14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A SEQUENTIAL MONTE CARLO METHOD FOR MOTIF DISCOVERY

Kuo-ching Liang, Xiaodong Wang, Dimitris Anastassiou†

Department of Electrical Engineering, Columbia University
500 West 120th Street, New York, NY 10027, USA
phone: + (002) 1-212-854-0609,
email: {kcliang,wangx,anastas}@ee.columbia.edu
†also with Columbia University Center for Computational Biology and Bioinformatics (C2B2)

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 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.

1. INTRODUCTION

Efforts by various genomic projects have steadily expanded the pool of sequenced DNA data. By seeking out similarities exhibited in these sequences, we can discover conserved sequence regions, or motifs, and further our knowledge on the functions and evolutions of these sequences. An important approach to motif discovery is the matrix-based approach where a position weight matrix (PWM) of size $4 \times w$ is used to describe the statistical distribution of the four possible nucleotides at every position in a motif of length w. The PWM is estimated in the various matrix-based algorithms and is used to estimate the most likely location of the motif within each sequence. In [1], *MEME*, an algorithm based on EM, is introduced with support for finding unknown number of motifs and unknown number of occurrences in the sequences. Based on [2], *AlignACE* is proposed using 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 [3], the Gibbs sampler-based *BioProspector* is proposed to treat the two-block motif model and palindromic patterns.

Using the MCMC-based algorithms, the sequences are batch-processed to estimate the PWM and the positions of the motifs. These algorithms become inefficient for datasets with large number of sequences. With the ever increasing amount of sequenced genomic data for various organisms, an algorithm that is better equipped to deal with large datasets is necessary. With this goal in mind, we propose a hidden Markov model (HMM) for the matrix-based approach to motif discovery, and proceed to estimate the PWM and the locations of the motifs using a sequential Monte Carlo (SMC) algorithm. The algorithm we propose 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. We show that the SMC-based algorithm can provide comparable performance in real data, and superior performance in synthesized data to the MCMC and EM-based algorithms. Furthermore, the SMC algorithm can also be used as a second-pass algorithm, taking the MCMC or EM-based results as inputs, and further improve those estimates.

2. SYSTEM MODEL

Let $\mathbf{S}_T = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_T\}$, with $\mathbf{s}_t = [s_{t,1}, \cdots, s_{t,L}]$, 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 Figure 1(a). The distribution of the motif is described by the $4 \times w$ position weight matrix $\mathbf{\Theta} = [\theta_1, \theta_2, \cdots, \theta_w]$, where the vector $\theta_j = [\theta_{j,1}, \cdots, \theta_{j,4}]^T$, $j = 1, \cdots, w$, is the probability distribution of the nucleotides $\{A, C, G, T\}$ at the j-th position of the motif. The remaining non-motif nucleotides are assumed to be drawn i.i.d. from the non-motif distribution vector $\theta_0 = [\theta_{0,1}, \cdots, \theta_{0,4}]^T$.

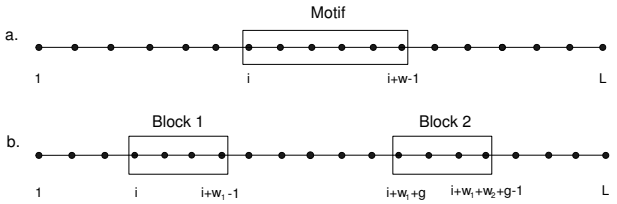

Figure 1: Position weight matrix models. (a) A model for a single-block motif with motif length w. (b) A two-block motif of lengths w_1 and w_2, and gap length g.

We implement an HMM to increment our observation by one full sequence at each step, and the state of the corresponding step is the location of the first nucleotide of the motif in the sequence. Since the last $w-1$ nucleotides in a sequence are not valid locations for the beginning of a motif with length w, at step $t, t = 1, \cdots, T$, the state, denoted as x_t, takes value from the set $\mathscr{X} = \{1, 2, \cdots, L_m\}$, where $L_m = L - w + 1$.

