doi:10.3969/j.issn.1672-5565.2015.02.09

利用位点特异性打分矩阵对大肠杆菌启动子的预测

闫 妍,万 平*
(首都师范大学生命科学学院,北京 100048)

摘 要:启动子是基因转录起始的一个关键性元件。本研究利用数据库中提供的大肠杆菌启动子数据,基于位点特异性打分矩阵(Position-specific scoring matrix,PSSM)算法建立了大肠杆菌启动子预测方法,并采用ROC曲线对预测结果进行评估。结果显示,本方法对大肠杆菌sigma24、sigma28、sigma32、sigma38、sigma54和sigma70启动子预测的准确度分别达到86%,96%,93%,96%,97%和74%。由于原核生物启动子序列的保守性,可将该方法推广至其他原核生物的启动子预测。

关键词:大肠杆菌;启动子;位点特异性打分矩阵(PSSM);预测
中图分类号:Q811.4 **文献标志码**:A **文章编号**:1672-5565(2015)-02-125-06

Prediction of *Escherichia coli* K-12 promoters using position-specific scoring matrix (PSSM) method

YAN Yan, WAN Ping*
(*College of Life Science*, *Capital Normal University*, *Beijing* 100048, *China*)

Abstract: Promoter is an essential element in transcription initiation. In this study, we proposed a method for the promoter prediction based on the position-specific scoring matrix (PSSM) constructed with the data from RegulonDB database, and evaluated the performance through the receiver operating characteristic (ROC). We predicted the *Escherichia coli* K-12 promoters, the accuracies of predictions for sigma24, sigma28, sigma32, sigma38, sigma54 and sigma70 are 86%, 96%, 93%, 96%, 97% and 74%, respectively. Since promoter sequences are conserved among prokaryotes, PSSM could be applied to the prediction of prokaryotic promoters.

Keywords: *E.coli*; Promoter; Position-specific Scoring matrix; Prediction

启动子是基因转录起始的一个关键性元件,位于基因转录起始点附近。在细菌中,启动子由RNA聚合酶核心酶与相应的sigma因子共同识别[1]。因子共有7种类型:sigma19、sigma24、sigma28、sigma32、sigma38、sigma54和sigma70,每种sigma因子所识别的序列都具有一定特征。除sigma54启动子外,启动子在转录起始位点上游-10和-35位附近都存在保守区域[2];而sigma54启动子的保守区域位于转录起始位点上游-12和-24位附近[3]。

对于特征结构域建模的算法有很多。例如,常用的有位点特异性打分矩阵(Position-specific scoring matrix,PSSM,也称PWM)、贪婪算法、EM算法和MCMC算法,这些算法都有各自的优缺点[4]。此外,近年内也报导了一些新型算法,如pHMM-ANN方法[5]、GLECLUBS算法[6]、BOBRO算法[7-8]、神经网络算法[9]、构建非传统的16列双核苷酸矩阵的PSSM算法[10]。

在众多算法中,PSSM仍然是最常用的算法,占据重要的地位。PSSM在发现例如启动元件或可变剪接等具有信号核酸序列方面有着广泛的应用[11]。有很多构建PSSM的方法,最常用的就是使用排列好且长度相等的具有已知类似功能的结构域构建打分矩阵。这个打分矩阵的行数由结构域中的元素种类决定,列数则由结构域的元素个数决定。构建好的打分矩阵能够搜索DNA序列或蛋白序列中的与已知序列相似的序列[10]。

收稿日期:2015-01-06;**修回日期**:2015-03-18.
作者简介:闫妍,女,硕士研究生,研究方向:生物信息学;E-mail:yanyan0108@163.com.
*通信作者:万平,男,博士,研究方向:生物信息学;E-mail:wanp_cnu@163.com.

