The value of prior knowledge in discovering motifs with MEME

Timothy L. Bailey* and **Charles Elkan**
Department of Computer Science and Engineering
University of California at San Diego
La Jolla, California 92093-0114
tbailey@cs.ucsd.edu and elkan@cs.ucsd.edu

Abstract

MEME is a tool for discovering motifs in sets of protein or DNA sequences. This paper describes several extensions to MEME which increase its ability to find motifs in a totally unsupervised fashion, but which also allow it to benefit when prior knowledge is available. When no background knowledge is asserted, MEME obtains increased robustness from a method for determining motif widths automatically, and from probabilistic models that allow motifs to be absent in some input sequences. On the other hand, MEME can exploit prior knowledge about a motif being present in all input sequences, about the length of a motif and whether it is a palindrome, and (using Dirichlet mixtures) about expected patterns in individual motif positions. Extensive experiments are reported which support the claim that MEME benefits from, but does not require, background knowledge. The experiments use seven previously studied DNA and protein sequence families and 75 of the protein families documented in the Prosite database of sites and patterns, Release 11.1.

Introduction

MEME is an unsupervised learning algorithm for discovering motifs in sets of protein or DNA sequences. This paper describes the third version of MEME. Earlier versions were described previously (Bailey & Elkan 1994), (Bailey & Elkan 1995a). The MEME extensions on which this paper focuses are methods of incorporating background knowledge, or coping with its lack. For incorporating background knowledge, these innovations include automatic detection of inverse-complement palindromes in DNA sequence datasets, and using Dirichlet mixture priors with protein sequence datasets. Dirichlet mixture priors bring information about which amino acids share common properties and thus are likely to be interchangeable in a given position in a protein motif. This paper also describes a new type of sequence model and a new heuristic for automatically determining the width of a motif which remove the need for the user to provide two types of information. The new sequence model type allows each each sequence in the training set to have exactly zero or one occurrences of each motif. This type of model is ideally suited to discovering multiple motifs in the majority of cases encountered in practice. The motif-width heuristic allows MEME to automatically discover several motifs of differing, unknown widths in a single DNA or protein dataset. We also describe an improved method of finding multiple, different motifs in a single dataset.

*Supported by NIH Genome Analysis Pre-Doctoral Training Grant No. HG00005.

Overview of MEME

The principal input to MEME is a set of DNA or protein sequences. Its principal output is a series of probabilistic sequence models, each corresponding to one motif, whose parameters have been estimated by expectation maximization (Dempster, Laird, & Rubin 1977). In a nutshell, MEME's algorithm is a combination of

- expectation maximization (EM),
- an EM-based heuristic for choosing the starting point for EM,
- a maximum likelihood ratio-based (LRT-based) heuristic for determining the best number of model free parameters,
- multistart for searching over possible motif widths, and
- greedy search for finding multiple motifs.

OOPS, ZOOPS, and TCM models

The different types of sequence model supported by MEME make differing assumptions about how and where motif occurrences appear in the dataset. We call the simplest model type OOPS since it assumes that there is exactly one occurrence per sequence of the motif in the dataset. This type of model was introduced by Lawrence & Reilly (1990). This paper describes for the first time a generalization of OOPS, called ZOOPS, which assumes zero or one motif occurrences per dataset sequence. Finally, TCM (two-component mixture) models assume that there are zero or more non-overlapping occurrences of the motif in each sequence in the dataset, as described by Bailey & Elkan (1994).

Each of these types of sequence model consists of two components which model, respectively, the motif and non-motif ("background") positions in sequences. A motif is modeled by a sequence of discrete random variables whose parameters give the probabilities of each of the different letters (4 in the case of DNA, 20 in the case of proteins)

occurring in each of the different positions in an occurrence of the motif. The background positions in the sequences are modeled by a single discrete random variable. If the width of the motif is W, and the alphabet for sequences is $\mathcal{L} = \{a, \ldots, z\}$, we can describe the parameters of the two components of each of the three model types in the same way as

$$\theta = [\ \theta_0 \quad \theta_1\ ] = [\ \mathbf{p}_0 \quad \mathbf{p}_1 \quad \mathbf{p}_2 \quad \cdots \quad \mathbf{p}_W\ ]$$
$$= \begin{bmatrix} P_{a,0} & P_{a,1} & P_{a,2} & \ldots & P_{a,W} \\ P_{b,0} & P_{b,1} & P_{b,2} & \ldots & P_{b,W} \\ \vdots & \vdots & \vdots & & \vdots \\ P_{z,0} & P_{z,1} & P_{z,2} & \ldots & P_{z,W} \end{bmatrix}.$$

Here, $P_{x,j}$ is the probability of letter x occurring at either a background position ($j = 0$) or at position j of a motif occurrence ($1 \leq j \leq W$), θ_0 is the parameters of the background component of the sequence model, and θ_1 is the parameters of the motif component.

Formally, the parameters of an OOPS model are the letter frequencies θ for the background and each column of the motif, and the width W of the motif. The ZOOPS model type adds a new parameter, γ, which is the prior probability of a sequence containing a motif occurrence. A TCM model, which allows any number of (non-overlapping) motif occurrences to exist within a sequence, replaces γ with λ, where λ is the prior probability that any position in a sequence is the start of a motif occurrence.

