)}|\boldsymbol{y}_t)} \quad (7)$$

where $W_t = \sum_{k=1}^{K} w_t^{(k)}$ and $\mathbb{I}(\cdot)$ is the indicator function such that $\mathbb{I}(x) = 1$ for $x = 0$ and $\mathbb{I}(x) = 0$ otherwise. The set $\{(\boldsymbol{x}_t^{(k)}, w_t^{(k)}), k = 1, \ldots, K\}$ is called a set of properly weighted samples with respect to the target distribution [13]. Furthermore, it is possible at time t to generate the set $\{(\boldsymbol{x}_t^{(k)}, w_t^{(k)}), k = 1, \ldots, K\}$, properly weighted with respect to $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{y}_t)$, recursively from the previous set of properly weighted samples $\{(\boldsymbol{x}_{t-1}^{(k)}, w_{t-1}^{(k)}), k = 1, \ldots, K\}$, properly weighted with respect to $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{t-1})$. By choosing the optimal trial distribution $q_{\boldsymbol{\theta}}(x_t|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t) = p_{\boldsymbol{\theta}}(x_t|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t)$ and suppose x_t takes values from $\mathcal{X} = \{1, \ldots, L_m\}$, then recursively, the SMC procedure proceeds at time t as follows [13].

- For $i = 1, \ldots, L_m$, compute

$$q_{\boldsymbol{\theta}}\left(x_t = i|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t\right) \propto p_{\boldsymbol{\theta}}(y_t|x_t = i) p_{\boldsymbol{\theta}}\left(x_t = i|x_{t-1}^{(k)}\right). \quad (8)$$

- Normalize these values such that $\sum_{i=1}^{L_m} q_{\boldsymbol{\theta}}(i|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t) = 1$.
- Draw $x_t^{(k)}$ from $q_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t)$ and let $\boldsymbol{x}_t^{(k)} = (\boldsymbol{x}_{t-1}^{(k)}, x_t^{(k)})$.
- Update the importance weight

$$\begin{aligned} w_t^{(k)} &\propto w_{t-1}^{(k)} p_{\boldsymbol{\theta}}\left(y_t|x_{t-1}^{(k)}\right) \\ &\propto w_{t-1}^{(k)} \sum_{i=1}^{L_m} p_{\boldsymbol{\theta}}(y_t|x_t = i) p_{\boldsymbol{\theta}}\left(x_t = i|x_{t-1}^{(k)}\right). \quad (9) \end{aligned}$$

- Normalize the importance weights so that they sum up to one.

Although powerful and simple to implement, it has been shown that in the above steps, the variance of the importance weights increases over time which causes the degeneracy problem [13]. Degeneracy occurs when too many samples have very small weights and become ineffective samples, in which case, the SMC algorithm becomes inefficient. Degeneracy of the samples can be measured by the effective sample size defined as

$$K_{\text{eff}} \triangleq K\left[1 + \text{Var}\left(\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t^{(k)}|\boldsymbol{y}_t\right)}{q_{\boldsymbol{\theta}}\left(x_t^{(k)}|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t\right)}\right)\right]^{-1} \quad (10)$$

which can be approximated by [14]

$$\widehat{K_{\text{eff}}} = \left(\sum_{k=1}^{K}\left(w_t^{(k)}\right)^2\right)^{-1}. \quad (11)$$

It is suggested that when the effective sample size is too small, e.g., $\widehat{K_{\text{eff}}} \leq (K)/(10)$, the following resampling steps can be performed to rejuvenate the samples [14], [15].

- Draw K sample streams $\{\overline{\boldsymbol{x}}_t^{(j)}, j = 1, \ldots, K\}$ from $\{\boldsymbol{x}_t^{(k)}, k = 1, \ldots, K\}$ with probabilities proportional to $\{w_t^{(k)}, k = 1, \ldots, K\}$.
- Assign equal weights to each stream, $\overline{w}_t^{(k)} = K^{-1}$.

B. SMC With Unknown Parameters

In our system model, the parameter $\boldsymbol{\Theta}$ is unknown and has to be estimated in the SMC process. As we will show later, the parameter $\boldsymbol{\Theta}$ is in a form which can be described by a sufficient statistic that is easily updated. To cope with the unknown static parameters with easily updated sufficient statistics such as the one in our motif discovery model, we consider the case where the distribution can be given as $p(\boldsymbol{\theta}|\boldsymbol{T}_t)$ where $\boldsymbol{T}_t = \boldsymbol{T}_t(\boldsymbol{x}_t, \boldsymbol{y}_t) = \boldsymbol{T}_t(\boldsymbol{T}_{t-1}, x_t, y_t)$ is some sufficient statistic at time t that can be easily updated from the sufficient statistic $\boldsymbol{T}_{t-1}$ at time $t-1$, and the current state and observation, x_t and y_t. Suppose we have available at time $t-1$ a set of properly weighted samples $\{(\boldsymbol{x}_{t-1}^{(k)}, w_{t-1}^{(k)}), k = 1, \ldots, K\}$ with respect to $p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{t-1})$. We have

$$\begin{aligned} &p(\boldsymbol{x}_t, \boldsymbol{\theta}|\boldsymbol{y}_t) \\ &\quad \propto p(\boldsymbol{x}_t, \boldsymbol{\theta}, y_t|\boldsymbol{y}_{t-1}) \\ &\quad \propto p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{t-1}) p(\boldsymbol{\theta}|\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) \\ &\qquad \times p(x_t|\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{\theta}) p(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{t-1}, \boldsymbol{\theta}) \\ &\quad \propto p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{t-1}) p(\boldsymbol{\theta}|\boldsymbol{T}_{t-1}) p(x_t|x_{t-1}, \boldsymbol{\theta}) p(y_t|x_t, \boldsymbol{\theta}). \quad (12) \end{aligned}$$

Keeping only the past simulated streams $\{x_{t-1}^{(k)}, w_{t-1}^{(k)}, k = 1, \ldots, K\}$, (12) can be approximated by drawing $(\boldsymbol{\theta}^{(k)}, x_t^{(k)})$ from a proposal distribution $q(\boldsymbol{\theta}, x_t|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t) = q_1(\boldsymbol{\theta}|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t) \cdot q_2(x_t|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t, \boldsymbol{\theta})$. The new weights can be updated by [16]

$$\begin{aligned} &w_t^{(k)} \\ &\propto w_{t-1}^{(k)} \frac{p\left(\boldsymbol{\theta}^{(k)}|\boldsymbol{T}_{t-1}^{(k)}\right) p\left(x_t^{(k)}|x_{t-1}^{(k)}, \boldsymbol{\theta}^{(k)}\right) p\left(y_t|x_t^{(k)}, \boldsymbol{\theta}^{(k)}\right)}{q_1\left(\boldsymbol{\theta}^{(k)}|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t\right) q_2\left(x_t^{(k)}|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{y}_t, \boldsymbol{\theta}^{(k)}\right)}. \end{aligned} \quad (13)$$

Hence, by obtaining a Monte Carlo approximation of $p(\boldsymbol{x}_t, \boldsymbol{\theta}|\boldsymbol{y}_t)$ from (12) and the set of sufficient statistics $\{\boldsymbol{T}_t^{(k)}, k = 1, \ldots, K\} = \{\boldsymbol{T}_t(\boldsymbol{T}_{t-1}, x_t^{(k)}, y_t), k = 1, \ldots, K\}$, the approximation of $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$ can then be obtained by discarding the samples $\boldsymbol{\theta}^{(k)}$. To simplify computations and achieve lower memory requirements, only the samples $\boldsymbol{x}_t^{(k)}$ and the corresponding sufficient statistics $\boldsymbol{T}_t^{(k)}$ are stored. Furthermore, the static parameters $\boldsymbol{\theta}$ can be estimated by Rao–Blackwellization [17]

$$\mathbb{E}\{\boldsymbol{\theta}|\boldsymbol{y}_t\} = \mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{y}_t}\{\mathbb{E}\{\boldsymbol{\theta}|\boldsymbol{y}_t,\boldsymbol{x}_t\}\} \approx \frac{1}{W_t}\sum_{k=1}^{K} w_t^{(k)}\mathbb{E}\left\{\boldsymbol{\theta}|\boldsymbol{T}_t^{(k)}\right\}. \tag{14}$$

C. The SMC Motif Discovery Algorithm

For the system states up to time $t, \boldsymbol{x}_t = [x_1,\ldots,x_t]$, and the corresponding sequences $\boldsymbol{S}_t = [\boldsymbol{s}_1,\ldots,\boldsymbol{s}_t]$, we will first present their prior distributions and their conditional posterior distributions, and then present the steps of the SMC motif discovery algorithm.

