中国科学: 生命科学 2023 年 第 53 卷 第 10 期: 1370 ~ 1382
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
评 述 南华大学衡阳医学院65周年庆专辑

增强子靶标基因的预测方法研究

徐晓强, 崔婷, 张涵, 尚德思, 李春权*

南华大学衡阳医学院附属第一医院, 人工智能与大数据中心, 衡阳 421001
* 联系人, E-mail: lcqbio@163.com

收稿日期: 2023-05-24; 接受日期: 2023-08-21; 网络版发表日期: 2023-10-07
湖南省自然科学基金(批准号: 2023JJ30536, 2023JJ30535)资助

摘要 增强子是作为基因组顺式调节元件的功能性DNA片段, 通过与靶基因启动子的相互作用调节基因表达. 识别增强子和靶标基因的相互作用对理解基因调控、细胞分化和疾病发展的机制有重要意义. 然而, 通过实验方法鉴定增强子靶标基因(enhancer-target gene, ETG)在时间、人力和金钱方面都花费太多. 因此, 越来越多的研究工作集中在开发计算方法来解决这一问题. 本文对用于ETG预测的计算方法进行系统的总结, 以促进其应用. 最后, 探讨了该领域提出的现有解决方案存在的限制和挑战并展望了未来的研究方向.

关键词 增强子, 基因表达, 增强子和靶标基因, 计算方法

增强子是在基因组上的DNA片段, 其作为远端顺式调控元件(*cis*-regulatory elements, CREs)通过与靶标基因启动子的相互作用调控基因的表达[1]. 增强子-靶标基因(enhancer-target gene, ETG)相互作用是基因调控的一个重要机制, 是阐明基因组功能背后的机制, 理解遗传分子如何决定细胞命运, 以及这种功能的破坏如何导致疾病的关键之一[2,3]. 因此, 识别ETG具有重要意义. 增强子和启动子一样, 通过结合转录因子(transcription factor, TF)发挥作用, 它们位于距离其调控的靶标基因的转录起始位点(transcription start site, TSS)更远的位置[4,5]. 事实上, 有多种原因导致阐明增强子的功能并识别ETG的复杂模式成为一个挑战. 首先, 一个增强子可以作用于一个或多个靶标基因或者启动子, 而一个或多个增强子也可以共同调节一个靶标基因或者启动子[6,7]. 同时, 增强子和启动子在位置上没有固定关系, 增强子可以位于靶标基因的上游, 也可以位于靶标基因的下游, 还可以位于靶标基因的附近, 但不调节最近的基因[8,9]. 这些原因都使得识别ETG变得更加复杂和困难. 其次, 哺乳动物基因组中有数以百万计的片段可能充当增强子, 但其活性在不同的细胞类型中差异很大; 事实上, 增强子活性是组织和细胞类型中变化很大的单一基因组特征[10,11]. 虽然一个特定的基因可能在多种细胞类型中具有活性, 但基因的激活在不同组织中可能由不同的增强子触发, 因此, ETG或者增强子-启动子相互作用(enhancer-promoter interaction, EPI)也是细胞类型特异性的[12,13]. 相比之下, EPI预测的计算方法更侧重于广泛的调控关系, ETGs预测方法更适合研究特定基因的调控元件, (下文ETG和EPI, 不再进行区分)[14]. 以上原因进一步增加了识别一组全面的增强子和ETG的复杂性.

引用格式: 徐晓强, 崔婷, 张涵, 等. 增强子靶标基因的预测方法研究. 中国科学: 生命科学, 2023, 53: 1370–1382
Xu X Q, Cui T, Zhang H, et al. Computational methods to predict Enhancer-target Gene Pairs (in Chinese). Sci Sin Vitae, 2023, 53: 1370–1382, doi: 10.1360/SSV-2023-0086

尽管存在这些挑战，但是识别ETG在基因组学和计算生物学领域越来越受到关注，因为研究者们可以利用全基因组的实验数据来尝试解决这一问题[14]. 增强子及其远端的靶标基因是通过染色质在各种结构蛋白，如中介蛋白和内聚蛋白复合物的协同作用下形成环状结构以促进它们的物理相互作用(图1)[15]. 已经发展的实验技术使人们更好地了解染色质结构，进而识别连接远端增强子与其靶标基因启动子的染色质相互作用. 这些实验技术的出现促进了这一领域的发展, 包括基于开创性的染色体构象捕获技术[16]及其染色体构象捕获芯片和染色体构象捕获碳拷贝的衍生物[17,18], 例如, Hi-C[19], ChIA-PET[20]和Capture HiC[21]已应用于几种人类细胞类型和组织[22~24]. 尽管实验技术已经大大扩展了远端染色质相互作用的注释目录，但仍有一些限制因素阻碍了对细胞类型特异性增强子-启动子相互作用的深入分析[25]. 由于这些限制，需要基于ENCODE[26]和Roadmap Epigenomics[12]等项目收集的大规模多组学数据资源，其中包含基因调控的多视图信息，包括基因表达、转录因子结合和组蛋白修饰的整合, 来定义增强子和预测ETG调控网络[27].

ETG全基因组表征的第一个实际问题是定义要考虑细胞类型的增强子区域集合. 尽管大型表观基因组联合体(如ENCODE和FANTOM)为增强子鉴定做出了努力，但增强子活性的动态和细胞类型特异性导致无法创建增强子的详尽参考列表[26,28]. 除了基因报告分析包括高通量版本的STARR-seq[29]、大规模平行报告分析[30]和数据库RAEdb[31]以及VISTA[32]等包含部分实验验证的增强子，对于全基因组范围的增强子的定义，人们通常采用多种类型的功能基因组数据包括表观基因组和转录组数据. 基于表观基因组数据的增强子区域注释，可以使用与增强子活性相关的功能数据，如转录因子或其他辅助因子的结合、特定的组蛋白修饰和染色质可及性. 比如, p300是一种组蛋白乙酰转移酶蛋白，作为转录共激活子，已知它与活性增强子结合[33]. 因此，针对p300的染色体免疫共沉淀测序(chromatin immunoprecipitation followed by sequencing, ChIP-seq)实验经常被用于增强子的全基因组注释；同样针对染色质标记的ChIP-seq实验，也可以达到相同目的[34,35]. 例如，高水平的H3K4me1和H3K27ac，通常会在与活性增强子相关的核小体中发现[36,37]；包括增强子和启动子在内的活跃调控区域通常是具有较高可及性的染色质[38~40]. 因此, 探索染色质可及性的基因组学方法，如DNase-seq和ATAC-seq，可作为全基因组识别增强子的替代方法[41~43]. 基于转录组数据的增强子区域注释，主要依赖于活性增强子产生的双向非编码RNA，也被称为eRNA，其长度通常为的0.5~2 kb[44~46]. eRNA的转录可以用如下测序方案检测，包括CAGE-seq, GRO-seq, GRO-cap, 5′GRO-seq, NET-Seq, PRO-cap, PRO-seq, Start-seq, TT-seq[47~56]. 比如, FANTOM项目联盟使用CAGE-seq鉴定的增强子在432个原代细胞、135个组织和241个细胞系样本中总计约60000个.

