Finding Motifs for Insufficient Number of Sequences with Strong Binding to Transcription Factor *

Francis Y.L. Chin †, Henry C.M. Leung †, S.M. Yiu †, T.W. Lam †
Roni Rosenfeld ‡, W.W. Tsang †, David K. Smith §, Y. Jiang §

ABSTRACT

Finding motifs is an important problem in computational biology. Our paper makes two major contributions to this problem. Firstly, we better characterize the types of problem instances that cannot be solved by most existing methods of finding motifs. Secondly, we introduce a different method, which is shown to succeed for various problem instances for which popular existing methods fail.

Most existing computational methods to finding motifs are based on the *strong-signal model* wherein only *strong-signal sequences* (i.e. those that are known to contain binding sites very similar to the motif) are considered as input and *weak-signal sequences* (i.e. those do not contain any sub-string similar to the motif) are disregarded.

Buhler and Tompa have studied the limitations of methods based on the strong-signal model. They characterized the problem instances for which the motif is unlikely to be found in terms of the number of input (strong-signal) sequences needed under the assumption that each input sequence contains exactly one binding site. They further gave a method to calculate the minimum number of input sequences required.

We re-characterize the limitations of the strong-signal model in terms of the minimum total number of binding sites, rather than the minimum number of strong-signal sequences, required to be in the input data set. We use a probability matrix to represent a motif instead of a string pattern to calculate the minimum total number of binding sites required. This new characterization is shown to be more general and realistic.

Next, we introduce a more general and realistic *energy-based model*, which considers all available sequences (including weak-signal sequences) with varying degrees of binding strength to the transcription factors (as measured experimentally by observed color intensity). Given varying degrees of binding strength, our model can consider sequences ranging from those that contain more than one binding site to those that are weak sequences. By treating sequences with different degrees of binding strength differently, we develop a heuristic algorithm called EBMF (Energy-Based Motif Finding algorithm) using an EM-like approach to find motifs under our model. This EBMF algorithm can find motifs for data sets that do not even have the required minimum number of binding sites as previously derived for the strong-signal model. Our algorithm compares favorably with common motif-finding programs AlignACE and MEME, which are based on the strong-signal model. In particular, for some simulated and real data sets, our algorithm finds the motif when both AlignACE and MEME fail to do so.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life And Medical Sciences—*Biology and genetics*

General Terms

Algorithms

Keywords

Motif Finding, Transcription Factor, DNA Microarray, Binding Energy

1. INTRODUCTION

One great challenge in molecular biology is to understand the regulation of *gene expression* - the process by which a segment of DNA is decoded to form a protein. Two main steps for gene expression are *transcription* and *translation*. During the transcription process, an mRNA molecule is formed by copying a gene from the DNA. During the translation process, the mRNA is decoded to produce a protein.

To start the transcription process for a particular gene, one or more corresponding proteins, called *transcription factors*,

*This research is supported in part by an RGC grant
†Department of Computer Science & Information Systems, The University of Hong Kong, Hong Kong. Email: {chin, smyiu, cmleung2, twlam, tsang}@csis.hku.hk
‡School of Computer Science, Carnegie Mellon University, USA. Email: Roni.Rosenfeldcs.cmu.edu
§Department of Biochemistry, The University of Hong Kong, Hong Kong. Email: {dsmith, yyjiang}@hkusua.hku.hk

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
RECOMB'04, March 27–31, 2004, San Diego, California, USA.
Copyright 2004 ACM 1-58113-755-9/04/0003 ...$5.00.

have to bind to several specific regions, called *binding sites*, in the promoter region of the gene. A transcription factor can bind to multiple binding sites, but these sites typically have similar length (usually about 8 to 20 bp) and a common DNA sequence pattern. For most transcription factors, the common patterns for their corresponding binding sites, simply referred to as the *motifs*, are still unknown. Many laboratory-based methods for motif identification have been developed. However, these experimental methods are both expensive and time-consuming.

A recent trend in motif-finding is to make use of computational methods based on microarray data. Most existing computational methods [1, 2, 6, 7, 10, 11, 13, 15] are based on having a set of sequences that are known to contain binding sites with strong signals (i.e. the *strong-signal model*) as input. These approaches assume that a sufficient number of such *strong-signal sequences* are available. However, this assumption may not be valid for some transcription factors, and the number of strong-signal sequences may be too small to successfully find the motif using existing methods. Some motif-finding algorithms also consider sequences that are known *not* to contain any binding sites, in addition to strong-signal sequences [3, 8]. However, for these algorithms, the number of weak-signal sequences with plausible binding sites is used in the hyper-geometric analysis in order to compute the probability of such occurrences under the null-hypothesis. The lower the probability, the more plausible that such occurrences are not artifacts. No attempt is made to exploit the patterns of sequences without binding sites in order to find the motifs more effectively. Sequences without binding sites should not contain any patterns similar to the motif, and this can be a useful form of information. In fact, all sequences, strong-signal or weak-signal, with multiple occurrences of binding sites or without binding sites contain different information about the motifs in various forms and can be useful for motif-finding.

In this paper, we focus on finding motifs for data sets that contain insufficient number of sequences with strong signals. We first study the limitations of existing methods that are based on the strong-signal model. Then we introduce a more general and realistic energy-based model for dealing with data sets containing insufficient number of sequences with strong signals. The approach we use is different from that in [3, 8] in the sense that our model can handle sequences containing a varying amount of signal, i.e. varying from sequences contain multiple binding sites to sequences without any binding sites. Last, we show how our algorithm finds the correct motif under those situations that algorithms based on strong-signal model fail to do so.

