利用计算生物学方法识别原核启动子的研究进展

苏 伟[1]，孙自杰[1]，岳 鹏[2]，林 昊[1*]

(1. 电子科技大学生命科学与技术学院 成都 611731；2. 成都东软学院健康医疗科技学院 成都 611844)

【摘要】原核启动子作为 DNA 中的一个关键区域，含有 RNA 聚合酶特异性结合和基因转录起始所需的保守序列，在转录调控中发挥着重要作用。然而，由于实验方法在实验周期和实验耗材上的限制，对启动子序列进行批量准确的鉴定仍然是分子生物学领域一项艰巨的任务。随着计算机技术的发展，出现了多个基于计算生物学的原核启动子预测方法，这些方法在数据质量、数据集大小、提取的特征、特征选择技术、分类算法以及评估策略方面表现出高度的多样性。该文系统地比较并总结这些方法，以便改进和进一步发展原核启动子识别技术。

关 键 词 生物信息； 机器学习； 预测器； 原核启动子

中图分类号 TP391；Q615 **文献标志码** A doi:10.12178/1001-0548.2021201

A Brief Review for Identifying Prokaryotic Promoters Based on Computational Biology

SU Wei[1], SUN Zijie[1], YUE Peng[2], and LIN Hao[1*]

(1. School of Life Science and Technology, University of Electronic Science and Technology of China Chengdu 611731;
2. School of Healthcare Technology, Chengdu Neusoft University Chengdu 611844)

Abstract As a key region of deoxyribonucleic acid (DNA), prokaryotic promoter contains the conserved sequence required for specific binding of ribonucleic acid (RNA) polymerase and transcription initiation, and plays an important role in transcription regulation. However, due to the limitations of experimental methods that are long experimental period and high cost, the identification of prokaryotic promoter sequences remains a major challenge. With the development of computer technology, dozens of prokaryotic promoter identification methods based on computational biology have emerged, which show a high degree of diversity in terms of data quality, dataset size, extracted features, feature selection techniques, classification algorithms and evaluation strategies. Thus, there is an urgent need to systematically compare and summarize these methods so as to improve and further develop prokaryotic promoter recognition techniques.

Key words bioinformatics; machine learning; predictor; prokaryotic promoter

启动子通常位于基因上游，能与 RNA 聚合酶特异性结合并起始转录的一段 DNA 序列，作为转录起始过程的关键元件，激活 RNA 聚合酶与模板 DNA 结合，是基因表达和转录调节的起始步骤[1]。

原核生物 RNA 聚合酶中的 σ 因子可以特异性识别并结合启动子。在大肠杆菌中，存在多种 σ 因子，根据分子量可以分为 7 类，σ70、σ54、σ38、σ32、σ28、σ24、σ19，在已知的 7 类 σ 因子中前 6 类保守性极强，而 σ19 在大多数基因组中是缺失的[2]。每一类 σ 因子具有特定的生物学功能[3-6]，σ70 主要负责持家基因的转录；σ54 被认为是参与氮代谢的调控因子以及控制一些辅助进程；σ38 参与稳定期基因的调节；σ32 是热休克 σ 因子 (热激因子)；σ28 参与鞭毛的合成；σ24 与极端热应激反应有关；σ19 则参与对铁离子转运系统的调控。根据 σ 因子的同源性，可将其大致分为两类：一类是 σ70 家族，包括 σ70、σ38、σ32、σ28、σ24、σ19；另一类是 σ54 家族。大肠杆菌基因组内的启动子类型依据与之结合的 σ 因子种类也可分为相应的类型。不同类型的启动子共有序列也有所差异。

收稿日期：2021－07－29；修回日期：2021－08－30
基金项目：国家自然科学基金 (61772119)，四川省杰出青年基金 (2020JDJQ0012)
作者简介：苏伟 (1996－)，男，博士生，主要从事生物信息学方面的研究.
*通信作者：林昊，E-mail：linhao@uestc.edu.cn

因此，启动子也依据被识别的片段分为 σ70 家族和 σ54 家族。如 σ70 启动子具有两个重要的基序区域，−10 区和−35 区，分别位于转录起始位点上游约 10 bp 和 35 bp 处。−10 区含有保守序列“TATAAT”，又被称为 Pribnow box 或 TATA box，富含腺嘌呤 (adenine, A) 和胸腺嘧啶 (thymine, T)，有助于 DNA 双链解螺旋分离；−35 区则由 6 个保守的核苷酸“TTGACA”组成[7]。除了 σ70 因子，−10 区和−35 区也是被 σ70 家族其他因子识别的重要片段。相比之下，σ54 启动子的共有序列及其位置与 σ70 启动子具有明显差异，在 σ54 启动子的−24 区和−12 区存在保守区域，其保守序列分别是“TGGCA[CT][GA]”和“TGC[AT][TA]”[8]。

启动子序列的鉴定对于研究基因表达、分析基因调控机制、研究基因结构以及注释基因信息至关重要。准确识别启动子的方法一般是依靠昂贵且耗时费力的实验检测方法，然而，在全基因组范围内进行检测是一项艰巨的任务。随着测序技术以及计算机技术的发展，越来越多生物的全基因组被测序出来，尤其是原核生物，因此出现了基于计算生物学的启动子预测方法，这些预测方法在不断地改进，有助于鉴别启动子序列。