ROC 曲线(Receiver operating characteristic curve)是一种坐标图式的分析工具,能描绘诊断中敏感性和特异性之间的制约关系[12]。

目前还未见采用 PSSM 方法预测原核生物启动子的报道。本研究采用 PSSM 方法预测大肠杆菌启动子,并且通过 ROC 曲线评估预测结果。

1 数据和方法

1.1 大肠杆菌 K-12 启动子核酸序列

大肠杆菌 K-12 sigma24、sigma28、sigma32、sigma38、sigma54 和 sigma70 启动子的核酸序列下载自 RegulonDB 数据库(http://regulondb.ccg.unam.mx/)。RegulonDB 收录了大肠杆菌 K-12 各种转录起始时的调控复合体和调控网络。除此之外,它还包括了各种功能的基因间的相互作用,如转录复合体、操纵子以及简单或复杂的调控子的基因[13]。

由于 sigma19 启动子在 RegulonDB 中只有一条序列,未列入本研究。对于下载的启动子序列,我们先对数据进行筛选。筛选包括去掉数据库中的冗余序列、无注释信息序列、以及属于多类启动子的序列。属于多类启动子的序列指可同时被多类启动子识别的序列,这些序列会影响 PSSM 的预测效果。经过筛选处理后,共得到 2 954 条启动子序列,其中 sigma24 有 511 条,sigma28 有 138 条,sigma32 有 285 条,sigma38 有 130 条,sigma54 有 92 条、sigma70 有 1 787条。

1.2 位点特异性打分矩阵(PSSM)的构建

对大肠杆菌 K-12 的每类启动子,分别构建位点特异性打分矩阵(PSSM)。

1.2.1 构建频数矩阵

从 RegulonDB 下载的启动子序列每一条的长度都为 81 个碱基。以 DNA 序列上的基因翻译起始点的碱基位置定为 0,将其坐标化,则启动子全长即为 -60~20。构建频数矩阵时,首先要统计每个坐标位置中 4 种核苷酸出现的次数,将结果填入 4 行 81 列矩阵中。该矩阵行的名称分别为 A、C、G、T,列名为启动子对应的位置坐标值。

1.2.2 构建伪计数矩阵

频数矩阵的某些元素的值可能为 0。一般认为,这是由于收集数据时的数据量不足造成的。为弥补这一缺陷,通常对频数矩阵中每个元素的值加一个正数(本研究中加 1),生成伪计数矩阵(Peudo count matrix)。

1.2.3 构建概率矩阵

与频数相比,概率被认为是一种更可靠的评判标准。将伪计数矩阵中的每一个元素都除以该列频数总和(公式(1))。其中 $F(x_{i,j})$ 代表伪计数矩阵中的第 i 行第 j 列的元素值,$\sum_{j=1}^{n}F(x_{i,j})$ 代表第 j 列矩阵元素总和。产生的 $P(x_{i,j})$ 就是概率矩阵。

$$P(x_{i,j}) = \frac{F(x_{i,j})}{\sum_{j=1}^{n} F(x_{i,j})} 。 \quad (1)$$

1.2.4 构建几率比(Odds ratio)矩阵

将概率矩阵中每个元素的值除以所对应的碱基在随机条件下出现的概率(均为 0.25),即得到几率比(Odds ratio)矩阵。如公式(2)所示,其中 M 代表实际观测情况,R 代表随机情况。

$$\text{oddsratio} = \frac{P(x \mid M)}{P(x \mid R)} 。 \quad (2)$$

1.2.5 构建对数几率比(Log-Odds ratio)矩阵,即位点特异性打分矩阵(PSSM)

将几率比矩阵中的每个元素取以 2 为底的对数,再取整数部分,即得到对数几率比矩阵(公式(3)),这就是最终的位点特异性打分矩阵(PSSM)。

$$\text{Log-oddsratio} = \text{round}(\log_2 \text{oddsratio}) 。 \quad (3)$$

1.3 利用 PSSM 预测大肠杆菌 K-12 的启动子

1.3.1 预测方法

对于给定的 DNA 序列,根据每个位置上出现的碱基,在 PSSM 中查出相应的得分,然后对各个位置的得分求和,得到总分。采用不同启动子的 PSSM 分别对同一 DNA 序列打分,得分最高者被视为此 DNA 序列所属的启动子类型。

1.3.2 分别对阳性数据集和阴性数据集进行预测

对于特定的启动子类型,阳性数据集指属于该类启动子的 DNA 序列,阴性数据集指不属于该类启动子的 DNA 序列。本研究中,阳性数据集和阴性数据集所包含的序列数目为 1∶1。阴性数据集由不属于某类启动子的其它 5 类启动子序列组成。

