PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2014, 41(7): 640~648
www.pibb.ac.cn

基于表观基因组学的 DNA 元件鉴定方法研究进展 *

卢一鸣　屈武斌　张成岗 **
(军事医学科学院放射与辐射医学研究所，蛋白质组学国家重点实验室，北京 100850)

摘要　在人类基因组测序已经完成的“后基因组”时代，对基因组序列的功能注释，尤其是各种 DNA 调控元件的鉴定，已成为进一步理解人类基因组复杂机制的瓶颈问题．最近，针对染色质状态图谱的大规模研究工作，揭示了各类 DNA 元件特征性的染色质修饰标记．这些研究结果推动了一系列基于有监督和无监督学习的 DNA 元件预测方法的产生，其中一些方法已经成功应用于多个基因组的 DNA 元件预测，并且已成为未知基因组的常规注释工具．这些预测方法因其算法特点和预测策略不同而适用于不同类型的 DNA 元件预测任务．大多数情况下，使用者需要联合使用多个预测方法来达到预测敏感性和特异性的平衡．尽管各类算法在 DNA 元件预测中都有一些成功的应用，但每一类算法都有其特有的弊端，需要使用者认真避免．本文回顾了前期和当下 DNA 元件预测方法的主要类型，全面分析了各类方法的优缺点，指出了下一步可以改进的方向．本综述中的分析和观点有助于读者深入理解 DNA 元件预测算法的主要原则，进而在相关研究中更好地应用这些方法．

关键词　DNA 元件，基因表达调控，表观基因组学，机器学习
学科分类号　Q751，Q03　　**DOI**: 10.3724/SP.J.1206.2013.00248

人类基因组的 DNA 序列编码着人体内每一个细胞的遗传信息，而其中只有约 1.5%的区域包含编码蛋白质的信息，其余 98.5%的人类基因组区域虽不编码任何蛋白质，却包含大量的具有生物学功能的调控序列(regulatory sequence)[1-3]．DNA 调控元件是 DNA 调控序列的最重要组成部分，控制着染色体上特定基因的转录和表达[4-5]．在“人类基因组计划”已经完成的后基因组时代，如何准确地鉴定这些 DNA 调控元件成为了进一步揭示人类基因组奥秘的首要问题．

人体中不同组织器官的细胞共享同一套基因组，却拥有截然不同的细胞形态和生理功能，其中主要原因要归结于染色质上广泛存在的表观遗传修饰[3]．表观遗传修饰不但在分化发育过程中细胞谱系的建立和维持起到关键作用，而且与 DNA 的复制、修复以及人类疾病密切相关[6-8]．表观遗传修饰虽然种类较多，但其发挥调控作用的机制比较一致，即通过改变局部染色质，特别是 DNA 调控元件的状态来发挥调控功能．DNA 调控元件广泛参与了基因转录等过程中的 DNA-DNA、DNA- 蛋白质相互作用，其状态的改变会直接影响其与这些分子之间的结合效率[9]．因此表观遗传修饰是除序列以外 DNA 调控元件最重要的生物学特征，是鉴定人类基因组中的各类 DNA 调控元件的重要识别信号．

已有研究表明，不同的 DNA 调控元件可以被特异的组蛋白修饰模式所表征．例如，启动子区域通常出现组蛋白 H3 第四位赖氨酸的三甲基化(H3K4me3)、组蛋白 H3 的乙酰化(H3ac)和组蛋白 H4 的乙酰化(H4ac)的富集，并伴随着组蛋白 H3 第四位赖氨酸的单甲基化(H3K4me1)的缺失[10]，而增强子区域则表现出 H3 第四位赖氨酸的单甲基化和

* 国家重点基础研究发展计划(973)(2012CB518200)，国家自然科学基金(30900862, 30973107, 81070741, 81172770)，蛋白质组学国家重点实验室自主研究及开放课题(SKLP-O201104，SKLP-K201004，SKLP-O201002)和国家科技重大专项(2012ZX09102301-016)资助项目．
** 通讯联系人．
Tel: 010-66931590, E-mail: zhangcg@bmi.ac.cn
收稿日期：2013-10-11，接受日期：2013-12-18

双甲基化(H3K4me1、H3K4me2)、H3ac 和 H4ac 的富集，伴随 H3K4me1 的缺失[10]，利用这些特异的表观遗传修饰特征可以用来准确地区分启动子和增强子两类 DNA 元件．He 等[11]对人类前列腺癌细胞的相关研究显示，H3K4me2 特异地出现在已知的雄激素受体结合位点，表明 H3K4me2 是增强子的重要表观修饰特征．Ernst 等[12]利用 9 个人类细胞系的 8 种组蛋白修饰的存在或缺失状态将人类基因组划分为 15 种不同状态，其中包含 8 种不同的 DNA 元件及其激活 / 抑制状态．基于这些重要的表观遗传修饰特征，一系列的 DNA 元件鉴定方法和工具被开发出来，实现了对基因组中 DNA 元件的准确鉴定，并且已经在人、小鼠、果蝇和秀丽线虫等多种模式生物的基因组注释项目中发挥了重要作用[1, 13–16].

