Sequence analysis

A sequential Monte Carlo EM approach to the transcription factor binding site identification problem

Edmund S. Jackson* and William J. Fitzgerald

Signal Processing Laboratory, Department of Engineering, Cambridge University, UK

Received on October 9, 2006; revised on January 21, 2007; accepted on February 9, 2007

Advance Access publication March 25, 2007

Associate Editor: 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 article, we derive and demonstrate a novel method utilizing a sequential Monte Carlo-based expectation-maximization (EM) optimization 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 Escherichia coli.

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

Contact: ej230@cam.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The fundamental element in the control of gene expression is a class of proteins called transcription factors (TFs) (Watson et al., 2004). 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 (TFBSs) 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 (Hu et al., 2005; Jensen et al., 2004). 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-maximization (EM) (Dempster et al., 1977) (such as MEME (Bailey and Elkan, 1994)), as well as EM’s stochastic counterpart, data augmentation (Tanner and Wong, 1987), (such as the ‘Gibbs Sampler’ method (Hughes et al., 2000; Lawrence et al., 1993)) are popular.

We wish to infer the model parameters, being the motif-binding parameters (often represented as a ‘logo’ such as in Fig. 3). The dimension of this parameter space is 4 · L, where L is the motif length, and often >20. This may certainly be considered high dimensional, which is widely known to complicate inference.

A more serious concern, and the focus of this article, is the multimodality 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’ (Celeux et al., 2000). 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’ (Robert and Casella, 2004).
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 (Andrieu and Doucet 2003). 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 Down and Hubbard (2005), who propose an alternative approach.

1.2 Development of the approach

In the statistics literature, SIR has long been shown useful for data augmentation (Rubin, 1987); however, there the sampling distribution is fixed. Given its importance for convergence, (Geyer, 1996) proposes a dynamic sampling distribution. They present a stochastic EM method with multiple parsing of the data set, optimizing 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 Andrieu and Doucet (2003), 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, Chopin (2002) 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 of this pervasive model.

2.1 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 utilized for eukaryotes. Recently switching Markov model based methods have been introduced (Down and Hubbard, 2005; Thompson, 2003) but these are beyond our scope.

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

\[
\Psi^T = \Phi^T \otimes \Phi.
\]

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

\[
\theta_0 \triangleq \{\theta_{0,\psi}\}_{\psi \in \Psi^T}.
\]

Let the set of such transitions observed in the string \( S \) be

\[
S^T \triangleq \{\psi \in \Psi^T : \psi \subset S\}.
\]

Using the notation \( S_i_{t+T-1} \rightarrow S_{i+T} \) to denote a transition of a \( T \)th order Markov chain, the probability of a sequence, \( S \), is given by

\[
P(S | \theta_0) = \prod_{i=1}^{N} P(S_i | S_{i\max(0, i-T), i-1}, \theta_0)
\]

\[
\triangleq \prod_{i=1}^{N} \theta_{0,S_{i\max(0, i-T), i-1} \rightarrow S_i} \\
\approx \prod_{i=1}^{N} \theta_{0,S_{i-T+1} \rightarrow S_i}
\]

where \( S_{0:0} \triangleq \emptyset \). The terms in \( S_i, i < T \) 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

\[
P(S | \theta_0) \propto \prod_{i \in T} \theta_{0, S_{i-1} \rightarrow S_i} \prod_{\psi \in \Psi^T} \theta_{0, \psi}^{S^T=\psi}
\]

\[
\approx \prod_{\psi \in \Psi^T} \theta_{0, \psi}^{S^T=\psi},
\]

where, the cardinality of the set \( S^T = \psi \) is simply the number of observations of the transition \( \psi \) in the observed sequence.

The probabilities, \( \theta_0 \), are the parameters of this model, and are found empirically on a per-organism basis. Fitting the maximum-likelihood (ML) estimation (Koski, 2001) on the full set of promoter regions of the organism of interest results in

\[
\hat{\theta}_0, S_{i+T-1} \rightarrow S_i = \frac{|\{S_{i+T-1} \otimes s \subset S\}|}{|\{S_{i+T-1} \subset S\}|}
\]

and hence the set of \( T \)th order transition probabilities, \( \theta_{0,\psi} \).