1 原核启动子识别方法

原核生物 RNA 聚合酶中的 σ 因子可以特异性识别并结合启动子，如图 1 所示。

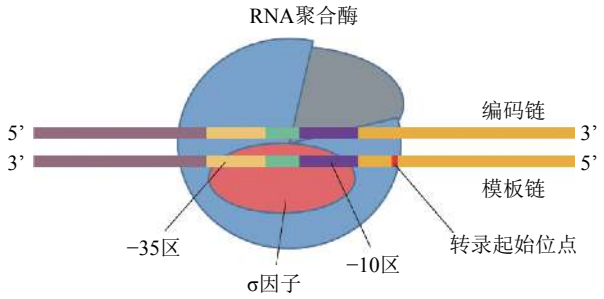

图 1 大肠杆菌 σ70 启动子与 RNA 聚合酶结合

2005 年至今已经开发了 30 多种计算方法来预测原核生物启动子，大致流程如图 2 所示。这些方法在许多方面有所不同，包括使用的基准数据集、特征提取方法、特征选择技术以及分类方法等。本文总结了 39 种原核启动子预测方法，从基准数据集信息、特征表示、特征选择、性能评估策略等多方面进行了比较和分析，如表 1 所示。

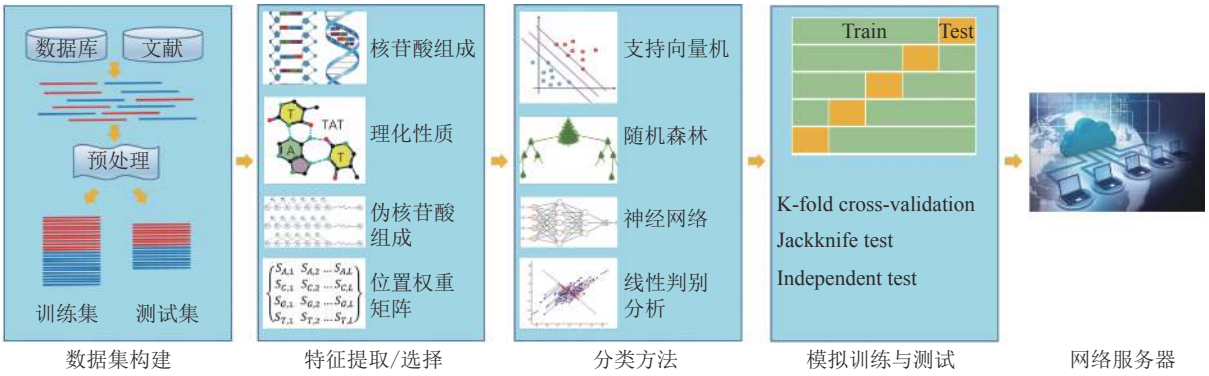

图 2 基于计算方法的原核启动子预测流程

39 个预测工具根据其功能可分为以下 3 类。

1) 普通启动子的识别。工具 1～9[9-17] 属于这一类，这些工具收集各种原核生物的启动子作为基准数据集，包含大肠杆菌、枯草芽孢杆菌、结核杆菌、乳酸乳球菌、天蓝色链霉菌、分枝杆菌以及假单胞菌等。并没有指出这一类启动子具体的类型，因此这些方法只是简单地对启动子序列进行预测。

2) 特殊类型启动子的预测。这一类方法包含工具 10～30[18-38]。这些工具以具体类型的启动子作为基准数据集，如大肠杆菌的 6 类启动子，原核生物的 σ54 启动子，蓝细菌的 5 类启动子等。不同类型的启动子在基因表达调控过程中起着不同且重要的作用，如目前已知的 σ54 启动子仅有数百条，而原核生物有 3 万多种，还有大量 σ54 启动子未被发现。σ54 启动子参与了氮代谢的调控，因此 σ54 启动子的预测对于了解原核生物氮代谢过程具有重要意义。

3) 启动子的预测与分类。剩余的 9 个方法[39-47] 均属于这一类，以大肠杆菌启动子作为数据集。这类方法具有一个典型的特征，即模型具有两层结构，第一层均是对启动子的预测，第二层是对启动子属性分类。工具 31～36 除了预测启动子和非启

动子，第二层还判断启动子的具体类型 (σ70, σ54, σ38, σ32, σ28, σ24)。实际上，启动子还有强弱之分。强启动子能增加转录频率从而提高基因的表达水平，所以预测启动子的强度也很重要。基于此，模型 37～39 的第二层鉴定启动子的强弱 (Strong, Weak)。