1.1 Better Characterization for Strong-Signal Model

Buhler and Tompa [6] have studied the limitations of computational approaches based on the strong-signal model. They proposed a method to calculate the minimum number of input sequences required and showed that, if the number of input sequences is less than the minimum requirement, it is unlikely that there exists a computational approach that can identify the motif.

One important assumption in their study is that each input sequence contains exactly one binding site. In real situations, there can be multiple occurrences of binding sites, or *multiple binding sites*, for the same transcription factor in one sequence [4, 5, 12]. In other words, even if the number of strong-signal sequences in the input data set is small, there may still be enough binding sites or *signals* to enable the discovery of the motif. This observation is supported by an experiment using only three very special sequences with strong signal as input to identify the motif for GAL4, where each of the three sequences contained multiple binding sites (see Section 2 for more details). We tested this input set on two common motif-finding programs, AlignACE [7, 15] and MEME [1], which are based on the strong-signal model. We found that both programs could successfully identify the motif. Some natural questions to ask are then: how do we decide whether an input data set has enough signals for motif recovery, and what are the limitations of strong-signal model if we allow multiple binding sites in each sequence?

Our first contribution is to improve Bulher and Tompa's results by allowing multiple binding sites in each sequence. We characterize the limitations of the strong-signal model in terms of the minimum total number of binding sites, rather than the minimum number of strong-signal sequences, required to be in the input data set. Bulher and Tompa represent a motif of length l by a length-l string. A more general representation, which is used by most existing approaches, makes use of a probability matrix. The probability matrix is a $4 \times l$ matrix where the rows are indexed by the nucleotides "A","C","G","T" and each entry in the k-th column of the matrix represents the probability of the nucleotide's occurrence at position k of the binding site. So we represent a motif by a matrix instead of a string. Our characterization on the limitation of the strong-signal model is confirmed by some data sets on programs AlignACE and MEME.

1.2 Energy-Based Model

Existing algorithms are not effective to identify the motif for input data sets that contain insufficient number of sequences with strong signals (see Section 2 for experimental results). Our main contribution is a new approach to solving this problem.

Existing algorithms have the following problems. They assume that each binding site in the strong-signal model contains the same amount of signals. However, in reality, different binding sites have different binding strengths with the transcription factor, thus contain different amounts of signals. Also, sequences having comparatively weak signals (including sequences with a weak binding to the transcription factor and sequences without binding sites) are not used. In fact, these ignored weak-signal sequences also carry useful information for identifying the motif.

In our model, we introduce a more general and realistic energy-based model to capture previously-ignored information. We make use of the additional information from experiments and consider the binding strength (as measured experimentally by observing color intensity) of each available sequence. Intuitively the binding strength should relate to the degree of similarity between the motif and the binding

site in each sequence. Based on the binding strength, our model considers the amount of signals that a sequence actually contains. This allows us to make use of sequences with not so strong or even weak signals.

We then formulate the motif-finding problem in a way that allows multiple occurrences of binding sites in each sequence. We develop a heuristic algorithm call EBMF (Energy-Based Motif Finding algorithm) to solve the problem. We compare the performance of EBMF with those of AlignACE and MEME. EBMF is shown to be effective on both simulated and real data when the data sets contain insufficient number of sequences with strong signals. In particular, in our test cases, EBMF is able to identify the motif while both AlignACE and MEME fail to do so.

Our paper is organized as follows. Section 2 discusses the limitations of the strong-signal model when given input sequences with multiple binding sites. Section 3 presents the energy-based model. We also show how to convert existing experimental data to fit our model. A heuristic algorithm EBMF is given in Section 4. Section 5 compares the performance of EBMF with AlignACE and MEME. A conclusion is given in Section 6.

2. THE LIMITATION OF THE STRONG-SIGNAL MODEL WITH MULTIPLE BINDING SITES

With the assumption that each sequence contains exactly one binding site (a substring which is close to the motif in hamming distance), Buhler and Tompa [6] have studied the minimum number of input sequences required for finding the motif based on strong-signal model. In this section, we use a probability matrix to represent a motif and improve their results by allowing multiple binding sites in a sequence.

Let a motif of length l be represented by a $4 \times l$ probability matrix M where $M(c, j)$ represents the occurrence probability of the nucleotide c in the j-th position of a binding site. Given t input sequences each of length n, those algorithms based on strong-signal model want to find a probability matrix M and a background probability P_0 (which represents the occurrence probabilities of "A", "C", "G", "T" in the non-binding regions), which maximize the log likelihood (see [1]) of the t sequences generated according to the background probability P_0 with implanted binding sites generated according to matrix M. Formally, the log likelihood of a binding site b generated according to matrix M is $L(b, M) = \sum_{i=1}^{l} \log M(b[i], i)$. Assuming the background occurrence probabilities of "A", "C", "G", "T" are equal, i.e., each with 0.25 occurrence probability, the log likelihood of t length-n input sequences generated according to M and P_0 is

$$L_{total}(M) = \max\left\{\sum_{k=1}^{B} L(b_k, M) + (tn - Bl)\log(0.25)\right\}$$

among all possible sets of B non-overlap binding sites $\{b_k\}$ in the t sequences.