2.2 Motif model

The binding-site motif is modelled as an independent, non-identically distributed multinomial distribution.
Thus, a motif of width $w$, $S = S_{[w]} = [S_1, \ldots, S_w] \in \Phi^w$ has likelihood
\[
P(S|\theta) = \prod_{i=1}^{w} P(S_i|\theta)
= \prod_{i=1}^{w} \theta_{S_i,i}
\]
where, $\theta$ is termed the ‘position weight matrix’ and defined as
\[
\theta = [\theta_{i,j}]_{i=1,...,w; j \in \Phi}
\]
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, $S = \{S_m\}_{m \in [1, M]}$, is found by exchangeability to be
\[
P(S|\theta) = \prod_{m=1}^{M} \prod_{i=1}^{w} \theta_{i,S_m,i}
= \prod_{\phi \in \Phi} \prod_{m=1}^{M} [\theta_{S_m,\phi,m \in [1, M]}]
\]
Similarly to (3) the ML solution of the parameters is
\[
\hat{\theta}_{i,\phi} = \left\{ S_{m,i} = \phi : m \in [1, M] \right\}
\]

### 2.3 Mixture model

Let $S = \{S_1, \ldots, S_K\}$ be a promoter set of $K$ promoters, each of which is $N$ nucleotides in length, $S_k = [S_1, \ldots, S_N] \in \Phi^N$. Consider $S$ as a single sequence modelled as a Markov chain with transition probabilities $\theta_0$ with a number of motif instances distributed according to an independent multinomial model with parameter $\theta$, inserted randomly at positions $A = \{A_{k,m}\}$ 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
\[
P(S|\theta_0, A, S) \approx P(S_{\setminus A}|\theta_0)P(S_{\setminus A}|\theta_0) \propto P(S_{\setminus A}|\theta_0)P(S_{\setminus A}|\theta_0)
\]
where subscripts indicate sub-strings; $S_{\setminus A}$ is the sequence with the motif instances at $A$ removed and $S_{\setminus A}$ 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
\[
S_{A_k,i} = \left\{ S_{k, A_{k,m},i} : m \in [1, M_k] \right\} \in \Phi^{M_k}
\]
and the set of all motif instances in promoter $k$ as
\[
S_{A_k} = \left\{ S_{A_{k,i}} : i \in [1, w] \right\} \in (\Phi^w)^{M_k}
\]
Substituting the models (1) and (5) into (7), we arrive at the mixture likelihood function
\[
P(S|\theta_0, A, \Theta) \propto \prod_{\phi \in \Phi} \prod_{i=1}^{w} \prod_{m=1}^{M} \theta_{S_{m,i} = \phi, m \in [1, M]}]
\]

### 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, $\hat{\theta}$, of a repeated TFBS motif supposed within a set of sequences $S$. This is achieved by inferring the latent variable, $A$, the set of motif instance positions, and calculating $\hat{\theta}$ 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
  \[
  Q(\theta | \hat{\theta}_{(t)}, S) = \mathbb{E}_{\hat{\theta}_{(t)}}[\log L'(\theta | S, \Theta)]
  \]
  \[
  = \int_{\Theta} \log f(S | \Theta) f(\Theta | \hat{\theta}_{(t)}, S) d\Theta
  \]
- **Maximization (M):** with respect to $\theta$
  \[
  \hat{\theta}_{(t+1)} = \arg \max_{\theta} Q(\theta | \hat{\theta}_{(t)}, S)
  \]
  which are iterated until $\hat{\theta}$ converges.

For the $E$ step, we propose a sequential Monte Carlo (SMC) method, using the natural segmentation of $S$ into promoters $S_k$ and hence sampling via SIS/R. We approximate (9) by
\[
\hat{Q}(\theta | \hat{\theta}_{(t)}, S) \propto \sum_{i=1}^{w} \log L'(\theta | S, A^{(i)})
\]
\[
A^{(i)} \leftarrow g(A^{(i)})
\]
The Monte Carlo aspect of this process is described in Section 3.2 below. In each EM iteration, once $\hat{Q}(\cdot, \cdot, \cdot)$ formed it must be maximized. This is done by first initializing $\theta$ to the weighted average of each particle’s $\hat{\theta}$ and proceeding with simulated annealing. Then $\hat{\theta}_{(t+1)}$ is set to the maximum value of $\theta$, and the next EM iteration commences. The resulting full algorithm is summarized 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 (Andrieu, 2004; Robert and Casella, 2004). Monte Carlo integration is the use of computational techniques to generate a set of samples $X^{(i)}$ in order to approximate a probability measure $\pi(dx)$, with an empirical measure $\hat{\pi}(dx)$. Under certain conditions the integral
\[
\mathbb{E}_{\pi}[h(X)] = \int h(x) \pi(dx),
\]
importance distribution) and weighting by \( w_k = \frac{\pi(x_k)}{\pi(x_{k-1})} g_k(x_{k-1} | x_{k-1}) \) sequentially by dimension. Decompose the problem by building samples from the trial distribution sequentially by dimension. This allows the importance weight to be written

\[
g(x_{1:k}) = g_1(x_1)g_2(x_2 | x_1)g_3(x_3 | x_1, x_2) \cdots g_d(x_d | x_{1:d-1}).
\]