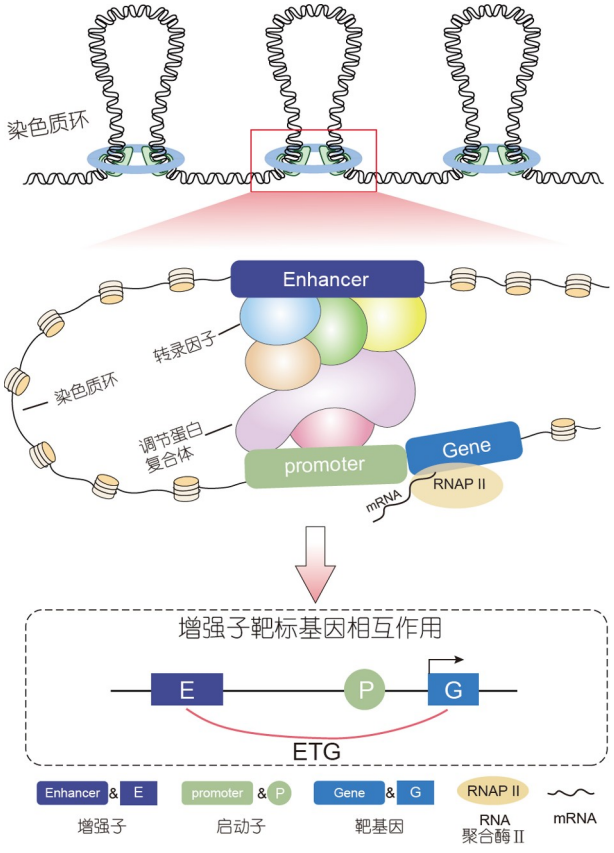

图 1　增强子影响基因转录表达模式图

Figure 1　The pattern of enhancer affecting gene transcription and expression

总之，从公共数据库公开获得的各种组织或细胞类型特异的基因组数据、基因组数据以及转录组数据可以作为识别ETG的有效信息，这使得以计算的方式研究ETG成为新的热点. 本文介绍了多种预测ETG的计算方法，并根据算法和模型类别和输入特征进行分类, 以作为帮助研究人员选择合适研究方法的参考.

1 基于监督学习的算法

监督学习方法包括基于经典的机器学习浅层模型，如随机森林、决策树、逻辑回归和线性回归分析的方法；也包括基于深度学习的深层模型，如卷积神经网网络(convolutional neural networks, CNN)[57,58]、注意力机制[59]、Transformer[60,61]、循环神经网络(recurrent neural networks, RNN)[62]等方法. 具体来说, 这些算法利用一组选定的已知真阳性ETG，来识别它们在基因组注释或功能数据中的相关模式，从而构建机器学习分类器，然后能够使用不同的具有特定表观基因组数据和序列信息的细胞系来训练分类器模型，最后应用该模型对另一个细胞系进行预测. 根据算法中应用的模型，监督学习的方法可分为基于回归的方法和基于训练分类器的方法. 值得注意的是，越来越多的研究团队开发了用户友好的web服务平台，用于高效预测、分析和可视化EPI交互. 比如EPIXplorer[63]，它集成了9种强大的预测算法，支持多种类型的三维接触数据和多组学数据作为输入. EPIXplorer的输出通过调节元件和风险单核苷酸多态性进行评分和充分注释. 此外, 可视化和下游模块提供了进一步的功能分析, 所有输出文件和高质量图像都可供下载. 其中提到的包括基于监督学习在内的相关算法会在下文分别介绍, 所以EPIXplorer并没有被归结为具体哪一类方法.

1.1 基于回归的方法

基于回归的方法通过将增强子的特征与启动子的特征或基因表达联系起来，确定增强子与启动子或靶标基因之间的相互关系，其基本原理是：因为多个增强子可以调节单个基因，所以这些方法使用组合而不是成对的方法进行ETG配对，识别增强子和目标基因之间的显著关系，同时评估多个增强子对其靶标的影响强度.

在这个类别中，JEME[64]，RIPPLE[65]和FOCS[66]就是在计算过程中应用回归方法的典型例子. JEME通过考虑多个增强子的联合效应以及综合全局和样本的信息来识别特定样本中的ETG. JEME使用每个转录起始位点1 Mb内的所有增强子作为候选的起始集；然后基于所有候选增强子的活性，使用多元线性回归结合Lasso正则化约束来评估预测TSS活性的误差项. 在回归步骤之后，基于每个细胞类型特定的数据训练随机森林分类器以预测ETG，即随机森林模型将基于回归模型估计的细胞类型特异性误差项，增强子、TSS和增强子-靶标基因对窗口处的染色质标记数据H3K4me1, H3K27ac, H3K27me3, DNase I以及增强子和候选靶标基因之间的距离作为输入特征，应用于重建935个人类原始细胞、组织和细胞系样品中的增强子-启动子网络，能够系统地研究正常和疾病状态下的基因调控. 从以上的步骤可以看出JEME实际上是一种混合方法，同时具有回归和基于训练分类器方法的特点.

类似地，RIPPLE也是一种结合了用于特征选择的回归模型的基于监督学习的方法，它以Sanyal等人的染色体构象捕获碳拷贝数据(即染色质环)为阳性集训练细胞类型特异性的随机森林分类器，并使用基于多任务学习和Group Lasso的方法在四个细胞系中进行联合特征选择. 然后，以细胞系特异的方式预测增强子和启动子的相互作用.

