LOGOS: a modular Bayesian model for *de novo* motif detection

Eric P. Xing[*]
Computer Science Division
University of California
Berkeley, CA 94720
`epxing@cs.berkeley.edu`

Wei Wu
Life Sciences Division
Lawrence Berkeley National Lab
Berkeley, CA 94720
`wwu@lbl.gov`

Michael I. Jordan
Computer Science and Statistics
University of California
Berkeley, CA 94720
`jordan@cs.berkeley.edu`

Richard M. Karp
Computer Science Division
University of California
Berkeley, CA 94720
`karp@cs.berkeley.edu`

Abstract

The complexity of the global organization and internal structures of motifs in higher eukaryotic organisms raises significant challenges for motif detection techniques. To achieve successful de novo motif detection it is necessary to model the complex dependencies within and among motifs and incorporate biological prior knowledge. In this paper, we present **LOGOS***, an integrated* **LO***cal and* **G***l***O***bal motif* **S***equence model for biopolymer sequences, which provides a principled framework for developing, modularizing, extending and computing expressive motif models for complex biopolymer sequence analysis.* **LOGOS** *consists of two interacting submodels: HMDM, a local alignment model capturing biological prior knowledge and positional dependence within the motif local structure; and HMM, a global motif distribution model modeling frequencies and dependencies of motif occurrences. Model parameters can be fit using training motifs within an empirical Bayesian framework. A variational EM algorithm is developed for de novo motif detection.* **LOGOS** *improves over existing models that ignore biological priors and dependencies in motif structures and motif occurrences, and demonstrates superior performance on both semi-realistic test data and cis-regulatory sequences from yeast and Drosophila sequences with regard to sensitivity, specificity, flexibility and extensibility.*

1 Introduction

The identification of motif structures within biopolymer sequences such as DNA and protein is an important task in computational biology and is essential in advancing our knowledge about biological systems. It is known that only a small fraction of the genomic sequences in multi-cellular higher organisms constitute the protein coding information of the genes (e.g., only 5% for mammalian genomes [18]), whereas the rest of the genome, besides playing purely structural roles such as forming the centromeres and telomeres of the chromosomes, contains a large number of short sequence motifs that make up an immensely rich codebook of the gene regulation program, known as the **cis-regulatory system**. It is believed that this regulatory program determines the level, location and chronology of gene expression, which significantly, if not predominantly, contributes to the developmental, morphological and behavioral diversity of complex organisms [4].

The problem of *de novo* motif detection[1] has been widely studied. Numerous algorithmic approaches have been proposed, most of which use probabilistic generative models to model motifs as stochastic string patterns randomly embedded in a simple background. In such a setting, motif detection can be formulated as a standard missing-value inference and parameter estimation problem (for motif locations and position weight matrices, respectively), and standard methods such as EM and Gibbs sampling can be applied. This literature is too large to survey here, but some relevant examples include MEME [1], BioProspector [15], and AlignACE [12]. A different framework based on word segmentation and dictionary construction was proposed in [3], which pointed out the importance of combinatorial analysis of a large set of potential motifs jointly, so that some dependencies among motifs can be captured. A similar 'word-enumeration' idea also appeared in [17]. Recently, Gupta and Liu [9] extended the dictionary model to a stochastic

---

[*] To whom correspondence should be addressed.

[1] Not to be confused with *motif scan*, the task of searching known motifs based on given position weight matrices, as addressed in [7].

dictionary (SD) model by replacing the words in the dictionary with *probabilistic word matrices*, allowing stochasticity of motif instances to be modeled. Many of these methods are widely used and have been successful empirically for motif detection in well curated bacterial or yeast gene regulatory sequences. However, generalization of these successful results to longer, more complex and weakly characterized input sequences such as those from higher eukaryotic genomes seems less immediate. A recent survey by Eisen [6] raises concerns over the inability of some contemporary motif models to incorporate biological knowledge of global motif distribution, motif structures and motif sequence composition.

Recent work has tried to address these concerns from several different perspective. For example, some authors have proposed better objective functions for motif detection, by scoring motifs based on the statistical significance of the information content [11] and considering cooperative motif binding between multiple transcription factors [8]. Van Helden et al. [10] recently suggested using a signature conservation pattern to constrain the motif patterns and incorporating gene expression data from microarrays. Frith et al. [7] used an HMM in their motif scanner to model the possible presence of clustered motif occurrences in complex *cis*-regulatory sequences. Though these attempts head in the direction of more expressive motif models, it is not clear whether these ideas can be integrated to assemble a powerful yet transparent and computationally efficient motif detection algorithm.

We are interested in developing a principled general framework for motif modeling, which is expressive (in terms of being able to describe internal structures, inter-motif relations, motif abundances, etc., and readily incorporates prior knowledge from experimental biology), yet mathematically and algorithmically transparent and well-structured, hence simplifying model construction, computation and extension. In a recent methodological paper, we briefly laid out a theoretical foundation for modular motif models where we made explicit the decomposition of a full motif model into the following two components: the *global distribution model*, which models the frequencies of different motifs and the dependencies between motif occurrences in a sequence; and the *local alignment model*, which captures the intrinsic properties within motifs, including characteristic position weight matrices (PWMs) and site dependencies [22]. Based on this framework, we extended the conventional motif-alignment model into a very expressive hierarchical Bayesian Markovian model, called a hidden Markov Dirichlet-Multinomial (HMDM) model, for local alignment, which successfully captures internal motif structure and incorporates prior knowledge from biologically known motifs using a structured Bayesian prior model for the PWMs of motifs. In the current paper, we integrate the HMDM model into a general framework for the modeling of motif-containing biopolymer sequences and present a fully implemented motif detector developed based on this framework. This framework uses the HMDM model as the local alignment submodel and uses a newly designed HMM that we describe here for the global submodel. A *variational EM algorithm* is developed for efficient Bayesian learning and prediction. We call our framework **LOGOS**, for integrated **LO**cal and **G**l**O**bal motif **S**equence model [2].

2 LOGOS: A Modular Generative Framework for Motif Sequences

2.1 Preliminaries

Motifs are short stochastic string patterns scattered in biopolymer sequences such as DNA and proteins. The characteristic sequence patterns of motifs and their locations often relate to potentially important biological functions such as serving as the *cis*-elements for gene regulation or the catalytic sites for protein activity. Numerous biological studies have revealed rich architecture in the global organization and the internal structures of motifs in higher eukaryotic organisms. Taking DNA motifs as an example, it is well known that the *cis*-regulatory elements often occur in clusters (referred to as *cis*-modules), possibly for eliciting synergistic or more robust regulatory signals [4]. The biophysical mechanisms of DNA-protein interactions at the motif-binding sites further suggest that the sites within the DNA motifs are not necessarily uniformly conserved [21, 20]. Rather, the conservation pattern may be subject to a constraint imposed by the structure of the binding protein, resulting in the so-called "shape" bias (Figure 1). These *meta-sequence features* of motif structure raise significant challenges to conventional motif-finding algorithms, which primarily rely on simplifying independence assumptions that decouple (potential) associations among sites within each single motif and among multiple instances of motifs. (For example, the conventional product multinomial model to be described shortly would assign equal probability to both the original motif and its permuted version in Figure 1).