DNA palindromes

A DNA palindrome is a sequence whose inverse complement is the same as the original sequence. DNA binding sites for proteins are often palindromes. MEME models a DNA palindrome by constraining the parameters of corresponding columns of a motif to be the same:

$$\theta_1 = \begin{bmatrix} P_{a,1} & P_{a,2} & \ldots & P_{t,2} & P_{t,1} \\ P_{c,1} & P_{c,2} & \ldots & P_{g,2} & P_{g,1} \\ P_{g,1} & P_{g,2} & \ldots & P_{c,2} & P_{c,1} \\ P_{t,1} & P_{t,2} & \ldots & P_{a,2} & P_{a,1} \end{bmatrix}.$$

That is,

$$\begin{aligned} P_{a,i} &= P_{t,W+1-i}, \\ P_{c,i} &= P_{g,W+1-i}, \\ P_{g,i} &= P_{t,W+1-i}, \\ P_{t,i} &= P_{t,W+1-i} \end{aligned}$$

for $i = 1, \ldots, \lfloor W/2 \rfloor$. The last column is an inverted version of the first column, the second to last column is an inverted version of the second column, and so on. As will be described below, MEME automatically chooses whether or not to enforce the palindrome constraint, doing so only if it improves the value of the LRT-based objective function.

Expectation maximization

Consider searching for a single motif in a set of sequences by fitting one of the three sequence model types to it. The dataset of n sequences, each of length L, will be referred to as $X = \{X_1, X_2, \ldots, X_n\}$.[1] There are $m = L - W + 1$ possible starting positions for a motif occurrence in each sequence. The starting point(s) of the occurrence(s) of the motif, if any, in each of the sequences are unknown and are represented by the the variables (called the "missing information") $Z = \{Z_{i,j} | 1 \leq i \leq n, 1 \leq j \leq m\}$ where $Z_{i,j} = 1$ if a motif occurrence starts in position j in sequence X_i, and $Z_{i,j} = 0$ otherwise. The user selects one of the three types of model and MEME attempts to maximize the likelihood function of a model of that type given the data, $Pr(X|\phi)$, where ϕ is a vector containing all the parameters of the model. MEME does this by using EM to maximize the expectation of the joint likelihood of the model given the data and the missing information, $Pr(X, Z|\phi)$. This is done iteratively by repeating the following two steps, in order, until a convergence criterion is met.

- E-step: compute

$$Z^{(t)} = \mathop{\mathrm{E}}_{(Z|X,\phi^{(t)})}[Z]$$

- M-step: solve

$$\phi^{(t+1)} = \mathop{\mathrm{argmax}}_{\phi} \mathop{\mathrm{E}}_{(Z|X,\phi^{(t)})}[\log Pr(X, Z|\phi)]$$

where ϕ is a vector containing all the parameters of the model. This process is known to converge (Dempster, Laird, & Rubin 1977) to a local maximum of the likelihood function $Pr(X|\phi)$.

Joint likelihood functions. MEME assumes each sequence in the training set is an independent sample from a member of either the OOPS, ZOOPS or TCM model families and uses EM to maximize one of the following likelihood functions. The logarithm of the joint likelihood for models of each of the three model types is as follows. For an OOPS model, the joint log likelihood is

$$\begin{aligned} &\log Pr(X, Z|\theta) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j} \log Pr(X_i|Z_{i,j} = 1, \theta) + n \log \frac{1}{m}. \end{aligned}$$

For a ZOOPS model, the joint log likelihood is

$$\begin{aligned} &\log Pr(X, Z|\theta, \gamma) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j} \log Pr(X_i|Z_{i,j} = 1, \theta) \\ &\quad + \sum_{i=1}^{n} (1 - Q_i) \log Pr(X_i|Q_i = 0, \theta) \\ &\quad + \sum_{i=1}^{n} (1 - Q_i) \log(1 - \gamma) + \sum_{i=1}^{n} Q_i \log \lambda. \end{aligned}$$

[1] It is not necessary that all of the sequences be of the same length, but this assumption will be made in what follows in order to simplify the exposition of the algorithm. In particular, under this assumption, $\lambda = \gamma/m$.

For a TCM model, the joint log likelihood is

$$\begin{aligned}\log Pr(X, Z|\theta, \lambda) \\ = \sum_{i=1}^{n}\sum_{j=1}^{m}(1 - Z_{i,j}) \log Pr(X_{i,j}|\theta_0) \\ + Z_{i,j} \log Pr(X_{i,j}|\theta_1) \\ + (1 - Z_{i,j}) \log(1 - \lambda) + (Z_{i,j}) \log \lambda.\end{aligned}$$

The variable Q_i used in the ZOOPS likelihood equation is defined as $Q_i = \sum_{j=1}^{m} Z_{i,j}$. Thus, $Q_i = 1$ if sequence X_i contains a motif occurrence, and $Q_i = 0$ otherwise. The conditional sequence probabilities for sequences containing a motif used by OOPS and ZOOPS models are defined as

$$\begin{aligned}\log Pr(X_i|Z_{i,j} = 1, \theta) \\ = \sum_{k=0}^{W-1} \mathbf{I}(i, j + k)^T \log \mathbf{p}_k + \sum_{k \in \Delta_{i,j}} \mathbf{I}(i, k)^T \log \mathbf{p}_0,\end{aligned}$$

where $\mathbf{I}(i, j)$ is a vector-valued indicator variable of length $A = |\mathcal{L}|$, whose entries are all zero except the one corresponding to the letter in sequence X_i at position j, $X_{i,j}$. $\Delta_{i,j} = \{1, 2, \ldots, j - 1, j + w, \ldots, L\}$ is the set of positions in sequence X_i which lie outside the occurrence of the motif when the motif starts at position j. The conditional probability of a sequence without a motif occurrence under a ZOOPS model is defined as

$$Pr(X_i|Q_i = 0, \theta) = \prod_{k=1}^{L} P_{X_{i,k},0}.$$

The conditional probability of a length-W subsequence generated according to the background or motif component of a TCM model is defined to be

$$\log Pr(X_{i,j}|\theta_c) = \sum_{k=0}^{W-1} \mathbf{I}(i, j + k)^T \log \mathbf{p}_{k'},$$

where $k' = 0$ if $c = 0$ (background), and $k' = k + 1$ if $c = 1$ (motif).

