Chemometrics and Intelligent Laboratory Systems 158 (2016) 69–79

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Prediction of human promoter with Least Square Support Vector Machine based on Kernel Locality Preserving Projection

Shuo Guo [a,*], Decheng Yuan [a], Ridong Zhang [b,*], Furong Gao [b]

[a] College of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China
[b] Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 20 April 2016
Received in revised form
22 August 2016
Accepted 26 August 2016
Available online 30 August 2016

Keywords:
Promoter
Kernel
Kernel Locality Preserving Projection
Gaussian Mixture Model
Fuzzy cluster
Least Square Support Vector Machine

ABSTRACT

The gene promoter region controls the transcription of a gene, so finding the gene promoter region is the most important step in gene regulation. Due to the huge amount and genetic diversity, although many algorithms have been proposed, the promoter recognition are rather complex with the performance still limited by low sensitivity and highly false positives. In this paper, we present a novel machine learning method for predicting promoter. First, the function motifs in different regions of Human promoter sequences have been recognized using Gaussian Mixture Model (GMM). The optimum number of GMM is given by the fuzzy cluster recognition algorithm based on fuzzy likelihood function without prior knowledge. Then the promoter sequences were mapped into the positional densities of oligonucleotides high dimension Bayes space. At last, Least Square Support Vector Machine classifier is built with Kernel Locality Preserving Projection to predict the promoter sequence, which simplifies the Least Square Support Vector Machine to form the Least Square model. Simulation results show that the performance is improved compared with other promoter classifiers and the proposed method can predict the unknown promoters with unknown similar genes in the database, and also the speed of the proposed method is significantly increased.

1. Introduction

Reconstruction of gene regulatory network is a research topic in bioinformatics study. With the development of DNA sequencing technology, DNA sequences can be gained and the gene region can be located easily. However, the promoter prediction is a difficult step in the reconstruction process. As the experimental approaches for finding DNA promoter are expensive and laborious, computational identification of promoter as the research foundation of transcriptional regulation is developing rapidly due to its high efficiency, wide applications, reliable results and low cost. For a gene sequence, promoter prediction services generally provide numerous possible promoter locations, which will causes more analysis for researchers to judge which one is the true promoter. Therefore, to develop a promoter prediction algorithm is meaningful for researchers. However, the complex structure and degeneracy of eukaryotic promoter impose a great challenge on the prediction of promoter in molecular biological study.

In research algorithms for modeling and prediction, many kinds of feature may be taken as the inputs of classification to improve the algorithms' accuracy [1–3]. Molecular structures, topology structural index, geometrical configuration, and quantum-chemical descriptors, etc., were taken as the inputs of system models [4,5]. The DNA duplex stability calculated by the nearest neighbors was taken into the repertoire of a neural network (NN) [6]. Local word content, CpG island (GC)-Skew and DNA geometric were taken as the inputs of support vector machine (SVM) to predict the two types of promoters [7].The feature selected by FS process with C4.5 decision tree rules were the inputs of the E. Coli promoter Fuzzy-AIRS classifier [8]. The pseudo-trinucleotide compositions extracted by discrete wavelets transform were the inputs of SVMs for the prediction of promoters [9]. The biological features and the maximum entropy markov model (MEMM) were used to recognize the promoter [10]. However, too many DNA features taken into the inputs of the promoter classifier may cause important information to submerge into large data and affect the identification accuracy.

The binding sites of promoter accomplished with transcription factors regulate the metabolism and transcription of DNA, or combined with RNA binding protein, influence the modification, localization, translation and degradation of RNA. Most regulatory elements interacted with transcription factors are unknown, including the compositions and the occurrence positions of the bases. Different occurrence positions of the binding sites will have

* Corresponding authors.
E-mail addresses: guo_shuo@163.com (S. Guo), kerzhang@ust.hk (R. Zhang).

http://dx.doi.org/10.1016/j.chemolab.2016.08.013
0169-7439/

different biological function significations, and the binding sites may interact with each other. The combination of upstream and downstream binding sites may also affect the transcription. Obtaining these kinds of knowledge is useful for the gene regulation mechanism research and promoter recognition. The occurrence position of binding site and the combination are very important features of promoter sequence [11,12]. Adjacent and non-adjacent positions' motifs all have very important correlations [13–15]. So analysis and identification of the transcription fact binding site is an important step for understanding and explaining the behavior of the entire genome.

A tagged mismatch string kernel used to code DNA sequences and SVM classifiers was trained on E. coli promoters at -35 db $\sim$ -10 db region to predict σ^A promoters in *B. subtilis* and σ^{66} promoters in Chlamydia trachomatis [16]. As the structure of eukaryotic promoter is more complex than that of procaryotic promoter, many useful information of upstream and downstream around transcription start site (TSS) may be omitted when modeling eukaryotic promoter at -35 db $\sim$ -10 db and uses the mismatch tree to find subsequence patterns that occur with mismatches. So the important motifs but less frequent cannot be identified. Paper [17] estimated collections of non-TSS locations (NTLs), extracted the statistical features around TSS, and distinguished genomic transcription initiation locations from those that are not likely to initiate transcription. Most algorithms have taken the motif PFM (Positional Frequency Matrix) as the DNA sequence features. This may take low occurrence with important motifs abandoned.

Due to the above problems, the positional densities of oligonucleotides (short DNA sequence, motif) in the promoter sequence are taken as the features and mapped into the Bayes space. Kernel Locality Preserving Projection (KLPP) with Gaussian Mixture Model (GMM) was used as the kernel of human promoter Least Square Support Vector Machines (LS-SVM) classifier, which simplifies the LS-SVM with Least square (LS). The algorithm can extract the promoter feature effectively and gain high accuracy.

