DNA Sequence Classification via an Expectation Maximization Algorithm and Neural Networks: A Case Study*

Qicheng Ma
Novartis Pharmaceuticals Corporation
556 Morris Avenue
Summit, NJ 07901, USA

Jason T. L. Wang†
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102, USA

Dennis Shasha
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012, USA

Cathy H. Wu
National Biomedical Research Foundation
Georgetown University Medical Center
3900 Reservoir Road, NW, Washington, DC 20007, USA

Running title: DNA Sequence Classification

Keywords: Bayesian inference, Bioinformatics, Data mining, Expectation maximization, Neural networks, Promoter recognition

*This work was supported in part by NSF grants IIS-9988345 and IIS-9988636.
†Corresponding author. Tel: (973) 596-3396; Fax: (973) 596-5777; Email: jason@cis.njit.edu.

Abstract

This paper presents new techniques for biosequence classification, with a focus on recognizing E. Coli promoters in DNA. Specifically, given an unlabeled DNA sequence S, we want to determine whether or not S is an E. Coli promoter. We use an expectation-maximization (EM) algorithm to locate the -35 and -10 binding sites in an E. Coli promoter sequence. The EM algorithm differs from previously published EM algorithms in that, instead of assuming a uniform distribution for the lengths of the spacer between the -35 binding site and the -10 binding site as well as the spacer between the -10 binding site and the transcriptional start site, our algorithm deduces the probability distribution for these lengths. Based on the located binding sites, we select features in each E. Coli promoter sequence according to their information contents and represent the features using an orthogonal encoding method. We then feed the features to a neural network for promoter recognition. Empirical studies show that the proposed approach achieves good performance on different datasets.

1 Introduction

Promoters are transcription signals, which regulate gene expressions. Characterization and recognition of such signals is an important research topic and has been studied by many researchers. [15], [18], [20], [27], for example, analyzed E. Coli promoters. [19] compiled and clustered a set of promoters recognized by E. Coli RNA polymerase. More recently, [6], [12], [23] considered eukaryotic promoters and presented techniques for detecting these signals.

In this paper we focus on the recognition of E. Coli promoters. Specifically, the problem we study here can be formulated as follows. Given an unlabeled DNA sequence S, we want to determine whether or not S is an E. Coli promoter. This is also known as the binary classification problem [28], [31] widely studied in the data mining (DM) field. In binary classification, one is given some training data including both positive and negative examples. The positive data belong to a target class (E. Coli promoters in our case), whereas the negative data belong to the non-target class. Based on the training data, the classifier will be able to assign unlabeled test data to either the target class or the non-target class. The importance of the binary classification problem has been addressed in the DM literature [9], [30].

In the past, several researchers have considered the binary classification problem for E. Coli promoters. In [27], Towell and Shavlik proposed to initialize the topology and weights of a neural network according to the characteristics of E. Coli promoters. They built a system, called KBANN, for recognizing the promoters. Later, Opitz [18] employed a genetic algorithm to search through the topology space of multiple neural networks. He developed a system, called REGENT, which created the initial population of the neural networks by utilizing KBANN. The fitness of each neural network was measured on a separate validation dataset. Recognizing and prediction of E. Coli promoters were performed using an ensemble of the neural networks. In both KBANN and REGENT, each promoter sequence was regarded as a 57 attribute tuple, where the number 57 is the length of the promoter sequence in their dataset. The authors employed an orthogonal encoding method to encode the E. Coli promoter sequences.

In [15], Mahadevan and Ghosh developed a three-phase process for recognizing E. Coli promoters. First, a neural network was employed to locate the two binding sites in each E. Coli promoter. The authors then aligned the promoters with respect to their binding sites and built another neural network for promoter recognition. In contrast to the previous work [15], [18], [27], we propose here to use an expectation-maximization (EM) algorithm [7] to locate the two binding sites of an E. Coli promoter. We then align training promoters with respect to their binding sites. Next, we choose features in each training promoter according to their information contents and represent the features based on an orthogonal encoding method. These features are then fed to a neural network, which is used to determine whether or not an unlabeled DNA sequence is an E. Coli promoter. While we focus on promoter classification here, our techniques and the framework of combining EM algorithms with neural networks should generalize to other domains for classifying other types of data.

The rest of the paper is organized as follows. Section 2.1 describes the characteristics of E. Coli promoters. Section 2.2 presents the EM algorithm for locating the binding sites of a promoter sequence. Section 2.3 presents techniques for selecting features according to their information contents and describes the neural network for promoter recognition. Section 3 presents experimental results. Section 4 concludes the paper.

2 Our Approach

2.1 Characteristics of E. Coli Promoters

An E. Coli promoter is located immediately before an E. Coli gene. Thus, successfully locating the E. Coli promoter conduces to identifying the E. Coli gene. The uncertain characteristics of E. Coli promoters contribute to the difficulty in promoter recognition. Each E. Coli promoter contains two binding sites to which the E. Coli RNA polymerase, a kind of protein, binds [14]. The two binding sites are the -35 hexamer box and the -10 hexamer box, respectively. Each binding site consists of 6 bases (nucleotides). The central nucleotides of the two binding sites are roughly 35 bases and 10 bases, respectively, upstream of the transcriptional start site. The transcriptional start site is the first nucleotide of a codon where the transcription begins; it serves as a reference point (position +1). The consensus sequences, i.e., the prototype sequences composed of the most frequently occurring nucleotide at each position, for the -35 binding site and the -10 binding site are `TTGACA` and `TATAAT`, respectively. However, very few of existing E. Coli promoters exactly contain the two consensus sequences. The average conservation is about 8 nucleotides, meaning that a promoter sequence can match, on average, 8 out of the 12 nucleotides in the two consensus sequences. Figure 1 shows an example E. Coli promoter with the -35 binding site being `TAGCGA` and the -10 binding site being `AAAGAT`. The conservation here includes only 6 nucleotides.