本文首先介绍了 DNA 调控元件的主要类型及其序列和功能特征等背景知识，并概括了表观遗传组学出现之前 DNA 调控元件的主要鉴定方法与存在问题，然后详细分析了近几年随表观遗传组学快速发展而出现的有监督学习和无监督学习两类 DNA 调控元件鉴定策略，最后分析了两类策略存在的问题和解决方法.

1　DNA 调控元件的主要类型及生物学特征

DNA 调控元件是指 DNA 上对特定基因的表达具有调控作用的区域．DNA 调控元件本身不编码任何蛋白，仅仅提供一个作用位点，通过与其他 DNA 区域或蛋白质分子相互作用而起作用．真核细胞的 DNA 调控元件按功能特性可以分为启动子、增强子和沉默子等．下面依次介绍这几类调控元件的结构和功能特点.

1.1　启动子(Promoter)

启动子是 RNA 聚合酶结合位点并启动转录的 DNA 序列，通常分布在转录起始位点(transcription start site，TSS)上下游 1 000 bp 以内．启动子区的核心结构包括一个转录起始位点和通用转录因子结合位点，如 TATA 盒等，此外有些启动子还包含特异转录因子结合位点(transcription factor binding sites，TFBS)．除了序列上的特征以外，启动子区通常还存在特异的表观遗传修饰特征，Heintzman 等[10]发现启动子区域通常伴随着 H3K4me3、H3ac 和 H4ac 的富集以及 H3K4me1 的缺失．同时，这些修饰类型的组合方式还与启动子激活或抑制的状态密切相关．Barski 等[17]和 Wang 等[18]在分别研究了 $CD4^+$ T 细胞中组蛋白的 21 种甲基化和 18 种乙酰化修饰分布后发现，一些表观遗传修饰类型高频出现在激活的启动子区，包括：H3K4me1、H3K4me2、H3K4me3、H3K9me1、H3K9ac 和 H3K27ac 等．而另一些修饰类型则特异性地与抑制的启动子相关，包括：H3K27me2、H3K27me3、H3K9me2 和 H3K9me3 等.

1.2　增强子(Enhancer)

增强子是 DNA 上一小段可与蛋白质(通常为转录因子)结合的区域，与蛋白质结合之后，可以增强启动子的转录活性，是决定基因的时间、空间特异性表达的重要调控元件．增强子可能位于基因上游，也可能位于下游．且不一定接近所要作用的基因，甚至不一定与基因位于同一染色体[19]．Heintzman 等[10]的发现表明，增强子区域也存在特异的表观遗传修饰特征，表现为 H3K4me1、H3K4me2、H3ac 和 H4ac 的富集以及 H3K4me1 的缺失．值得注意的是，最近的研究同时也显示基因组中增强子的功能和机制具有较大的变异性[20–21]，具体表现在表观遗传修饰模式的不同[17–18, 22]，提示增强子可能会对应多种表观遗传修饰特征.

1.3　沉默子(Insulator)

沉默子是一种可以阻断启动子和增强子相互作用的负性调控元件，当其结合特异蛋白因子时，对基因转录起阻遏作用．沉默子通常位于基因上游 20～2 000 bp 的位置，一小部分会出现在基因内部．沉默子能够使染色体上两个临近基因的转录过程不受到彼此的影响，从而具有截然不同的表达活性．已有研究表明，转录因子 CTCF(CCCTC-binding factor)所介导的特定 DNA 三维结构是沉默子发挥作用的必要条件[23–24]，其与基因组的结合状态可以用来准确地预测沉默子[12, 25].