Prior Distributions: Denote $\boldsymbol{\theta}_j \triangleq [\theta_{j1},\ldots,\theta_{j4}]^T, j = 1,\ldots,w$, as the jth column of the position weight matrix $\boldsymbol{\Theta}$. In Monte Carlo methods, the prior distribution is often chosen so that the posterior and the prior are conjugate pairs, i.e., they belong to the same functional family. It can be seen that for all of the motifs in the dataset $\boldsymbol{S}$, the nucleotide counts at each motif location are drawn from multinomial distributions. It is well known that the Dirichlet distribution provides conjugate pair for such distribution. Therefore, we use a multivariate Dirichlet distribution as the prior for $\boldsymbol{\theta}$. The prior distribution for the ith column of the PWM is then given by

$$\boldsymbol{\theta}_i \sim \mathcal{D}(\rho_{i1},\ldots,\rho_{i4}), \quad i = 1,2,\ldots,w. \tag{15}$$

Denote $\boldsymbol{\rho}_i \triangleq [\rho_{i1},\ldots,\rho_{i4}]$. Assuming independent priors, then the prior distribution for the PWM $\boldsymbol{\Theta}$ is the product Dirichlet distribution

$$\boldsymbol{\Theta} \sim \prod_{i=1}^{w}\mathcal{D}(\boldsymbol{\rho}_i). \tag{16}$$

Please refer to [18] for a detailed discussion on the Dirichlet distribution.

Conditional Posterior Distributions: Here we give the conditional posterior distributions that are used in the SMC algorithm:

1. The conditional posterior distribution of the PWM $\boldsymbol{\Theta}$:

$$\begin{aligned}
&p(\boldsymbol{\Theta}|\boldsymbol{S}_t,\boldsymbol{x}_{t-1},x_t=i)\\
&\quad\propto p(\boldsymbol{s}_t|\boldsymbol{\Theta},\boldsymbol{x}_{t-1},x_t=i,\boldsymbol{S}_{t-1})p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1},\boldsymbol{S}_{t-1})\\
&\quad\propto \prod_{j=1}^{w}\boldsymbol{\theta}_j^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(j))}\prod_{\ell=1}^{w}\boldsymbol{\theta}_\ell^{\boldsymbol{\rho}_\ell(t-1)-\mathbf{1}}\\
&\quad\propto \boldsymbol{\Lambda}_w\left(\boldsymbol{\Theta};\boldsymbol{\rho}_1(t-1)\right.\\
&\qquad\left.+\boldsymbol{n}(\boldsymbol{a}_{t,i}(1)),\ldots,\boldsymbol{\rho}_w(t-1)+\boldsymbol{n}(\boldsymbol{a}_{t,i}(w))\right)
\end{aligned} \tag{17}$$

where we denote $\boldsymbol{\Lambda}_w(\boldsymbol{\Theta};\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_w)$ as the product Dirichlet pdf for $\boldsymbol{\Theta}, \boldsymbol{\rho}_i(t) \triangleq [\rho_{i1}(t),\ldots,\rho_{i4}(t)], i = 1,\ldots,w$, as the parameters of the distribution of $\boldsymbol{\Theta}$ at time t, and $\boldsymbol{\theta}_k^{\boldsymbol{\rho}_k(t)-\mathbf{1}} \triangleq \prod_{\ell=1}^{4}\theta_{k\ell}^{(\rho_{k\ell}(t)-1)}$. Note that the posterior distribution of $\boldsymbol{\Theta}$ depends only on the sufficient statistics $\boldsymbol{T}_t \triangleq \{\rho_{ij}(t), 1 \le i \le w, 1 \le j \le 4\}$, which is easily updated based on $\boldsymbol{T}_{t-1}, x_t$, and $\boldsymbol{s}_t$ as given by (17), i.e., $\boldsymbol{T}_t = \boldsymbol{T}_t(\boldsymbol{T}_{t-1},x_t,\boldsymbol{s}_t)$.

2. The conditional posterior distribution of state x_t:

$$\begin{aligned}
p(x_t=i|\boldsymbol{S}_t,\boldsymbol{\Theta}) &= p(x_t=i|\boldsymbol{s}_t,\boldsymbol{\Theta})\\
&\propto \mathcal{B}(\boldsymbol{s}_t;i,\boldsymbol{\Theta}), \quad i = 1,2,\ldots,L_m.
\end{aligned} \tag{18}$$

Sequential Monte Carlo Estimator: We now outline the SMC algorithm for motif discovery when the PWM is unknown, assuming that there is only one motif of length w, and it is present in each of the sequences in the dataset. At time t, to draw random samples of $x_t^{(k)}$ we use the optimal proposal distribution

$$q_2(x_t=i|\boldsymbol{x}_{t-1}^{(k)},\boldsymbol{S}_t,\boldsymbol{\Theta}) = p(x_t=i|\boldsymbol{x}_{t-1}^{(k)},\boldsymbol{S}_t,\boldsymbol{\Theta}) \sim \mathcal{B}(\boldsymbol{s}_t;i,\boldsymbol{\Theta}). \tag{19}$$

To sample $\boldsymbol{\Theta}$, we use the following proposal distribution:

$$\begin{aligned}
&q_1(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}^{(k)},\boldsymbol{S}_t)\\
&\quad\propto \sum_{i=1}^{L_m} p(\boldsymbol{s}_t|x_t=i,\boldsymbol{\Theta},\boldsymbol{x}_{t-1},\boldsymbol{S}_{t-1})p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1},\boldsymbol{S}_{t-1})\\
&\quad\propto \sum_{i=1}^{L_m} P_{t,x_t}^3\prod_{k=1}^{w}\boldsymbol{\theta}_k^{\boldsymbol{\rho}_k(t-1)+\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))-\mathbf{1}}\\
&\quad\propto \sum_{i=1}^{L_m}\lambda_{i,t}\boldsymbol{\Lambda}_w\left(\boldsymbol{\Theta};\boldsymbol{\rho}_1(t-1)\right.\\
&\qquad\left.+\boldsymbol{n}(\boldsymbol{a}_{t,i}(1)),\ldots,\boldsymbol{\rho}_w(t-1)+\boldsymbol{n}(\boldsymbol{a}_{t,i}(w))\right)
\end{aligned} \tag{20}$$

where

$$\lambda_{i,t} \triangleq P_{t,x_t}^3\prod_{\ell=1}^{w}\boldsymbol{\rho}_\ell(t-1)^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(\ell))} \tag{21}$$

with $\boldsymbol{\rho}_\ell(t)^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(\ell))} \triangleq \prod_{j=1}^{4}\rho_{\ell j}(t)^{\mathbb{I}(s_{t,i+\ell-1}-j)}$. Detailed derivation of (20) can be found in the Appendix. The weight update formula (13) can be written as:

$$w_t \propto w_{t-1}\frac{\sum_{i=1}^{L_m}\lambda_{i,t}}{\prod_{k=1}^{w}\sum_{j=1}^{4}\rho_{kj}(t-1)} \tag{22}$$

where the derivation is also given in the Appendix.