The two binding sites are separated by a spacer. The length of the spacer has an effect on the relative orientation between the -35 region and the -10 region. A spacer of 17 nucleotides is most probable. The promoter sequence in Figure 1 has a spacer of 17 nucleotides. Another spacer between the -10 region and the transcriptional start site also has a variable length. The most probable length of this spacer is 7 nucleotides. The promoter sequence in Figure 1 has a ctttgtagcactttcacggTAGCGAaacgttagtttgaatggAAAGATgcctgCAgacacataa

-35 region -10 region +1 region

Figure 1: An example E. Coli promoter. Regions are highlighted by upper case letters. The -35 region, -10 region, and +1 region are `TAGCGA`, `AAAGAT` and `CA`, respectively.

spacer of 6 nucleotides. Notice that, in general, the distance between the -10 binding site and the transcriptional start site varies from 3 to 11 bases. The distance between the -35 binding site and the -10 binding site varies from 15 to 21 bases. These varying distances render promoter recognition difficult, as both the contents and positions of the binding sites are uncertain. Furthermore, because of the variable spacing, it is inappropriate to use orthogonal encoding directly to encode or view a promoter sequence as an n attribute tuple, where n is the length of the promoter sequence. For these reasons, we propose an expectation-maximization (EM) algorithm, to be described in the next subsection, to locate the binding sites of an E. Coli promoter.

Many E. Coli promoters have the pyrimidine (`C` or `T`) at the position -1 (one nucleotide upstream of the transcriptional start site), and the purine (`A` or `G`) at the transcriptional start site (position +1). The +1 region includes the nucleotides at the position -1 and the transcriptional start site. The E. Coli promoter in Figure 1 has a nucleotide `C` at the position -1 and a nucleotide `A` at the transcriptional start site.

In addition to these salient characteristics in the two binding sites and the transcriptional start site, there are some non-salient characteristics in other regions. Mengeritsky and Smith [17], and Galas *et al.* [8] applied pattern matching methods to the characterization of E. Coli promoters. Some weak motifs were found around the -44 and the -22 regions of a promoter sequence. A weak motif is a subsequence, which occurs frequently in a region. We use the term "weak", since the frequency of a base of the motif is not as significant as the frequency of a base of the consensus sequences occurring in the binding sites. In [5], as many as 8 nucleotides (weak motifs) within the spacer region between the two binding sites were found to have contributions to the specificity of promoter sequences. Recently, Pedersen and Engelbrecht [22] adopted a neural network to characterize E. Coli promoters. The significance of a weak motif was measured by the decrease in the maximum correlation coefficient when all motifs except that weak motif were fed into the neural network. By using this method, the authors found some weak motifs in the +1, -22, and -44 regions of an E. Coli promoter. It is interesting to observe that these weak motifs are spaced regularly with a period of 10–11 nucleotides corresponding to one helical turn. This phenomenon indicates that the RNA polymerase makes contact with the promoter on one face

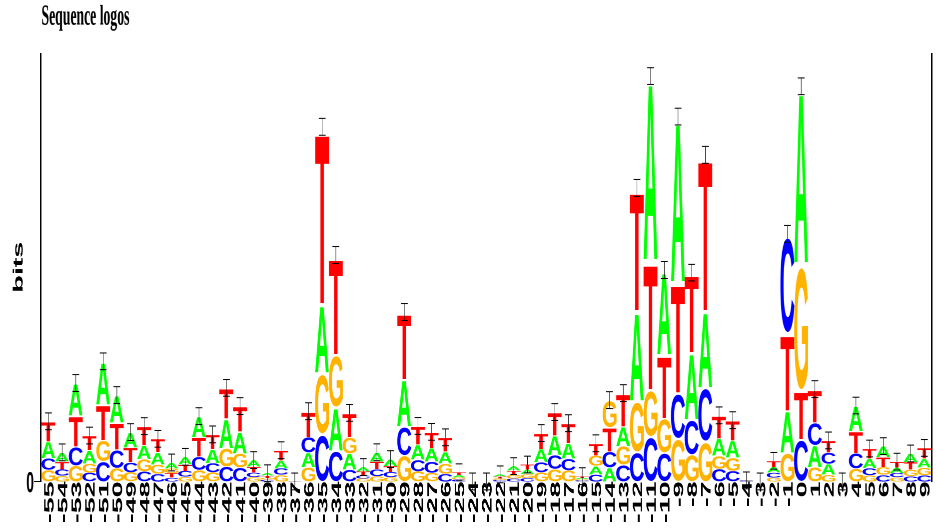

Figure 2: The sequence logos of 438 E. Coli promoter sequences. Position 0 in the figure is the transcriptional start site, which is equivalent to position +1 described in the text. The negative positions in the figure are consistent with those described in the text.

of the DNA. Subsequently, the authors carried out the characterization of E. Coli promoters by utilizing hidden Markov models [21]. It was observed that the position of the -35 binding site relative to the transcriptional start site is very flexible.