2　利用基因组序列鉴定 DNA 调控元件的方法分析与总结

早期使用计算方法对 DNA 调控元件进行鉴定的工作通常是基于 DNA 序列特征进行的，这些鉴定方法按其研究目的大致分为两类：启动子区域搜寻和启动子调控序列鉴定，两类方法中具有代表性的工具见表 1．识别启动子区域等同于识别转录起始位点的位置，其基本原理比较简单：即搜索核心序列如 TATA 盒或者已知的转录因子结合位点等．鉴定启动子中调控序列的流程基本如下：首先确定启动子区域的大致范围，通常取基因开放阅读

框的起始密码子上游 500～1 000 bp 的范围；然后确定一类具有共同特征(如相似的表达谱)的基因，并获取它们的启动子区序列；接着使用多重序列比对的方法鉴定上述序列中共同出现的模体(motif)，并计算模体每个位置的碱基比例，生成相应的权重矩阵(weight matrix). 模体的权重矩阵可以用来预测新的转录因子结合位点，常用的转录因子数据库有 TRANSFAC[26]和 JASPAR[27]等.

Table 1 List of promoter region searching and motif identification softwares

表 1 启动子区域搜寻和启动子调控序列鉴定软件列表

	软件名称	算法描述	网址	参考文献
启动子区域搜寻工具	Promoter2.0	人工神经网络	www.cbs.dtu.dk/services/Promoter/	[28]
	NNPP	时滞神经网络	www.fruitfly.org/seq_tools/promoter.html	[29]
	PromoterInspector or McPromoter	短片段聚类算法 随机片段网络模型	www.genomatix.de/online_help/help_gems/PromoterInspector_help.html tools.igsp.duke.edu/generegulation/McPromoter/	[30] [31]
	CorePromoter	判别式分析	rulai.cshl.org/tools/genefinder/CPROMOTER/index.htm	[32]
启动子调控序列鉴定工具	RSA Tools	强力搜索	rsat.ulb.ac.be/rsat/	[33]
	Gibbs sampler	序列比对	ccmbweb.ccv.brown.edu/gibbs/gibbs.html	[34]
	MEME	最大期望的比对	meme.nbcr.net/meme/	[35]
	BBA	系统发育足迹的贝叶斯比对	ccmbweb.ccv.brown.edu/cgi-bin/bayes_align_ccmb.pl	[36]
	PipMaker	系统发育足迹的本体图	pipmaker.bx.psu.edu/pipmaker/	[37]

对于其他类型 DNA 调控元件(如增强子)的预测原理与启动子区预测类似，都是通过搜寻一致性出现的模体序列(大部分为转录因子结合位点)，与启动子区不同的是，增强子没有固定的标志性序列，因此获得准确的位置信息更加困难，一般通过多物种的保守性分析方法来预测其位置范围，总体的鉴定效果并不能达到实用的水平. 此方面最常用的一个数据库是 VISTA Enhancer Browser[38]，该数据库整合了大量基于实验的增强子信息，具有一定的参考价值.

早期的 DNA 调控元件鉴定方法在表观遗传组学实验技术出现之前得到了较为充分的发展，取得了一定的研究成果，但由于这些算法能够利用的数据来源有限，仅限于基因组的序列信息，预测准确度并不理想. 一致性出现的模体序列虽然可能是潜在的转录因子结合位点，但该类序列也并非是所有启动子都具有的. 此外，大部分 DNA 调控元件鉴定算法都假设这些元件在进化中具有保守性的倾向，而相关的研究表明事实并非如此[39]. 由于以上所述诸多缺点，单一基于基因组序列的 DNA 元件鉴定方法的可靠性不能得到保证，在基因表达调控领域的应用受到了很大的限制. 最近，Lee 等[40]将基于序列预测增强子与表观遗传修饰数据结合起来，使对增强子的预测准确度有了大幅度的提升，为该基于基因组序列的 DNA 元件预测方法的未来发展方向提供了一个重要参考.

3 利用表观遗传组学数据鉴定 DNA 调控元件的方法介绍

染色质免疫共沉淀技术(chromatin immunoprecipitation，ChIP)是研究 DNA 与蛋白质相互作用最主要的实验方法. 染色质免疫共沉淀技术与高通量实验技术(如 DNA 芯片和 DNA 测序技术)结合起来，如 ChIP-chip 和 ChIP-seq 技术，可以大规模地检测染色质的蛋白质结合状态，成为了表观遗传组学中最重要的高通量技术手段. 如本文前言所述，DNA 调控元件上的各类表观遗传修饰是重要的识别标志，联合多种表观遗传修饰类型可以用来有效地鉴定 DNA 调控元件的位置和类型. 因此，随着近几年 ChIP-chip 和 ChIP-seq 实验数据的大量增加，特别是两个大规模合作项目 The Encyclopedia of DNA Elements (ENCODE) (网址：http://genome.ucsc.edu/ENCODE/)[2]和 NIH Roadmap Epigenomics(网址：http://www.roadmapepigenomics.org/)[41]计划对上百种人类细胞系的常见表观遗传修饰和转录因子进行 ChIP-seq 实验检测，并已经完成了其中的十几种细胞系主要修饰类型的检测. 在这些数据的驱动下，近几年发展出了一批基于表观