The E-step. The E-step of EM calculates the expected value of the missing information—the probability that a motif occurrence starts in position j of sequence X_i. The formulas used by MEME for the three types of model are given below. Derivations are given elsewhere (Bailey & Elkan 1995b). For an OOPS model,

$$Z_{i,j}^{(t)} = \frac{Pr(X_i|Z_{i,j} = 1, \theta^{(t)})}{\sum_{j=1}^{m} Pr(X_i|Z_{i,j} = 1, \theta^{(t)})}.$$

For a ZOOPS model,

$$Z_{i,j}^{(t)} = \frac{f_i}{f_0 + \sum_{k=1}^{m} f_k}, \text{ where}$$

$$\begin{aligned} f_0 &= Pr(X_i|Q_i = 0, \theta^{(t)})(1 - \gamma^{(t)}), \text{ and} \\ f_j &= Pr(X_i|Z_{i,j} = 1, \theta^{(t)})\lambda^{(t)},\ 1 \le j \le m. \end{aligned}$$

For a TCM model,

$$Z_{i,j}^{(t)} = \frac{Pr(X_{i,j}|\theta_1^{(t)})\lambda^{(t)}}{Pr(X_{i,j}|\theta_0^{(t)})(1 - \lambda^{(t)}) + Pr(X_{i,j}|\theta_1^{(t)})\lambda^{(t)}}.$$

The M-step. The M-step of EM in MEME reestimates θ using the following formula for models of all three types:

$$\mathbf{p}_k^{(t+1)} = \frac{\mathbf{c}_k + \mathbf{d}_k}{|\mathbf{c}_k + \mathbf{d}_k|},\ 0 \le k \le W, \text{ where}$$

$$\mathbf{c}_k = \begin{cases} \mathbf{t} - \sum_{j=1}^{W} \mathbf{c}_j & \text{if } k = 0, \\ \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j}^{(t)} \mathbf{I}(i, j + k - 1) & \text{otherwise.} \end{cases}$$

Here $\mathbf{d}_k$ is a vector of pseudo-counts which is used to incorporate background information into EM as will be described later, $\mathbf{t}$ is the length-A vector of total counts of each letter the dataset, and $|\mathbf{x}|$ is the sum of the components of vector $\mathbf{x}$. For ZOOPS and TCM models, parameters γ and λ are reestimated during the M-step by the formula

$$\lambda^{(t+1)} = \frac{\gamma^{(t+1)}}{m} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j}^{(t)}.$$

Finding multiple motifs

All three sequence model types supported by MEME model sequences containing a single motif (albeit a TCM model can describe sequences with multiple *occurrences* of the *same* motif). To find multiple, non-overlapping, different motifs in a single dataset, MEME uses greedy search. It incorporates information about the motifs already discovered into the current model to avoid rediscovering the same motif. The process of discovering one motif is called a pass of MEME.

The three sequence model types used by MEME assume, *a priori*, that motif occurrences are equally likely at each position j in sequence X_i. This translates into a uniform prior probability distribution on the missing data variables $Z_{i,j}$. That is, initially, MEME assumes that $Pr(Z_{i,j} = 1) = \lambda$ for all $Z_{i,j}$.[2] On the second and subsequent passes, MEME changes this assumption to approximate a multiple-motif sequence model. A new prior on each $Z_{i,j}$ is used during the E-step that takes into account the probability that a new width-W motif occurrence starting at position $X_{i,j}$ might overlap occurrences of the motifs found on previous passes of MEME.

To help compute the new prior on $Z_{i,j}$ we introduce variables $V_{i,j}$ where $V_{i,j} = 1$ if a width-W motif occurrence could start at position j in sequence X_i without overlapping an occurrence of a motif found on a previous pass. Otherwise $V_{i,j} = 0$.

$$V_{i,j} = \begin{cases} 1, \text{ if no old motifs in } [X_j, \ldots, X_{j+w-1}] \\ 0, \text{ otherwise} \end{cases}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, L$.

[2]For an OOPS model, $\lambda = 1/m$. For a ZOOPS model, $\lambda = \gamma/m$.

To compute $V_{i,j}$ we use another set of binary variables $U_{i,j}$ which encode which positions in the dataset are *not* contained in occurrences of previously found motifs. So, $U_{i,j}$ is defined as

$$U_{i,j} = \begin{cases} 1, \text{ if } X_{i,j} \notin \text{previous motif occurrence} \\ 0, \text{ otherwise} \end{cases}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

As with the missing information variables $Z_{i,j}$, MEME computes and stores the *expected values* of the variables $U_{i,j}$. Before the first pass of MEME, the probability that $X_{i,j}$ is *not* already contained in a motif, the expected value of $U_{i,j}$, is set to one: $U_{i,j}^{(0)} = 1$ for $i = 1, \ldots, n$ and $j = 1, \ldots, L$. These values are updated after each pass according to the formula

$$U_{i,j}^{(p)} = U_{i,j}^{(p-1)}(1 - \max_{k=j-W+1,\ldots,j} Z_{i,k}^{(t)}) \quad (1)$$

where $Z_{i,j}^{(t)}$ is the final estimate of the missing information at the end of the current pass, p. Intuitively, we change the estimate of $X_{i,j}$ not being part of some motif by multiplying it by the probability of it not being contained in an occurrence of the current motif. This we estimate using the most probable motif occurrence of the current width that would overlap it. We use the maximum of $Z_{i,j}^{(t)}$ because occurrences of the current motif cannot overlap themselves, hence the values of $Z_{i,j}^{(t)}$ are not independent, so the upper bound on the probability used here is appropriate. The value of $U_{i,j}^{(p)}$ is then used as the value for $Pr(U_{i,j} = 1)$ in equation (2) below during the next pass, $p + 1$.