We are now ready to give the SMC motif discovery algorithm:

Algorithm 1

```
[SMC motif discovery algorithm for single motif present
in all sequences]
  • For $k = 1,\ldots,K$
      — sample $\boldsymbol{\Theta}^{(k)}$ from the mixture Dirichlet distribution
        given by (20).
      — sample $x_t^{(k)}$ from (19).
      — update the sufficient statistics $\boldsymbol{T}_t^{(k)} =
        \boldsymbol{T}_t(\boldsymbol{T}_{t-1}^{(k)}, x_t^{(k)}, \boldsymbol{s}_t)$ from (17).
  • Compute the new weights according to (22).
  • Compute $\widehat{K_{\text{eff}}}$ according to (11). If $\widehat{K_{\text{eff}}} \le (K)/(10)$
    perform resampling.
```

Motif Scores: When searching for motifs in a dataset, it is often necessary to assign confidence scores to the motif locations estimated. A natural choice in this case will be to use the *a posteriori* probability

$$p(x_t|\boldsymbol{s}_t) \propto p(\boldsymbol{s}_t|x_t)p(x_t) \tag{23}$$

as the confidence score for our estimation, where $p(x_t)$, the prior probability of the starting location of the motif in sequence t is assumed to be uniformly distributed. Note that

$$p(\boldsymbol{s}_t|x_t) = \int p(\boldsymbol{s}_t|x_t, \boldsymbol{\Theta})p(\boldsymbol{\Theta})d\boldsymbol{\Theta}. \tag{24}$$

From [19] and [20], (24) can be approximated by

$$\begin{aligned} p(\boldsymbol{s}_t|x_t) &\approx p(\boldsymbol{s}_t|x_t, \hat{\boldsymbol{\Theta}})p(\hat{\boldsymbol{\Theta}}) \\ &= \mathcal{B}(\boldsymbol{s}_t; x_t, \hat{\boldsymbol{\Theta}})\boldsymbol{\Lambda}_w(\hat{\boldsymbol{\Theta}}; \boldsymbol{\rho}_1, t, \ldots, \boldsymbol{\rho}_w, t) \end{aligned} \tag{25}$$

where the estimated PWM $\hat{\boldsymbol{\Theta}}$ is computed from (14), and we denote (25) as the Bayesian score.

IV. Extensions

In this section, we present modifications to the SMC motif discovery algorithm introduced in the previous section to cope with more sophisticated scenarios including two-block model with unknown gap length, motifs of unknown lengths, motifs with unknown abundance, and sequences with multiple unique motifs.

A. Two-Block Model

In real datasets, the motifs are often highly conserved at both ends of the motif while showing little or no conservation in the middle. Such behavior is exhibited in the CRP dataset as discussed in [21]. For the two-block model, as shown in Fig. 1(b), we assume that the motif is segmented into two blocks of known lengths w_1 and w_2, separated by a gap of length $g \in [g_{\min}, g_{\max}]$. The statistics of the motif can be described by the $4 \times w$ PWM $\boldsymbol{\Theta}$, where now $w = w_1 + w_2$, and the first w_1 columns describe the statistics of the first block, and the remaining w_2 columns describe those of the second.

In order for the SMC motif discovery algorithm to be able to handle sequences with two-block motifs, we simply modify the state space. Instead of letting the state x_t be the location of the first nucleotide of the motif, we let the state be the number pair $x_t \triangleq (a_t, g_t)$ where $a_t \in \{1, \ldots, L_m\}, g_t \in \{g_{\min}, \ldots, g_{\max}\}$, and $a_t + g_t + w_1 + w_2 - 1 \leq L$. The proposal distributions q_1 and q_2, and the updates to the sufficient statistics and the weights are similar to those introduced in Section III-C for the single-block motif model, except that for the two-block model, after w_1 nucleotides, the index for the final w_2 nucleotides are advanced by g_t to account for the gap in the two-block model. We modify the proposal distributions as follows:

$$\begin{aligned} q_2(x_t = (i,j)|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{S}_t, \boldsymbol{\Theta}) &\propto \mathcal{B}(\boldsymbol{s}_t; (i,j), \boldsymbol{\Theta}) \\ &= P_{t,x_t}^3 \prod_{\ell=1}^{w} \boldsymbol{\theta}_\ell^{\boldsymbol{n}(\boldsymbol{a}_{t,(i,j)}(\ell))} \end{aligned} \tag{26}$$

where $\boldsymbol{a}_{t,(i,j)}(\ell)$ is the ℓth nucleotide of the two-block motif

$$\boldsymbol{a}_{t,(i,j)} \triangleq [s_{t,i}, \ldots, s_{t,i+w_1-1}, s_{t,i+j+w_1}, \ldots, s_{t,i+j+w-1}] \tag{27}$$

of the tth sequence. To sample $\boldsymbol{\Theta}$, we use the following proposal distribution:

$$\begin{aligned} &q_1\left(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{S}_t\right) \\ &\quad\propto \sum_{(i,j)\in\{(a_t,g_t)\}} P_{t,x_t}^3 \prod_{\ell=1}^{w} \boldsymbol{\rho}_\ell(t-1)^{\boldsymbol{n}(\boldsymbol{a}_{t,(i,j)}(\ell))} \\ &\quad\times \boldsymbol{\Lambda}_w\left(\boldsymbol{\Theta}; \boldsymbol{\rho}_1(t-1) + \boldsymbol{n}(\boldsymbol{a}_{t,(i,j)}(1)), \ldots,\right. \\ &\quad\left.\boldsymbol{\rho}_w(t-1) + \boldsymbol{n}(\boldsymbol{a}_{t,(i,j)}(w))\right). \end{aligned} \tag{28}$$

Finally, to update the sufficient statistics, we have

$$\begin{aligned} p(\boldsymbol{\Theta}|\boldsymbol{x}_t, \boldsymbol{S}_t) &\propto \boldsymbol{\Lambda}_w(\boldsymbol{\Theta}; \boldsymbol{\rho}_1(t-1) \\ &\quad+\boldsymbol{n}(\boldsymbol{a}_{t,(i,j)}(1)), \ldots, \boldsymbol{\rho}_w(t-1) + \boldsymbol{n}(\boldsymbol{a}_{t,(i,j)}(w))) \end{aligned} \tag{29}$$

and to update the weights

$$w_t \propto w_{t-1} \frac{\sum_{(i,j)\in\{(a_t,g_t)\}} \lambda_{(i,j),t}}{\prod_{k=1}^{w} \sum_{j=1}^{4} \rho_{kj}(t-1)} \tag{30}$$

where $\lambda_{(i,j),t} \triangleq P_{t,x_t}^3 \prod_{\ell=1}^{w} \boldsymbol{\rho}_\ell(t-1)^{\boldsymbol{n}(\boldsymbol{a}_{t,(i,j)}(\ell))}$.

The steps of the modified SMC algorithm for two-block model is as follows.

Algorithm 2

```
[SMC motif discovery algorithm for two-block model]
  • For $k = 1, \ldots, K$
      — Sample $\boldsymbol{\Theta}^{(k)}$ from (28).
      — Sample $x_t^{(k)}$ from (26).
      — Update the sufficient statistics $\boldsymbol{T}_t^{(k)} = \boldsymbol{T}_t(\boldsymbol{T}_{t-1}^{(k)}, x_t^{(k)}, \boldsymbol{s}_t)$ from (29).
  • Compute the new weights according to (30).
  • Compute $\widehat{K_{\text{eff}}}$ according to (11). If $\widehat{K_{\text{eff}}} \leq (K)/(10)$
    perform resampling.
```

B. Motif of Unknown Length

In the previous sections, we have assumed that the length of the motif is known, which is not always the case in practical applications. Assume that the dataset contains a motif of unknown length m^* that falls in the window $[m_{\min}, m_{\max}]$. Here, we modify the SMC algorithm so that the algorithm finds the unknown motif length from the window given. The basic idea is to associate with each sample k the quantity $m_t^{(k)}$, at time t, which is the length of the motif in sample k at time t. Corresponding to this length, we have for sample k the PWM $\boldsymbol{\Theta}^{(k)}$ with size $4 \times w_t^{(k)}$, where $m_t^{(k)} \in [m_{\min}, m_{\max}]$. At $t = 0, m_0^{(k)}$ is drawn uniformly from the set $\{m_{\min}, m_{\min}+1, \ldots, m_{\max}\}$. After updating the weights using the equation that will be introduced shortly, the resampling condition is checked. When resampling is performed, the motif length samples $m_t^{(k)}$ are replaced by the resampled values $\hat{m}_t^{(k)}, k = 1, \ldots, K$. Thus, adaptation to the optimum motif length is achieved through resampling [22].

Assume at time $t-1$ we have the weighted samples $\{(\boldsymbol{x}_{t-1}^{(k)}, m_{t-1}^{(k)}, w_{t-1}^{(k)}), k = 1, \ldots, K\}$. At time t, we let $m_t^{(k)} = m_{t-1}^{(k)}$, and obtain the weighted samples $\{(\boldsymbol{x}_t^{(k)}, m_t^{(k)}, w_t^{(k)}), k = 1, \ldots, K\}$ according to (19), (20), and (22), using $m_t^{(k)}$ as the length of the motif. Thus, following each time increment, the length of the motif for each sample is retained until resampling occurs.