遗传组学数据鉴定 DNA 调控元件的方法，并且已在基因组注释研究中发挥了重要作用[1, 13-16]. 这些方法通常基于类似的表观遗传修饰特征，其按照算法特点可以分成两类：无监督学习方法和有监督学习方法，其中代表性工作已在表 2 中列出. 下面将详细介绍这些方法所采用的机器学习特征，以及每类方法的主要代表性软件和相应特点.

Table 2 List of supervised and unsupervised DNA elements iddentification

表 2 无监督学习和有监督学习两类 DNA 元件鉴定软件列表

	软件名称	算法描述	调控元件类型	表观修饰特征类型	准确度 /%	网址	参考文献
无监督学习类	ChromHMM	隐马尔科夫模型	启动子、增强子、沉默子	组蛋白修饰、CTCF 位点	–	compbio.mit.edu/ChromHMM/	[48]
	Segway	动态贝叶斯网络	启动子、增强子、沉默子	组蛋白修饰、CTCF 位点	–	noble.gs.washington.edu/proj/segway/	[43]
	ChAT	动态编程	启动子、增强子	组蛋白修饰	41.7	–	[49]
	ChromaSig	似然函数聚类	启动子、增强子	组蛋白修饰、组蛋白分布	62.6	bioinformatics-renlab.ucsd.edu/rentrac/wiki/ChromaSig	[50]
	CoSBI	子空间聚类	启动子、增强子	组蛋白修饰、组蛋白变体	–	www.healthcare.uiowa.edu/labs/tan/CoSBIWebpage.html	[51]
有监督学习类	PM	分布谱相关性	启动子、增强子	组蛋白修饰	39.5	–	[10]
	CSI-ANN	人工神经网络	增强子	组蛋白修饰	66.3	www.medicine.uiowa.edu/Labs/tan/	[52]
	He's method	支持向量机	增强子	组蛋白修饰	–	–	[11]
	ChromaGenSVM	支持向量机	增强子	组蛋白修饰	～90	sysimm.ifrec.osaka-u.ac.jp/download/Diego/	[44]
	Won's method	隐马尔科夫模型	启动子、增强子	组蛋白修饰	～80	nash.ucsd.edu/chromatin.tar.gz.	[53]
	BNFinder	贝叶斯网络	增强子	组蛋白修饰、PolⅡ位点	78.0	bioputer.mimuw.edu.pl/software/bnf/	[54]
	CSS-RF	随机森林	增强子	p300 位点、Gli3 位点	58.0	–	[55]
	Yip's method	随机森林	启动子、增强子	组蛋白修饰	67.0	metatracks.encodenets.gersteinlab.org/	[56]
	RFECS	随机森林	增强子	组蛋白修饰	～90	enhancer.ucsd.edu/renlab/RFECS_enhancer_prediction/	[45]

3.1 表观遗传修饰的特征提取

目前，基于表观遗传学特征鉴定 DNA 元件的方法，一般采用表观遗传修饰在一个区域内的分布强度作为学习特征. 具体来讲，整个基因组序列通常被划分为一系列连续的小区段(称为 bin)，并统计每个 bin 中被覆盖的 ChIP-seq 读长数，作为某种表观遗传修饰在该区段中的分布强度的量化. 这些读长数通常被标准化，部分算法还将每个 bin 的数目根据 Poisson 分布进行二值化[42-43]. 这些定量的强度特征可以直接作为机器学习方法的特征参数. 有些方法采用滑动窗口策略(sliding-window strategy)对 DNA 元件进行训练和预测，这是通常

将每个 bin 对应的强度值在一个较大的窗口内进行平均后作为机器学习方法的特征参数[10, 18, 28]. 除此之外，一些方法还将这些 bin 的强度值进行一定的变换[28]. 在利用表观遗传修饰特征进行 DNA元件鉴定时，通常将多种修饰类型的强度特征组合起来进行鉴定，但由于表观遗传修饰类型繁多，寻找对应于每一种 DNA 元件的最优修饰类型的组合也是一个重要的问题. 一些算法通过对特征变量的筛选来寻找最优特征变量组合，取得了良好的预测效果[44-45].