In the following paragraph, we introduce the necessary notation for our presentation. Note that to simplify the presentation, we use DNA motifs as a running example, but it should be clear that our technique is readily applicable to protein motifs.

We denote a regulatory DNA sequence by a character string $y = (y_1, \ldots, y_T) \in \{\text{A,T,C,G}\}^T$. An indicator string x signals the locations of the motif occurrences. Following biological convention, we denote the *multi-alignment* of M instances of a motif of length L by an $M \times L$ matrix $\mathbf{A}$, of which each *column* corresponds to a *position* or *site* in the motif. The multi-alignment of all instances of motif

---

[2] Not to be confused with "*logo*," a graphic representation of an aligned set of biopolymer sequences first introduced by Tom Schneider [19] to help visualizing the consensus and the entropy (or "information") patterns of monomer frequencies. A *logo* is not a motif finding algorithm, but is often used as a way to present motifs visually.

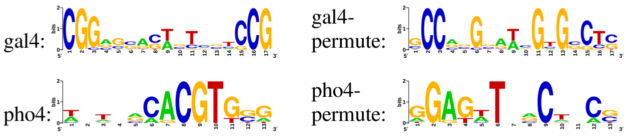


Figure 1. Shape bias. It is believed that for plausible biological motifs the conserved sites are more likely to occur consecutively and possibly followed (or preceded) by heterogeneous sites that are also consecutive (rather than interspersed). Such characteristic conservation patterns of the sites in a motif are often reflected in the "contour shape" of the motif *logo* (e.g., *U- or bell-shaped*, as exhibited by motifs *gal4* and *pho4*, respectively), which reflects the *spatial* pattern of information content over all sites. It is important to note that "shape" is only associated with the conservation pattern of a motif PWM, but **not** with any specific consensus sequences of the motif.

k specified by the indicator string x in sequence y is denoted by $\mathbf{A}^{(k)}(x,y)$. We define a *counting matrix* $h(\mathbf{A}^{(k)})$ (or $h^{(k)}(x,y)$) for each motif alignment, where each column $h_l = [h_{l1},\dots,h_{l4}]'$ is an integer vector with four elements, specifying the number of occurrences of each nucleotide (nt) at position l of the motif. (Similarly we define the *counting vector* h_{bk} for the background sequence $y-\mathbf{A}$, where the somewhat abusive use of the minus sign means excluding all motif sub-sequences in $\mathbf{A}$ from y.) We assume that the nucleotides at position l of motif k admit a *position-specific multinomial distribution*, $\theta_l^{(k)}{=}[\theta_{l1}^{(k)},\dots,\theta_{l4}^{(k)}]'$. The ordered set of position-specific multinomial parameters of all positions of motif k, $\theta^{(k)} = \{\theta_1^{(k)},\dots,\theta_{L^{(k)}}^{(k)}\}$, is referred to as a *position weight matrix* (PWM). It is clear that the counting matrix $h^{(k)}$ corresponds to the *sufficient statistics* of PWM $\theta^{(k)}$. Formally, the problem of motif detection is that of inferring $\mathbf{x} = \{x^{(1)},\dots,x^{(N)}\}$ and estimating $\boldsymbol{\theta} = \{\theta^{(1)},\dots,\theta^{(K)}\}$, given a set of sequences $\mathbf{y}{=}\{y^{(1)},\dots,y^{(N)}\}$ [3].

2.2 The modular motif model

Without loss of generality, assume that the occurrences of motifs in a DNA sequence, as indicated by x, are governed by a **global distribution model** $p(x|\Theta_g,\mathcal{M}_g)$, and for each type of motif, the nucleotide sequence pattern shared by all its instances admits a **local alignment model** $p(\mathbf{A}(x,y)|x,\Theta_l,\mathcal{M}_l)$. We further assume that the background non-motif sequences are modeled by a simple conditional model, $p(y-\mathbf{A}(y,x)|x,\Theta_{bk})$, where the background nt-distribution parameters Θ_{bk} are assumed to be estimated *a priori* from the entire sequence. The symbols Θ_g, Θ_l, $\mathcal{M}_g$, $\mathcal{M}_l$ stand for the parameters (e.g., the PWMs) and model classes (e.g., a product multinomial model) in the respective submodels. Thus, the likelihood of a regulatory sequence y is:

$$\begin{aligned} p(y|\Theta,\mathcal{M}) &= \sum_x p(x|\Theta_g,\mathcal{M}_g)p(y|x,\Theta_l,\mathcal{M}_l) \\ &= \sum_x p(x|\Theta_g,\mathcal{M}_g)p(\mathbf{A}|x,\Theta_l,\mathcal{M}_l)p(y{-}\mathbf{A}|x,\Theta_{bk}), \quad (1) \end{aligned}$$

where $\mathbf{A} \triangleq \mathbf{A}(x,y)$. Note that Θ_l here is not necessarily equivalent to the PWMs ($\boldsymbol{\theta}$) of the motifs, but is a generic symbol for parameters of a more general model of the aligned motif instances. (E.g., in the HMDM model to be defined shortly, Θ_l refers to the hyperparameters that describe a distribution of PWMs.)

Equation (1) makes explicit the modular structure of the **LOGOS** framework for generic motif models. The submodel $p(x|\Theta_g,\mathcal{M}_g)$ captures properties such as the frequencies of different motifs and the dependencies between motif occurrences. On the other hand, the submodel $p(\mathbf{A}|x,\Theta_l,\mathcal{M}_l)$ captures the intrinsic properties within motifs that can help to improve sensitivity and specificity to genuine motif patterns. Depending on the value of the latent indicator x_t (e.g., motif or not) at each position t, y_t admits different probabilistic distributions, such as a particular nucleotide distribution inside a motif or a background distribution.