The weak motifs mentioned above are also revealed by the sequence logos originated from Schneider and Stephens [25]. Figure 2 displays the sequence logos of 438 E. Coli promoters aligned according to their transcriptional start sites, where the sequence logos were produced using the software available at `http://www-lecb.ncifcrf.gov/~toms/delila.html`. Given a set of aligned sequences, the sequence logos measure the non-randomness of each position l independently by the Shannon entropy for that position:

$$R(l) = log_2(|\mathcal{D}|) - (-\sum_{b \in \mathcal{D}} f(b,l) log_2 f(b,l)), \tag{1}$$

where $|\mathcal{D}|$ is the cardinality of the 4-letter DNA alphabet $\mathcal{D} = \{\texttt{A, T, G, C}\}$, $log_2(|\mathcal{D}|) = 2$ is the maximum uncertainty at any given position, $-\sum_{b \in \mathcal{D}} f(b,l) log_2 f(b,l)$ is the Shannon entropy of position l, and $f(b,l)$ is the frequency of base b at position l.

The height at each position represents the information content of that position. The more the information content, the less random that position is. The size of each base at each position of the logos is proportional to the frequency of that base. Recall that a weak motif is a frequently occurring subsequence in a region. In the sequence logos, a weak motif consists of positions (bases) with non-zero information content. From Figure 2, it can be seen that some weak motifs exist in the +1, -22, -29, and -44 regions.

2.2 Locating Binding Sites by an EM Algorithm

Given a collection of E. Coli promoters, to align subsequences in their -44 region, -35 region, -29 region, -22 region and -10 region, we need to first locate the two binding sites in the promoters. Locating the binding sites can be done by an EM algorithm. In general, EM algorithms are applied to the maximum likelihood estimation problem when data are incomplete. Locating the binding sites of promoter sequences using EM algorithms was pioneered by Lawrence and Reilly [13] and adopted in [2], [30]. This approach was generalized by Cardon and Stormo [5] to allow for different spacers between the binding sites. In contrast to these published EM algorithms [2], [5], [13], we propose to use a Bayesian Maximum *A Posteriori* (MAP) EM algorithm and consider the binding sites separately from their spacers. Furthermore, our method does not assume a spacer length to be uniformly distributed.

Let T represent the set of training E. Coli promoters, i.e., T contains all positive training sequences. Let K denote the cardinality of T. For a promoter sequence $S_i \in T$, the length of the spacer between the -10 region and the transcriptional start site, denoted sp_{10}, and the length of the spacer between the -35 region and the -10 region, denoted sp_{35}, are unobserved, though S_i is observed. Specifically, we refer to the positive training sequences as "observed" data since they are given. These observed data are incomplete, because the lengths of the two spacers are not given (these lengths are referred to as "unobserved" or "missing" data).

In general, sp_{10} varies from 3 to 11 and sp_{35} varies from 15 to 21. For each $S_i \in T$, the missing data sp_{10} and sp_{35} are represented by a vector $z_i = (z_{i,1}, \ldots, z_{i,63})$, where

$$z_{i,f(m,n)} = \begin{cases} 1 & \text{if } m = sp_{10}, \text{ and } n = sp_{35} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $f(m,n) = (m-3) * 7 + n - 15$. Each binding site consists of 6 bases. Assume that the nucleotides at the two binding sites of a promoter sequence are independent. Then one can use the Position Weight Matrix (PWM) described in [26] to model nucleotides at each position of the two binding sites.

Input: The positive training set T, the negative training set G,
and the test set Q of DNA sequences.
Output: The position weight matrices $\mathbf{P}_{10}$, $\mathbf{P}_{35}$,
and the putative sp_{10}, sp_{35} of each DNA sequence.

Initialize probability distributions $\mathbf{P}^0_{10}$, $\mathbf{P}^0_{35}$, and $P^0(z_{i,f(sp_{10},sp_{35})})$;
repeat /* iterate to convergence */
  begin
    /* E step */
    for each promoter sequence $S_i \in T$ **do**
      begin
        for each possible value of sp_{10}, sp_{35} **do**
          calculate $P(S_i|z_{i,f(sp_{10},sp_{35})}=1,\theta^t)$ according to Equation (9);
        for each possible value of sp_{10}, sp_{35} **do**
          calculate $P(z_{i,f(sp_{10},sp_{35})}=1|S_i,\theta^t)$ according to Equation (10);
      end;
    calculate $f_{10,j}$, $f_{35,j}$, and f_s according to Equation (13);
    /* M step */
    calculate $\mathbf{P}^{t+1}_{10}$, $\mathbf{P}^{t+1}_{35}$, and $P^{t+1}(z_{i,f(sp_{10},sp_{35})})$ according to Equation (14);
  end
until the changes of the values of $\mathbf{P}^{t+1}_{10}$, $\mathbf{P}^{t+1}_{35}$, and $P^{t+1}(z_{i,f(sp_{10},sp_{35})}) \leq$ a predefined threshold;
for each DNA sequence $S_i \in T \cup G \cup Q$ **do**
  choose the values of sp_{10}, sp_{35} that maximize $P(S_i, z_{i,f(sp_{10},sp_{35})}=1|\theta)$;

Figure 3: The proposed EM algorithm.