2. The positional densities of oligonucleotides

The polymerizes with fewer bases are known as oligonucleotide (TATAAA, 6-length oligonucleotide). Some oligonucleotides at fixed positions are responsible for regulation and transcription [18]. The positional densities of oligonucleotides measuring the probability of oligonucleotides occurrence at various positions relative to the TSS within promoter sequences were measured by the positional densities, which is independent to the occurrence number of the subsequences.

There are two different statistical models: a promoter model, π and a non-promoter model, $\bar{\pi}$. The statistical models measure for each oligonucleotide, K_i, and its positional density, $f_i(p|\pi)$ and $f_i(p|\bar{\pi})$.

The positional density of oligonucleotides, K_i relative to the TSS in the promoter is approximated as a finite mixture of Gaussians [19]

$$f_i(p|G_i, \theta_i, \pi) = \sum_{s_i=1}^{G_i} \alpha_{s_i} \phi(p|\mu_{s_i}, \sigma_{s_i}^2)$$
$$\alpha_{s_i} \geq 0 \text{ and } \sum_{s_i=1}^{G_i} \alpha_{s_i} = 1 \tag{1}$$

where p is the random variable representing the position of occurrence K_i relative to the TSS. G_i is the optimal numbers of GMM; $\phi(p|\mu_{s_i}, \sigma_{s_i}^2)$ is a Gaussian distribution with parameters mean μ_{s_i}, variance $\sigma_{s_i}^2$, α_{s_i} are the mixing proportions, and $\theta = \{\alpha_{s_i}, \mu_{s_i}, \sigma_{s_i} | s_i = 1, 2, \ldots, G_i\}$ is the set of all model parameters.

The non-promoter model, on the other hand, is defined as

$$f_i(p|G_i', \theta_i', \bar{\pi}) = \sum_{s'_i=1}^{G_i'} \alpha'_{s'_i} \phi(p|\mu'_{s'_i}, \sigma'^2_{s'_i})$$
$$\alpha'_{s'_i} \geq 0 \text{ and } \sum_{s_i=1}^{G_i} \alpha_{s_i'} = 1 \tag{2}$$

The probability density functions of promoter and non-promoter are [19]

$$\Pr(p_1 < P_i < p_2|\pi) = \int_{p_1}^{p_2} f_i(p|\pi)dp \tag{3}$$

$$\Pr(p_1 < P_i < p_2|\bar{\pi}) = \int_{p_1}^{p_2} f_i(p|\bar{\pi})dp \tag{4}$$

where $_{Pi}$ is the random variable representing the position of occurrence of K_i relative to the TSS.

3. Method descriptions

3.1. The positional densities of oligonucleotides GMM modeling algorithm

At present, the optimum number of GMM is the key problem that needs solution. The sequences of the occurrence positions relative to TSS of oligonucleotide are clustered by the fuzzy recognition algorithm based on fuzzy likelihood function [20] without prior knowledge. The optimum number of components G_i and all of Gaussian distribution parameters: mean μ_{s_i}, variance $\sigma_{s_i}^2$ are obtained. The mixing proportions α_{s_i} are estimated by LS. The fuzzy recognition algorithm avoids the EM (Expectation Maximization) shortcoming of converging to some local maximum and the results' high dependence on the initial parameter values chosen by the EM algorithm. The algorithm is simple with structure identification and parameter identification accomplished simultaneously. The precision of modeling is improved and the computational time is reduced.

The observations of each oligonucleotide are obtained as, K_i, ~$p_i = [p_i^1, p_i^2, \ldots, p_i^{N_i}]$, $i=1, 2,\ldots, 4^k$, where p_i^j is the position of the jth occurrence of K_i relative to the TSS, and N_i is the total number of occurrences of K_i in all training sequences.

The first data sample from the oligonucleotide, K_i observation set is taken as the first cluster center. When the mth data sample (p_i^m), $m = 2, 3, \ldots, N_i$ is considered, and suppose M clusters exist, the distance between the mth data sample is

$$H_l = \frac{1}{S_h(A^l)} = \frac{1}{\mu_{A^l}(p_i^m)} \quad l = 1, 2, \ldots, M \tag{5}$$

Here $H_{\min}^m$ is the minimum distance of these distances. If $H_{\min}^m > R$, where R is clustering radius, then the mth data sample is added to a new cluster center. A^l is the fuzzy subset of input domain. The membership function is $\mu_{A^l}(p_i^m) = \exp(-(p_i^m - \rho_l)^2/\sigma_l^2)$, and ρ_l, σ_l are mean and variance. At last, LL clusters will be obtained. So the optimal numbers of GMM components for oligonucleotides K_i is LL. The mean μ_{s_i}, variance $\sigma_{s_i}^2$ of the cluster center set are calculated by data samples that belong to fuzzy subset A^{si}.

The probability position density functions of oligonucleotides in Eq. (2) can be rewritten into the vector form:

$$f_i(p|\pi) = \boldsymbol{\alpha}\boldsymbol{\phi}^{\mathrm{T}}(p) \tag{6}$$

where $\boldsymbol{\phi}(p) = \left[\phi(p|\mu_1, \sigma_1^2)\phi(p|\mu_2, \sigma_2^2 \cdots \phi(p|\mu_{G_i}, \sigma_{G_i}^2))\right]$, $\boldsymbol{\alpha} = \left[\alpha_1\alpha_2\cdots\alpha_{G_i}\right]$.