MEME estimates the probability of a width-W motif occurrence *not* overlapping an occurrence of *any* previous motif as the minimum of the probability of each position within the new motif occurrence *not* being part of an occurrence found on a previous pass. In other words, MEME estimates $Pr(V_{i,j} = 1)$ as

$$Pr(V_{i,j} = 1) = \min_{k=j,\ldots,j+W-1} Pr(U_{i,k} = 1). \quad (2)$$

The minimum is used because motif occurrences found on previous passes may not overlap (by assumption) so the values of $U_{i,j}$ are not independent. An approximate formula for reestimating $Z_{i,j}$ in the E-step of EM which takes motifs found on previous passes into account and thus approximates a multiple-motif model can be shown to be

$$\hat{Z}_{i,j}^{(t)} = \mathop{\mathrm{E}}_{(Z|X,\phi^{(t)})}[Z_{i,j}]\, Pr(V_{i,j} = 1).$$

MEME uses $\hat{Z}_{i,j}^{(t)}$ in place of $Z_{i,j}^{(t)}$ in the M-step of EM and in equation (1) above.

Using prior knowledge about motif columns

Applied to models of the forms described above, the EM method suffers from two problems. First, if any letter frequency parameter is ever estimated to be zero during EM, it remains zero. Second, if the dataset size is small, the maximum likelihood estimates of the letter frequency parameters tend to have high variance. Both these problems can be avoided by incorporating prior information about the possible values which the letter frequency parameters can take.

Using a mixture of Dirichlet densities as a prior in the estimation of the parameters of a model of biopolymer sequences has been proposed by Brown *et al.* (1993). This approach makes sense especially for proteins where many of the 20 letters in the sequence alphabet have similar chemical properties. Motif columns which give high probability to two (or more) letters representing similar amino acids are *a priori* more likely. A Dirichlet mixture density has the form $\rho = q_1\rho_1 + \ldots + q_R\rho_R$ where ρ_i is a Dirichlet probability density function with parameter $\beta^{(i)} = (\beta_a^{(i)}, \ldots, \beta_z^{(i)})$. A simple Dirichlet prior is the special case of a Dirichlet mixture prior where $R = 1$.

MEME uses Dirichlet mixture priors as follows. In the M-step, the mean posterior estimates of the parameter vectors $\mathbf{p}_i$, $i = 1$ to W, are computed instead of their maximum likelihood estimates. Let $\mathbf{c}_k = [c_a, \ldots, c_z]^T$ be the vector of expected counts of letters $a, \ldots, z$ in column k of the motif. We will consider this to be the "observed" letter counts in column k of the motif. The probability of component j in the Dirichlet mixture having generated the observed counts for column k is calculated using Bayes rule,

$$Pr(\beta^{(j)}|\mathbf{c}_k) = \frac{q_j\, Pr(\mathbf{c}_k|\beta^{(j)})}{\sum_{i=1}^{R} q_i\, Pr(\mathbf{c}_k|\beta^{(i)})}.$$

If we define $c = |\mathbf{c}_k| = \sum_{x\in\mathcal{L}} c_x$ and $b^{(j)} = |\beta^{(j)}| = \sum_{x\in\mathcal{L}} \beta_x^{(j)}$, then

$$Pr(\mathbf{c}_k|\beta^{(j)}) = \frac{\Gamma(c+1)\Gamma(b^{(j)})}{\Gamma(c+b^{(j)})} \prod_{x\in\mathcal{L}} \frac{\Gamma(c_x + b^{(j)})}{\Gamma(b^{(j)})}$$

where $\Gamma(\cdot)$ is the gamma function. We estimate the vector of pseudo-counts $\mathbf{d}_k = [d_a^k, d_b^k, \ldots, d_z^k]^T$ as

$$d_i^k = \sum_{j=1}^{R} Pr(\beta^{(j)}|\mathbf{c}_k)\beta_i^{(j)}$$

for $i = 1$ to A. The mean posterior estimate of the letter probabilities $\mathbf{p}_k$ is then

$$\mathbf{p}_k^{(t+1)} = \frac{\mathbf{c}_k + \mathbf{d}_k}{|\mathbf{c}_k + \mathbf{d}_k|}$$

for $k = 1$ to W. This gives the Bayes estimate of the letter probabilities for column k of the motif and is used to reestimate θ in the M-step.

Brown *et al.* (1993) have published several Dirichlet mixture densities that model well the underlying probability distribution of the letter frequencies observed in multiple alignments of protein sequences. The experiments reported in this paper use either their 30-component Dirichlet mixture prior or a 1-component prior where $\beta^{(1)}$ is just the average letter frequencies in the dataset.

Determining the number of model free parameters

The number of free parameters in a model of any of the MEME sequence model types depends on the width of the motif and on whether or not the DNA palindrome constraints are in force. When the width of the motifs is not specified by the user and/or when MEME is asked to check for DNA palindromes, MEME chooses the number of free parameters to use by optimizing a heuristic function based on the maximum likelihood ratio test (LRT).

The LRT is based upon the following fact (Kendall, Stuart, & Ord 1983). Suppose we successively apply constraints $C_1, \ldots, C_s$ to a model with parameters ϕ and let $\phi_{(s)}$ be the maximum likelihood estimator of ϕ when all constraints $C_1, \ldots, C_s$ have been applied. Then, under certain conditions, the asymptotic distribution of the statistic

$$\chi^2 = 2 \log \frac{Pr(X|\phi)}{Pr(X|\phi_{(s)})}$$

is central χ^2 with degrees of freedom equal to the number of independent constraints upon parameters imposed by $C_1, \ldots, C_s$.

