文章编号:1000－1638(2025)02-0166-08 DOI:10.13484/j.nmgdxxbzk.20250207

转录因子与组蛋白修饰对启动子-增强子相互作用的预测*

王云杰,李子涵,张利绒
(内蒙古大学物理科学与技术学院,呼和浩特 010021)

摘要:启动子和增强子之间的相互作用(Promoter-Enhancer interaction,PEI)与基因的转录与调控密切相关。本文以人B淋巴细胞系(GM12878)为研究对象,基于染色质环(Loop)数据库构建了启动子-增强子(Promoter-Enhancer,P-E)相互作用数据集,分析了P-E结构中149种转录因子(Transcription factor,TF)以及11种组蛋白修饰(Histone madification,HM)的相关性,筛选出与P-E结构具有较强关联的表观遗传修饰特征,并利用卷积神经网络(Convolutional neural network,CNN)和随机森林(Random forest,RF)算法预测了P-E相互作用对。结果显示RF预测的AUC值(Area under the curve)介于0.84至0.88之间,而CNN的AUC值在0.69至0.77之间,表明RF的预测性能略优于CNN。此外,仅使用TF信号作为特征的AUC值优于仅使用HM信号的情况,表明TF信号对P-E结构的识别具有更佳的效果。最后将TF和HM特征组合后,预测效果能够进一步提升,我们发现EGR1、H3K4me2、EP300等12种特征是预测PEI的重要特征。

关键词:染色质环; 启动子-增强子; 卷积神经网络; 随机森林算法
中图分类号:Q61 **文献标志码**:A

基因组的三维(3D)结构在基因的表达调控中起着关键作用,并决定着细胞的命运。真核生物细胞中的染色体高度折叠并在不同尺度下组织形成复杂的三维结构,包括染色质疆域(Chromosome territory)、活性/非活性染色质区室(A/B compartment)、拓扑关联结构域(Topologically associated domain,TAD)和染色质环(Loop)[1],其中染色质环是产生启动子-增强子远程相互作用(Promoter-Enhancer interaction,PEI)的主要机制[2-3]。启动子(Promoter)是位于基因上游区域特定的DNA片段[4],可被RNA聚合酶识别,并起始转录[5]。增强子(Enhancer)是位于基因上下游的DNA片段,是基因组上重要的顺式转录调控元件。增强子往往通过与启动子通信来激活基因的转录[6]。研究发现,loop中有一大部分是联系启动子和增强子的环。虽然,增强子与靶基因的一维线性距离很远,但由于增强子与靶基因启动子位于同一个loop,从而改变了增强子与启动子的物理距离,实现了增强子对靶基因的调控[7]。据估计,人类基因组中含有数十万个增强子,多个增强子会共调控一个启动子,单个增强子也可同时调控多个启动子。染色质空间的复杂性使得相互作用P-E对之间的距离变化很大,从千碱基到数百万碱基对不等[8-13]。同时,增强子loop结构的改变在几种病理中具有功能

* **收稿日期**:2024-05-22; **修回日期**:2024-07-08
基金项目:国家自然科学基金项目(61962041,62062053)
作者简介:王云杰(1997－),女,河北承德人,2021级硕士研究生。E-mail:2064381483@qq.com
通信作者:张利绒(1972－),女,内蒙古乌兰察布人,教授,博士。主要研究方向:生物信息学。E-mail:pyzlr@imu.edu.cn

性作用。在某些神经系统和免疫系统疾病中，如帕金森、阿尔茨海默、类风湿关节炎和系统性红斑狼疮等，发现了某些异常的增强子 loop 结构，从而引发病理生理过程[14]。因此，正确预测增强子与启动子相互作用，识别增强子与靶基因之间的调控网络具有重要的生物学意义。

