Bioinformatics Advance Access published March 25, 2007

BIOINFORMATICS

A Sequential Monte Carlo EM Approach to the Transcription Factor Binding Site Identification Problem

Edmund S. Jackson,* William J. Fitzgerald

Signal Processing Laboratory
Department of Engineering
Cambridge University

Associate Editor: Dr. Alex Bateman

ABSTRACT

Motivation A significant and stubbornly intractable problem in genome sequence analysis has been the *de-novo* identification of transcription factor binding sites in promoter regions. Although theoretically pleasing, probabilistic methods have faced difficulties due to model mismatch and the nature of the biological sequence. These problems result in inference in a high dimensional, highly multimodal space, and consequently often display only local convergence and hence unsatisfactory performance.

Algorithm In this paper we derive and demonstrate a novel method utilising a sequential Monte Carlo-based expectation-maximisation (EM) optimisation to improve performance in this scenario. The Monte Carlo element should increase the robustness of the algorithm compared to classical EM. Furthermore, the parallel nature of the sequential Monte Carlo algorithm should be more robust than Gibbs sampling approaches to multimodality problems.

Results We demonstrate the superior performance of this algorithm on both semi-synthetic and real data from *E. coli*.

Availability http://sigproc-eng.cam.ac.uk/∼ej230/smc_em_tfbsid.tar.gz

Contact ej230@cam.ac.uk

1 INTRODUCTION

The fundamental element in the control of gene expression is a class of proteins called transcription factors (TFs) [23]. These direct the timing, location and rate of gene expression by binding to the DNA double helix immediately upstream of specific genes. Should the complete set of transcription factor binding sites (TFBS) for an organism be found, it would reveal the fundamental topology of the gene regulatory network, a significant advancement of our understanding. Unfortunately, determining these TFBSs is still an open problem. Chemical assays exist, but are too costly and time consuming for a problem of this scale. The alternative is a computational approach. However, to date no approach has proven sufficiently accurate or precise and much work remains [14, 11].

There exist two main categories of computational methods: enumerative and probabilistic. Enumerative methods search for over-represented sequences by direct string comparison. Although highly effective, they are not physically or biologically well motivated and hence cannot model physical traits beyond conservation. Probabilistic methods are model based and are hence much more flexible and potentially powerful. These methods usually treat the promoter sequence as a set of binding sites embedded in a background of nucleotides. Inference is performed to determine the location of the binding sites and their parameters. The independent multinomial model is the standard for the binding site and a Markov chain model is usually used for the background.

1.1 Probabilistic Methods and Justification

The canonical treatment of this problem is as a missing data problem, with the binding site locations missing and the motif model parameters to be inferred. Hence, algorithms based on expectation maximisation (EM) [8] (such as MEME [3]), as well as EM's stochastic counterpart, data augmentation [21], (such as the "Gibbs Sampler" method [16, 12]) are popular.

We wish to infer the model parameters, being the motif binding parameters (often represented as a 'logo' such as in Figure 3). The dimension of this parameter space is $4 \cdot L$, where L is the motif length, and often greater than 20. This may certainly be considered high dimensional, which is widely known to complicate inference.

A more serious concern, and the focus of this paper, is the multimodatility of the parameter space. Unfortunately, in these problems there exist a multitude of suboptimal solutions, corresponding to the many (semi-)conserved patterns found among the promoters. This produces a highly multimodal, highly mode spread probability space. This is a well known pathological condition for both classical EM and Gibbs sampling. Our algorithm is designed specifically for these problems.

In conditions such as these "the Gibbs sampler is unable to move the Markov chain to another mode of equal importance because of its inability to step over valleys of low probability" [4]. Importance sampling approaches, due to the parallel nature, present a promising solution. "There are settings ... where Markov Chain Monte Carlo (MCMC) has a very hard time converging ... while importance sampling based on identical proposals manages to reach regions of interest for the target distribution," [18] For this reason we propose an algorithm based on importance sampling. Specifically, we propose a sequential importance sampling / resampling (SIR) algorithm in place of the usual E step of the EM algorithm, following [2]. This avoids the greediness problems of classical EM, and should produce a more robust representation of the probability landscape and hence superior results.

The convergence difficulties with the Gibbs sampler method for the TFBS identification problem have been recently and independently noted in [9], who propose an alternative approach.

*Thanks Trinity College and the Cambridge Commonwealth Trust for funding. Corresponding Author.

1.2 Development of the Approach

In the statistics literature SIR has long been shown useful for data augmentation [20]; however there the sampling distribution is fixed. Given its importance for convergence, [10] proposes a dynamic sampling distribution. They present a stochastic EM method with multiple parsing of the data set, optimising the sampling distribution after each parse. While an improvement, for large data sets this is still computationally burdensome, and does not produce as rapid a convergence as possible. In [2], which considers online estimation, the authors suggest even finer updating by segmenting the data set into blocks and updating the sampling distribution after each block. Similarly but for the batch case, [5] develops a particle filter which has a stepwise evolving sampling distribution. This fine data partition and rapid sequential incorporation allows the data to quickly influence inference, and produces a tempering effect. Our proposal is to use the already fine segmentation of the promoter set into individual promoters, which should achieve these advantages.

In summary, we propose an EM based algorithm due to its natural application to this latent variable mixture problem. However, we propose replacing the classical E step with a Monte Carlo approximation in order to gain robustness against multimodality. This is very similar to data-augmentation; however to gain further robustness we propose using an importance sampling method. This allows multiple, semi-independent explorations of the probability space. Finally, this importance sampling is performed sequentially (with resampling) as this allows a fine data partition and the itinerant advantages. This algorithm will be discussed in detail in Section 3 below.

2 MODELS

The promoter region of each gene is modelled as a mixture of background sequence and an unknown number of probabilistically conserved motif instances at unknown locations. The vector of motif instance locations is treated as a hidden variable which is inferred and from which the model of the motif calculated. This section presents concise representation representation of this pervasive model.