Suppose the input sequences are generated based on this model, that is, we generate t random sequences of length n based on the probability distribution P_0 and plant in them B^* instances of a motif randomly generated according to an arbitrary profile matrix M^*. Intuitively, if B^* is small or M^* looks too much like the background distribution, no algorithms can possibly pick out the B^* instances from the sequences without knowing M^*. It is because there exist many matrices M different from M^* (in the sense that the most probable strings generated according to M are quite different from those generated according to M^*), which have a log likelihood no less than $L_{total}(M^*)$. Therefore, the expected number of matrices, whose log likelihood are no less than $L_{total}(M^*)$, gives us an idea if it is possible to find the motif M^* from the input sequences. If the expected number of matrices is large, then finding the motif is impossible, otherwise it is highly probable.

Given a string Q of length l and a Hamming distance d, we define a probability matrix $M_{Q,d}$ such that for any j-th column of the matrix, the entry corresponding to the j-th character in Q is $(l - d)/l$ while the other entries in the same column are $d/3l$. We want to find the expected number of matrices in this format which have log likelihood no less than $L_{total}(M^*)$. If the expected number of matrices even in this restricted format and with log likelihood no less than $L_{total}(M^*)$ is large, it is impossible to find the motif M^* without extra information.

Assume the correct matrix is M^* and the expected log likelihood of a binding site b generated according to the matrix M^* is L_E. If the t sequences contain exactly B^* binding sites with respect to M^*, we can calculate the log likelihood of the t sequences generated according to M^* as $L_{total}(M^*) = B^*L_E + (nt - B^*l)\log(0.25)$. Now let us consider the log likelihood of a probability matrix $M_{Q,d}$. If the Hamming distance between a binding site b and Q is within d for $d \leq 3l/4$, then we can show that $L(b, M_{Q,d}) \geq (l-d)\log[(l-d)/l] + d\log(d/3l)$. The log likelihood of the t sequences generated according to $M_{Q,d}$ is $L_{total}(M_{Q,d})$ which is no less than $BL(b, M_{Q,d}) + (nt - Bl)\log(0.25)$ if the input sequences contain B non-overlap substrings whose Hamming distances from Q are within d (B can be different from B^*). Any $M_{Q,d}$ may be considered as a possible solution for the motif-finding algorithm if $L_{total}(M_{Q,d}) \geq L_{total}(M^*)$.

Given string Q and a random substring b (both of length-l and with equal occurrence probabilities for "A", "C", "G", "T"), the probability that the Hamming distance between Q and b is at most d where $0 \leq d \leq l$ is

$$p_d = \sum_{i=0}^{d} \binom{l}{i} \left(\frac{3}{4}\right)^i \left(\frac{1}{4}\right)^{l-i}$$

Let X be the sequence formed by concatenating the t input sequences (the length of X is nt) and b_i be the i-th substring in X such that the Hamming distance between b_i and Q is at most d.

We want to partition the sequence X into several non-overlap segments $X[k_{i-1} + 1 \ldots k_i]$ such that at the end of each segment, there exists exactly one substring $b_i = X[k_i - l + 1 \ldots k_i]$ whose Hamming distance with a fixed string Q is at most d. Let $B_{pos}(p, q)$ be the probability for the substring

$X[p \dots q]$ such that the Hamming distance between Q and $X[j \dots j+l-1]$, where $p \le j \le q-l$, is larger than d while the Hamming distance between Q and $b_i = X[q-l+1 \dots q]$ is at most d. We have $B_{pos}(p,q) = (1-p_d)^{q-p+1-l}p_d$.

Consider the probability $P_{Q,B}$ that X contains exactly B non-overlap substrings b_i at the positions $X[k_i - l + 1 \dots k_i]$ such that the Hamming distance between b_i and Q is no more than d while all other length-l substrings in X are of Hamming distance more than d from Q. Depending on the position of the last substring b_B, there are two cases to be considered.

Case I: $k_B > nt - l$ (the substring in X after the last binding site has length less than l, so it is impossible to have a binding site after k_B)

$$P_{Q,B} = \prod_{i=1}^{B} B_{pos}(k_{i-1}+1, k_i) = (1-p_d)^{k_B - Bl} p_d^B$$

Case II: $k_B \le nt - l$

$$\begin{aligned} P_{Q,B} &= (1-p_d)^{nt-k_B-l+1} \prod_{i=1}^{B} B_{pos}(k_{i-1}+1, k_i) \\ &= (1-p_d)^{nt-k_B-l+1}(1-p_d)^{k_B-Bl}p_d^B \end{aligned}$$

Note that the probability $P_{Q,B}$ is independent of the positions of the substrings b_i but depends on the ending position of the last binding site k_B. The probability $P_{Q,B}$ can then be expressed in term of the position of the last binding site j, the Hamming distance d and the number of binding sites B as follow,

$$PB(j,d,B) = \begin{cases} (1-p_d)^{j-Bl}p_d^B & j > nt-l \\ (1-p_d)^{nt-j-l+1}(1-p_d)^{j-Bl}p_d^B & \text{otherwise} \end{cases}$$

The probability of X that contains exactly B non-overlap substrings b_i (without considering the positions of the substrings) such that the Hamming distance between b_i and $Q \le d$ is the sum of probabilities $P_{Q,B}$ for all possible positions for the set of substrings $\{b_i\}$

$$\sum_{j=Bl}^{nt} \left[\begin{pmatrix} j-Bl+B-1 \\ B \quad 1 \end{pmatrix} PB(j,d,B) \right]$$

Assume X contains exactly B non-overlap substrings $\{b_i\}$ such that the Hamming distance between b_i and Q is no more than d. For each substring b_i, $L(b_i, M_{Q,d}) \ge (l-d)\log[(l-d)/l] + d\log(d/3l)$. Thus the log likelihood $L_{total}(M_{Q,d}) \ge B[(l-d)\log[(l-d)/l] + d\log(d/3l)] + (nt - Bl)\log(0.25)$.

