Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Promoter recognition based on the maximum entropy hidden Markov model ☆

Xiao-yu Zhao, Jin Zhang, Yuan-yuan Chen, Qiang Li, Tao Yang, Cong Pian, Liang-yun Zhang *

College of Science, Nanjing Agricultural University, China

ARTICLE INFO

Article history:
Received 16 September 2013
Accepted 9 April 2014

Keywords:
Promoter
MEMM
ME-HMM
Forward algorithm
IIS algorithm

ABSTRACT

Since the fast development of genome sequencing has produced large scale data, the current work uses the bioinformatics methods to recognize different gene regions, such as exon, intron and promoter, which play an important role in gene regulations. In this paper, we introduce a new method based on the maximum entropy Markov model (MEMM) to recognize the promoter, which utilizes the biological features of the promoter for the condition. However, it leads to a high false positive rate (FPR). In order to reduce the FPR, we provide another new method based on the maximum entropy hidden Markov model (ME-HMM) without the independence assumption, which could also accommodate the biological features effectively. To demonstrate the precision, the new methods are implemented by R language and the hidden Markov model (HMM) is introduced for comparison. The experimental results show that the new methods may not only overcome the shortcomings of HMM, but also have their own advantages. The results indicate that, MEMM is excellent for identifying the conserved signals, and ME-HMM can demonstrably improve the true positive rate.

1. Introduction

With the completion of human genome sequencing, one of the challenges in bioinformatics is to reconstruct gene regulatory networks. As a foundation of transcription regulation research, promoter recognition developed rapidly because of its low cost, wide application, reliable results and high efficiency.

At present, there are two kinds of promoter recognition method: the content-based method [1] and the signal-based method [2]. The content-based method is a data mining method according to different base preferences on DNA, which is suitable for the unknown sequences. For example, the hidden Markov model (HMM) [3,4] introduced by Pedersen and colleagues [5] is a content-based method. This method has been applied with success to recognize the promoter sequences. Its main idea is to mine the internal statistics rules of sequence by the ordinary discrete-time finite Markov chain. Content-based methods often bring some signal noise and ignore the biological features, which play an important role in promoter recognition. Signal-based methods overcome the shortages above and extract features from the local consensus regions and binding sites, such as TATA box, initiator regions, upstream activating elements, and downstream promoter elements. It is a powerful tool for dealing with promoters and suitable for the sequences whose biological features are given. Although the signal-based method has its own advantages, it also has some limitations. The signal-based method may lead to a high false positive rate (FPR). Utilizing the biological features effectively and at the same time reducing the FPR have become the challenge in promoter recognition research.

In order to overcome the above shortages and improve the recognition performance constantly, we introduce a new signal-based method to recognize promoters based on maximum entropy Markov model (MEMM), which was introduced by McCallum and colleagues [6] in 2000. The new method allows for limited sequences, which are difficult to enumerate, and can effectively recognize the promoters by biological features.

Furthermore, in order to reduce the FPR of signal-based methods, another new method is introduced based on the maximum entropy hidden Markov model (ME-HMM), which is not only the signal-based but also the content-based method. ME-HMM is a model developed from HMM and MEMM by Lin and colleagues [7] in 2005 in the field of text information extraction, and utilized by Sangwan and Hansen [8] in the field of speech recognition in 2009. In the next section, we will elaborate the details of the new method.

Both of the two methods above are applied in the field of promoter recognition for the first time. We implement the new

☆ This work is supported by the Research Fund for the Doctoral Program of Higher Education of China (20100097110040) and the Fundamental Research Funds for the Central Universities (KYZ201125).

* Corresponding author.
E-mail address: zlyun@njau.edu.cn (L.-y. Zhang).

http://dx.doi.org/10.1016/j.compbiomed.2014.04.003
0010-4825/© 2014 Elsevier Ltd.. All rights reserved.

methods with R language and take the HMM method as a comparison. The experimental results indicate that the sensitivity and specificity of ME-HMM methods are both up to 100%, and the performance of these new methods is much better than that of the popular content-based HMM method.

2. Methods

This section includes two parts. In the first part, we introduce the general definition of MEMM, and present the particular procedure of promoter recognition based on the MEMM method. In the second, we give a summary definition of ME-HMM, and provide the promoter recognition based on the ME-HMM method.

2.1. MEMM

The MEMM combines the advantages of the maximum entropy model (MEM) and Markov Model, using a combination of conditional probability $P_{y_i}(y_j|x_j)$ instead of the maximum entropy condition probability $P(y_j|x_j)$ in MEM and state transition probability $a_{ij} = P(y_j|y_i)$ in the Markov model:

$$P_{y_i}(y_j|x_j) = P(y_j|y_i, x_j) = P(q_t = y_j|q_{t-1} = y_i, o_t = x_j) \tag{1}$$

The joint conditional probability $P_{y_i}(y_j|x_j)$ means that the current state is y_j when its previous state is y_i and the current observation is x_j. Every $P_{y_i}(y_j|x_j)$ is an exponential model satisfying Markov properties.