高通量测序技术（如 Hi-c）能够有效地识别 PEI[15-16]，不足是耗时长且价格昂贵，亟须应用机器学习方法预测 PEI。Singh 等[17]提出 SPEID 方法，基于 DNA 序列应用卷积神经网络（Convolutional neural network，CNN）来预测 PEI，结果表明 SPEID 模型能够在全基因组范围内稳定而可靠地预测 PEI。随后，Zhuang 等[18]提出一种新的方法，使用 Singh 构建的 PEI 数据预训练 CNN，然后对目标细胞系继续训练 CNN，提高了 PEI 的预测性能。尽管仅使用 DNA 序列就能取得较好的预测结果，但关于 PEI 的表观特征分析和预测仍具有挑战。因此基于多种转录因子（Transcription factor，TF）和组蛋白修饰（Histone modification，HM）信号，提出新的数据处理和分析方法，运用机器学习算法成功预测 PEI，则具有实际的价值。

本文以人 B 淋巴细胞系（GM12878）作为研究对象，将基因的启动子区域和增强子区域分别与 loop 结构进行位点匹配，构建了 P-E 相互作用数据集。然后，分析了 P-E 结构中 149 种转录因子与 11 种组蛋白修饰之间的相关性，筛选出与 PEI 存在显著关联的表观特征。最后，运用 CNN 和随机森林（Random forest，RF）两种机器学习算法，对 PEI 进行了预测。

1　材料和方法

1.1　数据来源

本文选取的基因注释版本号为 Feb. 2009（hg19/CRCh37），下载自 UCSC Genome Browser（http://genome.ucsc.edu/cgi-bin/hgTables）。我们从 ENCODE 数据库中下载了 GM12878 细胞系 149 种转录因子和 11 种组蛋白修饰的 narrow peak 格式的 ChIP-seq 数据。另外，在 3D Genome Browser 数据库（http://3dgenome.org）中获取了该细胞系的 17228 个 loop 结构数据，在增强子数据库EnhancerAtlas 2.0 中下载了 49672 个增强子数据。最后，从 R 包 Homo.sapiens 中获取了 23056 个基因的转录起始位点（Transcriptional start site，TSS）。

1.2　PEI 数据集的建立

首先，根据基因注释定义启动子区域为基因 TSS 上下游 2 kb 区域，增强子区域为增强子中点上下游 2 kb 区域。利用 bedtools 软件，将 23056 个基因的启动子区域、49672 个增强子区域分别与 17228 个 loop 结构进行位点匹配，得到 3815 个具有 loop 结构的 PEI，其中包括多个增强子调控一个启动子、单个增强子调控多个启动子和一个增强子调控一个启动子 3 种情况，这些 PEI 构成了正集（表 1）。然后，删除正集中已包含的 2204 启动子和 2874 增强子，并将剩余的 20596 个 TSS 和 46798 个增强子在同一染色体中两两匹配，共得到 1665358 个 P-E 对。接着，我们将基因 TSS 到增强子中点之间的间隔定义为 P-E 对的距离。最后，在保证正负集 P-E 对具有相同距离分布的前提条件下，随机筛选出 19075 个非 loop 结构的 P-E 对作为负集（表 1）。

表 1　正负集中的 P-E 对个数

Table 1　Number of P-E pairs in positive and negative sets

数据集	启动子	增强子	P-E 对	数据集	启动子	增强子	P-E 对
正集	2204	2874	3815	负集	11570	11405	19075

1.3 特征参数的选取

对 PEI 中的启动子和增强子区域,我们分析了转录因子结合和组蛋白修饰的特征。首先,利用 GM12878 细胞系 149 种转录因子以及 11 种组蛋白修饰的 ChIP-seq 数据,对于任意一个 P-E 对的启动子和增强子区域,分别分为 20 个 bin,每个 bin 宽为 200 bp,应用 bedtools 中的 multicov 计算落入每个 bin 内表观信号的 count 数,最终在启动子和增强子区域各得到 20 维的表观修饰信号矢量。根据 Ouyang 等[19]提出的转录因子与靶基因的关联强度(TF associations strength,TFAS)得分定义,设第 i 个 bin 中 TF 或 HM 的强度为 x_i,第 i 个 bin 中点到启动子或增强子中点的距离为 d_i,则启动子或增强子表观信号的 TFAS 得分 A_i 定义为:

$$A_i = \sum\nolimits_i x_i \mathrm{e}^{\frac{-d_i}{d_0}},$$

式中,d_0 为常数,取 d_0 为 1000。

根据 Spearman 相关系数定义,利用关联强度计算了正负集中 P-E 对中各表观信号之间的相关系数 r,用于构建 PEI 相互作用网络。相关系数 r 是反映变量之间相关关系密切程度的统计指标,$-1 \leqslant r \leqslant 1$,$|r|$ 越趋近 1,表示 P-E 相关性越强,$r=0$,说明二者不相关。