除了JEME和RIPPLE外，FOCS被认为是基于回归的统计框架. 与JEME一样，FOCS采用eRNA作为增强子活性的标记, 它首先使用普通最小二乘回归模型, 根据k个最接近的增强子的活性预测每个启动子活性. 这个回归模型在多种细胞类型上进行训练，并进行留一交叉验证. 每个预测结果的可信程度则根据观察到的留下的样本中的启动子活性进行评估. 然后，为了进行精确预测，对整个模型进行训练并用Elastic-net正则化约束以选择与调控靶基因更相关的增强子. 增强子和启动子的活性最初是根据ENCODE的DNase-seq数据估计的，因此它可以用于推断来自ENCODE，Roadmap Epigenomics和FANTOM5样本中与活性模式相关的增强子-启动子相互作用. 将FOCS应用于大规模的基因组数据集，包括由Roadmap Epigenomic提供的DNase-seq数据, FANTOM5的CAGE-seq数据和一个定制的公开可获得的GRO-seq数据集，可以得到广泛的增强子-启动子图谱, 从而推导出生物假说.

1.2 基于训练分类器的方法

通过利用真阳性增强子-启动子或者靶标基因对的表观基因组特征或基因组特征，研究者们可以建立一个训练好的经典机器学习分类器以及神经网络算法来确定一个感兴趣的ETG是否参与相互作用. 对于单个基于浅层模型的预测工具，它们使用的特征数量和

类型变化很大，但是大多数预测工具都考虑了基因表达、染色质可及性(DNase I-hypersensitive sites, DHS或者ATAC)和组蛋白标记的组合等. 此外, 它们都在一定程度上考虑增强子与基因之间的距离. IM-PET[67]整合了一组判别特征，包括增强子和启动子活性、转录因子和目标启动子之间的相关性、相互作用的增强子-启动子对之间的进化约束以及增强子和目标启动子之间的距离约束来预测EPI; PETModule[68]是一种基于motif模块的ETGs预测方法，使用与IM-PET相似的特征. 具体来说, 它是一个集成的计算方法, 将之前方法使用的保守性、距离、增强子-启动子的相关性以及基于基序模块的全新特征集成到一个预测模型中. 一个基序模块是指在基因组短区域中显著共出现的一组DNA结合基序，长度为1 kb左右. TargetFinder[69]利用与增强子和启动子相关的多种特征作为集成的提升树模型的输入，来预测细胞特异性的增强子和启动子的相互作用；然而，由于需要多个输入并且过程复杂，计算这些特征并不容易. EPIP[70]则可以直接预测条件特异的EPIs，增强子和启动子的部分特征缺失或者带有不同的数据集合，EPIP使用特征分区策略将特征分组为11个分区或重叠的特征集，这样可以在一定程度上使用有缺失数据的特征集合. ProTECT[71]是通过利用转录因子的蛋白质互作网络(protein-protein interactions，PPIs)特征信息的一种新的细胞类型特异性增强子-启动子相互作用的预测模型. 这个模型的原理是基于最近的生物学研究发现, 转录因子之间的蛋白质-蛋白质相互作用参与了染色质环的调控. 为了进一步有效地将TF PPIs的特征整合到随机森林模型中, 该模型的独到之处是利用了一种基于图论的降维方法，基于PPIs网络的拓扑结构, 将单个TF分配于某一个TF PPIs模块中, 从而大幅度降低特征维度, 同时保持模型的可预测性且一定程度上解决了模型的过拟合问题. 通过一系列严格的性能比较进行评估，它取得了良好的效果 该模型还确定了可能介导远端调控相互作用的特定TF PPIs, 对增强子调控的新机制有了深刻理解.

除上述提到的方法外，其他基于表观基因组信号和表达数据，建立的预测EPI和ETG的综合分类器有McEnhancer[72]，EP_Bayes[73]，EAGLE[74]，DIRECT-NET[75]等. McEnhancer通过一种半监督学习算法将目标基因与假定的增强子联系起来，该算法基于丰富的序列特征预测基因表达模式. EP_Bayes使用ChIP-seq RNA聚合酶Ⅱ数据作为输入，但对于内含子增强子的识别并不理想. EAGLE可以应用于来自不同物种和许多细胞类型的数据, 使用少量的基因组特征, 具有很高的准确性. 值得注意的是，随着单细胞测序技术的发展，单细胞多组学数据的不断出现为详细研究控制细胞身份的转录调控机制提供了前所未有的机会，尤其是单细胞测序技术scATAC-seq检测染色质转座酶可及性的高通量分析的最新进展已经实现了数万个单细胞的染色质可及性景观分析. DIRECT-NET则是利用这些数据集合在单细胞水平上剖析了CREs, 包括增强子，与基因的关系；它是一种梯度提升的机器学习方法，从可同时获得的单细胞基因表达和染色质可及性数据，或单独从单细胞染色质可及性数据，通过如下过程识别全基因组CREs及其与靶标基因的关系. 首先, 基于真实功能CREs的开关状态应该显著改变其靶标基因表达模式的特点，它在所有可访问的染色质区域中识别功能CREs. 利用XGBoost模型对染色质可及性评分与基因表达值(或启动子可及性评分)之间的非线性预测模型进行描述，将CREs识别问题和CREs与靶标基因相互作用预测问题转化为模型选择问题. 使用独立功能基因组学数据(如ChIA-PET，Hi-C，Hi-ChIP和ChIP-seq)以及来自全基因组关联分析的疾病相关遗传变异的广泛基准分析表明，DIRECT-NET能够表征转录调控子并揭示细胞状态特异性调控机制.

近几年来，神经网络算法在计算机图像识别、语音识别、自然语言处理等领域都取得了非常大的进展，并被应用于生物学问题，如预测DNA可及性以及识别调控区域和蛋白质结合位点等[76,77]. 以往深度神经网络在生物领域的成功应用启发着越来越多的研究者设计一个深度学习模型来检测调控元件之间的相互作用，利用深度神经网络自动学习有意义的特征模式的优势, 捕捉高层次的情景依赖关系. DeepTACT[78]应用自助采样法深度学习模型来整合基因组序列和染色质可及性数据，在单个调控元件上预测染色质接触. DeepTACT不仅可以推断启动子-增强子相互作用，而且可以推断启动子-启动子相互作用. 从具体的模型设计来看，它是由一个基于CNN的特征提取模块和一个基于RNN的整合模块组成的深度学习模型. 首先，模型将一个增强子或者启动子序列信息的独热编码表示和对应的染色质可及性作为输入，通过一维卷积和池化操作得到增强子-启动子对的两种特征表示, 然后这

两种特征通过RNN整合模块的双向长短期记忆神经网络和注意力机制进行整合，最后通过全连接层输出最终的预测结果，即调控元件增强子和启动子之间存在相互作用的概率值.

