A marginal mixture model for discovering motifs in sequences

Elli Voudigari and **Konstantinos Blekas** [1]

Abstract. In this study we present a marginal mixture model for discovering probabilistic motifs in categorical sequences. The proposed method is based on a general framework for developing, extending and marginalizing expressive motif models that encapsulates spatial information. Two alternative schemes for constructing expressive models are described, the extend and the marginal approach. The EM algorithm is applied for estimating the model parameters, while an initialization procedure is used based on the known K-means algorithm. Numerical experiments on various simulated and real sets of sequences demonstrate the advantages of the proposed approach in comparison with the basic maximum likelihood with the EM algorithm scheme and the Gibbs sampling approach.

1 Introduction

Discovering motifs in sequences is an important and attractive pattern recognition problem found in several application areas, such as bioinformatics, web mining, etc. Given a set of discrete sequences, a probabilistic *motif* (or *pattern*) can be seen as a common substring that is noisily repeated in different locations in sequences. The motifs are highly conserved residues present in active sites of sequences and thus they have powerful discriminative abilities in the sense of creating features in classification tasks [4, 5].

In the literature there are various methods that have been introduced for solving this problem, classified according to the type of motifs they produce. The Gibbs sampling [9, 10], the MEME [1], the SAM [8], the BioProspector [11], the Greedy EM [3] and the LOGOS [13], constitute representative statistical approaches for discovering shared motifs in a set of sequences. Most of them use probabilistic generative models to model motifs as stochastic string patterns randomly embedded in a simple background. In such a setting, the motif discovery problem can be seen as a standard missing-value inference and being converted into a parameter estimation problem. In particular, all methods formulate the problem using either mixture of multinomial models or hidden Markov models, and apply standard methods such as the Expectation-Maximization (EM) algorithm [6, 12] or Gibbs sampling[9, 10] schemes to estimate the motif model parameters.

The most common way for stochastically representing a motif is through the position weight matrix (PWM) that gives the distribution of alphabet characters at every position assuming that the positions within a motif are independent. Nevertheless, the simple PWM model description is sensitive to noise and random or trivial recurrent patterns (repetitions of short substrings), and is unable to capture potential site dependencies inside the motif [13]. Various methods have been developed to incorporate spatial information in motifs but, mostly, they are motif specific and handle only special shape motifs [13]. Our aim in this study is to develop more compact expressive motif models that encloses more naturally the important spatial information of motif positions. This is done by introducing a *marginal* mixture model under two schemes that encompass the motif neighborhood. In the first approach (the extended model), the PWM is expanded for capturing the entire neighborhood of a motif and not a single K-mer. Alternatively (the marginal model), the PWM not only can fit the motif occurrences, but also can partially fit the substrings that overlap with them and belong to the same motif neighborhood. These new model capabilities strengthen the expression level of a motif model by indirectly introducing useful spatial operators. Finally, another advantage is that the proposed approach is less sensitive to the initialization that makes the EM algorithm applied for estimating model parameters to be less affected on parameter initialization.

We have made experiments with simulated sets of sequences as well as known biological benchmarks. Comparisons have been also made with rival methods: the basic maximum likelihood approach with the EM algorithm [1] and the Gibbs sampling [9, 10], where we set up three evaluation criteria.

In section 2 we present the basic methodology for discovering probabilistic motifs with mixture models and the EM algorithm for estimating model parameters, and then the proposed marginal mixture model and its expressive motif models. Special care is also given to the initialization where we present a scheme based on the k-means clustering algorithm over the model parametric space. Section 3 presents experimental results obtained by our method and comparative approaches in various datasets. Finally, we summarize and give some concluded remarks in section 4.

2 Applying mixture models for motifs discovering

Consider a finite set $\Sigma = \{c_1, \ldots, c_M\}$ consisting of M individual characters. An arbitrary string over the set Σ is any sequence $S_j = \{s_{jk}\}_{k=1}^{Lj}$ of length L_j, where $s_{jk} \in \Sigma$ denotes the character at the k-th position of the j-th sequence. Now, let $S = \{S_1, \ldots, S_N\}$ be a set of N strings of length $L_1, \ldots, L_N$, respectively. A common subsequence of length K that is repeated at different sites among the input sequences of set S is called *motif*. In our study we will considered that the motif length K is known.