In general, we have $\rho_{ij} > 1$ as t increases. From the definition of $\lambda_{i,t}$ for (21) we can see that each $\lambda_{i,t}$ is a product of L terms, w of which are the coefficients for the Dirichlet distributions, and the rest are probabilities of nucleotides from the nonmotif regions. It is clear that the longer is the motif length of a sample, the larger is the corresponding weight. Thus, the weight update needs to be normalized so that weights of different motif lengths can be compared fairly. From (17), we can see that for each Dirichlet parameter $\boldsymbol{\rho}_i^{(k)}(t)$, at every time step, the parameter is incremented by 1 at the position corresponding to the nucleotide observed at position i of the motif. Therefore, the sum $\sum_{j=1}^4 \rho_{ij}^{(k)}(t)$ is incremented by 1 from $\sum_{j=1}^4 \rho_{ij}^{(k)}(t-1)$, and at a given time t, the sum is the same for all $i, i = 1, \ldots, m_t^{(k)}$. In (22), the denominator is the product of the sums $\sum_{j=1}^4 \rho_{ij}^{(k)}(t-1), \ell = 1, \ldots, m_t^{(k)}$. Since all the sums have the same value, we have $\prod_{i=1}^{m_t^{(k)}} \sum_{j=1}^4 \rho_{ij}^{(k)}(t-1) = (\sum_{j=1}^4 \rho_{1j}^{(k)}(t-1))^{m_t^{(k)}}$. Notice that the product depends on the length of the motif for the given sample. In order to compare the weights of samples with different motif lengths, we need to normalize the parameters of the Dirichlet distribution for all samples such that

$$\left(\sum_{j=1}^{4} \rho_{1j}^{(m_{\min})}(t-1)\right)^{m_{\min}} = \beta_t^{(k)} \left(\sum_{j=1}^{4} \rho_{1j}^{(k)}(t-1)\right)^{m_t^{(k)}} \tag{31}$$

is true for all $m_t^{(k)}$, where $\boldsymbol{\rho}_i^{(m_{\min})}(t)$ and $\boldsymbol{\rho}_i^{(k)}(t)$ are the Dirichlet parameters at the ith position of the motif for the motif with minimum length and the motif of the kth sample, respectively, and $\beta_t^{(k)}$ is the normalizing constant for the kth sample at time t. Since both $\boldsymbol{\rho}_1^{(m_{\min})}(t-1)$ and $\boldsymbol{\rho}_1^{(k)}(t-1)$ are known, $\beta_t^{(k)}$ is simply

$$\beta_t^{(k)} \triangleq \frac{\left(\sum_{j=1}^4 \rho_{1j}^{(m_{\min})}(t-1)\right)^{m_{\min}}}{\left(\sum_{j=1}^4 \rho_{1j}^{(k)}(t-1)\right)^{m_t^{(k)}}}. \tag{32}$$

Similarly, since the computation of $\lambda_{i,t}^{(k)}$ involves the multiplication of $L_m^{(k)} \triangleq L - m_t^{(k)} + 1$ nonmotif nucleotide probabilities, whereas for a motif of minimum length, the multiplication only involves $L_m^{(m_{\min})} \triangleq L - m_{\min} + 1$ terms, we normalize $\lambda_{i,t}^{(k)}$ by

$$\lambda_{i,t}^{(k)} \triangleq (P_{t,x_t}^3)^{\gamma_t^{(k)}} \beta_t^{(k)} \prod_{\ell=1}^{m_t^{(k)}} \boldsymbol{\rho}_\ell^{(k)}(t-1)^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(\ell))} \tag{33}$$

where $\gamma_t^{(k)} = (L_m^{(m_{\min})})/(L_m^{(k)})$. We now use the following modified weight update formula in the SMC motif discovery algorithm for unknown motif length

$$w_t^{(k)} \propto w_{t-1}^{(k)} \frac{c_t^{(k)} \sum_{i=1}^{L_m^{(k)}} \lambda_{i,t}^{(k)}}{\beta_t^{(k)} \prod_{\ell=1}^{m_t^{(k)}} \sum_{j=1}^4 \rho_{\ell j}^{(k)}(t-1)} \tag{34}$$

where $c_t^{(k)} \triangleq (\sum_{i=1}^{L_m^{(k)}} P_{t,x_t}^3)/((\sum_{i=1}^{L_m^{(k)}} P_{t,x_t}^3)^{\gamma_t^{(k)}})$. We will show in the Appendix that this is properly weighted with respect to $p(\boldsymbol{x}_t|\boldsymbol{S}_t)$ for samples that have the same sampled motif lengths. The weights are now normalized so that they are equivalent to the weight for a minimum length motif so that the weights for different motif lengths can be compared fairly. Note that the set of weighted samples $\{(\boldsymbol{x}_t^{(k)}, m_t^{(k)}, w_t^{(k)}), k = 1, \ldots, K\}$ is not properly weighted with respect to the same posterior distribution due to the different motif lengths in the samples. However, the subset of samples with the same sampled motif length, m, is properly weighted with respect to $p(\boldsymbol{x}_t|\boldsymbol{S}_t, m)$. At each resampling, more and more samples with the true motif length are resampled. Eventually, most of the samples will become properly weighted with respect to $p(\boldsymbol{x}_t|\boldsymbol{S}_t, m = m^*)$.

We next summarize the SMC motif discovery algorithm for unknown motif length.

Algorithm 3

```
[SMC motif discovery algorithm for unknown motif length]
• Initialization: Sample $m_0^{(j)}$ uniformly from
  $[m_{\min}, m_{\max}]$.
• Importance Sampling: For $t = 1, 2, \ldots$
  — For $k = 1, \ldots, K$
     • set $m_t^{(k)} = m_{t-1}^{(k)}$.
     • sample $\boldsymbol{\Theta}^{(k)}$ from (20) using $m_t^{(k)}$ as length of
       motif;
     • sample $x_t^{(k)}$ from (19) using $m_t^{(k)}$ as length of
       motif;
     • update the sufficient statistics $\boldsymbol{T}_t^{(k)} = \boldsymbol{T}_t(\boldsymbol{T}_{t-1}^{(k)},$
       $x_t^{(k)}, \boldsymbol{s}_t)$ from (17) using $m_t^{(k)}$ as length of motif.
  — Compute the new weights according to (34).
  — Compute $\widehat{K_{\text{eff}}}$ according to (11). If $\widehat{K_{\text{eff}}} \leq (K)/(10)$
    perform resampling.
• At $T+1$, let $d$ be the number of sequences having
  estimated motif lengths that is different from the final
  converged motif length. For $t = T+1, \ldots, T+d$,
  repeat the Importance Sampling step for the $d$
  sequences to re-estimate motif location and motif
  length.
```

C. *Motif With Unknown Abundance*

To perform motif discovery on datasets where there exist an unknown number of the same motif, we can perform multiple passes of the SMC algorithm on the sequences. Before the subsequent pass, the motif fragment is removed from the sequences where they are found, and the remaining sequence fragments are appended to form a new sequence. By keeping an index on the locations in a sequence where the fragments are joined, we can determine the nucleotides that are possible locations for the starting point of a motif, and modify the state space of (19) accordingly. Note that the SMC algorithm presented so far finds the location in the given sequence that best matches the PWM samples, relative to the other locations in the sequence. However, the actual match may be very poor, thus we may assume that the motif does not exist in the given sequence. From (17), we can see that each Dirichlet parameter $\boldsymbol{\rho}_i$ should have a dominant value at one of the four nucleotides, which indicates the most likely nucleotide to occur at position i. If a motif is present in sequence t, these dominant values are more likely to be present in the computation of λ in (21). If a motif is not present, then the smaller values are more likely to be present in (21), thus the λ value for a sequence will be significantly different depending on the presence of a motif. To determine whether the motif being looked for in the current pass exists in any sequence, we use the following threshold:

$$\lambda_{\text{thresh}} \triangleq \frac{1}{L_m} \left\{ \sum_{i=1}^{L_m} \left[P_{t,x_t}^3 \prod_{m=1}^{w} \max\{\boldsymbol{\rho}_m\} \right.\right. \\ \left.\left. + \sum_{j \neq i} P_{t,x_t}^3 \prod_{m=1}^{w} \boldsymbol{\rho}_m^{\boldsymbol{n}(\boldsymbol{a}_{t,j}(m))} \right] \right\}. \quad (35)$$

This is simply the average of $\lambda_{i,t}$ over all possible starting position i for the starting location of the motif, assuming that a motif exists in the sequence. The sequence t can be declared not to contain a motif if $\sum_{i=1}^{L_m} \lambda_{i,t} < \alpha\lambda_{\text{thresh}}$ where $\alpha < 1$.