3.2 无监督学习方法类介绍

无监督学习方法的特点是不需要预先学习已知的各类 DNA 调控元件的特征，而是直接根据一段 DNA 区域的表观遗传修饰特征进行自动分类. ChromHMM 是这类算法中表现突出的一个软件，在多个研究中得到了很好的应用[12, 14, 42, 46]，并且其预测结果已作为人类基因组的参考注释在 UCSC Genome Browser[47]中的“Broad ChromHMM”一栏中进行了可视化展示(网址：http://genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18). 该软件的算法核心是一个多变量的隐马尔可夫模型(hidden markov model，HMM). 它首先在基因组 DNA 序列上扫描获取多种表观遗传修饰强度数据，同时监测 HMM 模型在不同状态之间的切换信息，并据此将染色质分为不同的状态，然后根据每种状态在基因组上的位置分布判断属于哪一种 DNA 调控元件或其他区域类型. 类似的软件还有 Segway，不同的是 Segway 采用动态贝叶斯网络(dynamic bayesian network，DBN)的算法，DBN 算法相比 HMM 模型可以更好地处理表观遗传修饰之间可能存在的复杂相互作用关系. 其他一些算法(如 ChAT、ChromaSig 和 CoSBI 等)大多采用无监督的聚类方法对染色体区域进行自动分类和注释.

无监督学习的鉴定方法最主要的特点是算法应用条件限制少，不需要提前借助已知各类 DNA 调控元件的表观遗传修饰信息，自动对染色质状态进行分类，可以对一些新的未知 DNA 调控元件类型进行预测. 当需要对基因组中的调控元件进行全面鉴定时，无监督学习方法是比较理想的选择.

3.3 有监督学习方法类介绍

有监督学习方法首先对已知的 DNA 调控元件进行特征提取和筛选，然后将其中有效的特征变量在已知的 DNA 调控元件中进行学习和训练，最后将学习后的特征用来预测新的基因组序列中的 DNA 调控元件. ChromaGenSVM 是这类方法中的典型代表，其对于增强子元件的预测准确性在同类软件中处于领先水平，该方法的主要优势在于采用了遗传算法对鉴定增强子的表观遗传修饰的最优化组合进行搜寻，并对算法参数进行了最优化处理，其对 $CD4^+$ T 细胞中增强子的预测准确度达到 90%，显著高于其他同类算法[44]. ChromaGenSVM 软件的基本预测流程为：首先，基因组 DNA 序列连续的 bin(长度为 100 bp)，并计算每个 bin 中各种表观遗传修饰的分布强度，具体如 **3.1** 中所述；接着，利用支持向量机(support vector machine，SVM)算法，针对不同表观遗传修饰类型，在增强子上下游 1 kb 区域内所有 bin 强度值的平均值进行训练，并使用遗传算法和交叉验证方法对表观遗传修饰的组合以及 SVM 的参数进行最优化；然后，在模型训练成功后，采用滑动窗口技术以相同窗宽(1 kb)在基因组 DNA 序列上进行滑动以预测潜在的 DNA 调控元件. 其他的同类算法(如 PM、CSI-ANN、RFECS 等)与 ChromaGenSVM 方法的区别主要体现在机器学习的具体算法类型上，而在特征的选取上较为一致，即主要关注表观遗传修饰的强度信息.

有监督学习鉴定方法的主要优点在于有较高的预测准确性，由于这类算法一般都对已知 DNA 调控元件的特征进行了充分的学习和训练，因此在鉴定时可以保证较高的准确性. 并且大部分有监督的学习方法还可以给出正确判断的概率值(*P*-value)或者错误判断的期望值(*E*-value). 当需要对一两种已知的 DNA 调控元件类型进行准确地鉴定时，有监督的学习方法通常是较为理想的选择. 有监督学习鉴定方法的缺点在于需要预知 DNA 调控元件的位置和表观遗传修饰等信息，在一定程度上限制了该类算法的应用.

4 DNA 调控元件鉴定中存在的问题和应对策略

以上介绍了无监督学习和有监督学习两类利用表观遗传组学数据鉴定 DNA 调控元件的方法及其各自的特点，下面将主要阐述这两类算法各自存在的主要问题、共有问题和相应的解决策略.

4.1 无监督学习方法存在的问题和对策

无监督学习类型的 DNA 调控元件鉴定方法主要存在的问题包括：a. 无监督学习类方法未对已知 DNA 调控元件的各类特征进行学习，因而在鉴

定准确率上不能达到较好的效果，这也是无监督学习算法的固有性质所决定的；b．上述无监督学习方法一般都需要人为设定模型的总状态数目，而目前尚没有对总状态数目估计的成熟方法；c．对无监督学习方法的分类结果进行解释时，仍需要借助已知的 DNA 调控元件的基因组分布特点和表观遗传修饰特征等信息.