The application of mixture models needs a necessary preprocessing step to be made first, where we collect all the possible substrings over S of length equal to K. This can be done by sliding a window of size K to every sequence S_j, obtaining a set of $L_j - K + 1$ substrings. Each substring indicates the starting position

[1] Department of Computer Science, University of Ioannina, 45110 Ioannina, Greece, email: {evoudiga, kblekas}@cs.uoi.gr

of a possible motif copy over sequences. In such way, we obtain a set of n substrings $X = \{x_i\}_{i=1}^n$, $n = \sum_{j=1}^N (L_j - K + 1)$, that constitutes the set of observations in our problem.

2.1 The basic approach

In the basic approach, we fit a two-component mixture of multinomials model to the set of input substrings X. Here, we assume that each substring x_i has been generated from either a motif model ($y_i = 1$) or a background model ($y_i = 0$), as given by the *hidden* (binary) variable y_i. The motif model component has a prior probability $P(y_i = 1) = \pi$, while the second component, the background, captures all the non-motif information with a prior probability $P(y_i = 0) = 1 - \pi$. The density function of the mixture model for an observation x_i is given by

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

where $\Theta = \{\pi, \theta, b\}$ is the set of the (unknown) model parameters: the prior probability ($0 \leq \pi \leq 1$) and the two component density parameters.

A flexible and most commonly used way for describing a motif model is through a position weight matrix (PWM) $\theta = [\theta_{kl}]$ of size $K \times M$. Each element θ_{kl} denotes the probability of character $c_l \in \Sigma$ to be found in the k-th position of the motif, where for each position k it holds $\sum_l \theta_{kl} = 1$. On the other hand, the background distribution is represented with an M-vector of probabilities $b = [b_1, \ldots, b_M]$ that is common for any substring position ($\sum_l b_l = 1$). Following the multinomial distribution and assuming independence among motif positions, the probability density functions of the motif $p_m(x_i|\theta)$ and the background model $p_b(x_i|b)$ are

$$\begin{aligned} p_m(x_i|\theta) &= p(x_i|y_i = 1, \theta) = \prod_{k=1}^{K}\prod_{l=1}^{M} \theta_{kl}^{\delta_{ikl}}\,, & (2)\\ p_b(x_i|b) &= p(x_i|y_i = 0, b) = \prod_{l=1}^{M} b_l^{\sum_{k=1}^{K}\delta_{ikl}}\,, & (3)\end{aligned}$$

where δ_{ikl} is the Kronecker delta, i.e. 1 if character c_l is found at the k-th position of substring x_i, 0 otherwise.

Based on the above formulation, the motif discovery problem can now be converted into a maximum likelihood (ML) estimation problem, where the log-likelihood function is given by

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

The EM algorithm [6] constitutes an efficient framework for estimating the model parameters $\Theta = \{\ \pi, \{\theta_{kl}\}$ and $\{b_l\}\ \}$. It iteratively performs two main steps. At first the E-step where the conditional expectation $z_i = p(y_i = 1|x_i, \Theta^{(t)})$ of the hidden variables y_i is computed

$$\begin{aligned} z_i^{(t)} &= p(y_i = 1|x_i, \Theta^{(t)}) = \\ &= \frac{\pi^{(t)} p_m(x_i|\theta^{(t)})}{\pi^{(t)} p_m(x_i|\theta^{(t)}) + (1-\pi^{(t)})p_b(x_i|b^{(t)})}\,, \end{aligned} \tag{5}$$

and also the expectation of the complete data log-likelihood is established:

$$\begin{aligned} Q(\Theta, \Theta^{(t)}) &= \sum_{i=1}^{n} z_i^{(t)}\{\log\pi + \log p_m(x_i|\theta)\} + \\ &+ (1 - z_i^{(t)})\{\log(1-\pi) + \log p_b(x_i|b)\}\,. \end{aligned} \tag{6}$$