The probability of X such that $L_{total}(M_{Q,d}) \ge L_{total}(M^*)$ is

$$\sum_{k=B'}^{\lfloor nt/l \rfloor} \left\{ \sum_{j=kl}^{nt} \left[\begin{pmatrix} j-kl+k-1 \\ k-1 \end{pmatrix} PB(j,d,k) \right] \right\}$$

where B' is the smallest number of binding sites for a matrix $M_{Q,d}$ such that the log likelihood of the t sequences generated according to $M_{Q,d}$ is no less than $L_{total}(M^*)$, i.e.

$$\begin{aligned} &B'\left[(l-d)\log\tfrac{l-d}{l} + d\log\tfrac{d}{3l}\right] \\ &\ge B^* L_E + (nt - B^* l)\log(0.25) \dots (1) \end{aligned}$$

By considering all possible substrings Q of length l and Hamming distance d, the expected number of matrices $M_{Q,d}$ such that $L_{total}(M_{Q,d}) \ge L_{total}(M^*)$ is approximately

$$\begin{aligned} &E(L_E, B^*) \\ &= 4^l \sum_{d=0}^{\lfloor 3l/4 \rfloor} \left\{ \sum_{k=B'}^{\lfloor nt/l \rfloor} \left\{ \sum_{j=kl}^{nt} \left[\begin{pmatrix} j-kl+k-1 \\ k-1 \end{pmatrix} PB(j,d,k) \right] \right\} \right\} \end{aligned}$$

According to Equation (1), B' is a function of L_E and B^*. (This is an approximation because the log likelihood of a given motif $M_{Q,d}$, $L_{total}(M_{Q,d}) \ge L_{total}(M^*)$ does not occur independently)

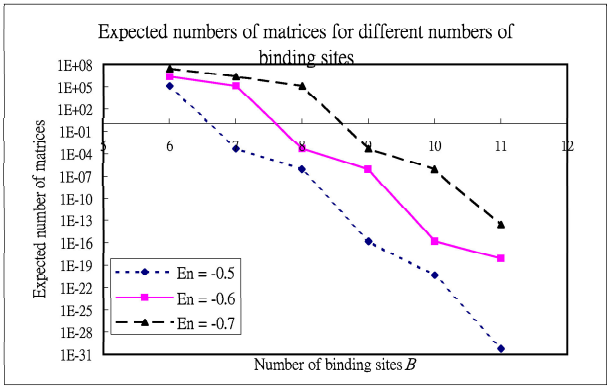


Figure 1: $E(L_E, B^*)$ **for different values of** B^* **and** En **where** $L_E = En \times l$**,** $t = 10$**,** $n = 700$**,** $l = 17$

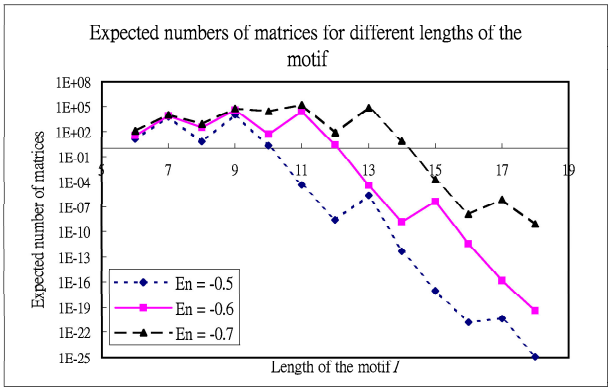


Figure 2: $E(L_E, B^*)$ **for different values of** l **and** En **where** $L_E = En \times l$**,** $t = 10$**,** $n = 700$**,** $B^* = 10$

Figure 1 shows the expected number $E(L_E, B^*)$ of matrices $M_{Q,d}$ with a log likelihood $L_{total}(M_{Q,d}) \ge L_{total}(M^*)$ for 10 input sequences, each of length 700 and the length of the motif is 17. It shows that the minimum required number of binding sites in the input sequences should be 7, 8, 9 (when the expected number of matrices $E(L_E, B^*) \le 1$) for En = -0.5, - 0.6, - 0.7 and L_E = -8.5, -10.2, -11.9 respectively, where L_E is the expected log likelihood of a binding site and $En = L_E/l$ is the expected log likelihood of a nucleotide in a binding site (note that it is negative of the entropy of a column in M^*). If the value of En increases, it means that