无监督学习方法鉴定准确性问题的解决，主要依赖于核心算法对于各类 DNA 调控元件生物学特征的把握程度，如果可以较为全面地将 DNA 调控元件的特征变量包含到模型中，将有利于进一步提高算法的鉴定准确性；关于总状态数的估计问题，Ernst 等[12]在一项相关研究中使用了多个细胞中模型预测的状态一致性指标来衡量总状态数目的方法，具有一定的参考借鉴意义.

4.2 有监督学习方法存在的问题和对策

有监督学习类型的 DNA 调控元件鉴定方法存在的主要问题包括：a．有监督学习方法的鉴定都是基于对已知 DNA 调控元件的特征学习，因此其对于新的未知 DNA 调控元件几乎没有鉴定能力；b．有监督学习方法进行特征学习时往往关注 DNA 调控元件周围固定长度范围的特征，鉴定时也是对固定长度的基因组片段进行预测，而且该固定长度一般较长，造成其对于 DNA 调控元件的边界判断不够准确；c．有监督学习方法在对已知 DNA 调控元件的特征进行学习时，容易出现“过学习”的现象，可能会给后期鉴定过程引入一定的偏性.

由于单一的有监督学习无法鉴定新 DNA 调控元件，因此实际研究中需要借助无监督学习方法的鉴定结果，两种方法联合使用可以有效地解决这一问题．对 DNA 调控元件边界的准确判断，需要今后的有监督学习鉴定方法考虑采用动态窗框技术或考虑更多的不依赖于区域长度的特征．另外，有监督学习在对模型的训练时需要采用多个独立数据集或者交叉验证的方法，减少“过学习”现象造成的影响.

4.3 两类鉴定方法共同存在的问题和对策

除此之外，一些问题在两类鉴定方法中是共同存在的．例如：目前所有的方法对表观遗传修饰的特征提取比较单一，都仅仅关注修饰的强度信息，但大量的研究结果表明，表观遗传修饰在 DNA 调控元件周围的分布形状也具有非常明显的特征．Schones 等[57]的研究显示，有转录因子结合的增强子区域会形成核小体移位，进而造成组蛋白修饰的 ChIP-seq 读长(reads)分布呈明显的双峰形状特征．另外，Heintzman 等[10]的研究表明，H3K4me1 和 H3K4me3 组蛋白修饰在启动子区和增强子区的 ChIP-seq 信号峰尖锐程度有明显的差别，H3K4me1 在启动子区呈宽而矮的形状，在增强子区呈窄而高的形状，而 H3K4me3 则恰好相反．这些证据表明，表观遗传修饰的分布形状与 DNA 调控元件有密切的关系．在 DNA 调控元件的鉴定中考虑 ChIP-seq 信号峰的形状分布特征将有可能提高无监督学习类和有监督学习类方法的鉴定表现.

5 结　　语

DNA 调控元件的鉴定作为研究复杂基因调控网络的前提和基础，是分子生物学长期以来致力研究的重要问题．DNA 调控元件的鉴定方法本身也经过了长期的发展和演化，在鉴定准确性和实用性上都得到了不断提高.

本文首先介绍了 DNA 调控元件的主要类型，以及各类型元件在 DNA 序列和表观遗传修饰上的特征，这些生物学特征是鉴定 DNA 调控元件的基本出发点和依据．本文接着分析了早期基于基因组 DNA 序列的鉴定方法，描述了此类方法的基本流程和原理，并深入分析了其缺点和问题．本文重点分析和总结了表观遗传组学实验技术出现之后发展出的无监督学习和有监督学习两类鉴定方法，介绍了各自的代表性工作及基本流程．最后，本文详细地分析了每类算法所存在的问题和共同存在的问题，并提出了相应的解决策略．本实验室在对 DNA 调控元件鉴定问题的前期调研、分析和总结的基础上，正在致力于发展表观遗传组学数据 ChIP-seq 信号峰的形状分布特征定量化描述体系，并将其应用在 DNA 调控元件的预测中，希望给基因表达调控领域引入全新的、基于实验数据的特征变量，使 DNA 调控元件的鉴定准确性得到进一步提高.