For example, the conventional *uniform and independent* (UI) model for motif start-positions used in many motif finding algorithms is an instance of a simple global model, where the motif instances are assumed to occur independently with uniform probability at all possible locations in a sequence. So, $p(x) = \prod_{m=1}^{M} p(x_m)$, where $p(x_m{=}t)$ is the marginal probability of the m-th motif at location t, which in this case is a uniform distribution over all t, and the same for all M instances. Note that there is no *model constraint* to prevent having overlapping motif instances [4]. The UI model does not appear to be problematic in *de novo* motif finding tasks involving bacterial or even simple yeast sequence sets, in which the input sequences are usually small in size and homogeneous in content (e.g., pre-screened according to mRNA co-expression) and the motif occurrences tend to be sparse. But some recent studies as well as our experiments suggest that the correctness of motif finding based on the UI assumption starts to break down for less well pre-screened input sequences or those with clustered motif occurrences, such as the Drosophila gene regulatory sequences [2].

An example of the local model is the standard product multinomial (PM) model, where the position-specific nt-distributions within a motif are assumed to be independent [15]. Thus the likelihood of a multi-alignment $\mathbf{A}$ is: $p(\mathbf{A}|\Theta) = \prod_{l=1}^{L}\prod_{j=1}^{4}\left[\theta_{lj}\right]^{h_{lj}}$. Although a popular model

---

[3] For simplicity, we omit the superscript k (motif type index) of variable θ and the superscript n (sequence index) of variable x and y in wherever it is clear from the context that we are focusing on a generic motif type or a generic sequence.

[4] Heuristics are generally employed—such as throwing away overlapping sampled motifs (in the Gibbs sampler) or rescaling the joint posterior of x (in MEME)—to enforce the *non-overlapping constraint*. Nevertheless, this results in inconsistencies between the computed motif distribution and the one defined by the model, and incurs a sizable overhead due to wasteful computations.

for many *de novo* motif finders, PM nevertheless is sensitive to noise and random or trivial recurrent patterns (e.g., poly-N or repetitions of short k-mers such as GC islands), and is unable to capture potential site dependencies inside the motifs. Various pattern-driven approaches (e.g., using a fragmentation model [16], splitting a "two-block" motif into two coupled sub-motifs [15, 1], or imposing explicit "shape" [10] or entropy constraints [14]), have been developed to handle special patterns such as the *U-shaped* motifs, but generalization to other "shapes" seen from known motifs is not very straightforward. Dirichlet priors for θ have been used in the PM setting [1, 16], but they are primarily used for smoothing rather than for explicitly incorporating prior knowledge about motifs.

Recently, Xing et al. [22] developed the HMDM model for motif alignment, which captures site dependencies inside the motifs and incorporates prior knowledge of nt distributions of all motif sites from biologically known motifs. It shows improved sensitivity (compared to PM) to true biological motifs in the presence of synthetic false motifs in the motif detection setting. Frith et al. [7] proposed an HMM model for *cis*-element clusters in higher eukaryotic DNA, which shows promising performance in motif scanning (for which the PWMs are given). Our goal in this paper is to develop an expressive modular motif model that builds on these previous lines of research.

We present a *de novo* motif detection algorithm using an HMM as the global distribution model and an HMDM as the local alignment model. The resulting composite **LOGOS** model is capable of: (1) performing formal and efficient inference of global motif occurrences under a flexible setting that allows clustered motif instances, multiple motif types, motifs on reverse complementary sequences; (2) correctly enforcing the non-overlapping constraint; (3) capturing site dependencies inside the motifs so as to bias prediction toward more biologically plausible motifs while remaining flexible with regards to motif shapes and lengths; and (4) incorporating prior knowledge of nt composition at each motif site to provide smoothed and robust Bayesian estimation of the PWMs.

2.3 The local model: an HMDM for motif alignment

The local alignment model is crucial for identifying the correct motif patterns in a noisy background. As mentioned before, many motifs are not uniformly well-conserved at all their sites[5] (e.g., *gal4* in Figure 1). Biological evidence shows that conserved sites are likely to occur consecutively [6]. This is called *site clustering*, one of the main motivations for the HMDM model. Obviously the PM model can not model such patterns: given a length L motif for which only $\frac{L}{2}$ positions are conserved, PM would assign the same probability regardless of the locations of the conserved sites.

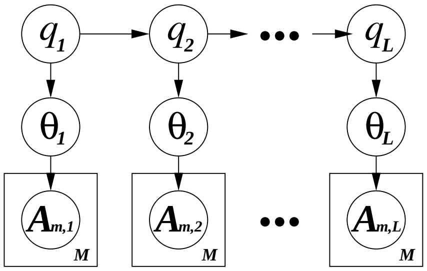


Figure 2. The HMDM model for motif instances specified by a given x. The circles are random variables and the boxes are plates representing replicates (i.e., M instances of a motif).

In the HMDM model (Figure 2), we assume that there are I underlying latent nt-distribution prototypes[6], according to which position-specific multinomial distributions of nt are determined, and that each prototype is represented by a Dirichlet distribution. Furthermore, the sequence of prototypes at consecutive positions in the motif is governed by a first-order Markov process.

More precisely, a multi-alignment $\mathbf{A}$ containing M motif instances is generated by the following process. First we sample a sequence of prototype indicators $q = (q_1, \ldots, q_L)$ from a first-order Markov chain with initial distribution π and transition matrix B. Then we repeat the following for each column $l \in \{1, \ldots, L\}$: (1) A component from a Dirichlet mixture $\alpha = \{\alpha_1, \ldots, \alpha_I\}$, where each $\alpha_i = [\alpha_{i1}, \ldots, \alpha_{i4}]'$ is the parameter vector for a Dirichlet distribution, is picked (deterministically) according to indicator q_l. Say, $q_l = i$, thus we picked α_i. (2) A multinomial distribution θ_l is sampled according to $p(\theta|\alpha_i)$, the probability defined by Dirichlet component i. (3) All the nucleotides in column l are generated *i.i.d.* according to the multinomial distribution parametrized by θ_l.

The complete likelihood of motif alignment $\mathbf{A}_{M\times L}$ characterized by a counting matrix h is:

$$p(\mathbf{A}, q, \theta | x, \Theta_l, \mathcal{M}_l) = p(h|x,\theta)p(\theta|q,\alpha)p(q|\pi,B), \quad (2)$$

where (using the update properties of the Dirichlet distribution and denoting $q_l^i = 1$ if q_l is in state i and 0 otherwise):

$$p(h|x,\theta)p(\theta|q,\alpha) = \prod_{l=1}^{L}\prod_{i=1}^{I} \mathrm{Dir}(\alpha_i + h_l)^{q_l^i}, \quad (3)$$

$$p(q|\pi,B) = \prod_{i=1}^{I}[\pi_i]^{q_1^i} \prod_{l=1}^{L-1} \prod_{i,j=1}^{I} [B_{i,j}]^{q_l^i q_{l+1}^j}. \quad (4)$$

---

[5] A possible reason could be that a binding protein only interacts with a DNA target through a few highly specific aa-nt interactions, but is tolerant of variation in other sites.