1.4 预测方法及效果的评估

卷积神经网络(CNN)是一种主要用于图像和视觉任务的深度学习模型,典型结构包括输入层、卷积层、激活函数、池化层、全连接层和输出层。在训练过程中,CNN 通过反向传播算法自动学习卷积核的权重参数,从而使网络能够在输入图像上进行有效特征提取和分类。随机森林(RF)是一种集成学习(Ensemble learning)算法,它通过构建多个决策树(Decision tree)来提高预测性能。本文采用三折法交叉验证预测结果,并利用敏感性(Sn)、特异性(Sp)、准确度(ACC)、Precision-Recall(PR)、Area under curve(AUC)5 个参数来衡量预测的精度。其中,AUC 是受试者工作特征(Receiver operating characteristic,ROC)曲线下方的面积,ROC 曲线可以直观地看出模型预测的效果,曲线越靠近左上角,说明模型预测效果越好。相关公式如下:

$$Sn = \frac{TP}{TP + FN},$$

$$Sp = \frac{TN}{TN + FP},$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}。$$

式中,TP 为正集预测正确的个数;TN 为负集的个数;FN 为被判断为负集实际上是正集的个数;FP 为被判断为正集实际上是负集的个数。

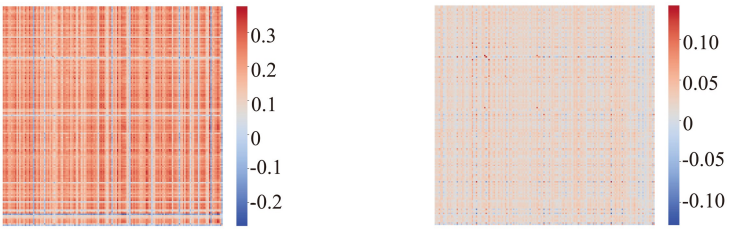
2 结果

2.1 P-E 对中 TF 和 HM 的关联强度

根据 P-E 对中 TF 或 HM 的关联强度得分,我们进一步分析了正负集合 P-E 对中 TF 或 HM 之间的相关性,分析了各 P-E 对之间表观修饰的内在联系。

首先,对 149 种 TF 与 11 种 HM,分别计算了它们在启动子和增强子区域的关联强度得分。其次,通过计算启动子和增强子之间每一对表观修饰信号关联强度得分之间的 Spearman 相关系数,在正负集中分别生成了一个 160×160 的相关性矩阵。该矩阵中的每个元素代表了启动子区域的一个 TF/HM 信号与增强子区域的另一个 TF/HM 之间相关程度,正负集合中 P-E 对的 TF/HM 相关性如图 1 所示。

图 2(a)为正集 P-E 对中 TF/HM 之间的相关性，图 2(b)为负集 P-E 对中 TF/HM 之间的相关性。从图中可以看到，正集中有 loop 结构支持的 P-E 对之间的相关系数有半数达到 0.2～0.3 之间，而负集中随机组合的 P-E 对之间的相关系数大部分接近 0，表明在实际发生的 PEI 之间，TF 与 HM 之间存在着较显著的相关性。以上结果表明，TF 和 HM 在正负集中的相关性存在差异，有助于 PEI 的识别和预测。