事实上，人们还开发了一些方法来探索预测仅基于序列数据的调控元件之间相互作用的可能性. 例如, Yang等人开发了一种使用词嵌入直接从基因组序列生成特征的预测算法PEP，并训练了一个提升树集成模型. 他们证明该模型可以捕捉真实EPI和非EPI之间的顺序特征[79]. 他们还表明, EPI的序列特征与一些基因组特征互补, 有助于提高性能. 基于深度学习的预测模型SPEID[80], 与DeepTACT类似, 同样将CNN与RNN相结合并只基于增强子和启动子区域的序列特征来预测EPI. 但由于缺乏对通用的基于序列的EPI预测机制的设计预期，SPEID只能有效地预测训练细胞系中的EPI，这可能使得它不适用于其他数据集. EPIANN[81]是一种基于注意力的神经网络模型，可以更加关注有助于EPI的特征，更准确地预测EPI; EP2vec[82]利用自然语言处理的无监督深度学习方法将增强子和启动子序列转化为序列嵌入特征, 并利用有监督分类器对EPI进行预测. EPIVAN[83]则使用预先训练的DNA序列嵌入特征来编码增强子和启动子，也就是说将生物序列分析与预训练技术相结合. 预训练方法在图像处理和自然语言处理中已经非常成功，例如，预训练的词向量包含更丰富和更准确的信息，这有助于模型节省训练时间和计算资源，提高性能，特别是在大规模预测中. 然后，利用一维卷积和门控循环单元提取局部特征和全局特征; 最后，通过注意力机制提高关键特征的贡献, 进一步提高EPIVAN的性能. 尽管上述这些深度学习模型已经取得了实质性的进展，但仍存在一定的局限性. 首先，提到的深度学习模型大多使用独热编码表示将人们感兴趣的启动子和增强子对转换为网络的输入, 这种顺序特征的表示类型存在缺陷, 可能会导致维数灾难. 其次, 现有的模型架构对于处理的问题可能过于简单，无法在特征提取过程中进一步学习鉴别复杂特征. 再次, 在大多数模型中, 启动子和增强子的学习特征被直接连接起来用于后续预测，从而丢失了潜在的相互作用信息. 最后，由于现有深度学习模型参数众多，训练效率相对较低. 为了进一步解决上述存在的问题，研究人员在EPIVAN的基础上进一步改进模型，提出了一种新的深度学习预测框架EPI-DLMH[84]. EPI-DLMH由三个主要步骤组成. 首先，使用两层卷积神经网络学习局部特征，使用双向门控循环单元网络捕获启动子和增强子序列的长期依赖关系. 其次, 注意机制用于关注相对重要的特征. 最后, 引入一种匹配启发式机制来探索增强子和启动子之间的相互作用. 总的来说, 本段提到的基于序列的这些研究表明，基因组序列本身包含了关于增强子是否与启动子在基因组中长期相互作用的关键信息. 利用深度学习或自然语言处理技术，很有可能探索增强子及其远程目标启动子之间的隐藏信息.

越来越多的证据表明，表观基因组修饰和序列特征可以作为重要的特征信息来预测和识别单个调控元件之间的相互作用，由于它们通过控制DNA可及性和特定蛋白质的募集在转录调控和染色质的折叠过程中起作用. 例如，TransEPI[85]直接从包含增强子-启动子对的大量基因组背景集中获取基因组信号的输入，并使用Transformer编码器捕获远程的依赖关系，以此来尝试解决基于CNN的模型架构在捕获长距离依赖关系方面的效率低于循环神经网络，尤其是低于Transformer的问题. TransEPI证实，大规模基因组背景特征的精细表征的整合对于EPI的预测是至关重要的.

2 基于无监督学习的方法

无监督的学习方法揭示了基于调控元件之间的相关性或者距离等方式预测染色质相作用的自然模式, 这是客观的, 不使用“人工选择”特征. 基于将远端增强子连接到靶标基因的策略，无监督学习方法具体分为基于相关性的方法，以及包括基于距离的策略、矩阵分解的策略，或者关注染色质三维结构等方法的其他打分方法.

2.1 基于相关性的方法

基于相关性的增强子靶标基因相互作用预测算法的基本原理是: 假设增强子的活性随细胞类型变化, 同时在多种细胞类型中，增强子及其靶标基因的活性状态是相关的. 因此，这些算法依赖于涵盖多种条件的大量表观基因组学或转录组数据，来估计描述增强子或基因活性状态的定量评分. 例如, Thurman等人[86]和Ernst等人[87]通过组蛋白修饰与增强子和启动子DHS或给定基因组域中的启动子转录水平的相关性，提出

的两个方法都改进了之前的方法，提高了增强子靶标基因相互作用预测的性能. 具体地，前者根据调控区域的DHS确定了EPI. 在1454901个远端DHS区域使用DNase-seq读取覆盖率, 将其与79个细胞类型的启动子处的DNase-seq信号进行相关分析，以识别靶标基因, 然后使用染色体构象捕获碳拷贝技术进行验证；后者利用TFs表达和基序富集之间的相关性，Ernst等人预测了调节推定靶标基因的细胞类型特异性激活因子和抑制因子. Cicero[88]基于单细胞ATAC-seq数据将增强子链接到其靶标基因. 该算法缓解了技术和基因组距离效应，同时从单细胞染色质可及性图谱中构建全局顺式调控图. 简言之，用户向其提供已经聚类或者伪时间组织的细胞作为输入，最后它计算500 KB以内的所有位点对之间可及性的相关性. 类似地, C3D[89]基于开放染色质区域之间的相关性预测CRE之间的染色质相互作用. 基于染色质可及性、组蛋白修饰和TF结合之间的相关性, Naville等人[90]估计了增强子-靶标基因关联.

事实上，基于活性调节元件的表观基因组修饰与基因表达之间的相关性的方法也已开发出来，以识别增强子和靶标基因的相互作用. 例如，如Shen等人[91]所述，增强子的活性可以用ChIP-seq H3K4me1检测, 靶标基因的活性可以用Pol2 ChIP-seq检测，所以他们分析了19种小鼠细胞类型的增强子和靶标基因的相互作用. 类似地，PreSTIGE[92]方法分别使用H3K4me1 ChIP-seq和RNA-seq来估计增强子和靶标基因的活性. 此外，PreSTIGE是一种多线性结构域模型，将细胞类型特异性增强子与其靶基因联系起来，将CTCF结合位点视为绝缘体. 它可以通过Galaxy作为在线应用程序提供(http://prestige.case.edu/). 虽然它基于简单的相关性策略，但是它采用了一种不同但在概念上相关的方法, 并不是直接衡量相关性. 相反, PreSTIGE专注于选择基于Shannon熵的增强子和基因的细胞类型特异性模式. 然后, 如果在增强子和基因都活跃的细胞类型中也存在匹配项, 则称为ETG. 因此, PreSTIGE与常规的基于相关性的方法不同，因为它明确地优先考虑细胞类型特异性模式. 同时，这个方法描述了增强子基因的相互作用，且重点关注与特定疾病相关的变异增强子. 除PreSTIGE外, 还有一个预测方法ELMER与所有其他基于相关性的方法不同，其通过将甲基化影响的增强子与附近基因的表达相关来识别转录目标. ELMER方法寻找表达水平和增强子的甲基化水平最接近的20个基因，即每个不同甲基化增强子的10个上游基因和10个下游基因的表达水平之间的负相关模式. 对于每个增强子-靶标基因对，通过非参数Mann-Whitney U检验来确定显著的反相关性, 比较根据增强子甲基化水平分组(最高与最低20%患者的增强子甲基化水平)的癌症患者的表达水平. 后来，Silva等人提出了ELMER的修订版本2, 该版本提供了可选的基于web界面和新的监督分析模式, 显示出更好的性能[93,94].