[6] We can roughly imagine that the set of prototypes should include prototypes corresponding to four possible conserved nt-distributions (i.e., those having most of the probability mass at A, C, G, T, respectively), as well as other prototypes corresponding to distributions that are less conserved or even heterogeneous in different ways.

The major role of the HMDM model is to impose dynamic priors for modeling data whose distributions exhibit spatial dependencies.

As Figure 2 makes clear, this model is *not* a simple HMM for discrete sequences. In an HMM model the transitions would be between the emission models (i.e., multinomials) themselves, and the output at each time would be a single data instance in the sequence. In HMDM, the transitions are between different priors for the emission models, and the direct output of the HMM is the parameter vector of a generative model, which will be sampled multiple times at each position to generate random instances. This approach is especially useful when we have either empirical or learned prior knowledge (e.g., from training motifs) about motif properties such as *site clustering* or other positional dependencies.

2.4 The global model: an HMM for motif indicators

The HMDM generative process only creates aligned multiple instances of a motif, but does not complete the generation of the observed sequence set. We need a model for the background sequences and another process that generates the positions of the motif instances. For this we need a global model for the indicator variable sequence x that can specify the locations of all motif instances.

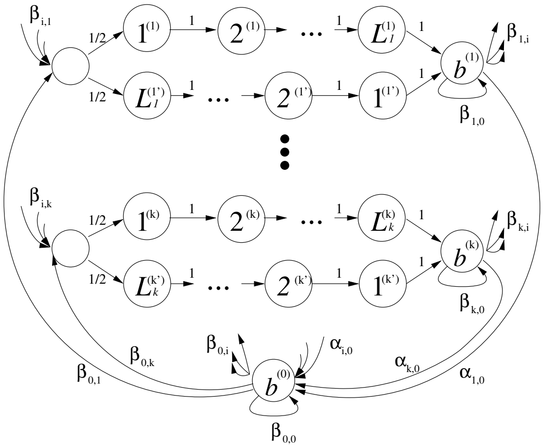


Figure 3. The global HMM (unlabeled circles are dummy states).

Let x be the indicator variable sequence specifying whether each y_t in a DNA sequence is in the background or in a motif, and if in a motif, which motif and where in the motif: $x = (x_1, \ldots, x_T)$, where $x_t \in \mathbb{S}$. The indicator state space $\mathbb{S}$ includes all possible identity labels of a monomer (nt) in a sequence: $\mathbb{S} = \mathbb{M} \cup \mathbb{M}' \cup \{b^0, b^1, \ldots, b^k, d\}$, where $\mathbb{M} = \{1^{(1)} \ldots L_1^{(1)}, 1^{(2)} \ldots L_2^{(2)}, \ldots, 1^{(K)} \ldots L_k^{(K)}\}$ is the set of all possible sites within a motif on the forward strand (i.e., states $1^{(1)}$ to $L_1^{(1)}$ correspond to the sites in motif type 1 on the forward strand, and so on); $\mathbb{M}'$ is the set of all possible sites within a motif on the reverse complementary strand; b^0 corresponds to the inter-cluster background state; $b^k, k \neq 0$ corresponds to the intra-cluster background states; and d represents dummy states. We model the distribution of x with the first-order Markov process depicted in Figure 3.

The motivation for this Markov model is that we expect to see occasional motif clusters in a large ocean of global background sequences (represented by state b^0), and each motif instance in a cluster is embedded in a corresponding sea of intra-cluster background sequences (b^i). The model assumes that the distance between clusters is geometrically distributed with mean $1/(1-\beta_{0,0})$, and the distance between motif instances within cluster k is also geometrically distributed with mean $1/(1-\beta_{k,0})$. As shown in Figure 3, with equal probability $\beta_{k,k}/2$, an intra-background state b^k reaches the start states $1^{(k)}$ and $L_k^{(k')}$ of motif k on the forward or reverse strand, deterministically passes through all internal sites of motif k (thus avoiding motif overlapping), and transits back to the same background state b^k, thereby stochastically generating a cluster of occurrences of motif k; b^k also has a small probability $\beta_{k,i}/2$ of transiting to the start state of another motif i, which terminates cluster k and leads into cluster i; all intra-background states also have probability $\alpha_{k,0}$ of returning to the global background state. These parameters can in principle be fitted using a training set, or just specified empirically based on a rough estimation of the motif or *cis*-module frequencies [7]. Note that these parameters do not impose rigid constraints on the number of motif instances or modules; the actual number of instances is determined by the posterior distribution of the indicator sequence $p(x|y)$.

In accordance with the above framework, we introduce multinomial parameters $\theta_{b_0} = [\theta_{b_0,1}, \ldots, \theta_{b_0,4}]'$ and $\theta_{b_1} = [\theta_{b_1,1}, \ldots, \theta_{b_1,4}]'$ for the inter- and intra-cluster background nt distributions, respectively (assuming all intra-cluster backgrounds use the same multinomial model). Thus, given the PWMs Θ of motifs and the background parameters $\Theta_{bk} = \{\theta_{b_0}, \theta_{b_1}\}$, we have the usual joint probability for a conditional HMM model of a motif-containing sequence:

$$\begin{aligned} & p(x, y|\Theta, \Theta_{bk}, \Theta_g, \mathcal{M}_g) \\ & \quad = \quad p(x_1) \prod_{t=2}^{T} p(x_t|x_{t-1}) \prod_{t=1}^{T} p(y_t|x_t, \Theta, \Theta_{bk}) \qquad (5) \end{aligned}$$

where $$p(y_t|x_t, \Theta, \Theta_{bk}) = \prod_{i \in \mathbb{S}} \prod_{j=1}^{4} [\theta_{ij}]^{\delta(y_t,j)\delta(x_t,i)}.$$

The locations of all motif instances encoded in x can be inferred from the global model using Bayes rule.

---

[7] When no strong knowledge about modular dependencies is available, it is better to just set all bk-to-mt transitions $\beta_{i,j}, i, j \neq 0$, to the same small constant reflecting motif frequency, and similarly for $\beta_{0,k}$ and $\alpha_{k,0}$ reflecting cluster frequency, to avoid overfitting. In our experiment, we parametrize our HMM model in such a fashion. This reduced model is very similar to the one used in Cister [7], but with unknown PWMs in our case.

The HMM model we proposed is not meant to capture fine details of the global motif dependencies, because without a sufficiently large and well-characterized training data set, we could risk overfitting to hypothetical structures and fail to generalize to sequences bearing unknown (and possibly simpler) structures. But within the **LOGOS** framework, if so desired, we can easily generalize to more elaborate models, such as one that models higher order dependencies, or one which uses a more complex background (e.g., a higher-order Markov model) in a principled way. All that is needed is to simply expand the state space $\mathbb{S}$, and either train or empirically parametrize a more expressive initial and transition model in the global HMM.