Denote $\mathbf{a}_{t,i}$ as a sequence motif fragment of length w from $\mathbf{s}_t$ beginning from position i, and denote $\mathbf{a}_{t,i}^c$ as the remaining fragment from $\mathbf{s}_t$ with $\mathbf{a}_{t,i}$ removed. Let us further define a vector $\mathbf{n}(\mathbf{a}) = [n_1, n_2, n_3, n_4]$ where $n_i, i = 1, \cdots, 4$, denotes the number of different nucleotides in the sequence fragment $\mathbf{a}$. Given the vectors $\theta = [\theta_1, \cdots, \theta_4]$ and $\mathbf{n} = [n_1, \cdots, n_4]$, we define

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

Since the non-motif nucleotides are assumed to be i.i.d. with the probability of each nucleotide given by θ_0, and the motif nucleotides are independent with the probability of the j-th nucleotide given by θ_j, the distribution of the observed sequence $\mathbf{s}_t$ conditioned on the state at time t and the PWM is then given as follows:

$$p(\mathbf{s}_t \mid x_t = i, \mathbf{\Theta}) = \theta_0^{\mathbf{n}(\mathbf{a}_{t,i}^c)} \prod_{k=1}^{w} \theta_k^{\mathbf{n}(a_{t,i}(k))} \triangleq \mathscr{B}(\mathbf{s}_t; i, \mathbf{\Theta}), \tag{2}$$

where $a_{t,i}(k)$ is the k-th element of the sequence fragment $\mathbf{a}_{t,i}$, and $\mathbf{n}(a_{t,i}(k))$ is a 1×4 vector of zeros except at the position corresponding to the nucleotide $a_{t,i}(k)$, where it is a one. Furthermore, rearranging the order of the sequences in the dataset does not change the statistical properties of the motif, and the location of the motif in sequence $\mathbf{s}_t$ does not affect the location of the motif in sequence $\mathbf{s}_{t+1}$. Therefore, we will assume the transition probability $p(x_{t+1}=j|x_t=i)=1/L_m$.

2.1 Inference Problem

From the discussion above, we formulate our problem as one of filtering an HMM with unknown parameters:

$$x_t \sim \mathscr{MC}\left(\frac{1}{L_m}\mathbf{1}_{1\times L_m},\frac{1}{L_m}\mathbf{1}_{L_m\times L_m}\right), \quad (3)$$

$$(\mathbf{s}_t \mid x_t) \sim \mathscr{B}(\mathbf{s}_t; x_t, \mathbf{\Theta}), \quad (4)$$

where $\mathscr{MC}(\pi,\mathbf{A})$ denotes a discrete-time Markov chain with initial probability distribution π and state transition probability matrix $\mathbf{A}$, $\mathbf{1}_{n\times m}$ denotes an $n\times m$ matrix of ones; and $\mathscr{B}(\mathbf{s}_t;x_t,\mathbf{\Theta})$ is the probability distribution given by (2).

Let us denote the state realizations up to time T as $\mathbf{x}_T \triangleq [x_1,x_2,\cdots,x_T]$ and similarly the sequences up to time T as $\mathbf{S}_T \triangleq [\mathbf{s}_1,\mathbf{s}_2,\cdots,\mathbf{s}_T]$. The unknown parameter of the HMM is $\mathbf{\Theta}$, i.e., the position weight matrix. Given the sequences $\mathbf{S}_T$ we wish to estimate the state realizations $\mathbf{x}_T$, which are the starting locations of the motif in each sequence, and the position weight matrix $\mathbf{\Theta}$, which describes the statistics of the motif. In the next section, we derive the SMC algorithm to solve this inference problem.

3. SMC MOTIF DISCOVERY ALGORITHM

In this section, we 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 4 we will extend this algorithm to additional models.

