University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Supervised Detection of Regulatory Motifs in DNA Sequences

Sunduz Keles[*]   Mark J. van der Laan[†]   Sandrine Dudoit[‡]

Biao Xing[**]   Michael B. Eisen[††]

[*]Division of Biostatistics, School of Public Health, University of California, Berkeley , keles@stat.wisc.edu

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley

[‡]Division of Biostatistics, School of Public Health, University of California, Berkeley

[**]Division of Biostatistics, School of Public Health, University of California, Berkeley

[††]Dept. of Molecular and Cell Biology, UC Berkeley and Life Sciences Division, Ernest Orlando Lawrence Berkeley National Lab

Supervised Detection of Regulatory Motifs in DNA Sequences

Sunduz Keles, Mark J. van der Laan, Sandrine Dudoit, Biao Xing, and Michael B. Eisen

Abstract

Identification of transcription factor binding sites (regulatory motifs) is a major interest in contemporary biology. We propose a new likelihood based method, COMODE, for identifying structural motifs in DNA sequences. Commonly used methods (e.g. MEME, Gibbs sampler) model binding sites as families of sequences described by a position weight matrix (PWM) and identify PWMs that maximize the likelihood of observed sequence data under a simple multinomial mixture model. This model assumes that the positions of the PWM correspond to independent multinomial distributions with four cell probabilities. We address supervising the search for DNA binding sites using the information derived from structural characteristics of protein-DNA interactions. We extend the simple multinomial mixture model by incorporating constraints on the information content profiles or on specific parameters of the motif PWMs. The parameters of this extended model are estimated by maximum likelihood using a nonlinear constraint optimization method. Likelihood-based cross-validation is used to select model parameters such as motif width and constraint type. The performance of COMODE is compared with existing motif detection methods on simulated data that incorporate real motif examples from Saccharomyces cerevisiae. The proposed method is especially effective when the motif of interest appears as a weak signal in the data. Some of the transcription factor binding data of Lee et al. (2002) were also analyzed using COMODE and biologically verified sites were identified.

1 Introduction

Transcription factors control the expression of specific genes by binding to short unique families of sequences - referred to here as *binding sites* or *motifs*. Description of these binding sites is critical in understanding the biological content of genome sequences, and is an important problem in contemporary computational biology. Many methods have been developed for this purpose (reviewed in Stormo (2000)). Here, we focus on methods that use a two-component multinomial mixture model (first introduced by Lawrence & Reilly (1990)) for the description of nucleotides (A, C, G, T) in a DNA sequence. In this model, one component corresponds to bases in the background sequences and the second component corresponds to bases in the motif. In such methods (e.g. Lawrence & Reilly (1990); MEME of Bailey & Elkan (1995a); the Gibbs sampler of Lawrence et al. (1993); Hertz & Stormo (1999); Tavazoie et al. (1999); Hughes et al. (2000); Liu et al. (2001)), transcription factor binding sites are generally represented by a *position weight matrix (PWM)*- a $4 \times W$ matrix where position (j, w) represents the frequency of observing nucleotide j at position w of the DNA motif.

PWMs are attractive for use in modeling motifs because they are mathematically and computationally simple to handle. In essence, each column in a PWM corresponds to a multinomial distribution with four cell probabilities that represent the nucleotides {A,C,G,T}. A PWM can be easily estimated from a set of aligned sequences by simply counting the occurrences of each nucleotide at each position in the aligned sequences. Methods such as those listed above can be used to infer PWMs from unaligned sequences. There is also a straightforward relationship between the score of a given sequence against a particular PWM (the product of the probabilities of the observed nucleotides at each position of the motif) and the binding energy of the associated protein bound to the sequence (Berg & von Hippel, 1987). Although there are some obvious limitations in PWMs, such as the inability to consider interactions among positions in computing the likelihood that a particular sequence is bound by a transcription factor, they have proven to be very effective in describing the families of sequences bound by a given transcription factor and have considerable predictive power (Stormo, 2000).

In addition to the base specificities represented by the matrix values, PWMs can also be summarized by the *information content profile* (Schneider et al., 1986). The information content at a position w is given by

$$IC(w) = \log_2 J + \sum_{j=1}^{J} p_{wj} \log_2 p_{wj} = \log_2 J - Entropy(w),$$

and can be thought as a measure of how conserved position w is. The information content is measured in bits and for $J = 4$, it takes on values between 0 and 2 bits,
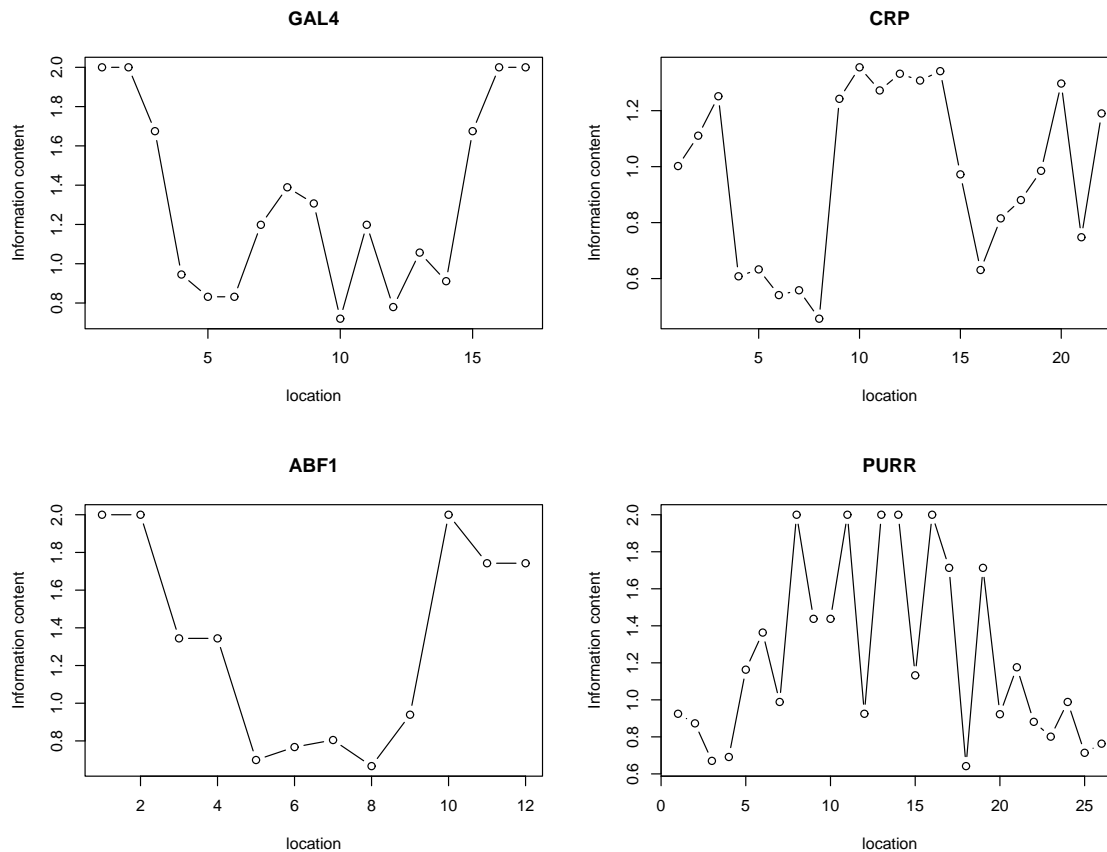
Figure 1: *Examples of information content profiles.* Information content across positions of GAL4 and ABF1 binding sites in *Saccharomyces cerevisiae* and PURR and CRP binding sites in *Escherichia coli.*

which correspond to a random site (all nucleotides have equal probability at that position) and to a deterministic site (probability of one of the nucleotides at that position is 1 and the rest are 0), respectively. Examples of information content profiles for several transcription factors are shown in Figure 1. For our purposes, it is important to note that PWMs describing factors with very different base specificities can have similar information content profile - see, for example, GAL4 and ABF1 in Figure 1. A recent paper by Mirny & Gelfand (2002) motivates further consideration of information content profiles. Using proteins for which both a three-dimensional structure of the protein bound to DNA and a family of experimentally verified bound sequences were available, these authors showed that the information content at each position in the motif computed from the bound sequences is proportional to the number of contacts between the protein and that base pair observed in the crystal structure. Or, more simply put, there is a direct relationship between the structural footprint of a transcription factor on DNA and the information content profile of

2

the corresponding motif. An important corollary of this intuitively sensible (and we believe general) observation is that the motifs bound by proteins with structurally similar DNA binding domains should have similar information content profiles, since structurally related transcription factors generally have similar structural footprints on DNA (Eisen, 2003).

In this paper, we develop a new motif detection method, COMODE, that builds on this notion of family-specific information content profile. COMODE stands for **CO**nstrained **MO**tif **DE**tection. Specifically, our method performs supervised searches for motifs whose information content profiles are constrained to match a particular user-specified profile or family of profiles. The proposed framework is quite general in the sense that it allows any type of constraints on motif PWMs. Some examples include enforcing nucleotide biases, or high versus low information contents at various positions. The underlying statistical model is an extension of the standard two-component multinomial mixture model, and the parameters of the model, including the PWM and parameters of the information content profile, are estimated with maximum likelihood, either using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) or Sequential Quadratic Programming methods (SQP) (see Bazaraa et al. (1979) for an overview). Simulations are performed to compare the performance of the proposed method COMODE with simple unconstrained motif methods that are building blocks of the popular motif detecting methods such as MEME (Bailey & Elkan, 1995a). The results illustrate many advantages of CO-MODE including detection of weak but structured motifs, robustness against model misspecification, and high small sample size relative efficiency. We have also analyzed some of the ChIP (chromatin immunoprecipitation) data by Lee et al. (2002). In this work, the unsupervised motif finding method MEME (Bailey & Elkan, 1995a) was used to analyze sequences that show evidence of binding by specific transcription factors. By our supervised approach, we were able to identify some of experimentally verified sites that MEME missed. In addition, we discuss extensions of COMODE that adapt to the true nature of biological datasets such as multiple motif occurences and illustrate these on two *cis*-regulatory regions from *Drosophila*.

The paper is organized as follows: Section 2 reviews the basic multinomial mixture model for sequence data and Section 3 introduces the constrained motif model and also addresses parameter estimation and model selection in this model. Results from simulation studies and data analysis are presented in Sections 4 and 5. We conclude with a brief discussion of our method.

2 Motif finding using mixture models: *oops* and *zoops*

We propose two constrained motif models for sequence data that employ constraints on the information content profile or individual positions of the motif PWM. The first model is based on the basic one motif per sequence (oops) model introduced by Lawrence & Reilly (1990). The latter is based on the zero or one motif per sequence (zoops) model of Bailey & Elkan (1994) and is an extension of the first model. In this section, we give a brief review of the two basic unconstrained multinomial mixture models and then we extend them to allow constraints on the motif PWM in the following section.