Let $P_{10,j}(x), j=1,\ldots,6$, denote the probability of x, $x \in \mathcal{D}$, occurring at position j in the -10 region. Let $\mathbf{P}_{10}$ denote $(P_{10,1},\ldots,P_{10,6})$. Let $P_{35,j}(x), j=1,\ldots,6$, denote the probability of x occurring at position j in the -35 region. Let $\mathbf{P_{35}}$ denote $(P_{35,1},\ldots,P_{35,6})$. Thus, $P_{10,j}$ and $P_{35,j}$, $1 \leq j \leq 6$ (from upstream nucleotides to downstream nucleotides) are in the multinomial distribution. Let θ denote the PWM model parameter ($\mathbf{P_{10}}$, $\mathbf{P_{35}}$). For each E. Coli promoter sequence, had we known the lengths of the two spacers, it would be easy to calculate the model parameter θ. The proposed EM algorithm can estimate the model parameter θ from the incomplete data. Based on the estimates of the model parameter, we are able to determine the locations of the two putative binding sites for any DNA sequence. Figure 3 shows the algorithm.

The EM algorithm proceeds iteratively to converge. Each iteration consists of two steps: an expectation step (E step) and a maximization step (M step). In general, the EM algorithm can not guarantee to reach global maxima; it may be trapped in local maxima. We use a MAP EM algorithm to make the objective function more concave [16]. The prior probabilities of $P_{10,j}$ and $P_{35,j}$, $j=1,\ldots,6$, are in the Dirichlet distribution, conjugate to the multinomial distribution, which means the posterior probabilities are also in the Dirichlet distribution [3], [24]. The Dirichlet distribution on the probability vector $P=(p(\mathtt{A}),p(\mathtt{C}),p(\mathtt{G}),p(\mathtt{T}))$ (P could be

$P_{10,j}$ or $P_{35,j}$, $j = 1, \ldots, 6$) has the form:

$$P(p(\mathtt{A}), p(\mathtt{C}), p(\mathtt{G}), p(\mathtt{T}) | \alpha_{\mathtt{A}}, \alpha_{\mathtt{C}}, \alpha_{\mathtt{G}}, \alpha_{\mathtt{T}}) = \frac{\Gamma(\alpha_0)}{\Pi_{x=\mathtt{A}}^{\mathtt{T}} \Gamma(\alpha_x)} \Pi_{x=\mathtt{A}}^{\mathtt{T}} p(x)^{\alpha_x - 1} \tag{3}$$

where $\alpha_0 = \sum_{x=\mathtt{A}}^{\mathtt{T}} \alpha_x$, $0 \leq p(x) \leq 1$, $\sum_{x=\mathtt{A}}^{\mathtt{T}} p(x) = 1$, $\alpha_x > 0$, and $\Gamma()$ is a Gamma function. The mean of $p(x)$, $x \in \mathcal{D}$, is

$$E(p(x)) = \frac{\alpha_x}{\alpha_0} \tag{4}$$

The variance of $p(x)$, $x \in \mathcal{D}$, is

$$Var(p(x)) = \frac{(\alpha_0 - \alpha_x)\alpha_x}{\alpha_0^2(\alpha_0 + 1)} \tag{5}$$

The mean values of the Dirichlet distribution on the probability vectors $P_{10,j}$ and $P_{35,j}$, $1 \leq j \leq 6$, are taken from [11]. Thus, the α_x, $x \in \mathcal{D}$, of the Dirichlet distribution can be calculated from Equation (4) given α_0 of the Dirichlet distribution, which is regarded as a parameter. (In the study presented here, we set $\alpha_0 = 20$. Our experimental results indicated that the performance of the proposed method is insensitive to the value of α_0 in general.)

The E step calculates the sum of log of the prior probability of θ, Pr_θ, and the expected complete-data log likelihood, where the expectation is over the distribution of the missing data given the observed data and current estimates of θ. Thus, the E step calculates

$$E_{Z|T,\theta^t} log P(T, Z|\theta) + log Pr_\theta. \tag{6}$$

Assume that all $S_i \in T$, $1 \leq i \leq K$, are independent, and $P(Z|\theta) = P(Z)$, i.e., the probability distribution of unobserved data is independent of θ. Then

$$\begin{aligned} & E_{Z|T,\theta^t} log P(T, Z|\theta) \\ = \; & E_{Z|T,\theta^t} log(P(T|Z,\theta)P(Z)) \\ = \; & \textstyle\sum_{i=1}^{K} \sum_{m=3}^{11} \sum_{n=15}^{21} P(z_{i,f(m,n)}=1|S_i,\theta^t) log(P(S_i|z_{i,f(m,n)}=1,\theta)P(z_{i,f(m,n)}=1)) \end{aligned} \tag{7}$$

Suppose that all promoter sequences in the positive training dataset T are 65 nucleotides long (the position 1 is now at the upstream end and the position 65 now is at the downstream end). Furthermore, these sequences are aligned with respect to their transcriptional start sites, which are at position 56. Let $S_{i,j}$ denote the nucleotide at position j of the promoter sequence S_i. Define

$$I_{i,j,x} = \begin{cases} 1 & \text{if } S_{i,j} = x \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

For each S_i, given θ^t and $z_{i,f(m,n)} = 1$, the likelihood of S_i is

$$P(S_i|z_{i,f(m,n)} = 1, \theta^t) = \Pi_{j=1}^{6} P_{10,j}^t(S_{i,49-m+j}) \Pi_{j=1}^{6} P_{35,j}^t(S_{i,43-m-n+j}) \tag{9}$$