The above Q-function is maximized next during the M-step over the mixture model parameters. This gives the following update equations:

$$\begin{aligned} \pi^{(t+1)} &= \frac{\sum_{i=1}^{n} z_i^{(t)}}{n}\,, & (7)\\ \theta_{kl}^{(t+1)} &= \frac{\sum_{i=1}^{n} z_i^{(t)}\delta_{ikl}}{\sum_{i=1}^{n} z_i^{(t)} \sum_{l=1}^{M}\delta_{ikl}}\,, & (8)\\ b_l^{(t+1)} &= \frac{\sum_{i=1}^{n}(1 - z_i^{(t)})\sum_{k=1}^{K}\delta_{ikl}}{K\sum_{i=1}^{n}(1 - z_i^{(t)})}\,. & (9)\end{aligned}$$

The EM algorithm guarantees the convergence of the likelihood function to a local maximum where simultaneously satisfies all the constraints of the parameters. At the end of the process, the occurrences of the estimated motif can be found by selecting the substrings x_i whose posterior probability value z_i is above a threshold (e.g. 0.8).

2.2 The proposed approach

Nevertheless, a significant drawback from the above basic approach appears due to the convenient i.i.d. assumption of the observations x_i. This prevents into taking into account the valuable spatial information of data. As a result substrings that are found in the neighborhood of a motif occurrence may be considered as motif copies i.e. to have high posterior probability values z_i to be a motif. This phenomenon, which is mostly common in recurrent type of motifs where a short part of the motif is repeated, may be substantially modify the expressive motif multinomial model to a non-informative uniform model.

Suppose, for example, that the motif we want to discover is the $****TATATATATA****$, where the 2-gram TA is repeated 5 times ($K = 10$). Suppose also a copy of this motif $x_i = TATATATATA$. Then, as shown in Table 1, there are 5 substrings in its neighborhood (which overlap with the x_i) that will contain a significant part of the motif. This will make their posterior probability values (Eq. 5) to be significant large, and thus their contribution to the update rule of motif model parameters θ_{kl} of Eq. 8 will be high. This will cause into a step by step dramatic reduction of the higher probabilities values in every motif position (that correspond to the characters of the motif) during the EM procedure. Under this circumstance, the EM algorithm will end up to a uniform distribution of M characters in K motif positions and thus the motif discovery will be failed.

In the literature there are various methods that try to overcome this occurrence by indirectly or directly adopting spatial information to the model. In [1] for example, a normalization of the posterior value z_i of the adjacent sequences is performed so that guarantees in any window of length K the sum of z_i values remains less than or equal to 1. Another approach was presented in [2] by introducing spatial constraints to the model through the use of a Markov Random Field (MRF) prior over the motif labels. In this study we work more systematically over spatiality by incorporating this useful kind

Table 1. Substrings in the neighborhood of a motif occurrence x_i. Since they all have a significant overlap with the real motif and thus high posterior values z_i, they will contribute considerably to the EM updated rule of the motif parameters.

x_{i-4}	****TATATA
x_{i-2}	**TATATATA
x_i	TATATATATA
x_{i+2}	TATATATA**
x_{i+4}	TATATA****

of information from observations more naturally and directly to the expressive motif model.

2.2.1 *The extended model*

In the first scheme, the original position weight matrix θ is extended by $K-1$ more lines to both sides, so as to include the entire neighborhood of a motif. The motif model is now described with a new matrix of dimension $(3K-2)\times M$ and the fitting procedure can now be seen as searching over the new matrix model θ to find the best suited block matrix of K size (number of lines). Therefore, this matrix expansion suggests $2K-1$ (overlapping) multinomial distributions equal to the number of all possible starting positions within motif neighborhood (first $2K-1$ lines of matrix θ), with the following density function ($\forall\, j=1,\ldots,2K-1$)

$$\phi_j(x_i|\theta) = p(x_i|y_i=j,\theta) = \prod_{k=1}^{K}\prod_{l=1}^{M}\theta_{j+k-1,l}^{\delta_{ikl}} \,. \qquad (10)$$

The hidden variable y_i now defines the starting position of the substring x_i within the neighborhood of the motif. Note that when there is not any such position ($y_i = 2K$), the substring is generated by the simplest background model as previous. The mixture model density function now becomes

$$f(x_i|\Theta) = \sum_{j=1}^{2K-1}\pi_j\phi_j(x_i|\theta) + (1-\sum_{j=1}^{2K-1}\pi_j)p_b(x_i|b)\,. \qquad (11)$$

In the example of Table 1 the introduction of the new matrix θ allow the neighboring substrings ($x_{i-4}, x_{i-2}, x_{i+2}, x_{i+4}$) that are overlapping with the real motif copy (x_i) to be fitted with one of these multinomial distribution and not destroy the distribution of the original motif. We will refer to this scheme as the *extended* model.