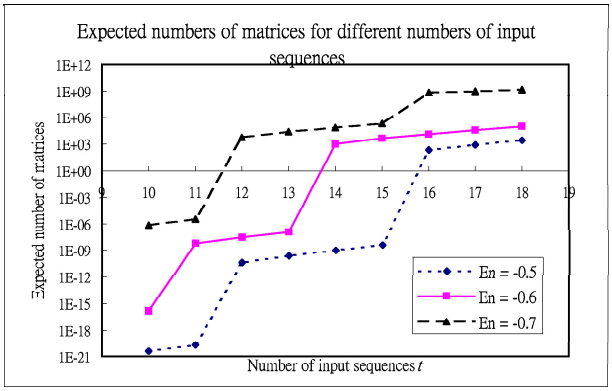


Figure 3: $E(L_E, B^*)$ **for different values of** t **and** En **where** $L_E = En \times l$, $n = 700$, $l = 17$, $B^* = 10$

each binding site contains more signal and less binding sites are required for finding motif. In other words, if the input sequences do not contain the least amount of binding sites, it is unlikely that any motif-finding algorithms based on strong-signal model can identify the real motif without extra information. Figure 2 shows the minimum required length of the motif for 10 input sequences of length 700 with 10 binding sites in total. As indicated in Figure 2, the shorter the motif, the less likely that the motif can be identified. For En = -0.5, -0.6, -0.7, the minimum lengths of the motif are 11, 13 and 15 respectively. Figure 3 shows the tendency of the values of $E(L_E, B^*)$ for different numbers of sequences of length 700 when the length of the motif is 17 and there are 10 binding sites in total. As indicated in Figure 3, if the total number of binding sites is fixed, the more the number of sequences in the input, the more noise in the data and the more difficult to find the motif.

We can also confirm our analysis by experiments which illustrate the limitations of existing programs, such as AlignACE and MEME. Gal4 is a well-studied transcription factor which activates genes necessary for galactose metabolism. Bing Ren et al.[14] found 10 genes to be bound by Gal4 and induced in galactose. The exact binding sites for most of these genes can be found in [4, 5, 12]. Given the 9 sequences of the intergenic regions (the gene Gal1 and Gal10 share one intergenic region), we want to test whether MEME and AlignACE can find the published motif pattern CGGN11CCG of Gal4 in different input sequences with different values of B^*. From the published binding sites, we calculate the expected log likelihood L_E of a binding site which is -11.47 ($En = -0.67$). Table 1 confirms our analysis that motif can be found in the first three cases and definitely not in the last case. In the first three cases, the values of $E(L_E, B^*)$ are very small and the numbers of binding sites in the input data are more than the minimum number required. On the other hand, in the last case, $E(L_E, B^*)$ is much larger than 1 and the number of binding sites is less than the minimum number required, so it is difficult to find the correct motif pattern.

Table 1: Results of AlignACE and MEME on Gal4

	$\overline{n}$	B	Min B	$E(L_E, B)$	AlignACE		MEME	
					Find?	rank	Find?	rank
9 seq.	762	18	9	3.055×10^{-52}	yes	1	yes	1
8 seq.	736	13	9	2.301×10^{-22}	yes	1	yes	1
7 seq.	746	9	9	2.298×10^{-7}	yes	1	no	-
6 seq.	749	7	9	2532	no	-	no	-

Min B is the minimum value of B such that $E(L_E, B) \leq 1$.

3. OUR ENERGY-BASED MODEL AND PROBLEM DEFINITION

In order to make use of the information contained in weak-signal sequences for motif finding, we propose a more general energy-based model in this section. In the next subsection, we show an example how to estimate the binding energy between a sequence and a transcription factor from a real experiment.

3.1 Applying the Model to a Real Case

In the microarray experiment, multiple copies of a transcription factor TF are added to a solution of genome of a particular organism. At the equilibrium state, some copies of transcription factor may bind to the binding sites of different sequences s_i in the genome while some copies of the transcription factor and sequences are free. Let e_i be the average binding energy between the transcription factor TF and sequence s_i, then $e_i = -ln(K_e)$ where the binding constant $K_e = [TF \bullet s_i]/[TF][s_i]$ (number of sequences bound to the transcription factors over number of free sequences) with the binding reaction modeled by $TF + s_i \Longleftrightarrow TF \bullet s_i$ [9]. Note that the unit of e_i is in (RT) where T is the constant temperature throughout the experiment and R is a constant.

For each sequence in the genome of a particular organism, we will get from the experiment a color intensity which represents the value of K_e. Therefore, the color intensity can be used to estimate the binding energy between the transcription factor and each input sequence.

3.2 Energy-Based Model

In our model, we do not treat the input sequences equally. Each sequence is associated with a value e_i which represents the binding energy between the transcription factor and its binding sites (which can be multiple). Let sequence s_i contain B_i binding sites and $E(b_{ij}, M)$ be the binding energy between the transcription factor and the j-th binding site b_{ij} in sequence s_i. The probability that the transcription factor binds to b_{ij} [9] is

$$P_{ij} = \frac{e^{-E(b_{ij},M)}}{\sum_{k=1}^{B_i} e^{-E(b_{ik},M)}} \ldots (2)$$

We use a $4 \times l$ energy matrix M to represent the motif where the row of this matrix is indexed by "A","C","G","T". $M(c, j)$ represents the binding energy of the transcription factor and the nucleotide c at the j-th position of the binding site. The total binding energy between binding site b and the transcription factor can be approximated by $E(b, M) = \sum_{j=1}^{l} M(b[j], j)$ where $b[j]$ is the j-th character of b.

