计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100027

基于多特征融合的增强子-启动子相互作用预测综述

胡宇佳　甘　伟　朱　敏
四川大学计算机学院　成都 610065
(543574831@qq.com)

摘　要　研究增强子-启动子相互作用机理有助于人们理解基因调控关系,进而揭示与疾病相关的基因,为疾病诊疗提供新思路和新方法。传统的生物检测方法的实验成本高、耗时长,且受分辨率的限制,难以精确鉴定单个增强子-启动子的相互作用。通过计算方法来解决生物问题已成为近年来的研究热点,此类方法可以通过复杂的网络结构主动学习序列特征和空间结构,进而准确预测增强子-启动子的作用。首先介绍了传统生物实验检测方法的研究现状;然后从序列特征的角度出发,围绕多特征融合的基本思想,对统计学和深度学习方法在增强子-启动子相互作用预测上的应用进行归纳整理;最后对该领域的研究热点和挑战进行总结分析。

关键词:增强子-启动子相互作用;多特征融合;序列特征;应用综述;疾病诊疗
中图法分类号　TP391

Enhancer-Promoter Interaction Prediction Based on Multi-feature Fusion

HU Yu-jia,GAN Wei and ZHU Min
College of Computer Science,Sichuan University,Chengdu 610065,China

Abstract　The study of the mechanism of Enhancer-Promoter Interaction is helpful to understand gene regulations,thus revealing specific genes that are relevant to diseases as well as providing new clinical methods and ideas for disease diagnosis and treatment. Compared to traditional biological analysis methods which are always more expensive,time-consuming and more difficult to precisely identify specific interactions due to limited resolution,computational methods to solve biological problems have become a hot research topic in recent years. This method can actively learn sequence features and spatial structures through complex network structures,so as to precisely and accurately predict the interactions of enhancers and promoters. This paper firstly introduces the research status of traditional biological detection methods. Then,from the perspective of sequence features,the application of statistics and deep learning method in the prediction of enhancer - promoter interaction is summarized and sorted out based on the basic idea of multi-feature fusion. Finally,the research hotspots and challenges in this field are summarized and analyzed.

Keywords　Enhancer-promoter interaction,Multi-feature fusion,Sequence feature,Application overview,Disease diagnosis and treatment

1　引言

增强子和启动子是DNA分子中具有转录调节功能的特异序列,它们通过与转录因子相结合来调控基因转录的精确起始位点和转录效率[1]。启动子一般位于转录起始位点上游,用于启动下游基因表达[2]。增强子是一段50～1500bp的非编码DNA序列[3],能够提高基因转录频率。增强子-启动子相互作用(Enhancer-Promoter Interactions,EPIs)是指增强子与活性蛋白相结合,协同靶启动子以驱动组织特异性基因表达的过程[4-5]。

研究EPIs的机理有助于人们理解基因调控关系,进而揭示与疾病相关的基因[6]。Davision等证明,EPIs作为一种媒介,会导致Ⅰ型糖尿病、多发性硬化症的发生,以此为基础能预测与疾病相关的新基因[7];Smeno等分别在小鼠和人类中发现了FTO基因的第一内含子区域,且发现同源基因Irx3存在远程EPIs。其中,第一内含子区域富含与肥胖相关的单核苷酸多态性(Single Nucleotide Polymorphism,SNPs);而Irx3基因是人类脑、心、肺中高表达的一个转录因子,对控制体重非常重要[8]。因此,研究EPIs,尤其是跨越不同细胞系的EPIs,对理解精准基因表达调控、细胞分化以及疾病机理有着重要作用,最终可为疾病诊疗、生物医学以及药物研发等领域提供新方法和新思路。

EPIs远比想象中复杂:增强子可以在距离靶启动子1Mbp外的转录起始位点上游或下游发挥作用[9];一个增强子

到稿日期:2019-11-05　返修日期:2020-03-25　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:"十三五"国家科技重大专项(2018ZX10201002-002-004)
This work was supported by the National Major Scientific and Technologic Project During the Thirtieth Five-Year Plan (2018ZX10201002-002-004).
通信作者:朱敏(zhumin@scu.edu.cn)

可以作用于一个或多个靶启动子；多个增强子也可以协同调控靶启动子；有些增强子甚至能在不同组织中作用于相同的靶启动子；而更多的 EPIs 则表现出了很强的细胞特异性，由多个增强子协同控制的启动子具有更高的组织特异性[10]。增强子与启动子通过复杂的空间结构[11]相互靠近并发生相互作用。

传统的生物实验方法成本高、耗时长，且数据分辨率低，难以精确鉴定单个 EPIs。利用统计学方法和深度学习方法从数据中学习事先未被定义的功能关系[12-14]，这类方法也被广泛应用于计算生物学，尤其是序列预测问题。它无需事先对结果做出任何假设，直接从样本中学习特征，就可获得准确性较高的预测模型[15]。

因此，本文首先阐述传统的 EPIs 实验检测方法；然后简单介绍了一种基于组学特征预测 EPIs 的典型方法；接着从序列特征出发，考虑序列的语义特征和空间特征，围绕多特征融合的基本思想，对统计学方法以及深度学习方法在 EPIs 预测上的应用进行总结分析；最后对当前领域的研究热点及挑战进行归纳。

2 传统 EPIs 检测方法

2.1 荧光原位杂交技术

荧光原位杂交技术（Fluorescence In Situ Hybridization，FISH）[16]通过标记特定序列的荧光探针来识别邻近的区域[17]。该方法经济安全、实验周期短、探针稳定性高，能在短时间内得到结果；但只能在有限数量的 DNA 位点使用，且该方法为非高通量方法，分辨率有限。