2.0.1 Background Model Very little biological knowledge exists on the nature of this sequence; and hence given the linear nature of the data a Markov model of particular order is employed. Early efforts focused on order 0 models while increasingly higher order models are utilised for eukaryotes. Recently switching Markov model based have been introduced [22, 9] but these are beyond our scope.

We begin with some notation. Define the set of nucleotide symbols $\Phi \triangleq \{a, c, t, g\}$. Furthermore, define the sequence of N nucleotides $\mathbf{S} \in \Phi^N$, and let sub-scripting indicate a sub-string operation, $\mathbf{S}_{1:N} \triangleq [S_1, \ldots, S_N], S_i \in \Phi \ \forall i$. Let the space of T^{th} order Markov transitions be denoted by

$$\Psi^T \triangleq \Phi^T \otimes \Phi.$$

Let the probabilities of a particular transition, $\psi \in \Psi^T$, be written as $\theta_{0,\psi} \triangleq \mathbb{P}(\psi)$, with the set of all such probabilities given by

$$\boldsymbol{\theta}_0 \triangleq \left\{\theta_{0,\psi}\right\}_{\psi \in \Psi^T}.$$

Let the set of such transitions observed in the string $\mathbf{S}$ be

$$\mathbf{S}^T \triangleq \left\{\psi \in \Psi^T : \psi \subset \mathbf{S}\right\}.$$

Using the notation $\mathbf{S}_{a:a+T-1} \rightarrow \mathbf{S}_{a+T}$ to denote a transition of a T^{th} order Markov chain, the probability of a sequence, $\mathbf{S}$, is given by

$$\begin{aligned} P(\mathbf{S}|\,\boldsymbol{\theta}_0) &= \prod_{i=1}^{N} P\left(\mathbf{S}_i|\,\mathbf{S}_{\max(0,i-T):i-1}, \boldsymbol{\theta}_0\right) \\ &\triangleq \prod_{i=1}^{N} \theta_{0,\mathbf{S}_{\max(0,i-T):i-1}\rightarrow \mathbf{S}_i} \\ &\approx \prod_{i=T}^{N} \theta_{0,\mathbf{S}_{i-T:i-1}\rightarrow \mathbf{S}_i} \end{aligned}$$

where $\mathbf{S}_{0:0} \triangleq \emptyset$. The terms in $\mathbf{S}_i, i < T$ are calculated by appropriate lower order Markov chains, or usually simply ignored, as in the given approximation, as they contribute only a small edge effect. Thus, by exchangeability we have

$$P(\mathbf{S}|\,\boldsymbol{\theta}_0) \propto \prod_{i \leq T} \theta_{0,\mathbf{S}_{1:i-1}\rightarrow \mathbf{S}_i} \prod_{\psi \in \Psi^T} \theta_{0,\psi}^{|\mathbf{S}^T=\psi|} \tag{1}$$

$$\approx \prod_{\psi \in \Psi^T} \theta_{0,\psi}^{|\mathbf{S}^T=\psi|}, \tag{2}$$

where the cardinality of the set $\mathbf{S}^T = \psi$ is simply the number of observations of the transition ψ in the observed sequence.

The probabilities, $\boldsymbol{\theta}_0$, are the parameters of this model, and are found empirically on a per-organism basis. Fitting the maximum likelihood (ML) estimation [15] on the full set of promoter regions of the organism of interest results in

$$\hat{\theta}_{0,\mathbf{S}_{i:i+T-1}\rightarrow s} = \frac{|\{\mathbf{S}_{i:i+T-1} \otimes s \subset \mathbf{S}\}|}{|\{\mathbf{S}_{i:i+T-1} \subset \mathbf{S}\}|} \tag{3}$$

and hence the set of T^{th} order transition probabilities, $\boldsymbol{\theta}_{0,\psi}$.

2.0.2 Motif Model The binding site motif is modelled as an independent, non-identically distributed multinomial distribution. Thus a motif of width w, $\mathbf{S} = \mathbf{S}_{1:w} = [S_1, \ldots, S_w] \in \Phi^w$ has likelihood

$$\begin{aligned} P(\mathbf{S}|\,\boldsymbol{\theta}) &= \prod_{i=1}^{w} P(S_i|\,\boldsymbol{\theta}) \\ &= \prod_{i=1}^{w} \theta_{S_i,i} \end{aligned}$$

where $\boldsymbol{\theta}$ is termed the 'position weight matrix' and defined as

$$\boldsymbol{\theta} = [\theta_{i,j}]_{i \in \{1,\ldots,w\}; j \in \Phi} \tag{4}$$

where $\theta_{i,j}$ is the probability of observing a nucleotide of type j in position i of the motif. This matrix is the subject of subsequent inference.

In extension, the likelihood of a set of M motif instances, $\mathcal{S} = \{\mathbf{S}_m\}_{m\in[1,M]}$, is found by exchangeability to be

$$\begin{aligned} P(\mathcal{S}|\boldsymbol{\theta}) &= \prod_{m=1}^{M}\prod_{i=1}^{w}\theta_{i,\mathbf{S}_{m,i}} \\ &= \prod_{\phi\in\Phi}\prod_{i=1}^{w}\theta_{i,\phi}^{|\{\mathbf{S}_{m,i}=\phi:m\in[1,M]\}|} \end{aligned} \quad (5)$$

Similarly to (3) the ML solution of the parameters is

$$\hat{\theta}_{i,\phi} = \frac{|\{\mathbf{S}_{m,i}=\phi : m\in[1,M]\}|}{M}. \quad (6)$$

2.0.3 Mixture Model Let $\mathcal{S} = \{\mathbf{S}_1,\ldots,\mathbf{S}_K\}$ be a promoter set of K promoters, each of which is N nucleotides in length, $\mathbf{S}_k = [S_1,\ldots,S_N] \in \Phi^N$. Consider $\mathcal{S}$ as a single sequence modeled as a Markov chain with transition probabilities $\boldsymbol{\theta}_0$, with a number of motif instances distributed according to an independent multinomial model with parameter $\boldsymbol{\theta}$, inserted randomly at positions $\mathbf{A} = \{A_{k,m}\}$ at positions m in motifs k, where $m \in [1, M_k]$ and M_k is the number of motif instances in promoter k.