The following gives the SMC algorithm for datasets with unknown motif abundance and/or multiple unique motifs.

Algorithm 4

```
[SMC motif discovery algorithm for unknown motif abundance]
  • If there are sequences remaining in the dataset, perform the following steps.
  • Importance Sampling: For $t = 1, 2, \ldots$
    — If motif determined to be present in previous pass, remove motif and append fragments. Mark the location where the fragments are appended. If motif determined not to be present in the previous pass, remove sequence from dataset. For the first pass, assume motif is present in the previous pass.
    — If motif is present in the previous pass, for $k = 1, \ldots, K$
      • sample $\boldsymbol{\Theta}^{(k)}$ from (20);
      • sample $x_t^{(k)}$ from (19);
      • compute $\lambda_{\text{thresh}}$ according to (35);
      • if $\sum_{i=1}^{L_m} \lambda_i > \alpha\lambda_{\text{thresh}}$, declare motif to be present;
      • if $\sum_{i=1}^{L_m} \lambda_i > \alpha\lambda_{\text{thresh}}$, update the sufficient statistics $\boldsymbol{T}_t^{(k)} = \boldsymbol{T}_t(\boldsymbol{T}_{t-1}^{(k)}, x_t^{(k)}, \boldsymbol{s}_t)$ according to (17).
    — If $\sum_{i=1}^{L_m} \lambda_{i,t} > \alpha\lambda_{\text{thresh}}$, compute the new weights according to (22).
    — Compute $\widehat{K}_{\text{eff}}$ according to (11). If $\widehat{K}_{\text{eff}} \leq (K)/(10)$ perform resampling.
```

D. Multiple Unique Motifs

The SMC motif discovery algorithm can very easily be adapted to search a dataset for a large number of unique motifs, while this functionality may be supported by other algorithms in some form or other, the nature of the SMC motif discovery algorithm allows for very efficient implementation. Similar to the extension discussed in Section IV-C, the SMC algorithm is used in multiple passes through the dataset to locate the different motifs. After each pass, the motif that has been located is removed from each sequence, and the remaining fragments are appended to form a new sequence. Before performing a new pass over the modified dataset, the parameters for the priors are reset to their initial values, and the state space is modified to correspond to the possible starting locations of the new motif. Whereas in Section IV-C, the updated parameters are retained to locate the same motif which may occur multiple times in a sequence, by resetting the parameters here we allow the algorithm to look for motifs that may be different from the one that was just found.

E. Using Results From Another Algorithm as Prior to SMC

While the SMC algorithm can be used as a stand-alone algorithm for motif discovery, it can also be used as a second pass algorithm to refine and improve the results of other motif discovery algorithms. Note from (19)–(21), the starting location of a motif is drawn using a PWM sample drawn from a mixture product Dirichlet distribution, which depends on the parameters $\boldsymbol{\rho}_i, i = 1, \ldots, w$. Having parameters that better represent the statistical structure of the motif will allow the algorithm to better recognize the location of the motif inside the sequence. From (17), we can see that the Dirichlet parameters can be easily updated if we have the sequences and the estimated starting locations of the motifs in those sequences by some other motif discovery algorithms. When initiating the SMC algorithm, we simply increment the Dirichlet parameters according to (17) using the sequences and their corresponding estimated starting locations as indexes. This procedure works even if some of the estimated starting locations are incorrect, since the cumulative effect will still allow the SMC algorithm to draw a sample PWM which closely agrees with the statistical structure of the motif.

V. Experimental Results

We have implemented the proposed SMC motif discovery algorithms and evaluated their performance on real and synthetic data. The results are compared to those of MEME, *AlingACE*, *BioProspector*, and UPMF.

A. Results for Real Data

We use two sets of real DNA sequences to evaluate the performance of the SMC motif discovery algorithm. The first dataset used is the cyclic-AMP receptor protein (CRP) from *Escherichia coli* which contains 18 sequences [23]. Each sequence is 105 nucleotides long, and the dataset contains 23 motifs of length 22 that have been experimentally determined. The results of motif discovery using *MEME, AlignACE*, and *BioProspector* are given in [21]. The second set of real data used consists of 200 sequences, each of which contains a

TABLE I
PERFORMANCE COEFFICIENT AND COMPUTATIONAL COMPLEXITY COMPARISON FOR SMC AND GIBBS SAMPLING-BASED ALGORITHMS

Dataset/Algorithm	SMC	Gibbs Sampling
TATA-box perfor. coeff	1.0000	1.0000
TATA-Box running time	7 min.	10 min.
CRP perf. coeff.	0.6700	0.6563
CRP running time	4 min.	8 min.

TATA-box binding site. The TATA-box binding site is usually found as the binding site for the RNA polymerase II and is usually located approximately 25 nucleotides upstream from the transcription start site [24] with experimentally determined length of 8 nucleotides [25]. We have chosen fragments of 75 nucleotides long from upstream of 200 RNA polymerase II binding sites for this dataset.

Basic SMC Algorithm: The performance results of the SMC algorithm, *MEME, AlignACE*, and *BioProspector* on the CRP and TATA-Box datasets are given in Tables I –V.

To demonstrate the performance of the SMC Algorithm 1, we have implemented the proposed SMC algorithms and the Gibbs sampling-based algorithm proposed in [7] in MATLAB, and compared the accuracy and the time needed for both algorithms to process the TATA-box and CRP datasets. For the SMC algorithm, the results are obtained from a single pass with the first 3 sequences estimated again using the updated priors after the first pass. For the Gibbs sampling-based algorithm, we ran the algorithm until the predictions have converged. For the TATA-box dataset we selected the top 200 motifs and for the CRP dataset the top 23. MATLAB simulations are performed on a machine with Pentium IV 2.56-GHz processor. For the TATA-box and the CRP datasets, we used fixed lengths of $w = 8$ and $w = 22$, respectively, for both Algorithm 1 and the Gibbs sampling-based algorithm. Table I shows the performance coefficient [1] and the running time for each algorithm for the two datasets.

For the TATA-box dataset, we can see that both algorithms are able to locate all the motifs correctly, but Algorithm 1 has the advantage in terms of computational time required. For the CRP dataset, Algorithm 1 again requires less running time, and also a higher performance coefficient due to higher degree of overlap with the correct motifs for those motifs that are incorrectly predicted. As we can see from this example, the SMC algorithm can perform better or comparably with the Gibbs sampling-based algorithm in terms of prediction accuracy, and for the cases where the performances are comparable, the SMC algorithms are also less computationally intensive.

Motif of Unknown Length: For the CRP dataset, the length of the motifs has been experimentally determined to be 22 nucleotides long [23]. Here, we employed Algorithm 3 to adaptively determine the optimum length of the CRP motif. For *AlignACE* and *BioProspector*, both of which require a specific motif length as an input, several runs using different motif lengths were performed. Table II shows the estimated motif length, number of potential motifs found by each algorithm, the number of correct site predictions, and the performance coefficient of the predictions. Both *MEME* and *AlignACE* have predicted motifs that contain a consistent shift with respect to the known starting locations. For these two algorithms, we consider the predicted sites that have a consistent shift from the known locations as correct predictions.

For the CRP dataset, we can see that the SMC algorithm outperforms *AlignACE* and *MEME*, and has comparable accuracy to *BioProspector* in terms of performance coefficient. The length of the CRP motif predicted by Algorithm 3 is also only 1 less than the known length. For the results of the CRP dataset, while *MEME* was able to identify more correct instances of the motif, the estimated length and locations predicted by *MEME* are different from the experimentally determined results. From footprinting methods [26] we know that the CRP motif is 22 nucleotides long with the consensus sequence "TTATGT-GATCGAGGTCACACTT". Table III gives the consensus sequence of the CRP motif discovered by each algorithm. We can see that only Algorithm 3 and *BioProspector* have predicted motifs having the same starting location that matches those of the known sites. For *MEME*, not only are the predicted sites shifted downstream by three nucleotides with respect to the known starting locations, the incorrectly predicted sites also have less overlap with the known sites, thus *MEME* has a much lower performance coefficient despite having accurately predicted more motifs.