The set of substrings in a sequence s_i, which are likely to be bound by the transcription factor, is said to be the binding sites of s_i. For a sequence s_i, the binding sites b_{ij} are those substrings with $E(b_{ij}, M) \leq \alpha$ where α is a determined threshold. If s_i does not contain any substring b such that $E(b, M) \leq \alpha$, the substring b with the lowest $E(b, M)$ will be chosen as its binding site. As for those binding sites that are too close to each other, i.e., the distance between each of two binding sites is less than some determined value d_{min}, we assume that there will not be two or more transcription factors bound to these binding sites simultaneously. While for those binding sites whose distances are larger than d_{min}, each of them can be bound by a transcription factor at the same time. We define $E_{total}(s_i, M)$ to be the expected binding energy between the transcription factor and sequence s_i given that at least one binding site in s_i is bound by the transcription factor.

3.3 Problem Definition

Given the length of binding sites l, an energy threshold α, a distance threshold d_{min}, t sequences $S = \{s_i\}$ in which each sequence s_i has a corresponding binding energy e_i, we want to find a $4 \times l$ energy matrix M to minimize the *prediction error*

$$\sum_{i=1}^{t} (E_{total}(s_i, M) - e_i)^2$$

4. ENERGY-BASED FINDING MOTIF ALGORITHM

EBMF tries to predict the $4 \times l$ energy matrix M from the input sequences using two steps. In the first step, we identify a set of candidate matrices based on the strings that occur frequently in the input sequences of strong signal. In the second step, we refine each candidate matrix using an EM-like iteration, which can be described as follows. Based on the candidate matrix, find the best possible binding sites for each sequence (see Section 4.2). These binding sites together with the given binding energy for each sequence are used to calculate another energy matrix so as to minimize the prediction error. The iteration process is repeated until there is no further decrease in the prediction error or the number of iterations reaches a certain value. After processing all candidate matrices, the top 10 matrices that give the smallest prediction errors are considered as the actual energy matrices. We first describe the details of an EM-like step in refining the candidate matrix.

4.1 Refine the Candidate Motif

Let the B_i best possible binding sites be $b_{i1}, \ldots, b_{iB_i}$ for each sequence s_i with respect to candidate matrix M. Based on the user input d_{min}, we estimate the expected binding energy $E_{total}\{s_i, M'\}$ for an arbitrary matrix M' as follows. We divide the B_i binding sites b_{ij} into several groups described by subsets $BS_{i1}, \ldots, BS_{ip}$ of binding sites b_{ij} where $BS_{i1} \cup \ldots \cup BS_{ip} = \{b_{i1}, \ldots, b_{iB_i}\}$. For any two binding sites in the same group BS_{ik}, the distance between them is within d_{min} (i.e. if $b_{im}, b_{in} \in BS_{ik}$ then the distance between b_{im} and $b_{in} \leq d_{min}$) while the distance between any two binding sites in different groups is larger than d_{min}. Note that $BS_{i1}, \ldots, BS_{ip}$ may not be disjoint; however, in practice, they are disjoint in most cases. The expected binding energy of a transcription factor bound to a binding site in BS_{ik} is $\sum_{b_{ij} \in BS_{ik}} P_{ij} E(b_{ij}, M')$ where P_{ij} is given in Equation (2). Given that at least one binding site is bound by the transcription factor, the expected binding energy between the transcription factor and sequence s_i can be calculated as follows:

$$\begin{aligned} E_{total}(s_i, M') &= \sum_{\text{all } BS_{ik}} \left[\frac{\sum_{b_{ij} \in BS_{ik}} P_{ij} E(b_{ij}, M')}{1 - \prod_{\text{all } BS_{ik}} \left(1 - \sum_{b_{ij} \in BS_{jk}} P_{ij}\right)} \right] \\ &= \frac{\sum_{j \in \{1,\ldots,B_i\}} P_{ij} E(b_{ij}, M')}{1 - \prod_{\text{all } BS_{ik}} \left(1 - \sum_{b_{ij} \in BS_{ik}} P_{ij}\right)} \end{aligned}$$

We then formulate an equation by setting this expected binding energy equal to the given binding energy of that sequence, that is, $E_{total}(s_i, M') = e_i$. With t input sequences, we have a system of t equations. We use QR decomposition to solve this system of equations to obtain all $4l$ entries of the new energy matrix M' that minimizes the predication error.

Technically, we convert each character in b_{ij} for any j in BS_{ik} to a 4-dimensional vector by using (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) to represent "A", "C", "G" and "T" respectively. The resultant $4l$-dimensional vector v_{ij} is used to represent the binding site b_{ij} of length l. For example, we convert "ATC" to a 12-dimensional vector (1,0,0,0,0,0,0,1,0,1, 0,0). Then, the equation for sequence s_i can be represented as follows,

$$\left\{ \sum_{j=1}^{B_i} \left[\frac{P_{ij}}{1 - \prod_{\text{all } BS_{ik}} \left(1 - \sum_{b_{im} \in BS_{ik}} P_{im}\right)} \times v_{ij} \right] \right\} \times V(M')^T = e_i$$

where P_{ij} is the probability that the transcription factor is bound to b_{ij} w.r.t. M (see Section 2.2) and $V(M') = (M'(1,1), M'(2,1), M'(3,1), M'(4,1), M'(1,2), \ldots, M'(4,l))$ represents the vector formed by concatenating the column entries of M'.