2.2 染色体构象捕获技术及其衍生技术

染色体构象捕获技术（Chromosome Conformation Capture，3C）及其衍生技术 4C（Chromosome Conformation Capture-on-Chip）和 5C（Chromosome Conformation Capture Carbon Copy）[18]是用于分析细胞中染色质空间组织的一组分子生物学方法，量化了在三维空间中靠近的基因之间的相互作用数量[19]。3C 方法[20]用于量化一对基因组基因座之间的相互作用；4C 方法[21]用于捕获一个基因座和所有其他基因组基因座之间的作用；5C 方法[22]用于检测给定区域内所有限制片段间的相互作用。这些技术通过高通量方法检测了三维空间中的物理接触，但未能捕获全基因组的复杂相互作用关系。

高通量染色体构象捕获技术（High-Throughput Chromosome Conformation Capture，Hi-C）[23]通过高通量测序方法找到核苷酸序列片段；通过配对末端测序方法，从连接片段末端检索出短序列[24]；用于检测片段之间所有可能的成对相互作用。该方法允许在全基因组范围内检测，但分辨率不高（大于 5 kb），难以捕获单个 EPIs。

2.3 配对末端标签测序分析技术

配对末端标签测序分析染色质相互作用技术（Chromatin Interaction Analysis using Paired End Tag Sequencing，ChIA-PET）[25]结合了染色质免疫沉淀富集、染色质邻近选择、成对末端标签检测及高通量测序技术，用于识别全基因组的长距离染色质相互作用[26]；但其仅能检测目标蛋白介导的相互作用。

2.4 生物学方法总结

总体来说，这些传统的生物实验方法成本高、耗时长，且由于分辨率的限制，检测到的数据会存在很多不相关因素，因此难以捕获准确的单个 EPIs。

3 EPIs 预测方法

本节首先描述该领域使用的数据集、数据预处理方式以及序列 motif 的相关概念；然后从数据处理、特征提取及融合和分类预测 3 个角度来归纳现有的 EPIs 预测方法；最后对这些方法进行总结。

3.1 基本介绍

3.1.1 数据集

本文涉及的数据有以下 6 种：红白血病细胞（K562）、人类 B 淋巴细胞（GM12878）、宫颈癌细胞（HeLa-S3）、人脐静脉表皮细胞（HUVEC）、人表皮角质（NHEK）和人胚肺成纤维细胞（IMR90）。这 6 种数据集的详细信息如表 1 所列。其中，增强子和启动子数据通过以下几种方式得到：从 DNA 元素百科全书[27]中获取，经过染色质状态发现[28]，以及表观基因组图谱联盟[29]筛选得到；正负样本数据通过 Hi-C 技术检测得到，仅考虑远端 EPIs。上述 6 个细胞系的数据来源准确，目前被认定为基准数据集，被广泛用于各类实验中来进行模型性能的比较。

表 1 常用 EPIs 数据集

Table 1 EPIs datasets

Dataset	enhancer	promoter	true EPIs	false EPIs
K562	82 806	8 196	1 977	39 500
GM12878	100 036	8 453	2 113	42 200
HeLa-S3	103 460	7 794	1 740	34 800
HUVEC	65 358	8 180	1 524	30 400
NHEK	144 302	5 254	1 291	25 600
IMR90	108 996	5 253	1 254	25 000

3.1.2 数据预处理

DNA 序列由腺嘌呤（Alanine，A）、胞嘧啶（Cytosine，C）、鸟嘌呤（Guanine，G）和胸腺嘧啶（Thymine，T）组成。数据预处理的目标是将 DNA 序列转化为模型能够理解的语言。本文主要介绍 one-hot 编码和 k-mer 分析。

1）one-hot 编码

One-hot 编码[30]的思想是有多少种状态就用多少位表示，每一位都只有 0 或 1 两种取值。设 $S=s_1 s_2 \cdots s_j \cdots s_m$ 是长度为 m 的 DNA 序列，其中 $s_j=\{A,G,C,T\}$，通过 one-hot 编码将其转换为 $4*m$ 的矩阵 $\boldsymbol{S}$，其中行数 i 和列数 j 满足 $1\leqslant i\leqslant 4, 1\leqslant j>m$，矩阵中的每个值 $S_{i,j}$ 是根据 S_j 位点上碱基的不同取 0 或者 1。

$$S_{i,j}=\begin{cases}1, & S_j=\{A,G,C,T\}_i\\ 0, & \text{other}\end{cases} \tag{1}$$

该编码方式可以将一维的序列转化为二维的矩阵。以 AGCTTTAC 为例，编码结果如图 1 所示。

$$\begin{array}{c}A\\G\\C\\T\end{array}\begin{bmatrix}1&0&0&0&0&0&1&0\\0&1&0&0&0&0&0&0\\0&0&1&0&0&0&0&1\\0&0&0&1&1&1&0&0\end{bmatrix}$$

图 1 DNA 序列的 one-hot 编码示意图

Fig. 1 One-hot encoding of DNA sequence

2)k-mer 分析

设 S 是长度为 m 的 DNA 序列，$S=S_1S_2\cdots S_m$，其中 $S_i=\{A,T,C,G\}$。一个长度为 k 的子串是指从序列 S 中的任意位置 i 开始的 k 个连续符号，称为 k-mer。基因组序列中的 k-mer 存在一定的规律，利用 k-mer 组装 DNA 序列，可以提高基因的异源性表达[31]，还可以识别基因组样品中的特定品种[32-34]。图 2 为 k-mer 示意图。