Ignoring minor edge effects, the likelihood for this sequence is given through independence arguments by

$$P(\mathcal{S}|\boldsymbol{\theta}_0,\boldsymbol{\theta},\mathbf{A}) \approx P(\mathcal{S}_{-\mathbf{A}}|\boldsymbol{\theta}_0)P(\mathcal{S}_{\mathbf{A}}|\boldsymbol{\theta}) \propto \frac{P(\mathcal{S}_{\mathbf{A}}|\boldsymbol{\theta})}{P(\mathcal{S}_{\mathbf{A}}|\boldsymbol{\theta}_0)} \quad (7)$$

where subscripts indicate sub-strings; $\mathcal{S}_{-\mathbf{A}}$ is the sequence with the motif instances at $\mathbf{A}$ removed and $\mathcal{S}_{\mathbf{A}}$ is the set of motif instances. With slight abuse of notation we may define the set of nucleotides at position i of all motif instances in promoter k as

$$\mathbf{S}_{\mathbf{A}_k,i} \triangleq \left\{\mathbf{S}_{k,A_{k,m}+i-1} : m\in[1,M_k]\right\} \in \Phi^{M_k}$$

and the set of all motif instances in promoter k as

$$\mathbf{S}_{\mathbf{A}_k} \triangleq \left\{\mathbf{S}_{\mathbf{A}_{k,i}} : i\in[1,w]\right\} \in (\Phi^w)^{M_k}$$

Substituting the models, (1) and (5) into (7) we arrive at the mixture likelihood function

$$P(\mathcal{S}|\boldsymbol{\theta}_0,\boldsymbol{\theta},\mathbf{A}) \propto \frac{\prod_\Phi \prod_{i=1}^{w}\theta_{i,\phi}^{|\{\mathbf{S}_{\mathbf{A}_k,i}=\phi:k\in[1,K]\}|}}{\prod_{\Psi^T}\boldsymbol{\theta}_{0,\psi}^{|\{\mathbf{S}^T_{\mathbf{A}_k}=\psi:k\in[1,K]\}|}}. \quad (8)$$

3 INFERENCE ALGORITHM

Having developed the models and likelihood functions in the previous section it remains to perform inference based on them. The goal of this algorithm is to find the most probable model, $\hat{\boldsymbol{\theta}}$, of a repeated TFBS motif supposed within a set of sequences $\mathcal{S}$. This is achieved by inferring the latent variable, $\mathbf{A}$, the set of motif instance positions, and calculating $\hat{\boldsymbol{\theta}}$ based on this. The calculation proceeds on two levels; the upper level is a stochastic EM algorithm to be described in Section 3.1 within which the E step is performed with a Monte Carlo algorithm described in Section 3.2.

3.1 Sequential Monte Carlo EM

The EM algorithm for the latent variable problem consists of two steps:

- **Expectation (E):** of the log-likelihood function

$$\begin{aligned} \mathcal{Q}\left(\boldsymbol{\theta}\left|\hat{\boldsymbol{\theta}}_{(m)},\mathcal{S}\right.\right) &= \mathbb{E}_{\hat{\boldsymbol{\theta}}_{(m)}}\left[\log L^c(\boldsymbol{\theta}|\mathcal{S},\mathbf{A})\right] \\ &= \int_{\mathcal{A}} \log f(\mathcal{S},\mathbf{A}|\boldsymbol{\theta}) f(\mathbf{A}|\hat{\boldsymbol{\theta}}_{(m)},\mathcal{S}) d\mathbf{A} \end{aligned} \quad (9)$$

- **Maximisation (M):** with respect to $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}}_{(m+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta}\left|\hat{\boldsymbol{\theta}}_{(m)},\mathcal{S}\right.\right) \quad (10)$$

which are iterated until $\hat{\boldsymbol{\theta}}$ converges.

For the E step we propose a sequential Monte Carlo (SMC) EM method, using the natural segmentation of $\mathbf{S}$ into promoters $\mathbf{S}_k$ and hence sampling via SIS/R. We approximate (9) by

$$\hat{\mathcal{Q}}\left(\boldsymbol{\theta}\left|\hat{\boldsymbol{\theta}}_{(m)},\mathcal{S}\right.\right) \propto \textstyle\sum_{i=1}^{P} w^{(i)} \log L^c\left(\boldsymbol{\theta}\left|\mathcal{S},\mathbf{A}^{(i)}\right.\right) \quad (11)$$
$$\mathbf{A}^{(i)} \leftarrow g(\mathbf{A}^{(i)})$$

The Monte Carlo aspect of this process is described in Section 3.2 below. In each EM iteration, once $\hat{\mathcal{Q}}(\cdot|\cdot,\cdot)$ formed it must be maximised. This is done by first initialising $\boldsymbol{\theta}$ to the weighted average of each particle's $\hat{\boldsymbol{\theta}}$ and proceeding with simulated annealing. Then $\hat{\boldsymbol{\theta}}_{(m+1)}$ is set to the maximum value of $\boldsymbol{\theta}$, and the next EM iteration commences. The resulting full algorithm is summarised in Algorithm 1.