2.1 One motif per sequence model: *oops*

Let $X_i = \{X_{i,k}\}_{k=1}^{L_i}$ denote the data on the i-th sequence with length L_i and let $X_{i,k} \in \{A, C, G, T\}$ denote the base pair at the k-th position of the i-th sequence. The observed data will be represented by N i.i.d. random variables $\{X_1, \cdots, X_N\}$. Both oops and zoops models assume that sequence data come from a two component multinomial mixture model. The first component is the background model, which assumes that nucleotides at sites that do not contribute to the motif, hence fall into the background, are independent and identically distributed. The second component is the motif model and assumes that each nucleotide position of the motif is independent but not identical, thus each position comes from a different multinomial distribution. Let J be the number of letters in the sequence alphabet (i.e. $J = 4$ for DNA as used here, and $J = 20$ for protein sequences). We will denote parameters of the multinomial background distribution by $\overrightarrow{P}_0 = (p_{01}, \cdots, p_{0J})$ and the parameters of the multiniomial distributions of motif positions by $\mathcal{P}_W = (\overrightarrow{P}_1, \cdots, \overrightarrow{P}_W)$ where $\overrightarrow{P}_w = (p_{w1}, \cdots, p_{wJ})$ and W represents the width of the motif. We will use $\mathcal{P}$ to denote the set of all parameters in the model, i.e., $\mathcal{P}$ includes $\overrightarrow{P}_0$ and $\mathcal{P}_W$.

Let $Y_i = \{Y_{i,l}\}_{l=1}^{L_i}$ be the set of indicator variables representing the start site of the motif in sequence i. We have that $\sum_l^{L_i} Y_{i,l} = 1, \forall i$ in the oops model. Here, Y_i is a hidden random variable. The start sites are assumed to be uniformly distributed, i.e., $Pr(Y_{i,l} = 1) = 1/(L_i - W + 1), l \in \{1, \cdots, L_i - W + 1\}$. The conditional likelihood of sequence i given the hidden variables is as follows:

oops:

$$Pr(X_i \mid Y_{i,l} = 1, \mathcal{P}) = \prod_{k \in T_i^l} \prod_{j=1}^{J} p_{0j}^{I(X_{i,k}=j)} \prod_{w=1}^{W} \prod_{j=1}^{J} p_{wj}^{I(X_{i,w+l-1}=j)},$$

4

where $T_i^l = \{1, \cdots, L_i\} - \{l, l+1, \cdots, l+W-1\}$ denotes the background sites and $I(.)$ is the indicator variable.

Since the random variable Y is unobserved, maximum likelihood estimation can be done by maximizing the expectation of the full data log likelihood given the observed data with the EM algorithm (Dempster et al., 1977). Note that the full data for the oops model is $(\mathcal{X}, \mathcal{Y}) \equiv \{(X_i, Y_i), i = \{1, \cdots, N\}\}$. Let $\mathcal{P}^r$ be the parameter estimates after r-th EM iteration, then expected full data log likelihood given the observed data is as follows up to a constant:

$$\mathcal{Q}(\mathcal{P} \mid \mathcal{P}^r)$$
$$= \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E[I(Y_{i,l} = 1) \mid X_i, \mathcal{P}^r] \log Pr(X_i \mid Y_{i,l} = 1, \mathcal{P})$$
$$= \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E[I(Y_{i,l} = 1) \mid X_i, \mathcal{P}^r] \left\{ \sum_{k \in T_i^l} \sum_{j=1}^{J} I(X_{i,k} = j) \log p_{0j} + \sum_{w=1}^{W} \sum_{j=1}^{J} I(X_{i,w+l-1} = j) \log p_{wj} \right\}$$

Define

$$N_{wj} = \begin{cases} \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E[I(Y_{i,l} = 1) \mid X_i, \mathcal{P}^r] I(X_{i,(l+w-1)} = j), & w \neq 0, \\ \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E[I(Y_{i,l} = 1) \mid X_i, \mathcal{P}^r] \sum_{k \in T_i^l} I(X_{i,k} = j), & w = 0. \end{cases}$$

Then, we have that

$$\mathcal{Q}(\mathcal{P} \mid \mathcal{P}^r) = \sum_{j=1}^{4} N_{0j} \log p_{0j} + \sum_{w=1}^{W} \sum_{j=1}^{4} N_{wj} \log p_{wj}.$$

The EM update steps in the $(r+1)$-th iteration are given by

E-step:

$$E[I(Y_{i,l} = 1) \mid X_i, \mathcal{P}^r] = Pr(Y_{i,l} = 1 \mid X_i, \mathcal{P}^r) = \frac{Pr(X_i \mid Y_{i,l} = 1, \mathcal{P}^r)}{\sum_{s=1}^{L_i-W+1} Pr(X_i \mid Y_{is} = 1, \mathcal{P}^r)}.$$

M-step:

$$p_{0j} = \frac{N_{0j}}{\sum_{j=1}^{4} N_{0j}}$$
$$= \frac{\sum_{i=1}^{N} \sum_{l}^{L_i-W+1} Pr(Y_{i,l} = 1 \mid X_i, \mathcal{P}^r) \sum_{k \in T_i^l} I(X_{i,k} = j)}{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l} = 1 \mid X_i, \mathcal{P}^r) |T_i^l|}, \quad j \in \{1, \cdots, J\},$$
$$p_{wj} = \frac{N_{wj}}{\sum_{j=1}^{4} N_{wj}}$$
$$= \frac{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l} = 1 \mid X_i, \mathcal{P}^r) I(X_{i,l+w-1} = j)}{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l} = 1 \mid X_i, \mathcal{P}^r)}, \quad j \in \{1, \cdots, J\}, w \in \{1, \cdots, W\}.$$

5

2.2 Zero or one motif per sequences: *zoops*

The zoops model extends the oops model by allowing zero or one occurrence of the motif in each sequence. This extension is obtained by introducing another hidden random variable Z_i which is the indicator variable representing whether sequence i has exactly one copy of the motif. As in the oops model, let $Y_i = \{Y_{i,l}\}_{l=1}^{L_i}$ be the set of indicator variables representing the start site of the motif in sequence i. We have that $\sum_l^{L_i} Y_{i,l} \in [0,1], \forall i$ in the zoops model. Note that both Z_i and Y_i are hidden random variables. The start sites are again assumed to be uniformly distributed, i.e., $Pr(Y_{i,l} = 1 \mid Z_i = 1) = 1/(L_i - W + 1), l \in \{1, \cdots, L_i - W + 1\}$. The conditional likelihood of sequence i given the hidden variables is as follows:

zoops:

$$Pr(X_i \mid Z_i = 1, Y_{i,l} = 1, \mathcal{P}) = \prod_{k \in T_i^l} \prod_{j=1}^{J} p_{0j}^{I(X_{i,k}=j)} \prod_{w=1}^{W} \prod_{j=1}^{J} p_{wj}^{I(X_{i,w+l-1}=j)},$$

$$Pr(X_i \mid Z_i = 0, \mathcal{P}) = \prod_{k=1}^{L_i} \prod_{j=1}^{J} p_{0j}^{I(X_{i,k}=j)}.$$

where $T_i^l = \{1, \cdots, L_i\} - \{l, l+1, \cdots, l+W-1\}$ denotes the background sites and $I(.)$ is the indicator variable. The full data equals $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) \equiv \{(X_i, Y_i, Z_i), i = \{1, \cdots, N\}\}$ for the zoops model. Let π denote the mixing proportion, i.e., proportion of the sequences with the motif. The parameter vector $\mathcal{P}$ also includes π in the zoops model. Let $\mathcal{P}^r$ denote the parameter estimates after r-th EM iteration, then the expected full data likelihood conditional on the observed data is as follows up to a constant:

$$\mathcal{Q}(\mathcal{P} \mid \mathcal{P}^r) = \sum_{i=1}^{N} E[I(Z_i = 1) \mid X_i, \mathcal{P}] \log \pi_1$$

$$+ \sum_{i=1}^{N} \sum_{l=1}^{L_i - W + 1} E[I(Y_{i,l} = 1, Z_i = 1) \mid X_i, \mathcal{P}^r] \log Pr(X_i \mid Z_i = 1, Y_{i,l} = 1, \mathcal{P})$$

$$+ \sum_{i=1}^{N} (1 - E[I(Z_i = 1) \mid X_i, \mathcal{P}^r]) \log(1 - \pi_1)$$

$$+ \sum_{i=1}^{N} (1 - E[I(Z_i = 1) \mid X_i, \mathcal{P}^r]) \log Pr(X_i \mid Z_i = 0, \mathcal{P}).$$

Define

$$
N_{wj} = \begin{cases}
\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E[I(Y_{i,l}=1, Z_i=1) \mid X_i, \mathcal{P}^r], I(X_{i,(l+w-1)}=j), & w \neq 0, \\
\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E[I(Y_{i,l}=1, Z_i=1) \mid X_i, \mathcal{P}^r], \sum_{k \in T_i^l} I(X_{i,k}=j) & \\
+ \sum_{i=1}^{N}(1 - E[I(Z_i=1) \mid X_i, \mathcal{P}^r]) \sum_{k=1}^{L_i} I(X_{i,k}=j), & w = 0.
\end{cases}
$$

Then, we have

$$
\begin{aligned}
\mathcal{Q}(\mathcal{P} \mid \mathcal{P}^r) &= \sum_{i=1}^{N} E[I(Z_i=1) \mid X_i, \mathcal{P}^r] \log \pi_1 + \sum_{i=1}^{N}(1 - E[I(Z_i=1) \mid X_i, \mathcal{P}^r]) \log(1 - \pi_1) \\
&+ \sum_{j=1}^{4} N_{0j} \log p_{0j} + \sum_{j=1}^{4} \sum_{w=1}^{W} N_{wj} \log p_{wj}.
\end{aligned}
$$

The $(r+1)$-th update steps are as follows:

E-step:

$$
\begin{aligned}
E[I(Y_{il=1}, Z_i=1) \mid X_i, \mathcal{P}^r] &= Pr(Y_{i,l}=1, Z_i=1 \mid X_i, \mathcal{P}^r) \\
&= \frac{Pr(X_i \mid Y_{i,l}=1, Z_i=1, \mathcal{P}^r) \frac{\pi_1}{1-L_i-W+1}}{\sum_{s=1}^{L_i-W+1} Pr(X_i \mid Y_{is}=1, Z_i=1) \frac{\pi_1}{L_i-W+1} + Pr(X_i \mid Z_i=0, \mathcal{P}^r)(1-\pi_1)}, \\
E(I(Z_i=1) \mid X_i, \mathcal{P}^r) &= Pr(Z_i=1 \mid X_i, \mathcal{P}^r) \\
&= \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l}=1, Z_i=1 \mid X_i, \mathcal{P}^r).
\end{aligned}
$$