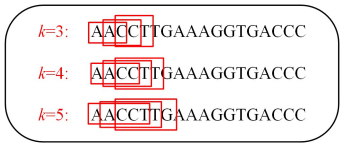

图 2 DNA 序列 k-mer 分析示意图

Fig. 2 Analysis diagram of k-mer in DNA sequence

3.1.3 基因序列模体

基因序列模体(motif)是指序列中的局部保守区域，是核苷酸或氨基酸的一种短序列模式。它常与特定功能联系，是无偏见、通用且完整的基于数学统计的序列特征集。k-mer 结果可视为 motif。研究人员可以通过模型参数提取出不同的特征映射，挖掘出相关的 motifs，如图 3 所示。

图 3 模序 motif 示意图

Fig. 3 Illustration of motif

3.2 多特征融合的 EPIs 预测方法

为了突破实验检测方法的局限性，研究者尝试使用统计学方法以及深度学习手段来解决生物问题。本文对 EPIs 预测方法进行归纳总结，得到如图 4 所示的基于特征融合的 EPIs 预测框架。

将原始序列数据作为输入，经过数据处理、特征提取及融合和分类器预测 3 个模块之后得到最终的预测结果。

1)数据处理模块：根据模型选择不同的数据处理方式。数据的处理方式不同，特征提取的方式也不同。该模块得到的是每个增强子和启动子的独立表示。

2)特征提取及融合模块：特征提取模块参考数据的编码形式，选择模型提取序列特征；特征融合模块则是将增强子和启动子的特征融合成一个整体。

3)分类器预测模块：将特征融合之后的数据作为输入，学习正负样本序列的特征，通过迭代学习调整模型参数，构建分类器。

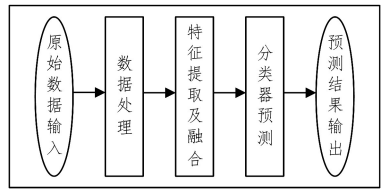

图 4 基于特征融合的 EPIs 预测框架

Fig. 4 Feature fusion based EPIs prediction framework

3.2.1 方法介绍

TargetFinder[35]利用生物实验中富集的大量细胞系特异性信息，如组蛋白修饰、表观遗传修饰、转录因子结合位点、染色质开放性[36]和基因表达等实验检测数据来识别 EPIs。该方法证明，发生相互作用的转录起始位点会有 Pol II 和 H3K4me3 富集，同时还会有大量的 CTCF 和 RAD21 富集在启动子附近，结构蛋白也会在增强子周围富集。这样一来，通过检测细胞区域内特征的富集程度即可判断是否发生相互作用。但该方法的数据主要来源于实验，成本高且耗时长；同时，细胞系的特征受到了有限实验数据的限制。

为了打破 TargetFinder 方法的局限性，Lee 提出了一种判别性计算框架 kmer-SVM[37]，其可以单独检测不依赖于保守性或已知转录因子结合特异性的 DNA 序列中的增强子。该方法使用支持向量机(Support Vector Machine，SVM)，并以 DNA 序列为特征，旨在找到一个决策边界，该边界能够最大程度地区分增强子数据(正样本)和随机基因组(负样本)数据。作者首先对 DNA 序列进行 k-mer 分析，找出与增强子关联的 motif；然后将这些 motif 输入 SVM，得到分类结果。kmer-SVM 方法的示意图如图 5 所示。

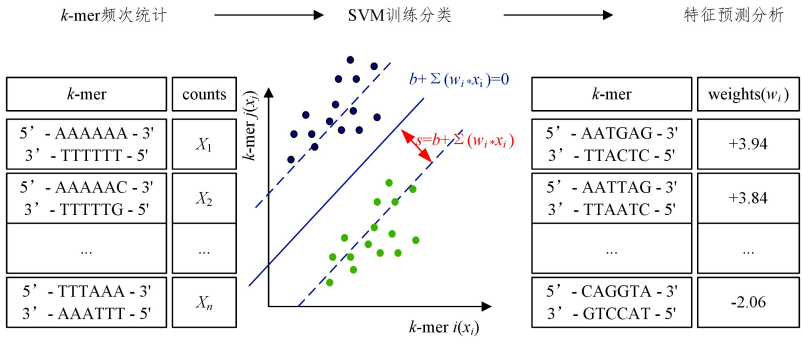

图 5 kmer-SVM 方法概览

Fig. 5 Workflow of kmer-SVM

k-mer 是基于数据统计模型得到的无偏差特征集合。k 值较大(序列较长)时，每种序列出现的频次低，难以表现特征的重要程度。因此，Ghandi 等在此基础上提出了 gkmSVM[38]，并引入带间隙的 k-mer(gapped k-mer，gkm)编码方法。SVM 核函数的计算方法如下：

$$K(S_1,S_2)=\frac{\langle f^{S_1},f^{S_2}\rangle}{\|f^{S_1}\|\|f^{S_2}\|} \tag{2}$$

其中，S_1 和 S_2 为两序列。给定序列 S 可以用 $f^S=[y_1^S,$

$y_2^S,\cdots,y_M^S]^T$ 表示，其中 M 是带间隙的 k-mer 的个数，y_i^S 是每个带间隙的 k-mer 在序列 S 中出现的频次。其他计算公式如式(3)和式(4)所示：

$$\langle f^{S_1},f^{S_2}\rangle=\sum_{m=0}^{l}N_m(S_1,S_2)h_{lk}(m) \tag{3}$$

$$h_m=\begin{cases}\binom{l-m}{k}, & l-m\geqslant k\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