Likewise, the EM algorithm can now be applied for estimating the model parameters. It requires the calculation of the posterior probabilities of hidden variables during the E-step

$$z_{ij}^{(t)} = P(y_i=j|x_i,\Theta^{(t)}) = \frac{\pi_j^{(t)}\phi_j(x_i|\theta^{(t)})}{f(x_i|\Theta^{(t)})}\,, \qquad (12)$$

and the maximization of the corresponding Q-function during the M-step

$$Q(\Theta,\Theta^{(t)}) = \sum_{i=1}^{n}\{\sum_{j=1}^{2K-1} z_{ij}^{(t)}\{\log\pi_j + \log\phi_j(x_i|\theta)\} + \\ +(1-\sum_{j=1}^{2K-1} z_{ij}^{(t)})\{\log(1-\sum_{j=1}^{2K-1}\pi_j) + \log p_b(x_i|b)\}\}\,. \qquad (13)$$

This gives the following update rules

$$\pi_j = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}\,, \qquad (14)$$

$$\theta_{kl}^{(t+1)} = \begin{cases} \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}^{(t)}\delta_{i,k-j+1,l}}{\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}^{(t)}} & \text{, if } 1\le k<K \\ \dfrac{\sum_{i=1}^{n}\sum_{j=k-K+1}^{k} z_{ij}^{(t)}\delta_{i,k-j+1,l}}{\sum_{i=1}^{n}\sum_{j=k-K+1}^{k} z_{ij}^{(t)}} & \text{, if } K\le k<2K \\ \dfrac{\sum_{i=1}^{n}\sum_{j=k-K+1}^{2K-1} z_{ij}^{(t)}\delta_{i,k-j+1,l}}{\sum_{i=1}^{n}\sum_{j=k-K+1}^{2K-1} z_{ij}^{(t)}} & \text{, if } 2K\le k<3K-1 \end{cases} , \qquad (15)$$

$$b_l^{(t+1)} = \frac{\sum_{i=1}^{n}(1-\sum_{j=1}^{2K-1} z_{ij}^{(t)})\sum_{k=1}^{K}\delta_{ikl}}{K\sum_{i=1}^{n}(1-\sum_{j=1}^{2K-1} z_{ij}^{(t)})}\,. \qquad (16)$$

As it is clear in the update equation for the motif extended matrix values θ_{kl} (Eq. 15), every line contributes more than one time (except for the borders) to the computation of the posterior probabilities z_{ij}), since they belong to the $2K-1$ multinomials several times. Moreover, it must be noted that in practice it does not always happen the real motif to be found in the central block matrix (between the K and $2K-1$ lines) of the extended matrix θ. After convergence of the EM algorithm, the final motif will be found by searching for the K continuous lines of the matrix that have the greater sum of their maximum probability values θ_{kl}.

Finally, there are two other important advantages of the above scheme that must be discussed. At first it is less sensitivity to the initialization of the motif parameters. Even if we have not perfectly initialized the motif and a part of it is missing, the matrix extension will finally manage to fix it and incorporate the missing part to its neighboring lines. Also, the extended model is not very sensitive to the length of motif K, since the new expressive matrix has a sufficient space for motifs of overestimated or underestimated length.