3 Inference and Learning Algorithm

3.1 Variational Bayesian learning

In order to do Bayesian estimation of the motif parameter θ, and to predict the locations of motif instances via the indicator sequence x, we need to be able to compute the posterior distribution $p(\theta|y)$, which is infeasible in closed form for a complex motif model (because we have to marginalize out q and x in the joint posterior $p(\theta, q, x|y, \mathcal{M})$). A possible approach is to use a Markov Chain Monte Carlo (MCMC) method, such as a Gibbs sampler, which performs "asymptotically exact inference." However, concerns over likely slow mixing and difficulties in detecting convergence motivate us to use *variational Bayesian inference*, which has a more deterministic flavor similar to that of EM and is computationally more efficient.

The variational Bayesian inference method developed in [22] for the HMDM model is a special instance of the *generalized mean field* (GMF) algorithm [23]. Briefly, in the GMF framework, a complex joint distribution p, such as the joint posterior $p(\theta, q, x|y, \mathcal{M})$, is approximated with a simpler distribution Q defined by the product of interdependent local marginals over disjoint subsets of all domain variables, e.g., $Q(\theta, q, x) = Q_l(\theta, q)Q_g(x)$. The optimal form of each local marginal can be obtained via minimizing the Kullback-Leibler (KL) divergence between Q and p with respect to free distributions Q_l and Q_g [23]. Omitting mathematical details, this optimization results in the following coupled updates:

$$\begin{aligned} Q_g(x) &= p(x|\mathcal{M}_g)p(y|x, \bar{\phi}(\theta), \mathcal{M}_g) \quad (6) \\ Q_l(\theta, q) &= p(q|\mathcal{M}_l)p(\theta|q, \alpha, \bar{h}(y), \mathcal{M}_l) \quad (7) \end{aligned}$$

where, E_D denotes expectation with respect to the distribution D, and $\bar{h}(y) = E_{Q_g}[h(x, y)]$, $\bar{\phi}(\theta) = E_{Q_l}[\ln\theta]$ [8], which are referred to as the *generalized mean field messages* exchanged between submodels conducting probabilistic influences of the respective submodel.

A key property revealed in Eqs. (6) and (7) is the isomorphism of their right-hand sides to those of the Eqs. (2) and (5). Essentially, the variational marginals $Q_g(x)$ and $Q_l(\theta, q)$ recover exactly the same form of the original global and local submodels, except that the motif parameters θ on which the global submodel is conditioned are replaced by their Bayesian estimates (in the natural parameter form), and the sufficient statistics h propagated from the global submodel to the local submodel are replaced by their posterior expectations. This means that the locality of inference and marginalization in the composite **LOGOS** model is preserved in both local and global submodels. We can easily obtain the optimal approximate posterior distribution of θ by marginalizing $Q_l(\theta, q)$ over q, and that of x using $Q_g(x)$. It can be further proved that the coupled updates (6) and (7) actually optimize a lower bound of the likelihood $p(y|\Theta, \mathcal{M})$ and are guaranteed to converge to a local maximum (as in standard EM) [13]. In the following section we summarize the computation procedure involved in **LOGOS**, which we call a *"variational EM" algorithm* (VEM), after its operational resemblance to a conventional EM.

3.2 The variational EM algorithm

Due to the locality of variational Bayesian inference, we can perform inference in the local alignment model HMDM as if we have "observations" $\bar{h}$ (to obtain a distribution $Q_l(\theta, q)$ that approximates the marginalized conditional $p(\theta, q|y)$), and in the global HMM model as if the position-specific multinomial distribution of a motif $\bar{\phi}(\theta)$ is given (to obtain $Q_g(x)$ that approximates $p(x|y)$). Therefore, Bayesian estimates of the multinomial parameters can be obtained via fixed-point iteration through the following EM-like procedure:

Variational E step:

Compute the expected sufficient statistics, the count matrix $\bar{h} = E_{Q_g(x)}[h]$, via inference in the global motif model given $\bar{\phi}(\theta)$ and sequence set $\mathbf{y}$:

$$\bar{h} = \sum_{n=1}^{N} \sum_{t=1}^{T_n-L+1} h(y^{(n)}_{t:t+L-1}) p(x^{(n)}_t = 1|y^{(n)}, \bar{\phi}(\theta), \Theta_{bk}), \quad (8)$$

where superscript n indicates the n^{th} DNA sequence; $p(x^{(n)}_t = 1|y^{(n)}, \bar{\phi}(\theta), \Theta_{bk})$ is the posterior probability of position t in sequence n being the start site of a motif given sequence $y^{(n)}$, Bayesian estimate of motif PWMs, and the background, which can be computed using the standard forward-backward algorithm for HMMs on $Q_g(x)$.

Variational M step:

Compute the posterior mean of the natural parameter, $\bar{\phi}(\theta) = E_{Q_l(\theta,q)}[\phi(\theta)]$, via inference in the local motif

---

[8] $\ln\theta$, where $\ln(\cdot)$ is a componentwise operation, is called the *natural parameterization* of a multinomial.

alignment model given $\bar{h}$:

$$\bar{\phi}(\theta_{l,j}) = \sum_{i=1}^{I} p(q_l = i|\bar{h})\big(\Psi(\alpha_{ij} + \bar{h}_{lj}) - \Psi(|\alpha_i| + |\bar{h}_l|)\big), \quad (9)$$

where $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x} = \frac{\Gamma^{'}(x)}{\Gamma(x)}$ is the digamma function, $p(q_l = i|\bar{h})$ is the posterior probability of hidden state q given "observation" $\bar{h}$, which can be computed using the forward-backward algorithm on $Q_l(\theta, q)$.

This *modular inference* procedure provides a framework that scales readily to more complex models. For example, the motif distribution model $p(x)$ can be made more sophisticated so as to model complex properties of multiple motifs such as motif-level dependencies (e.g., co-occurrence, overlaps and concentration within regulatory modules) without complicating the inference in the local alignment model. Similarly, the motif alignment model can also be more expressive (e.g., a mixture of HMDMs) without interfering with inference in the motif distribution model.

The Dirichlet parameters and HMM transition matrix of the HMDM are fitted from a training dataset via empirical Bayes estimation [5].

4 Experiments

In [22], we systematically examined the performance of the HMDM model by implementing a prototype motif detector using HMDM as the local model and testing it on semi-realistic datasets in which biologically identified motifs are planted in a random background, possibly in the presence of artificially produced "false motifs" as decoys. The major advantage of using such a test system is that we know the ground truth, i.e., the true locations and PWMs of the motifs to be detected, and hence can reliably compare performance of different models. We showed that HMDM has a notably higher specificity (than PM) to the genuine motifs in the presence of an artificial decoy, and significantly out-performs the PM-based MEME algorithm in the one-motif-per-sequence scenario.