为了使该方法适用于大规模基因组，作者设计了一种用于计算内核矩阵的有效树形数据结构。$S_1S_2S_3$ 的 3-mer 表示形式如图 6 所示。

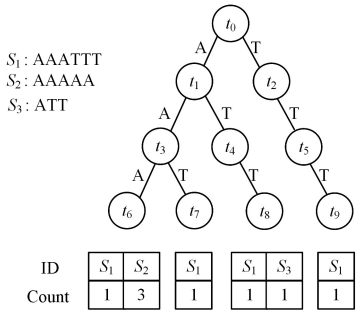

图 6 树结构示意图

Fig. 6 Tree structure

实验证明，gkmSVM 方法预测功能基因组调控元件以及组织特异性增强子的性能有显著提升，准确度是 kmerSVM 的两倍。此外，该方法还进一步使用朴素贝叶斯分类器证明了模型的通用性，其适用于任何序列分类问题。

上述两种方法主要从数据处理角度出发，通过 k-mer 分析得到序列特征来实现分类预测功能。Singh 等提出的 SPEID[39] 方法，不考虑基因功能信号，仅从序列特征角度出发。首先，通过 one-hot 编码将一维的增强子和启动子序列分别编码为二维的矩阵，并将其作为卷积神经网络(Convolutional Neural Networks，CNN)的输入；然后，对从增强子和启动子中分别提取到的高维特征进行特征融合；接着，将融合后的特征传入长短期记忆网络(Long Short-Term Memory，LSTM)模型中；最后，通过全连接层输出预测结果。SPEID 方法的整体流程如图 7 所示。

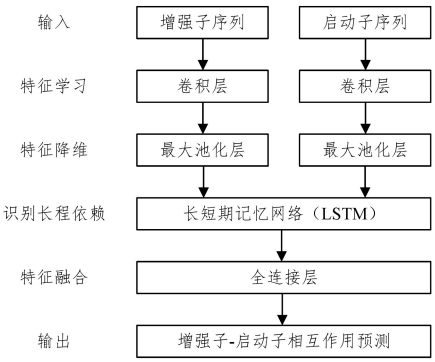

图 7 SPEID 方法概览

Fig. 7 Workflow of SPEID

CNN[40-41] 是一类包含卷积计算的深度前馈神经网络，具有表征学习能力和平移不变性[42]。该模型共享卷积核，善于处理高维数据，并能自动提取特征。循环神经网络(Recurrent Neural Network，RNN)[43] 的变体 LSTM[44]，主要解决了 RNN 可能遇到的梯度爆炸和梯度消失问题[45]，并在其基础上增加了输入门、输出门和遗忘门 3 个控制单元，解决了神经网络的长序列依赖问题。模型通过 CNN 可以学习到比 k-mer 数量更多且维度更高的特征。而 LSTM 考虑了增强子和启动子的复杂作用关系，保留了上下游基因的位置关系。这样的结合，使 SPEID 能够取得较好的预测效果。

SPEID 模型主要通过神经网络学习序列特征，而 gkmSVM 方法用 k-mer 来表示特征，上述两类方法都在 EPIs 预测方面取得了较好的结果。为了验证融合上述两种方式得到的特征能否进一步提高预测准确率，Yang 等提出了多序列特征融合的 PEP[46] 方法，其架构如图 8 所示。

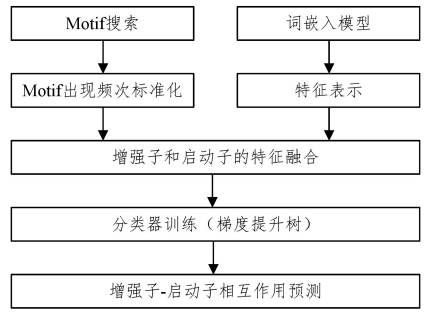

图 8 PEP 方法概览

Fig. 8 Workflow of PEP

分析 k-mer 时，k 的取值通常在 6～8 之间，取值过大会给实验数据带来更多的噪声，从而影响结果的准确性。由于转录因子结合位点(Transcription Factor Binding Site，TFBS)的长度通常在 10～20 bp 之间，因此 k-mer 无法较好地表示这类信息。word2vec 方法[47] 可使所有文本以词向量的形式存在于向量空间中，使在语料库中共享上下文的词在空间中彼此靠近。此方法兼顾了语义信息和上下文关系，因此也适用于 EPIs 预测。

PEP 方法的核心思想是将长短 motif 分开表示：首先，提取 TFBS 这类长序列作为 motif 特征；然后，通过 word2vec 从原始序列中提取短的词向量[48] 作为另外的特征；接着，进行特征融合；最终，通过梯度提升决策树 GTB 预测 EPIs。

在一维染色质状态预测方面，gkmSVM 将可变长度的序列用固定长度的 k-mer 频次表示。k-mer 特征是无偏差的、完整的特征集，但在相邻的 k-mer 之间丢失了文本上下文信息。在三维基因相互作用预测方面，SPEID 的效果较好，但该方法要求输入固定长度的序列信息，而序列的填充和拆分可能会影响序列的功能。将长度不一的序列转化为固定长度的表达，在序列预测问题中至关重要。doc2vec[49] 是在 word2vec 的基础之上提出的，可以用来学习一个句子的嵌入向量。其本质就是训练一个能够自动将句子转换为向量并对其语义进行编码的模型。段落向量就是将文本映射成统一向量表示的方法。

基于以上几点，Zeng 等提出了三维基因组相互作用预测方法 EP2vec[50]。EP2vec 方法主要分为两步。1)数据预处

理：通过 k-mer 将增强子和启动子序列分别拆分为长度固定的短序列；将每条完整的序列视为文本，将拆分出来的短序列视为词语，通过非监督学习的 doc2vec 方法来进行训练，并学习特征。2）利用监督学习方法预测 EPIs：给定增强子-启动子序列对，分别将增强子和启动子通过 doc2vec 方法训练得到的特征进行表示，再将编码后的增强子和启动子特征融合，并通过梯度提升回归树 GBRT 来预测这对序列是否发生相互作用。EP2vec 方法的流程图如图 9 所示。