3.1 SMC with Unknown Parameters

In our model (3)-(4), the parameter $\mathbf{\Theta}$ is unknown and has to be estimated in the SMC process. As we will show later, the parameter $\mathbf{\Theta}$ is in a form which can be described by a sufficient statistic that is easily updated, i.e., the distribution can be given as $p(\mathbf{\Theta}|\mathbf{T}_t)$ where $\mathbf{T}_t=\mathbf{T}_t(\mathbf{x}_t,\mathbf{S}_t)=\mathbf{T}_t(\mathbf{T}_{t-1},x_t,\mathbf{s}_t)$ is some sufficient statistic at time t that can be easily updated from the sufficient statistic $\mathbf{T}_{t-1}$ at time $t-1$, and the current state and observation, x_t and $\mathbf{s}_t$. Suppose we have available at time $t-1$ a set of K properly weighted samples $\left\{\left(\mathbf{x}_{t-1}^{(k)},\omega_{t-1}^{(k)}\right),k=1,\cdots,K\right\}$ with respect to $p(\mathbf{x}_{t-1}|\mathbf{S}_{t-1})$. The posterior distribution $p(\mathbf{x}_t,\theta\mid\mathbf{S}_t)$ can be approximated by drawing $(\theta^{(k)},x_t^{(k)})$ from a proposal distribution $q(\mathbf{\Theta},x_t|\mathbf{x}_{t-1}^{(k)},\mathbf{S}_t)=q_1(\mathbf{\Theta}|\mathbf{x}_{t-1}^{(k)},\mathbf{S}_t)\cdot q_2(x_t|\mathbf{x}_{t-1}^{(k)},\mathbf{S}_t,\theta)$. The new weights can be updated by [4]

$$\omega_t^{(k)} \propto \omega_{t-1}^{(k)} \frac{p\left(\mathbf{\Theta}^{(k)}\mid\mathbf{T}_{t-1}^{(k)}\right)p\left(x_t^{(k)}\mid x_{t-1}^{(k)},\mathbf{\Theta}^{(k)}\right)p\left(\mathbf{s}_t\mid x_t^{(k)},\mathbf{\Theta}^{(k)}\right)}{q_1\left(\mathbf{\Theta}^{(k)}\mid\mathbf{x}_{t-1}^{(k)},\mathbf{S}_t\right)q_2\left(x_t^{(k)}\mid\mathbf{x}_{t-1}^{(k)},\mathbf{S}_t,\mathbf{\Theta}^{(k)}\right)}. \quad (5)$$

Hence a Monte Carlo approximation of $p(\mathbf{x}_t|\mathbf{S}_t)$ can be obtained by

$$\hat{p}_\theta(\mathbf{x}_t\mid\mathbf{S}_t)=\frac{1}{\Omega_t}\sum_{k=1}^{K}\omega_t^{(k)}\mathbb{I}(\mathbf{x}_t-\mathbf{x}_t^{(k)}), \quad (6)$$

where $\Omega_t=\sum_{k=1}^{K}\omega_t^{(k)}$ and $\mathbb{I}(\cdot)$ is the indicator function such that $\mathbb{I}(x)=1$ for $x=0$ and $\mathbb{I}(x)=0$ otherwise, and we update the set of sufficient statistics $\left\{\mathbf{T}_t^{(k)},k=1,\cdots,K\right\}=\left\{\mathbf{T}_t(\mathbf{T}_{t-1},x_t^{(k)},\mathbf{s}_t),k=1,\cdots,K\right\}$. Furthermore, the static parameters θ can be estimated by Rao-Blackwellization [5],

$$\mathbb{E}\{\theta\mid\mathbf{S}_t\}=\mathbb{E}_{\mathbf{x}_t|\mathbf{S}_t}\{\mathbb{E}\{\theta\mid\mathbf{S}_t,\mathbf{x}_t\}\}\approx\frac{1}{\Omega_t}\sum_{k=1}^{K}\omega_t^{(k)}\mathbb{E}\left\{\theta\mid\mathbf{T}_t^{(k)}\right\}. \quad (7)$$

It turns out that the variance of the importance weights increases over time which causes too many samples to 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 which can be estimated by [6]

$$\widehat{K_{eff}}=\left(\sum_{k=1}^{K}(\omega_t^{(k)})^2\right)^{-1}. \quad (8)$$

It is suggested that when the effective sample size is too small, e.g., $\widehat{K_{eff}}\le\frac{K}{10}$, the following resampling steps can be performed to rejuvenate the samples [7]:

- Draw K sample streams $\{\overline{\mathbf{x}}_t^{(j)},j=1,\cdots,K\}$ from $\{\mathbf{x}_t^{(k)},k=1,\cdots,K\}$ with probabilities proportional to $\{\omega_t^{(k)},k=1,\cdots,K\}$.
- Assign equal weights to each stream, $\overline{\omega}_t^{(k)}=K^{-1}$.

