© 2004 Schattauer GmbH

A Sequential Method for Discovering Probabilistic Motifs in Proteins

K. Blekas, D. I. Fotiadis, A. Likas
Department of Computer Science, University of Ioannina,
and Biomedical Research Institute, Foundation for Research and Technology – Hellas, Ioannina,
Greece

Summary

Objectives: This paper proposes a greedy algorithm for learning a mixture of motifs model through likelihood maximization, in order to discover common substrings, known as *motifs*, from a given collection of related biosequences.

Methods: The approach sequentially adds a new motif component to a mixture model by performing a combined scheme of global and local search for appropriately initializing the component parameters. A hierarchical clustering scheme is also applied initially which leads to the identification of candidate motif models and speeds up the global searching procedure.

Results: The performance of the proposed algorithm has been studied in both artificial and real biological datasets. In comparison with the well-known MEME approach, the algorithm is advantageous since it identifies motifs with significant conservation and produces larger protein fingerprints.

Conclusion: The proposed greedy algorithm constitutes a promising approach for discovering multiple probabilistic motifs in biological sequences. By using an effective incremental mixture modeling strategy, our technique manages to successfully overcome the limitation of the MEME scheme which erases motif occurrences each time a new motif is discovered.

Keywords

Motif discovery, mixture of motifs, EM algorithm, protein fingerprints, MEME algorithm

Methods Inf Med 2004; 43: 9–12

1. Introduction

The motif identification is one of the most important problems in protein sequence analysis covering many application areas. It concerns the discovery of portions of protein strands of major biological interest with important structural and functional features. Motifs can also be used for characterizing biological families and searching for new family members. This leads to the development of diagnostic signatures (*fingerprints*) that contain groups of conserved motifs used to characterize a family. The PRINTS (or PRINT-S) database [1] is an example of a protein fingerprints database containing ungapped motifs.

Usually, patterns or motifs can be either *deterministic or probabilistic* [2]. A simplified way of modeling a probabilistic ungapped motif is the *position weight matrix* (PWM) representing the relative frequency of each character at each motif position. The Gibbs sampling [3] and MEME [4] represent probabilistic methods for finding multiple shared motifs within a set of unaligned biological sequences. The MEME algorithm fits a two-component finite mixture model to a set of sequences using the *Expectation Maximization* (EM) algorithm [5], where one component describes the motif and the other describes the background (other positions in the sequences). Multiple motifs are discovered by sequentially applying a new mixture model with two components to the sequences remaining after erasing the occurrences of the already identified motifs.

In this paper we present an innovative approach for discovering significant motifs in a set of sequences based on recently developed incremental schemes for Gaussian mixture learning [6]. Our method learns a mixture of motifs model in a greedy fashion by incrementally adding components (motifs) to the mixture. Starting with one component that models the background, at each step a new component is added which corresponds to a candidate motif. The algorithm tries to identify a good initialization for the parameters of the new motif by performing *global* search over the input substrings together with *local* search for fine tuning of the parameters of the new component. In addition, a hierarchical clustering procedure is proposed based on *k*d-tree techniques [7] for partitioning the input dataset of substrings, which can reduce the time complexity for global searching.

2. Greedy EM Algorithm for Motifs Discovery

2.1 The Mixture of Motifs Model

Consider a finite set of characters $\Sigma = \{\alpha_1, \Lambda, \alpha_\Omega\}$ where $\Omega = |\Sigma|$. Any sequence $S = a_1\, a_2\, \Lambda\, a_L$ of length L, such that $L \geq 1$ and $a_i \in \Sigma$, is called a string (or *sequence*) over the character set Σ. The consecutive characters $a_i\, \Lambda\, a_{i+w-1}$ form a *substring* x_i of length W, identified by the starting position i over the string S. There are $n = L - W + 1$ such possible substrings of length W generated from sequence S. We assume a set of N unaligned sequences $\mathrm{S} = \{S_1, \Lambda, S_N\}$ of length L_1, Λ, L_N, respectively. In order to deal with the problem of motif discovery of length W we construct a new dataset containing all substrings of length W in S. Therefore, we obtain a training dataset $X = \{x_1,\Lambda ,x_n\}$ of n substrings ($n = \sum_{s=1}^{N} \{L_s - W + 1\}$) for the learning problem.