MEME uses the LRT in an unusual way to compute a measure of statistical significance for a single model by comparing it to a "universal" null model. The null model is designed to be the simplest possible model of a given type. Let ϕ be the parameters of a model discovered by MEME using EM. Then, ϕ is the maximum likelihood estimate (MLE) for the parameters of the model.[3] Likewise, let ϕ_0 be the maximum likelihood estimate for the parameters of the null model. Since both ϕ and ϕ_0 are maximum likelihood estimates, the LRT can be applied to these two models. At some significance level between 0 and 1, the LRT would reject the null model in favor of the more complicated model. We define $LRT(\phi)$ to be this significance level, so

$$LRT(\phi) = Q(\chi^2|\nu), \text{ where}$$

$$Q(\chi^2|\nu) \approx Q(x_2), \quad x_2 = \frac{(\chi^2/\nu)^{1/3} - (1 - \frac{2}{9\nu})}{\sqrt{2/(9\nu)}}$$

(Abramowitz & Stegun 1972). $Q(x_2)$ is the Q function for the standard normal distribution (i.e., size of the right tail), and ν is the difference between the number of free parameters in the model used with EM and the null model. There are $A - 1$ free parameters per column of θ, so the difference in free parameters is $\nu = W(A - 1)$ for all three model types. If the DNA palindrome constraints are in force, half the parameters in θ_1 are no longer free and $\nu = (W/2)(A - 1)$.

To compute the value of $LRT(\phi)$ we need values of the likelihood functions for the given and null models and the difference in the number of free parameters between them. For the likelihood of the given model, MEME uses the value of the joint likelihood function maximized by EM. For the null model, it is easy to show that the maximum likelihood estimate has all columns describing motif and background positions equal to μ where $\mu = [\mu_a, \ldots, \mu_z]^T$ is the vector of average letter frequencies in the dataset. The log likelihood of the null model is

$$\log Pr(X|\phi_0) = nL \sum_{x \in \mathcal{L}} \mu_x \log \mu_x.$$

The criterion function which MEME minimizes is

$$G(\phi) = LRT(\phi)^{1/\nu}.$$

This criterion is related to the Bonferroni heuristic (Seber 1984) for correcting significance levels when multiple hypotheses are tested together. Suppose we only want to accept the hypothesis that ϕ is superior if it is superior to every model with fewer degrees of freedom. There are ν such models so the Bonferroni adjustment heuristic suggests to replace $LRT(\phi)$ by $LRT(\phi)\nu$. The function $G(\cdot)$ applies a much higher penalty for additional free parameters and yields motif widths much closer to those chosen by human experts than either $LRT(\phi)$ or $LRT(\phi)\nu$.

The MEME algorithm

The complete MEME algorithm is sketched below. The number of passes and maximum and minimum values of motif widths to try are input by the user. If the model type being used is OOPS, the inner loop is iterated only once since λ is not relevant. For a ZOOPS model, $\lambda_{min} = 1/(m\sqrt{n})$ and $\lambda_{max} = 1/(mn)$. For a TCM model, $\lambda_{min} = 1/(m\sqrt{n})$ and $\lambda_{max} = nm/(2w)$. The dynamic programming implementation of the inner loop, the EM-based heuristic for choosing a good value of $\theta^{(0)}$ as a starting point for EM, and the algorithms for shortening motifs and applying the DNA palindrome constraints are omitted here due to space limitations. They are described in a longer version of this paper (Bailey & Elkan 1995b). The time complexity of MEME is roughly quadratic in the size of the dataset.

```
procedure MEME ( X: dataset of sequences )
  for pass = 1 to pass_max do
    for W = W_min to W_max by ×√2 do
      for λ^(0) = λ_min to λ_max by ×2 do
        Choose good θ^(0) given W and λ^(0).
        Run EM to convergence from chosen
        value of φ^(0) = (θ^(0), λ^(0), W).
        Remove outer columns of motif
        and/or apply palindrome constraints
        to maximize G(φ).
      end
    end
    Report model which maximizes G(φ).
    Update prior probabilities U_{i,j} to
    approximate multiple-motif model.
  end
end
```

[3] We overlook the possibility that EM converged to a local maximum of the likelihood function. We note also that ϕ is actually the mean posterior estimate of the parameters, not the MLE, when a prior is used. In practice, the value of the likelihood function at ϕ is close to the value at the MLE.

name	*type*	*N*	*L*	*W*	*sites*	
					proven	*total*
lip	protein	5	182	16	5	5
					5	5
hth	protein	30	239	18	30	30
farn	protein	5	380	12	0	30
					0	26
					0	28
crp	DNA	18	105	20	18	24
lex	DNA	16	200	20	11	21
crplex	DNA	34	150	20	18	25
					11	21
hrp	DNA	231	58	29	231	231

Table 1: Overview of the datasets used in developing MEME showing sequence type, number of sequences (N), average sequence length (L), and motif width (W). Proven sites have been shown to be occurrences of the motif by laboratory experiment (footprinting, mutagenesis, or structural analysis). Total sites include the proven sites and sites reported in the literature based primarily on sequence similarity with known sites.

quantity	*mean*	(*sd*)
sequences per dataset	34	(36)
dataset size	12945	(11922)
sequence length	386	(306)
shortest sequence	256	(180)
longest sequence	841	(585)
pattern width	12.45	(5.42)

Table 2: Overview of the 75 Prosite datasets. Each dataset contains all protein sequences in SWISS-PROT (Release 11.1) annotated in the Prosite database as true positives or false negatives for the Prosite pattern characterizing a given family. Dataset size and sequence length count the total number of amino acids in the protein sequence(s).