M-step

$$
\begin{aligned}
\pi_1 &= \frac{1}{N} \sum_{i=1}^{N} Pr(Z_i=1 \mid X_i, \mathcal{P}^r) = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l}=1, Z_i=1 \mid X_i, \mathcal{P}^r), \quad \pi_0 = 1 - \pi_1, \\
p_{0j} &= \frac{N_{0j}}{\sum_{j=1}^{4} N_{0j}} \\
&= \frac{\sum_{i=1}^{N} Pr(Z_i=0 \mid X, \mathcal{P}^r) \sum_{k=1}^{L_i} I(X_{i,k}=j)}{\sum_{i=1}^{N} Pr(Z_i=0 \mid X_i, \mathcal{P}^r) L_i + \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l}=1, Z_i=1 \mid X, \mathcal{P}^r)(L_i-W+1)} \\
&+ \frac{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l}=1, Z_i=1 \mid X, \mathcal{P}^r) \sum_{k \in T_i^l} I(X_{i,k}=j)}{\sum_{i=1}^{N} Pr(Z_i=0 \mid X_i, \mathcal{P}^r) L_i + \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l}=1, Z_i=1 \mid X, \mathcal{P}^r)(L_i-W+1)}, \\
p_{wj} &= \frac{N_{wj}}{\sum_{j=1}^{4} N_{wj}} \\
&= \frac{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l}=1, Z_i=1 \mid X_i, \mathcal{P}^r) I(X_{i,l+w-1}=j)}{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l}=1, Z_i=1 \mid X, \mathcal{P}^r)}, \quad j \in \{1, \cdots, J\}, w \in \{1, \cdots, W\}.
\end{aligned}
$$

7

Note that

$$\sum_{i=1}^{N} Pr(Z_i = 0 \mid X_i, \mathcal{P}^r) = \sum_{i=1}^{N}(1 - Pr(Z_i = 1 \mid X_i, \mathcal{P}^r)) = N - \sum_{i=1}^{N} Pr(Z_i = 1 \mid X_i, \mathcal{P}^r)$$

$$= N - \sum_{i=1}^{N} \sum_{l=1}^{L_i - W + 1} Pr(Y_{i,l} = 1, Z_i = 1 \mid X_i, \mathcal{P}^r).$$

The outputs of the EM algorithm are the estimates of the model parameters, $\hat{\mathcal{P}}$, and the posterior location probabilities, i.e., $Pr(Y_{i,l} = 1 \mid X_i, \hat{\mathcal{P}}), i = \{1, \cdots, N\}, l = \{1, \cdots, L_i - W + 1\}$ for the oops model and $Pr(Y_{i,l} = 1, Z_i = 1 \mid X_i, \hat{\mathcal{P}})$, and $Pr(Z_i = 1 \mid X_i, \hat{\mathcal{P}}), i = \{1, ..., N\}, l = \{1, ..., L_i - W + 1\}$ for the zoops model. These posterior location probability estimates point out the most likely motif start site in each sequence.

2.3   Constraints in existing motif detection algorithms

The notion of supervising the motif searches has not been much explored in the regulatory motif finding literature. In the early Lawrence & Reilly (1990) paper, authors consider different motif models for the CRP binding site in *E.coli*. In particular, they employ constraints on the parameters of the PWM to enforce palindromicity, AT specific regions and some other simple constraints in the binding site. Frech et al. (1993) develop a method that operates under the observation that biological significance is concentrated in a distinct region which includes a highly conserved consensus core and extends beyond the core in one or both directions. Even though they do not use multinomial mixture models to model sequence data, their method roughly corresponds to searching for motifs with a high information core and lower information positions beyond the core in one or more directions. Some of the recent methods also provide ad hoc ways of enforcing palindromicity to the motifs (Bailey & Elkan, 1995a; Liu et al., 2001) or allowing motifs to have two conserved blocks separated by short background site (Liu et al., 2001).

3   Motif finding using constrained mixture models: *c.oops* and *c.zoops*

3.1   Mixture of constrained multinomial models

In this section, we extend the unconstrained motif models oops and zoops to allow constraints on the motif PWMs. We incorporate structural information of the binding
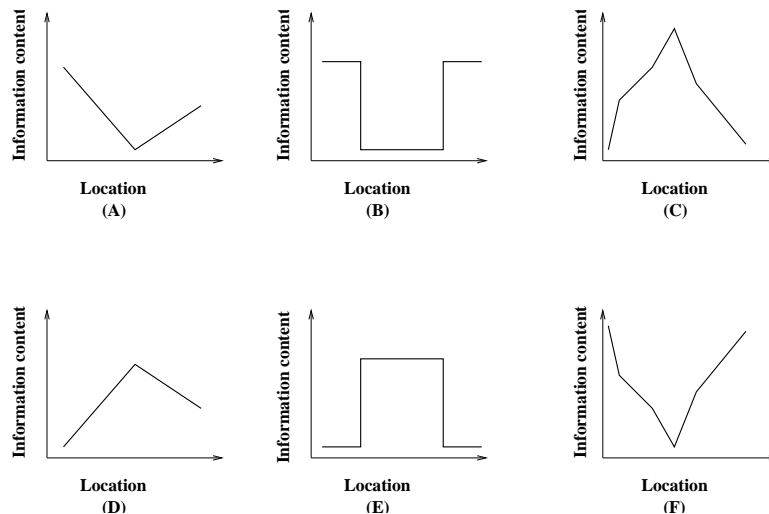
8

Figure 2: *Examples of information content profiles.* (A) and (B) roughly correspond with `high-low-high` information content profiles and mirror images (D) and (E) of these correspond to `low-high-low` information content profiles. Figures (C) and (F) put order constraints on the information content at various positions.

site by constraining the information content profile of the motif PWM. Recall that the information content at a position w is given by

$$IC(w) = \log_2 J + \sum_{j=1}^{J} p_{wj} \log_2 p_{wj}.$$

We will assume a specific model, $IC(w; \overrightarrow{\theta}), w = \{1, \cdots, W\}$, for the information content $IC(w)$ and parameterize it by $\overrightarrow{\theta}$. The functional form of $IC(w; \overrightarrow{\theta})$ is solely determined by the constraints we are enforcing on the information content profile. Figure 2 gives examples of various information content profiles, each of which can be formulated as constraints on $IC(w)$. For instance Figure 3 is an example from a family of motifs with high information in the middle and low information towards the ends. It is parameterized by θ_1 and θ_2 which denote the information content of the highest information position and the angle of the line that passes through the information content of each position, respectively. We refer to motifs with such structured information content profiles as *structured motifs*, and motifs that have random information content profiles (no specific ordering or clustering of information content across positions) as *unstructured motifs*. In particular, one could also have a completely deterministic information content profile, where the actual values of $IC(w), \forall w$ are known a priori or one could have a specified ordering of information content at various positions. As an example, the information content profile of the ABF1 site given in Figure 1 can be viewed as a `high-low-high` information content

9

profile or can be represented by more restrictive constraints such as

$$IC(1) = IC(2) = IC(10) = IC(11) = IC(12) = \theta_1 \geq IC(3) = IC(4) = \theta_2$$
$$\geq IC(5) = IC(6) = IC(7) = IC(8) = IC(9) = \theta_3,$$

where the whole information content profile is parameterized by three parameters, $(\theta_1, \theta_2, \theta_3)$. This particular ordering corresponds to equally conserved first two and last three positions and less conserved middle positions.
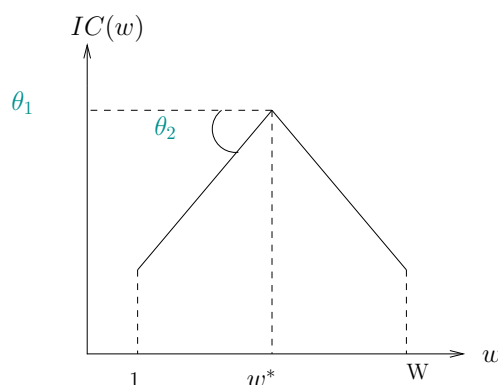

Figure 3: *Example of a parametrized information content profile for a structured motif.* Motif families with this type of information content profiles contain high information in the middle and lower information towards the ends. Information content at position w equals: $IC(w; \theta_1, \theta_2) = \theta_1 - |w - w^*| \tan \theta_2$.

Note that if we take a particular PWM with any of the information content profiles in Figure 2 and permute its columns, the total information content will not change; however, the specific shape of the information content profile would be destroyed. The oops and zoops models are unsupervised methods and do not enforce any particular information content profile on the motif and as a result the are unable to distinguish between motifs with specific information content profiles. Our supervised method COMODE extends these models in a way that allows incorporation of information regarding the structure of the information content profile. It is worth noting that constraints on the information content profile are global constraints, however there might be cases where simply biasing of certain positions in the motif towards certain nucleotides is required. Specifically, if one is expecting, say, a T rich region in the motif, then the probabilities of nucleotide T at these positions could be forced to be greater than a threshold, i.e., $p_{w4} > 0.7$. Similarly, it is straightforward to enforce symmetry or palindromicity constraints with this method.

In the previous section, we reviewed the unconstrained oops and zoops that model background sites using a 0th order Markov chain. Recent trend is to model background sites by a higher order Markov chain and make the order flexible by specifying it as a user defined parameter (Liu et al., 2001). Moreover, more flexible approaches such as allowing the background distribution to be estimated from an independent dataset (such as all intergenic regions of the organism of the interest) with the specified order on the Markov chain or from the dataset to be searched for motifs a priori are pursued (Liu et al., 2001) and shown to be useful. We also follow this approach and allow the background model to be a Markov chain with user defined order and estimate the corresponding transition matrix and the starting state probabilities separately based on either the input sequences or the intergenic regions of the organism of interest. Then, the parameters of the background distribution are fixed at these estimated values.

We will refer to these resulting new models with constraints on the PWM as *c.oops* and *c.zoops*. Note that with these new models we still have the same likelihood representation, however the parameters of the PWM now lie in a reduced space and this reduction is determined by the type of constraints employed either on the information content profile or the individual multinomial probabilities of PWM.

3.2 Maximum likelihood estimation in constrained motif models

Maximum likelihood estimation in the unconstrained motif models involves a standard application of the EM algorithm to a two component multinomial mixture model. Both E and M steps of the EM algorithm have closed form solutions in such models. However, in our constrained motif model the M-step does not have a closed form solution. For this purpose, we discuss two approaches for maximizing the likelihood. The first one relies on the EM algorithm and solves a nonlinear optimization problem at every M-step. The second approach directly applies a nonlinear optimization method to maximize the the observed data likelihood. We noticed that for large problems (i.e. number of sequences > 20 and length of each sequence > 100), the latter method is substantially faster. In particular, when the enforced constraints are complex (i.e. specific ordering on the information content profile), the latter approach performs about 5 times as fast as the former approach (based on 30 sequences of size 600 base pairs). Next, we describe these two approaches in details.