随着后基因组时代的到来以及计算机的发展，对于原核启动子的预测方法也不局限于初步的分类，还增加了对启动子类型和强度的鉴定，为了解基因调控过程提供新信息。

表 1　39 个原核启动子预测工具比较

Tools	Benchmark dataset size (promoter)	Sequence similarity	Feature extraction/ selection	Classification algorithm	Evaluation strategy	AUC
1.TLS-NNPP[9]	771 (*E.coli*)	/	The empirical probability distribution of TSS-TLS distance	ANN	Independent test	/
2.SIDD[10]	500 (*E.coli*)	/	SIDD	FLD	Independent test	/
3.FS_LSSVM[11]	53 (*E.coli*)	/	A domain theory for promoters/ C4.5 decision tree	LSSVM	10-fold cross-validation	/
4.Free energy[12]	1 044 (*E.coli*) 879 (*B.subtilis*)	/	Free energy	Modified scoring function	Independent test	/
5.PromPredict[13]	1 145 (*E.coli*) 615 (*B.subtilis*) 82 (*M.tuberculosis*)	/	GC content; Average free energy	difference between the average free energy	Training and validation	/
6.SIDD-ANN[14]	1 648 (*E.coli*)	/	SIDD profile data	ANN	Independent test	/
7.PePPER[15]	L.lactis	/	PWM	HMM	/	/
8.G4PromFinder[16]	3 570 (*S.coelicolor*) 2 117 (*P.aeruginosa*)	/	AT-rich element and G-quadruplex motif-based algorithm	/	Independent test	/
9.LN-QSAR[17]	135 (*M.bovis*)	/	Pseudo-folding 2D lattice graph	LDA	Independent test	/
10.Ensemble-SVM[18]	450 (*E.coli* σ70)	/	k-mer with location with respect to the TSS/ Symmetric uncertainty	Ensemble-SVM	10-fold cross-validation	/
11.TSS-PREDICT[19]	450 (*E.coli* σ70) 205 (*B.subtilis*) 26 (*C.trachomatis*)	/	Information Content; PWM	Ensemble-SVM	Independent test	/
12.TSS-SLP[20]	669 (*E.coli* σ70)	/	Dinucleotide Frequency Features	SLP	5-fold cross-validation; Independent test	/
13.PCSF[21]	683 (*E.coli* σ70)	/	Conversation of sequence segments; PCSF	Score function	10-fold cross-validation	/
14.IPMD[22]	270 (*B.subtilis* σ43) 741 (*E.coli* σ70)	/	PCSF; ID	Modified MD	10-fold cross-validation	0.847 (*B.subtilis*) 0.920 (*E.coli*)
15.70ProPred[23]	741 (*E.coli* σ70)	/	PSTNPss; PseEIIP	SVM	5-fold cross-validation; Jackknife test	0.990
16.iProEP[24]	270 (*B.subtilis*) 741 (*E.coli*)	≤80%	PseKNC; PCSF/ mRMR; IFS	SVM	10-fold cross-validation	0.988 (*B.subtilis*) 0.976 (*E.coli*)
17.IPWM[25]	683 (*E.coli* σ70)	/	Entropy-based conservative characteristics; Improved PWM	Score function	10-fold cross-validation	/
18.BacPP[26]	1 034 (*E.coli*)	/	Binary digits	ANN	(2,3,10)-fold cross-validation; Independent test	/
19.vw Z-curve[27]	1 401 (*E.coli*) 660 (*B.subtilis*)	/	variable-window Z-curve/ IFS	PLS	10-fold cross-validation	/
20.Stability[28]	1 035 (*E.coli*)	/	DNA duplex stability	ANN	(2,3,10)-fold cross-validation	/
21.iPro54-PseKNC[29]	161 (*prokaryotic* σ54)	≤75%	PseKNC/ F-score; IFS	SVM	Jackknife test	/
22.Promote Predictor[30]	161 (*prokaryotic* σ54)	≤75%	Motif profile-based ANF/ MRMD	Bagging; RF; SVM	10-fold cross-validation; Independent test	/

续表

Tools	Benchmark dataset size (promoter)	Sequence similarity	Feature extraction/ selection	Classification algorithm	Evaluation strategy	AUC
23.meta-predictior[31]	579 (*E.coli* σ70)	⩽45%	sequence-based features; structure-based features	Meta-predictor	Independent test	0.850
24.bTSSfinder[32]	3 597 (*E.coli*) 12 797 (*Nostoc*) 351 (*Synechocystis*) 1 471 (*S.elongatus*)	/	PWM; Physicochemical properties/ Mahalanobis distance	ANN	Independent test	/
25.iPro70-PseZNC[33]	741 (*E.coli* σ70)	/	PseZNC/ F-score; IFS	SVM	5-fold cross-validation	0.909
26.iPromoter-FSEn[34]	741 (*E.coli* σ70)	/	Nucleotide Statistics; k-mer; g-gapped k-mer; Approximate signal pattern count; Position specific occurences; Distribution of nucleotides/ Feature subspace	Ensemble learning	10-fold cross-validation	0.932
27.iPro70-FMWin[35]	741 (*E.coli* σ70)	/	k-mer; g-gapped k-mer; Pattern finding; Positioning distance count/ Adaboost	LR	10-fold cross-validation	0.959
28.CNNProm[36]	839 (*E.coli* σ70) 746 (*B.subtilis*)	/	one-hot	CNN	5-fold cross-validation	/
29.IBBP[37]	1888 (*E.coli* σ70)	/	Image-based and evolutionary approach	SVM	Independent test	/
30.SAPPHIRE[38]	170 (*P. aeruginosa* and *P. putida* σ70)	/	one-hot	ANN	5-fold cross-validation; Independent test	/
31.iPromoter-2L[39]	2 860 (*E.coli*)	⩽80%	Multi-window-based PseKNC	RF	5-fold cross-validation; Jackknife test	/
32.iPromoter-2L2.0[40]	2 860 (*E.coli*)	⩽80%	Smoothing Cutting Window algorithm; k-mer; PseKNC	SVM; Ensemble learning	5-fold cross-validation	/
33.MULTiPly[41]	2 860 (*E.coli*)	⩽80%	Bi-profile bayes; KNN; k-mer; DAC/ F-score	SVM	5-fold cross-validation; Jackknife test; Independent test	/
34.pcPromoter-CNN[42]	2 860 (*E.coli*)	⩽80%	one-hot	CNN	5-fold cross-validation; Independent test	0.957
35.iPromoter-BnCNN[43]	2 860 (*E.coli*)	⩽80%	one-hot; k-mer; Structural properties	CNN	5-fold cross-validation; Independent test	/
36.SELECTOR[44]	2 860 (*E.coli*)	⩽80%	CKSNAP; PCPseDNC; PSTNPss; DNA strand	Ensemble learning	5-fold cross-validation; Independent test	0.984
37.iPSW(2L)-PseKNC[45]	3 382 (*E.coli*)	⩽85%	NCP; ANF	SVM	5-fold cross-validation	0.905
38.deepPromoter[46]	3 382 (*E.coli*)	⩽85%	Combination of Continuous FastText N-Grams/ MRMD	CNN	5-fold cross-validation	0.885
39.iPSW(PseDNC-DL)[47]	3 382 (*E.coli*)	⩽85%	one-hot; PseDNC	CNN	5-fold cross-validation	0.925