1.4 利用 ROC 曲线对预测结果进行评估

使用 RStudio 中的 ROCR 包[14]绘制 ROC 曲线。

ROC 曲线中,AUC 代表“曲线下面积”,该值越趋近 1,说明预测效果越好。

敏感度(Sensitivity, Sens)、特异性(Specificity, Spec)和准确度(Accuracy, Acc)评估预测效果[15]。公式(4)~(6),式中 TP 为真阳性,FN 为假阴性,TN 为真阴性,FP 为假阳性。

$$Sens = \frac{TP}{TP + FN} 。 \quad (4)$$

$$Spec = \frac{TN}{FP + TN} 。 \quad (5)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}。\quad (6)$$

2　结　果

2.1　大肠杆菌 K-12 6 种启动子的 PSSM

我们计算了大肠杆菌 K-12 的 6 种启动子的 PSSM 图 1 是 6 类启动子相应的 logo。为做 ROC 评估提供阳性数据集和阴性数据集,采用如下方法处理 *E.coli* K-12 启动子的 perl 脚本:

```
#! /usr/bin/perl-w
use strict;
#从 RegulonDB 上下载的启动子的原始数据放在一个文件夹下
my @file=glob"PromoterSigma*Set.txt";
my (%sigma_all_sequences, %matrices_total, %prediction_all);
# Score Matrices Computation
foreach my $file(@file){
my ($promoter_index) = $file =~ /PromoterSigma(.*)Set.txt/;
my (%sigma_sequences, %sigma_transposition_sequences);
my ($score, @matrix_score);

# Store the promoters of current file into %sigma_sequences and then
# store all types of the promoter sequences into %sigma_all_sequences.
open(DATA, $file)||die"Can´t open file \"$file\"! \n";
while(<DATA>){
chomp;
(/^#/ or /^\s*$/) and next;
my @fields = split("\t");
($fields[5]! ~/^[acgtACGT]+$/ or $fields[4]=~/,/) and next;
$sigma_sequences{lc($fields[5])} = $fields[4];
}
$sigma_all_sequences{$promoter_index} = \%sigma_sequences;

# Transposit the %sigma_sequences to %sigma_transposition_sequences
# in order to calculate the score matrices.
foreach my $key(keys %sigma_sequences){
my @fields=split("", $key);
for(my $i=0; $i<@fields; $i++){
$sigma_transposition_sequences{$i}.=$fields[$i];
}
}

# Calculate score matrices and store the results in %matrices_total.
foreach my $key(sort {$a <=> $b} keys %sigma_transposition_sequences){
my $promoter_num = length $sigma_transposition_sequences{$key};
my @fields = split("", $sigma_transposition_sequences{$key});
my %acgt=();
foreach my $base(@fields){
(exists $acgt{$base}) ? ($acgt{$base}++) : ($acgt{$base}=1);}
foreach my $base(keys %acgt){
my $base_score = ($acgt{$base}+.1)*4/($promoter_num+.4);
$base_score = sprintf"%.0f", log($base_score)/log(2);
$matrix_score[$key][judge($base)] = int($base_score);
}
}
$matrices_total{$promoter_index} = \@matrix_score;
}

# True or False Promoters Prediction
foreach my $file(@file){
my ($promoter_index) = $file =~ /PromoterSigma(.*)Set.txt/;
my %prediction;

# True promoters score
foreach my $key(keys%{$sigma_all_sequences{$promoter_index}}){
my $score;
for(my $i = 0; $i < length $key; $i++){
$score += ${$matrices_total{$promoter_index}}[$i][judge(substr($key, $i, 1))];
```

```
    }
    $prediction{$key} = $score."\tT";
    }

    # False promoters score
    foreach my $key(keys %sigma_all_sequences){
    $key =~ $promoter_index and next;
    foreach my $false_key(keys %{$sigma_all_sequences{$key}}){
    my $score;
    for(my $i = 0; $i < length $false_key; $i++){
    $score += ${$matrices_total{$promoter_index}}[$i][judge(substr($false_key, $i, 1))];
    }
    $prediction{$false_key} = $score."\tF";
    }
    }
    $prediction_all{$promoter_index} = \%prediction;
    }

    # Print the score into files.
    foreach my $key(keys %prediction_all){
    open(RS,">score_sigma".$key.".txt");
    foreach my $sub_key(keys %{$prediction_all{$key}}){
    print RS $sub_key,"\t",${$prediction_all{$key}}{$sub_key},"\n";
    }
    close RS;
    }

    sub judge{
    my($string)=@_;
    my $num;
    $string=~/a/ and $num=0;
    $string=~/c/ and $num=1;
    $string=~/g/ and $num=2;
    $string=~/t/ and $num=3;
    return $num;
    }
```

