Mixture Model based MAP Motif Discovering

Konstantinos Blekas
Department of Computer Science, University of Ioannina
GR-45110 Ioannina, Greece
Tel. +30 26510 98816, Fax. +30 26510 98890
E-mail: kblekas@cs.uoi.gr

Abstract

In this paper a new maximum a posteriori (MAP) approach based on mixtures of multinomials is proposed for discovering probabilistic motifs in sequences. The main advantage of the proposed methodology is the ability to bypass the problem of overlapping motif occurrences among neighborhood positions in sequences through the use of a Markov Random Field (MRF) as a prior. This model consists of two components, the first is responsible for modeling a motif and the second corresponds to the background. The Expectation-Maximization (EM) algorithm is used to estimate the model parameters and provides closed form update rules. Special care is also taken to produce good initial values for the motif multinomial model, in order to overcome the known dependence of the EM algorithm to initialization. This is done by applying an adaptive agglomerative clustering procedure that provides candidate initial models. Experiments with both artificial and real sets of biological sequences show the advantages of the proposed approach in discovering qualitatively better motifs, in comparison with the classical maximum likelihood (ML) approach and the Multiple EM for Motif Elicitation (MEME) method which uses also an ML-based mixture model.

Keywords: Motif discovering, Markov Random Field (MRF), mixture of multinomials model, Expectation-Maximization (EM), agglomerative clustering.

1 Introduction

Discovering motifs (or patterns) in biological sequences is an important problem in computational biology. Given a set of sequences, such as a DNA or a protein sequence, a motif can be represented as a common substring that is repeated in the set. The motif discovering problem is related to other problems in Biology, such as the multiple sequence alignment problem, and can be also found in other application areas apart from Biology. Sequence motifs are focused on highly conserved residues present in active sites of sequences and can be used to assign functions to newly sequenced genes or proteins [1, 2]. Motifs can also enclose diagnostic features for families in the sense of generating rules for classification purposes.

Various methods have been introduced for solving this problem that are distinguished according to the model of the motif [2, 3]. Under the Bayesian framework, a motif can be modeled using independent multinomial distributions for its positions. Gibbs sampling [4], the MEME [5], the SAM [6], the BioProspector [7], the Greedy EM [8, 9] and the LOGOS [10] represent statistical methods for discovering shared motifs in a set of (unaligned) sequences. They all formulate the problem using either mixture models or hidden Markov models, and use the Expectation-Maximization (EM) algorithm [11, 12] or variational EM schemes to estimate the model parameters.

The application of statistical methods to discovering sequence-motifs usually forces the assumption that all the possible starting positions in sequences are independent. Nevertheless, the problem has the particular characteristic that spatial information should be taken into account. That is, apart from the content of a subsequence, its location must be also used in order to determine its posterior probability for matching it as motif given the subsequence. In other words, it is not desired to identify overlapping motifs. In most of these methods that discover motifs, the common framework used is the maximum likelihood (ML). Under this prism, the motif model parameters are estimated by maximizing the likelihood of the observations, while the spatial constraints are indirectly enforced to the model. This is done by either renormalizing the estimated posterior probability values during each EM step [5], or simply by throwing away any overlapping motif samples when using Gibbs sampling strategies [13]. Therefore, in a sense, there is an inconsistency between the computed motif distribution and the one defined by the model [10].

In this paper we present a *maximum a posteriori* (MAP) approach that provides a direct method to implement these ideas. The basic scheme is a two-component mixture of multinomials model, where one component models the motif and the other the remaining non-motif regions (background). Following this framework, a likelihood term is used to capture the content information of the data, while a bias term is also used to capture the spatial information of the neighborhood locations. This is accomplished by considering the motif labels of each starting position of sequences through a Markov Random Field (MRF) [14, 15] as *a prior* model. This constrains the local characteristics of the sequences and thus provides useful information to the motif estimation process. Furthermore, we consider Dirichlet priors for the multinomial parameters that mostly act as smoothers, since they are conjugate. The EM algorithm is used to estimate the model parameters which provides closed form update equations for all parameters. Since the EM algorithm is very sensitive to the initial parameter values, we also present an agglomerative hierarchical clustering scheme for producing candidate multinomial models for initializing motifs. Finally, borrowing this technique from the MEME approach, multiple motifs are discovered by iteratively applying the two-component mixture model after erasing old motif occurrences. As will be demonstrated in the experimental study of this paper, in contrast to the classical unconstrained mixture model the proposed one overcomes the problem of overlapping subsequences. It also estimates qualitatively better motif models when treating motifs as diagnostic features for classifying sequence families, as compared with the ML and the MEME approaches.

Section 2 presents the two-components mixture of multinomials model that is used for discovering a single motif in two methods: the classical maximum likelihood and the proposed maximum a posteriori approach. Experimental results are given in section 3 using both artificial and real sets of sequences, while section 4 presents conclusions and discussion.