2.2.2 *The marginal model*

In the second approach, the position weight matrix θ remains the same as the basic approach (size $K\times M$), but now it has a different behavior. Here, we consider that in this model only a part of it is used for fitting input substrings x_i, while the remaining part is treated as background. This can be seen as applying a shifting operator to the examined substring x_i across the matrix θ so as to find the optimal alignment between them in terms of fitting. There are two cases. Either the first j positions of x_i are modeled by the background and the rest $K-j$ positions are modeled from the first j lines of the matrix θ, or the opposite, i.e. the last j lines of θ are responsible for modeling the first j motif positions while the rest are considered as background. In the negative case, the substring is considered entirely as background information (Eq. 3). Therefore, we have again $2K-1$

multinomial distributions, equal to the number of all possible combinations of fitting the observation with the marginal model. Every density function is now given by

$$\phi_j(x_i|\theta) = \begin{cases} \prod_{l=1}^{M} b_l^{\sum_{k=1}^{K-j} \delta_{ikl}} \prod_{k=1}^{j}\prod_{l=1}^{M} \theta_{kl}^{\delta_{i,k+K-j,l}} & ,1 \leq j \leq K \\ \prod_{k=1}^{2K-j}\prod_{l=1}^{M} \theta_{j-K+k,l}^{\delta_{ikl}} \prod_{l=1}^{M} b_l^{\sum_{k=2K-j+1}^{K} \delta_{ikl}} & ,K+1 \leq j \leq 2K-1 \end{cases} \tag{17}$$

In this case the hidden variable y_i defines the part (number of last or first positions) of any observation x_i that belongs to the motif model. This scheme will be referred next as *marginal* model.

The application of the EM algorithm for estimating the parameters of the marginal model leads to different update equations at the M-step, in comparison with the extended model. These are:

$$\theta_{kl}^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{j=k}^{K+k-1} z_{ij}^{(t)} \delta_{i,K-j+k,l}}{\sum_{i=1}^{n}\sum_{j=k}^{K+k-1} z_{ij}^{(t)}} \tag{18}$$

$$b_l^{(t+1)} = \frac{\sum_{i=1}^{n}\{\sum_{j=1}^{K} z_{ij}^{(t)} \sum_{k=1}^{K-j}\delta_{ikl} + \sum_{j=K+1}^{2K} z_{ij}^{(t)} \sum_{k=2K-j+1}^{K} \delta_{ikl}\}}{\sum_{i=1}^{n}\{\sum_{j=1}^{K}(K-j)z_{ij}^{(t)} + \sum_{j=K+1}^{2K}(j-K)z_{ij}^{(t)}\}} \tag{19}$$

At the end of the EM algorithm, the motif can be found by searching for the first (k_1) and the last (k_2) position (line of the matrix θ) among the totally K positions, where their maximum probability value are above a threshold (e.g. 0.8). This segment $[k_1 - k_2]$ corresponds only to a part of the true motif. Its rest part can be substantially obtained by first finding the substrings x_i with high posterior probability values ($z_{i\,k_2} > 0.8$) that capture the segment $[k_1 - k_2]$, and then by the sufficient statistics of their neighboring substrings in either direction.

2.3 Initialization of motif models

The major drawback of the EM algorithm is its great dependence on the initial values of the model parameters that may cause into getting stuck in local maxima of the likelihood function [12]. In our study, this problem is mainly concentrated on the initialization of the motif model parameters θ, since the background density b is always initialized by the relative frequencies of characters in sequences. Various approaches have been proposed in the literature to overcome this problem. In [1] for example, a dynamic programming approach is used which estimates the goodness of many possible starting points based on the likelihood of the model after one EM iteration. Another method presented in [3] applies a divisive hierarchical clustering approach that generates a motif models parametric search space. In this study we have applied an initialization strategy that is based on a clustering scheme using the classical k-means algorithm.

In the general case the k-means algorithm aims at finding a partition of M disjoint clusters to a set of n observations, so as to minimize the overall sum of their distances with the cluster centers μ_j. Depending on the type of observations, one must determine an appropriate distance function and also a method for calculating cluster centers. In our study, we initially map every input substring x_i into a position weight matrix ϑ_i, where its values were taken from the Kronecker delta values $\vartheta_{ikl} = \delta_{ikl}$. In this model parametric search space we perform then a clustering procedure by using the Manhattan distance function between two subsequences x_i, x_j:

$$d(i,j) = \sum_{k=1}^{K}\sum_{l=1}^{M} |\vartheta_{ikl} - \mu_{jkl}| = \|\vartheta_i - \mu_j\|_1 \,.$$

Finally, the cluster centers $\mu_j = \{\mu_{jkl}\}$ are estimated by the sufficient statistics of the substrings that currently belong to jth cluster (relative frequencies of characters c_l in every position k). The number of clusters for searching was set to $k = n/N$. When finishing, we initialize the motif model with the center μ_{j^*} from the cluster j^* that has the minimum *intracluster* distance, i.e. average distance between all cluster members and its center. The experimental study has shown that this clustering scheme provides satisfactory initial values of model parameters θ very fast.