2.2　其他打分方法

一些算法已经实现了其他自定义定量评分或者采用了新的设计思路, 以将靶标基因分配给增强子. 这些方法的共同思想是使用自己设计的打分方式来定义增强子和靶标基因的关联强度，同时考虑多种类型的基因组特征信息. 它们包括基于距离的方法，基于分解的方法, 基于统计的方法. 其中, 基于距离的方法原理简单, 通过计算CREs之间的线性距离将距离增强子最近的基因作为其靶标基因，所以在许多研究中被用作基线方法. 比如，ABC[95]通过将距离效应与增强子-启动子对的接触频率和增强子活性相结合(通过DHS和H3K27ac ChIP-seq数据), 同时为基于表观基因组数据集的多种细胞类型的增强子-基因连接的全基因组图谱提供了框架. GeneHancer[96]利用的是eQTLS、TF靶标基因共表达、eRNA、Capture Hi-C和增强子与靶基因之间的基因组距离的得分. 由于这些指标具有不同的分布和值范围，因此它们与各种数据转换和权重结合在一起.

基于分解的方法使用不同的分解策略从高维信号中提取潜在特征，以基于这些特征之间的关系来识别增强子和靶标基因的相互作用. EpiTensor[97]通过利用更高阶张量分解, 将5种细胞类型的16个染色质标记、RNA-seq和DNase-seq组合在一起, 从中导出特征向量, 并用于计算“关联得分”. 这个分数基本上解释了功能基因组数据在远端基因组位点上模式的相似性，并用于调用相关的峰值. 该分数显示与增强子-启动子对的Hi-C衍生相互作用具有良好的一致性. SWIPE-NMF[98]则构建了一个整合异质数据的矩阵分解框架, 并已用于重建127个人类细胞系中的增强子-启动子网络.

除了基于距离和基于分解的方法, Salviato等人[99]

提出了一个基于统计的方法来定义增强子靶标基因对. 与同为基于统计框架的FOCS不同的是, 考虑了定义染色质三维体系结构的结构域层次结构. 该论文提到迄今为止, 在文献中提出的ETG网络重建算法中, 染色质三维架构并没有得到优化. 该方法假设, 染色质三维结构数据的有效结合将提高ETG的泛化全基因组定义的准确性. 因此, 研究者开发了一个计算和统计框架, 利用功能基因组数据重建ETG的明细图. 他们使用大量的表观基因组学数据集来定义跨多种细胞和组织类型的增强子活性, 以及高分辨率的Hi-C数据. 然后, 该方法同时证明, 染色质三维结构信息的合并提高了定义ETG的准确性. 相反, PEGASUS[100]不依赖任何功能基因组学数据, 因为它完全依赖于进化保守性, 即PEGASUS首先基于序列保守性定义调控元件, 然后使用同基因保守性评分将它们与靶基因联系起来. 因此这个方法可以归于这个类别.

3 总结与展望

增强子调控基因表达对于破译细胞分化和疾病发生发展的机制至关重要. 基于基因组、表观基因组和转录组的计算方法极大地促进了研究者对增强子在转录调控中作用的理解. 本文描述了多种基于多组学数据预测ETG的计算方法(图2, 表1). 尽管截至目前, 这些文献提出的解决方案多种多样, 但仍有一些关键的限制和挑战.

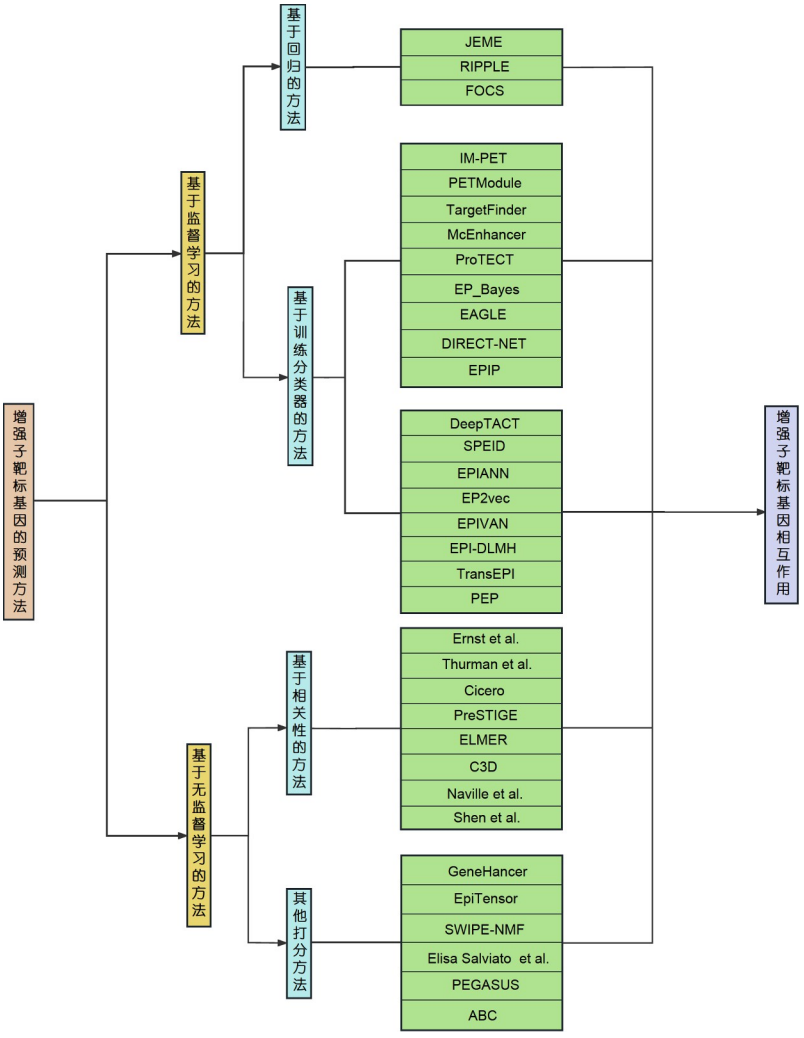

图 2 增强子靶标基因的预测方法分类
Figure 2 Classification of prediction methods for enhancer target genes