From the Bayes' law, we have

$$\begin{aligned} & P(z_{i,f(m,n)=1}|S_i,\theta^t) \\ = & \frac{P(S_i|z_{i,f(m,n)}=1,\theta^t)P^t(z_{i,f(m,n)}=1)}{P(S_i|\theta^t)} \\ = & \frac{P(S_i|z_{i,f(m,n)}=1,\theta^t)P^t(z_{i,f(m,n)}=1)}{\sum_{m=3}^{11}\sum_{n=15}^{21}P(S_i|z_{i,f(m,n)}=1,\theta^t)P^t(z_{i,f(m,n)}=1)} \end{aligned} \tag{10}$$

Leaving out the terms not involving θ, we get log of the prior of θ, Pr_θ, as follows:

$$logPr_\theta = \sum_{j=1}^{6}\sum_{x=\mathtt{A}}^{\mathtt{T}}(\alpha_x^{10,j}-1)logP_{10,j}(x) + \sum_{j=1}^{6}\sum_{x=\mathtt{A}}^{\mathtt{T}}(\alpha_x^{35,j}-1)logP_{35,j}(x) \tag{11}$$

Substituting (9) and (10) into (7), we have

$$\begin{aligned} & E_{Z|T,\theta^t}logP(T,Z|\theta) + logPr_\theta \\ = & \sum_{j=1}^{6}(K+\alpha_0^{10,j}-4)\sum_{x=\mathtt{A}}^{\mathtt{T}} f_{10,j}(x)logP_{10,j}(x) + \sum_{j=1}^{6}(K+\alpha_0^{35,j}-4)\sum_{x=\mathtt{A}}^{\mathtt{T}} \\ & f_{35,j}(x)logP_{35,j}(x) + K\sum_{m=3}^{11}\sum_{n=15}^{21} f_s(m,n)logP(z_{i,f(m,n)}=1) \end{aligned} \tag{12}$$

where

$$\begin{aligned} f_{10,j}(x) &= \frac{1}{K+\alpha_0^{10,j}-4}(\alpha_x^{10,j}-1+\sum_{i=1}^{K}\sum_{m=3}^{11}\sum_{n=15}^{21} I_{i,49-m+j,x}P(z_{i,f(m,n)}=1|S_i,\theta^t)) \\ f_{35,j}(x) &= \frac{1}{K+\alpha_0^{35,j}-4}(\alpha_x^{35,j}-1+\sum_{i=1}^{K}\sum_{m=3}^{11}\sum_{n=15}^{21} I_{i,43-m-n+j,x}P(z_{i,f(m,n)}=1|S_i,\theta^t)) \\ f_s(m,n) &= \frac{1}{K}\sum_{i=1}^{K}\sum_{m=3}^{11}\sum_{n=15}^{21} P(z_{i,f(m,n)}=1|S_i,\theta^t) \end{aligned} \tag{13}$$

Let θ^0 denote the value of θ at the beginning of the first iteration. θ^0 was initialized to a random value so that the E step can proceed. In each iteration, we use the current estimate θ^t to calculate the sum of log of the prior probability of θ and the expected complete data log likelihood.

The M step maximizes Equation (12) with respect to θ. According to the information theory (Lemma 1.4.1 of [1]), $\sum_{x=\mathtt{A}}^{\mathtt{T}} f_{10,1}(x)logP_{10,1}(x)$ is maximized when $P_{10,1}(x)$ equals $f_{10,1}(x)$, where $f_{10,1}(x)$ is a constant. For instance, when $f_{10,1}(\mathtt{A})$, $f_{10,1}(\mathtt{C})$, $f_{10,1}(\mathtt{G})$, and $f_{10,1}(\mathtt{T})$ are 0.4, 0.3, 0.2, and 0.1, respectively, $\sum_{x=\mathtt{A}}^{\mathtt{T}} f_{10,1}(x)logP_{10,1}(x)$ can be maximized when $P_{10,1}(\mathtt{A})$, $P_{10,1}(\mathtt{C})$, $P_{10,1}(\mathtt{G})$, and $P_{10,1}(\mathtt{T})$ are 0.4, 0.3, 0.2, and 0.1, respectively. Thus, the maximum likelihood estimate of θ includes sample frequencies $f_{10,j}$, $f_{35,j}$, and f_s, $j=1,\ldots,6$. That is,

$$P_{10,j}^{t+1}(x) = f_{10,j}(x), x \in \mathcal{D}$$

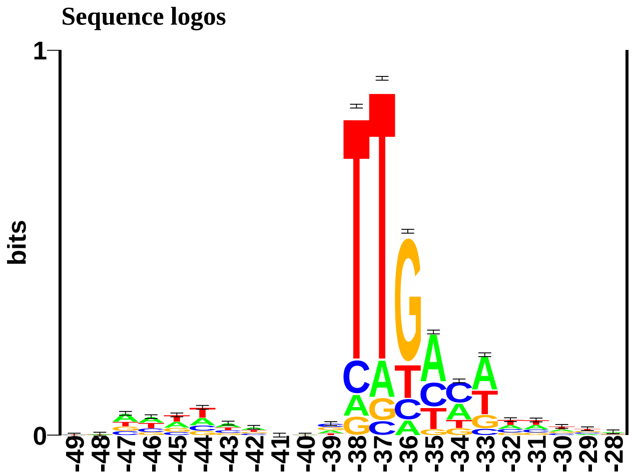

Figure 4: The sequence logos around the -35 binding site.

$$P_{35,j}^{t+1}(x) = f_{35,j}(x), x \in \mathcal{D}$$
$$P^{t+1}(z_{i,f(m,n)} = 1) = f_s(m,n) \tag{14}$$

The new value of θ can be used in the next iteration. The process iterates to convergence. Given the model parameters calculated from the positive training sequences (i.e., the promoter sequences in the training dataset T), we can determine the locations of the two putative binding sites of any DNA sequence S_i, where S_i could be a positive or negative training sequence or an unlabeled test sequence, by choosing the two spacer lengths sp_{10} and sp_{35} that are calculated by $\max_{3\leq m\leq 11, 15\leq n\leq 21}\{P(S_i, z_{i,f(m,n)} = 1|\theta)\}$.