4.2 Finding Candidate Matrices

When the algorithm based on the energy model is applied to find the motif, not all the initial matrices can converge to the correct matrix M^*. The success of the algorithm depends very much on the set of candidate matrices chosen as "seed". For example, if we use a random string Q of length l to construct a $4 \times l$ matrix M as the seed where $M(Q[i], i) = -1$ for $1 \leq i \leq l$ and 0 for all other entries, it can be confirmed from experiments that the success rate is

very low at about 0.3%. In the following, we show a better method of finding the seeds.

4.2.1 Improved Method for Finding Seed

Our approach to find a seed matrix is to select the most likely string Q among the 4^l possible strings by voting. Each substring σ of length l appearing in the input sequences will give a score to every string Q with similar pattern (that is, the hamming distance between σ and Q is within a given threshold). The set of strings received the highest scores will be chosen for converting to seed matrices. However, the votes should carry different weights depending on the binding energy e_i of the sequence from where σ is derived. In our experiment, we have defined the score function as follows.

$$\text{Score}(s_i, Q) = \begin{cases} -e_i / \prod_{k=1}^{l/2} P_0(Q[k]) \\ \text{if } \exists \text{ a substring } \sigma \text{ in } s_i \text{ s.t. } H(\sigma, Q) \leq \lfloor l/8 \rfloor \\ \\ 0 \text{ otherwise} \end{cases}$$

where $H(\sigma, Q)$ defines the hamming distance between σ and Q and $P_0(c)$ the occurrence probability of c in X where c is "A", "C", "G", or "T".

In general, it is very time-consuming to find the highest scoring Q among the 4^l ($= 2^{34}$ if $l = 17$) possible strings. In order to reduce the number of tests, we need to reduce the length of the "seed". One way to do this is the following. Given a string Q of length l, we project the $l/2$ characters at the odd positions of C to form a representative string of length $l/2$. For example, when $l = 8$, we will use "ACAC" to represent "ATCGATCG". We modify the scoring function such that $H(\sigma, Q)$ is the hamming distance between the representative string of σ and Q, and we calculate the product of $P_0(Q[k])$ for odd number k only. Instead of finding the scores of all the 4^l possible strings of length l, we find the scores for the $4^{l/2}$ representative strings of length $l/2$ and use those representative strings with high scores to predict the candidate matrices. Similarly, we can get another set of candidate matrices if we project the even positions of a string to form the representative string.

5. EXPERIMENTAL RESULTS

We have implemented EBMF in C++ and tested it on both real data and simulated data. We compare EBMF with common motif-finding programs AlignACE and MEME. The results show that EBMF is effective and compares favorably with these programs.

5.1 Simulated Data

Let m be the total number of sequences, n be the length of each sequence, t be the number of sequences with binding sites and B^* be the number of binding sites in the t sequences, we generate the simulated data as follow. A $4 \times l$ energy matrix E^* is generated randomly and a corresponding probability matrix M^* is constructed such that for each column j in M^*, the probability of the occurrence of a nucleotide c is directly proportional to $e^{-E^*(c,j)}$. Then we generate m sequences of length n where each nucleotide occurs with equal probability, and plant B^* binding sites (generated according to the probability matrix) in these t sequences at random positions. Finally, we use the energy matrix E^* to calculate the energy level $e_i = E_{total}(s_i, E^*)$ of each sequence s_i.

Table 2: Results of the algorithms on simulated data for 200 sequences of length 700 where 10 of them contain B binding sites of length 17 with expected likelihood -10

	$E(L_E, B)$	EBMF		AlignACE		MEME	
		Find?	rank	Find?	rank	Find?	rank
$B = 7$	149475	yes	1	no	-	no	-
$B = 8$	0.000439	yes	1	no	-	yes	1
$B = 9$	7.70349×10^{-7}	yes	1	yes	1	yes	1

Table 3: Results of the algorithms on simulated data for 200 sequences of length 700 where 10 of them contain B binding sites of length 17 with expected likelihood -8.8

	$E(L_E, B)$	EBMF		AlignACE		MEME	
		Find?	rank	Find?	rank	Find?	rank
$B = 6$	619609	yes	1	no	-	no	-
$B = 7$	0.000439	yes	1	no	-	yes	1
$B = 8$	7.70353×10^{-7}	yes	1	yes	1	no	-

Tables 2 and 3 show the results of AlignACE, MEME and EBMF on the simulated data. There are situations in which EBMF finds the motif while AlignACE and MEME fail to do so. This is because when the number of binding sites in the sequences is small, there exist many matrices whose log likelihoods are no smaller than that of matrix M^*. In fact, there is an infinite number of such matrices. When these matrices in turn represent many different strings, AlignACE and MEME will fail. The EBMF algorithm can help in these situations by using weak-signal sequences to eliminate the number of matrices and, more importantly, the number of different strings they represent, to the extent that the motif can be found.

5.2 Real Data

Using Gal4 as an example, we know from Section 2 that once we remove several sequences containing multiple binding sites, both MEME and AlignACE cannot find the motif pattern CGGN11CCG. [4, 5, 12]. In this section, we test whether our algorithm can discover the correct pattern in similar situation.

From the mircoarray experiment (data from [14]), we obtained 6000 intergenic regions (the length of the sequences is in the range $[100, 1000]$), each with a color intensity. After sorting the sequences according to their color intensities in decreasing order, we remove the 2,3,4 and 6 sequences from the data set, which contain multiple binding sites with strong signal. We want to find the motif using this weak data set.