3.2.1 Using the EM algorithm

For the new constrained models c.oops and c.zoops, the E-step stays the same as the unconstrained oops and zoops E-step. The only modification required is in the M-step

when updating the parameters of the motif PWM. Recall that we maximize the expected conditional log likelihood over PWM parameters $\mathcal{P}_W$ and information content profile parameters, $\overrightarrow{\theta}$, and we are not dealing with the background distribution. Let r denote the r-th EM iteration and define

$$\zeta_i^l = \begin{cases} E[I(Y_{i,l} = 1) \mid X_i, \mathcal{P}^r] & \text{in c.oops,} \\ E[I(Y_{i,l} = 1, Z_i = 1) \mid X_i, \mathcal{P}^r] & \text{in c.zoops,} \end{cases}$$

where $\mathcal{P}^r$ refers to the parameter estimates after the r-th iteration. Let

$$N_{wj} = \sum_{i=1}^{N} \sum_{l=1}^{L_i - W + 1} \zeta_i^l I(X_{i,l+w-1} = j), \quad \text{for} \quad w = 1 \cdots, W.$$

Then, the M-step at iteration $(r + 1)$ for the c.oops and c.zoops models requires solving the following maximization problem (G):

$$\max_{p_{wj}} \quad \sum_{w=1}^{W} \sum_{j=1}^{J} N_{wj} \log p_{wj}$$

$$subject \quad to$$

$$2 + \sum_{j=1}^{J} p_{wj} \log_2 p_{wj} = IC(w, \overrightarrow{\theta}), \quad w = 1, \cdots, W, \quad (1)$$

$$\sum_{j=1}^{J} p_{wj} = 1, \quad w = 1, \cdots, W,$$

$$p_{wj} \geq 0, \quad j = 1, \cdots, J; w = 1, \cdots, W.$$

Here, constraint (1) is an information content profile specific constraint, hence this constraint will change depending on the information content profile that the motif model assumes. The last two constraints are typical multinomial probability constraints that one also has in the unconstrained motif models. This is the most generic form of the M-step for these constrained information content profile models. To be more concrete, we will give two examples of such constrained information content profiles. Let the motif model follow the information content profile given in Figure 3,
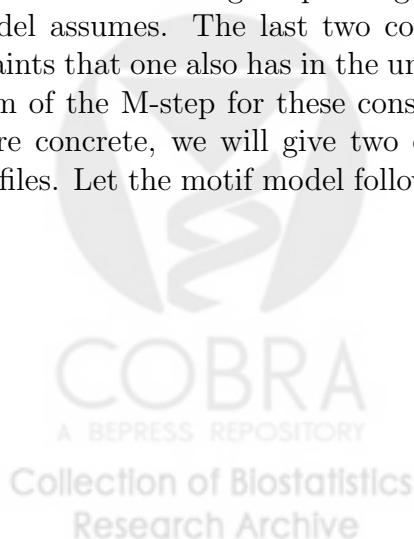
then the M-step maximization problem can be formulated as (A):

$$\max_{p_{wj}} \quad \sum_{w=1}^{W}\sum_{j=1}^{J} N_{wj} \log p_{wj}$$

$subject \quad to$

$$2 + \sum_{j=1}^{J} p_{wj} \log_2 p_{wj} = \theta_1 - \delta(w, w^*)\tan\theta_2, \quad w = 1, \cdots, W,$$

$$\sum_{j=1}^{J} p_{wj} = 1, \quad w = 1, \cdots, W,$$

$$p_{wj} \geq 0, \quad j = 1, \cdots, J; w = 1, \cdots, W.$$

where $\delta(w, w^*) = |w - w^*|$. Note that if we want to enforce this information content profile and not allow mirror images of it, it is sufficient to ensure that θ_2 is strictly positive. In this maximization problem we have to solve for all motif parameters simultaneously, namely for $\mathcal{P}_W, \theta_1, \theta_2$. As a second example, we will assume that the information content at all positions are known, i.e., we have a priori given information content profile. Then, the corresponding M-step is an easier optimization problem since we can maximize over each position in the motif separately. We have the following maximization problem for each position w separately (B):

$$\max_{p_{wj}} \quad \sum_{j=1}^{J} N_{wj} \log p_{wj}$$

$subject \quad to$

$$2 + \sum_{j=1}^{J} p_{wj} \log p_{wj} = IC(w),$$

$$\sum_{j=1}^{J} p_{wj} = 1,$$

$$p_{wj} \geq 0, \quad j = 1, \cdots, J,$$

where $IC(w)$ is just a constant, $w = 1, \cdots, W$. The maximization problems (A) and (B) are nonlinear constrained optimization problems without closed form solutions. We use a state of the art nonlinear optimization method called *Sequential Quadratic Programming (SQP)* to solve these. Details of this method will be given in the next subsection.

3.2.2 Using optimization techniques for nonlinear constraint problems

The EM algorithm is generally appealing since it guarantees convergence to a local maximum as long as the likelihood is increasing at every iteration. Although the

13

convergence is slow, this is still favorable if the expectation and maximization steps are easy to compute, i.e., the M-step has a closed form solution. We have shown in the previous subsection that maximum likelihood estimation in the constrained motif models can be done with the EM algorithm by solving a nonlinear constraint problem at every M-step. However, this approach could get computationally intensive and slow since it boils down to solving as many nonlinear problems as the number of iterations it takes the EM algorithm to converge. As an alternative, one can directly work with the observed data likelihood and solve a single nonlinear constraint problem. Define

$$
B(i,l) = \frac{1}{L_i - W + 1} \prod_{k \in T_i^l} \prod_{j=1}^{J} p_{0j}^{I(X_{i,k}=j)},
$$

$$
C(i) = \prod_{k=1}^{L_i} \prod_{j=1}^{J} p_{0j}^{I(X_{i,k}=j)}.
$$

Here, $B(i,l)$ represents the likelihood of the sites contributing to the background given that the motif start site is l in sequence i, and $C(i)$ represents the likelihood of sequence i under the background model. This representation is given for a 0-th order Markov background model, however it is straightforward to adapt it to higher order background models. Then, the likelihood of observation i in the c.zoops model is given by

$$
Pr(X_i \mid \mathcal{P}) = \pi_i \sum_{l=1}^{L_i - W + 1} B(i,l) \prod_{w=1}^{W} \prod_{j=1}^{J} p_{w,j}^{I(X_{i,w+l-1}=j)} + (1 - \pi_1)C(i),
$$

Note that we do not have the term $(1 - \pi_1)C(i)$ in the c.oops model. Then, we have the following nonlinear constraint maximization problem

$$
\max_{p_{wj}, \pi_1} \quad \sum_{i=1}^{N} \log P(X_i \mid \mathcal{P})
$$

$$
subject \quad to
$$

$$
2 + \sum_{j=1}^{J} p_{wj} \log_2 p_{wj} = IC(w, \overrightarrow{\theta}), \quad w = 1, \cdots, W, \quad (1)
$$

$$
\sum_{j=1}^{J} p_{wj} = 1, \quad w = 1, \cdots, W,
$$

$$
p_{wj} \geq 0, \quad j = 1, \cdots, J; \quad w = 1, \cdots, W,
$$

$$
0 \leq \pi_1 \leq 1.
$$

Nonlinear constraint problems are a well studied topic of the optimization literature. We refer to Bazaraa et al. (1979) for an overview and analysis of such methods.

14

In our application, we used the NAG Fortran subroutine `E04UCF` to solve the encountered nonlinear constraint problems. This subroutine minimizes (hence maximizes - of) a given smooth function subject to constraints. The constraints that can be handled are quite general and include simple upper and lower bounds, linear constraints or smooth nonlinear constraints on the parameters of the PWM. The underlying method that `E04UCF` employs is a Sequential Quadratic Programming (SQP) algorithm. SQP is a generalization of Newton's method for unconstrained optimization and is applied to a Lagrangian function in the context of constrained optimization. Briefly, the main idea is to replace the objective function with its quadratic approximation and the constraint functions by their linear approximations. A detailed overview on sequential quadratic programming can be found in Boggs & Tolle (1995). Moreover, details on the specific implementation of `E04UCF` are available in the NAG documentation (NAG, 1998).

3.2.3 Starting values for the algorithms

We will firstly discuss the rationale for our method of choosing starting values for the EM algorithm. The same method will be employed for choosing starting values of the SQP algorithm when maximizing the observed data likelihood directly.

The EM algorithm is guaranteed to converge to a local maximum from a given starting value. Generally, it is wise to run the EM algorithm from multiple starting values and to choose the result with the highest final likelihood. Bailey & Elkan (1995b) suggest running the EM algorithm for one step with various starting values that are constructed by converting width W oligos into PWMs. At each position of the these starting PWMs, the corresponding nucleotide in the oligo gets probability p_c (typically equal to 0.5) and the rest of the nucleotide probabilities are set to $(1-p_c)/3$. This is a quick systematic way of constructing starting PWMs and the oligos used are obtained from the dataset itself. After running the EM algorithm for one step from all starting values, EM is run till convergence from the starting value with the highest one step likelihood. In c.oops and c.zoops, the M-step is time wise more expensive than the oops and zoops M-step since there does not exist a closed form solution. Hence, we tried to avoid running one step EM for a large number of starting values. Instead, we considered the following initial likelihood based approach. The likelihood is evaluated at each of 4^W possible starting values (all sequences of width W), and we refer to these likelihoods as the initial likelihoods. Then, we run the EM algorithm till convergence for the `k` starting values that have the highest initial likelihoods, where `k` is a user supplied parameter (e.g. a value of `k` between 10 and 100 works well in practice).

We compared this initial likelihood based approach with the one-step EM likelihood based approach on 10 simulated datasets. These datasets consisted of $N = 50$

sequences of length $L = 100$, with a structured motif of width $W = 6$ inserted in each of them. The simulations were run on 200Mhz Ultrasparc SUN workstations. Table

I.1	I.5	I.10	I.20	I.50	I.100	OS.1	OS.5	OS.10	OS.20
0	0	0	1	3	8	0	0	0	3
14.4	18.86	26.33	51.35	123.54	234.28	152.28	155.94	164.09	183.33
0	0	0	0	2	4	0	0	0	2
13.8	22.23	28.04	42.4	102.08	213.15	150.47	160.75	165.51	179.03
0	0	0	1	2	7	1	2	2	3
14.67	20.44	26.75	52.82	139.51	274.28	147.18	153.1	171.11	190.55
0	0	1	1	7	12	1	2	5	11
21.1	37.75	57.8	98.32	194.59	341.31	152.22	162.87	172.68	183.78
0	1	1	1	1	10	1	2	3	3
14.69	19.7	29.88	44.72	101.72	205.01	149.37	154.33	159.86	172.06
0	0	0	1	2	1	1	4	5	6
14.32	19.36	27.34	49.07	110.27	210.44	156.5	160.73	165.63	177.72
0	1	1	3	4	5	1	1	1	3
14.89	21.22	28.36	49.47	121.88	231.06	144.05	150.62	158.94	181.34
1	1	5	7	14	24	1	1	3	5
14.8	19.03	25.86	40.14	92.59	195.06	153.6	158.67	163.14	177.64
0	0	0	1	4	10	0	3	4	5
14.9	21.78	28.84	43.5	92.02	182.39	155.41	161.58	166.72	182.13
0	0	1	1	4	7	0	1	2	4
14.02	25.34	33.19	52.41	119.66	225.84	152.2	157.27	166.86	185.75

Table 1: *Performance of the two starting value strategies on 10 simulated datasets.* I.k refers to the strategy of running c.zoops till convergence from k of the starting values with the highest initial likelihood. OS.k is the strategy that runs c.zoops till convergence from k of the starting values with the highest one-step likelihood. Each row of the table refers to a different data set. The first line of each dataset row reports the number of starting points that converged to the global maximum out of k starting points that are used. The second line refers to the run time required for the implementation of that strategy (in minutes).