Measuring performance

We measured the performance of the motifs discovered by MEME by using the final sequence model output after each pass of as a classifier. The parameters, ϕ, of the sequence model discovered on a particular pass are converted by MEME into a log-odds scoring matrix LO and a threshold t where $LO_{x,j} = \log(p_{x,j}/p_{x,0})$ for $j = 1, \ldots, W$ and $x \in \mathcal{L}$, and $t = \log((1-\lambda)/\lambda)$. The scoring matrix and threshold was used to score the sequences in a test set of sequences for which the positions of motif occurrences are known. Each subsequence whose score using LO as a position-dependent scoring matrix exceeds the threshold t is considered a hit. For each known motif in the test set, the positions of the hits were compared to the positions of the known occurrences. The number of true positive (tp), false positive (fp), true negative (tn) and false negative (fn) hits was tallied. From these, recall $= tp/(tp+fn)$ and precision $= tp/(tp+fp)$ were computed.

We also calculated the receiver operating characteristic (ROC) (Swets 1988) of the MEME motifs. The ROC statistic is the integral of the ROC curve, which plots the true positive proportion, $tpp = recall = tp/(tp+fn)$, versus the false positive proportion, $fpp = fp/(fp+tn)$. The ROC statistic was calculated by scoring all the positions in the test set using the log-odds matrix, LO, sorting the positions by score, and then numerically integrating tpp over fpp using the trapezoid rule.

MEME motifs which were shifted versions of a known motif were detected by shifting all the known motif positions left or right the same number of positions and repeating the above calculations of recall, precision and ROC. All shifts such that all predicted occurrences overlap the known occurrences (by exactly the same amount) were tried. The performance values reported are those for the best shift. For datasets with multiple known motifs, recall, precision and ROC were calculated separately for each known motif using each of the sequence models discovered during the passes of MEME.

Experimental datasets

We studied the performance of MEME on a number of datasets with different characteristics. Seven datasets which were used in the development of MEME are summarized in Table 1. Another 75 datasets each consisting of all the members of a Prosite family are summarized in Table 2.

Development datasets. The protein datasets lip, hth, and farn, were created by Lawrence *et al.* (1993) and used to test their Gibbs sampling algorithm. Very briefly, the lip dataset contains the five most divergent lipocalins with known 3D structure. They contain two known motifs, each occurring once in each sequence. The hth proteins contain DNA-binding features involved in gene regulation. The farn dataset contains isoprenyl-protein transferases, each with multiple appearances of three motifs.

The *E. coli* DNA datasets, crp, lex and crplex, are described in detail in (Bailey & Elkan 1995a). The crp sequences contain binding sites for CRP (Lawrence & Reilly 1990), while the lex sequences contain binding sites for LexA; the crplex dataset is the union of the crp and lex datasets. The *E. coli* promoter dataset hrp (Harley & Reynolds 1987) contains a single motif which consists of two submotifs with a varying number of positions (usually about 17) between them.

Prosite datasets. The 75 Prosite families described in general terms in Table 2 correspond approximately to the 10% of fixed-width Prosite patterns with worst combined (summed) recall and precision. Fixed-width patterns such as `D-[SGN]-D-P-[LIVM]-D-[LIVMC]` are a proper subset of the patterns expressible by MEME motifs, and they form a majority in Prosite. Recall and precision for Prosite patterns and for corresponding MEME motifs were calculated using information in the Prosite database about matches found when searching the large (36000 sequence) SWISS-PROT Release 11.1 database of protein sequences (Bairoch 1994).

Performance of different model types

Table 3 shows the ROC motifs found by MEME in the development datasets when MEME was run with the motif width set at $W \leq 100$ for 5 passes. The first lines for each of the three model types shows the performance of MEME without background information—DNA palindromes were not searched for and the one-component Dirichlet prior was used. As expected, the ZOOPS model type outperforms both the OOPS and TCM model types on those datasets which conform to the ZOOPS assumptions, as seen from the higher values of ROC for the ZOOPS model type (line 4) compared with the OOPS model type (line 1) for datasets hrp and crplex in Table 3. Accuracy is not sacrificed when *all* of the sequences contain a motif occurrence: the performances of the OOPS and ZOOPS model types are virtually identical on the first four datasets. The TCM model type outperforms the other two model types on the farn dataset whose sequences contain multiple occurrences of multiple motifs.

For comparison, the last line in Table 3 shows the performance of the motifs discovered using the Gibbs sampler (Lawrence *et al.* 1993). The conditions of the tests were made as close as possible to those for the MEME tests using the OOPS model type, except that *the Gibbs sampler was told the correct width of the motifs* since it requires the user to specificy the width of all motifs. With each Prosite dataset, the Gibbs sampler was told to search for 5 motifs, each of the width of the Prosite signature for the family, and that each sequence contained one occurrence of each motif. It was run with 100 independent starts (10 times the default) to maximize its chances of finding good motifs. Note that we did *not* tell either the Gibbs sampler or MEME how many occurrences of a particular motif a particular sequence has as was done in (Lawrence *et al.* 1993).

The ROC of the MEME motifs found using the ZOOPS model type without background information is as good or better than that of the sampler motifs for five of seven datasets. The MEME motifs found using the OOPS model type perform as well or better than those found by the Gibbs sampler with four of the seven datasets. Note once again that the Gibbs sampler was told the correct motif widths to use, whereas MEME was not. MEME using the ZOOPS model type does significantly better than the Gibbs sampler on the two ZOOPS-like datasets.