3.2 Monte Carlo Integration

As it is a non standard technique we very briefly review the Monte Carlo method here; further details are provided in [1, 18]. Monte Carlo integration is the use of computational techniques to generate a set of samples $X^{(j)}$ in order to approximate a probability measure $\pi(dx)$, with an empirical measure $\tilde{\pi}(dx)$. Under certain conditions the integral

$$\mathbb{E}_\pi[h(X)] = \int_{\mathcal{X}} h(x)\pi(dx), \quad (12)$$

may be approximated with

$$\bar{h}_N = \lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} h(x^{(j)}), \quad (13)$$

where $x^{(j)}$ are independently sampled from $\pi(x)$. Thus if one can sample $x^{(j)} \leftarrow \pi(x)$ and evaluate $h(x), \forall x \in \mathcal{X}$ one may perform this integral numerically.

3.2.1 Importance Sampling A particular method of sampling is to recast (12) as

$$\begin{aligned} \mathbb{E}_\pi[h(X)] &= \int_{\mathcal{X}} h(x)\pi(dx) \\ &= \int_{\mathcal{X}} h(x)\frac{\pi(x)}{g(x)}g(dx). \end{aligned} \quad (14)$$

It is apparent that by sampling $x^{(j)} \leftarrow g(x)$ (termed the importance distribution) and weighting by $w(x) \triangleq \frac{\pi(x)}{g(x)}$ (the importance

weights) one may approximate

$$\mathbb{E}_\pi[h(X)] \approx \bar{h}_N = \frac{1}{N}\sum_{j=1}^{N}\frac{\pi(x^{(j)})}{g(x^{(j)})}h(x^{(j)}) \quad (15)$$

provided that $\text{supp}(g) \supset \text{supp}(\pi)$. This technique is called importance sampling.

3.2.2 Sequential Importance Sampling (SIS) In high dimensional problems it is difficult to find a good importance distribution. Sequential importance sampling solves this problem by building samples from the trial distribution sequentially by dimension. Decompose $\mathbf{x}$ into its dimensions $\mathbf{x} = (x_1, \ldots, x_d)$ and let $\mathbf{x}_{1:k} = (x_1, \cdots, x_k)$, where $k \leq d$ indicate all the dimensions up to and including k. Then the importance distribution may be written

$$g(\mathbf{x}_{1:d}) = g_1(x_1)g_2(x_2|x_1)g_3(x_3|\mathbf{x}_{1:2})\cdots g_d(x_d|\mathbf{x}_{1:d-1}).$$

This allows the importance weight to be written

$$w_{d-1} = \frac{\pi(\mathbf{x}_{1:d-1})}{\prod_{k=1}^{d} g_k(x_k|\mathbf{x}_{1:k-1})}, \quad (16)$$

and hence to be calculated sequentially according to

$$w_k = w_{k-1}\frac{\pi(\mathbf{x}_{1:k})}{\pi(\mathbf{x}_{1:k-1})g_k(x_k|\mathbf{x}_{1:k-1})}. \quad (17)$$

3.3 Sequential Latent Variable Algorithm

Consider sequential importance sampling for inference in the case of latent variables. Following [17], let the dataset comprise observed and missing data $\mathbf{x} = (\mathbf{x}_{obs}, \mathbf{x}_{mis}) \sim f(\mathbf{x}_{obs}, \mathbf{x}_{mis}|\theta)$ with a parameter prior $f(\theta)$. The problem at hand is to maximise the posterior $f(\theta|\mathbf{x}_{obs})$. By marginalisation

$$f(\theta|\mathbf{x}_{obs}) = \int f(\theta|\mathbf{x}_{mis}, \mathbf{x}_{obs})f(\mathbf{x}_{mis}|\mathbf{x}_{obs})d\mathbf{x}_{mis} \quad (18)$$

This integral is a clear candidate for Monte Carlo solution with

$$\begin{aligned} f(\theta|\mathbf{x}_{mis}, \mathbf{x}_{obs}) &= h(\mathbf{x}) \\ f(\mathbf{x}_{mis}|\mathbf{x}_{obs}) &= \pi(\mathbf{x}). \end{aligned}$$

So by sampling $\mathbf{x}_{mis} \leftarrow f(\mathbf{x}_{mis}|\mathbf{x}_{obs})$ we may solve the integral (18), producing an empirical posterior distribution function. Given the dimensionality, dimension-wise sequential importance sampling is appropriate. If we choose the sampling function to be

$$g(x_{mis,k}) \triangleq f(x_{mis,k}|\mathbf{x}_{obs,1:k}, \mathbf{x}_{mis,1:k-1}), \quad (19)$$

then from equation (17) the weight recursion is solved as

$$\begin{aligned} w_k &= w_{k-1}\frac{\pi_k(\mathbf{x}_{1:k})}{\pi_{k-1}(\mathbf{x}_{1:k-1})g_k(x_k|\mathbf{x}_{1:k-1})} \\ &= w_{k-1}\frac{f\left(x_{obs,k}|\mathbf{x}_{mis,1:k-1}, \mathbf{x}_{obs,1:k-1}\right)}{f(x_{obs,k}|\mathbf{x}_{obs,1:k-1})} \\ &\propto w_{k-1}f\left(x_{obs,k}|\mathbf{x}_{mis,1:k-1}, \mathbf{x}_{obs,1:k-1}\right). \end{aligned}$$

Thus the weight is updated sequentially as

$$w_k = w_{k-1}f(x_{obs,k}|\mathbf{x}_{obs,k-1}, \mathbf{x}_{mis,k-1}). \quad (20)$$

3.4 Framing of TFBS-ID as a Sequential Latent Variable Problem

The TFBS-ID problem is clearly amenable to the sequential latent variable treatment outlined above. Beginning with the parameter estimation for a fixed motif model $\mathcal{M}_i$

$$p\left(\boldsymbol{\theta}_i|\mathcal{M}_i, \mathbf{S}\right) = \int p\left(\boldsymbol{\theta}_i|\mathcal{M}_i, \mathbf{S}, \mathbf{A}\right)p\left(\mathbf{A}|\mathcal{M}_i, \mathbf{S}\right)d\mathbf{A}. \quad (21)$$