1 reports the detailed results for these 10 datasets. I.k refers to the initial likelihood based method where k is the number of highest initial likelihood starting values for which EM is run till convergence. Similarly, OS.k refers to the one-step EM likelihood based approach. For each dataset, the number of starting values that reach the global maximum out of k and the time requirement in minutes is reported. The first thing to notice in this table is that running the EM algorithm till convergence from the starting value with the highest one-step likelihood does not always achieve the global maximum. Having noticed this, we also observe that with $k = 20$ the one-step EM

16

likelihood approach finds the global maximum for all of the datasets, and the initial likelihood based approach with $k = 50$ achieves the same. The initial likelihood based approach misses the global maximum for one dataset with $k = 20$. A summary of the results on 10 datasets is given in Table 2. The first row in this table represents

	I.1	I.5	I.10	I.20	I.50	I.100	OS.1	OS.5	OS.10	OS.20
# datasets	1	3	5	9	10	10	6	8	8	10
time req.	15.2	22.6	31.2	52.4	119.8	231.3	151.3	157.6	165.5	181.3

Table 2: *Summary of the performances of the two starting value strategies.* The first row reports the number of datasets (among 10) for which the global maximum is found by the corresponding strategy. Second row is the average time requirement (in minutes) for that strategy.

the number of datasets for which the global maximum was found by each strategy, and second row reports the average time requirement in minutes for that strategy. When we compare the time requirements for the two approaches for various k values, we notice that the initial likelihood based approach with $k = 50$ achieves the global maximum for all of the ten datasets and requires approximately 30% less computing time than the one-step EM likelihood based approach with $k = 20$. Based on this limited but suggestive study, we employ the initial likelihood based approach for selecting starting values for the EM algorithm and also allow k to be a user specified parameter for flexibility.

Similarly, when maximizing the observed data likelihood directly by an SQP method, we use k starting values determined by the initial likelihood evaluation.

3.3   Model selection with likelihood based cross-validation

One of the main challenges in motif finding is that the width of the motifs are not known a priori. Most of the available methods search through a specified range of motif widths and then choose the best width by optimizing a model selection criterion. This criterion cannot simply be the likelihood value since the likelihood increases with motif width as the number of free parameters change. In MEME, motif width is chosen by minimizing a heuristic function based on the likelihood ratio test. As mentioned in Bailey & Elkan (1995b), this method ignores the fact that the EM algorithm might have converged to a local minimum and hence the distributional assumptions on the test statistics that the method relies on will not hold. Moreover, their criteria tries to adjust for multiple testing in an unusual way by replacing the p-value of the likelihood ratio test statistics by p-value to the power $1/(\#$ of degrees of freedom$)$.

As our model selection method, we chose likelihood-based cross-validation. Likelihood-

17

based cross-validation was used in the context of mixture of normals by Smyth (2000) to select the number of components in the mixture model. The simulation studies of Pavlic & van der Laan (2003) showed that likelihood based cross-validation performed well compared to common approaches based on validity functionals such as Akaike's information criterion (AIC) (Akaike, 1973; Bozdogan, 2000), Bayesian Information criterion (BIC) (Schwartz, 1978) and ICOMP (Bozdogan, 1993). Recently, van der Laan et al. (2003) derived a finite sample result that implied the asymptotic optimality of likelihood-based cross-validation.

The general idea of cross-validation is to divide the total number of observations into a *training* and *validation* set. The training set is used to learn the parameters of a given model and the validation set is used to evaluate the performance of this trained model. In the context of likelihood-based cross validation, the performance assessment is based on the Kulback-Leibler divergence (KL-divergence) that is used as the loss function. The KL-divergence is a measure of distance between two densities and is given by

$$DKL(f, g) = \int \log \left(\frac{f(x)}{g(x)} \right) f(x) dx,$$

for two densities f and g. We refer to van der Laan et al. (2003) for an overview of the methodology and briefly outline $V-$fold likelihood-based cross-validation in the motif finding context:

- STEP I: Divide the dataset into V disjoint sets of approximately the same size (n_v).

- STEP II: For each motif model $k \in \{1, \cdots, K\}$ (e.g. K different motif widths), perform the following:

  - For each $v \in \{1, \cdots, V\}$, let $\mathcal{V}_v$ denote the observations in the v-th set and $\mathcal{T}_v$ denote the remaining $N - n_v$ observations. $\mathcal{V}_v$ and $\mathcal{T}_v$ represent the v-th validation and training set, respectively. Estimate the model parameters $\mathcal{P}_k$ based on the training sample $\mathcal{T}_v$ and denote the parameter estimates by $\mathcal{P}_k^v$. Compute the average KL-divergence on the validation sample $\mathcal{V}_v$:

    $$\hat{\theta}(k)^v = - \sum_{i \in \mathcal{V}_v} \frac{1}{n_v} \log Pr(X_i \mid \mathcal{P}_k^v),$$

  - Compute $\hat{\theta}(k) = \sum_{v=1}^{V} \hat{\theta}(k)^v / V$.

- STEP III: Report $\hat{k} = \text{argmin } \hat{\theta}(k)$ as the best model chosen by the likelihood-based cross-validation.

18

Depending on the fold V and the training procedure, cross-validation is known to be a computationally intensive procedure. van der Laan et al. (2003) compare the practical performance of various fold cross-validation schemes in the context of mixture of multinomials and illustrate that 2-fold performs comparably with other fold schemes. Hence, we use 2-fold cross-validation in our application.

4   Simulation studies

We performed simulation studies to investigate various properties of our proposed method COMODE using c.oops and c.zoops models for the identification of structured motifs. We compared the results with those obtained by the standard oops and zoops methods. We will present the results for these four different models in three sections since each simulation study addresses a different question. In these simulations, COMODE obtains the maximum likelihood estimates of the parameters using the EM algorithm as described in section 3.2.1 since we used few short sequences.

4.1   Detection of motifs with low information content

The main goal in this simulation study is to assess the performance of COMODE when the motif appears as a weak signal in the data but has a characteristic information content profile. For this purpose, we generated $N = 30$ sequences of length $L = 100$ from an i.i.d. background model and inserted an instance of a motif with an information content profile given in the sequence logo of Figure 4 in a varying percentage of the sequences. Sequence logos are a way of visualizing the information content profile together with the sequence consensus. The height of each nucleotide letter is proportional to the probability of that nucleotide at that position and the total height represents the information content at that position. All sequence logos in this paper are generated through the website

http://genio.informatik.uni-stuttgart.de/GENIO/logo/logo.cgi.

As mentioned previously, the EM algorithm returns estimated location posterior probabilities of the sequence sites, $P(Y_{i,l} = 1 \mid X_i, \hat{\mathcal{P}}), i \in \{i, \cdots, N\}, l \in \{1, \cdots, L_i - W + 1\}$, where $\hat{\mathcal{P}}$ is the final estimate of the model parameters. These estimated location posterior probabilities are effective in identifying the most likely location of the motif in each sequence. Let K_b denote the true set of motif sites in sample b and let $\hat{K}_b$ denote the set of predicted motif sites in sample b. As a performance measure we use the percentage of the true sites that are predicted by the used method (i.e., c.zoops, c.oops, oops or zoops), $100 \times |K_b \cap \hat{K}_b|/|K_b|$ (percentage sensitivity) and also
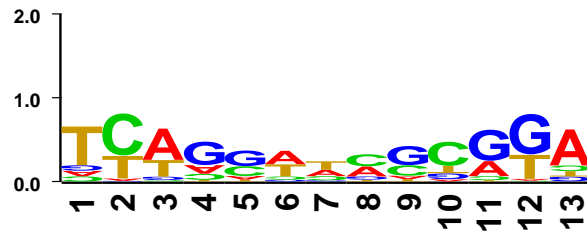
19

Figure 4: *Weak structured motif.* Sequence logo of a structured motif with low information content.

allow the estimated sites to deviate from the corresponding true sites by at most 3 base pairs. When all the sequences have a motif occurrence ($F = 100\%$), oops and zoops and similarly c.oops and c.zoops perform virtually the same. This was also observed by Bailey & Elkan (1995a) for oops and zoops, and it is an important observation since it practically means that going for the bigger zoops model is not causing any loss in the prediction power. When the proportion of the sequences with the motif gets smaller, zoops and c.zoops outperform the corresponding oops versions, however at $F = 75\%$, c.oops performs as well as zoops since it is incorporating structure information.

In Figure 5, we present boxplots of the performance measures over 100 datasets for zoops and c.zoops models. At $F = 100\%$, c.zoops is performing dramatically better than zoops which indicates that even though the motif signal is weak, incorporating knowledge about the information content profile (supervising the search) helps to discriminate it from the background. Moreover, c.zoops remains superior to zoops as F decreases. All four methods had high specificity (between 0.92 and 0.95) and typically did not predict sites on sequences which did not have a motif occurence.

4.2 Performance in the presence of a competing unstructured motif

It is quite common for biological sequences to share common motifs which are not biologically very interesting along with the biologically interesting motifs. In a way, the one or zero motif per sequence assumption is over simplistic and the underlying model of the sequences is often misspecified. For this reason, it is important to assess how in practice the simple model responds to misspecifications.