2.3 Feature Extraction

After locating the two binding sites of each training promoter sequence in T, we can align all the training promoter sequences with respect to their binding sites and transcriptional start sites. Figures 4, 5 and 6 show the sequence logos of regions around the -35 binding site, the -10 binding site and the transcriptional start site, respectively, for the same 438 E. Coli promoters used to produce Figure 2. Here, the promoters are aligned with respect to their binding sites found by the proposed EM algorithm. Compared to Figure 2, it is easier to observe consensus sequences from Figures 4, 5 and 6, indicating that the proposed EM algorithm can precisely locate the

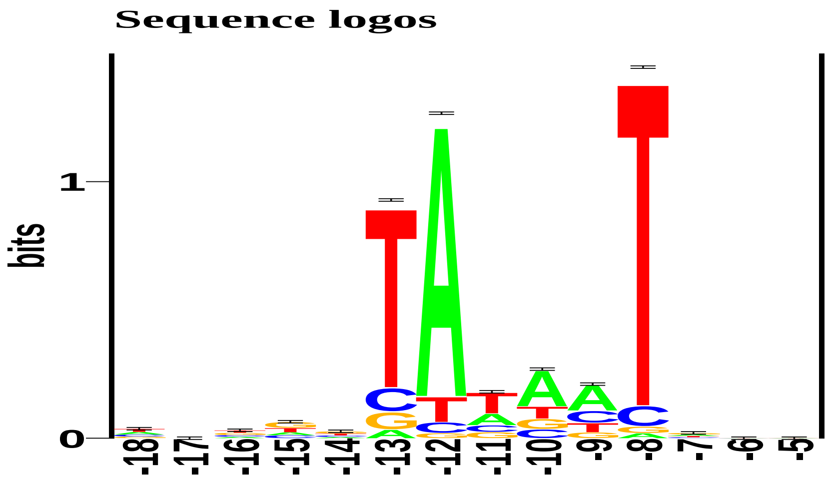

Figure 5: The sequence logos around the -10 binding site.

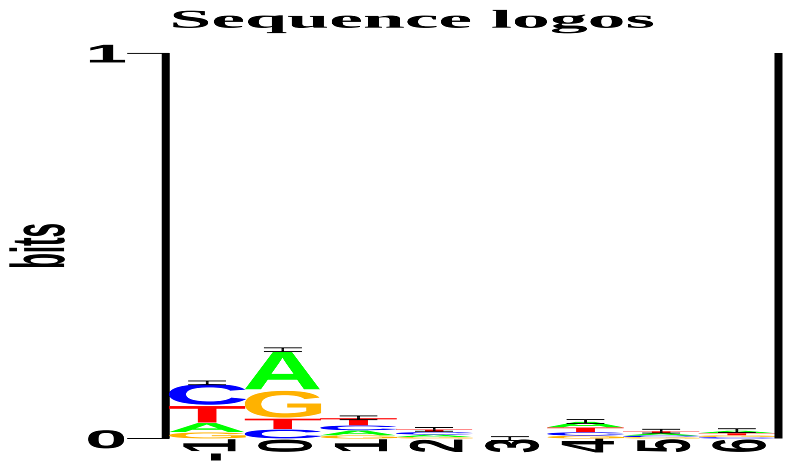

Figure 6: The sequence logos around the transcriptional start site.