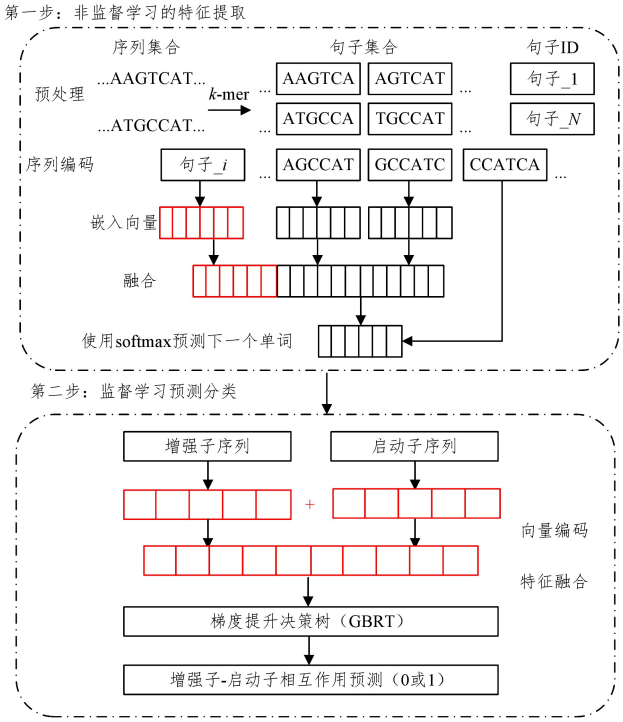

图 9 EP2vec 方法概览

Fig. 9 Workflow of EP2vec

针对不同特征对序列含义的贡献度不同的问题，EP2vec 引入了注意力机制来判断哪些 motifs 对表征某类增强子或者启动子的含义至关重要。以第 i 个增强子 x_i 为例，它包含了 T_i 个单词 $\boldsymbol{W}^{C_i,1}$，$\boldsymbol{W}^{C_i,2}$，…，$\boldsymbol{W}^{C_i,T_i}$，首先计算词向量 $\boldsymbol{W}^{C_i,t}$ 和句向量 $\boldsymbol{x}_i$ 之间的相似性，即权重，如式(5)所示：

$$\alpha_{it}=\frac{\exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}^{C_{i,t}}}{\sum_j \exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}^{C_{i,j}}} \tag{5}$$

然后，将高权重的 motifs 进行可视化表示，并将结果与通过生物方法检测的 motif 相比较，最终发现大多数 motif 都具有实际的生物学意义，从而证明了该方法的有效性。另外，该方法在 FANTOM 数据集上的预测效果也很好，该数据集仅包含 EPIs 正样本，负样本通过序列随机匹配生成。

Zhuang 等在研究 SPEID 之后，假设 DNA 序列的独立结构和长程依赖关系对 EPIs 预测的影响不大，进而提出了 EPIsCNN[51] 模型。作者不考虑上下文中包含的位置信息，同时还假设仅使用简单的模型结构就能达到较好的预测效果，因此只使用了与 SPEID 相同的 CNN 结构：一个卷积层、一个最大池化层和一个全连接层，并且得到了一个与 SPEID 类似的结果。模型结构如图 10 所示。

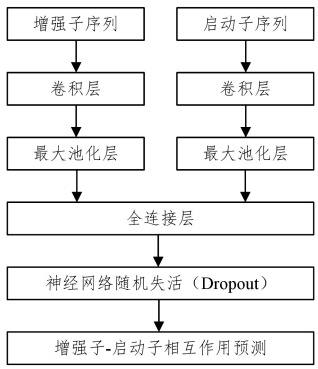

图 10 EPIsCNN 方法概览

Fig. 10 Workflow of EPIsCNN

模型的输入为 one-hot 编码后的定长序列。同时，该方法在训练模型的过程中引入了转移学习方法：首先使用除目标细胞系以外的其他细胞系预训练模型，并使用目标细胞系训练模型；接着使用所有细胞系数据训练模型并提取特征，再使用目标细胞系的数据到全连接层中去更新特征以进行下一步 EPIs 预测[52]，进一步提高了预测精度。

3.2.2 EPIs 预测方法小结

表 2 从原始输入、数据处理、特征融合和模型分类这几个角度对上述 EPIs 预测方法进行了对比分析。

表 2 EPIs 预测方法的对比

Table 2 Comparison of EPIs prediction methods

方法名称	原始输入	数据处理	特征融合	分类模型	优点	缺点
TargetFinder[35]	增强子、启动子 基因特征	—	基因特征	Boosted Tree	学习组学特征 预测结果准确	组学特征的获取成本高 特征受生物实验限制
kmer-SVM[37]	增强子 (EP300-bound)	k-mer	k-mer 频次	SVM	准确区分增强子和 其他基因元件	k 值固定 分类结果受其影响
gkm-SVM[38]	增强子 (EP300-bound)	gapped k-mer	k-mer 频次	SVM Naive Bayes	k 值可变 效果优于 kmer-SVM	分类模型较简单
PEP[46]	增强子、启动子 TFBS	k-mer+word2vec motifs	word2vec motifs	GTB	融合更多特征 (短 k-mer+长 TFBS) 效果优于 SVM 类方法	学习到重复特征
EP2vec[50]	增强子、启动子	k-mer + doc2vec	word2vec doc2vec	GBRT	从原始序列中学习特征 保留了特征的长程依赖关系 效果在数据集上达到最优	模型性能受 k 值和 特征数量的影响
SPEID[39]	增强子(3 kbp) 启动子(2 kbp)	one-hot	高维特征 位置关系	CNN LSTM	保留了特征的 长程依赖关系	序列长度固定 部分信息丢失
EPIsCNN[51]	增强子、启动子	one-hot	高维特征	CNN	直接学习高维特征转移 学习方法获取更多特征	未考虑 EP 在空间上相互 靠近的相互作用方式