2.1.1. Promoter recognition based on MEMM

First, we need to establish the model of DNA. If the length of a DNA sequence is N, the model can be built as follows:

1. Observation state value is

   $x_i \in X = A, T, G, C;\quad N_X = 4,\quad i = 1, 2, \cdots, N_x$

2. Time sequence of observation state is

   $O = o_1, o_2, \cdots, o_T,\qquad \text{in which } o_t = x_i,\quad t = 1, 2, \ldots, T;$

3. Implicit state set is

   $y_i \in Y = Start, m_1, m_2, \ldots, m_N, i_0, i_1, \ldots, i_N, d_1, d_2, \ldots, d_N,$
   $End,\quad N_Y = 3N+3,\quad i = 1, 2, \cdots, N_Y,$

   which includes one start state, N main states (m), $N+1$ insert states (i), N delete states (d) and one end state; N_Y is the total length of state Y.
4. Time sequence of implicit state is $Q = q_1 q_2 \ldots q_T$, in which $q_t = y_i$. Sequence Q can be shown by the state transition diagram in Fig. 1.

   The chart shows how the state transits from the current state to the next state, assuming that the current state is the start state and the next state should be the main state. If no match, the main state may be removed and the next state is the delete state. Otherwise, an insert state may exist between the start state and the main state.

   Second, we suppose that for all state values $y_i \in Y$, an observation value exists $x_i \in X$ in a period of observation sequence $O_i \in O$, which is named the feature sequence.

Then we can define the feature function as follows:

$$f_i(x, y) = \begin{cases} 1, & x = O_i y = y_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

which means that if the observation sequence is a feature sequence Q_i and the state value is y_i, then the value of the feature function is 1. Otherwise, the value is 0.

Last but not least, we bring in a series of feature functions $f_a, a = 1, 2, \ldots, k$, in which a is the number of feature functions, and k the number of features and then import the weight parameters λ_a for each f_a.

We can then recognize the promoter based on the MEMM as follows:

Step 1. *Feature selection*: First, analyse the signal features which are commonly used; second, establish feature templates based on the feature function; finally, select features based on the frequency selection method.

Step 2. *Parameter training*: Estimate parameters $P_{y_i}(y_j|x_j)$ and λ_a by the generalized iterative scaling (GIS) algorithm.

Step 3. *Threshold training*: Calculate the scores of the training datasets using the GISmemm.R program, which implements the GIS algorithm and the MEMM-forward algorithm together. Then compute the threshold from the box plot of scores $P(O|\lambda)$.

Step 4. *Recognition*: Calculate the scores of the test datasets by the MEMM-forward algorithm, and compare it with the threshold to identify the promoters.

Above all, the flowchart of promoter recognition based on MEMM is shown in Fig. 2.

The diamond in Fig. 2 represents the logic judge module. If the condition $P(O|\lambda) > threshold$ is satisfied, the sequence can be recognized as a promoter, otherwise not.

2.1.2. Feature selection

It is important to extract the correct features to evaluate the particular datasets, because the selection of features directly influences the performance of the MEMM methods. However, it is a hard work to choose an available feature. In this section, we introduce an easier way to select suitable features for the dataset.

First, we need to know that promoters contain various combinations of binding sites, but only parts of the transcription factor binding sites are useful for modelling. In order to reduce the time complexity of feature extraction, we select three important transcription factor binding sites: TATA box, GC box, and CAAT box, which commonly exist in eukaryotic promoters.

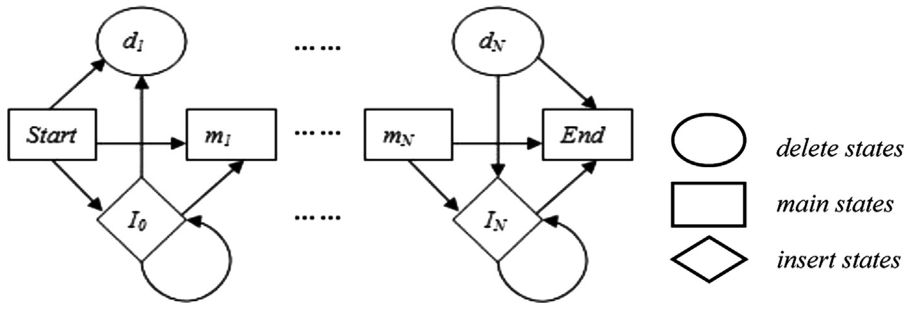

Fig. 1. State transition diagram of implicit state sequence.

Please cite this article as: X.-y. Zhao, et al., Promoter recognition based on the maximum entropy hidden Markov model, Comput. Biol. Med. (2014), http://dx.doi.org/10.1016/j.compbiomed.2014.04.003