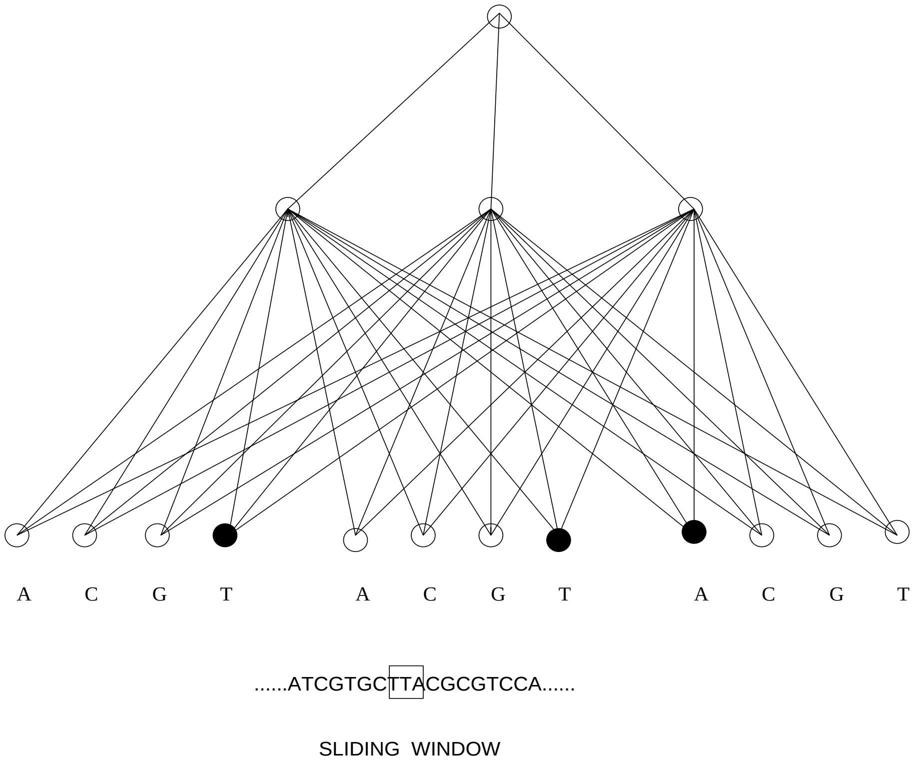

Figure 7: An example of the orthogonal encoding of a DNA sequence.

binding sites for each E. Coli promoter.

Referring to Figures 4, 5 and 6, we select those positions with high information contents as features. Specifically, we choose 17 positions around the -35 binding site, 11 positions around the -10 binding site, and 7 positions around the transcriptional start site, and use them as features. These 35 nucleotides are encoded by an orthogonal encoding method.

In orthogonal encoding, nucleotides in a DNA sequence are viewed as unordered categorical values, which are represented by 4-dimensional orthogonal binary vectors. The number 4 here is the cardinality of the 4-letter DNA alphabet $\mathcal{D}$. That is, we use 4 binary (0/1) variables, among which only one binary variable is set to 1 to represent one of the 4 possible categorical values and the rest are all set to 0. For instance, we represent the nucleotide `A` by "1000". Figure 7 shows an example of the orthogonal encoding method. When there is an uncertain nucleotide, denoted by '`X`', we use "1111" to represent it. Besides these 35 nucleotides, the two spacer lengths are also chosen as features.

We feed these features to the Matlab neural network toolbox version 3.0 [4], [10], [32], [33], run on a Sun workstation with the operating system Solaris version 2.6. The neural network has one hidden layer with sigmoid activation functions. The output layer has one output unit; the output value is bounded between 0 and 1. The neural network is fully connected and trained with a scaled conjugate gradient algorithm [4]. We tested the neural network with different numbers of hidden units and found that the system was most effective with 20 hidden units, which were used in our experimental study.

3 Experimental Results

We carried out a series of experiments to evaluate the performance of our approach. Three measures were used: precision, specificity, and sensitivity. Precision is defined as

$$\frac{C}{N} * 100\% \tag{15}$$

where C is the number of test sequences classified correctly and N is the total number of test sequences. Specificity is defined as

$$(1 - \frac{N_{fp}}{N_{ng}}) * 100\% \tag{16}$$

where N_{fp} is the number of false positives and N_{ng} is the total number of negative test sequences. A false positive is a non-promoter test sequence that was misclassified as a promoter sequence. Sensitivity is defined as

$$\frac{N_{tp}}{N_{po}} * 100\% \tag{17}$$

where N_{tp} is the number of true positives and N_{po} is the total number of positive test sequences. A true positive is a promoter test sequence that was also classified as a promoter sequence.

In the first experiment we compared our approach with the closely related method developed by Mahadevan and Ghosh [15]. As far as we know, this method is the best one published in the literature with clearly documented datasets and software. The positive training dataset included 362 promoter sequences. The negative training dataset contained 4,500 random sequences with 60% `AT` composition, i.e., the sum of probabilities of `A` and `T` was 0.6. The test dataset included 126 promoter sequences and 5,000 random sequences with 60% `AT` composition. These data are the same as those used in [15]. Table 1 shows the result. Clearly, our approach outperforms the previously published method.

In the next experiment, we used the E. Coli promoter sequences taken from the latest E. Coli promoter compilation [19]. No previous work performed promoter recognition on this dataset. There were 441 E. Coli promoters aligned with respect to their transcriptional start sites. We trimmed each promoter sequence to get a sequence of 65 nucleotides including nucleotides from the -55 position (55 nucleotides upstream of the transcriptional start site) to the +10 position (10 nucleotides downstream of the transcriptional start site). This yielded 438 promoter sequences.

	Our approach	Mahadevan and Ghosh
Precision	91.94%	90.40%
Specificity	91.76%	90.20%
Sensitivity	99.20%	98.00%

Table 1: Comparison of our approach with Mahadevan and Ghosh.

As in [15], the negative data (i.e., non-promoter sequences) was randomly generated with 60% `AT` composition. Each negative sequence was also 65 nucleotides long. There were 5,000 negative sequences in total.

We used ten-fold cross validation to evaluate the performance of our approach. The dataset containing both the positive data (promoters) and the negative data (non-promoters) was randomly split into ten mutually exclusive folds $D_1, D_2, \ldots, D_{10}$ of approximately equal size. The neural network was trained and tested ten times. During the ith time, it was trained on $D - D_i$, and tested on D_i. We allocated the data in such a way that the training dataset $D - D_i$ (the test dataset D_i, respectively) has approximately $\frac{9}{10}$ ($\frac{1}{10}$, respectively) positive data and $\frac{9}{10}$ ($\frac{1}{10}$, respectively) negative data. The average over the ten tests was calculated.

Experimental results indicated that the proposed approach performs well on the dataset, with precision 96.29%, specificity 96.68% and sensitivity 91.78%. This happens mainly due to the fact that our EM algorithm is able to precisely locate the binding sites of the promoter sequences.