2 Mixture models for discovering motifs

Consider a finite set $\Sigma = \{c_1, \dots, c_\Omega\}$ consisting of $\Omega = |\Sigma|$ individual characters. An arbitrary string over the set Σ is any sequence $S_j = \{s_{jk}\}_{k=1}^{L_j}$ of length L_j, where $s_{jk} \in \Sigma$ denotes the character at the k-th position of the sequence S_j. Now, let $S = \{S_1, \dots, S_N\}$ be a set of N strings of length $L_1, \dots, L_N$, respectively. The motif discovering problem is to find a common subsequence of length K that is repeated at different sites among the sequences of set S.

In order to deal with this, we collect all the possible substrings of set S having length equal to K. This can be done by sliding a window of size K in every sequence S_j, obtaining a set of $L_j - K + 1$ substrings of equal length K. Each substring indicates the starting position of a possible motif occurrence in sequences. Therefore, we finally construct a set of n substrings $X = \{x_i\}_{i=1}^{n}$, $n = \sum_{j=1}^{N}(L_j - K + 1)$, that constitute the observation data.

The assumption that the observations x_i are i.i.d. results in discovering overlapping motifs. However, by enforcing spatial constraints one avoids this problem and estimates better motifs [5]. In the next subsections, two mixture model based approaches will be presented: the classical maximum likelihood without any constraint, as well as, a new proposed maximum a posteriori approach that also uses spatial information.

2.1 The ML approach

Lets assume that the set X has been generated from a two-component mixture of multinomials. The first component models the motif with a prior probability of π, while the second one models the background and represents all the subsequences which do not contribute to the motif, with a prior probability equal to $1 - \pi$. The density function $f(x_i|\pi, \Theta)$ of the model for an observation x_i is given by

$$f(x_i|\pi, \Theta) = \pi p(x_i|\theta) + (1 - \pi) p(x_i|b) \ , \tag{1}$$

where $\Theta = \{\theta, b\}$ is the set of parameters for the multinomial density. To parameterize the motif we use a position weight matrix $[\theta_{kl}]$ of size $\Omega \times K$, where each value θ_{kl} denotes the probability that character $c_l \in \Sigma$ is at the k-th position of the motif. For each position k we have $\sum_l \theta_{kl} = 1$. Parameters $b = [b_1, \dots, b_\Omega]$ of the multinomial background distribution are represented with a vector of probabilities ($\sum_l b_l = 1$) with dimension equal to the alphabet size Ω.

Following the multinomial distribution, and assuming independence among positions of the motif, the probability density function of the motif model is

$$p(x_i|\theta) = \prod_{k=1}^{K} \prod_{l=1}^{\Omega} \theta_{kl}^{\delta_{ikl}} \ , \tag{2}$$

where δ_{ikl} is the Kronecker delta function (1 if character c_l is at the k-th position of substring x_i, 0 otherwise). Likewise, the density function of the background model is given by

$$p(x_i|b) = \prod_{l=1}^{\Omega} b_l^{\sum_{k=1}^{K} \delta_{ikl}} \ . \tag{3}$$

Based on the above formulation, the model parameters can be estimated through maximum likelihood (ML). The log-likelihood function is then given by

$$L(X|\pi, \Theta) = \sum_{i=1}^{n} \log f(x_i|\pi, \Theta) \ . \tag{4}$$

The EM algorithm [11, 12] is an efficient framework to estimate the parameters π, $\{\theta_{kl}\}$ and $\{b_l\}$. It requires the computation of conditional expectation z_i of the hidden variables at the E-step, which are given by

$$z_i^{(t)} = \frac{\pi^{(t)} p(x_i|\theta^{(t)})}{\pi^{(t)} p(x_i|\theta^{(t)}) + (1 - \pi^{(t)}) p(x_i|b^{(t)})} \ , \tag{5}$$

while at the M-step the complete log-likelihood is maximized over the model parameters. This gives the following update equations

$$\begin{aligned} \pi^{(t+1)} &= \frac{\sum_{i=1}^{n} z_i^{(t)}}{n} \ , \\ \theta_{kl}^{(t+1)} &= \frac{\sum_{i=1}^{n} z_i^{(t)} \delta_{ikl}}{\sum_{i=1}^{n} z_i^{(t)} \sum_{l=1}^{\Omega} \delta_{ikl}} \ , \\ b_l^{(t+1)} &= \frac{\sum_{i=1}^{n} (1 - z_i^{(t)}) \sum_{k=1}^{K} \delta_{ikl}}{K \sum_{i=1}^{n} (1 - z_i^{(t)})} \ . \end{aligned} \tag{6}$$

The EM algorithm is guaranteed to convergence to a local maximum of the likelihood function and also satisfies all the constraints of the parameters.

Nevertheless, a significant drawback of the ML approach is the fact that the spatial information of the subsequences is not taken into account. This results in the estimation of overlapping subsequences as motif occurrences of the set X, especially in cases where a motif consists of repeated strings of one or two characters. To avoid this problem, the MEME algorithm [5] performs a normalization of the posterior value z_i of the adjacent sequences. This is an "ad-hoc" step so that guarantees in any window of length K the sum of z_i values remains less than or equal to 1. Next, we introduce a new approach that deals with this problem in a more systematic way by modeling the spatial arrangements of a motif using a Markov Random Field (MRF) prior.