3.2 The SMC Motif Discovery Algorithm

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

3.2.1 Prior Distributions:

Let us denote $\theta_j\triangleq[\theta_{j1},\cdots,\theta_{j4}]^T$, $j=1,\cdots,w$, as the j-th column of the position weight matrix $\mathbf{\Theta}$. It can be seen that for all of the motifs in the dataset $\mathbf{S}_T$, the nucleotide counts at each motif location are drawn from multinomial distributions. Therefore, we use a multivariate Dirichlet distribution as the prior for θ_j to obtain a conjugate pair. The Dirichlet distribution is defined as follows. If $\mathbf{u}=[u_1,\cdots,u_N]$, $u_i\ge 0$, $\sum_{i=1}^{N}u_i=1$, and $\mathbf{u}$ has a multivariate Dirichlet distribution $\mathbf{u}\sim\mathscr{D}(\gamma_1,\cdots,\gamma_N)$ with $\gamma_i>0$, then,

$$p(\mathbf{u})=\frac{\Gamma\left(\sum_{i=1}^{N}\gamma_i\right)}{\prod_{i=1}^{N}\Gamma(\gamma_i)}\prod_{i=1}^{N}u_i^{\gamma_i-1}, \quad (9)$$

where $\Gamma(\cdot)$ is the Gamma function. The prior distribution for the i-th column of the PWM is then given by

$$\theta_i\sim\mathscr{D}(\rho_{i1},\cdots,\rho_{i4}),\quad i=1,2,\cdots,w. \quad (10)$$

Let us define $\rho_i\triangleq[\rho_{i1},\cdots,\rho_{i4}]$. Assuming independent priors, then the prior distribution for the PWM $\mathbf{\Theta}$ is the product Dirichlet distribution

$$\mathbf{\Theta}\sim\prod_{i=1}^{w}\mathscr{D}(\rho_i). \quad (11)$$

3.2.2 Conditional Posterior Distributions:

The conditional posterior distribution of the PWM $\mathbf{\Theta}$ can be given as

$$\begin{aligned} p(\mathbf{\Theta}\mid\mathbf{S}_t,\mathbf{x}_t) &\propto p(\mathbf{s}_t\mid\mathbf{\Theta},\mathbf{x}_t,\mathbf{S}_{t-1})\,p(\mathbf{\Theta}\mid\mathbf{x}_{t-1},\mathbf{S}_{t-1}) \\ &\propto \prod_{j=1}^{w}\theta_j^{\mathbf{n}(a_{t,i}(j))}\prod_{i=1}^{w}\theta_i^{\rho_i(t-1)-\mathbf{1}} \\ &\propto \mathscr{D}\left(\mathbf{\Theta};\rho_1(t-1)+\mathbf{n}(a_{t,i}(1)),\cdots,\rho_w(t-1)+\mathbf{n}(a_{t,i}(w))\right) \end{aligned} \quad (12)$$

where we denote $\rho_i(t) \triangleq [\rho_{i1}(t), \cdots, \rho_{i4}(t)], i = 1, \cdots, w$, as the parameters of the distribution of $\mathbf{\Theta}$ at time t, and $\theta_k^{\rho_k(t)-\mathbf{1}} \triangleq \prod_{\ell=1}^{4} \theta_{k\ell}^{(\rho_{k\ell}(t)-1)}$. The conditional posterior distribution of state x_t can be given as

$$p(x_t = i \mid \mathbf{S}_t, \mathbf{\Theta}) = p(x_t = i \mid \mathbf{s}_t, \mathbf{\Theta}) \propto \mathscr{B}(\mathbf{s}_t; i, \mathbf{\Theta}). \tag{13}$$

3.2.3 Sequential Monte Carlo Estimator:

We now outline the SMC algorithm for motif discovery. At time t, to draw random samples of $x_t^{(k)}$ we use the optimal proposal distribution

$$q_2\left(x_t = i \mid \mathbf{x}_{t-1}^{(k)}, \mathbf{S}_t, \mathbf{\Theta}\right) = p\left(x_t = i \mid \mathbf{x}_{t-1}^{(k)}, \mathbf{S}_t, \mathbf{\Theta}\right) \sim \mathscr{B}(\mathbf{s}_t; i, \mathbf{\Theta}). \tag{14}$$