表 1　增强子靶标基因的预测方法

Table 1　Prediction methods for enhancer target genes

方法	年份	算法	输入特征	预测结果	参考文献
监督学习					
JEME	2017	回归	表观基因组, RNA	ETGs	[64]
RIPPLE	2015		表观基因组	EPI	[65]
FOCS	2018		表观基因组	EPI	[66]
IM-PET	2014	训练分类器	DNA, RNA, 表观基因组	EPI	[67]
PETModule	2016		表观基因组	ETGs	[68]
TargetFinder	2016		表观基因组	EPI	[69]
McEnhancer	2017		表观基因组	ETGs	[72]
ProTECT	2021		表观基因组, RNA	EPI	[71]
EP_Bayes	2017		表观基因组, RNA	EPI	[73]
EAGLE	2019		表观基因组, RNA	EPI	[74]
DIRECT-NET	2022		表观基因组, RNA	ETGs/EPI	[75]
EPIP	2019		表观基因组	EPI	[70]
DeepTACT	2019		表观基因组, DNA	EPI	[78]
SPEID	2019		DNA	EPI	[80]
EPIANN	2017		DNA	EPI	[81]
EP2vec	2018		DNA	EPI	[82]
EPIVAN	2020		DNA	EPI	[83]
EPI-DLMH	2021		DNA	EPI	[84]
TransEPI	2022		DNA, 表观基因组	EPI	[85]
PEP	2017		DNA	EPI	[79]
无监督学习					
Ernst等人	2011	相关性	表观基因组	ETGs	[87]
Thurman等人	2012		表观基因组	EPI	[86]
Cicero	2018		表观基因组, RNA	ETGs	[88]
PreSTIGE	2018		表观基因组, RNA	ETGs	[92]
ELMER	2018		表观基因组, RNA	ETGs	[94]
C3D	2019		表观基因组	染色质相互作用	[89]
Naville等人	2015		表观基因组	ETGs	[90]
Shen等人	2012		表观基因组	ETGs	[91]
GeneHancer	2017	其他打分	DNA, RNA, 表观基因组	ETGs	[96]
EpiTensor	2016		表观基因组, RNA	ETGs	[97]
SWIPE-NMF	2016		DNA, 表观基因组	EPI	[98]
Salviato等人	2021		表观基因组, DNA	EPI	[99]
PEGASUS	2018		DNA	ETGs	[100]
ABC	2019		表观基因组	ETGs	[95]

第一，缺乏给定生物体基因组中所有非编码区的全基因组详尽参考列表，这些非编码区可以作为增强子. 比如，在整个人类基因组中已经确定了数十万个假定的增强子，特别是在非编码区域，突出了增强子

调控的生物学影响. 虽然已经开发了一系列计算方法来预测细胞类型特异性增强子的基因组位置, 但在不同细胞类型或组织中识别增强子调控的特定靶标基因仍然具有挑战性. 第二, 缺乏大量经过实验验证的真阳性和真阴性ETG集合作为方法开发和基准测试的参考标准. 尽管基于染色体构象捕获分析等尖端技术为人类基因组和其他模式物种中的细胞类型或者组织生成了大规模染色质接触图谱, 高分辨率空间邻近数据仍然有限, 在支出和敏感性方面仍然困难. 数据分辨率限制了对细胞类型特异性增强子-启动子相互作用的深入分析. 具体来说, 首先Hi-C和Capture Hi-C所描述的相互作用基因组锚的分辨率相对较低(约5~10 kb基因组片段), 这使得很难确定参与远端调控的特定增强子. 其次, 虽然Capture Hi-C和ChIA-PET实验可以发现细胞类型或组织特异性增强子调控, 但Hi-C实验产生的数据在不同细胞类型或不同组织中基本不变. 然后, Hi-C和Capture Hi-C数据集的背景噪声水平很高, 导致许多假阳性发现. 最后, 由于对特定蛋白抗体的依赖性, 如CTCF或RNA Pol II, 每个ChIA-PET实验只能描述一组远端相互作用, 导致大量未识别的假阴性相互作用. 第三, 在最先进的计算方法中所用到的策略仍然存在局限性. 基于监督学习的方法主要局限性如下: 分类器的训练需要一组已知的正例和负例样本, 但是因为技术限制, ETG的阳性和阴性集合的定义依赖更多的假设, 从而变得复杂. 接下来面临的问题是, 如果分类器模型训练好, 原则上它们可以用于任何其他细胞类型或组织的ETG. 然而, 实际应用在不同的细胞类型时, 可能因为增强子调控靶标基因的细胞类型特异性, 分类器的性能会差别很大. 同时, 基于回归的方法也依赖于一些超参数的选择. 比如, 窗口的定义以及在每个TSS周围考虑增强子数量的最大值等. 基于无监督学习的方法也存在着诸多限制, 比如基于相关性的方法主要面临着多种细胞的基因组学数据是否在所有条件下都有可以比较的质量和分辨率的问题; 此外, 也存在着细胞类型特异性、评估增强子活性等问题. 其他打分方法的性能主要取决于过多假设和任意定义的参数和权重, 以便能够将多源的信息组合成一个单一的向量或者值. 比如基于距离的方法则忽略了远端调控相互作用, 以及多个增强子针对同一启动子的情况.

综上所述, 在过去的几年中, 人们为了实现增强子和靶标基因对的全面匹配提出了大量的计算生物学解决方案. 随着基因组、表观基因组和转录组数据可用性日益提升, 数据处理能力的进步以及三维基因组研究的深入, 人们将会在面对的挑战方面达成共识, 这将会有助于推动未来几年增强子靶标基因的预测方法研究.