2.2 The MAP approach

In the proposed model, the labels $\pi_i = P(motif|x_i)$, the probabilities that the substring x_i is a motif, are considered as model parameters that satisfy the constraint $0 \leq \pi_i \leq 1$. By letting $\Pi = \{\pi_1, \ldots, \pi_n\}$ be the set of label parameters, this model assumes that the density function $f(x_i|\Pi, \Theta)$ at an observation x_i is given by

$$f(x_i|\Pi, \Theta) = \pi_i p(x_i|\theta) + (1 - \pi_i) p(x_i|b) \ . \tag{7}$$

Spatial constraints for the label parameters Π can be introduced based on prior knowledge. A suitable prior that captures this knowledge is the Gibbs distribution function [14, 15] which is given by

$$p(\Pi) = \frac{1}{Z} \exp(-U(\Pi)) \ , \tag{8}$$

where, Z is a normalization constant called the partition function, and $U(\Pi)$ is an energy function given by

$$U(\Pi) = \beta \sum_{i=1}^{n} V_{\mathcal{N}_i}(\Pi) \ . \tag{9}$$

The parameter β is often called regularization parameter. The energy function is a sum of *clique potentials* $V_{\mathcal{N}_i}$ over all possible cliques, where a clique is defined as the set of label parameters $\{\pi_m\}$ within the neighborhood $\mathcal{N}_i$ of the i-th position of a sequence. A similar in principle spatially-constrained model has been also used for solving the image segmentation problems [16].

In this study, we consider as neighborhood $\mathcal{N}_i$ all the m positions around the position i whose corresponding substrings x_m overlaps with the substring x_i. In the general case, there are $2(K-1)$ such sites around each position which are mutually dependent. When a motif is found at position i ($\pi_i \approx 1$), it is desired that all overlapping substrings x_m that belong to the neighborhood $\mathcal{N}_i$ not to be labeled as motifs ($\pi_m \approx 0$). A potential function that guarantees this behavior is the following simple inner product

$$V_{\mathcal{N}_i}(\Pi) = \sum_{m \in \mathcal{N}_i} \pi_i \pi_m \ , \tag{10}$$

because the inner product of similar subsequences is high.

Moreover, we treat motif model parameters θ as random variables and introduce priors for them. Since Dirichlet densities are conjugate to multinomial densities, it is convenient to use them. Thus, for every motif position k we consider a Dirichlet prior of the form

$$p(\theta_k|\alpha_k) = \frac{\Gamma(\sum_{l=1}^{\Omega} \alpha_{kl})}{\prod_{l=1}^{\Omega} \Gamma(\alpha_{kl})} \prod_{l=1}^{\Omega} \theta_{kl}^{\alpha_{kl}-1} \ , \tag{11}$$

where the parameter α_k is a Ω-vector with components $\alpha_{kl} > 0$ and $\Gamma(x)$ is the Gamma function. Adding Dirichlet priors in effect introduces pseudo-counts to every character at each position of a motif. As it will shown later, setting all $\alpha_{kl} > 1$ regularizes the estimation and prevents the estimates of θ_{kl} from approaching the boundary value 0. During the experimental study, the Dirichlet prior parameters were the same for every motif position k and were set equal to $1 + \epsilon_l$, where ϵ_l was some low percentage (e.g. 10%) of the total predefined relevant frequency of character c_l in each examined dataset X.

Given the above prior densities of Eqs. (8), (11) for the model parameters Π and θ, we can formulate the problem as a *maximum a posteriori* (MAP) approach. Therefore, the posteriori log-density function is

$$p(\Pi, \Theta|X) = \sum_{i=1}^{n} \log f(x_i|\Pi, \Theta) + \log p(\Pi) + \sum_{k=1}^{K} \log p(\theta_k|\alpha_k) \ . \tag{12}$$

The use of EM algorithm for MAP estimation of the parameters requires at each step t the computation of the conditional expectation values z_i of the hidden parameters during the E-step

$$z_i^{(t)} = \frac{\pi_i^{(t)} p(x_i|\theta^{(t)})}{\pi_i^{(t)} p(x_i|\theta^{(t)}) + (1 - \pi_i^{(t)}) p(x_i|b^{(t)})} \ , \tag{13}$$

while in the M-step the maximization of the following log-likelihood function of the complete data is performed

$$\begin{aligned} Q(\Pi, \Theta|\Pi^{(t)}, \Theta^{(t)}) = &\sum_{i=1}^{n} z_i^{(t)} \{\log(\pi_i) + \log(p(x_i|\theta))\} + (1 - z_i^{(t)})\{\log(1 - \pi_i) + \log(p(x_i|b))\} \\ &- \beta \sum_{i=1}^{n} \pi_i \sum_{m \in \mathcal{N}_i} \pi_m + \sum_{k=1}^{K} \sum_{l=1}^{\Omega} (\alpha_{kl} - 1) \log(\theta_{kl}) \ . \end{aligned} \tag{14}$$