By setting

$$\begin{aligned} p\left(\boldsymbol{\theta}_i|\mathcal{M}_i, \mathbf{S}, \mathbf{A}\right) &= f(\theta|\mathbf{y}_{mis}, \mathbf{y}_{obs}) = h(\mathbf{x}) \\ p\left(\mathbf{A}|\mathcal{M}_i, \mathbf{S}\right) &= f(\mathbf{y}_{mis}|\mathbf{y}_{obs}) = \pi(\mathbf{x}) \end{aligned}$$

we may invoke the latent variable framework, if we can sample $\mathbf{A} \leftarrow p\left(\mathbf{A}|\mathcal{M}_i, \mathbf{S}\right)$. Using the sequential framework we may sample from equation (19)

$$\begin{aligned} x_{mis,k} \leftarrow g(x_{mis,k}) &\triangleq f(x_{mis,k}|\mathbf{x}_{obs,1:k}, \mathbf{x}_{mis,1:k-1}) \\ &= f(\mathbf{A}_k|\mathbf{S}_{1:k}, \mathbf{A}_{1:k-1}), \end{aligned} \quad (22)$$

and evaluate the weighting recursion (20)

$$\begin{aligned} w_k &= w_{k-1}f(x_{obs,k}|\mathbf{x}_{obs,1:k-1}, \mathbf{x}_{mis,1:k-1}) \\ &= w_{k-1}f(\mathbf{S}_k|\mathbf{S}_{1:k-1}, \mathbf{A}_{1:k-1}). \end{aligned} \quad (23)$$

In this section we will derive these equations.

3.4.1 The Weighting Function Consider marginalising the weighting function developed as equation (23) with respect to the discrete latent variable,

$$\begin{aligned} w_k &= w_{k-1}p\left(\mathbf{S}_k|\mathbf{A}_{1:k-1}, \mathbf{S}_{1:k-1}\right) \quad (24) \\ &= w_{k-1}\int p(\mathbf{S}_k, \mathbf{A}_k|\mathbf{S}_{1:k-1}, \mathbf{A}_{1:k-1})d\mathbf{A}_k \\ &= w_{k-1}\sum_{\mathbf{A}_k \in \mathcal{A}_k} p(\mathbf{A}_k|\mathbf{S}_{1:k-1}, \mathbf{A}_{1:k-1})p(\mathbf{S}_k|\mathbf{S}_{1:k-1}, \mathbf{A}_{1:k}), \end{aligned}$$

where $\mathcal{A}_k$ is the set of all possible alignments in promoter k.

We assume that $\mathbf{A}_k$ is independent of $\mathbf{A}_{k-1}$ as there is no reason to believe that in any given collection of promoters, the transcription factors will bind in a related way between the promoters. Some transcription factors do display a position dependant binding with respect to the transcription start site; however this may be accounted for in the $p(\mathbf{A}_k)$ term in (29). Thus, in the absence of the promoter data $\mathbf{S}_k$, we may take the first term in equation (24) as flat and absorb it into proportionality.

$$w_k \propto w_{k-1}\sum_{\mathbf{A}_k \in \mathcal{A}_k} p(\mathbf{S}_k|\mathbf{S}_{1:k-1}, \mathbf{A}_{1:k}), \quad (25)$$

The term under summation may be solved by realising that the only information that $(\mathbf{S}_{1:k-1}, \mathbf{A}_{1:k-1})$ contributes to inference at dimension k is an estimate of the model parameters $\hat{\boldsymbol{\theta}}$. The maximum likelihood (ML), $\hat{\boldsymbol{\theta}}$ of (6) where $\mathbf{S} = \mathbf{S}_{\mathbf{A}_k}$ is

$$\hat{\theta}_{i,\phi} = \frac{|\{\mathbf{S}_{A_{k,i}} = \phi : A_k \in \mathbf{A}_{1:k}\}| + N_{(m)}\hat{\theta}_{(m),i,\phi}}{|\mathbf{A}_{1:k}| + N_{(m)}} \quad (26)$$

By substituting this into (8) we arrive at the likelihood function $P(S_k|\hat{\boldsymbol{\theta}}, A_k)$, which is substituted into equation (25) allowing us

to write the unnormalised weighting function as

$$w_k = w_{k-1} \sum_{\mathbf{A}_k \in \mathcal{A}_k} \frac{\prod_{\phi \in \Phi} \prod_{i=1}^{w} \hat{\theta}_{i,\phi}^{|\{\mathbf{S}_{\mathbf{A}_k}, i=\phi\}|}}{\prod_{\psi \in \Psi^T} \boldsymbol{\theta}_{0,\psi}^{|\{\mathbf{S}^T_{\mathbf{A}_k}=\psi\}|}}. \quad (27)$$

For simple models (a low number of motifs considered per promoter) this summation may be performed exhaustively, and this is the technique adopted in the proposed algorithm.