3.3 Sequential latent variable algorithm

Consider sequential importance sampling for inference in the case of latent variables. Following (Liu, 2001), let the data set comprise observed and missing data \( x = (x_{obs}, x_{mis}) \sim f(x_{obs}, x_{mis} | \theta) \) with a parameter prior \( f(\theta) \). The problem at hand is to maximize the posterior \( f(\theta | x_{obs}) \).

By marginalization

\[
f(\theta | x_{obs}) = \int f(\theta | x_{mis}, x_{obs}) f(x_{mis} | x_{obs}) dx_{mis}.
\]

This integral is a clear candidate for Monte Carlo solution with

\[
\begin{align*}
    f(\theta | x_{mis}, x_{obs}) &= h(x) \\
    f(x_{mis} | x_{obs}) &= \pi(x).
\end{align*}
\]

So by sampling \( x_{mis} \leftarrow f(x_{mis} | x_{obs}) \) 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

\[
g(x_{mis,k}) \overset{\Delta}{=} f(x_{mis,k} | x_{obs,k-1}) \pi(x_{mis,k} | x_{obs,k-1})
\]

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

\[
w_k = w_{k-1} f(x_{mis,k} | x_{obs,k-1}) g_k(x_{k-1} | x_{k-1}) / f(x_{mis,k-1} | x_{obs,k-1})
\]

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 \( M_i \)

\[
p(\theta | M_i, S) = \int p(\theta | M_i, S, A)p(A | M_i, S)dA.
\]
By setting
\[ p(\theta | M_k, S, A) = f(\theta | y_{\text{mis}}, y_{\text{obs}}) = h(x) \]
\[ p(A | M_k, S) = f(y_{\text{mis}} | y_{\text{obs}}) = \pi(x) \]
we may invoke the latent variable framework, if we can sample
\[ A \leftarrow p(A|M_k, S). \]
Using the sequential framework we may sample from Equation (19)
\[ x_{\text{mis}, k} \leftarrow g(x_{\text{mis}, k}) = f(x_{\text{mis}, k} | x_{\text{obs}, 1:k-1}, x_{\text{mis}, 1:k-1}) = f(A_k | S_{1:k-1}, A_{1:k-1}), \tag{22} \]
and evaluate the weighting recursion (20)
\[ w_k = w_{k-1} f(x_{\text{mis}, k} | x_{\text{obs}, 1:k-1}, x_{\text{mis}, 1:k-1}) = w_{k-1} f(S_k | S_{1:k-1}, A_{1:k-1}). \tag{23} \]
In this section we will derive these equations.