The function Q is maximized independently for each parameter. Hence, the following update equation for the motif multinomial parameters is obtained

$$\theta_{kl}^{(t+1)} = \frac{\sum_{i=1}^{n} z_i^{(t)} \delta_{ikl} + (\alpha_{kl} - 1)}{\sum_{i=1}^{n} z_i^{(t)} \sum_{l=1}^{\Omega} \delta_{ikl} + \sum_{l=1}^{\Omega} (\alpha_{kl} - 1)} , \tag{15}$$

while for the background model the update rules are the same as in the case of the ML approach (Eq. 7).

In order to maximize the complete log-likelihood function with respect to the label parameters π_i, we set the derivative of Q equal to zero. This gives the following quadratic expression

$$\mathcal{B}_i (\pi_i^{(t+1)})^2 - (1 + \mathcal{B}_i)(\pi_i^{(t+1)}) + z_i^{(t)} = 0 , \tag{16}$$

where we have substituted for simplicity the contribution of the Gibbs function $\beta \sum_{m \in \mathcal{N}_i} \pi_m$ with the term $\mathcal{B}_i$. It must be noted that this term can also include positions with updated labels ($\pi_m^{(t+1)}$), as well as positions whose labels have not yet been updated ($\pi_m^{(t)}$). The above equation has two roots

$$\pi_i^{(t+1)} = \frac{(1 + \mathcal{B}_i) \pm \sqrt{(1 + \mathcal{B}_i)^2 - 4\mathcal{B}_i z_i^{(t)}}}{2\mathcal{B}_i} . \tag{17}$$

It can be easily shown that only the root with the negative sign is valid, since the other one is discarded due to the constraint $0 \leq \pi_i \leq 1$. Therefore, the above equation provides a simple update of the label parameters π_i during the M-step of the EM algorithm and ensures that the solution is unique and satisfies the constraints.

Looking carefully at Eq. 17 we can make some useful observations. In the case where a motif occurrence starts from position i, and thus the substring x_i has high posterior probability value ($z_i^{(t)} \approx 1$), the following two things may be happening within the neighborhood $\mathcal{N}_i$:

- None of sites $m \in \mathcal{N}_i$ is labeled as a motif, i.e. $\mathcal{B}_i = \beta \sum_m \pi_m \lessapprox 1$, and thus, following Eq. 17, this site will be labeled as motif ($\pi_i^{(t+1)} \approx 1$).

- There is at least one motif in the neighborhood $\mathcal{N}_i$, i.e. $\mathcal{B}_i = \beta \sum_m \pi_m \gtrapprox 1$, and thus, from the update rule of Eq. 17, the new label value will be approximately $\pi_i^{(t+1)} \approx \frac{1}{\mathcal{B}_i}$. The larger the value of $\mathcal{B}_i$, the smaller the update label values of π_i. In this overlapping neighborhood, only one occurrence will be the most probable to survive, that having the higher posterior value z_i.

On the other hand, when a substring x_i has small posterior value of being a motif ($z_i^{(t)} \approx 0$) it will continue being labeled as background ($\pi_i^{(t+1)} \approx 0$), independently of its neighborhood $\mathcal{N}_i$.

From the above analysis it is clear that the regularization parameter β of the Gibbs distribution function plays a significant role. Only large values of this parameter ($\beta \gg 1$) are acceptable in order to discourage overlapping substrings being labeled as motifs. However, in our experiments, a large range of values of β seems to yield similar behavior. This implies that the proposed method is not sensitive to the value of this parameter. A typical value that has been used with success in the experimental study is $\beta = 100$.

2.3 An agglomerative clustering method for initialization

The two previous subsections described two model-based approaches for discovering motifs in sequences based on mixture models: the classical ML approach, and a new spatially-constrained approach through a MAP framework. Both schemes use the EM algorithm to estimate the model parameters.

A common problem of the EM algorithm is its dependence on the initial values of the density parameters. This may cause it to get stuck of local maxima of the likelihood function [12]. In our study, the problem of poor initialization is concentrated in the selection of motif multinomial parameters θ, since for the background density we can use the relative frequencies of each character c_l in the dataset X. To overcome this problem a number of different approaches have been proposed in the literature. The MEME algorithm, for example, uses dynamic programming which estimates the goodness of many possible starting points based on the likelihood of the model after one EM iteration [5]. Another method proposed in [8] applies a divisive hierarchical clustering approach based on kd-trees that generates candidate motif models. Here, we use an adaptation of the *agglomerative clustering* (AC) approach [17, 18] over the set X which is based on a bottom-up generation of a tree-like structure.