3 Experimental results

Several experiments were performed using both artificial and real datasets in an attempt to study the effectiveness of the proposed approach. During all experiments the clustering scheme of k-means was first executed twenty (20) times and the optimum solution was kept to initialize the motif model. The prior probabilities π_j were all initially set to $\pi_i = 1/(2K)$. We have tested both versions of the proposed marginal mixture model, the extend (*Ext*) and the marginal (*Marg*). Comparative results have been also obtained using the basic ML approach (*basic*), as well as the Gibbs sampling (*GS*) [9, 10]. The GS method iteratively performs two steps until likelihood convergence: It randomly selects first a sequence S_i and re-estimates the motif model $\theta^{(t+1)}$ using the current motif positions of all sequences but S_i. Then a new starting position of the motif model is selected in S_i (among the $L_i - K + 1$ possible) by sampling from the posterior distribution. Obviously, this version of the GS assumes that each sequence has a unique occurrence of the pattern, while our approach suggests an arbitrary number of motif copies in each sequence. Although this is not fair, in our study we have selected datasets having a single motif copy in every sequence. Finally, it must be mentioned that all methods were initialized with the same way, except from the GS method which is better to randomly initialized. The last was executed 20 successive times with different seed value and the best solution found was kept.

The generation mechanism of the artificial sets used in our experiments was the following: Using a discrete alphabet Σ with M characters, we uniformly produce a number $N = 20$ sequences of variable length (with a mean length $\overline{L} = 100$). Then, in each sequence we randomly select a position for placing a noisy copy of a preselected seed motif of length K, according to a mutation probability value p_m (common to every motif position). Eight different values for the noise parameter p_m were used, and for each value we generated 50 different sets of artificial sequences and kept statistics (mean values and stds) of the performance of all the comparative methods.

We have used the following three performance criteria for evaluating each method:

- $\Delta\theta$: Manhattan distance between the estimated and the true motif model (distance distributions)

$$\Delta\theta = ||\theta - \theta^{true}||_1 = \sum_{k=1}^{K}\sum_{l=1}^{M} |\theta_{kl} - \theta_{kl}^{true}| \, ,$$

where true pattern density θ^{true} was estimated from the relative frequencies of characters of the noisy motif copies.

- Sn: percentage of the predicted positions of all true motif copies in the set of input sequences X.
- Sp: percentage of the real copies of a motif found having detected at least 33% of their positions.

In must be noted that for the calculation of Sn and Sp quantities, when a method found overlapping substrings in a motif copy neighborhood, only their mean value of the predicted number of positions would have been kept for that copy.