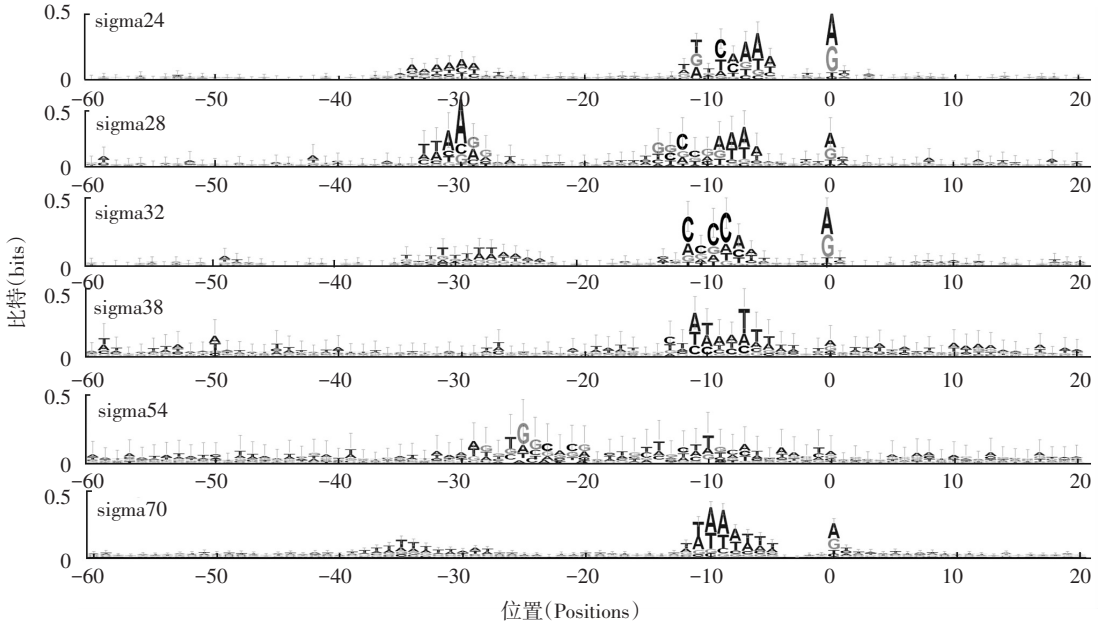

图 1 大肠杆菌 K-12 6 种启动子的 Logo

Fig.1 Logos for six kinds of promoters in *E.coli* K-12

3.2 ROC 曲线

根据对 6 种启动子预测结果,我们使用 R 语言的 ROCR 包绘制了相应的 ROC 曲线(图 2)。表 1 显示了 PSSM 对每一种启动子预测的敏感度(Sensitivity)和特异性(Specificity)。绘制 ROC 曲线的 Rscript 如下:

```
library(ROCR)
setwd("")          # 将工作目录设在原始数据在的地方
par(mfrow=c(2,3),bg="white",mai=c(.6,.
```

```
6,.6,.6))
    for(i in c("24","28","32","38","54","70")){
    data=read.table(paste0("score_sigma",i,".txt"))
    pred <- prediction(data[,2],data[,3])
    perf <- performance(pred,"tpr","fpr")
    sens <- performance(pred,"spec","sens")@x.values
    spec <- performance(pred,"spec","sens")@y.values
    print(paste0("sigma",i,sens,spec))
    auc <- format(performance(pred,"auc")@y.values,digits=2)
    plot(perf,main=paste0("sigma",i),colorize=FALSE,lwd=2,xaxis.cex.axis=1,
    yaxis.cex.axis=1,yaxis.las=1,cex.main=1.5)
    segments(0,0,1,1,lty=2)
    text(0.6,0.5,paste0("AUC=",auc),cex=1.2)
    }
```