(a) 正集　　(b) 负集

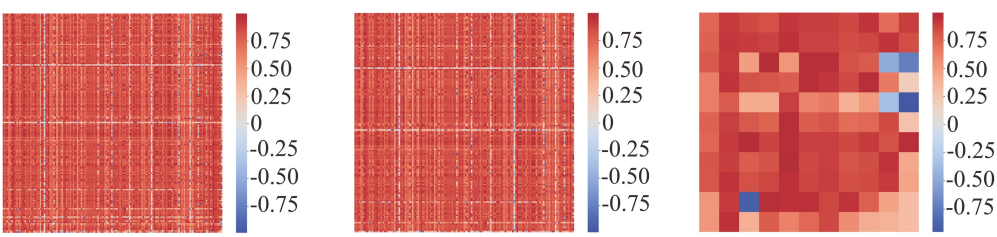
图 1　P-E 对中 TF/HM 信号的相关性

Fig. 1　Correlation of TF/HM signals in P-E pairs

为了比较正负集合中 P-E 对之间 TF/HM 之间的相关性差异，我们定义

$$R=\frac{|r|-|r'|}{|r|+|r'|},$$

式中，r 为正集中的相关系数；r'为负集中的相关系数；R 为相互作用与非相互作用 P-E 对之间的相关性差异。图 2(a)是所有 TF 和 HM 的相关性差异热图，图 2(b)是 TF 的相关性差异热图，图 2(c)是 HM 的相关性差异热图。由图 2 可知，正负集之间 P-E 对 TF/HM 信号的相关性差异分布在－1～1之间，热图中大部分的颜色几乎表现为红色，一半以上差异趋近于 1 和－1，这表明转录因子信号或组蛋白修饰信号作为特征时，PEI 对于基因表达的影响有显著差别。



(a) TF 和 HM　　(b) TF　　(c) HM

图 2　正负集之间 P-E 对 TF/HM 信号的相关性差异

Fig. 2　Correlation difference of TF/HM signals in P-E pairs between the positive and negative sets

2.2　特征筛选

通过对比正负集 P-E 对之间 TF 和 HM 相关性的差值，我们筛选差异显著的表观信号作为识别 PEI 的特征。

(1)差异值最大的 12 组 TF 对和 HM 对，如表 2 所示。

(2)差异值 R 趋近于－1 的 3 对 TF 和趋近于 1 的 3 对 TF，如表 3 所示。

(3)差异值 R 趋近于－1 的 3 对 HM 和趋近于 1 的 3 对 HM，如表 4 所示。

2.3　预测结果

采用三折交叉验证法，利用 CNN 和 RF 两种预测方法，以启动子和增强子区域 4 kb 中 20 个 bin 的信号作为输入，对 P-E 对进行了预测，并利用 Sn、Sp、ACC、PR 和 AUC 这 5 个参数来衡量预测的精度，结果如表 5 所示。

表 2 正负集之间相关性差异显著的前 12 组 TF/HM(TF/HM-Top 12)

Table 2 Top 12 TFs/HMs with significant correlation differences between the positive and negative sets(TF/HM-Top 12)

编号	*R* 趋近于 1			*R* 趋近于−1		
	信号 1	信号 2	差值 *R*	信号 1	信号 2	差值 *R*
1	EGR1	ZFP36	1	SIX5	NR2C2	−0.99147
2	H3K4me2	EP300	0.999962	H3K27me3	ZZZ3	−0.9873
3	MTA3	PAX8	0.999909	JUND	EGR1	−0.98722
4	JUNB	H3K9ac	0.999874	HDGF	SMC3	−0.98408
5	DPF2	CUX1	0.999854	NR2C2	E2F4	−0.98042
6	RBBP5	IKZF1	0.999828	LARP7	NR2C2	−0.97889
7	H3K36me3	LARP7	0.999808	CREB1	USF1	−0.97761
8	TAF1	JUNB	0.999775	H3K4me1	H4K20me1	−0.95246
9	FOXM1	NFYB	0.999712	NBN	EGR1	−0.94979
10	FOS	ELF1	0.999642	EGR1	EZH2	−0.94736
11	NR2F1	CEBPZ	0.999555	RBBP5	MYC	−0.94515
12	RAD51	POLR2AphosphoS5	0.999467	FOXK2	USF1	−0.94117

表 3 正负集之间相关性差异显著的前 3 组 TF 对(TF-Top 3)

Table 3 Top 3 TFs with significant correlation differences between the positive and negative sets (TF-Top 3)