3.4.2 The Sampling Function Upon substitution of the model, the sampling distribution for the missing data in (22) becomes:

$$A_k \leftarrow g(A_k) = f(A_k|\mathbf{S}_{1:k}, \mathbf{A}_{1:k-1}). \quad (28)$$

Through Bayes' rule we can manipulate

$$\begin{aligned} g(A_k) &= f(A_k|\mathbf{S}_{1:k}, A_{1:k-1}) \\ &= \frac{f(\mathbf{S}_k|\mathbf{S}_{1:k-1}, A_{1:k}) f(A_k|\mathbf{S}_{1:k-1}, \mathbf{A}_{1:k-1})}{f(\mathbf{S}_k|\mathbf{S}_{1:k-1}, A_{1:k-1})} \\ &\propto f(A_k) f(\mathbf{S}_k|\mathbf{S}_{1:k-1}, A_{1:k}). \end{aligned} \quad (29)$$

The first term is the prior on the binding site, which can reasonably be taken as flat. The second term is again the likelihood function (8), thus

$$g(A_k) \propto \frac{\prod_{\Phi} \prod_{i=1}^{w} \hat{\theta}_{i,\phi}^{|\{\mathbf{S}_{\mathbf{A}_k}, i=\phi\}|}}{\prod_{\Psi^T} \boldsymbol{\theta}_{0,\psi}^{|\{\mathbf{S}^T_{\mathbf{A}_k}=\psi\}|}}. \quad (30)$$

As it is discrete, sampling this function is trivial through inversion sampling. Furthermore, it is computationally convenient, as it is required in the particle weighting (27) and hence need only be calculated once.

Algorithm 1: The SMC EM Algorithm

```
// Setup
```
Set for P particles: $\hat{\theta}_{(0)}^{(p)}, w_0^{(p)}, N_{(0)} = 0$
```
// EM Iterations
```
for $m \in [1, M]$ **do**
```
  // Iteration over particles
```
  for $p \in [1, P]$ **do**
```
    // Iteration over promoters
```
    for $k \in [1, |\mathbf{S}_k|]$ **do**
```
      // Sample latent var. with Eq. (28)
```
      $\mathbf{z}_k^{(p)} \triangleq \mathbf{A}_k \sim \frac{\prod_{\Phi} \prod_{i=1}^{w} \hat{\theta}_{i,\phi}^{|\{\mathbf{S}_{\mathbf{A}_k}, i=\phi\}|}}{\prod_{\Psi^T} \boldsymbol{\theta}_{0,\psi}^{|\{\mathbf{S}^T_{\mathbf{A}_k}=\psi\}|}}$
```
      // Weight particle using Eq. (27)
```
      $u_k^{(p)} \propto \sum_{\mathbf{A}_k \in \mathcal{A}_k} \frac{\prod_{\phi \in \Phi} \prod_{i=1}^{w} \hat{\theta}_{i,\phi}^{|\{\mathbf{S}_{\mathbf{A}_k}, i=\phi\}|}}{\prod_{\psi \in \Psi^T} \boldsymbol{\theta}_{0,\psi}^{|\{\mathbf{S}^T_{\mathbf{A}_k}=\psi\}|}}$
      $w_k^{(p)} = w_{k-1}^{(p)} u_k^{(p)}$
      Update $\hat{\theta}_{(m)}^{(p)}$
    Resample if neccessary
  Maximize
  $\hat{\theta}_m = \arg\max_\theta \sum_{i=1}^{P} w_K^{(i)} \log L^c \left(\boldsymbol{\theta} \middle| \mathcal{S}, \mathbf{A}_K^{(i)}\right)$
Return $\hat{\theta}_M$

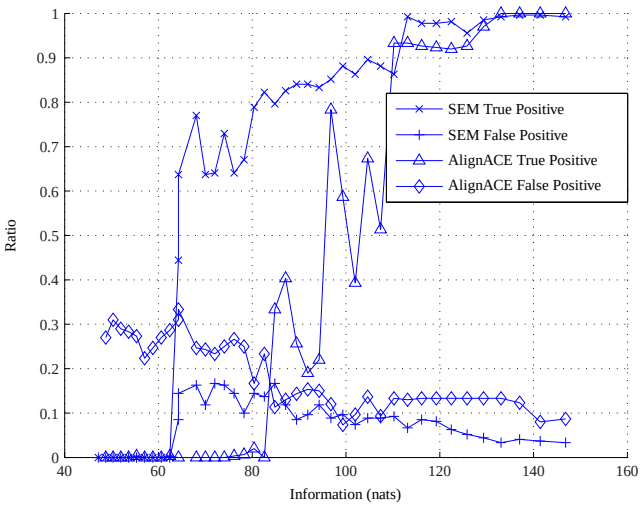

Fig. 1. TFBS Semi-Synthetic Data Results This graph is plotted with information contents (in nats) on the x-axis and performance metric on the y-axis. It demonstrates the superior performance of the SEM algorithm in terms of higher true positive ratio and lower false positive ratio, especially at low information content.

4 RESULTS AND DISCUSSION

4.1 Semi Synthetic

In order to observe the properties of the proposed algorithm in a controlled setting, we have constructed a semi-synthetic test-set. This allows the algorithm to be tested in an environment where the model assumptions are met, thus examining the algorithm, rather than the models. Furthermore, as the ground truth for this scenario is known it is possible to assign true performance measures to the algorithm.

The test-set consists of several subsets, each of which is a collection of artificial promoters, with a background of nucleotides generated according to human IID frequencies, into which are inserted one, or two real binding sites for CRP (catabolite receptor protein [24]) at random locations. Each subset of the test-set is differentiated by an artificial reduction of the information content of binding site. This is achieved by sequentially altering the binding site model until it is the background model. Thus the test-set consists of sets of promoters, each of which has a different information content.

We tested this semi-synthetic set with both our algorithm (SEM) and AlignAce [13] which makes identical model assumptions but is Gibbs sampler based. This is done as the thesis of this paper is that the Gibbs sampler should present inferior convergence to the true mode, due to multimodality. The results of this test are shown in Figure 1, which display the true positive and true negative ratios of the two algorithms as a function of the varying information content. Here the true and false counts are with respect to the nucleotides.

This graph highlights two important characteristics of the algorithms. The first concerns the robustness of the true positive ratio. At information content above approximately 120 nats (information content with logs taken to the natural number base) the algorithms

are indistinguishable. However, the AlignAce algorithm collapses sharply, such that by 80 nats it makes no true predictions, while the SEM algorithm is still 80% correct, and only falls to 0% true positive by 60 nats. At virtually all points the SEM algorithm has a higher true positive ratio than the Gibbs sampler.