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

$$\begin{aligned} & q_1\left(\mathbf{\Theta} \mid \mathbf{x}_{t-1}^{(k)}, \mathbf{S}_t\right) \\ \propto & \sum_{i=1}^{L_m} p(\mathbf{s}_t \mid x_t = i, \mathbf{\Theta}, \mathbf{x}_{t-1}, \mathbf{S}_{t-1})\, p(\mathbf{\Theta} \mid \mathbf{x}_{t-1}, \mathbf{S}_{t-1}) \\ \propto & \sum_{i=1}^{L_m} \theta_0^{\mathbf{n}(\mathbf{a}_{t,i}^c)} \prod_{k=1}^{w} \theta_k^{\rho_k(t-1)+\mathbf{n}(a_{t,i}(k))-\mathbf{1}} \\ \propto & \sum_{i=1}^{L_m} \lambda_{i,t}\, \mathscr{D}\Big(\mathbf{\Theta}; \rho_1(t-1) + \mathbf{n}(a_{t,i}(1)), \cdots, \rho_w(t-1) + \mathbf{n}(a_{t,i}(w))\Big) \end{aligned} \tag{15}$$

where

$$\lambda_{i,t} \triangleq \theta_0^{\mathbf{n}(\mathbf{a}_{t,i}^c)} \prod_{\ell=1}^{w} \rho_\ell(t-1)^{\mathbf{n}(a_{t,i}(\ell))}, \tag{16}$$

with $\rho_\ell(t)^{\mathbf{n}(a_{t,i}(\ell))} \triangleq \prod_{j=1}^{4} \rho_{\ell j}(t)^{\mathbb{I}(s_{t,i+\ell-1}-j)}$. The weight update formula (5) can be written as:

$$\omega_t \propto \omega_{t-1} \frac{\sum_{i=1}^{L_m} \lambda_{i,t}}{\prod_{k=1}^{w} \sum_{j=1}^{4} \rho_{kj}(t-1)}. \tag{17}$$

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

- For $k = 1, \cdots, K$
  - Sample $\mathbf{\Theta}^{(k)}$ from the mixture Dirichlet distribution given by (15).
  - Sample $x_t^{(k)}$ from (14).
  - Update the sufficient statistics $\mathbf{T}_t^{(k)} = \mathbf{T}_t(\mathbf{T}_{t-1}^{(k)}, x_t^{(k)}, \mathbf{s}_t)$ from (12).
- Compute the new weights according to (17).
- Compute $\widehat{K_{eff}}$ according to (8). If $\widehat{K_{eff}} \leq \frac{K}{10}$ perform resampling.

4. EXTENSIONS

In this section, we present modifications to the basic SMC motif discovery algorithm to support different motif models.

4.1 Two-block Model

For the two-block model, as shown in Figure 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 $\mathbf{\Theta}$, 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 of the HMM. 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 $a_t \in \{1, \cdots, L_m\}$, $g_t \in \{g_{\min}, \cdots, g_{\max}\}$, 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 3.2 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 therefore have the following modified sequence fragment

$$\mathbf{a}_{t,(i,j)} \triangleq [s_{t,i}, \cdots, s_{t,i+w_1-1}, s_{t,i+j+w_1}, \cdots, s_{t,i+j+w-1}]. \tag{18}$$

The samples x_t and $\mathbf{\Theta}$ are drawn using (14) and (15) with $\mathbf{a}_{t,i}$ replaced by $\mathbf{a}_{t,(i,j)}$, and $\mathbf{a}_{t,i}^c$ by $\mathbf{a}_{t,(i,j)}$. The sufficient statistics and weight updates also follow the basic SMC algorithm with similar replacements by $\mathbf{a}_{t,(i,j)}$.