To mimic such a situation, we have extracted 14 ABF1 sites from the Promoter Database of *Saccharomyces cerevisiae* (SCPD) (Zhu & Zhang, 1999). We inserted these sites (structured motif with the sequence logo given in 6) into 14 sequences of
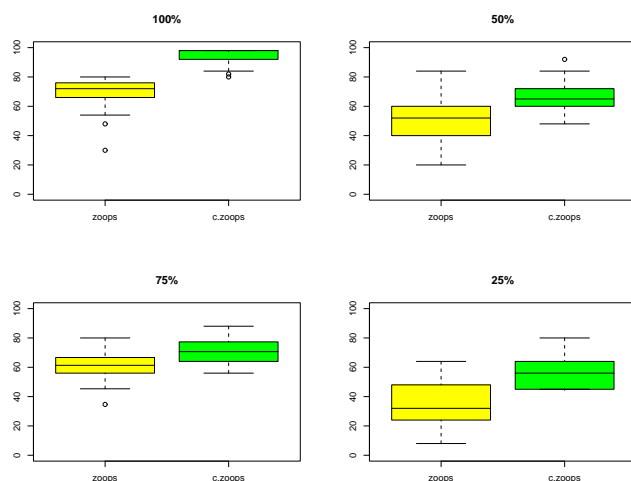
20

Figure 5: *Detecting weak structured motifs.* Boxplot of the percentage of true sites predicted by zoops and c.zoops methods as the percentage of the sequences with the motif occurrence decreases. The average sensitivity decreases as the number of sequences with the motif decreases in the dataset.

length $L = 100$ that were generated from an i.i.d background. A column permuted version of these sites (unstructured motif with the sequence logo given in Figure 7) were also inserted into each of the sequences. Note that by permuting the columns of a PWM, we are not changing the total information content but just perturbating the structure of the information content profile. We applied c.zoops employing a piecewise
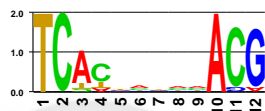

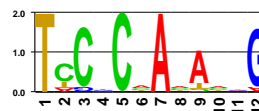Figure 6: Sequence logo of the true ABF1 PWM

Figure 7: Sequence logo of the permuted ABF1 PWM

linear information content profile (mirror image of Figure 3) together with zoops. Boxplots of the performance measures over 50 such datasets are given in Figure 8. We observe that, as expected, c.zoops is performing substantially better than zoops. Since zoops is an unsupervised method, it finds the structured motif as the maximum likelihood estimate in half of the datasets and finds the unstructured motif in the other half, whereas c.zoops almost always finds the structured motif (about 95% of the time). Similar result holds when oops and c.oops are compared. This is a nice property of the constrained models since it implies that supervising is working in the direction that it should even though the imposed piecewise linear information
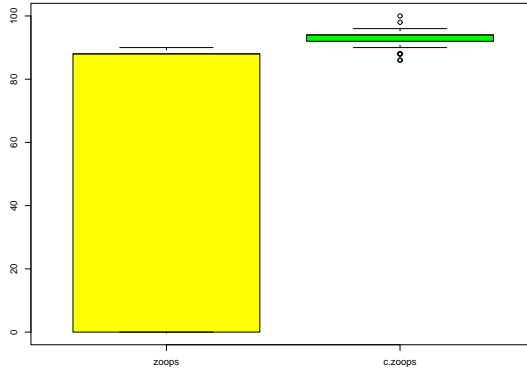
21

Figure 8: *Sensitivity in detecting structured motifs in the presence of a competing unstructured motif.* Boxplot of the percentage of true predicted sites over 50 datasets for zoops and c.zoops models when the sequences contain an unstructured (column permuted ABF1 site) and a structured motif (ABF1 site).

content profile is just an approximation to the true information content profile. This concludes that c.oops and c.zoops are robust against model misspecification and they successfully search for motifs of specific information content profiles.

4.3 Relative efficiency comparisons

In this last simulation study, we compare the small sample size performances of the four methods. Sequences in the simulated datasets have only a structured motif with the information content profile given in Figure 3, and the percentage of the sequences with the motif are varied. This simulation model is more realistic since the datasets generated mimics the biological datasets. We generated $B = 400$ datasets of each with $N = 50$, sequences of length $L = 100$. We varied the percentage of sequences with the motif in the set, $F \in \{100, 75, 50, 25\}$. Let $\hat{\mathcal{P}}^{o,b}$, $\hat{\mathcal{P}}^{zo,b}$, $\hat{\mathcal{P}}^{c.o,b}$, $\hat{\mathcal{P}}^{c.zo,b}$ denote the PWM estimates obtained in sample b by oops, zoops, c.oops, and c.zoops, respectively. We will refer to the (j, w)-th entry of these matrices as $\hat{p}_{wj}^{k,b}$, $k \in \{o, zo, c.o, c.zo\}$. Define

$$MSE_b^k = \sum_{w=1}^{W} \sum_{j=1}^{J} (\hat{p}_{wj}^{k,b} - p_{wj})^2, \quad k \in \{o, zo, c.o, c.zo\},$$

as the squared euclidean distance of the estimated PWM estimated by method k to the true motif PWM, $\mathcal{P}_W$, in sample b. Note that $1/B \sum_{b=1}^{B} (\hat{p}_{wj}^{k,b} - p_{wj})^2$ is the estimated mean squared error for position (j, w) of the PWM obtained by method k. The ratio of the estimated mean squared errors of the two estimates reflects the
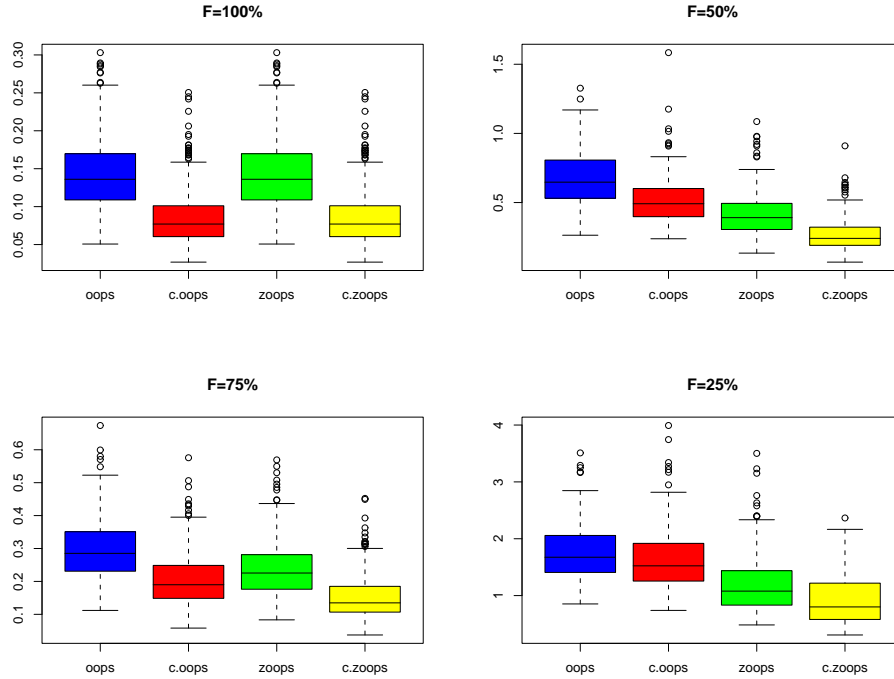
22

Figure 9: *Mean squared error comparisons.* Boxplots of the squared euclidean distance of PWM estimates obtained by oops, zoops, c.oops, and c.zoops methods over $B = 400$ datasets as the percentage of sequences (F) with the structured motif decreases.

relative efficiency of the two estimators, which then can be interpreted as the the ratio of sample size requirements for both methods to achieve the same accuracy. When comparing zoops and c.zoops, we find that the estimated relative efficiency of the two can be as high as 8 for some positions of the PWM, which then implies that zoops needs 8 times as many sequences as c.zoops to achieve the same accuracy at that particular position. To summarize the results in a more compact form we present boxplots of MSE_b^k for each method as F varies in Figure 9. The mean $1/B \sum_{b=1}^{B} MSE_b^k$ can be interpreted as a collapsed mean squared error.

Note again that oops and zoops, and c.oops and c.zoops perform exactly the same at $F = 100\%$. The constrained motif methods c.oops and c.zoops perform better than the unconstrained motif methods at this percentage. As F decreases, c.zoops become superior to all other methods. The main conclusion of this simulation study is that c.zoops provides more accurate PWM estimates than the other 3 methods when the motif of interest is a structured motif and it occurs only in a small proportion of the sequences in the dataset.

23

4.4 Misspecification of the information content profile

As emphasized throughout this paper, the method we are proposing is a supervised method for motif searching which uses additional information about motifs that may not be utilized by other available methods. However, if this information is flawed, then obviously the findings of the method will be flawed, too. We have investigated a few cases of information content profile misspecification including enforcement of a piecewise linear `low-high-low` information content profile as in Figure 3 when in fact the true information content profile is almost random. In this particular case, the resulting PWM had an almost flat information content profile which is still an acceptable example of the profiles given in Figure 3. The typical behavior of the method under such misspecifications is to project the true PWM onto the space defined by the enforced constraints. If one does not have a clear idea about the structure of the motif, different motif models including the unconstrained motif model can be applied and the best one could be selected by likelihood-based cross-validation as described in subsection 3.3.

5 Data analysis

5.1 Comparisons with MEME and BioProspector on transcription factor binding data of *Saccharomyces cerevisiae*

We analyzed the binding data from Lee et al. (2002) for three transcription factors: ARO80, SWI5, and BAS1. In Lee et al. (2002), the sequences of intergenic regions bound with p-values less than 0.001 for each transcription factor were used in motif search by MEME. MEME was run using the zoops model with a 6th order Markov background model and a motif width range of 6 to 18 bases. For each transcription factor, two motifs were reported. The first motif reported is selected based on the likelihood ratio test statistics and the second motif reported is based on a specificity score described in Hughes et al. (2000). The details of the final selection criteria are available in the supplementary material of Lee et al. (2002). We specified the following information for supervising the motif search for each transcription factor:

- `ARO80`: a zinc binuclear cluster - probably bi-modal (3 bp conserved in each region).

- `BAS1`: a tryp cluster - probably one main conserved region (5-6 bp) possibly with smaller sub-peaks.

- `SWI5`: a C2H2 zinc finger - probably one main conserved region (5-6 bp) possibly with smaller sub-peaks.

24

We formulated this information into constraints as in Figure 10. ARO80 has two high information regions and a low information region whereas BAS1 and SWI5 have a single high information region. More specifically, in `ARO80`, positions $\{1, \cdots, w_1\}$ and $\{w_2, \cdots, W\}$ are enforced to have information content at least as large as some θ_1 and θ_3 and the middle positions are enforced to have information content at most θ_2. Similarly, for `SWI5` and `BAS1`, the information content of all positions is enforced to be at least as large as some θ_4 and θ_5. The summary of the constraints employed and the resulting number of models are given in Table 3.

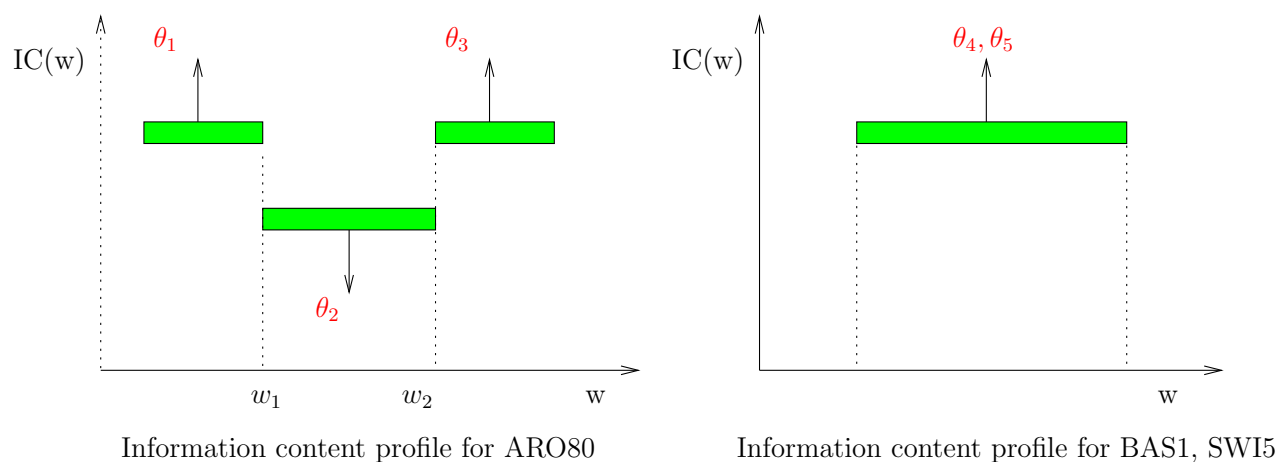Information content profile for ARO80 Information content profile for BAS1, SWI5