目前效果较好的模型有 PEP，EP2vec 和 EPIsCNN 3 种。下面将从不同角度来阐述这几种方法的优缺点以及使用场景。

1）PEP：考虑到 k-mer 容易受长度 k 的影响，该方法将长短序列 motifs 区分开，即短序列特征通过模型学习得到，长序列则直接使用实验得出的结果。此方式打破了模型只能学习固定长度特征的局限性，从数据本身出发，融合了更多特征，进一步提高了预测精度。其缺点在于，两种考虑会学习到重复的特征。

2）EP2vec：DNA 序列的拆分和填充都可能会影响序列本身的信息。doc2vec 在处理可变长度文本的同时还能兼顾上下文信息。该方法处理 EPIs 问题的难点在于，k-mer 分析中 k 值的选取以及特征数量的选择。特征数量的多少会影响分类的结果，而 GBRT 决策树能较好地对特征进行无监督学习。该方法在序列相互作用预测以及文本处理情感分析领域应用广泛。

3）EPIsCNN：该方法忽略了多数增强子通过空间上的接近和启动子发生作用这一事实，不考虑所有的空间关系，仅通过 CNN 学习序列的高维特征并进行分类。其预测效果虽然与 EP2vec 有一定差距，但它考虑到 6 类细胞系可能共同存在某些特征，采取转移学习的方法来学习全部数据集的特征，提高了预测效果。深度学习的样本数据量较少时，可以考虑使用转移学习方法来获得更多特征。

4 结论与展望

“高通量测序与组学大数据研究”中的转录序列分析[53]是生物医学、农学等领域的重要部分[54]。研究者通过研究增强子的活化结构与活化剂之间的关系来理解基因的调控机理[55]，将超级增强子作为靶点的研究可以为抗肿瘤药物的研发提供理论基础和借鉴[56]。EPIs 的相关研究越来越深入，应用也越来越广泛。下面对 EPIs 预测工作的未来发展趋势和挑战进行了总结。

1）数据预处理：现有的数据预处理方法主要是 one-hot 编码和 k-mer 分析，前者将一维的 DNA 数据拆分成固定长度，再将其转化为二维的图片形式，以此作为 CNN 模型的输入；k-mer 分析默认可以将出现频次高的序列视为序列的特征（motif）。这两种方法都无法很好地体现序列在空间上的靠近。因此，如何找到新的序列编码方式，使得既能保持序列的空间关系，又能较好地体现特征，还使得机器能够快速理解和处理这些信息，将是我们下一步亟待解决的问题。

2）类不平衡问题[57]的处理方法：序列预测问题中普遍存在正样本数量远远少于负样本数量的情形。若把全部样本作为模型的输入，模型会过多地学习到负样本的特征[58]，从而影响预测结果。EPIs 预测问题中的常用方法包括过采样和欠采样[59]，它们的最终目的都是确保正负样本数量一致。为了消除随机选择或随机修改数据带来的影响，后续工作可以考虑给正负样本动态设置权重[60]、样本聚类[61]等方法，使模型能够学到数据的全部特征，也可考虑更有效的样本生成方法。

3）特征数量选择问题：特征数量多的模型容易造成过拟合；特征数量少的模型会丢失某些重要信息。目前，RIFS 方法[62]常用于特征选择[63]，后续研究也可以关注这方面的问题。

4）模型选择和参数调整：目前 CNN 和 doc2vec 模型已经在 EPIs 预测中达到了较好的预测结果，后续若想进一步提升精度，应根据数据的特点进行针对性选择，如利用善于处理文本数据的 BERT[64]和 Transformer 模型[65]、自编码器[66]、深度置信网络 DBN[67]及深度信念网络 DBM[68]等去学习特征。另外，神经网络模型的参数调整过于依赖人的经验；使用网格搜索[69]等参数遍历寻优的方法又会耗费大量的时间，在模型训练过程中也极易出现过拟合[70]问题。因此，如何学习到最优的模型参数[71-72]值得研究人员思考。

5）跨细胞系预测的准确性：目前的模型仅仅在已有的细胞系上取得了较好的效果，若将该模型用于其他数据集，预测效果就会很差。因此，如何训练出合适的模型，保证其跨细胞系预测的效果，亦是未来的研究重点。

结束语 本文对 EPIs 生物检测方法做了简单的概述。针对生物检测方法的局限性，总结对比了基于多特征融合的 EPIs 预测方法，分析说明了该研究方法的不足以及该研究在未来可能的发展方向。当前，应用在 EPIs 中的深度学习模型虽然结构简单，但仍取得了比较好的效果。其缺点是深度学习模型学习到的特征并非全都具有合理的生物学解释。在未来，深度学习方法必然会成为解决 EPIs 预测以及类似序列相互作用预测问题的主流方法。

参 考 文 献

[1] ESTELLER M. Non-coding RNAs in human disease[J]. Nature Reviews Genetics, 2011, 12(12): 861-874.

[2] YANG F. Research on piRNA and promoter based on sequence information[D]. Harbin: Harbin Institute of Technology, 2018.