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

- For $k = 1, \cdots, K$
  - Sample $\mathbf{\Theta}^{(k)}$ from (15) using $\mathbf{a}_{t,(i,j)}$ and $\mathbf{a}_{t,(i,j)}$.
  - Sample $x_t^{(k)}$ from (14) using $\mathbf{a}_{t,(i,j)}$ and $\mathbf{a}_{t,(i,j)}$.
  - Update the sufficient statistics $\mathbf{T}_t^{(k)} = \mathbf{T}_t(\mathbf{T}_{t-1}^{(k)}, x_t^{(k)}, \mathbf{s}_t)$ from (12) using $\mathbf{a}_{t,(i,j)}$ and $\mathbf{a}_{t,(i,j)}$.
- Compute the new weights according to (16) and (17) using $\mathbf{a}_{t,(i,j)}$ and $\mathbf{a}_{t,(i,j)}$.
- Compute $\widehat{K_{eff}}$ according to (8). If $\widehat{K_{eff}} \leq \frac{K}{10}$ perform resampling.

4.2 Motif of Unknown Length

In this extension we assume that the dataset contains a motif of unknown length m^* that falls in the window $[m_{\min}, m_{\max}]$ and modify the SMC algorithm to adaptively estimate the unknown length. 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 $\mathbf{\Theta}^{(k)}$ with size $4 \times m_t^{(k)}$, where $m_t^{(k)} \in [m_{\min}, m_{\max}]$. At $t = 0$, $m_0^{(k)}$ is drawn uniformly from the set $\{m_{\min}, m_{\min}+1, \cdots, m_{\max}\}$. 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 $\hat{m}_t^{(k)}$, $k = 1, \cdots, K$. Thus adaptation to the optimum motif length is achieved through resampling [8].

When comparing weights with different motif lengths, the weight with the longer motif length is usually favored. Thus, it is necessary to normalize the weights so that they can be compared fairly. First we normalize the Dirichlet mixture coefficient $\lambda_{i,t}^{(k)}$ as

$$\lambda_{i,t}^{(k)} \triangleq \left(\theta_0^{\mathbf{n}(\mathbf{a}_{t,i}^c)}\right)^{\gamma_t^{(k)}} \beta_t^{(k)} \prod_{\ell=1}^{m_t^{(k)}} \rho_\ell^{(k)}(t-1)^{\mathbf{n}(a_{t,i}(\ell))}, \tag{19}$$

and the weight update formula as

$$\omega_t^{(k)} \propto \omega_{t-1}^{(k)} \frac{c_t^{(k)} \sum_{i=1}^{L_m^{(k)}} \lambda_{i,t}^{(k)}}{\beta_t^{(k)} \prod_{\ell=1}^{m_t^{(k)}} \sum_{j=1}^{4} \rho_{\ell j}^{(k)}(t-1)}, \tag{20}$$

where $\beta_t^{(k)} \triangleq \frac{\left(\sum_{j=1}^{4} \rho_{1j}^{(m_{\min})}(t-1)\right)^{m_{\min}}}{\left(\sum_{j=1}^{4} \rho_{1j}^{(k)}(t-1)\right)^{m_t^{(k)}}}$, $\gamma_t^{(k)} = \frac{L_m^{(m_{\min})}}{L_m^{(k)}}$, and $c_t^{(k)} \triangleq \frac{\sum_{i=1}^{L_m^{(k)}} \theta_0^{\mathbf{n}(\mathbf{a}_{t,i}^c)}}{\left(\sum_{i=1}^{L_m^{(k)}} \theta_0^{\mathbf{n}(\mathbf{a}_{t,i}^c)}\right)^{\gamma_t^{(k)}}}$.

Thus the weights are 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 $\left\{(\mathbf{x}_t^{(k)}, m_t^{(k)}, \omega_t^{(k)}), k = 1, \cdots, K\right\}$ 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 $p(\mathbf{x}_t|\mathbf{S}_t, m)$. 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 $p(\mathbf{x}_t|\mathbf{S}_t, m = m^*)$.