3.4.1 The weighting function Consider, marginalizing the weighting function developed as Equation (23) with respect to the discrete latent variable,
\[ w_k = w_{k-1} \sum_{A_k} p(S_k | A_{1:k-1}, S_{1:k-1}) = w_{k-1} \int p(S_k, A_k | S_{1:k-1}, A_{1:k-1}) \, dA_k \tag{24} \]
\[ = w_{k-1} \sum_{A_k \in A_k} p(A_k | S_{1:k-1}, A_{1:k-1}) p(S_k | S_{1:k-1}, A_{1:k-1}), \]
where, \( A_k \) is the set of all possible alignments in promoter \( k \).
We assume that \( A_k \) is independent of \( A_{k-1} \) as there is no reason to believe that in any given collection of promoters, the TFs will bind in a related way between the promoters. Some TFs do display a position-dependent binding with respect to the transcription start site; however, this may be accounted for in the \( p(A_k) \) term in (29). Thus, in the absence of the promoter data \( S_k \), we may take the first term in Equation (24) as flat and absorb it into proportionality.
\[ w_k \propto w_{k-1} \sum_{A_k \in A_k} p(S_k | S_{1:k-1}, A_{1:k-1}), \tag{25} \]
The term under summation may be solved by realizing that the only information that \( (S_{1:k-1}, A_{1:k-1}) \) contributes to inference at dimension \( k \) is an estimate of the model parameters \( \theta \). The ML, \( \hat{\theta} \) of (6) where \( S = S_{1:k} \) is
\[ \hat{\theta}_{\phi} = \frac{|\{S_{A_k} : A_k \in A_k \}| + N_{(m)} \hat{\theta}_{\phi}(m), \, \phi}{|A_k| + N_{(m)}} \tag{26} \]
By substituting this into (8) we arrive at the likelihood function \( L(S_k | \Theta, A_k) \), which is substituted into Equation (25) allowing us to write the unnormalized weighting function as
\[ w_k = w_{k-1} \sum_{A_k \in A_k} \prod_{\phi \in \Phi} \prod_{i \in \phi} \theta_{\phi, i}^{[S_{A_k}, i=\phi]} / \prod_{\psi \in \Psi} \theta_{\psi}^{[S_{A_k, i=\psi}]} \tag{27} \]
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:
\[ A_k \leftarrow g(A_k) = f(A_k | S_{1:k}, A_{1:k-1}). \tag{28} \]
Through Bayes’ rule we can manipulate
\[ g(A_k) = f(A_k | S_{1:k}, A_{1:k-1}) = f(S_k | S_{1:k-1}, A_{1:k-1}) / f(S_k | S_{1:k-1}, A_{1:k}), \tag{29} \]
\[ \propto f(A_k) / f(S_k | S_{1:k-1}, A_{1:k}). \]
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
\[ g(A_k) \propto \prod_{\phi} \prod_{i \in \phi} \theta_{\phi, i}^{[S_{A_k}, i=\phi]} / \prod_{\psi \in \Psi} \theta_{\psi}^{[S_{A_k, i=\psi}]} \tag{30} \]
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 to be calculated once.

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 IBD frequencies, into which are inserted one, or two real binding sites for CRP (catabolite receptor protein (Wingender et al., 2000)) 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 (Hughes et al., 2000), which makes identical model assumptions but is Gibbs sampler based. This is done as the thesis of this article is that the Gibbs sampler should present inferior convergence to the true model 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 \( \sim 120 \) nats
A recent comparative study of motif-finding algorithms (Hu et al., 2005) 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 Full E.coli K12 set

A recent comparative study of motif-finding algorithms (Hu et al., 2005) 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.
Table 1. Performance Summary. This table compares the specificity (Sp) and sensitivity (Sn) across the set between existing algorithms, and the proposed algorithm using zeroth- and third-order Markov background models.

<table>
<thead>
<tr>
<th>Margin length</th>
<th>Sn</th>
<th>Sp</th>
<th>Sn</th>
<th>Sp</th>
<th>Sn</th>
<th>Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.20</td>
<td>0.22</td>
<td>0.14</td>
<td>0.17</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>300</td>
<td>0.35</td>
<td>0.28</td>
<td>0.26</td>
<td>0.21</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td>500</td>
<td>0.35</td>
<td>0.26</td>
<td>0.28</td>
<td>0.22</td>
<td>0.24</td>
<td>0.18</td>
</tr>
<tr>
<td>AlignAce</td>
<td>0.28</td>
<td>0.34</td>
<td>0.20</td>
<td>0.28</td>
<td>0.15</td>
<td>0.22</td>
</tr>
<tr>
<td>BioProspector</td>
<td>0.32</td>
<td>0.24</td>
<td>0.18</td>
<td>0.14</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>MEME</td>
<td>0.39</td>
<td>0.35</td>
<td>0.34</td>
<td>0.30</td>
<td>0.21</td>
<td>0.19</td>
</tr>
<tr>
<td>MotifSampler</td>
<td>0.38</td>
<td>0.34</td>
<td>0.11</td>
<td>0.10</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>SMC O(0)</td>
<td>0.43</td>
<td>0.39</td>
<td>0.15</td>
<td>0.14</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>SMC O(3)</td>
<td>0.38</td>
<td>0.34</td>
<td>0.11</td>
<td>0.10</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Fig. 2. Specificity comparison. This graph shows the specificity for each TF individually, under two background models Flat 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.

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 θ, 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 TFBS 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. 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 mammals, will require the solution to this problem.

ACKNOWLEDGEMENTS

The authors thank Mr Adam Johansen for his frequent discussions and corrections, and Dr Arnaud Doucet for his valuable input. Thanks Trinity College and the Cambridge Commonwealth Trust for funding.

Conflict of Interest: none declared.

REFERENCES