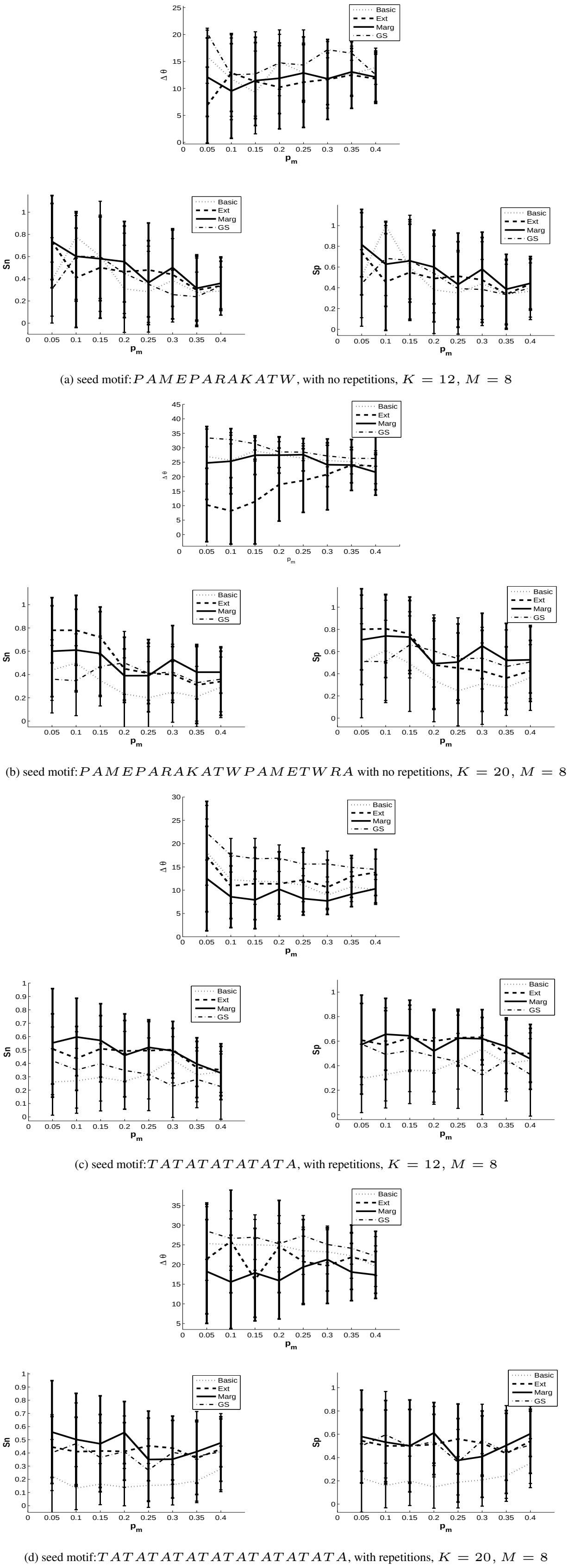
We have made experiments in a variety of simulated sets of sequences (various seed motifs, length K, alphabet size M and noise parameter p_m). Figure 1 illustrates the depicted results we obtained with four seed motifs and an alphabet of size $M = 8$. Each diagram represents the mean value and the standard deviation of the evaluation measurements ($\Delta\theta, Sn, Sp$) as calculated by each one of the four comparative approaches using different noisy levels (p_m). Both methods, extend and marginal, showed a better performance in comparison with the other two approaches. In almost all cases both schemes of our method had managed to estimate better the true model according to the $\Delta\theta$ criterion. And this is of great interest. Also, as it was expected, in the case of recurrent motifs (Fig. 1 (c), (d)) the basic approach completely fails to discover them. On the other hand, our approach yielded satisfactory results and comparable to these of the GS method. The results were the same with several other experiments with simulated data created from alphabet of different size.

We have also evaluated our method in a real dataset generated by the motif information of *Escherichia coli* RegulonDB [7]. This database consists of various sets of DNA sequences [2] having a single motif copy per sequence with symmetric margins (*background*) on both site of it. The study of these benchmarks is very interesting and attractive, since the ground truth (locations of motif occurrences) is known a priori and also the degree of similarity of motif occurrences is very low (extremely noisy motifs). More details about these data can be found in [7]. In our experiments we have used a part of eight (8) such data sets presented in Table 2, where we have considered two margin sizes: 20 and 50.

Table 2. The description of the real data used in our experimental study.

Filename	Motif length (K)	Number of sequences (N)
DnaA	9	7
FruR	14	10
Fur	19	20
LexA	20	10
CytR	40	10
PurR	76	15
SoxS	78	17
TyrR	82	17

Figure 2 summarizes the results obtained by the four comparing methods. In particular, it illustrates the mean values of both measurements Sn and Sp, as calculated after 20 different runs in any of the