A mixture of motifs model f for an arbitrary substring x_i assuming g components can be written as:

$$f(x_i; \Psi_g) = \sum_{j=1}^{M} \pi_j \varphi_j (x_i; \Theta_j), \tag{1}$$

where Ψ_g is the vector of all unknown parameters in the mixture model of g components, i.e. $\Psi_g = [\pi_1,...,\pi_{g-1}, \Theta_1,...,\Theta_g]$. The mixing proportion π_j ($\pi_j \geq 0$) can be viewed as the prior probability that data x_i has been generated by the j-th component of the mixture and they satisfy $\sum_{j=1}^{g} \pi_j = 1$.

Each one of the g components corresponds to either a motif or the background. Following the position weight matrix representation, a motif j can be modeled by $\text{PWM}_j = [p_{l,k}^{j}]$ of size $[\Omega \times W]$, where each value $p_{l,k}^{j}$ denotes the probability that the letter α_1 is located in motif position k. On the other hand, a background component j is represented using a probability vector BPM_j (of length Ω), where each parameter value ρ_l^j denotes the probability of letter α_1 to occur at an arbitrary position. The probability that a substring $x_i = \alpha_{i1} \Lambda\, \alpha_{iW}$ has been generated by the component j is

$$\varphi_j(x_i;\theta_j) = \begin{cases} \prod_{k=1}^{W} p_{a_{ik},k}^{j} & \text{if } j \text{ is motif} \\ \prod_{k=1}^{W} \rho_{a_{ik}}^{j} & \text{if } j \text{ is background.} \end{cases} \tag{2}$$

The log-likelihood of the observed dataset X corresponding to the above model is

$$L(\Psi_g) = \sum_{i=1}^{n} \log f(x_i; \Psi_g). \tag{3}$$

Formulating the problem as an incomplete-data problem (5), each substring x_i can be considered as having arisen from one of the g components of the mixture model of Equation 1. Therefore, we can introduce the parameters $z_{ij} = 1$ or 0 that indicate whether x_i has been generated by the j-th component of the mixture. The EM algorithm can be applied for the log-likelihood maximization problem by treating the z_{ij} as missing data. The following update equations are obtained for each component j (4, 5)

$$z_{ij}^{(t+1)} = \Pr(z_{ij} = 1 \mid x_i, \Psi_g^{(t)}) = \frac{\pi_j^{(t)} \varphi_j(x_i;\theta_j^{(t)})}{f(x_i;\Psi_g^{(t)})}, \tag{4}$$

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(t+1)}, \tag{5}$$

$$\theta_j^{(t+1)} = \begin{cases} p_{l,k}^{j} = \dfrac{c_{l,k}^{j}}{\sum_{l=1}^{\Omega} c_{l,k}^{j}} & \text{if } j \text{ is motif} \\ \rho_l^j = \dfrac{c_l^j}{\sum_{l=1}^{\Omega} c_l^j} & \text{if } j \text{ is background} \end{cases} \tag{6}$$

where the elements $c_{l,k}^{j}$ (c_l^j) correspond to the observed frequency of letter α_1 at position k of motif j occurrences (at background j arbitrary positions) and can be formally expressed as

$$\begin{aligned} c_{l,k}^{j} &= \sum_{i=1}^{n} z_{ij}^{(t+1)} \mathbf{I}(a_{ik}, l) && \text{if } j \text{ is motif} \\ c_l^j &= \sum_{i=1}^{n} z_{ij}^{(t+1)} \sum_{k=1}^{W} \mathbf{I}(a_{ik}, l) && \text{if } j \text{ is background} \end{aligned} \tag{6a}$$

The indicator I (α_{ik}, l) denotes a binary function which takes value 1 if the substring x_i contains letter α_l at position k ($\alpha_{ik} \equiv \alpha_l$) and *0* otherwise.

Equations 4–6 can be used to estimate the parameter values Ψ_g of the g-component mixture model which maximize the log-likelihood function and ensure the convergence of the algorithm to a local maximum of the likelihood function (5). However, its great dependence on parameter initialization and its local nature (it gets stuck in local maxima of the likelihood function) do not allow us to directly apply the EM algorithm to a g component mixture of motifs model. To overcome the problem of *poor initialization*, we next propose an efficient combined scheme of global searching over appropriate defined candidate motifs, followed by a local searching for fine tuning the parameters of a new motif.