This method starts with a set of n clusters (nodes), each one containing one data sample x_i from the set X. Each cluster v that contains n^v samples generates a multinomial distribution ϑ^v based on their sufficient statistics. In particular, multinomial parameters are estimated as $\vartheta^v_{kl} = n^v_{kl}/n^v$, where $n^v_{kl} = \sum_{x_i \in v} \delta_{ikl}$, and $\delta_{ikl} = 1$ if the character c_l is found at the position k of the substring x_i and 0 otherwise. At each step the algorithm searches the set of current clusters to identify the two closest clusters (v, u) that are merged into a new cluster denoted by $v \cup u$. This is accomplished by using an intercluster distance $D(v, u)$ that is defined as [17, 18]

$$D(v, u) = L_v(\vartheta^v) + L_u(\vartheta^u) - L_{v\cup u}(\vartheta^{v\cup u}) \;, \qquad (18)$$

where the quantity $L_v(\vartheta^v)$ represents the log-likelihood value that characterizes the cluster v and is given by

$$L_v(\vartheta^v) = \sum_{x_i \in v} \sum_{k=1}^{K} \sum_{l=1}^{\Omega} \delta_{ikl} \log \vartheta_{kl}^{v} \ . \tag{19}$$

The algorithm terminates when a predetermined number C of clusters (parents) are found. In our study we used $C = n/2$. Among the C final clusters, we find then the one v^* whose multinomial distribution function differs more than the (global) background b of alphabet characters distribution, based on the *Kullback Liebler* (KL) distance metric, i.e.

$$v^* = \arg\max_{v \in C} \{\mathrm{KL}(\vartheta^v || b)\} = \arg\max_{v \in C} \{\sum_{k=1}^{K} \sum_{l=1}^{\Omega} \vartheta_{kl}^{v} \log \frac{\vartheta_{kl}^{v}}{b_l}\} \ . \tag{20}$$

Thus, we can initialize the motif model parameters $\theta^{(0)}$ with those of the multinomial ϑ^{v^*}.

It must be also noted that during the reduced tree construction, it is not allowed to merge two clusters that contain overlapping substrings x_i. Therefore, the merging criterion is not only the distance measure of Eq. 18, but also the corresponding positions of substrings between the two clusters examined. This is a modification of the original AC scheme that takes also into account spatial information. As will be experimentally shown, this strategy is very efficient for constructing better clusters and independent candidate multinomial densities.

The drawback of the AC method is its large computational complexity (order of n^2) especially when the number of subsequences n is huge. In order to reduce it a possible solution is to iteratively apply the AC procedure to smaller portions of the set of sequences S in order to produce candidate multinomial models.

2.4 Discovering multiple motifs

The two previously presented approaches use a two-component mixture of multinomials model to discover a single motif. We can extend this framework for the identification of multiple motifs in a family of sequences S. This can be accomplished by iteratively applying the mixture model to the set of observations, after erasing from S the motifs that were already found. It must be noted that a similar strategy has been also proposed in [5].

In particular, after convergence of the EM algorithm and estimation of the motif multinomial model parameters, all substrings x_i whose label parameters π_i surpass a threshold value T (e.g. $T = 0.9$) are deleted from the set S. A new set S' is then created, $S' = S - \{x_i : \pi_i > T\}$. Finally, the two-component mixture model is repeatedly applied to the new set of observations X' (constructed from the set S') to discover other motifs.

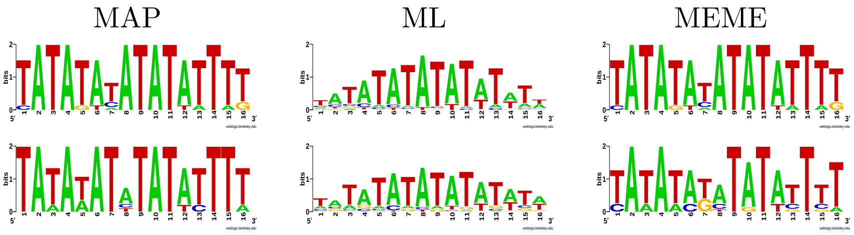

Figure 1: The sequence logos of the DNA motif TATATATATATTTT are shown as estimated by the three methods.

3 Experimental results

Several experiments were performed using both artificial and real sets of sequences in an attempt to study the effectiveness of the proposed MAP approach. During all experiments the AC method was first applied to generate candidate multinomial models used for initializing the motif model and then the EM algorithm was applied for MAP estimation of the model parameters. The label parameters π_i of the model were all initialized to $\pi_i = 0.5$.

Comparative results have been also obtained using the ML approach without spatial constraints, as well as the MEME[1] method using the same sets of sequences. The motifs discovered by all three methods were evaluated using information content-based and diagnostic criteria. It must be noted that the motif multinomial parameters of both the ML and the MAP approaches were initialized identically.

3.1 Experiments with artificial sequences

For the artificial data used in our experiments, we generated sets of sequences with artificial motifs of equal length K. Every sequence had a mutant copy of each motif according to a probability p_m for position-specific mutation, while the rest (non-motif) positions were filled with arbitrary characters using a uniform distribution over the alphabet Σ. In this way, $N = 10$ number of artificial sequences were generated in all cases with mean length 100 characters from the alphabet Ω.