编号	*R* 趋近于−1			*R* 趋近于 1		
	TF1	TF2	差值 *R*	TF1	TF2	差值 *R*
1	SIX5	NR2C2	−0.991	EGR1	ZFP36	1
2	JUND	EGR1	−0.987	MTA3	PAX8	1
3	HDGF	SMC3	−0.984	DPF2	CUX1	1

表 4 正负集之间相关性差异显著的前 3 组 HM 对(HM-Top 3)

Table 4 Top 3 HMs with significant correlation differences between the positive and negative sets (HM-Top 3)

编号	*R* 趋近于−1			*R* 趋近于 1		
	HM1	HM2	差值 *R*	HM1	HM2	差值 *R*
1	H3K4me1	H4K20me1	−0.952	H3K79me2	H3K4me1	0.994
2	H3K9me3	H3K27me3	−0.911	H3K4me3	H3K27me3	0.993
3	H3K27me3	H4K20me1	−0.723	H3K27me3	H3K36me3	0.986

由表 5 的预测结果可见,RF 预测的 PR 值在不同特征输入时在 0.668 到 0.680 之间,没有明显的提升或下降趋势; CNN 预测的 PR 值随着特征数量的增加,从 0.740 逐步下降至 0.709,但 Top9 特征和 Top12 特征的预测结果一致,说明 RF 的预测性能相对稳定。RF 预测的 AUC 值介于 0.84 至 0.88 之间,而 CNN 预测的 AUC 值在 0.69 至0.77之间,表明 RF 的预测性能略优于 CNN。另外,从 Top3 到 Top12 的预测结果观察到,随着特征信号数量的增加,AUC 值也呈上升趋势,表明特征数量与预测效果呈正相关关系。然而,从 Top6 到 Top12 的特征信号中,AUC 的增长速度放缓,这表明

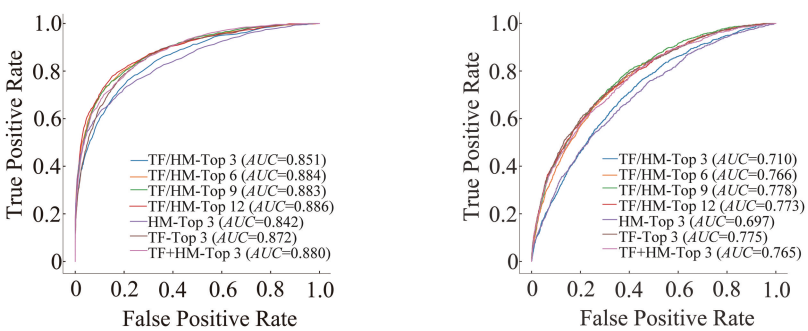
存在一定程度的特征冗余。此外，单独引入 TF 信号作为特征时，AUC 值优于单独引入 HM 信号的情况，表明 TF 信号对 P-E 结构的识别具有更佳的效果。将 TF 和 HM 特征组合后，预测效果进一步提升。最后，我们发现 EGR1、H3K4me2、EP300 等 12 种特征是预测 PEI 的重要特征。Zheng 等[20]在同一细胞系中的研究结果也表明，仅利用 H3K27ac、ATAC-seq、RAD21 和距离这 4 个特征时，RF 方法的 PR 值达到了0.7，与我们的结果一致。

表 5　基于七个特征组合的预测结果

Table 5　Prediction results based on seven feature combinations

特征	RF					CNN				
	Sn	*Sp*	*ACC*	*PR*	*AUC*	*Sn*	*Sp*	*ACC*	*PR*	*AUC*
TF/HM-Top3	0.786	0.761	0.774	0.679	0.851	0.631	0.684	0.658	0.740	0.710
TF/HM-Top6	0.818	0.790	0.804	0.668	0.884	0.743	0.652	0.698	0.715	0.766
TF/HM-Top9	0.874	0.736	0.805	0.669	0.883	0.689	0.722	0.705	0.709	0.778
TF/HM-Top12	0.848	0.779	0.814	0.668	0.886	0.746	0.659	0.702	0.709	0.773
HM-Top3	0.891	0.638	0.764	0.680	0.842	0.636	0.648	0.642	0.741	0.697
TF-Top3	0.813	0.774	0.793	0.673	0.872	0.727	0.678	0.702	0.708	0.775
TF＋HM-Top3	0.883	0.700	0.792	0.670	0.880	0.739	0.645	0.692	0.712	0.765