The **LOGOS** model developed in the current paper integrates HMDM as a subcomponent, which models the motif alignments, accompanied with an expressive HMM model, which models the global distribution of motifs in a biologically more realistic way than the UI model. In the following sections, we examine the performance of **LOGOS** using both semi-realistic datasets and real genomic sequences from yeast. All yeast motif sequences are obtained from the *Promoter Database of Saccharomyces cerevisiae* (SCPD), 15 of which are used to fit the hyperparameters of the HMDM, and others (independent of the training set) are used for testing. We compare three variants of **LOGOS**, ordered with decreasing model expressiveness, HMDM+HMM (**LOGOS**$_{hh}$), PM+HMM (**LOGOS**$_{ph}$) and PM+UI (**LOGOS**$_{pu}$), as well as the MEME and AlignACE program (both of which are essentially the same as **LOGOS**$_{pu}$ in terms of model assumptions, but are enhanced by additional pattern-driven submodels, i.e. gapped motifs, and a more sophisticated implementation).

4.1 Learning the HMDM parameters

We learn our HMDM model using a motif collection from the *Promoter Database of Saccharomyces cerevisiae* (SCPD). Our dataset contains twenty motifs. Each has 6 to 32 instances all of which have been identified via biological experiments.

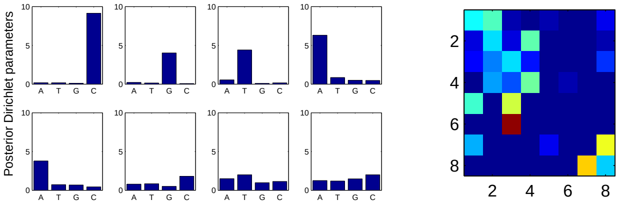


Figure 4. (a) Dirichlet hyperparameters (of components 1-8 from left to right, top to bottom). (b) Markov transition matrix.

We begin with an experiment showing how HMDM can capture intrinsic properties of the motifs. The prior distribution of the position-specific multinomial parameters θ, reflected in the parameters of the Dirichlet mixtures learned from data, can reveal the nt-distribution patterns of the motifs. Examining the transition probabilities between different Dirichlet components further tells us the about dependences between adjacent positions (which indirectly reveals the "shape" information). We set the total number of Dirichlet components to be 8 based on an intelligent guess, and Figure 4a shows the Dirichlet parameters fitted from the dataset via empirical Bayes estimation. Among the 8 Dirichlet components, nos. 1-4 favor pure distribution of single nucleotides A, T, G, and C, respectively, suggesting they correspond to "homogeneous" prototypes, whereas nos. 7 and 8 favor a near uniform distribution of all 4 nt-types, hence "heterogeneous" prototypes. Components 5 and 6 are somewhat in between. Such patterns agree well with the biological definition of motifs. Interestingly, from the learned transition model of the HMM (Figure 4b), it can be seen that the transition probability from a homogeneous prototype to a heterogeneous prototype is significantly less than that between two homogeneous or two heterogeneous prototypes, confirming an empirical speculation in biology that motifs have the so-called *site clustering* property.

4.2 Performance on *semi-realistic* sequence data

4.2.1 Single motif, and multiple instances per sequence

Under a realistic motif detection condition, the number of motif instances is unknown. Rather than trying all possible numbers of occurrences suggested by the user or decided by the algorithm and reporting a heuristically determined plausible number, **LOGOS** uses the global HMM model to describe a posterior distribution of motif instances, which depends on both the prespecified indicator state transition probabilities (which can by empirically supplied by the user to reflect her rough estimation of motif frequencies, or estimated from a training dataset) and the actual sequence y to be analyzed. Due to modularity of variational inference in **LOGOS**, the locations of all instances, which are specified by the indicator sequence x, can be efficiently inferred from the variational marginal distribution $Q(x)$, a standard HMM, using *posterior decoding*, which computes the posterior expectation of x.

Table 1. Performance of **LOGOS** for single motif detection, with unknown number of instances per sequence.

motif	**LOGOS**$_{hh}$		**LOGOS**$_{ph}$		**LOGOS**$_{pu}$	
name	FP	FN	FP	FN	FP	FN
abf1	**0.3115**	**0.2116**	0.6774	0.1957	0.7917	0.9123
gal4	**0.1569**	**0.1569**	0.1895	0.1534	0.2917	0.7939
gcn4	**0.1820**	**0.2355**	0.6142	0.2821	0	0.9594
gcr1	**0.1962**	**0.2134**	0.3371	0.2038	0.3333	0.9437
mat	**0.0723**	**0.0337**	0.3563	0	0.5000	0.9643
mcb	0.3734	0.0910	**0.3628**	**0.0792**	0.3333	0.9431
mig1	**0.0774**	**0**	0.0854	0	0.9764	0.1000
crp	**0.3768**	**0.3398**	0.2727	0.5294	0	0.9487

Table 1 summarizes the performance of three variants of **LOGOS** for single motif detection, with an unknown number of instances per sequence. We present the median false positive (FP) and false negative (FN) rates (in terms of finding each instance of the motifs within an offset of 3bp) of motif detection experiments over 20 test datasets. Each test dataset consists of 20 sequences, each generated by planting 0-7 instances of a motif, together with its permuted "decoy," in a 300-400 bp random background sequence. As Table 1 shows, **LOGOS**$_{pu}$ yields the weakest results, losing in all 8 motif detections (in terms of (FP+FN)/2), suggesting that the conventional PM+UI model, which is used in MEME, and with slight variation, in AlignACE and BioProspector, is not powerful enough to handle non-trivial detection tasks as posed by our testset. **LOGOS**$_{ph}$ improves significantly over **LOGOS**$_{pu}$, even yielding the best performance in one case (for *mcb*), suggesting that the HMM global model we introduced indeed strengthens the motif detector. Finally, as hoped, **LOGOS**$_{hh}$ yields the strongest results, performing best on 7 of the 8 motifs, convincingly showing that capturing the internal structures of motifs and making use of prior knowledge from known motifs, combined with the use of the HMM global model, can yield substantially improved performance. Our results are reasonably robust under different choices of the global HMM parameters, but due to space limitations, we omit details.