At first we examined the capability of the proposed MAP approach to clearly identify motifs containing repeated characters without estimating overlapping copies of them. For this purpose we created two such motifs, one from the DNA alphabet ($\Omega = 4$) and the other from the protein alphabet ($\Omega = 20$). Assuming two values for the mutation probability $p_m = \{0.1, 0.2\}$, four sets of artificial sequences were generated, respectively. The results of

[1] Experiments with the MEME have been made using a related software that can be downloaded from the site: meme.sdsc.edu/meme/website/meme-download.html

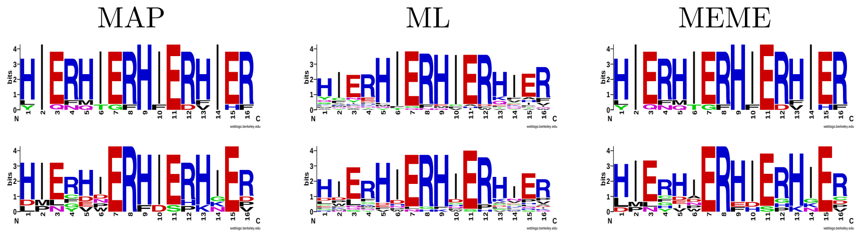

Figure 2: The sequence logos of the protein motif HIERHIERHIERHIER are shown as estimated by the three methods.

the three methods which are compared are shown in Figures 1,2. They illustrate the logo of the discovered motif occurrences by each method, using the WebLogo tool [19], which also shows the information content of a motif. In other words, the size of the plotted character is analogous to its information content. As it is clear from these figures the proposed method achieves the proper identification of the motifs and its results are similar to the MEME method. On the other hand, as expected, the unconstrained ML approach fails to distinguishing overlapping copies of motifs in both sets from the two alphabets.

In another series of experiments a set of two motifs was used. Here half of their sites were identical and thus their discovery is hard. Figures 3, 4 illustrate the logo of the patterns found by each method in the case of the two alphabets using identical mutation probability p_m values as previously. These results show the weakness of the MEME approach to separate motifs with high degree of homology, since it discovered only one complex motif obtained from the synthesis of both. On the other hand, the MAP method manages to estimate both of them as two different motifs in all cases. Also, the ML approach, although it yields good performance in the protein case, is failed in the case of DNA motifs where the overlapping is higher. The different results between the MEME and the (unconstrained) ML approaches for the protein motifs (even if they use the same model) may be explained by the initialization strategy used. In all experiments with artificial datasets the agglomerative clustering scheme seeded both ML and MAP mixture models with a proper motif multinomial model and the estimation process of the ML approach depended on the degree of overlapping among the characters of the motif. Finally, similar observations can be made in the logos of Figure 5 that show the results of another experiment, where two protein motifs are used that have identical odd positions.

3.2 Experiments with real sequences

We have also tested the proposed MAP approach using real biological sequences selected from the PROSITE database that contains families of protein sequences [20]. Four such datasets

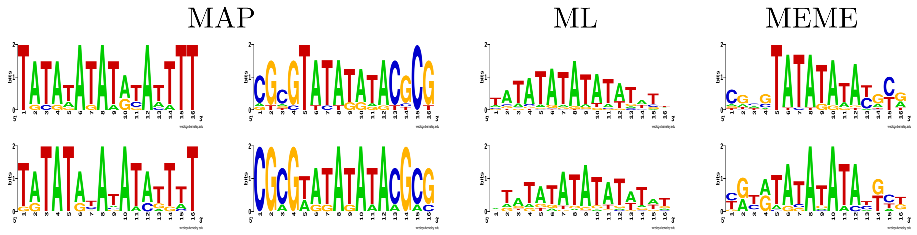

Figure 3: The estimated sequence logos of two half-identical DNA motifs: {TATATATATATATTTT, CGCGTATATATACGCG} are shown.

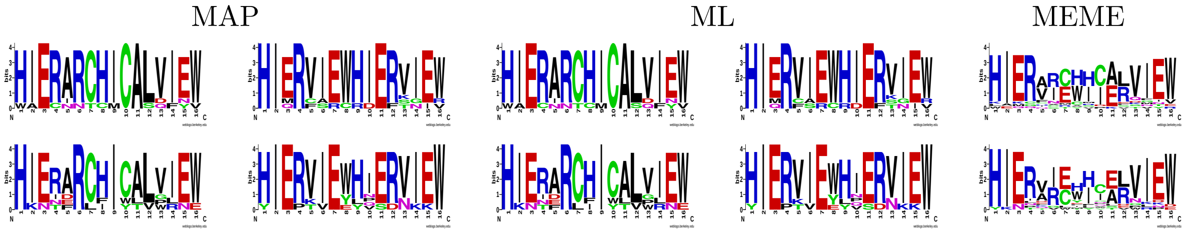

Figure 4: The estimated logos of two half-identical protein motifs: {HIERARCHICALVIEW, HIERVIEWHIERVIEW} are shown.