2.2 Greedy Mixture Learning

Assume that a new component φ_{g+1} (x_i; Θ_{g+1}) that corresponds to a motif is added to a g-component mixture model f (x_i; Ψ_g). Then the resulting mixture has the following form

$$f(x_i;\Psi_{g+1}) = (1-a) f(x_i;\Psi_g) + a \varphi_{g+1}(x_i;\theta_{g+1}), \tag{7}$$

with $a \in (0,1)$. The vector Ψ_{g+1} specifies the new parameter vector and consists of the parameter vector Ψ_g of the g-component mixture, the weight a and the parameter vector f (x_i; Ψ_{g+1}). This formulation proposes a two-component likelihood maximization problem, where the first component is described by the old mixture f (x_i; Ψ_g) and the second one is the motif component φ_{g+1} (x_i; Θ_{g+1}). If we consider that the parameters Ψ_g of f (x_i; Ψ_g) remain fixed during maximization of $L(\Psi_{g+1})$, the problem can be treated by applying searching techniques to optimally specify the parameters a and Θ_{g+1} which maximize $L(\Psi_{g+1})$.

As presented in (6), an EM algorithm can be applied where the learning procedure is applied only to the mixing weight a and the probabilistic quantities $p_{l,k}^{g+1}$ of the newly inserted component. Following Equations 5-7, the next update procedures can be derived

$$z_{i,g+1}^{(t+1)} = \frac{a^{(t)} \varphi_{g+1}(x_i;\theta_{g+1}^{(t)})}{(1-a^{(t)}) f(x_i;\Psi_{g+1}) + a^{(t)} \varphi_{g+1}(x_i;\theta_{g+1}^{(t)})}, \tag{8}$$

$$a^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{i,g+1}^{(t+1)}, \tag{9}$$

$$\theta_{g+1}^{(t+1)} = \lfloor p_{l,k}^{g+1} \rfloor, \text{ where } p_{l,k}^{g+1} = \frac{c_{l,k}^{g+1}}{\sum_{l=1}^{\Omega} c_{l,k}^{g+1}}, \tag{10}$$

where $c_{l,k}^{g+1} = \sum_{i=1}^{n} z_{i,g+1}^{(t+1)} \mathbf{I}(a_{ik}, l)$.

The above *partial EM* steps constitute a simple and fast method for locally searching the maxima of $L(\Psi_{g+1})$. However, the problem of poor initialization still remains since this scheme is very sensitive to the proper initialization of the two parameters a and Θ_{g+1}. For this reason a global search strategy has been developed (6) which substitutes the log-likelihood function using a Taylor approximation about a point $a = a_0$, and then using the resulting estimate to search for the optimal Θ_{g+1} value. Therefore we expand $L(\Psi_{g+1})$ by second order Taylor expansion about point $a_0 = 0{,}5$ and then the resulting quadratic function is maximized with respect to a. It can be shown (6) that, for a given parameter vector Θ_τ, a local maximum of $L(\Psi_{g+1})$ near $a_0 = 0.5$ is given by

$$\hat{L}(\theta_\tau) = \sum_{i=1}^{n} \log \frac{f(x_i;\Psi_g) + \varphi_{g+1}(x_i;\theta_\tau)}{2} + \frac{1}{2} \frac{\left[\sum_{i=1}^{n} \delta(x_i,\theta_\tau)\right]^2}{\sum_{i=1}^{n} \delta^2(x_i,\theta_\tau)} \quad (11)$$

and is obtained for

$$\hat{a} = \frac{1}{2} - \frac{1}{2} \frac{[\sum_{i=1}^{n} \delta(x_i,\theta_\tau)]^2}{\sum_{i=1}^{n} \delta^2(x_i,\theta_\tau)} , \quad (12)$$

where

$$\delta(x_i,\theta_\tau) = \frac{f(x_i;\Psi_g) - \varphi_{g+1}(x_i;\theta_\tau)}{f(x_i;\Psi_g) + \varphi_{g+1}(x_i;\theta_\tau)} \cdot \quad (13)$$

The above methodology has the benefit of modifying the problem of maximizing the likelihood function to become independent on the selection of initial value for the mixing weight *a*. The only problem is now the identification of *candidate* values Θ_τ so as to properly initialize the motif parameters and to conduct partial EM steps.