参 考 文 献

[1] Bernstein B E, Birney E, Dunham I, *et al*. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, **489**(7414): 57−74

[2] The ENCODE Project Consortium (ENCyclopedia Of DNA Elements) Project. Science, 2004, **306**(5696): 636−640

[3] Bernstein B E, Meissner A, Lander E S. The mammalian epigenome. Cell, 2007, **128**(4): 669−681

[4] Bejerano G, Pheasant M, Makunin I, *et al*. Ultraconserved elements in the human genome. Science, 2004, **304**(5675): 1321−1325

[5] Maston G A, Evans S K, Green M R. Transcriptional regulatory elements in the human genome. Annu Rev Genom Human Genet, 2006, **7**: 29-59

[6] Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet, 2007, **8**(4): 286-298

[7] Chi P, Allis C D, Wang G G. Covalent histone modifications - miswritten, misinterpreted and mis-erased in human cancers. Nat Rev Cancer, 2010, **10**(7): 457-469

[8] Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol, 2010, **28**(10): 1057-1068

[9] Kouzarides T. Chromatin modifications and their function. Cell, 2007, **128**(4): 693-705

[10] 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**(3): 311-318

[11] He H H, Meyer C A, Shin H, *et al*. Nucleosome dynamics define transcriptional enhancers. Nat Genet, 2010, **42**(4): 343-347

[12] Ernst J, Kheradpour P, Mikkelsen T S, *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 2011, **473**(7345): 43-49

[13] Shen Y, Yue F, McCleary D F, *et al*. A map of the cis-regulatory sequences in the mouse genome. Nature, 2012, **488**(7409): 116-120

[14] Roy S, Ernst J, Kharchenko P V, *et al*. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science, 2010, **330**(6012): 1787-1797

[15] Negre N, Brown C D, Ma L, *et al*. A cis-regulatory map of the *Drosophila* genome. Nature, 2011, **471**(7339): 527-531

[16] Gerstein M B, Lu Z J, Van Nostrand E L, *et al*. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science, 2010, **330**(6012): 1775-1787

[17] Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. Cell, 2007, **129**(4): 823-837

[18] Wang Z, Zang C, Rosenfeld J A, *et al*. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet, 2008, **40**(7): 897-903

[19] Spilianakis C G, Lalioti M D, Town T, *et al*. Interchromosomal associations between alternatively expressed loci. Nature, 2005, **435**(7042): 637-645

[20] Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell, 2011, **144**(3): 327-339

[21] Jin F, Li Y, Ren B, *et al*. Enhancers: multi-dimensional signal integrators. Transcription, 2011, **2**(5): 226-230

[22] Rada-Iglesias A, Bajpai R, Swigut T, *et al*. A unique chromatin signature uncovers early developmental enhancers in humans. Nature, 2011, **470**(7333): 279-283

[23] Phillips J E, Corces V G. CTCF: master weaver of the genome. Cell, 2009, **137**(7): 1194-1211

[24] Mishiro T, Ishihara K, Hino S, *et al*. Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J, 2009, **28**(9): 1234-1245

[25] Kim T H, Abdullaev Z K, Smith A D, *et al*. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell, 2007, **128**(6): 1231-1245

[26] Wingender E, Dietze P, Karas H, *et al*. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res, 1996, **24**(1): 238-241

[27] Portales-Casamar E, Thongjuea S, Kwon A T, *et al*. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res, 2010, **38** (Database issue): D105-110

[28] Knudsen S. Promoter2.0: for the recognition of Pol Ⅱ promoter sequences. Bioinformatics, 1999, **15**(5): 356-361

[29] Reese M G. Application of a time-delay neural network to promoter annotation in the *Drosophila* melanogaster genome. Comput Chem, 2001, **26**(1): 51-56

[30] Klingenhoff A, Frech K, Quandt K, *et al*. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. Bioinformatics, 1999, **15**(3): 180-186

[31] Ohler U. Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. Nucleic Acids Res, 2006, **34**(20): 5943-5950

[32] Zhang M Q. Identification of human gene core promoters in silico. Genome Res, 1998, **8**(3): 319-326

[33] Thomas-Chollier M, Sand O, Turatsinze J V, *et al*. RSAT: regulatory sequence analysis tools. Nucleic Acids Res, 2008, **36**(Web Server issue): W119-127

[34] Thompson W, Rouchka E C, Lawrence C E. Gibbs recursive sampler: finding transcription factor binding sites. Nucleic Acids Res, 2003, **31**(13): 3580-3585

[35] Bailey T L, Boden M, Buske F A, *et al*. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res, 2009, **37**(Web Server issue): W202-208

[36] Zhu J, Liu J S, Lawrence C E. Bayesian adaptive sequence alignment algorithms. Bioinformatics, 1998, **14**(1): 25-39