For AlignACE and MEME, no matter how we set the threshold for selecting the top strong-signal sequences, the motif cannot be found. However, since the EBMF algorithm take advantage of weak-signal sequences, we can find the CGGN11CCG pattern using the top 100 sequences(Table 4).

Table 4: Results of the algorithms on Gal4

	Our Algorithm		AlignACE		MEME	
	Find?	rank	Find?	rank	Find?	rank
Using the top 100 sequences in the original data	yes	2	yes	1	yes	1
Using the top 100 sequences except sequences 2,3,4 and 6	yes	1	no	-	no	-
Using the top 100 sequences except sequences 1 to 6	yes	10	no	-	no	-
Using the top 100 sequences except sequences 1 to 8	yes	5	no	-	no	-

We set the numbers of input sequences be different values for AlignACE and MEME. We say AlignACE and MEME can find the motif if they can find the CGGN11CCG pattern in at least one setting.

6. CONCLUSION

In this paper, we have characterized data sets for which existing motif-finding algorithms, which are based on the strong-signal model, succeed to find the motif in terms of the minimum number of binding sites the data set must have. This characterization provides a better description of the data set for which we can expect success.

Commonly-used motif-finding programs, such as AlignACE and MEME, are based on strong-signal model, where the patterns of weak-signal sequences are ignored. Clearly, weak-signal sequences, such as sequences without binding sites, also contain information about motif in the negative sense, although possibly less than information from strong-signal sequences. For data sets which do not have the minimum number of binding sites, we have proposed a new EMBF algorithm for finding motifs, which makes use the information of weak-signal sequences in order to outperform AlignACE and MEME. However, our EBMF algorithm in its present state has two shortcomings which require attention and will be addressed in our future papers.

1. Comparatively, our EBMF algorithm is rather slow and takes a much longer time to identify the motif than other motif-finding algorithms. We believe, however, time improvement can be realized through a more efficient way of finding "seed" matrices (Section 4.2.1).

2. For most data sets, exact information about each sequence's binding energy is not available. It is then desirable to devise another approach to address data sets with only two groups of sequences - those with and those without binding sites.

7. REFERENCES

[1] Timothy L. Bailey and Charles Elkan. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proceedings of Second International Conference, pp 28-36, 1994.

[2] Timothy L. Bailey and Charles Elkan. *Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization.* Machine Learning Journal, 21, pp 51-83, 1995.

[3] Yoseph Barash, Gill Bejerano and Nir Friedman. *A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites.* Proceedings of WABI, pp 278-293, 2001.

[4] Richard J. Bram and Roger D. Kornberg. *Specific protein binding to far upstream activating sequences in polymerase II promoters.* Proceedings of the National Academy of Sciences, 82, 43-47, 1984.

[5] Richard J. Bram, Neal F. Lue and Roger D. Kornberg. *A GAL family of upstream activating sequences in yeast: roles in both induction and repression of transcription.* The EMBO Journal, 5 , pp 603-608, 1986.

[6] Jeremy Bulher and Martin Tompa. *Finding Motifs Using Random Projections.* Journal of Computional Biology, 9, pp 225-242, 2002.

[7] J.D. Hughes, P.W. Estep, S. Tavazoie and G.M. Church. *Computational identification of cis-regulatory elements associated with groups.* Journal of Molecular Biology, 296(5), pp 1205-14, 2000.

[8] Lars M. Jakt, Liang Cao, Kathryn S.E. Cheah and David K.Smith. *Assessing Clusters and Motifs from Gene Expression Data.* Genome Research, 11, 112-123, 2001.

[9] I.M. Klotz. *Introduction to Biomolecular Energetics.* London UK:Academic Press Inc, 1986.

[10] Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., and Wootton, J. *Detecting subtle sequence signals: A Gibbs sampling strategy.* Science, 262, pp 208-214, 1993.

[11] Jun S. Liu, Andrew F. Neuwald, Charles E. Lawrence. *Bayesian Motifs for Multiple Local Sequence Alignment and Gibbs Sampling Strategies.* Journal of the American Statistical Association, Volume 90, issue 432, pp 1156-1170, 1995.

[12] Viktor Magdolen, Ulrich Occhsner, Paul Trommler and Wolfhard Bandlow. *Transcriptional control by galactose of a yeast gene encoding a protein homologous to mammalian aldo/keto reductases.* Gene, 90, pp 105-114, 1990.

[13] Pavel A. Pevzner and Sing-Hoi Sze. *Combinatorial approaches to finding subtle signals in DNA sequences.* Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp 269-278, 2000.

[14] Bing Ren, Francois Robert, John J. Wyrick, Oscar Aparicio, Ezra G. Jennings, Itamar Simon, Julia Zeitlinger, Jorg Schreiber, Nancy hannett, Elenita Kanin, Thomas L. Volkert, Christopher J. Wilson, Stephen P. Bell, Richard A. Young. *Genome-Wide Location and Function of DNA Binding Proteins.* Science, 290, 1993.

[15] F.R. Roth, J.D. Hughes, P.E. Estep and G.M. Church. *Finding DNA Regulatory Motifs within Unaligned Non-Coding Sequences Clustered by Whole-Genome mRNA Quantitation.* Nature Biotechnology, 16(10), pp 939-45, 1998.