PWM: position weight matrix; SIDD: stress-induced DNA duplex destabilization; PCSF: position-correlation scoring function; ID: increment of diversity; PSTNPss: position-specific trinucleotide propensity based on single-strand; PseEIIP: electron-ion interaction pseudo-potentials of trinucleotide; PseKNC: pseudo k-tuple nucleotide composition; ANF: accumulated nucleotide frequency; PseZNC: pseudo multi-window Z-curve nucleotide composition; KNN: k-nearest neighbors; DAC: dinucleotide-based auto-covariance; PCPseDNC: parallel correlation pseudo dinucleotide composition; NCP: nucleotide chemical property; PseDNC: pseudo dinucleotide composition; mRMR: minimum redundancy maximum relevance; IFS: incremental feature selection; MRMD: maximum-relevance-maximum-distance; ANN: artificial neural network; SVM: support vector machine; FLD: fisher linear discriminant; SLP: single-layer perceptron; LSSVM: least square support vector machine; MD: mahalanobis discriminant; PLS: partial least squares; HMM: hidden markov models; RF: random forest; LR: logistic regression; CNN: convolution neural network; LDA: linear discriminant analysis.

2 数据集构建

建立原核启动子预测模型的第一步需要构建一个高质量的基准数据集。大肠杆菌 (*E.coli*) 作为原核生物中被广泛使用、研究的模式生物，其经过实

验验证的转录调控信息已被系统地收录在 RegulonDB 数据库[48] 中。DBTBS 数据库[49] 则收集整理了关于枯草芽孢杆菌 (*B.subtilis*) 的启动子数据。因此，RegulonDB 和 DBTBS 数据库为预测方法提供了数据基础。39 个工具中共有 35 个工具的数据集包含大肠杆菌和枯草芽孢杆菌启动子。

另外，为了减少由序列同源性引起的潜在误差，通常会使用 CD-HIT[50] 工具以 75%～85% 的序列相似性阈值来去除掉数据集中序列冗余。原核启动子相较真核启动子，其结构相对较为简单、功能元件也相对较少，因此一般选择转录起始位点 (transcriptional start site, TSS) 上游 60 bp 以及下游 20 bp 作为原核启动子序列，不仅包含了重要的共有序列，如−35 区、−10 区、起始位点等，也避免了序列过长导致引入不必要的信息，具体数据可见原核启动子数据库 (prokaryotic promoter database, PPD)[51]。

3　特征提取

几乎所有的机器学习方法是以数值向量作为输入，因此需要一个合适的特征描述方法将数据集中的每一个样本转换为能够反映序列信息的数值向量。在原核启动子识别工作中，这些特征大致可以分为 5 类：核苷酸组成、核苷酸理化性质、伪核苷酸组成、二进制编码以及位置权重矩阵，以下对这 5 类特征进行简单的介绍。

3.1　核苷酸组成

核苷酸组成，也叫 k-mer，统计了 DNA 序列片段的所有可能组合的 k 长度子串出现频率，其计算公式为：

$$f_i = \frac{N(i)}{L-k+1} \tag{1}$$

式中，i 代表某一 k 联体，有 4^k 种可能性；$N(t)$ 表示 DNA 序列中某一 k 联体出现的次数；L 表示 DNA 序列的长度。随着 k 值的增加，DNA 序列的局部或短程信息也会逐渐增加。

此外，核苷酸组成还包括了 g-gapped k-mer，GC 含量，累积核苷酸频率 (accumulated nucleotide frequency, ANF) 等。ANF 表示了每一个碱基在序列中的分布密度，表达式为：

$$d_i = \frac{1}{|s_i|}\sum_{i=1}^{L} N(s_i) \qquad N(s_i) = \begin{cases} 1 & s_i = q \\ 0 & \text{其他} \end{cases} \tag{2}$$

式中，$|s_i|$代表第 i 个碱基的位置；$N(s_i)$表示某一碱基出现频数；$q \in \{A, C, G, T\}$。

3.2　理化性质

DNA 序列中碱基的理化性质也可作为启动子预测的重要特征，包括核苷酸的化学性质、双链的稳定性、自由能、应激诱导的 DNA 双链不稳定性等。

根据表 2 中对不同核苷酸的分类，DNA 序列中第 i 个核苷酸可以表示为：

表 2　核苷酸化学性质

Chemical property	Class	Nucleotides
Ring Structure	Purine	A, G
	Pyrimidine	C, T
Functional Group	Amino	A, C
	Keto	G, T
Hydrogen Bond	Strong	C, G
	Weak	A, T

$$N_i = (x_i, y_i, z_i) \tag{3}$$

式中，x_i, y_i, z_i 分别表示指环结构 (ring structure)，功能组别 (function group)，以及氢键 (hydrogen bond)，如：

$$\begin{aligned} x_i &= \begin{cases} 1 & N_i \in \{A, G\} \\ 0 & N_i \in \{C, T\} \end{cases} \\ y_i &= \begin{cases} 1 & N_i \in \{A, C\} \\ 0 & N_i \in \{G, T\} \end{cases} \\ z_i &= \begin{cases} 1 & N_i \in \{A, T\} \\ 0 & N_i \in \{C, G\} \end{cases} \end{aligned} \tag{4}$$