[37] Schwartz S, Zhang Z, Frazer KA, *et al*. PipMaker——a web server for aligning two genomic DNA sequences. Genome Res, 2000, **10**(4): 577-586

[38] Visel A, Minovitsky S, Dubchak I, *et al*. VISTA Enhancer Browser——a database of tissue-specific human enhancers. Nucleic Acids Res, 2007, **35**(Database issue): D88-92

[39] Visel A, Bristow J, Pennacchio L A. Enhancer identification through comparative genomics. Semin Cell Dev Biol, 2007, **18**(1): 140-152

[40] Lee D, Karchin R, Beer M A. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res, 2011, **21**(12): 2167-2180

[41] Bernstein B E, Stamatoyannopoulos J A, Costello J F, *et al*. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol, 2010, **28**(10): 1045-1048

[42] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat

Biotechnol, 2010, **28**(8): 817−825

[43] Hoffman M M, Buske O J, Wang J, *et al*. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods, 2012, **9**(5): 473−476

[44] Fernandez M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res, 2012, **40**(10): e77

[45] Rajagopal N, Xie W, Li Y, *et al*. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol, 2013, **9**(3): e1002968

[46] Dunham I, Kundaje A, Aldred S F, *et al*. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, **489**(7414): 57−74

[47] Kent W J, Sugnet C W, Furey T S, *et al*. The human genome browser at UCSC. Genome Res, 2002, **12**(6): 996−1006

[48] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods, 2012, **9**(3): 215−216

[49] Wang J, Lunyak V V, Jordan I K. Chromatin signature discovery *via* histone modification profile alignments. Nucleic Acids Res, 2012, **40**(21): 10642−10656

[50] Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol, 2008, **4**(10): e1000201

[51] Ucar D, Hu Q, Tan K. Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. Nucleic Acids Res, 2011, **39**(10): 4063−4075

[52] Firpi H A, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics, 2010, **26**(13): 1579−1586

[53] Won K J, Chepelev I, Ren B, *et al*. Prediction of regulatory elements in mammalian genomes using chromatin signatures. BMC Bioinformatics, 2008, **9**: 547

[54] 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**(2): 148−156

[55] Rodelsperger C, Guo G, Kolanczyk M, *et al*. Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. Nucleic Acids Res, 2011, **39**(7): 2492−2502

[56] Yip K Y, Cheng C, Bhardwaj N, *et al*. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol, 2012, **13**(9): R48

[57] Schones D E, Cui K, Cuddapah S, *et al*. Dynamic regulation of nucleosome positioning in the human genome. Cell, 2008, **132**(5): 887−898

Methods of DNA Elements Identification in Epigenomics*

LU Yi-Ming, QU Wu-Bin, ZHANG Cheng-Gang**

(*Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Beijing* 100850, China)

Abstract In the post-genomic era after human whole-genome sequencing has been completed, accurate functional annotation of genomic sequences, especially the DNA regulatory elements, has become an urgent need of further in-depth understanding of the complex mechanisms of human genome. Recent large-scale chromatin states mapping efforts have revealed characteristic chromatin modification signatures for various types of functional DNA elements. The conclusions drew in these studies have promoted the emergence of a series of supervised and unsupervised methods of DNA elements identification, some of which have been successfully applied to identify functional DNA elements in a number of genomes and have become the regular tools to decode an unknown genome. These methods are adept at different aspects of genomic studies, varied by the embedded algorithms and the identification strategies. In most cases users should consider joint application of different types of methods to obtain a balance between identification sensitivity and specificity. Despite the successful applications of current methods, each type of methods has its own disadvantages which users should scrupulously avoid. In this paper, we not only reviewed the main types of previous and current DNA elements identification methods, and comprehensively analyzed the advantages and disadvantages of each type of methods, but also pointed out the next possible directions of method improvement. We anticipated the analysis and views put forward in this review could help readers to deepen the understanding of principles of DNA elements identification methods, thus better applied them in their own studies.

Key words DNA elements, regulation of gene expression, epigenomics, machine learning

DOI: 10.3724/SP.J.1206.2013.00248

* This work was supported by grants from The National Basic Research Program of China (2012CB518200), The National Natural Science Foundation of China (30900862, 30973107, 81070741, 81172770), The State Key Laboratory of Proteomics of China (SKLP-O201104, SKLP-K201004, SKLP-O201002), and National Science and Technology Major Project of China (2012ZX09102301-016).

**Corresponding author.

Tel: 86-10-66931590, E-mail: zhangcg@bmi.ac.cn

Received: October 11, 2013 Accepted: December 18, 2013