如图 3 所示，对比 CNN 与 RF 预测的 ROC 曲线，可进一步发现两种算法在不同假阳性率(FPR)条件下对真阳性率(TPR)的区分能力。



(a) RF　　(b) CNN

图 3　两种方法预测的 ROC 曲线

Fig. 3　ROC curves of two methods

图 3(a)绘制了使用 RF 算法预测正负样本时所得到的 ROC 曲线，而图 3(b)则展示了 CNN 算法所得出的 ROC 曲线。从这两组曲线的对比可以发现，RF 的 ROC 曲线的 AUC 值整体相较于 CNN 曲线的更大，直观地验证了 RF 的分类性能优于 CNN。

3　讨论

启动子-增强子相互作用是一个复杂的过程，对基因的表达和调控有重要的影响，基因组序列特征和表观修饰特征都与 PEI 有关，尽管已经开发了一些深度学习方法来预测 PEI，但特征的选取和跨细胞系问题仍然是一项很大的挑战。

基于 loop 结构数据，我们已经成功地划分出相互作用和非相互作用的 P-E 对数据集。在此基础

之上,针对 149 种转录因子以及 11 种组蛋白修饰信号进行了深入分析,计算了它们在正负集P-E对间的关联强度得分,以揭示 TF 与 HM 的潜在相互作用模式。研究发现,在正集的 P-E 对中,TF/HM 间展现出较强的关联性,而在负集的 P-E 对中,这种关联性几乎不存在。通过比较正负集 P-E 对间 TF 与 HM 相关性的差异,我们筛选出了具有显著区分能力、可用于识别 PEI 的表观遗传修饰信号。

结果显示,采用 RF 算法进行预测时,得到的 AUC 值区间为 0.84 至 0.88,表现出较高的预测性能;相比之下,使用 CNN 模型的 AUC 值则位于 0.69 至 0.77 之间,说明在本次 P-E 对预测中,RF 的预测性能略优于 CNN。此外,通过对 Top3 至 Top12 的预测结果进行比较,我们注意到随着特征信号数量的增加,AUC 值呈现同步上升的趋势,证明了特征数量与预测效果之间存在正相关联系。然而,在 Top6 至 Top12 的特征信号区间内,AUC 增长速率有所减缓,暗示着特征信号存在一定程度的特征冗余问题。同时,研究还揭示了单独引入 TF 信号作为预测特征时,其 AUC 值表现优于仅引入 HM 信号的情况,表明 TF 信号对 P-E 结构的识别具有更佳的效果。但当我们将 TF 和 HM 两种特征信号有效结合后,预测性能得到了更进一步的提升。

本研究利用 GM12878 细胞系的数据来预测 PEI,尚未拓展至其他细胞系进行同类预测分析。然而,为了深化对基因表达及调控机制的理解,未来开展针对其他细胞系中 PEI 的预测分析至关重要。这一步骤将有助于我们在更广泛的生物学背景下探究基因表达调控的复杂性和多样性,从而推动相关领域的理论研究与实践应用进展。

参考文献:

[1] LI R F,LIU Y T,HOU Y P,et al. 3D genome and its disorganization in diseases[J]. Cell Biology and Toxicology,2018,34(5):351-365.

[2] RUBTSOV M A,POLIKANOV Y S,BONDARENKO V A,et al. Chromatin structure can strongly facilitate enhancer action over a distance[J]. Proceedings of the National Academy of Sciences of the United States of America,2006,103(47):17690-17695.

[3] MIELE A,DEKKER J. Long-range chromosomal interactions and gene regulation[J]. Molecular BioSystems,2008,4(11):1046-1057.

[4] FULLWOOD M J,RUAN Y J. ChIP-based methods for the identification of long-range chromatin interactions[J]. Journal of Cellular Biochemistry,2009,107(1):30-39.