因此 4 种碱基 (A, C, G, T) 可以分别表示为 (1,1,1)，(0,1,0)，(1,0,0) 和 (0,0,1)。

3.3　伪核苷酸组成

伪核苷酸组成 (pseudo k-tuple nucleotide composition, PseKNC) 最初是由文献 [52] 提出，分为 I 型和 II 型。这两种方法基于核苷酸的物化性质引入了 DNA 序列的全局或长程顺序信息。

I 型 PseKNC，也叫平行相关伪核苷酸组成，将每一条 DNA 序列转化为 $4^k + \lambda$ 维的向量，具体表示为：

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_u + \omega\sum_{j=1}^{\lambda}\tau_j} & 1 \leqslant u \leqslant 4^k \\ \dfrac{\omega\tau_{u-4^k}}{\sum_{i=1}^{4^k} f_u + \omega\sum_{j=1}^{\lambda}\tau_j} & 4^k + 1 \leqslant u \leqslant 4^k + \lambda \end{cases} \tag{5}$$

II 型 PseKNC，也叫串联相关伪核苷酸组成，可产生 $4^k + \lambda\Lambda$维向量：

$$d_u=\begin{cases}\dfrac{f_u}{\sum\limits_{i=1}^{4^k}f_u+\omega\sum\limits_{j=1}^{\lambda\Lambda}\tau_j} & 1\leqslant u\leqslant 4^k\\ \dfrac{\omega\tau_{u-4^k}}{\sum\limits_{i=1}^{4^k}f_u+\omega\sum\limits_{j=1}^{\lambda\Lambda}\tau_j} & 4^k+1\leqslant u\leqslant 4^k+\lambda\Lambda\end{cases} \tag{6}$$

式 (5) 和式 (6) 中的f_u与式 (1) 相同；前 4^k 个元素是核苷酸组成特征，后面的元素是伪核苷酸组成特征；λ是一个正整数，反映序列顺序关联阶数；ω是权重因子，用于权衡核苷酸组分和 DNA 序列局部结构性质的影响；τ_j代表的是 m 阶关联因子，反映了每条 DNA 序列所有二核苷酸的 m 阶顺序关联性。

3.4 二进制编码

二进制编码通过将 4 种核苷酸转换成包含 4 个元素的向量作为特征，其中一个元素为 1，其余为 0，既 A、C、G 和 T 分别表示为 (1,0,0,0)，(0,1,0,0)，(0,0,1,0) 以及 (0,0,0,1)。因此，一段长为 L 的 DNA 序列可以用 $L\times4$ 的二维矩阵表示。

3.5 位置权重矩阵

位置权重矩阵 (position weight matrix, PWM) 可用来表示序列的保守片段，以序列每一位置的碱基保守程度为参量，分别计算每种碱基的保守指数，以此作为特征，具体表示为：

$$S_{i,j}=\log\frac{q_{i,j}}{b_i} \tag{7}$$

式中，$S_{i,j}$表示碱基 i 在第 j 个位置的保守指数；$q_{i,j}$是指在背景序列中碱基 i 出现在第 j 个位置的频率；b_i是背景概率。

因此，PWM 是一个 $4\times L$ 的二维矩阵：

$$\boldsymbol{P}=\begin{Bmatrix}S_{A,1}S_{A,2}\cdots S_{A,L}\\ S_{C,1}S_{C,2}\cdots S_{C,L}\\ S_{G,1}S_{G,2}\cdots S_{G,L}\\ S_{T,1}S_{T,2}\cdots S_{T,L}\end{Bmatrix} \tag{8}$$

4 特征选择

从式 (1) 以及式 (5)、式 (6) 可以看出，随着 k 值的增加，特征维度呈指数级增长，会导致“维度灾难”以及过拟合问题，而且由不同特征提取方法整合形成的融合特征集合往往会夹杂一些冗余或不相关的信息，所以为了避免出现上述问题并且提高计算效率，筛选有用的特征也是必不可少的步骤。

4.1 最小冗余最大相关

最小冗余最大相关 (minimum redundancy maximum relevance, mRMR)[53] 是一种通过筛选相关性最大的特征来减少信息冗余的方法。mRMR 的应用大大减少了特征维数和模型训练的时间，几乎不丢失有效信息。

对于两个随机变量 x 和 y，其互信息为：

$$I(x,y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}\mathrm{d}x\mathrm{d}y \tag{9}$$

式中，$p()$ 表示概率密度函数。

最大相关性为：

$$\max D(S,c)\quad D=\frac{1}{|S|}\sum_{x_i\in S}I(x_i;c) \tag{10}$$

式中，c 为类别变量；S 为特征子集。

最小冗余度则表示为：

$$\min R(S)\quad R=\frac{1}{|S|^2}\sum_{x_i,x_j\in S}I(x_i;x_j) \tag{11}$$

最后的评选标准如式 (12) 所示：

$$\max\phi(D,R)\quad \phi=D-R \tag{12}$$

mRMR 会将所有特征的最大相关最小冗余打分按从大从小排序，值越大表明该特征越重要。

4.2 最大相关最大距离

当两个特征高度依赖时，它们对模型的贡献不能叠加，文献 [54] 基于距离函数提出了最大相关最大距离 (max-relevance-max-distance, MRMD) 来衡量每个特征的独立性。