4.2.2 Simultaneous detection of multiple motifs

Detecting multiple motifs simultaneously is arguably a better strategy than detecting one at a time followed by deleting or masking the detected motifs, especially when motif concentrations are high, because the latter strategy mistakenly treats the other motifs as background, causing potentially suboptimal estimation of both motif and background parameters. The global HMM model we propose readily handles simultaneous multiple motif detection: we only need to encode all motif states into the state space $\mathbb{S}$ of the motif indicator x, and perform standard HMM inference. The locations of all motifs can be directly read off from the state configuration of x. Table 2 summarizes the results on 20 testsets each containing 20 sequences harboring motifs *abf1*, *gal4* and *mig1*. The upper panels show the predictive performance based on the optimal (in terms of maximal log-likelihood of y from 50 independent runs of the VEM) posterior expectation of x. Note that with a HMDM local model, **LOGOS**$_{hh}$ exhibits better performance. In the lower panels, we show the best FP-FN results in the top three predictions made by **LOGOS** (note the 'k-at-a-time' prediction yields a total of $3k$ possibly redundant motif patterns). This is close to the stochastic dictionary scenario where the predicted motif is to be identified from optimal dictionary of patterns resulting from the motif detection program [9]. It is expected that a human observer could easily pick out the biologically more plausible motifs when given a visual presentation of the most likely motifs suggested by a motif finder.

Table 2. Simultaneous multiple motif detection (median FP-FN rate over 20 testsets containing three motifs.)

	LOGOS$_{hh}$		**LOGOS**$_{ph}$	
	FP	FN	FP	FN
abf1	0.3591	0.3274	0.7778	0.7434
gal4	0.1259	0.1714	0.3751	0.1491
mig1	0.3849	0.2243	0.3481	0
abf1	0.3841	0.2400	0.4721	0.3972
gal4	0.0926	0.0986	0.2609	0.1255
mig1	0.1250	0.0333	0.2318	0

4.2.3 Detecting motifs of uncertain lengths

A useful property of the HMDM submodel is that it actually does not need to know the exact lengths of the motifs to be detected, since HMDM allows a motif to start (and end) with consecutive heterogeneous sites. Thus, a blurred motif boundary is permissible, and as a result, we do not have to know the exact length of the motif. This is another appealing feature of **LOGOS**, which extends its flexibility. As shown in Table 3, even in simultaneous multiple motif detection, with improperly specified motif lengths, HMDM+HMM performs nearly as well as when motif lengths are precisely specified, whereas PM+HMM is not as good.

4.3 Performance on real genomic sequence data

4.3.1 Motif detection in yeast promoter regions

In this section we report a performance comparison of **LOGOS** (HMM+HMDM) with two popular motif detection

[redacted]

	LOGOS$_{hh}$		LOGOS$_{ph}$	
	FP	FN	FP	FN
abf1	0.7295	0.6667	0.8021	0.7680
gal4	0.1167	0.2042	0.2357	0.1325
mig1	0.4183	0.2128	0.8150	0.8381
abf1	0.3310	0.2804	0.5742	0.4821
gal4	0.0955	0.1222	0.1882	0.1250
mig1	0.2124	0.1327	0.3218	0.1623

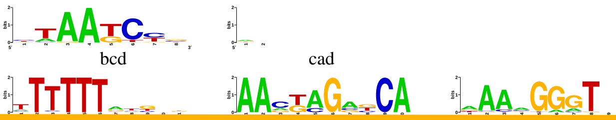
programs, MEME and AlignACE, on 12 yeast genomic sequence sets gathered from the SCPD database (the selection is based on having at least a total of 5 motif instances in all sequences and the motif being independent of our training set). Each sequence set consists of multiple yeast pro- [redacted] rightmost column gives the number of sequences in each [redacted] input sequences and the possible presence of motifs other than what has been annotated make the motif finding task significantly more difficult than a semi-realistic test data or [redacted] [redacted] tcm -revcomp -nmotifs 1 ". In practice, this means that we search for a DNA sequence on both strands for at most one motif, which can occur zero or more times in any given sequence. AlignACE is run with default command-line arguments nearly identical to those for MEME, with the only difference that AlignACE can return multiple predicted motifs (of which we select the best match from the top five MAP predictions). **LOGOS** is set in the multiple-detection mode and is used to make two motif predictions simultaneously. As shown in Table 4, for this non-trivial *de novo* motif detection task, **LOGOS** outperforms the other two programs by a significant margin.

Table 4. Comparison of motif detectors on yeast promoter sequences.

set name	LOGOS		MEME		AlignACE		seq no.
	FP	FN	FP	FN	FP	FN	
abf1	0.7949	0.6522	1.0000	1.0000	**0.5294**	**0.6087**	20
csre	**0.4444**	**0.1667**	0.7778	0.5000	0.8000	0.5000	4
gal4	**0.1333**	**0.0714**	0.1667	0.2857	0.3333	0.1429	6
gcn4	**0.3529**	**0.1852**	1.0000	1.0000	0.3333	0.5556	9
gcr1	**0.2859**	**0.6154**	1.0000	1.0000	0.4545	0.4615	6
hstf	0.8571	0.5556	**0.6000**	**0.5556**	0.8500	0.6667	6
mat	**0.4194**	**0**	0.3750	0.5625	0.2500	0.2500	7
mcb	0.4706	0.2500	0.2000	0.3333	**0.2500**	**0.2500**	6
mig1	**0.8077**	**0.2857**	1.0000	1.0000	0.8333	0.7857	22
pho2	**0.9024**	**0.5000**	1.0000	1.0000	1.0000	1.0000	3
swi5	**0.7647**	**0.5000**	1.0000	1.0000	0.9412	0.7500	2
uash	**0.8250**	**0.6818**	1.0000	1.0000	0.9231	0.9545	18

4.3.2 Motif detection in Drosophila regulatory DNAs

In this section we report on a preliminary *de novo* motif discovery analysis of the regulatory regions of the 9 Drosophila [redacted] PWMs postulated by Berman et al. for these five motifs, which were used in their *motif scan* analysis (Figure 5). The sources of all PWMs are biologically identified sequence segments from the literature (which are unaligned, ranging from 5 to 93 instances per motif, and about 20 ∼ 40 bases in length). The PWMs are derived from an alignment of all these identified motif sequences.



[redacted]

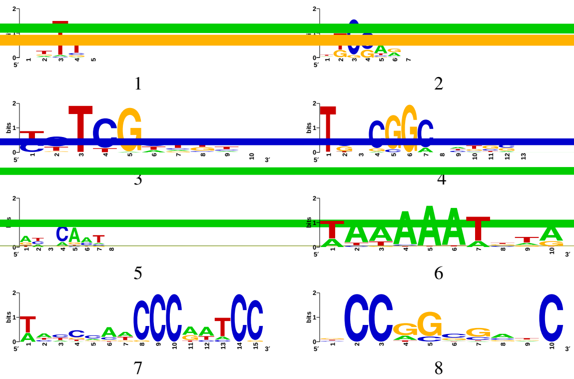


Figure 6. Motif patterns detected by **LOGOS** in the regulatory regions of 9 Drosophila genes.