[5] WHALEN S,TRUTY R M,POLLARD K S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin[J]. Nature Genetics,2016,48(5):488-496.

[6] MARSMAN J,HORSFIELD J A. Long distance relationships:Enhancer-promoter communication and dynamic gene transcription[J]. Biochimica et Biophysica Acta,2012,1819(11/12):1217-1227.

[7] SCHOENFELDER S,FRASER P. Long-range enhancer-promoter contacts in gene expression control[J]. Nature Reviews Genetics,2019,20(8):437-455.

[8] ROWLEY M J,CORCES V G. The three-dimensional genome:Principles and roles of long-distance interactions[J]. Current Opinion in Cell Biology,2016,40:8-14.

[9] DEKKER J,MIRNY L. The 3D genome as moderator of chromosomal communication[J]. Cell,2016,164(6):1110-1121.

[10] BICKMORE W A,VAN STEENSEL B. Genome architecture:Domain organization of interphase chromosomes[J]. Cell,2013,152(6):1270-1284.

[11] VAN STEENSEL B,DEKKER J. Genomics tools for unraveling chromosome architecture[J]. Nature Biotechnology,2010,28(10):1089-1095.

[12] VISEL A,RUBIN E M,PENNACCHIO L A. Genomic views of distant-acting enhancers[J]. Nature,2009,461(7261):199-205.

[13] JING F, ZHANG S W, ZHANG S H. Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network[J]. BMC Bioinformatics, 2020, 21(1): 507.

[14] SMITH E, SHILATIFARD A. Enhancer biology and enhanceropathies[J]. Nature Structural & Molecular Biology, 2014, 21(3): 210-219.

[15] DOSTIE J, RICHMOND T A, ARNAOUT R A, et al. Chromosome conformation capture carbon copy (5C): A massively parallel solution for mapping interactions between genomic elements[J]. Genome Research, 2006, 16(10): 1299-1309.

[16] SPLINTER E, DE WIT E, VAN DE WERKEN H J G, et al. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation[J]. Methods, 2012, 58(3): 221-230.

[17] SINGH S, YANG Y, PÓCZOS B, et al. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks[J]. Quantitative Biology, 2019, 7(2): 122-137.

[18] ZHUANG Z, SHEN X T, PAN W. A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data[J]. Bioinformatics, 2019, 35(17): 2899-2906.

[19] OUYANG Z, ZHOU Q, WONG W H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells[J]. Proceedings of the National Academy of Sciences, 2009, 106(51): 21521-21526.

[20] ZHENG L Q, LIU L, ZHU W, et al. Predicting enhancer-promoter interaction based on epigenomic signals[J]. Frontiers in Genetics, 2023, 14: 1133775.

（责任编委　刘俊杰）

Prediction of Promoter-Enhancer Interaction Based on Transcription Factors and Histone Modifications

WANG Yunjie, LI Zihan, ZHANG Lirong

(*School of Physical Science and Technology, Inner Mongolia University, Hohhot* 010021, *China*)

Abstract: The interaction between promoter and enhancer (PEI) is closely related to gene expression and regulation. In human B-lymphoid (GM12878) cell line, a P-E interaction dataset was constructed based on the chromatin loop database, and the correlation of 149 transcription factors and 11 histone modifications in the P-E structure was analyzed. Since epigenetic signals strongly associated with the P-E structure, the interacting P-E pairs can be identified using the Convolutional Neural Network (CNN) and Random Forest (RF) algorithms. The results showed that the AUC (Area Under the ROC Curve) values by RF ranged from 0.84 to 0.88, while the AUC values by CNN ranged from 0.69 to 0.77, indicating that the performance of the RF method is slightly better than the CNN's. In addition, the AUC values with only TF signals as input, are better than only HM signals as input, indicating that TF signals have better predictive effects on the recognition of P-E interactions. Combining TF and HM features, the prediction effects can be further improved. As a result, we found 12 features, including EGR1, H3K4me2 and EP300, are important for PEI structures.

Key words: chromatin loop; promoter-enhancer; convolution neural netwook; random forest algorithm