4 Conclusion

In this paper, we presented new techniques for recognizing E. Coli promoters in DNA. We first used a Bayesian MAP EM algorithm to locate the binding sites of the promoter sequences. We then aligned the promoters with respect to their binding sites and transcriptional start sites. This alignment helps to identify features in the sequences. Next, we extracted the features according to their information contents. These features were then represented by an orthogonal encoding method and fed to a neural network. Our experimental results indicated that the proposed approach achieves good performance. This happens mainly because our EM algorithm is able to precisely locate the binding sites of the promoter sequences. The program developed from this research can be obtained from the authors. Currently we are integrating the techniques presented here into a web-based genome mining system for DNA and protein sequence classification [29].

Acknowledgments

The sequence logos software was developed by Dr. Thomas Schneider. We thank Dr. O. N. Ozoline for providing the E. Coli promoter sequences used in the paper.

References

[1] R. Ash. *Information Theory.* Interscience Publishers, New York, 1965.

[2] T. L. Bailey and C. P. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21**, 51–83, 1995.

[3] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, New York, 1985.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, New York, 1995.

[5] L. R. Cardon and G. D. Stormo. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *Journal of Molecular Biology* **223**(1), 159–170, 1992.

[6] E. M. Crowley, K. Roeder, and M. Bina. A statistical model for locating regulatory regions in genomic DNA. *Journal of Molecular Biology* **268**(1), 8–14, 1997.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, **39**, 1–38, 1977.

[8] D. J. Galas, M. Eggert, and M. S. Waterman. Rigorous pattern-recognition methods for DNA sequences: Analysis of promoter sequences from E. Coli. *Journal of Molecular Biology* **186**(1), 117–128, 1985.

[9] J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 2000.

[10] D. C. Hanselman. Mastering MATLAB 5: A comprehensive tutorial and reference. Prentice Hall, Upper Saddle River, NJ, 1998.

[11] C. B. Harley and R. P. Reynolds. Analysis of E. Coli promoter sequences. *Nucleic Acids Research* **15**(5), 2343–2361, 1987.

[12] S. Knudsen. Promoter2.0: For the recognition of PolII promoter sequences. *Bioinformatics* **15**(5), 356–361, 1999.

[13] C. E. Lawrence and A. A. Reilly. 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, 1990.

[14] S. Lisser and H. Margalit. Compilation of E. Coli mRNA promoter sequences. *Nucleic Acids Research* **21**(7), 1507–1516, 1993.

[15] I. Mahadevan and I. Ghosh. Analysis of E. Coli promoter structures using neural networks. *Nucleic Acids Research* **22**(11), 2158–2165, 1994.

[16] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley, New York, New York, 1997.

[17] G. Mengeritsky and T. F. Smith. Recognition of characteristic patterns in sets of functionally equivalent DNA sequences. *Computer Applications in the Biosciences* **3**(3), 223–227, 1987.

[18] D. W. Opitz and J. W. Shavlik. Connectionist theory refinement: Genetically searching the space of network topologies. *Journal of Artificial Intelligence Research* **6**, 177–209, 1997.

[19] O. N. Ozoline, A. A. Deev, and M. V. Arkhipova. Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by E. Coli RNA polymerase. *Nucleic Acids Research* **25**(23), 4703–4709, 1997.

[20] O. N. Ozoline, A. A. Deev, and E. N. Trifonov. DNA bendability – A novel feature in E. Coli promoter recognition. *Journal of Biomolecular Structure and Dynamics* **16**(4), 825–831, 1999.

[21] A. G. Pedersen, P. Baldi, S. Brunak, and Y. Chauvin. Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 1996, pp. 182–191.

[22] A. G. Pedersen and J. Engelbrecht. Investigations of E. Coli promoter sequences with artificial neural networks: New signals discovered upstream of the transcriptional start point. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 1995, pp. 292–299.

[23] L. Pickert, I. Reuter, F. Klawonn, and E. Wingender. Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics* **14**(3), 244–251, 1998.

[24] T. J. Santner. *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York, New York, 1989.

[25] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research* **18**(20), 6097–6100, 1990.

[26] R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research* **12**(1), 505–519, 1984.

[27] G. G. Towell and J. W. Shavlik. Knowledge-based artificial neural networks. *Artificial Intelligence* **70**, 119–165, 1994.

[28] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu. New techniques for extracting features from protein sequences. *IBM Systems Journal* (Special Issue on Deep Computing for the Life Sciences) **40**(2), 426–441, 2001.

[29] J. T. L. Wang, T. G. Marr, D. Shasha, B. A. Shapiro, G. Chirn, and T. Y. Lee. Complementary classification approaches for protein sequences. *Protein Engineering* **9**(5), 381–386, 1996.

[30] J. T. L. Wang, B. A. Shapiro, and D. Shasha (eds.). *Pattern Discovery in Biomolecular Data: Tools, Techniques and Applications*. Oxford University Press, New York, New York, 1999.

[31] X. Wang, J. T. L. Wang, D. Shasha, B. A. Shapiro, I. Rigoutsos, and K. Zhang. Finding patterns in three dimensional graphs: Algorithms and applications to scientific data mining. *IEEE Transactions on Knowledge and Data Engineering*, in press.

[32] C. H. Wu. Artificial neural networks for molecular sequence analysis. *Computers and Chemistry* **21**(4), 237–256, 1997.

[33] C. H. Wu and J. McLarty. *Neural Networks and Genome Informatics*. Elsevier Science, New York, 2000.