The benefit of background knowledge

The efficacy of using the DNA palindrome bias and the Dirichlet mixture prior can be seen in Table 3. ROC improves in 9 out of 21 cases and stays the same with another 5. The improvements are substantial in the case of the least constrained model type, TCM. For five of seven datasets, using the background information results in the model with the best or equal-best overall ROC.

The LRT-based heuristic does a good job at selecting the "right" width for the motifs in the seven non-Prosite datasets, especially when the DNA palindrome or Dirichlet mixture prior background information is used. The widths of the best motifs found by MEME are shown in Table 4. With background information and the model type appropriate to the dataset, the motif widths chosen by MEME are close to the correct widths with the exception of the lip dataset. That dataset is extremely small and the motifs are faint, which explains why MEME underestimates their widths.

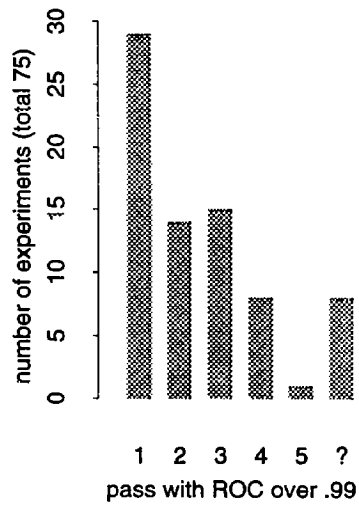

Figure 1: The pass where MEME finds the known Prosite motif is shown. MEME was run for five passes using the OOPS model without any background information. '?' means the known motif(s) were not found by MEME within five passes.

Performance on the Prosite datasets

MEME does an excellent job of discovering the Prosite motifs in training sets consisting of entire families. This is true with both the OOPS and ZOOPS model types and with or without the background information provided by the Dirichlet mixture prior. For 91% of the 75 Prosite families, one of the motifs found by MEME run for five passes using the OOPS model type and the simple prior corresponds to the known Prosite signature (i.e., identifies the same sites in the dataset). MEME finds multiple known motifs in many of the Prosite families. The criterion we use for saying that a MEME motif identifies a known Prosite pattern is that it have ROC of at least 0.99. MEME usually discovers the known motifs on early passes, as shown in Figure 1.

Of the 75 Prosite families we studied, 45 significantly overlap other families. We define significant overlap to mean two families share five or more sequences in common. If we include the motifs contained in these overlapping families, there are 135 known motifs present in the 75 Prosite family datasets. When run for 5 passes using the OOPS model type with the simple Dirichlet prior, MEME discovers 112 of these known motifs. The ZOOPS model type does better, discovering 117 of the 135 motifs. With the Dirichlet mixture prior, MEME does even better, discovering 119 out of 135 known motifs using either the OOPS or ZOOPS model types.

model type	*dataset*						
	OOPS-*like*				ZOOPS-*like*		TCM-*like*
	crp	lex	hth	lip	hrp	crplex	farn
OOPS	**0.9798**	0.9998	0.9979	**1.0000**	0.9123	0.9615	0.9446
OOPS_PAL	0.9792	**1.0000**			0.9123	0.9565	
OOPS_DMIX			**1.0000**	**1.0000**			0.9336
ZOOPS	**0.9798**	0.9999	0.9992	**1.0000**	**0.9244**	**0.9881**	0.9112
ZOOPS_PAL	0.9792	**1.0000**			**0.9244**	0.9867	
ZOOPS_DMIX			**1.0000**	**1.0000**			0.9324
TCM	0.9240	0.9895	0.9888	0.9842	0.8772	0.9764	0.9707
TCM_PAL	0.9786	0.9811			0.8772	0.9792	
TCM_DMIX			0.9841	0.9952			**0.9880**
OOPS_GIBBS	0.9709	1.0000	1.0000	0.9999	0.8881	0.9672	0.9291

Table 3: Average ROC of the best motif discovered by MEME for all known motifs contained in dataset. Highest ROC for each dataset is printed in boldface type. Blank fields indicate that the model type is not applicable to the dataset.

	dataset										
	OOPS-*like*					ZOOPS-*like*			TCM-*like*		
	crp	lex	hth		lip	hrp	crplex		farn		
known width	20	20	18	16	16	29	20	20	12	12	12
OOPS	**15**	18	15	**5**	**6**	46	29	18	7	9	10
OOPS_PAL	16	**16**				46	24	24			
OOPS_DMIX			**18**	**7**	**6**				8	16	11
ZOOPS	**15**	18	21	**5**	**6**	**46**	**21**	**18**	12	12	9
ZOOPS_PAL	16	**16**				**46**	22	20			
ZOOPS_DMIX			**18**	**7**	**6**				7	8	12
TCM	11	11	10	8	8	29	21	12	10	7	10
TCM_PAL	16	9				29	20	11			
TCM_DMIX			11	7	7				**11**	**7**	**8**

Table 4: Width of the best motif discovered by MEME for all known motifs contained in dataset. Blank fields indicate that the model type is not applicable to the dataset. A width in boldface indicates that this model type has the best average ROC for this dataset.

model type	*ROC*		*recall*		*precision*		*relative width*		*shift*	
OOPS	0.991	(0.025)	0.805	(0.356)	0.751	(0.328)	1.297	(0.753)	-0.978	(5.608)
OOPS_DMIX	0.992	(0.031)	0.815	(0.349)	0.758	(0.325)	1.210	(0.677)	-0.637	(5.337)
ZOOPS	0.992	(0.024)	0.823	(0.335)	0.775	(0.307)	1.307	(0.774)	-0.696	(5.575)
ZOOPS_DMIX	0.993	(0.026)	0.821	(0.340)	0.768	(0.314)	1.220	(0.715)	-0.585	(4.890)