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

- Initialization: Sample $m_0^{(j)}$ uniformly from $[m_{\min}, m_{\max}]$.
- Importance Sampling: For $t = 1, 2, \cdots$
  - For $k = 1, \cdots, K$
    - Set $m_t^{(k)} = m_{t-1}^{(k)}$.
    - Sample $\mathbf{\Theta}^{(k)}$ from (15) using $m_t^{(k)}$ as length of motif.
    - Sample $x_t^{(k)}$ from (14) using $m_t^{(k)}$ as length of motif.
    - Update the sufficient statistics $\mathbf{T}_t^{(k)} = \mathbf{T}_t(\mathbf{T}_{t-1}^{(k)}, x_t^{(k)}, \mathbf{s}_t)$ from (12) using $m_t^{(k)}$ as length of motif.
  - Compute the new weights according to (20).
  - Compute $\widehat{K_{eff}}$ according to (8). If $\widehat{K_{eff}} \leq \frac{K}{10}$ 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 $t = T+1, \cdots, T+d$, repeat the Importance Sampling step for the d sequences to re-estimate motif location and motif length.

4.3 Motif with Unknown Abundance

To perform motif discovery on datasets where the sequences can contain any number of the motif, we can perform multiple passes of the SMC algorithm on the dataset. 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 indices on the locations in a sequence where the fragments are joined, we can determine the remaining possible locations for the starting point of a motif, and modify the state space of (14) accordingly. Note that in this case, a threshold is needed to determine the presence of a motif in the sequence. The algorithm is terminated when all the sequences have been determined not to contain any motifs. To determine whether the motif being looked for in the current pass exists in any sequence, we use the following threshold:

$$\lambda_{thresh} \triangleq \frac{1}{L_m}\left\{\sum_{i=1}^{L_m}\left[\theta_0^{\mathbf{n}(\mathbf{a}_{t,i}^c)}\prod_{m=1}^{w}\max\{\rho_m\} + \sum_{j\neq i}\theta_0^{\mathbf{n}(\mathbf{a}_{t,j}^c)}\prod_{m=1}^{w}\rho_m^{\mathbf{n}(a_{t,j}(m))}\right]\right\} \tag{21}$$

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 $\sum_{i=1}^{L_m}\lambda_{i,t} < \alpha\lambda_{thresh}$ where $\alpha < 1$.

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

- If there are sequences remaining in the dataset, perform the following steps.
- Importance Sampling: For $t = 1, 2, \cdots$
  - 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 $k = 1, \cdots, K$
    - Sample $\mathbf{\Theta}^{(k)}$ from (15).
    - Sample $x_t^{(k)}$ from (14).
    - Compute λ_{thresh} according to (21).
    - If $\sum_{i=1}^{L_m}\lambda_i > \alpha\lambda_{thresh}$, declare motif to be present.
    - If $\sum_{i=1}^{L_m}\lambda_i > \alpha\lambda_{thresh}$, update the sufficient statistics $\mathbf{T}_t^{(k)} = \mathbf{T}_t(\mathbf{T}_{t-1}^{(k)}, x_t^{(k)}, \mathbf{s}_t)$ according to (12).
  - If $\sum_{i=1}^{L_m}\lambda_{i,t} > \alpha\lambda_{thresh}$, compute the new weights according to (17).
  - Compute $\widehat{K_{eff}}$ according to (8). If $\widehat{K_{eff}} \leq \frac{K}{10}$ perform resampling.

4.4 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 (14)-(16), the starting location of a motif is drawn using a PWM sample drawn from a mixture product Dirichlet distribution, which depends on the parameters $\rho_i, i = 1, \cdots, w$. From (12) 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 (12) using the sequences and their corresponding estimated starting locations as indexes.

5. 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 that of existing motif discovery algorithms *MEME* and *BioProspector*.

5.1 Results for Real Data

We evaluate the performance of the SMC algorithm using the cyclic-AMP receptor protein (CRP) from *Escherichia coli* which contains 23 motifs in 18 sequences. The performance results of the SMC algorithm, *MEME*, and *BioProspector* on the CRP dataset are given in Tables 1 and 2.

For the CRP dataset, we adaptively determined the optimum length using the extension to the SMC algorithm. For *AlignACE* and *BioProspector*, several runs using different motif lengths were performed. The results are shown in Table 1 for each algorithm. For the CRP dataset, we can see that the SMC algorithm outperforms *MEME*, and has comparable accuracy to that of *BioProspector*. Only *BioProspector*'s estimated motif length matches that of the experimental result. However, the SMC found motifs with starting locations that match the experimental and *BioProspector* results, whereas the motifs found by *MEME* have starting locations that are different from those determined both experimentally and by other algorithms.