Figure 10: *Constraint formulation for the selected transcription factors.*

In the analysis of SWI5, we encountered many A repeats in the results despite the higher order Markov background model used. To eliminate such repeats without physically altering the sequences, we added the following *consecutive sum* constraint into our model

$$\sum_{w=k}^{w=k+4} p_{w1} \leq 5 \times 0.5, \quad k = 1, \cdots, W - 4. \tag{1}$$

This constraint enforces any 5 consecutive positions in the PWM to have a total nucleotide A probability less than 0.5×5. We chose 5 as the number of constrained consecutive positions since 5 is the smallest motif width that we used. Note that this constraint allows at most an A repeat of size 2. The results of our analyses for these three factors are shown in Table 4 together with MEME (Bailey & Elkan, 1995a) and BioProspector (Liu et al., 2001) results, and the true consensus of the sites from the literature. BioProspector does not automatically provide a method for choosing among different motif widths. For this reason, we set the motif width to the true motif width (from the literature) when using the BioProspector. Additionally,

25

Factor	Constraints	# of different models
ARO80	$IC(w) \geq \theta_1,\ w \in \{1, \cdots, w_1\}$ $IC(w) \geq \theta_3,\ w \in \{w_2, \cdots, W\}$ $IC(w) \leq \theta_2,\ w \in \{w_1 + 1, \cdots, w_2 - 1\}$ $\theta_1 = \theta_3 \in \{0.6, 1.2, 1.8\}$ $\theta_2 \in \{0.6, 1.2, 1.8\},\ w_1 = w_2 = 3$	$13 \times 3 \times 3$
SWI5	$IC(w) \geq \theta_1,\ w \in \{1, \cdots, W\}$ $\theta_4 \in \{0.6, 1.2, 1.8\}$ Consecutive sum constraint (1)	$14^* \times 3$
BAS1	$IC(w) \geq \theta_1,\ w \in \{1, \cdots, W\}$ $\theta_5 \in \{0.6, 1.2, 1.8\}$	$14^* \times 3$

Table 3: *Constraints employed for ARO80, SWI5 and BAS1 motif detection.* For each factor, 13 different motif widths in the range $\{6, \cdots, 18\}$ are considered. The last column is the different number of models that are used in likelihood-based cross-validation. * For BAS1 and SWI5, we also consider width of 5.

BioProspector allows the background distribution to be at most 3rd order Markov model which prevented us from comparing it to our method and MEME in a fair manner. BioProspector was run using the following options

```
ARO80: BioProspector -i infile -f yeast_int.bg -W 4 -w 4 -G 7
-g 7 -a 1 -r 25 -o outfile &
BAS1: BioProspector -i infile -f yeast_int.bg -W 7 -a 1 -r 25
-o outfile &
SWI5: BioProspector -i infile -f yeast_int.bg -W 6 -a 1 -r 25
-o outfile &
```

BioProspector allows the smallest motif width to be 4 bases. Hence, when searching for ARO80 sites, we asked for two blocks of size 4 instead of two blocks of size 3. COMODE uses 2-fold likelihood cross-validation for model selection. BioProspector and COMODE identify the true binding site for BAS1 whereas MEME seems to be stuck at some local maximum with T repeats and overfits the motif width. COMODE overfits the motif by one base, however the resulting motif includes the true consensus. For ARO80, COMODE again overfits the motif by one base, however the rest of the consensus matches closely to the true consensus. BioProspector identifies one of the conserved blocks correctly and gives one base mismatch in the second block. MEME seems to be getting a completely different motif than the true consensus with the likelihood ratio test based scoring whereas the specificity score identifies a long motif with two blocks [CCG,SSG] that matches to the true consensus. In the analysis of SWI5, all the three methods generate different results. MEME gets stuck

26

at CA repeats and Bioprospector results in 3 different binding sites that do not match to the true consensus. When COMODE is used without the additional cumulative sum constraint given in equation (1), it reports an A repeat site. However with this additional constraint, the resulting binding site matches closely to the true consensus. In summary, MEME succeeds in one, BioProspector in two, and COMODE in three of the examples. In these examples BioProspector was provided the true motif length and COMODE used 2-fold likelihood based cross-validation. COMODE seemed to slightly overfit the motifs by choosing a one base pair longer motifs, however the chosen motif contained the true consensus.

These limited real data examples show that simple constraints on the information content profiles might be used to supervise motif searches. The three transcription factors we analyzed have sites that can be represented by simple constraints and the three methods (MEME, BioProspector and COMODE) work competitively on at least two of these. However, COMODE has the generality of allowing any type of constraints. BioProspector works pretty well in identifying two conserved blocks by assuming that these two blocks are separated by background sites. However, there are biological examples where some of the positions between the blocks are well conserved and different than the background sites. In such cases, COMODE has the flexibility to provide a more precise information content profile for the binding site. We also note that the constraints employed by COMODE in these three examples do not result in estimating extra information content profile specific parameters. Instead, the additional complexity comes from having to compare more motif models with different size parameter spaces.

5.2 Application to even skipped gene of *Drosophila*

The current version of COMODE implements motif detection with the c.oops and c.zoops models, which allow at most one occurrence of the motif of interest in each input sequence. However, there are many biological examples where a long sequence contains multiple copies of multiple binding sites. We applied our method on two of the *cis*-regulatory regions of the even skipped (*eve*) gene of *Drosophila*. These *cis*-regulatory regions are stripe 2 and stripe 3-7 which are 670 and 511 base pairs long, respectively. Multiple transcription factors are observed to bind to these *cis*-regulatory regions and there might be more than one site for each of these factors (Davidson, 2001). We treat each of these *cis*-regulatory modules independently and identify binding sites for each of them separately. In this application, we have to account for the fact that we have a long stretch of regulatory region with multiple motif occurrences. One empirical way to deal with this phenomena using the c.zoops model is to divide the regulatory module into subsequences of cut length U base pairs and call the new dataset $\mathcal{D}_U$. Several values of the cut parameter U are considered in this application. We then run c.zoops on the new dataset $\mathcal{D}_U$ and search for

motifs of `high-low-high` or `low-high-low` information content profiles (Figure 3). We repeat the analysis with motif widths of $8, 9, 10, 11, 12$. The explicit formulation of the M-step for this constrained model is given by the maximization problem (A). As mentioned previously, starting values of the SQP algorithm are based on the initial likelihood approach and we use 100 starting starting values with the highest initial likelihood. w^* of each width W motif is set to $\lceil w/2 \rceil + 1$, and π, the proportion of the sequences with the motif occurrence, is set to 0.5. We perform the analysis for both stripe 2 and stripe 3-7 in this fashion. As a result of this analysis, we firstly observe that the motif estimates obtained for various motif widths are extensions of each other (i.e., motif estimate obtained for W=12 extends motif estimate obtained with $W = 8$). Further selection of the motif width is performed using the likelihood based cross-validation method with 2-fold cross-validation scheme. Secondly, the results obtained for various cut lengths turn out to be quite similar in the sense that about 96% of the sites identified with different cut lengths overlap. The sequence logos obtained by aligning the positions with posterior location probability greater than 0.7 are given in Figures 11 and 12.
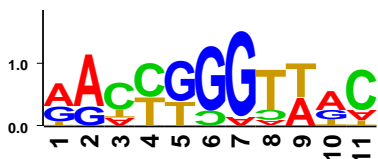

Figure 11: *Sequence logo of the aligned sites for stripe 2.* Sequence logo of the motif obtained with $U = 100$, $W = 11$ for stripe 2.
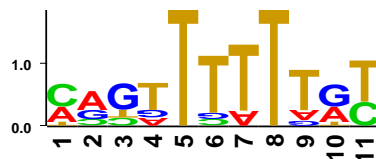

Figure 12: *Sequence logo of the aligned sites for stripe 3-7.* Sequence logo of the motif obtained with $U = 30$, $W = 11$ for stripe 3-7.

The motif identified on stripe 2 is the *Kruppel (Kr)* binding site and the site identified on stripe 3.7 is the *Hunchback (Hb)* binding site. Most of the sites identified on stripe 2 and stripe 3-7 correspond to experimentally verified targets of *Kr* and *Hb* (comparisons were made with the compiled Figure 5 of Berman et al. (2002)) on these stripes. In stripe 2, 4 out of 6 experimentally verified *Kr* sites are correctly identified. In stripe 3-7, 7 out of 11 experimentally verified sites were correctly identified. These sites together with other predicted sites are listed in Tables 5 and 6; experimentally known sites are marked with ∗. The results summarized for stripe 2 and stripe 3-7 correspond to cut lengths $U = 100$ and $U = 30$, respectively.

28

Current available methods use different heuristics to deal with multiple occurrences of a single motif in a single sequence since exact methods increase the computational complexity a great deal. MEME (Bailey & Elkan, 1995a) deals with multiple occurrences of the same motif by modifying the E-step of the EM algorithm. In the E-step, MEME sets

$$\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} Pr(Y_{i,l} = 1 \mid X_i, \mathcal{P}) = NSITES,$$

where each $Pr(Y_{i,l} = 1 \mid X_i, \mathcal{P})$ is between 0 and 1 and $NSITES$ is a user supplied value referring to the expected number of occurrences of the motif in all of the sequences under consideration. Under this constraint, $\sum_{l=1}^{L_i-W+1} P(Y_{i,l} = 1 \mid X_i, \mathcal{P})$ for any given sequence does not necessarily sum to 1. BioProspector (Liu et al., 2001) allows multiple occurrences of the same motif using a heuristic called *threshold sampling*, where more than one subsequence is sampled from each sequence based on some rules.

In the analysis of *cis*-regulatory region of the even skipped gene of *Drosophila*, we used a cutting heuristic to deal with multiple occurrences of the same motif and cut each sequence into subsequences of length U to form a new dataset to apply COMODE. We did not observe any apparent sensitiveness of the results to the cut length U; however cut length is another parameter that can be cross-validated. In essence, different cut lengths index different models by allowing different number of motif occurrences in the dataset.