We apply **LOGOS** (which is set to identify 4 motifs at a time) to the Drosophila dataset and Figure 6 gives a partial list of the top-scoring (data likelihood under the **LOGOS** model) motif patterns. Note that the *logos* shown here are not the conventional sequence *logos* based on counts of aligned nucleotides; instead we use the *logo* visualization software to graphically present the **Bayesian estimate** of the position-specific multinomial parameters θ of each motif, so they are not necessarily equal to the usual nt frequencies of aligned sequences, but represent a more robust probabilistic model of the motif sequences. A visual inspection reveals that patterns 1 and 5 correspond to the *hb* and *cad* binding sites (as confirmed by the matching of the locations of our results and the sequence annotations). Part of pattern 2 agrees with the reverse complement of the *kr* motif (containing -CCCxTT-), but this motif seem to be actually a "two-block" motif because the pattern we detected under a longer estimated motif length contains an additional co-occurring conserved pattern a few bases upstream. Part of pattern 7 is close to the *bcd* motif (containing -AATCC-)

but also contains additional sites (i.e., the three highly conserved C's upstream). A careful examination of pattern 6 suggests that it may be actually derived from putative motif subsequences that correspond to the *kni* binding site. This is not obvious at first because it appears quite different from the *kni logo* in Figure 5. But after seeing an example *kni* site in stripe2/7: 5'agaaaactagatca3', starting at position 35, we realized that the answer might be plausible. The discrepancy is likely due to the artifacts in the original generation of the alignment data supporting the *kni logo*: only 5 biologically identified instances were used and they are quite diverse; the resulting multiple alignment is visually suboptimal in that homogeneous sites are severely interspersed with heterogeneous sites. Patterns 3, 4, and 8 are putative motifs not annotated in the input sequences. We also ran the same dataset through MEME and the output (not shown here) is much weaker and harder to interpret. Note that the motif logos given in Figure 5 are based on the nucleotide-frequency profiles of biologically identified instances from many sources. Thus it is not surprising that some of the patterns we found are similar but do not match the *logos* in Figure 5 exactly since our logos are derived from Bayesian estimates of the motif parameters and our data source consists of the 4 regulatory regions of the *even-skipped* gene, which might be smaller and less representative compared to the data source underlying Figure 5 (except for *kni*).

5 Conclusions

We have presented a principled generative probabilistic framework for modeling motifs in biopolymer sequences. A modular architecture is proposed, which consists of a local submodel of motif alignment, and a global submodel of motif distribution.

We use an HMDM model for local motif alignment, which captures site dependencies inside motifs and incorporates learnable prior knowledge from known motifs for Bayesian estimation of the PWMs of novel motifs in unseen sequences. We use an HMM model for the global motif distribution, which introduces simple dependencies among motif instances and allows efficient and consistent inference of motif locations. A deterministic algorithm, variational EM, is developed to solve the complex missing value and Bayesian learning problems associated with our model. VEM allows probabilistic inference in the local alignment and the global distribution submodel to be carried out virtually separately with a proper Bayesian interface connecting the two processes. This *divide and conquer* strategy makes it much easier to develop more sophisticated models for various aspects of motif analysis without being overburdened by the somewhat daunting complexity of the full motif problem.

As discussed at length in [22], the HMDM model describes a rich continuous distribution of PWMs whose position-specific parameters follow first-order Markov dependences, and which are not captured in most contemporary motif detection algorithms. Nevertheless, although it is a more expressive model, we realize that the actual dependencies inside the motif could be even more complex; further investigations into these properties and more powerful models are needed. Similarly, the HMM global model we propose is only a first step beyond the conventional UI model, and is only able to capture dependencies between motifs and motif clusters at a very limited level (e.g., it can not model higher order dependencies such as hierarchical structures and long-distance influence between motifs). More expressive models are needed to achieve these goals. Nevertheless, under the **LOGOS** architecture, extensions from baseline models are modular and the probabilistic calculations involved can also be handled in a *divide-and-conquer* fashion via generalized variational inference. We are in the process of developing more expressive versions of **LOGOS**. In particular, recent work by Liu et al. [17] and Gupta et al. [9] have motivated us to pursue combination of the dictionary-based models with our approach to capture richer motif properties in complex sequences. We are optimistic that **LOGOS** can serve as a flexible framework for motif analysis in biopolymer sequences.

Acknowledgments

We thank Prof. Michael Eisen for helpful discussions on motif structures.

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learning*, 21:51–80, 1995.

[2] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. USA*, 99:757–762, 2002.

[3] H. Bussemaker, H. Li, and E. Siggia. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA*, 97, 2000.

[4] E. H. Davidson. *Genomic Regulatory Systems*. Academic Press, 2001.

[5] B. Efron. Empirical Bayes methods for combining likelihoods (with discussion). *J. Amer. Statist. Assoc.*, 91:538–565, 1996.

[6] M. Eisen. Structural properties of transcription factor-DNA interactions and the inference of sequence specificity. submitted, 2003.

[7] M. C. Frith, U. Hansen, and Z. Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17:878–889, 2001.

[8] D. GuhaThakurta and G. D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinform.*, 17:608–621, 2001.

[9] M. Gupta and J. Liu. Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Amer. Statist. Assoc.*, 98, 2003.

[10] J. V. Helden, A. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 28:1808–1818, 2000.

[11] G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinform.*, 15:563–577, 1999.

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

[13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, 1999.

[14] K. Kechris, E. van Zwet, P. Bickel, and M. Eisen. Detecting DNA regulatory motifs by incoporating position-specific base conservation. *submitted*, 2003.

[15] J. Liu, X. Liu, and D. Brutlag. Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Proc.of PSB*, 2001.

[16] J. Liu, A. Neuwald, and C. Lawrence. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc*, 90:1156–1169, 1995.

[17] X. S. Liu, D. L. Brutlag, and J. S. Liu. An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–9, 2002.

[18] L. A. Pennacchio and E. M. Rubin. Genomic strategies to identify mammalian regulatory sequences. *Nature Reviews Genetics*, 2(2):100–109, 2001.

[19] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucl. Acids Res.*, 18:6097–6100, 1990.

[20] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends in Biochemical Sciences*, 23:109–113, 1998.

[21] L. Stryer. *Biochemistry (4th. edition)*. W. H. Freeman and Company, 1995.

[22] E. P. Xing, M. I. Jordan, R. M. Karp, and S. Russell. A hierarchical Bayesian Markovian model for motifs in biopolymer sequences. In *Proc. of Advances in Neural Information Processing Systems 16*, 2003.

[23] E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the 19th Annual Conference on Uncertainty in AI*, 2003.