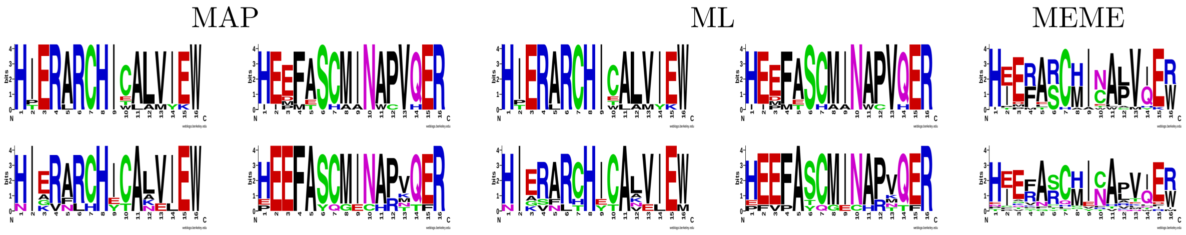

Figure 5: The estimated logos of two protein motifs that have their odd positions identical are shown.

Prosite family	Family size	Training sequences	Number of motifs found by MAP	 ML	 MEME
PS00651 (L9)	83	10 (172 bps)	4	4	3
PS00359 (L11)	215	10 (180 bps)	9	7	7
PS00783 (L13)	66	10 (167 bps)	11	10	8
PS00049 (L14)	92	10 (122 bps)	5	5	3

Table 1: Four PROSITE families used for the experimental study on real sequences.

are summarized in Table 1 that correspond to signatures from different ribosomal proteins. For each family only a small portion of sequences N was selected as the training set for our experiments. The MAP approach, as well as the ML and the MEME methods were then applied to each training set to discover multiple motifs. Here the size of motifs was set fixed $K = 15$. The number of the motifs discovered by each method is shown in Table 1. From the results of this table the capability of the proposed method to identify larger groups of motifs is apparent.

To evaluate the discovered motifs by each method we used the Motif Alignment and Search Tool (MAST) algorithm [21] that constitutes a sequence homology searching tool for matching multiple motifs against a set of sequences. The MAST method uses the multinomial motif models (log-odds scores) and computes the statistical significance (E-value) of the matches of the input group of motifs to a target sequences. In our experiments we have used the whole current set of the SWISS-PROT protein sequences database (number of entries 163200 sequences) as the target set of sequences. By specifying a threshold for the calculated E-value, all the target sequences having E-values lower than this threshold are considered as positives. Therefore, for each experiment we measured the number of true and false positives at several E-value thresholds in order to estimate the sensitivity and specificity of the methods compared.

Figure 6 illustrates the performance of the three methods in each PROSITE family using ROC curves (plots of the true positives as a function of false positives for various thresholds). The classification results show improved performance for the proposed method. In all of these plots the curves of the MAP approach were equal or higher than those drawn by the other two approaches. Among the three methods the ML approach showed the poorest performance. Since the proposed method discovers more motifs in sets of sequences it enhances our ability to discriminate between biological families. This experimental study also proves the biological importance of the motifs in classifying sequences. Only a very small training set of sequences seems to be necessary for the composition of family signatures and the provision of high homology searching characteristics.

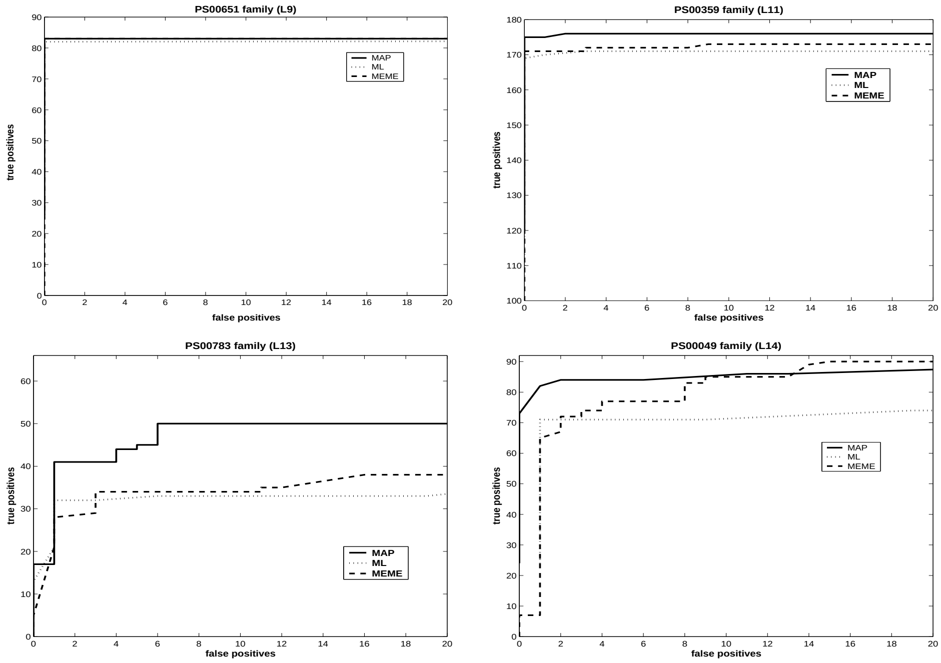

Figure 6: Comparative classification performance in each experimental PROSITE family using ROC curves