We next applied the extension proposed in Section 4.3 to the CRP dataset. The results are shown in Table 1. In our experiment, the SMC algorithm found the same ratio of true sites as that of *BioProspector*. For *MEME*, although more motifs were found than by any other algorithm, the motifs found by *MEME* have different starting locations, as discussed earlier.

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 6 to 8 nucleotides long. We performed motif discovery again on the CRP dataset, this time using the two-block model. We chose as parameters $w_1 = w_2 = 6$, $g_{\min} = 6$, and $g_{\max} = 8$ for both the *BioProspector* and the modified SMC algorithm as described in Section 4.1. As we can see in Table 2, both the *BioProspector* and SMC algorithm have similar performance, and the results for both algorithms using the two-block model outperform the results for both algorithms using the single-block model.

5.2 Results for Synthetic Data

We used the following rules to generate synthetic data for performance comparisons. 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. Each dataset used contains 50 sequences.

We compared the performance of basic SMC algorithm, *MEME*, *AlignACE*, and *BioProspector* using synthesized datasets at various motif lengths. The performance comparisons are given in Figure 2. We can see that the SMC algorithm outperforms the other three algorithm for all motif lengths tested. It is clear by looking at (15), motifs with greater length will allow the SMC algorithm to draw more samples with the correct starting location.

Employing the SMC algorithm described in Section 4.4, we can improve upon the results of other algorithms by using the SMC algorithm to perform a second pass through the dataset. In Table 3, the results of first pass results from various algorithms are compared to results after using the SMC algorithm as the second pass algorithm. We can see that the second pass results improve over the first pass results for each of the algorithms tested. Notice that the SMC first pass results are also improved after the second pass.

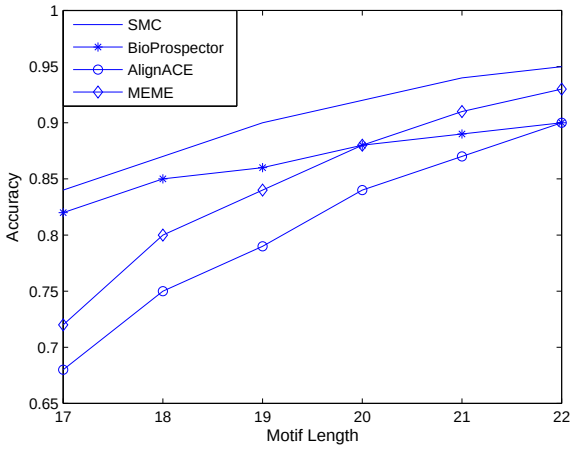

Figure 2: Accuracy for synthetic motifs of various length.

6. CONCLUSIONS

The SMC algorithm proposed in this paper performs motif discovery in sequences by jointly estimating the position weight matrix that describes the statistical properties of the motif and the motif location. The SMC algorithm provides a more accurate and efficient solution to datasets with large amount of sequences, which is crucial due to the increasing number of sequenced genomes and the growing numbers of paralogous and orthologous sequences.

Dataset/Algorithm	SMC	BioProspector	MEME
CRP Estimated Length	21	22	20
Potential CRP Motif Found	14	13	18
CRP Accuracy	12/23	12/23	16/23

Table 1: Motif discovery results using CRP dataset.

	Motif found	Accuracy
SMC	18	16/23
BioProspector	17	16/23

Table 2: Two-block model accuracy comparison for CRP dataset.

Passes/Algorithm	SMC	BioProspector	MEME
First Pass	91%	87%	86%
Second Pass	93%	93%	93%

Table 3: First pass accuracy for each algorithm and their second pass results using SMC algorithm.

REFERENCES

[1] T. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proc. 2nd In't Conf. on Intelligent Systems for Molecular Biology*. AAAI Press, 1994, pp. 28–36.

[2] 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.

[3] X. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," in *Pacific Symposium on Biocomputing 2001*, Mauna Lani, HI, January 3-7. 2001, pp. 127–138.

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

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

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

[7] 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. Sig. Proc.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

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