[3] KARNUTA J M, SCACHERI P C. Enhancers: bridging the gap between gene control and human disease[J]. Human Molecular Genetics, 2018, 27(R2): R219-R227.

[4] BLACKWOOD E M, KADONAGA J T. Going the Distance: A Current View of Enhancer Action[J]. Science, 1998, 281(5373): 60-63.

[5] PENNACCHIO L A, BICKMORE W, DEAN A, et al. Enhancers: five essential questions[J]. Nature Reviews Genetics, 2013, 14(4): 288.

[6] JIANG R. Walking on multiple disease-gene networks to prioritize candidate genes[J]. Journal of Molecular Cell Biology, 2015, 7(3): 214-230.

[7] DAVISON L J, WALLACE C, COOPER J D, et al. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene[J]. Human Molecular Genetics, 2012, 21(2): 322-333.

[8] SMEMO S, TENA J J, KIM K H, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3[J]. Nature, 2014, 507(7492): 371-375.

[9] MASTON G A, EVANS S K, GREEN M R. Transcriptional regulatory elements in the human genome[J]. Annual Review of Genomics & Human Genetics, 2006, 7(1): 29.

[10] HE B, CHEN C, TENG L, et al. Global view of enhancer-promoter interactome in human cells[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(21).

[11] YU Z, ZHAO Y X, YI Z L, et al. Research on folding diversity in statistical learning methods for RNA secondary structure prediction[J]. International Journal of Biological Sciences, 2018, 14(8): 872-882.

[12] DAVID R. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[J]. Journal of the American Statistical Association, 2004, 99(466): 567-567.

[13] ROBERT C. Machine Learning, a Probabilistic Perspective [M]// Machine learning: a probabilistic perspective. 2012.

[14] MICHALSKI R S, CARBONELL J G, MITCHELL T M. Machine Learning[M]//Symbolic Computation. 1994: 3-61.

[15] ANGERMUELLER C, PARNAMAA T, PARTS L, et al. Deep learning for computational biology[J]. Molecular Systems Biology, 2016, 12(7): 878.

[16] PRICE C M. Fluorescence in situ hybridization[J]. Blood Reviews, 1993, 7(2): 127-134.

[17] LI G, RUAN X, AUERBACH R K, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation[J]. Cell, 2012, 148(1/2): 84-98.

[18] DE WIT E, DE LAAT W. A decade of 3C technologies: insights into nuclear organization [J]. Genes & Development, 2012, 26(1): 11-24.

[19] HAKIM O, MISTELI T. SnapShot: Chromosome confirmation capture[J]. Cell, 2012, 148(5): 1068. e1.

[20] DEKKER J, RIPPE K, DEKKER M, et al. Capturing chromosome conformation[J]. Science, 2002, 295(5558): 1306-1311.

[21] SIMONIS M, KLOVS P, SPLINTER E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)[J]. Nature Genetics, 2006, 38(11): 1348-1354.

[22] 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.

[23] RAO S S, HUNTLEY M H, DURAND N C, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping[J]. Cell, 2014, 159(7): 1665-1680.

[24] LIEBERMAN-AIDEN E, VAN BERKUM N L, WILLIAMS L, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome[J]. Science, 2009, 326(5950): 289-293.

[25] HEIDARI N, PHANSTIEL D H, He C, et al. Genome-wide map of regulatory interactions in the human genome[J]. Genome Research, 2014, 24(12): 1905-1917.

[26] FULLWOOD M J, RUAN Y. ChIP-Based Methods for the Identification of Long-Range Chromatin Interactions[J]. Journal of Cellular Biochemistry, 2009, 107(1): 30-39.

[27] HOFFMAN M M, BUSKE O J, WANG J, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation[J]. Nature Methods, 2012, 9(5): 473-476.

[28] ERNST J, KELLIS M. ChromHMM: automating chromatin-state discovery and characterization[J]. Nature Methods, 2012, 9(3): 215-216.

[29] BERNSTEIN B E, STAMATOYANNOPOULOS J A, COSTELLO J F, et al. The NIH roadmap Epigenomics mapping consortium[J]. Nat Biotechnol, 2010, 28(10): 1045-1048.

[30] HARRIS D M, HARRIS S H. Digital design and computer architecture[M]. Chian Machine Press, 2014.

[31] COMPEAU P E, PEVZNER P A, TESLER G, et al. How to apply de Bruijn graphs to genome assembly[J]. Nature Biotechnology, 2011, 29(11): 987-991.

[32] WELCH M, GOVINDARAJAN S, NESS J E, et al. Design Parameters to Control Synthetic Gene Expression in Escherichia coli[J]. PLOS ONE, 2009, 4(9): e7002.

[33] GUSTAFSSON C, GOVINDARAJAN S, MINSHULL J. Codon bias and heterologous protein expression[J]. Trends in Biotechnology, 2004, 22(7): 346-353.

[34] ESCHKE K, TRIMPERT J, OSTERRIEDER N, et al. Attenuation of a very virulent Marek's disease herpesvirus (MDV) by codon pair bias deoptimization [J]. PLOS Pathogens, 2018, 14(1).

[35] 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.

[36] JOHN S, SABO P J, THURMAN R E, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns [J]. Nature Genetics, 2011, 43(3): 264-268.

[37] LEE D. Discriminative prediction of mammalian enhancers from DNA sequence[J]. Genome Research, 2011, 21(12): 2167-2180.

[38] GHANDI M, LEE D, MOHAMMADNOORI M, et al. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features [J]. PLoS Computational Biology, 2014, 10(12): e1003711.

[39] SINGH S, YANG Y, POCZOS B, et al. Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks[J]. Quantitative Biology, 2019, 7: 122-137.