(a) seed motif:$PAMEPARAKATW$, with no repetitions, $K = 12$, $M = 8$

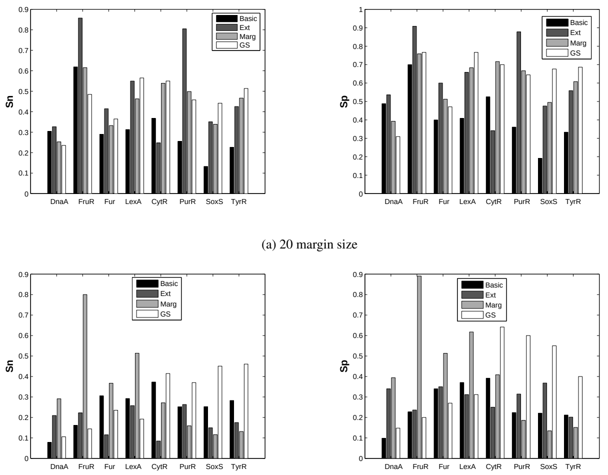
(b) seed motif:$PAMEPARAKATWPAMETWRA$ with no repetitions, $K = 20$, $M = 8$

(c) seed motif:$TATATATATATA$, with repetitions, $K = 12$, $M = 8$

(d) seed motif:$TATATATATATATATATATA$, with repetitions, $K = 20$, $M = 8$

Figure 1. Comparative results with simulated data. For each problem we present three diagrams with the calculated mean values and stds of the three performance measurements.

[2] Data can be downloaded from http://dragon.bio.purdue.edu/pmotif/



(a) 20 margin size

(b) 50 margin size

Figure 2. Comparative results on the real datasets of Table 2 in the case of 20 (a) and 50 (b) margin size. The vertical bars represent the mean values of the two calculated measurements Sn, Sp after 50 executions of each method.

selected data sets. Although these kind of motifs are not recurrent, on average, we obtained better results with our approach in comparison with the basic ML method. This reinforce the proposed marginal mixture model since it manages to detect qualitatively better motifs in extremely noise parameter. Finally, although the comparison with the version of the GS method in unfair, our method yielded comparable results especially in motifs of small length K. An explanation of this is that when searching for motifs of large length, for example $K = 76$, the selected margin size of 20 or 50 is not enough for capturing the neighborhood of motifs within the expressive model and thus the proposed approaches cannot be normally applied.

4 Conclusions

We have presented a marginal mixture model for discovering probabilistic motifs in categorical sequences that incorporates an advanced and more informative expressive motif model. Two similar versions of that model were proposed, one with an expansion and another one with a delimitation of bounds within position weight matrix. This allows to fit the neighborhood of a motif that leads to an efficient and consistent inference of motif locations. Another significant advantage of our method is that the EM algorithm that is used to estimate the model parameters, is less depended on the model parameters initialization than with the classical approach. Experiments on a variety of artificial and real data sets have shown improved performance of the proposed scheme and its ability to identify qualitatively better motifs, in comparison with the basic ML approach and the GS method. We are planning to further investigate the performance of the proposed method to other experimental data sets and also to design more complex expressive motif models that can simultaneously handle gaps among motif positions.

REFERENCES

[1] T.L. Bailey and C. Elkan, 'Unsupervised learning of multiple motifs in Biopolymers using Expectation Maximization', *Machine Learning*, **21**, 51–83, (1995).

[2] K. Blekas, 'A mixture model based Markov random fields for discovering probabilistic patterns in sequences', in *Panhellenic Conference in Artificial Intelligence (SETN-2006) (Lecture Notes in Artificial Intelligence)*, volume 3955, pp. 25–34, (2006).

[3] K. Blekas, D.I. Fotiadis, and A. Likas, 'Greedy mixture learning for multiple motif discovering in biological sequences', *Bioinformatics*, **19**(5), 607–617, (2003).

[4] A. Brāzma, I. Jonasses, I. Eidhammer, and D. Gilbert, 'Approaches to the automatic discovery of patterns in biosequences', *Journal of Computational Biology*, **5**(2), 277–303, (1998).

[5] 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.

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

[7] J. Hu, B. Li, and D. Kihara, 'Limitations and potentials of current motif discovery algorithms', *Nucleic Acids Research*, **33**(15), 4899–4913, (2005).

[8] R. Hughey and A. Krogh, 'Hidden Markov models for sequence analysis: Extension and analysis of the basic method', *CABIOS*, **12**(2), 95–107, (1996).

[9] 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*, **226**, 208–214, (1993).

[10] 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*, **90**, 1156–1169, (1995).

[11] 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).

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

[13] 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*, **2**(1), 127–154, (2004).