4 Conclusions

This paper presents a new spatially-constrained approach which uses MRF priors for discovering probabilistic motifs in sequences. The method uses a mixture of multinomials model with two components for modeling the motif and the background of sequences. The spatial information is embodied in the model by treating the labels of the starting positions as random variables that follow a Gibbs distribution function. The EM algorithm is used for estimating the model parameters, where it is initialized with an agglomerative clustering algorithm that provides candidate multinomial models. Multiple motifs can be found by iteratively apply the basic scheme to the set of substrings after erasing the motif copies found. Experiments have been performed using artificial and real sets of sequences where we evaluated the proposed method and compared it with the classical ML approach without constraints [1] and the known MEME approach [5]. The MAP method was able to clearly identify motifs with repeated characters, similar to the MEME. Moreover, it estimates qualitatively better motif models with noticeable performance, when considering classification tasks, in terms of the diagnostic capabilities of the discovered motifs. Further research can be used to design more complex motif models that can also take into account gaps among sites. This may be useful in cases of low homologies. Considering also variable length motifs is another interesting topic for future study.

References

[1] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach.* MIT Press, Cambridge, 1998.

[2] A. Brāzma, I. Jonasses, I. Eidhammer, and D. Gilbert, "Approaches to the automatic discovery of patterns in biosequences," *Journal of Computational Biology*, vol. 5, no. 2, pp. 277–303, 1998.

[3] B. Bréjova, C. DiMarco, T. Vinař, S. R. Hidalgo, G. Holguin, and C. Patten, "Finding patterns in biological sequences. Project Report for CS798g, University of Waterloo," 2000.

[4] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwland, and J. C. Wootton, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science*, vol. 226, pp. 208–214, 1993.

[5] T. L. Bailey and C. Elkan, "Unsupervised learning of multiple motifs in Biopolymers using Expectation Maximization," *Machine Learning*, vol. 21, pp. 51–83, 1995.

[6] R. Hughey and A. Krogh, "Hidden Markov models for sequence analysis: Extension and analysis of the basic method," *CABIOS*, vol. 12, no. 2, pp. 95–107, 1996.

[7] X. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," in *Pac. Symp. Biocomput*, pp. 127–138, 2001.

[8] K. Blekas, D. I. Fotiadis, and A. Likas, "Greedy mixture learning for multiple motif discovering in biological sequences," *Bioinformatics*, vol. 19, no. 5, pp. 607–617, 2003.

[9] K. Blekas, D. I. Fotiadis, and A. Likas, "A sequential method for discovering probabilistic motifs in proteins," *Methods of Information in Medicine*, vol. 43, no. 1, pp. 9–12, 2004.

[10] E. P. Xing, W. Wu, M. I. Jordan, and R. M. Karp, "LOGOS: A modular Bayesian model for *de novo* motif detection," *Journal of Bioinformatics and Computational Biology*, vol. 2, no. 1, pp. 127–154, 2004.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[12] G. M. McLachlan and D. Peel, *Finite Mixture Models.* New York: John Wiley & Sons, Inc., 2001.

[13] J. S. Liu, A. F. Neuwald, and C. E. Lawrence, "Bayesian models for multiple local sequence alignment and Gibbs sampling strategies," *J. Amer. Statistical Assoc*, vol. 90, pp. 1156–1169, 1995.

[14] J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. Roy. Stat. Soc., ser. B*, vol. 36, no. 2, pp. 192–326, 1975.

[15] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[16] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, "A Spatially-Constrained Mixture Model for Image Segmentation," *IEEE Trans. on Neural Networks (to appear)*, 2005.

[17] M. Meilă and D. Hecherman, "An experimental comparison of model-based clustering methods," *Machine Learning*, vol. 42, pp. 9–29, 2001.

[18] K. Blekas and A. Likas, "Incremental mixture learning for clustering discrete data," in *3d Hellenic Conf. on Artificial Intelligence, (Lecture Notes in Artificial Intelligence)*, vol. 3025, pp. 210–219, Springer-Verlag, 2004.

[19] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: A sequence logo generator," *Genome Research*, vol. 14, pp. 1188–1190, 2004.

[20] C. Sigrist, L. Cerutti, N. Hulo, and et. al., "PROSITE: a documented database using patterns as motif descriptors," *Brief Bioinform.*, vol. 3, pp. 265–274, 2002.

[21] T. L. Bailey and M. Gribskov, "Combining evidence using p-values: application to sequence homology searches," *Bioinformatics*, vol. 14, pp. 48–54, 1998.