[40] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551.

[41] LECUN Y. Generalization and network design strategies [C]// Connectionism in Perspective. 1989: 143-155.

[42] Zhang W, Itoh K, Tanida J, et al. Parallel distributed processing model with local space-invariant interconnections and its optical architecture[J]. Applied Optics, 1990, 29(32): 4790-4797.

[43] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[44] GREFF K, SRIVASTAVA R K, KOUTNiK J, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 28(10): 2222-2232.

[45] SALEHINEJAD H, SANKAR S, BARFETT J, et al. Recent Advances in Recurrent Neural Networks [J]. arXiv: 1801.01078.

[46] YANG Y, ZHANG R, SINGH S, et al. Exploiting sequence-

based features for predicting enhancer-promoter interactions [J]. Bioinformatics,2017,33(14):i252-i260.

[47] MIKOLOV T,CHEN K,CORRADO G S,et al. Efficient Estimation of Word Representations in Vector Space[C]// International Conference on Learning Representations. 2013.

[48] GOLDBERG Y,LEVY O. word2vec Explained:deriving Mikolov et al.'s negative-sampling word-embedding method [J]. arXiv:1402.3722.

[49] LE Q V,MIKOLOV T. Distributed Representations of Sentences and Documents[C]// International Conference on Machine Learning. 2014:1188-1196.

[50] ZENG W,WU M,JIANG R. Prediction of enhancer-promoter interactions via natural language processing[J]. BMC Genomics,2018,19(S2):84.

[51] ZHUANG Z,SHEN X,PAN W,et al. A Simple Convolutional Neural Network for Prediction of Enhancer-Promoter Interactions with DNA Sequence Data [J]. Bioinformatics,2019,35(17):2899-2906.

[52] PAN S J,YANG Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering,2010,22(10):1345-1359.

[53] ØROM U A. Enhancer RNAs[M]. New York:Humana Press,2017.

[54] XIE J H,SUN Y,WANG S,et al. Functional Identification of Enhancer and Its Research Progress in Agricultural Animals [J]. Chinese Journal of Cell Biology,2019,41(7):1395-1400.

[55] BENABDALLAH N S,WILLIAMSON I,ILLINGWORTH R S,et al. Decreased enhancer-promoter proximity accompanying enhancer activation[J]. Molecular cell,2019,76(3):473.

[56] WU Z Q,MI Z Y. Research progress of super enhancer in cancer [J]. Hereditas,2019,41(1):41-51.

[57] HE H,GARCIA E A. Learning from Imbalanced Data [J]. IEEE Transactions on Knowledge & Data Engineering,2009,21(9):1263-1284.

[58] CHAWLA N V,JAPKOWICZ N,KOTCZ A. Editorial:special issue on learning from imbalanced data sets[J]. Acm Sigkdd Explorations Newsletter,2004,6(1):1-6.

[59] KANG P,CHO S. EUS SVMs:Ensemble of Under-Sampled SVMs for Data Imbalance Problems[C]// International Conference on Neural Information Processing. 2006.

[60] LU Y,CHEUNG Y M,TANG Y Y. Dynamic Weighted Majority for Incremental Learning of Imbalanced Data Streams with Concept Drift[C]// Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017.

[61] EISEN M B,SPELLMAN P T,BROWN P O,et al. Cluster analysis and display of genome-wide expression patterns[J]. Proceedings of the National Academy of Sciences of the United States of America,1998,95(25):14863-14868.

[62] YE Y,ZHANG R,ZHENG W,et al. RIFS:a randomly restarted incremental feature selection algorithm[J]. Scientific Reports,2017,7(1):13013.

[63] RAO H,SHI X,RODRIGUE A K,et al. Feature selection based on artificial bee colony and gradient boosting decision tree[J]. Applied Soft Computing,2019,74:634-642.

[64] DEVLIN J,CHANG M,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// North American Chapter of the Association for Computational Linguistics. 2019:4171-4186.

[65] VASWANI A,SHAZEER N,PARMAR N,et al. Attention Is All You Need[J]. arXiv:1706.03762.

[66] VINCENT P,LAROCHELLE H,BENGIO Y,et al. Extracting and composing robust features with denoising autoencoders [C]// Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML 2008). ACM,2008.

[67] MOHAMED A,DAHL G E,HINTON G E,et al. Acoustic Modeling Using Deep Belief Networks[J]. IEEE Transactions on Audio,Speech,and Language Processing,2012,20(1):14-22.

[68] SRIVASTAVA N,SALAKHUTDINOV R R,HINTON G E. Modeling Documents with Deep Boltzmann Machines[J]. arXiv:1309.6865.

[69] LAVALLE S M,BRANICKY M S. On the Relationship between Classical Grid Search and Probabilistic Roadmaps[J]. International Journal of Robotics Research,2003,23(23):673-692.

[70] REUNANEN J. Overfitting in Making Comparisons Between Variable Selection Methods[J]. Journal of Machine Learning Research,2003,3(3):1371-1382.

[71] WRIGHT A H. Genetic Algorithms for Real Parameter Optimization[J]. Foundations of Genetic Algorithms,1991,1:205-218.

[72] HAO S,WANG X,XIE J,et al. Rigid framework section parameter optimization and optimization algorithm research [J]. Transactions of the Canadian Society for Mechanical Engineering,2019,43(8):398-404.

HU Yu-jia, born in 1995, postgraduate, is a member of China Computer Federation. Her main research interests include data mining and bioinformatics.

ZHU Min, born in 1971, Ph.D, professor, is a senior member of China Compu- ter Federation. Her main research inte-rests include bioinformatics, information visualization and visual analytics.