The second important feature is the false positive ratio. Again, at virtually every point the SEM has a lower false positive ratio than AlignAce. More important is the failure mode of the algorithm. Below 80 nats AlignAce makes no true calls, yet continues to make false calls. SEM however fails into silence - when the true positive falls to zero, so does the false positive. This is because the algorithm will favour the case of no binding sites present, and hence will make no (true or false) predictions.

This is intentionally the simplest data case, in order to clearly demonstrate the superior ability of the proposed algorithm to find a solution that is known to exist. Although the data set is contrived, this example illustrates the superior performance of the proposed algorithm by controlling for obscuring effects that the more realistic tests to follow will introduce.

4.2 Full *E coli* K12 Set

A recent comparative study of motif finding algorithms [11] provides an interesting benchmark of competitive algorithms. Included in this study is a data set consisting of laboratory verified binding sites for 82 motifs from the *E coli* K12 organism. Each example of each motif is aligned, with a margin of the flanking nucleotides retained on each end. There are several sets where the length of the margin varies from 100 to 800 nucleotides. This provides a good test data set as the true position of the binding site in each sequence is known, yet the data are not synthetic.

4.2.1 Results The dataset is initially filtered to remove binding sites with fewer than two examples and to remove overlapping sequences of the genome. Each of the remaining sets has been run through the proposed algorithm and in each case five motifs have been requested. The results are reported for that with the highest specificity. The algorithm is tested with

Table 1 below reports the algorithmic performance on this data set at various margin lengths and with a background model of varying order. Sensitivity and specificity are as usual defined by

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \qquad \text{Specificity} = \frac{TP}{TP+FP}.$$

where TP, FN and FP are the nucleotide true positive, false negative and false positive counts. The performance of the comparative algorithms is reproduced from [11]. This table shows the favourable performance of the algorithm with respect to competing algorithms when the SMC algorithm uses a flat background model. The higher order models obtained from [6] display rapidly decreasing performance. This is unexpected and explored below. The effect of the background model is displayed in Figures 2 below, which displays the specificity of the algorithm with a 300 nucleotide margin with a flat, and 3^{rd} order Markov background for each transcription factor. In addition the percentage of A and T nucleotides in the true binding site is plotted, which we call the AT content henceforth.

The performance of the flat background model is unexpectedly superior to that of the 3^{rd} order Markov background. The reason, as figure 2 shows, is the AT content, which we now explain. There are several (mostly unknown) factors that allow a motif to be recognised

	Margin Length					
	100		300		500	
	Sn	Sp	Sn	Sp	Sn	Sp
AlignAce	0.20	0.22	0.14	0.17	0.10	0.12
BioProspector	0.35	0.28	0.26	0.21	0.22	0.16
MDScan	0.35	0.26	0.28	0.22	0.24	0.18
MEME	0.28	0.34	0.20	0.28	0.15	0.22
MotifSampler	0.32	0.24	0.18	0.14	0.15	0.08
SMC Flat	0.39	0.35	0.34	0.30	0.21	0.19
SMC O(0)	0.38	0.34	0.11	0.10	0.04	0.04
SMC O(3)	0.43	0.39	0.15	0.14	0.07	0.07

Table 1. Performance Summary This table compares the specificity (Sp) and sensitivity (Sn) across the set between existing algorithms and the proposed algorithm using 0^{th} and 3^{rd} order Markov background models

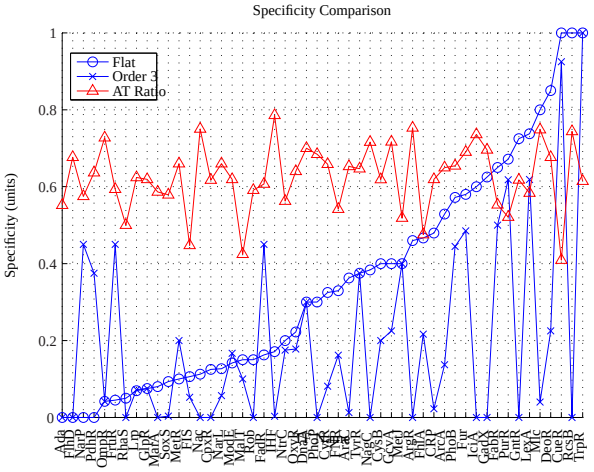

Fig. 2. Specificity Comparison This graph shows the specificity for each transcription factor individually, under two background modelsFlat and SMC O(3). It also plots the AT content of the true motif. It demonstrates the deleterious effect that high AT content exerts of specificity.

well by the algorithm. By ordering the transcription factors by specificity with 0^{th} order background we group them according to some such set of characteristics. Hence neighbouring transcription factors on the abscissa of this plot are in some sense similarly difficult to find according to this large set unknown set of factors. Plotting the specificity of the model with a 3^{rd} order background with the same ordering we see that above approximately 20% specificity the performance of the algorithm is similar for most transcription factors, except for being unable to find many. The final curve on this plot shows the AT content and it is seen that transcription factors with high AT content relative to nearby transcription factors are identically those that the algorithm is unable to find with a 3^{rd} order background.

This is because poly-A, poly-T sequences and poly-AT repeats occur often in intergenic regions. Hence these transitions probabilities are highly weighted in the 3^{rd} order background model that has been fitted to intergenic sequences. Consequently, when calculating the motif likelihood according to equation (8), true motifs with these patterns will score a low probability. This explains this behaviour.