MRMD 包含两个方面的特征排序度量：1) 特征子集与目标类别的相关性；2) 特征子集的冗余度。采用皮尔逊相关系数来衡量相关性、多种距离函数来计算冗余度。皮尔逊相关系数越大，特征与目标类别之间的相关性越高；特征距离越大，特征子集的冗余度越低；相关性与距离之和大的特征被选入最终的特征子集。因此，MRMD 生成的特征子集冗余度最低，与目标类别的相关性最强。

4.3 F-score

F-score 是一种基于 filter 的特征选择方法，对每一个特征进行重要性打分，其具体计算方法为：

$$F_{(i)}=\frac{\left(\bar{x}_i^{(+)}-\bar{x}_i\right)^2+\left(\bar{x}_i^{(-)}-\bar{x}_i\right)^2}{\dfrac{1}{n^+-1}\sum\limits_{k=1}^{n^+}\left(x_{k,i}^{(+)}-\bar{x}_i^{(+)}\right)^2+\dfrac{1}{n^--1}\sum\limits_{k=1}^{n^-}\left(x_{k,i}^{(-)}-\bar{x}_i^{(-)}\right)^2} \tag{13}$$

式中，n^+、n^-分别表示正负样本的数量；$\bar{x}_i^{(+)}$、$\bar{x}_i^{(-)}$、$\bar{x}_i$分别指第 i 个特征在正样本、负样本以及所有样本中的平均值；$x_{k,i}^{(+)}$、$x_{k,i}^{(-)}$分别指的是正负样本

中第 k 条序列的第 i 个特征的数值。

F-score 通常与增量特征选择技术相结合来确定最优特征子集。

4.4　增量特征选择

增量特征选择 (incremental feature selection, IFS) 方法适用于确定最优特征子集。该方法的核心思想是将按重要性评分降序的特征依次加入到特征子集中，形成新的子集，将每一个子集输入至模型中，从而根据结果决策出最优特征子集。

5　分类方法

选择合适的算法可以使最终的模型具有良好的性能和泛化能力，各种监督学习方法已经被广泛应用于预测原核启动子，大致有以下 4 类。

5.1　支持向量机

支持向量机 (support vector machine, SVM)[55] 是基于监督学习方式对数据进行二元分类，在样本空间中寻找最优分类超平面使得两类的间隔最大。

对于线性可分的情况，存在一个分类超平面能将训练样本正确分类。而对于线性不可分的情况，需要使用核函数将低维不可分样本映射到更高维的特征空间，使得样本在高维空间中线性可分。

5.2　神经网络

神经网络 (neural networks, NN) 学习是一种模拟生物大脑神经网络的自适应计算模型。随着近年来人工智能的快速发展，人工神经网络 (artificial neural network, ANN) 及其卷积神经网络 (convolutional neural network, CNN) 已成为研究生物信息学问题的重要方法。

基本的 ANN 结构包括输入层、隐藏层和输出层，主要特点是信号正向传播，误差反向传播。通过最小化误差函数，修正神经元间的连接权重，当其误差小于一定阈值的时候，即停止训练。

CNN 目前在很多研究领域都取得了巨大的成功，如语音识别、图像识别、自然语言处理等，是深度学习的代表算法之一。CNN 通常由输入层、卷积层、激活函数、池化层、全连接层和输出层组成。与传统的神经网络不同的是 CNN 采用局部连接和权值共享，使得网络易于优化并且降低了模型的复杂度，减小过拟合风险。

5.3　集成学习

集成学习 (ensemble learning, EL) 通过构建并结合多个学习器来完成学习任务。在预测原核启动子的方法中，集成学习也是被广泛应用的，如随机森林 (random forest, RF)。

RF 是一种基于决策树的集成学习方法，在决策树的训练过程中引入了随机属性选择。对于基决策树的每个结点，随机选择该结点属性集合中的一个子集，再从这个子集中选择一个最优属性用于划分。RF 的每一个决策树都会产生一个分类结果，通过投票决定最终输出。与单一的决策树相比，RF 具有较强的鲁棒性，并且对大数据具有较好的处理效果。

5.4　线性判别分析

线性判别分析 (linear discriminant analysis, LDA) 在二分类问题上最初是由文献 [56] 提出的，亦称为“Fisher 判别分析”。

LDA 的核心思想相对简单：首先将训练集中的样本投影到一条直线上，使得同一类样本尽可能靠近，不同类样本尽可能远离；当新样本进来时，将其投影到同一直线上，从而根据投影点的位置判断其类别。

6　性能评估

在统计分析中，独立测试集和 K 折叠交叉验证已经被广泛地应用于验证分类器性能。当样本数量足够多时，会将基准数据集划分为训练集和独立测试集。独立测试集由于未参与模型的训练，可以更好地评价模型性能。在原核启动子识别模型中，K 折叠交叉验证的应用最为广泛，其基本思想是重复利用数据，每一个样本既可以作为训练集参与模型训练，也会作为测试集参与模型评估。方法是将数据平均分成 K 份，K−1 个子集用作训练，剩余一份用作测试，重复 K 次，最后返回 K 次结果的平均值。K 折叠交叉验证最大程度上利用了每一个数据，能更好地反应模型的预测性能。

另外，受试者工作特征曲线 (receiver operating characteristic curve, ROC) 下面积 AUC 值也可以反应模型性能，其值越接近于 1，表明模型性能越好。

7　结 束 语

近年来，基于生物信息学的原核启动子预测方法备受学者关注，已有多种方法被提出。为了充分了解这个领域的发展现状，本文收集并系统地分析了 2005 年至今共计 39 个原核启动子预测方法，详细阐述了这些方法的数据集构建、特征选择、特征提取、分类算法以及性能评估，详细信息如表 1 所示。

目前，对原核启动子预测的研究取得了令人满意的结果。随着更多原核生物的基因组被测序出来，被研究的物种也不局限于少数几个模式生物，使用这些预测算法有助于了解原核生物基因调控机制。本文系统地比较了原核启动子预测方法，为研究此问题提供新思路、新角度。

参 考 文 献

[1] PERRON G G, WHYTE L, TURNBAUGH P J, et al. Functional characterization of bacteria isolated from ancient arctic soil exposes diverse resistance mechanisms to modern antibiotics[J]. Plos One, 2015, 10(3): e0069533.