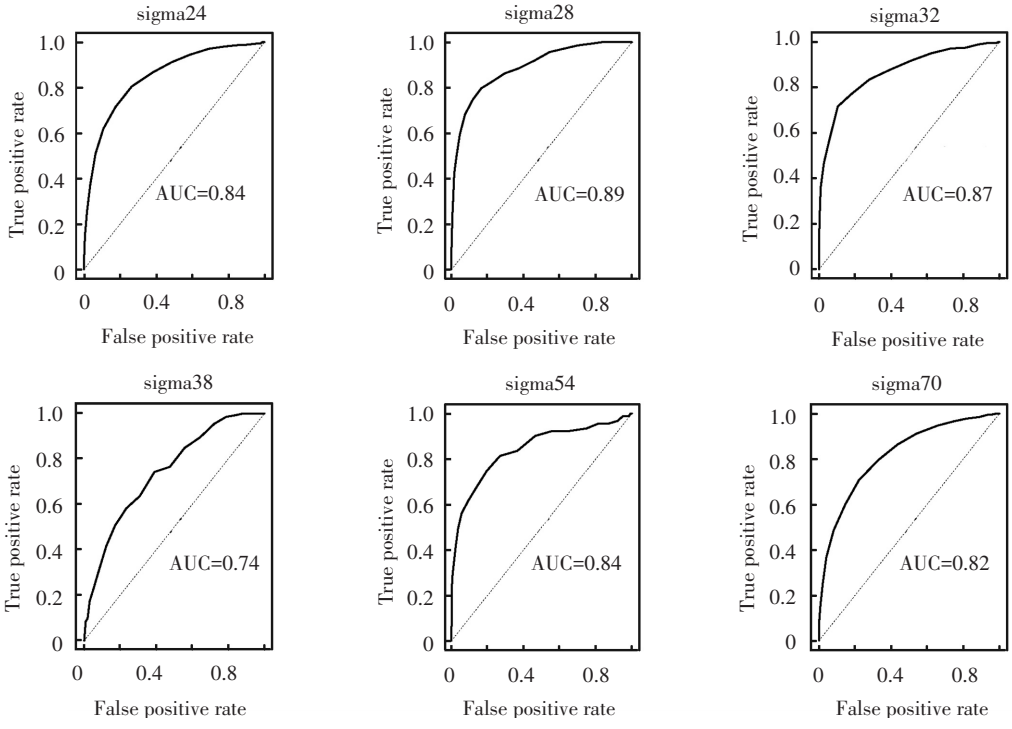

图 2　PSSM 预测的 6 种启动子的 ROC 曲线图

Fig.2　The ROC curves of PSSMs in prediction of *E.coli* promoters

表 1　PSSM 方法预测启动子的灵敏度和特异性

Table 1　Sensitivities and specificities of promoter predictions using PSSM

	σ24	σ28	σ32	σ38	σ54	σ70
Sensitivity	0.71	0.93	0.74	0.74	0.73	0.73
Specificity	0.84	0.74	0.88	0.64	0.77	0.68

表 2　PSSM 与 BacPP 方法的准确度(Acc)比较

Table2　Comparison of Acc between PSSM and BacPP

	σ24	σ28	σ32	σ38	σ54	σ70
PSSM	0.86	0.96	0.93	0.96	0.97	0.74
BacPP	0.87	0.93	0.92	0.89	0.97	0.84

4　讨　论

通过比较 PSSM 与 BacPP 方法[16]的准确度(Acc)(见表 2)可以看出,PSSM 方法在预测 6 种类型的 sigma 因子时,有 3 种启动子(sigma28、sigma32、sigma38)的预测准确度优于 BacPP 方法;一种启动子(sigma54)的预测准确度与 BacPP 方法持平,均为 0.97。