Motif With Unknown Abundance: In the CRP dataset, 23 motifs are known to exist in the 18 sequences. We treated the number of motifs as an unknown, and employed Algorithm 4 to locate all possible motifs of length 22 within the dataset. The number of motifs found by each algorithm, and the accuracy of the motifs found are shown in Table II. In our experiment, the SMC algorithm found the same ratio of true sites as that of *BioProspector* and exceeded that of *AlignACE*. For *MEME*, although more true motifs were found than by any other algorithm, the motifs found by *MEME* have different starting locations, as discussed earlier.

Two-Block Model: As can be observed from the consensus sequence of the CRP dataset, the CRP motif can also be seen as two blocks of conserved motifs with a gap around six to eight nucleotides long. To see the effect of using a two-block model, we performed the simulations again on the CRP dataset, this time using the two-block model and Algorithm 2. We chose as parameters $w_1 = w_2 = 6, g_{\min} = 6$, and $g_{\max} = 8$ for both the *BioProspector* and Algorithm 2. As we can see in Table IV, both the *BioProspector* and SMC algorithm have similar performances in terms of the number of sites predicted, with the SMC algorithm having higher performance coefficient, and the results for both algorithms using the two-block model outperform the results for both algorithms using the single-block model.

Second Pass Accuracy: Employing the SMC algorithm described in Section IV-E, we can improve upon the results of other algorithms by using the SMC algorithm to perform a second pass through the dataset. In the first row of Table V, using the CRP dataset, we first show the results of motif discovery using the SMC Algorithm 1, *MEME, AlignACE, BioProspector*, and UPMF as first pass algorithms. As the second

TABLE II
MOTIF DISCOVERY RESULTS USING CRP DATASET

Dataset/Algorithm	SMC	BioProspector	MEME	AlignACE
CRP Estimated Length	21	22	20	24
Potential CRP Motif Found	14	13	18	10
CRP Accuracy	12/23	12/23	16/23	10/23
Perf. Coeff.	0.6566	0.6563	0.4816	0.4536

TABLE III
PREDICTED CONSENSUS SEQUENCE FOR THE CRP MOTIF

Algorithm	Estim. Length	Consensus Seq.
Known	22	TTATGTGATCGAGGTCACACTT
SMC	21	TTATGTGATCGAGGTCACACT
BioProspector	22	TTATGTGATCGAGGTCACACTT
AlignACE	24	ATTTATGTGATCGAGGTCACACTT
MEME	20	TGTGATCGAGGTCACACTTT

TABLE IV
TWO-BLOCK MODEL ACCURACY COMPARISON FOR CRP DATASET

	Motif found	Accuracy	Perf. Coeff
SMC	18	16/23	0.7738
BioProspector	17	16/23	0.7432

TABLE V
FIRST PASS ACCURACY FOR EACH ALGORITHM AND THEIR SECOND PASS RESULTS USING SMC ALGORITHM USING CRP DATASET

Pass/Alg.	SMC	BioP.	MEME	AlignACE
1st	12/23	12/23	16/23	10/23
2nd	14/23	14/23	16/23	12/23

row of Table V shows, the second pass results are improved from the first pass results for each of the algorithms tested.

Note that for the *UPMF*, no improvements can be made since the authors have been unable to find the motifs in the CRP dataset in the first pass. To the authors' best knowledge, applying different parameters to *UPMF* results in either no motifs found, or incorrect motifs found. This phenomenon may be due to the basic assumption of the *UPMF*, which is to solve the (ℓ, d) motif problems proposed in [1]. The (ℓ, d) motif problems assumption which is not always held in real motifs. While it has been shown that projection based algorithms can work with real datasets [2], the range of the number of deviations from the consensus sequence exhibited by the motifs in the CRP dataset may be too large, and presents a particular difficult problem for the UPMF.

B. *Results for Synthetic Data*

We used the following rules to generate synthetic data of different levels of conservation for performance comparisons. For highly conserved motifs, the dominant nucleotide at each position in the motif is assigned probability of 91%, where as the remaining nucleotides are assigned probability of 3% each. For mildly conserved motifs, the dominant nucleotide at each position in the motif is assigned probability of 70%, where as the remaining nucleotides are assigned probability of 10%. Non-motif frequency is assigned as 25% for each nucleotide.

Basic SMC Algorithm: We compared the performance of SMC Algorithm 1, *MEME, AlignACE*, and *BioProspector* using synthesized datasets of highly conserved and mildly conserved motifs at various motif lengths. We generated highly conserved motif datasets with motif lengths between 8 and 12, and mildly conserved motif datasets with motif lengths between 17 and 22. Each dataset contains 50 sequences, each of which is 200 nucleotides long. A motif of corresponding length is present in each of the sequences. The performance comparisons using synthetic data of highly conserved and mildly conserved motifs are given in Figs. 2 and 3, respectively. For both highly conserved and mildly conserved motifs, the accuracy of each algorithm is plotted against the length of motif. In both figures, we can see that the SMC algorithm outperforms the other three algorithm for all motif lengths tested. It is clear by looking at (20), motifs with greater length will allow the SMC algorithm to draw more samples with the correct starting location. For mildly conserved motifs, longer motif length is needed to have more nucleotide matches to the true motif so that the correct starting location can be drawn.

Unknown Motif Abundance: In addition to higher accuracy in motif site predictions, the SMC algorithm also has higher sensitivity to possible motif locations. We used 3 datasets of 50 se-

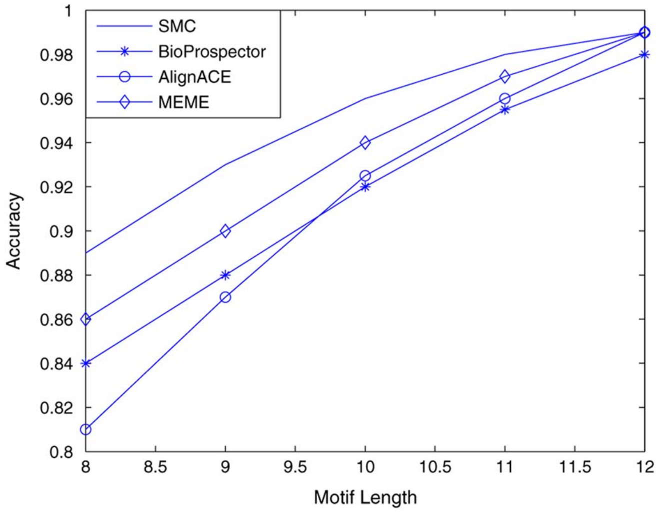


Fig. 2. Accuracy for highly conserved motifs.

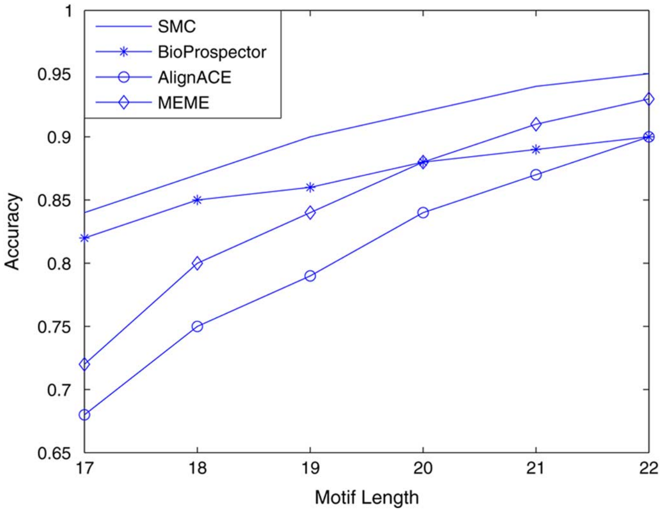


Fig. 3. Accuracy for mildly conserved motifs.

TABLE VI
PERCENTAGE OF SEQUENCES WITH MOTIF THAT ARE CORRECTLY IDENTIFIED

Con./Alg.	SMC	BioP.	MEME	AlignACE
0.91	100%	100%	100%	100%
0.80	92%	90%	91%	88%
0.70	68%	60%	61%	59%

quences each, with conservation of 0.90, 0.80, and 0.71 for the dominant nucleotides in the motifs, and motif length of 12, 16, and 18 respectively. For SMC Algorithm 4, the threshold multiplier α is set to 0.05. In each dataset, only half of the sequences contain a single motif. Table VI tabulates the percentage of sequences in each dataset that is correctly identified with a motif by each algorithm. It is seen that the SMC algorithm suffers less from false negative errors than the other algorithms.