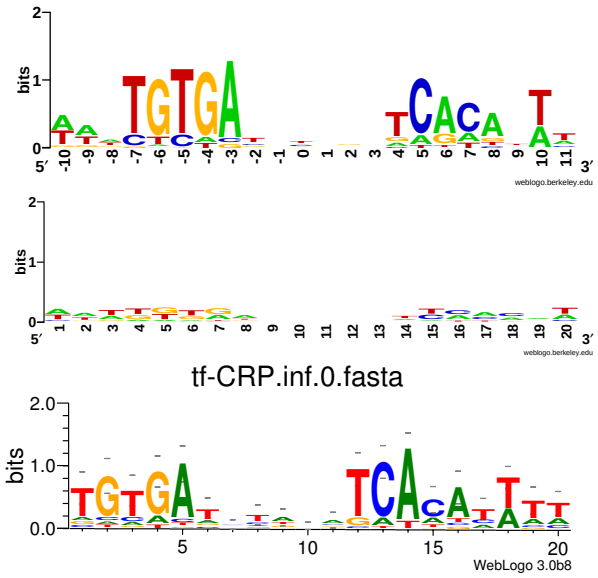

Fig. 3. CRP Logos. On the top is the reference CRP Motif. This is the motif logo generated from a different reference data set [19] of 49 binding sites for the CRP TF in *E. coli* K12. In the centre is the true motif in this test set. Although the same in form, it is clearly less conserved. The lower motif is that inferred by the algorithm. The algorithm is visibly only able to find well conserved motifs. Produced with WebLogo [7].

In addition, the signal strength of the logos is of interest. In the supplementary materials we provide the Kullback-Leibler divergence, or cross entropy between the empirical true- and inferred-motifs. We also plot the motif logos, which present the same information graphically, for CRP only below.

From these logos, which are a graphical representation of $\hat{\theta}$ several points may be made. Although CRP has a very well defined and clear reference logo in the literature, it is not sufficiently conserved in this data set for the logo to be very clear, Figure 3. As all the binding sites in this set represent actual binding sites, this speaks to the faintness of the sought after signal in practice.

5 CONCLUSION

We have identified a problem in the current algorithmic approaches to transcription factor binding site identification arising from the irregularity of the probability space. The effect of this problem is that single line of search algorithms, such as the Gibbs' sampler, may easily become trapped in local modes, leading to poor performance. To overcome this problem we have proposed a batch setting SMC EM algorithm that benefits from having many independent, parallel searches, which by covering the alignment space more comprehensively return higher performance. We have demonstrated this increased performance with both semi-synthetic as well as true sequence data from *E. coli*. A strong observation from this work is that the main problem in inference problems in this field is the modelling. Specifically, there is difficulty in defining and thus fitting the background model, and hence inference in general is severely inhibited. The successful extension of this TFBS identification to higher organisms, such as mammalians, will require the solution to this problem.

6 ACKNOWLEDGEMENTS

The authors thank Mr Adam Johansen for his frequent discussions and corrections, and Dr Arnaud Doucet for his valuable input.

REFERENCES

[1]C. Andrieu. Monte Carlo Methods for Absolute Beginners. Lecture Notes in Computer Science, pages 113–145, 2004.

[2]C. Andrieu and A. Doucet. Online expectation-maximization type algorithms for parameter estimation in general state space models. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2003.

[3]T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International conference on intelligent systems for molecular biology, 1994.

[4]G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association, 95, 2000.

[5]N. Chopin. A sequential particle filter method for static models. Biometrika, 89:539–552, 2002.

[6]B. Coessens, G. Thijs, S. Aerts, K. Marchal, F. De Smet, K. Engelen, P. Glenisson, Y. Moreau, J. Mathys, and B. De Moor. INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. Nucleic Acids Research, 31(13):3468–3470, 2003.

[7]G.E. Crooks, G. Hon, J.M. Chandonia, and S.E. Brenner. WebLogo: A Sequence Logo Generator. Genome Research, 14:1188–1190, 2004.

[8]A.P. Dempster, N.M Laird, and D.B Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39:1–38, 1977.

[9]T.A. Down and T.J.P. Hubbard. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Research, 33(5):1445–1453, 2005.

[10]C.J. Geyer. Estimation and optimization of functions. In Markov Chain Monte Carlo in Practice, 1996.

[11]J. Hu, B. Li, and D. Kihara. Limitations and potentials of current motif discovery algorithms. Bioinformatics, 33, 2005.

[12]J.D Hughes, P.W. Estep, S. Tavazoie, and G.M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. Journal of Molecular Biology, 296(5):1205–14, 2000.

[13]J.D. Hughes, P.W. 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, 296(5):1205–1214, 2000.

[14]S.T. Jensen, X.S. Liu, Q. Zhou, and J.S. Liu. Computational discovery of gene regulatory binding motifs: A Bayesian perspective. Statistical Science, 19:188–204, 2004.

[15]T. Koski. Hidden Markov Models for Bioinformatics. Kluwer Academic Publishers, 2001.

[16]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, 262:208–214, 1993.

[17]J.S. Liu. Monte Carlo Strategies in Scientific Computing. Springer, 2001.

[18]C.P. Robert and G. Casella. Monte Carlo statistical Methods. Springer, 2004.

[19]K. Robison, A.M. Mcguire, and G.M. Church. A Comprehensive Library of DNA-binding Site Matrices for 55 Proteins Applied to the Complete Escherichia coli K-12 Genome. Journal of Molecular Biology, 284(2):241–254, 1998.

[20]D.B. Rubin. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. Journal of the American Statistical Association, 82, 1987.

[21]M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82:528–550, 1987.

[22]W. Thompson, E. Roucha, and C.E Lawrence. Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Research, 31(13):3580–3585, 2003.

[23]J.D. Watson, T.A. Baker, S.P. Bell, A. Gann, M. Levine, and R. Losick. Molecular Biology of the Gene. Cold Spring Harbour Laboratory Press, 2004.

[24]E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüß, I. Reuter, and F. Schacherer. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Research, 28(1):316–319, 2000.