参考文献

1 De Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature, 2013, 502: 499–506
2 Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet, 2019, 20: 437–455
3 Lupiáñez D G, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell, 2015, 161: 1012–1025
4 Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet, 2014, 15: 272–286
5 Buecker C, Wysocka J. Enhancers as information integration hubs in development: lessons from genomics. Trends Genet, 2012, 28: 276–284
6 van Arensbergen J, van Steensel B, Bussemaker H J. In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol, 2014, 24: 695–702
7 Sanyal A, Lajoie B R, Jain G, et al. The long-range interaction landscape of gene promoters. Nature, 2012, 489: 109–113
8 Pennacchio L A, Bickmore W, Dean A, et al. Enhancers: five essential questions. Nat Rev Genet, 2013, 14: 288–295
9 Mumbach M R, Satpathy A T, Boyle E A, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet, 2017, 49: 1602–1612
10 Heinz S, Romanoski C E, Benner C, et al. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol, 2015, 16: 144–154
11 Moore J E, Pratt H E, Purcaro M J, et al. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol, 2020, 21: 17
12 Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. Nature, 2015, 518: 317–330
13 Visel A, Akiyama J A, Shoukry M, et al. Functional autonomy of distant-acting human enhancers. Genomics, 2009, 93: 509–513

14 Tao H, Li H, Xu K, et al. Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. Brief Bioinform, 2021, 22: bbaa405

15 Visel A, Rubin E M, Pennacchio L A. Genomic views of distant-acting enhancers. Nature, 2009, 461: 199–205

16 Dekker J, Rippe K, Dekker M, et al. Capturing chromosome conformation. Science, 2002, 295: 1306–1311

17 Zhao Z, Tavoosidana G, Sjölinder M, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet, 2006, 38: 1341–1347

18 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. Genome Res, 2006, 16: 1299–1309

19 Lieberman-Aiden E, van Berkum N L, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 2009, 326: 289–293

20 Fullwood M J, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem, 2009, 107: 30–39

21 Schoenfelder S, Javierre B M, Furlan-Magaril M, et al. Promoter capture Hi-C: high-resolution, genome-wide profiling of promoter interactions. J Vis Exp, 2018, 136: 57320

22 Capurso D, Tang Z, Ruan Y. Methods for comparative ChIA-PET and Hi-C data analysis. Methods, 2020, 170: 69–74

23 Rao S S P, Huntley M H, Durand N C, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 2014, 159: 1665–1680

24 Jung I, Schmitt A, Diao Y, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet, 2019, 51: 1442–1449

25 Li X, Luo O J, Wang P, et al. Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. Nat Protoc, 2017, 12: 899–915

26 Consortium ENCODE Project. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, 489: 57–74

27 Meuleman W, Muratov A, Rynes E, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature, 2020, 584: 244–251

28 Forrest A R, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. Nature, 2014, 507: 462–470

29 Arnold C D, Gerlach D, Stelzer C, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science, 2013, 339: 1074–1077

30 Melnikov A, Murugan A, Zhang X L, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol, 2012, 30: 271–277

31 Cai Z, Cui Y, Tan Z, et al. RAEdb: a database of enhancers identifified by high-throughput reporter assays. Database, 2019, 2019: bay140

32 Visel A, Minovitsky S, Dubchak I, et al. VISTA enhancer browser—a database of tissue-specifific human enhancers. Nucleic Acids Res, 2007, 35: D88–D92

33 Raisner R, Kharbanda S, Jin L, et al. Enhancer activity requires CBP/P300 bromodomain-dependent histone H3K27 acetylation. Cell Rep, 2018, 24: 1722–1729

34 Bonn S, Zinzen R P, Girardot C, et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet, 2012, 44: 148–156

35 Rada-Iglesias A, Bajpai R, Swigut T, et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature, 2011, 470: 279–283

36 Koch F, Andrau J C. Initiating RNA polymerase II and TIPs as hallmarks of enhancer activity and tissue-specifificity. Transcription, 2011, 2: 263–268

37 Heintzman N D, Stuart R K, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet, 2007, 39: 311–318

38 Tsompana M, Buck M J. Chromatin accessibility: a window into the genome. Epigenet Chromatin, 2014, 7: 33

39 Boyle A P, Davis S, Shulha H P, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell, 2008, 132: 311–322

40 Buenrostro J D, Wu B, Chang H Y, et al. ATAC-seq: a method for assaying chromatin accessibility genome-wide. CP Mol Biol, 2015, 109: 21.29.1–21.29.9

41 Vierstra J, Stamatoyannopoulos J A. Genomic footprinting. Nat Methods, 2016, 13: 213–221

42 Brenowitz M, Senear D F, Kingston R E. DNase I footprint analysis of protein-DNA binding. Curr Protoc Mol Biol, 2001, Chapter 12: Unit 12.4

43 Li Z, Schulz M H, Look T, et al. Identification of transcription factor binding sites using ATAC-seq. Genome Biol, 2019, 20: 45

44 Kim T K, Hemberg M, Gray J M, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature, 2010, 465: 182–187

45 De Santa F, Barozzi I, Mietton F, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol, 2010, 8: e1000384

46 Arner E, Daub C O, Vitting-Seerup K, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science, 2015, 347: 1010–1014

47 Kodzius R, Kojima M, Nishiyori H, et al. CAGE: cap analysis of gene expression. Nat Methods, 2006, 3: 211–222

48 Takahashi H, Lassmann T, Murata M, et al. 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc, 2012, 7: 542–561

49 Valen E, Pascarella G, Chalk A, et al. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res, 2009, 19: 255–265

50 Core L J, Waterfall J J, Lis J T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 2008, 322: 1845–1848

51 Lam M T Y, Cho H, Lesch H P, et al. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature, 2013, 498: 511–515

52 Churchman L S, Weissman J S. Native elongating transcript sequencing (NET-seq). Curr Protoc Mol Biol, 2012, 98: unit 4.14.1–unit 4.14.17

53 Kwak H, Fuda N J, Core L J, et al. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science, 2013, 339: 950–953

54 Mahat D B, Kwak H, Booth G T, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc, 2016, 11: 1455–1476

55 Nechaev S, Fargo D C, dos Santos G, et al. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. Science, 2010, 327: 335–338

56 Schwalb B, Michel M, Zacher B, et al. TT-seq maps the human transient transcriptome. Science, 2016, 352: 1225–1228

57 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv, 2014, 1409.1556

58 Shen X, Wang Y, Lin M, et al. DeepMAD: mathematical architecture design for deep convolutional neural network. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver. 2023. New York: IEEE, 2023. 6163–6173

59 Zhu L, Wang X J, Ke Z H, et al. BiFormer: vision transformer with bi-level routing attention. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver. 2023. New York: IEEE, 2023. 10323–10333

60 Zhang Y, Guo X, Poggi M, et al. Completionformer: depth completion with convolutions and vision transformers. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver. 2023. New York: IEEE, 2023. 18527–18536