Second Pass Accuracy: In Table VII, we show the second pass results of motif discovery using the SMC algorithm proposed in Section IV-E, *MEME, AlignACE, BioProspector*, and *UPMF* as first pass algorithms. The results are averaged over ten datasets, each of which contains 50 sequences with mildly conserved motifs of 20 nucleotides long. Similarly to the results using the CRP dataset, the results in Table VII show that the SMC algorithm is able to improve upon the results of the first pass algorithms.

TABLE VII
FIRST PASS ACCURACY FOR EACH ALGORITHM AND THEIR SECOND PASS RESULTS USING SMC ALGORITHM USING MILDLY CONSERVED SYNTHETIC DATASETS

Pass/Alg.	SMC	BioP.	MEME	AlignACE	UPMF
1st	89%	87%	86%	83%	92%
2nd	91%	93%	93%	87%	94%

TABLE VIII
PERFORMANCE OF DIFFERENT ALGORITHMS FOR SYNTHETIC $(16, 5)$ MOTIFS EMBEDDED IN SEQUENCES OF 250 AND 600 NUCLEOTIDES

	Sequence Length	Perf. Coeff.
BioProspector	250	0.34
	600	0.04
UPMF	250	0.93
	600	0.70
SMC	250	0.35
	600	0.04
SMC/SMC	250	0.68
	600	0.04
UPMF/SMC	250	0.95
	600	0.73

In Table VIII, we give the performance comparisons of SMC Algorithm 1, *BioProspector*, and *UPMF* on two types of synthetic datasets generated based on the (ℓ, d) motif problem. The datasets contain $(16, 5)$ motifs embedded in sequences of 250 and 600 nucleotides long. As we can see in Table VIII, as first pass algorithms, UPMF significantly outperforms both the SMC Algorithm 1 and *BioProspector*. This is not surprising since the projection-based algorithms are designed specifically for these types of problems. However, the SMC algorithm is still able to improve those results as a second pass algorithm. These results also show that the SMC algorithm has similar performance to other Gibbs sampling-based algorithms in the twilight zones, where the motifs are often too mildly conserved that it is difficult for the statistical-based algorithms to find a starting point.

VI. CONCLUSION

In this paper, we have proposed a sequential Monte Carlo solution to the hidden Markov model of motif discovery problem. We have shown that the SMC algorithm can provide in many cases better performance than those of other algorithms. Even in cases where it does not offer the best performance, it is still valuable as a refining tool for those superior results. The scope of this paper focuses on improving the performance of traditional models where the sequences are assumed to be independent. The current system model can be modified to support models where the locations of the motifs are correlated between adjacent sequences by casting the current model in to a HMM problem. The states and observations in the HMM model remain the same as the current model, and the state transition probabilities can be estimated with some metric based on data provided by microarray experiment results. Finally, we note that in a follow-up work [11], we have proposed a deterministic tree-based search method to discover motifs of unknown length, and motifs with insertion/deletion mutations.

APPENDIX

Derivation of (20):

$$
\begin{aligned}
p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{S}_t)
&\propto p(\boldsymbol{S}_t|\boldsymbol{\Theta}, \boldsymbol{x}_{t-1}, \boldsymbol{S}_{t-1}) p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{S}_{t-1}) \\
&\propto \sum_{i=1}^{L_m} p(\boldsymbol{S}_t|\boldsymbol{\Theta}, x_t = i, \boldsymbol{x}_{t-1}, \boldsymbol{S}_{t-1}) \\
&\quad \times p(x_t = i|\boldsymbol{\Theta}, \boldsymbol{x}_{t-1}, \boldsymbol{S}_{t-1}) p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{S}_{t-1}) \\
&\propto \sum_{i=1}^{L_m} p(\boldsymbol{S}_t|\boldsymbol{\Theta}, x_t = i, \boldsymbol{x}_{t-1}, \boldsymbol{S}_{t-1})\, p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{S}_{t-1}) \\
&\propto \sum_{i=1}^{L_m} P_{t,x_t}^3 \prod_{\ell=1}^{w} \boldsymbol{\theta}_\ell^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(\ell))} \prod_{k=1}^{w} \boldsymbol{\theta}_k^{\boldsymbol{\rho}_k(t-1)-\boldsymbol{1}} \\
&\propto \sum_{i=1}^{L_m} P_{t,x_t}^3 \prod_{\ell=1}^{w} \boldsymbol{\rho}_\ell(t-1)^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(\ell))} \boldsymbol{\Lambda}_w \left(\boldsymbol{\Theta}; \boldsymbol{\rho}_1(t-1)\right. \\
&\quad \left. + \boldsymbol{n}(\boldsymbol{a}_{t,i}(1)), \ldots, \boldsymbol{\rho}_w(t-1) + \boldsymbol{n}(\boldsymbol{a}_{t,i}(w))\right). \qquad (36)
\end{aligned}
$$

The proposal distribution of $\boldsymbol{\Theta}$ is a mixture of Dirichlet distributions which can also be rewritten as

$$
\begin{aligned}
p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{S}_t) &= \frac{1}{\sum_{m=1}^{N} \lambda_{m,t}} \\
&\quad \times \sum_{i=1}^{L_m} \lambda_{i,t} p(\boldsymbol{\Theta}|\boldsymbol{T}_t(\boldsymbol{T}_{t-1}, x_t = i, \boldsymbol{x}_{t-1}, \boldsymbol{S}_t)). \qquad (37)
\end{aligned}
$$

Derivation of (22): From (37), we can see that

$$
\begin{aligned}
&\frac{p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{y}_t)}{p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})} \\
&= \frac{1}{\sum_{m=1}^{L_m} \lambda_{m,t}} \sum_{i=1}^{L_m} \lambda_{i,t} \\
&\quad \times \frac{\prod_{k=1}^{w} \sum_{j=1}^{4} \rho_{kj}(t-1)}{\prod_{k=1}^{w} \boldsymbol{\rho}_k(t-1)^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))}} \prod_{k=1}^{w} \boldsymbol{\theta}_k^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))} \\
&= \frac{\prod_{k=1}^{w} \sum_{j=1}^{4} \rho_{kj}(t-1)}{\sum_{m=1}^{L_m} \lambda_{m,t}} \sum_{i=1}^{L_m} P_{t,x_t}^3 \prod_{k=1}^{w} \boldsymbol{\theta}_k^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))}. \qquad (38)
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\frac{p(x_t|x_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{s}_t|x_t, \boldsymbol{\Theta})}{p(x_t|\boldsymbol{x}_{t-1}, \boldsymbol{S}_t, \boldsymbol{\Theta})} &= p(\boldsymbol{s}_t|x_{t-1}, \boldsymbol{\Theta}) \\
&= \sum_{i=1}^{L_m} P_{t,x_t}^3 \prod_{k=1}^{w} \boldsymbol{\theta}_k^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))}. \qquad (39)
\end{aligned}
$$

The weight update is thus given by

$$
w_t \propto w_{t-1} \frac{\sum_{m=1}^{L_m} \lambda_{m,t}}{\prod_{k=1}^{w} \sum_{j=1}^{4} \rho_{kj}(t-1)}. \qquad (40)
$$