A reasonable approach is to search for candidates directly over the total dataset of substrings $X = \{x_\tau\}, (\tau = 1,\Lambda ,n)$. For this reason we associate with each substring $x_\tau = a_{\tau 1}\Lambda ,a_{\tau W}$ a position weight matrix Θ_τ constructed as follows:

$$\Theta_\tau = \left\lfloor p_{l,k}^{\tau} \right\rfloor, \text{ where } p_{l,k}^{\tau} = \begin{cases} \lambda & \text{if } a_{\tau k} \equiv a_l \\ \dfrac{1-\lambda}{\Omega - 1} & \text{otherwise} \end{cases}, \quad (14)$$

where the parameter λ has a fixed value in the range (0,1). Therefore, the log-likelihood $\hat{L}(\Theta_\tau)$ is determined by selecting among the Θ_τ matrices the one which maximizes the right hand size of Equation 11.

The drawback of the above approach is the increasing time complexity $(O(n^2))$ of the search procedure. In order to reduce the complexity, we perform a *hierarchical clustering* technique based on the notion of *k*d-trees [7], by proposing a modified approach in order to deal with sequential symbolic data. In particular, using an appropriate criterion based on maximum character variance, we apply a partitioning scheme that divides the original set *X* into a set of *C* << *n* clusters. The position weight matrices (Equation 14) corresponding to the *centroids* of the clusters (*consensus substrings*) constitute the candidate matrices $\Theta_\tau (\tau = 1,\Lambda ,C)$ used in global search (Equations 11-13).

Special treatment has also been given to avoid overlappings with the already discovered motifs during the selection of a candidate motif instance. This is achieved by excluding, from the set *C* of consensus substrings (candidate motifs), those substrings that overlap with motif occurrences.

Table 1 The seed substrings in the artificial datasets

6 seed motifs

positions

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	V	E	R	Y	I	N	T	E	R	E	S	T	I	N	G	V	I	E	W
D	E	S	T	I	N	A	T	E	D	Y	D	E	T	I	C	A	T	E	D
P	R	I	M	E	R	S	T	A	G	E	D	E	S	I	G	N	I	N	G
W	A	I	T	I	N	G	G	T	H	E	G	I	A	N	T	T	R	A	M
S	T	E	W	A	R	T	E	L	E	N	A	M	I	L	E	N	A	E	M
A	R	I	S	T	I	D	I	S	K	N	A	M	I	L	E	N	A	E	M

3. Experimental Results

In order to evaluate the effectiveness of our method we have conducted a series of experiments considering both artificial and real sets of biological sequences. In all cases the width *W* of the motifs is considered constant, while good values for the λ parameter used to initialize the candidate position weight matrices (Equation 14), were found to be in the range [0.6, 0.8]. For all the experimental datasets we have also applied the MEME approach using the available software from the corresponding Web site*.

3.1 Experiments with Artificial Datasets

In the artificial datasets used in our experiments each motif has an associated randomly generated "seed substring" and copies of the motif are created by randomly performing a number of substitutions (*mutations*) with a mutation probability p_m. We created artificial sequences of variable length (between 310 and 330) by randomly locating (ensuring no overlapping) and mutating copies of six (6) different seed substrings of length *W* = 20 (Table 1). The rest positions were filled with characters from the amino acids (AA) alphabet (Ω = 20). As illustrated in Table 1 the last two seed substrings (5 and 6) are exactly the same in half of their length (from position 11 to 20). Assuming three different values of the mutation probability (p_m = {0,0.1,0,2}), three different datasets of twenty (*N* = 20) artificial protein sequences were constructed.

The comparative results with the MEME algorithm have shown the superiority of the greedy EM algorithm in discovering all the incorporated motifs in the three datasets. The MEME approach was unable to identify the motifs 5 and 6 and considered them as one motif.