61 Takashima R, Hayamizu N, Inoue H, et al. Visual atoms: pre-training vision transformers with sinusoidal waves. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver. 2023. New York: IEEE, 2023. 18579–18588

62 Zhou W, Jiang Y, Cui P, et al. RecurrentGPT: interactive generation of (arbitrarily) long text. arXiv, 2023, 2305.13304

63 Tang L, Zhong Z, Lin Y, et al. EPIXplorer: a web server for prediction, analysis and visualization of enhancer-promoter interactions. Nucleic Acids Res, 2022, 50: W290–W297

64 Cao Q, Anyansi C, Hu X, et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet, 2017, 49: 1428–1436

65 Roy S, Siahpirani A F, Chasman D, et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res, 2015, 43: 8694–8712

66 Hait T A, Amar D, Shamir R, et al. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol, 2018, 19: 56

67 He B, Chen C, Teng L, et al. Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci USA, 2014, 111: E2191–E2199

68 Zhao C, Li X, Hu H. PETModule: a motif module based approach for enhancer target gene prediction. Sci Rep, 2016, 6: 30043

69 Whalen S, Truty R M, Pollard K S. EPIP: a novel approach for condition-specific enhancer-promoter interaction prediction. Nat Genet, 2016, 48: 488–496

70 Talukder A, Saadat S, Li X, et al. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Bioinformatics, 2019, 35: 3877–3883

71 Wang H, Huang B, Wang J. Predict long-range enhancer regulation based on protein-protein interactions between transcription factors. Nucleic Acids Res, 2021, 49: 10347–10368

72 Hafez D, Karabacak A, Krueger S, et al. McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. Genome Biol, 2017, 18: 199

73 Dzida T, Iqbal M, Charapitsa I, et al. Predicting stimulation-dependent enhancer-promoter interactions from ChIP-Seq time course data. PeerJ, 2017, 5: e3742

74 Gao T, Qian J, Charapitsa I, et al. EAGLE: an algorithm that utilizes a small number of genomic features to predict tissue/cell type specific enhancer-gene interactions. PLoS Comput Biol, 2019, 15: e1007436

75 Zhang L, Zhang J, Nie Q. DIRECT-NET: an efficient method to discover *cis*-regulatory elements and construct regulatory networks from single-cell multiomics data. Sci Adv, 2022, 8: eabl7393

76 Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods, 2015, 12: 931–934

77 Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 2015, 33: 831–838

78 Li W, Wong W H, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res, 2019, 47: e60

79 Yang Y, Zhang R, Singh S, et al. Exploiting sequence-based features for predicting enhancer-promoter interactions. Bioinformatics, 2017, 33: i252–i260

80 Singh S, Yang Y, Póczos B, et al. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Biol, 2019, 7: 122–137

81 Mao W, Kostka D, Chikina M. Modeling enhancer-promoter interactions with attention-based neural networks. bioRxiv, 2017, 219667

82 Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics, 2018, 19: 84

83 Hong Z, Zeng X, Wei L, et al. Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics, 2020, 36: 1037–1043

84 Min X, Ye C, Liu X, et al. Predicting enhancer-promoter interactions by deep learning and matching heuristic. Brief Bioinform, 2021, 22: bbaa254

85 Chen K, Zhao H, Yang Y. Capturing large genomic contexts for accurately predicting enhancer-promoter interactions. Brief Bioinform, 2022, 23: bbab577

86 Thurman R E, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. Nature, 2012, 489: 75–82

87 Ernst J, Kheradpour P, Mikkelsen T S, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 2011, 473: 43–49

88 Pliner H A, Packer J S, McFaline-Figueroa J L, et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell, 2018, 71: 858–871.e8

89 Mehdi T, Bailey S D, Guilhamon P, et al. C3D: a tool to predict 3D genomic interactions between *cis*-regulatory elements. Bioinformatics, 2019, 35: 877–879

90 Naville M, Ishibashi M, Ferg M, et al. Long-range evolutionary constraints reveal *cis*-regulatory interactions on the human X chromosome. Nat Commun, 2015, 6: 6904

91 Shen Y, Yue F, McCleary D F, et al. A map of the *cis*-regulatory sequences in the mouse genome. Nature, 2012, 488: 116–120

92 Yizhar-Barnea O, Valensisi C, Jayavelu N D, et al. DNA methylation dynamics during embryonic development and postnatal maturation of the mouse auditory sensory epithelium. Sci Rep, 2018, 8: 17348

93 Yao L, Shen H, Laird P W, et al. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol, 2015, 16: 105

94 Silva T C, Coetzee S G, Gull N, et al. ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. Bioinformatics, 2019, 35: 1974–1977

95 Fulco C P, Nasser J, Jones T R, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet, 2019, 51: 1664–1669

96 Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database, 2017, 2017: bax028
97 Zhu Y, Chen Z, Zhang K, et al. Constructing 3D interaction maps from 1D epigenomes. Nat Commun, 2016, 7: 10812
98 Chen Y, Wang Y, Xuan Z, et al. *De novo* deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. Nucleic Acids Res, 2016, 44: e106
99 Salviato E, Djordjilović V, Hariprakash J M, et al. Leveraging three-dimensional chromatin architecture for effective reconstruction of enhancer-target gene regulatory interactions. Nucleic Acids Res, 2021, 49: e97
100 Clément Y, Torbey P, Gilardi-Hebenstreit P, et al. Genome-wide enhancer-gene regulatory maps in two vertebrate genomes. bioRxiv, 2018, 244475

Computational methods to predict Enhancer-target Gene Pairs

XU XiaoQiang, CUI Ting, ZHANG Han, SHANG DeSi & LI ChunQuan

Artificial Intelligence and Big Data Center, The First Affiliated Hospital of Hengyang Medical School, University of South China, Hengyang 421001, China

Enhancers are functional DNA fragments that act as cis regulatory elements in the genome and that regulate gene expression through interactions with promoters of target gene. Identifying the interaction between enhancer and target gene is important for understanding the mechanisms of gene regulation, cell differentiation, and disease development. However, identification of enhancer-target gene (ETG) by experimental methods is a time-consuming and labor-intensive work. Therefore, more and more research is focused on developing computational methods to solve this problem. This review systematically summarizes the computational methods used for ETG prediction to promote their application. Finally, we discussed the limitations and challenges of existing solutions proposed in this field and looked forward to future research.

enhancers, gene expression, enhancer and target gene, computational method

doi: 10.1360/SSV-2023-0086