Derivation of (34): For the motif of unknown length model, we use the following proposal distribution to sample $\boldsymbol{\Theta}$ for the kth sample:

$$
\begin{aligned}
q_1(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}^{(k)}, \boldsymbol{S}_t) &\propto \sum_{i=1}^{L_m^{(k)}} \lambda_{i,t}^{(k)} \boldsymbol{\Lambda}_{m_t^{(k)}} \left(\boldsymbol{\Theta}; \boldsymbol{\rho}_1(t-1)\right. \\
&\quad \left. + \boldsymbol{n}(\boldsymbol{a}_{t,i}(1)), \ldots, \boldsymbol{\rho}_{m_t^{(k)}}(t-1) + \boldsymbol{n}(\boldsymbol{a}_{t,i}(m_t^{(k)}))\right) \qquad (41)
\end{aligned}
$$

where the Dirichlet mixture coefficient $\lambda_{i,t}^{(k)}$ is given by (33). Now from (37) we have

$$
\begin{aligned}
&\frac{p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{y}_t)}{p(\boldsymbol{\Theta}|\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})} \\
&= \frac{1}{\sum_{m=1}^{L_m^{(k)}} \lambda_{m,t}^{(k)}} \sum_{i=1}^{L_m^{(k)}} \lambda_{i,t}^{(k)} \frac{\prod_{k=1}^{m_t^{(k)}} \sum_{j=1}^{4} \rho_{kj}(t-1)}{\prod_{k=1}^{m_t^{(k)}} \boldsymbol{\rho}_k(t-1)^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))}} \\
&\quad \times \prod_{k=1}^{m_t^{(k)}} \boldsymbol{\theta}_k^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))} \\
&= \frac{\beta_t^{(k)} \prod_{k=1}^{m_t^{(k)}} \sum_{j=1}^{4} \rho_{kj}(t-1)}{\sum_{m=1}^{L_m^{(k)}} \lambda_{m,t}^{(k)}} (\sum_{i=1}^{L_m^{(k)}} P_{t,x_t}^3)^{\gamma_t^{(k)}} \\
&\quad \times \prod_{k=1}^{m_t^{(k)}} \boldsymbol{\theta}_k^{\boldsymbol{n}(\boldsymbol{a}_{t,i}(k))}. \qquad (42)
\end{aligned}
$$

From (39) and (42), (13) can now be written as

$$
\begin{aligned}
w_t^{(k)} &\propto w_{t-1}^{(k)} \frac{\sum_{i=1}^{L_m^{(k)}} P_{t,x_t}^3}{\left(\sum_{i=1}^{L_m^{(k)}} P_{t,x_t}^3\right)^{\gamma_t^{(k)}}} \frac{\sum_{m=1}^{L_m^{(k)}} \lambda_{m,t}^{(k)}}{\beta_t^{(k)} \prod_{k=1}^{m_t^{(k)}} \sum_{j=1}^{4} \rho_{kj}(t-1)} \\
&\propto \frac{c_t^{(k)} \sum_{m=1}^{L_m^{(k)}} \lambda_{m,t}^{(k)}}{\beta_t^{(k)} \prod_{k=1}^{m_t^{(k)}} \sum_{j=1}^{4} \rho_{kj}(t-1)}. \qquad (43)
\end{aligned}
$$

REFERENCES

[1] P. A. Pevzner and S. H. Sze, "Combinatorial approaches to finding subtle signals in DNA sequences," in *Proc. Int. Conf. Intelligent Systems for Molecular Biology*, 2000, pp. 269–278.

[2] J. Buhler and M. Tompa, "Finding motifs using random projections," *J. Comput. Biol.*, vol. 9, no. 2, pp. 225–242, 2002.

[3] B. Raphael, L. T. Liu, and G. Varghese, "A uniform projection method for motif discovery in DNA sequences," *IEEE Trans. Comp. Biol. Bioinf.*, vol. 1, no. 2, pp. 91–94, 2004.

[4] C. E. Lawrence and A. A. Reilly, "An expectation maximization (EM) algorithm for the identication and characterization of common sites in unaligned biopolymer sequences," *Proteins: Struct., Funct., Genet.*, vol. 7, pp. 41–51, 1990.

[5] T. Bailey and C. Elkan, "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," Univ. of California, San Diego, Tech. Rep. CS93-302, 1993.

[6] T. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proc. 2nd Int. Conf. Intelligent Systems for Molecular Biology*, 1994, pp. 28–36.

[7] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment," *Science*, vol. 262, pp. 208–214, 1993.

[8] F. R. Roth, J. D. Hughes, P. E. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome m RNA quantitation," *Nature Biotechnol.*, vol. 10, pp. 939–945, Oct. 1998.

[9] J. D. Hughes, P. E. Estep, S. Tavazoie, and G. M. Church, "Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae," *J. Mol. Biol*, vol. 296, pp. 1205–1214, Mar. 2000.

[10] X. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," presented at the Pacific Symp. Biocomputing 200, Mauna, Lani, HI, 2001.

[11] K. Liang, X. Wang, and D. Anastassiou, "A profile-based deterministic sequential Monte Carlo algorithm for motif discovery," *Bioinformatics*, vol. 24, no. 1, pp. 46–55, 2008.

[12] P. Fearnhead, "Particle filters for mixture models with an unknown number of components," *J. Stat. Comput.*, vol. 14, pp. 11–21, 2004.

[13] A. Doucet and X. Wang, "Monte Carlo methods for signal processing: A review in the statistical signal processing context," *IEEE Signal Process. Mag.*, vol. 22, pp. 152–170, 2005.

[14] J. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems," *J. Amer. Stat. Assoc.*, vol. 93, no. 443, pp. 1032–1044, 1998.

[15] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[16] G. Storvik, "Particle filters for state-space models with the presence of unknown static parameters," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 281–289, Feb. 2002.

[17] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Stat. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.

[18] B. P. M. Evans and N. Hastings, *Statistical Distributions*, 3rd ed. New York: Wiley-Interscience, 2002.

[19] X. Zhou, X. Wang, R. Pal, I. Ivanov, M. Bittner, and E. R. Dougherty, "A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks," *Bioinformatics*, vol. 20, no. 17, pp. 2918–2927, Nov. 2004.

[20] C. Andrieu, J. Freitas, and A. Doucet, "Robust full Bayesian learning for neural networks," *Neural Comput.*, vol. 13, pp. 2359–2407, 2001.

[21] S. T. Jensen, X. S. Liu, Q. Zhou, and J. S. Liu, "Computational discovery of gene regulatory binding motifs: A Bayesian perspective," *Stat. Sci.*, vol. 19, no. 1, pp. 188–204, 2004.

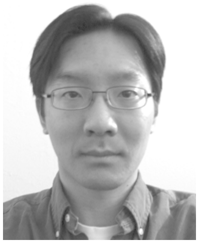
[22] D. Guo, X. Wang, and R. Chen, "Wavelet-based sequential Monte Carlo blind receivers in fading channels with unknown channel statistics," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 227–239, Jan. 2004.

[23] G. D. Stormo and G. W. Hartzell, "Identifying protein-binding sites from unaligned DNA fragments," in *Proc. Nat. Acad. Sci. 1183*, USA, 1989, vol. 86, pp. 1183–1187.

[24] B. Alberts *et al., Essential Cell Biology*. New York: Garland Science, 2003.

[25] Z. S. Juo *et al.*, "How proteins recognize the TATA box," *J. Mol. Biol*, vol. 261, pp. 239–254, Aug. 1996.

[26] J. Liui, M. Gupta, X. Liu, L. Mayerhofere, and C. Lawrence, "Statistical models for biological sequence motif discovery," in *Case Studies in Bayesian Statistics*. New York: Springer, 2004, vol. VI.



Kuo-ching Liang received the B.S. degree in electrical engineering and computer science from the University of California, Berkeley, in 1999, the M.Eng. degree in electrical engineering from Cornell University, Ithaca, NY, in 2000, and the M.S. degree in electrical engineering from the University of Pennsylvania, Philadelphia, in 2003. He is currently working towards the Ph.D. degree at the Department of Electrical Engineering at Columbia University, New York.

His research interests are in the areas of digital communication, statistical signal processing, and its applications in the area of computational biology and bioinformatics.

Xiaodong Wang (S'98–M'98–SM'04–F'08) received the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ.

He is currently on the faculty of the Department of Electrical Engineering, Columbia University, New York.

His research interests are focused in the general areas of computing, signal processing, and communications, and he has published extensively in these areas. Among his publications is a recent book entitled *Wireless Communication Systems: Advanced Techniques for Signal Reception* (Prentice-Hall, 2003). His current research interests include wireless communications, statistical signal processing, and genomic signal processing.

Dr. Wang received the 1999 NSF CAREER Award, and the 2001 IEEE Communications Society and Information Theory Society Joint Paper Award. He has served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE TRANSACTIONS ON INFORMATION THEORY.

Dimitris Anastassiou (S'77–M'78–SM'94–F'98))received the Diploma degree in electrical engineering from the National Technical University of Athens, Greece, in 1974 and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1975 and 1979, respectively.

Prior to joining Columbia University, he was with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. He is currently Professor and Director of Columbia's Genomic Information Systems Laboratory. His research interests focus on computational biology with emphasis on systems-based gene expression data analysis and comparative genomics.

Dr. Anastassiou has received an IBM Outstanding Innovation Award, an NSF Presidential Young Investigator Award, and a Columbia University great teacher award.