[2] COOK H, USSERY D W. Sigma factors in a thousand E. coli genomes[J]. Environ Microbiol, 2013, 15(12): 3121-3129.

[3] BARRIOS H, VALDERRAMA B, MORETT E. Compilation and analysis of sigma(54)-dependent promoter sequences[J]. Nucleic Acids Res, 1999, 27(22): 4305-4313.

[4] JANGA S C, COLLADO-VIDES J. Structure and evolution of gene regulatory networks in microbial genomes[J]. Res Microbiol, 2007, 158(10): 787-794.

[5] LEE S J, GRALLA J D. Sigma38 (rpoS) RNA polymerase promoter engagement via -10 region nucleotides[J]. J Biol Chem, 2001, 276(32): 30064-30071.

[6] POTVIN E, SANSCHAGRIN F, LEVESQUE R C. Sigma factors in Pseudomonas aeruginosa[J]. Fems Microbiol Rev, 2008, 32(1): 38-55.

[7] ABRIL A G, RAMA J L R, SANCHEZ-PEREZ A, et al. Prokaryotic sigma factors and their transcriptional counterparts in Archaea and Eukarya[J]. Appl Microbiol Biot, 2020, 104(10): 4289-4302.

[8] 丁辉, 邓恩泽, 陈伟, 等. 细菌 σ54 启动子序列分析与预测[J]. 电子科技大学学报, 2015, 44(1): 147-149.
DING H, DENG E Z, CHEN W, et al. The sequence analysis and prediction of σ54 promoter in bacteria[J]. Journal of University of Electronic Science and Technology of China, 2015, 44(1): 147-149.

[9] BURDEN S, LIN Y X, ZHANG R. Improving promoter prediction for the NNPP2.2 algorithm: A case study using Escherichia coli DNA sequences[J]. Bioinformatics, 2005, 21(5): 601-607.

[10] WANG H, BENHAM C J. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress[J]. Bmc Bioinformatics, 2006, 7: 248.

[11] POLAT K, GUNES S. A novel approach to estimation of E-coli promoter gene sequences: Combining feature selection and least square support vector machine (FS_LSSVM)[J]. Appl Math Comput, 2007, 190(2): 1574-1582.

[12] RANGANNAN V, BANSAL M. Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability[J]. J Biosciences, 2007, 32(5): 851-862.

[13] RANGANNAN V, BANSAL M. Relative stability of DNA as a generic criterion for promoter prediction: Whole genome annotation of microbial genomes with varying nucleotide base composition[J]. Mol Biosyst, 2009, 5(12): 1758-1769.

[14] BLAND C, NEWSOME A S, MARKOVETS A A. Promoter prediction in E-coli based on SIDD profiles and artificial neural networks[J]. Bmc Bioinformatics, 2010, 11: S17.

[15] DE J A, PIETERSMA H, CORDES M, et al. PePPER: A webserver for prediction of prokaryote promoter elements and regulons[J]. Bmc Genomics, 2012, 13: 299.

[16] DI S M, PINATEL E, TALA A, et al. G4PromFinder: An algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs[J]. Bmc Bioinformatics, 2018, 19(1): 36.

[17] PEREZ-BELLO A, MUNTEANU C R, UBEIRA F M, et al. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices[J]. J Theor Biol, 2009, 256(3): 458-466.

[18] GORDON J J, TOWSEY M W, HOGAN J M, et al. Improved prediction of bacterial transcription start sites[J]. Bioinformatics, 2006, 22(2): 142-148.

[19] TOWSEY M, TIMMS P, HOGAN J, et al. The cross-species prediction of bacterial promoters using a support vector machine[J]. Comput Biol Chem, 2008, 32(5): 359-366.

[20] RANI T S, BHAVANI S D, BAPI R S. Analysis of E. coli promoter recognition problem in dinucleotide feature space[J]. Bioinformatics, 2007, 23(5): 582-588.

[21] LI Q Z, LIN H. The recognition and prediction of sigma70 promoters in Escherichia coli K-12[J]. J Theor Biol, 2006, 242(1): 135-141.

[22] LIN H, LI Q Z. Eukaryotic and prokaryotic promoter prediction using hybrid approach[J]. Theory Biosci, 2011, 130(2): 91-100.

[23] HE W, JIA C, DUAN Y, et al. 70ProPred: A predictor for discovering sigma70 promoters based on combining multiple features[J]. BMC Syst Biol, 2018, 12(Suppl 4): 44.

[24] LAI H Y, ZHANG Z Y, SU Z D, et al. iProEP: A computational predictor for predicting promoter[J]. Mol Ther-Nucl Acids, 2019, 17: 337-346.

[25] WU Q Q, WANG J J, YAN H. An improved position weight matrix method based on an entropy measure for the recognition of prokaryotic promoters[J]. Int J Data Min Bioin, 2011, 5(1): 22-37.

[26] DE A E, 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.