Table 5: Average (standard deviation) performance and width of best motifs found by MEME in the 75 Prosite datasets. All 135 known motifs contained in the datasets are considered.

model type	*ROC*		*recall*		*precision*		*relative width*	
OOPS_DMIX $W \leq 100$	0.971	(0.065)	0.738	(0.288)	0.725	(0.310)	1.170	(0.840)
ZOOPS_DMIX, $W \leq 100$	0.960	(0.090)	0.728	(0.305)	0.699	(0.327)	1.141	(0.815)
OOPS_DMIX, $W = 20$	0.987	(0.029)	0.820	(0.211)	0.840	(0.228)	1.896	(0.785)
OOPS_GIBBS, $W = 20$	0.980	(0.053)	0.781	(0.242)	0.884	(0.169)	1.896	(0.785)

Table 6: Average (standard deviation) two-fold cross-validated performance of MEME and the Gibbs sampler on the 75 Prosite families. The training set consisted of half of the sequences in a given family. The test set consisted of the other half plus half of the 36000 sequences in SWISS-PROT Release 11.1.

Small improvements are seen in the performance of MEME motifs discovered in the Prosite datasets when the Dirichlet mixture prior is used. This is especially true for the datasets containing few (under 20) sequences. For the 36 Prosite datasets we used which meet this criterion and would thus be most likely to benefit from the background information contained in the Dirichlet mixture prior, the improvement in ROC is statistically significant at the 5% level for the OOPS model type according to a paired t-test. The motifs discovered using the ZOOPS model type are slightly superior to those found with the OOPS model type. Table 5 shows the average performance results on the Prosite datasets when MEME is run for five passes with various model types, with or without Dirichlet priors, and required to choose the motif width in the range $5 \leq W \leq 100$. The performance values are for all 135 known motifs contained in the 75 datasets, as described above. The difference in ROC between the OOPS and ZOOPS model types when the simple Dirichlet prior is used is significant at the 5% level. When the Dirichlet mixture prior is used, the difference in ROC between the two model types is not statistically significant. For both model types, whether or not the Dirichlet mixture prior is used does not make a statistically significant difference in the ROC of the discovered motifs.

The MEME motifs are extremely similar to the Prosite signatures. In general, they identify almost exactly the same positions in the sequences in the families. This fact can be seen in Table 5 from the high ROC, relative width close to 1, and small shift of the MEME motifs.

Generalization

Cross-validation experiments show that the motifs discovered by MEME on the Prosite datasets can be expected to correctly identify new members of the protein families. Table 6 shows the results of 2-fold cross-validation experiments on the 75 Prosite families using MEME and the Gibbs sampler. The first two lines of the table show the results when MEME is forced to choose the motif width. The performance of the OOPS model type is slightly better than that of the ZOOPS model type (ROC better at 5% significance level). Performance is better if MEME is given background information in the form of being told a good width ($W = 20$), as seen in the third line in Table 6. Then the generalization performance (cross-validated ROC) of the MEME motifs is better than that of sampler motifs at the 5% significance level. In these experiments, both MEME and the Gibbs sampler were allowed to generate only one motif per training set. The Gibbs sampler was instructed to use motif width $W = 20$ and 250 (25 times the default) independent starts to ensure that the two algorithms got approximately the same number of CPU cycles. The performance figures in Table 6 are based on the number of hits scored on sequences in SWISS-PROT known to be in the family, and do *not* require the hit to be at any particular position within the sequence. We used a threshold of 18 bits for determining if scores were hits.

A direct comparison of the predicted generalization performance of motifs discovered by learning algorithms such as MEME and the Gibbs sampler with that of the Prosite signatures is not possible. The Prosite signatures were created by hand and cannot easily be cross-validated, so their generalization performance is not known. However, the average performance of the Prosite signatures *on their own training sets*, ROC = 0.99(0.02), is the same as the *cross-validated performance* of the MEME OOPS-model motifs found when the algorithm is given a hint about the width of the motifs. This is impressive since the MEME motifs were learned from only half of the members of the families so the cross-validated ROC is likely to be an underestimate of the actual ROC of the motifs. The non-cross-validated estimate of the Prosite signature performance is likely to overestimate their actual performance on new sequences.

References

Abramowitz, M., and Stegun, I. A., eds. 1972. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover Publications, Inc.

Bailey, T. L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28–36. AAAI Press.

Bailey, T. L., and Elkan, C. 1995a. Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learning*. In press.

Bailey, T. L., and Elkan, C. 1995b. The value of prior knowledge in discovering motifs with MEME++. Technical Report CS95-413, Department of Computer Science, University of California, San Diego.

Bairoch, A. 1994. The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Research* 22(17):3578–3580.

Brown, M.; Hughey, R.; Krogh, A.; Mian, I. S.; Sjolander, K.; and Haussler, D. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In *Intelligent Systems for Molecular Biology*, 47–55. AAAI Press.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39(1):1–38.

Harley, C. B., and Reynolds, R. P. 1987. Analysis of *E. coli* promoter sequences. *Nucleic Acids Research* 15:2343–2361.

Kendall, S. M.; Stuart, A.; and Ord, J. K. 1983. *The Advanced Theory of Statistics*. Charles Griffin & Company Limited.

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

Lawrence, C. E.; Altschul, S. F.; Boguski, M. S.; Liu, J. S.; Neuwald, A. F.; and Wootton, J. C. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262(5131):208–214.

Seber, G. A. F. 1984. *Multivariate observations*. John Wiley & Sons, Inc.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. *Science* 270:1285–1293.