Discussion

We have introduced a novel supervised motif finding method, COMODE, to detect motifs with specific structural constraints. This method is motivated by the observation that transcription factor binding sites, especially in bacteria genomes, have characteristic information content profiles that distinguish them from random pattern repeats. Through the classification of motifs of DNA binding proteins, a catalog of information content profiles could be obtained. Consequently, the search for regulatory motifs could be focused by utilizing the specific characteristics of the motifs. Our method incorporates structural constraints on the motifs by enforcing constraints on either the information content profile or specific parameters of the corresponding PWM. Therefore, parameters of the resulting PWM lie in a smaller space than the parameter space of unconstrained PWMs. Through simulations, we compared our method with simple unconstrained motif finding methods (Lawrence & Reilly, 1990; Bailey & Elkan, 1994). Advantages of COMODE include improved performance in detecting motifs with weak structured information content profiles, better small sample size performance in PWM estimates (measured by mean squared error). When

there is no prior knowledge about the searched motif in a dataset, this method can be applied using various information content profiles, including the unstructured information content profile and the best fit to the data can be selected by likelihood-based cross-validation. We have compared our method with commonly used methods such as MEME (Bailey & Elkan, 1995a) and BioProspector (Liu et al., 2001) on limited examples. We observed that our method successfully utilizes the available information. The method described here is one specific way of incorporating biological knowledge into motif finding and it is not based on heuristics. As a result the computational complexity is higher compared to unsupervised methods. However, the formulation of the constraints, in general, are efficient since they mostly correspond to searching a smaller parameter space for the PWMs and/or estimating a few extra parameters.

Acknowledgments

We would like to thank Peter Bickel, Derek Chiang, Katherina Kechris, Alan Moses, Erik van Zwet for many helpful discussions.

Appendix

ABF1 sites

The ABF1 sites that are used in competing unstructured motif simulations are given in Table 7. SCPD reports 23 ABF1 sites under the `get sites` section, however the reported position specific weight matrix of the section `get matrix` only incorporates 14 of these sites. Thus, we have used only the sites that contribute to the specified PWM.

Software

The described methodology is implemented in the C programming language in a software package called COMODE. Maximization steps are performed using NAG Fortran subroutine `EOUCF`. The major inputs of the program are

1. Set of sequences to search for motif.

2. Motif width.

3. Background distribution file. (A C function to estimate the parameters of the user supplied order of Markov chain from a given data set is also provided.)

4. A C function evaluating the user supplied constraints and their first derivatives with respect to the PWM parameters (and other parameters if applicable) for a given PWM. This C function will change depending on the type of constraints employed. We have assembled a library of common types of constraints.

Moreover, a function implementing $V-$fold likelihood based cross-validation to choose among different motif widths and information content profiles is also available. This version, using NAG library, is available from the first author (`sunduz@stat.berkeley.edu`). A NAG free version in matlab is in progress.

References

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. Petrov & F. Csaki, eds. Budapest: Academiai Kiado.

BAILEY, T. & ELKAN, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* , 28–36.

BAILEY, T. & ELKAN, C. (1995a). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21**, 51–80.

BAILEY, T. & ELKAN, C. (1995b). The value of prior knowledge in discovering motifs with meme. *Technical Report CS95-413 Department of Computer Science, University of California, San Diego* .

BAZARAA, M., SHERALI, H. & SHETTY, C. (1979). *Nonlinear programming: Theory and Algorithms*. New York: John Wiley & Sons, Inc.

BERG, O. G. & VON HIPPEL, P. H. (1987). Selection of dna binding sites by regulatory proteins, statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology* **193**, 723–750.

BERMAN, B., NUBU, Y., PFEIFFER, B., TOMANCAK, P., CELNIKER, S., LEVINE, M., RUBIN, G. & EISEN, M. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. In *Proceedings of National Academic of Sciences*, vol. 99.

BOGGS, P. & TOLLE, J. (1995). Sequential quadratic programming. *Acta Numerica* , 1–52.

31

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse fisher informati on matrix. In *Information and Classification*, O. Opitz, B. Lausen & R. Klar, eds. Heidelberg: Springer Verlag.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in info rmation complexity. *Journal of Mathematical Psychology* **44**, 62–91.

Brazas, R. M., Bhoite, L. T., Murphy, M. D., Yu, Y., Chen, Y., Neklason, D. W. & Stillman, D. J. (1995). Determining the requirements for cooperative DNA binding by Swi5p and Pho2p (grf10p/bas2p) at the HO promoter. *Journal of Biological Chemistry* **270**, 29151–29161.

Daignan-Fornier, B. & Fink, G. (1992). Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. *Proceedings of the National Academy of Sciences* **89**, 6746–6750.

Davidson, E. (2001). *Genomic Regulatory Systems: Development and Evolution.* San Diego: Academi Press.

Dempster, A. P., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JRSSB* **39**, 1–38.

Eisen, M. B. (2003). Structural properties of Transcription Factor-DNA interactions and the inference of sequence specificity. Submitted.

Frech, K., Herrmann, G. & Werner, T. (1993). Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Research* **21**, 1655–1664.

Hertz, G. & Stormo, G. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577.

Hughes, J., Estep, P., Tavazoie, S. & Church, G. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *Journal of Molecular Biology* **296**, 1205–1214.

Iraqui, I., Vissers, S., Andre, B. & Urrestarazu, A. (1999). Transcriptional induction by aromatic amino acids in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **19**, 3360–3371.

Lawrence, C., Altschul, S., Boguski, M., J.S. Liu, A. N. & Wootton, J. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.

32

Lawrence, C. & Reilly, A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function and Genetics* **7**, 41–51.

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., I., S., J., Z., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., T.L., V., Fraenkel, E., D.K., G. & Young, R. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804.

Liu, X., Brutlag, D. & Liu, J. (2001). BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Proceedings of the Pacific Symposium of Biocomputing*.

Mirny, L. A. & Gelfand, M. S. (2002). Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Research* **30**, 1704–1711.

NAG (1998). *NAG Fortran Library routine documentation: E04UCF/E04UCA*. www.nag.co.uk/numeric/fl/manual/pdf/E04/e04ucf.pdf.

Pavlic, M. & van der Laan, M. (2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics and Data Analysis* **41**, 413–428.

Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology* **188**, 415–431.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Smyth, P. (2000). Model selection of probabilistic clustering using cross-valida ted likelihood. *Statistics and Computing* **10**, 63–72.

Stormo, G. (2000). DNA binding sites: representation and discovery. *Bioinformatics.* **16**, 16–23.

Tavazoie, S., Hughes, J., Campbell, M., Cho, R. & Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285.

van der Laan, M. J., Dudoit, S. & Keleş, S. (2003). Asymptotic optimality of likelihood based cross-validation. Tech. Rep. 125, Division of Biostatistics, University of California, Berkeley.

Zhu, J. & Zhang, M. (1999). SCPD: A promoter database of yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15**, 607–611.

33

BAS1	
METHOD	MOTIF
MEME	TTTTYYTTYTTKYNTYNT/ANRANRMAARAARRAAAA
	CSNCCAATGKNNCS/SGNNMCATTGGNSG
BIOPROSPECTOR	TGACTC/GAGTCA
	TGACTC/GAGTCA
	TGACTC/GAGTCA
COMODE	TGACTCY/RGAGTCA
LITERATURE	TGACTC/GAGTCA
(Daignan-Fornier & Fink, 1992)	

ARO80	
METHOD	MOTIF
MEME	YKYTYTTYTTNNNNKY/RMNNNNNAARAARARMR
	TRCCGAGRYWNSSSGCGS/SCGCSSSNWWRYCTCGCYA
BIOPROSPECTOR	(TTCG/CGAA, TCGG/CCGA)
	(CCGA/TCGG, CGAA/TTCG)
	(ATAA/TTAT, AAGC/GCTT)
COMODE	WCCGMSNNNNNCCG/CGGNNNNNSKCGGW
LITERATURE	CCGNNNNNNCCG/CGGNNNNNNCGG
(Iraqui et al., 1999)	

| SWI5	
METHOD	MOTIF
MEME	CACACACACACACACACA/TGTGTGTGTGTGTGTGTG
BIOPROSPECTOR	CATACA/TGTATG
	TGTATG/CATACA
	TGTGTG/CACACA
COMODE	RCCAGCR/YGCTGGY
LITERATURE	ACCAGC/GCTGGT
(Brazas et al., 1995)	

Table 4: *Summary of results for* `ARO80`*,* `BAS1`*, and* `SWI5` *motif detection.* Resulting motifs and their reverse complements are given together, i.e., `TGACTCY/RGAGTCA` represents the regulatory motif `TGACTCY` and its reverse complement `RGAGTCA` for `BAS1`. The top 3 motifs are reported for BioProspector. The degenerate nucleotide symbols are as follows: R={A,G}, Y={C,T}, M={A,C}, K={G,T}, S={C,G}, W={A,T}, N={A,C,G,T}.

stripe 2	gctGGCCTGGTTTCtcg
stripe 2	cgcAGTTTGGTAACacg
stripe 2	cgaGACCGGGTTGCgaa*
stripe 2	cttGACTTGGAATCcaa
stripe 2	gcgAACTGGGTTATttt*
stripe 2	ttgAGCCGGGCAGCagg
stripe 2	tcaAAACGGGTTAAgct*
stripe 2	gttAATTGCGTTGCctg
stripe 2	cctGACTTCGCAACccg
stripe 2	ggcAAACGGATTAAcac*
stripe 2	cagTACCGGGTAACcag

Table 5: *Aligned sites with $U = 100$, $W = 11$ in stripe 2* . Sites with posterior location probability ≥ 0.7. * refers to experimentally verified sites.

stripe 3.7	gatCAGTTTTTTGTttt*
stripe 3.7	cgaCCGATTTTTGTgcc*
stripe 3.7	ttaCGGTTTATGGCcgc
stripe 3.7	cccAGCTTCTTTGTtcc
stripe 3.7	atgCAGATTTTTATggg*
stripe 3.7	atcACGTTTTTTGTtcc*
stripe 3.7	cgcTAGTTTTTTTCccc*
stripe 3.7	tctAATTTTTTAATtct*
stripe 3.7	gacAAGGTTATAACgct
stripe 3.7	atcCGTTTGTTTGTgtt
stripe 3.7	attCACGTTTTTACgag*

Table 6: *Aligned sites with $U = 30$, $W = 11$ in stripe 3-7*. Sites with posterior location probability ≥ 0.7. * refers to experimentally verified sites.

35
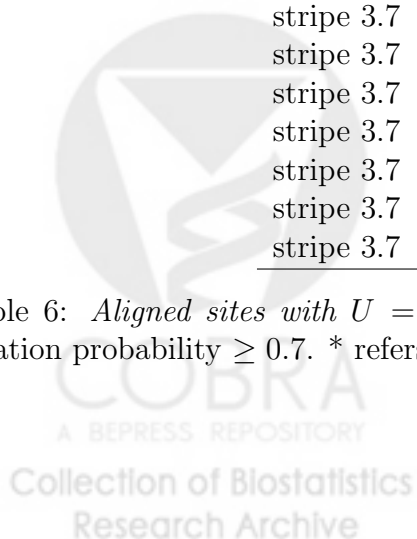
TCTCTCGCAACG
TCTCTCGCAACG
TCACGTCACACG
TCACCGCGAACG
TCATAAAGCACG
TCACTAAAGACG
TCAAAATTAACG
TCACTGTACACG
TCACTAACGACG
TCCCCATTAACG
TCACGATACACG
TCATGCGCTACG
TCATGCGCTACG
TCAAATAACAGA

Table 7: *Aligned ABF1 sites extracted from the Promoter Database of Saccharomyces cerevisiae (SCPD).* 14 ABF1 sites (out of 23 reported sites in SCPD) used in constructing the position specific weight matrix in SCPD.