[27] SONG K. Recognition of prokaryotic promoters based on a novel variable-window Z-curve method[J]. Nucleic Acids Res, 2012, 40(3): 963-971.

[28] SILVA S A, FORTE F, SARTOR I T, et al. DNA duplex stability as discriminative characteristic for Escherichia coli sigma(54)- and sigma(28)- dependent promoter sequences[J]. Biologicals, 2014, 42(1): 22-28.

[29] LIN H, DENG E Z, DING H, et al. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide

composition[J]. Nucleic Acids Res, 2014, 42(21): 12961-12972.
[30] LIU B, HAN L, LIU X, et al. Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies[J]. IEEE/ACM Trans Comput Biol Bioinform, 2019, 16(4): 1211-1218.
[31] ABBAS M M, MOHIE-ELDIN M M, EL-MANZALAWY Y. Assessing the effects of data selection and representation on the development of reliable E. coli sigma 70 promoter region predictors[J]. Plos One, 2015, 10(3): e0119721.
[32] SHAHMURADOV I A, MOHAMAD RAZALI R, BOUGOUFFA S, et al. bTSSfinder: A novel tool for the prediction of promoters in cyanobacteria and Escherichia coli[J]. Bioinformatics, 2017, 33(3): 334-340.
[33] LIN H, LIANG Z Y, TANG H, et al. Identifying sigma70 promoters with novel pseudo nucleotide composition[J]. IEEE ACM T Comput Bi, 2019, 16(4): 1316-1321.
[34] RAHMAN M S, AKTAR U, JANI M R, et al. iPromoter-FSEn: Identification of bacterial sigma(70) promoter sequences using feature subspace based ensemble classifier[J]. Genomics, 2019, 111(5): 1160-1166.
[35] RAHMAN M S, AKTAR U, JANI M R, et al. iPro70-FMWin: Identifying Sigma70 promoters using multiple windowing and minimal features[J]. Mol Genet Genomics, 2019, 294(1): 69-84.
[36] UMAROV R K, SOLOVYEV V V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks[J]. Plos One, 2017, 12(2): e0171410.
[37] WANG S, CHENG X, LI Y, et al. Image-based promoter prediction: A promoter prediction method based on evolutionarily generated patterns[J]. Sci Rep, 2018, 8(1): 17695.
[38] COPPENS L, LAVIGNE R. SAPPHIRE: A neural network based classifier for sigma70 promoter prediction in Pseudomonas[J]. Bmc Bioinformatics, 2020, 21(1): 415.
[39] LIU B, YANG F, HUANG D S, et al. iPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC[J]. Bioinformatics, 2018, 34(1): 33-40.
[40] LIU B, LI K. iPromoter-2L2.0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features[J]. Mol Ther Nucleic Acids, 2019, 18: 80-87.
[41] ZHANG M, LI F, MARQUEZ-LAGO T T, et al. MULTiPly: A novel multi-layer predictor for discovering general and specific types of promoters[J]. Bioinformatics, 2019, 35(17): 2957-2965.
[42] SHUJAAT M, WAHAB A, TAYARA H, et al. pcPromoter-CNN: A CNN-based prediction and classification of promoters[J]. Genes (Basel), 2020, 11(12): 1529.
[43] AMIN R, RAHMAN C R, AHMED S, et al. iPromoter-BnCNN: A novel branched CNN-based predictor for identifying and classifying sigma promoters[J]. Bioinformatics, 2020, 36(19): 4869-4875.
[44] LI F, CHEN J, GE Z, et al. Computational prediction and interpretation of both general and specific types of promoters in Escherichia coli by exploiting a stacked ensemble-learning framework[J]. Brief Bioinform, 2021, 22(2): 2126-2140.
[45] XIAO X, XU Z C, QIU W R, et al. IPSW(2L)-PseKNC: A two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition[J]. Genomics, 2019, 111(6): 1785-1793.
[46] LE N Q K, YAPP E K Y, NAGASUNDARAM N, et al. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext N-grams[J]. Front Bioeng Biotechnol, 2019, 7: 305.
[47] TAYARA H, TAHIR M, CHONG K T. Identification of prokaryotic promoters and their strength by integrating heterogeneous features[J]. Genomics, 2020, 112(2): 1396-1403.
[48] SANTOS-ZAVALETA A, SALGADO H, GAMA-CASTRO S, et al. RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12[J]. Nucleic Acids Research, 2019, 47(D1): D212-D220.
[49] ISHII T, YOSHIDA K, TERAI G, et al. DBTBS: A database of Bacillus subtilis promoters and transcription factors[J]. Nucleic Acids Research, 2001, 29(1): 278-280.
[50] HUANG Y, NIU B F, GAO Y, et al. CD-HIT suite: A web server for clustering and comparing biological sequences[J]. Bioinformatics, 2010, 26(5): 680-682.
[51] SU W, LIU M L, YANG Y H, et al. PPD: A manually curated database for experimentally verified prokaryotic promoters[J]. Journal of Molecular Biology, 2021, 433(11): 166860.
[52] CHEN W, LEI T Y, JIN D C, et al. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition[J]. Anal Biochem, 2014, 456: 53-60.
[53] PENG H C, LONG F H, DING C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE T Pattern Anal, 2005, 27(8): 1226-1238.
[54] ZOU Q, ZENG J, CAO L, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification[J]. Neurocomputing, 2016, 173: 346-354.
[55] CHANG C C, LIN C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.
[56] FISHER R A. The use of multiple measurements in taxonomic problems[J]. Annals of Human Genetics, 2012, 7(7): 179-188.

编　辑　刘飞阳