从结果我们可以判断,用 PSSM 模型预测原核生物启动子是一种较为准确的算法。

首先,图 1 中的 ROC 曲线都处于坐标对角虚线的上方,这说明使用 PSSM 预测启动子的概率比随机概率要高。

其次,根据 AUC 的值判断 PSSM 方法的可信性。图中的 AUCs 只有 sigma38 为 0.74,其余均大于 0.8,说明 PSSM 的可信度很高。

再次,预测方法的敏感性和特异性是评价一种预测方法最具说明力的指标。在表 1 中,PSSM 的敏感性和特异性均大于 0.6。另外,sigma28 的敏感性达到了 0.93,达到了相当高的水平。

PSSM 算法为大肠杆菌 K-12 启动子的预测提供了一种较为准确的可靠方法。从 ROC 曲线的形状、AUC 值,以及敏感度、特异性和准确度值均表明 PSSM 在预测启动子方面的有效性。由于原核生物的启动子具有较大的保守性,PSSM 可以作为原核生物启动子预测的一种有效方法。PSSM 方法缺陷在于,使用 PSSM 方法需要指定打分矩阵的窗口大小,该缺陷可以通过采用隐马尔科夫模型(HMM)方法得以克服。另外,我们还将采用多重交叉验证的方法进一步提高预测的准确度。

参考文献(References)

[1] 杨明,李权胜.原核生物的 sigma 因子[J].河南医学研究,1999,8(1):88-90.
YANG Ming,LI Quansheng.Prokaryotic sigma factors [J]. Henan medical research,1999,8(1):88-90.

[2] HAWLEY D K , MCCLURE W R.Compilation and analysis of Escherichia coli promoter DNA sequences[J].Nucleic Acids Rresearch,1983,11(8):2237-2255.

[3] THÖNY B,HENNECKE H.The -24/-12 promoter comes of age [J].FEMS Microbiol. Rev,1989,63:341-357.

[4] GUHATHAKURTA D.Computational identification of transcriptional regulatory elements in DNA sequence [J].Nucleic Acids Research,2006,34(12):3585-3598.

[5] MANN S,LI J,CHEN Y P P.A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts[J].Nucleic Acids Res,2007,35:e12.

[6] ZHANG S,XU M,LI S,et al.Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes[J].Nucleic Acids Research,2009,37(10):e72.

[7] LI G,LIU B,MA Q,et al.A new framework for identifying cis-regulatory motifs in prokaryotes[J].Nucleic Acids Research,2011,39(7):e42.

[8] MA Q,LIU B,ZHOU C,et al.An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale[J].Bioinformatics,2013,29(18):2261-2268.

[9] AHMAD S,SARAI A.PSSM-based prediction of DNA binding sites in proteins[J].BMC Bioinformatics,2005,6:33.

[10] GERSHENZON N I,STORMO G D,IOSHIKHES I P.Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites[J].Nucleic Acids Research,2005,33(7):2290-2301.

[11] CLAVERIE J M,AUDIC S.The statistical significance of nucleotide position-weight matrix matches[J].Computer Applications in the Biosciences,1996,12(5):431-439.

[12] 韦修喜,周永权.基于 ROC 曲线的两类分类问题性能评估方法[J].计算机技术与发展,2010,20(11):47-50.
WEI Xiuxi,ZHOU Yongquan.Assess of performance of two types of classification methods based on ROS[J].Computer Technology and Development,2010,20(11):47 -50.

[13] SALGADO H,PERALTA-GIL M,GAMA-CASTRO S, et al.RegulonDB(version 8.0): Omics data sets,evolutionary conservation, regulatory phrases, cross-validated gold standards and more[J].Nucleic Acids Research,2012,41(D1):D203-213.

[14] SING T,SANDER O,BEERENWINKEL N,et al.ROCR: visualizing classifier performance in R[J].Bioinformatics,2005,21(20):7881.

[15] ZHOU X,LI Z,DAI Z,et al.Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform[J].Journal of Theoretical Biology,2013,319:1-7.

[16] DE AVILA E SILVA S,ECHEVERRIGARAY S,GERHARDT G J.BacPP: bacterial promoter prediction-a tool for accurate sigma-factor specific assignment in enterobacteria[J].J.Theor Biol,2011,287:92-99.