3.2 Experiments with Real Datasets

The real datasets used in our experiments were obtained from the PRINTS database [1] which contains protein motif fingerprints. Three families from the PRINTS database were selected, describing fingerprints for L5 ribosomal proteins (PR00058), secretion pathway protein C (PR00810) and pi-class glutathione S-transferases (PR001268). Each motif discovered was evaluated in terms of the information content (IC) (4), specified as follows

$$IC_j = \sum_{k=1}^{W} \sum_{\alpha_l \in \Sigma} p_{lk}^{j} \log_2 \frac{p_{lk}^{j}}{\rho_l^1} , \quad (15)$$

where ρ_l^1 indicates the overall background probability of letter a_1 in the dataset. This score becomes maximal if the motif is well conserved.

Table 2 summarizes the comparative results obtained using the three protein families. The superiority of the greedy EM algorithm over MEME is obvious not only in terms of the greater number of real-motifs discovered but also in terms of the degree of motif conservations as indicated by the *IC* scores. In all cases, the number of the discovered motifs is also greater than the

* The Web site of MEME/MAST system version 3.0 can be found at http://meme.sdsc.edu/meme/website/

Table 2
Comparative results

Problem *Acc. Number*	*Greedy EM* Motif	 IC	*MEME* Motif	 IC
PR00058 (W=20) 16 seqs of 297 AA 6 motifs	I	72.4434	I'	63.2093
	II	72.4559	II'	56.9837
	III	72.1314	III'	49.4185
	IV	67.3621	IV'	46.0057
	V	69.4684		
	VI	69.7300		
	VII	65.3427		
	VIII	57.0378		
PR00810 (W=10) 6 seqs of 286 AA 2 motifs	I	37.2183	I'	34.3275
	II	35.2812	II'	33.9981
	III	34.7850	III'	31.8698
	IV	33.0999	IV'	30.0946
	V	33.2078		
	VI	29.8507		
PR01268 (W=17) 19 seqs of 209 AA 3 motifs	I	59.7625	I'	57.5664
	II	57.0125	II'	56.2638
	III	58.7029	III'	53.1302
	IV	57.5660		
	V	56.2634		
	VI	54.2335		
	VII	51.4565		
	VIII	55.0629		
	IX	53.1297		

number of motifs specified in the PRINTS database (Table 2). This means that the proposed method has led to the discovery of larger fingerprints (containing more motifs) and thus constitutes a promising tool for biological sequence analysis.

4. Conclusions

In this paper we have proposed a greedy EM algorithm for solving the multiple motif discovery problem in biological sequences. Our approach learns a mixture of motifs model in a greedy fashion by iteratively adding new components, through a combined scheme of local and global search which ensures fine tuning of the parameter vector of the new component.

The main difference with the MEME technique is the way that the mixture models are applied. Although both methods treat the same problem through mixture learning using the EM algorithm, our approach is able to effectively fit multiple-component mixture models, overcoming the problem of poor initialization of EM that frequently gets stuck on local maxima of the likelihood function. This results in exploring the input dataset efficiently and the discovery of greater number of motifs. The MEME scheme of erasing motif occurrences, pruning in such way the input dataset, does not allow the parameters of the discovered motifs to be re-estimated, and thus future discovered motifs cannot contribute to possible re-allocation of the character distribution in the motif positions. As the results indicate, this drawback becomes significant in cases where motifs exist that partially match, since these motifs are recognized by the MEME algorithm as one "composite" motif that cannot be further analyzed.

References

1. Attwood TK, Croning MDR, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley J, Wright W. PRINT-S: the database formerly known as PRINTS. Nucleic Acids Research 2000; 28 (1): 225-7.
2. Rigoutsos I, Floratos A, Parida L, Gao Y, Platt D. The Emergency of Pattern Discovery Techniques in Computational Biology. Metabolic Engineering 2000; (2):159-77.
3. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwland AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 1993; 226: 208-14.
4. Bailey TL, Elkan C. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. Machine Learning 1995; 21: 51-83.
5. McLachlan GM, Peel P. Finite Mixture Models. New York: John Wiley & Sons, Inc.; 2001.
6. Vlassis N, Likas A. A greedy EM algorithm for Gaussian mixture learning. Neural Processing Letters 2002, 15 (1): 77-87.
7. Bentley JL. Multidimensional binary search trees used for associative searching. Commun ACM 1975; 18 (9): 509-17.

Correspondence to:
Konstantinos Blekas
Department of Computer Science
University of Ioannina
P.O. Box 1186
GR-45110 Ioannina
Greece
E-mail: kblekas@cs.uoi.gr