

DNA 甲基化的生物信息学研究进展*

凡时财^{1, 2)} 张学工^{1)**}

¹⁾生物信息学教育部重点实验室, 清华信息国家实验室生物信息学研究部, 清华大学自动化系, 北京 100084;

²⁾电子科技大学自动化工程学院, 成都 610054)

摘要 作为重要的表观遗传学现象之一, DNA 甲基化对基因的表达发挥重要的调控功能. 随着高通量检测技术的不断发展, 对 DNA 甲基化的生物信息学研究也成为 DNA 甲基化研究中的一个非常活跃的热点. 对生物信息学在 DNA 甲基化状态的预测、CpG 岛不易被甲基化的机制研究、探索 DNA 甲基化同其他表观遗传学现象之间的关系以及 DNA 异常甲基化同癌症的发生和发展之间的关系等方面的研究进展进行综述.

关键词 DNA 甲基化, 生物信息学研究, CpG 岛, 预测, 机制, 表观遗传学

学科分类号 Q3, Q6, Q7

DOI: 10.3724/SP.J.1206.2008.00426

在多细胞生物的发育过程中, 细胞如何分化很大程度上取决于其内部的基因表达模式, 而基因的表达模式不仅仅依赖于细胞内转录因子的作用, 还同表观遗传学因素密切相关. 表观遗传学是指在不改变 DNA 序列的条件下所发生的可遗传基因表达的变化^[1]. 表观遗传学主要包括 DNA 甲基化、组蛋白修饰和染色质结构. 脊椎动物中, DNA 甲基化表现为在 DNA 甲基化转移酶 (DNA methyltransferase, DNMT) 作用下, 甲基基团合成到 5'-CpG-3' 中胞嘧啶的第五位碳原子上^[2]. DNA 甲基化参与了细胞的多种生理活动, 比如基因的时空特异性表达、X 染色体失活、衰老以及癌症的发生^[3, 4]. 甲基化的 CpG 双核苷酸通过募集转录抑制因子^[5]或者阻碍转录激活因子的结合抑制基因的表达^[6]. 由于 DNA 甲基化在基因调控中的重要功能, 以 Sanger 研究所为代表的人类表观基因组联盟, 于 2000 年 10 月发起了人类表观基因组计划 (human epigenome project, HEP), 试图获得人类多个组织中的 DNA 甲基化模式^[7]. 目前该计划已经完成了来自 12 个组织的第 6 号、20 号和 22 号染色体的检测工作^[8].

人类基因组中, 70%~80% 的 CpG 双核苷酸处于 DNA 甲基化状态^[9]. 非甲基化的 CpG 不是均匀分布, 而是呈现局部聚集倾向, 形成一些 GC 含量较高、CpG 双核苷酸相对聚集的区域, 即 CpG 岛^[4].

根据 Gardiner-Garden 等^[10]的定义, CpG 岛是一段长度不小于 200 bp、GC 含量不小于 50%、CpG 含量与期望含量之比不小于 0.6 的区域. 由于该定义将一些重复片段也包含其中, Takai 和 Jones^[11]将 CpG 岛重新定义为长度不小于 500 bp、GC 含量不小于 55%、CpG 含量与期望含量之比不小于 0.65 的区域. 据统计, 多于 50% 的基因的启动子区含有 CpG 岛^[12, 13]. 在早期的认识中, CpG 岛都是非甲基化的^[4], 但是随着研究的不断深入, 人们发现在印记基因^[9]、失活的 X 染色体^[14]甚至是正常的体细胞中都存在甲基化的 CpG 岛^[15]. 部分 CpG 岛的异常甲基化常常伴随着癌症等疾病的发生^[16].

近年来, 随着高通量的 DNA 甲基化检测技术的出现, DNA 甲基化的生物信息学研究得到了很大的发展. 一系列具有较好精度的 DNA 甲基化预测工具不仅成为实验检测技术的补充, 也反映了 DNA 甲基化本身是有规律可寻的, 这启发了研究者们对于 DNA 甲基化内在机制的探索. DNA 甲基化具有序列偏好性的观点得到进一步证实, CpG

* 国家自然科学基金资助项目 (30625012, 60721003), 国家重点基础研究发展计划项目 (973) (2004CB518605) 和国家高技术研究发展计划项目 (863) (2006AA02Z325).

** 通讯联系人.

Tel: 010-62794919, E-mail: zhangxg@tsinghua.edu.cn

收稿日期: 2008-06-13, 接受日期: 2008-08-18

岛边界序列通过结合特定转录因子使 CpG 岛不易被甲基化的机制假说也有了更多的证据。同时, 研究疾病特别是癌症中 DNA 甲基化的特点和规律、发现与特定癌症相关的甲基化生物标记(biomarker), 成为 DNA 甲基化研究中新的热点问题。

本文综述了目前关于人的 DNA 甲基化研究中生物信息学研究的最新进展。首先介绍了 DNA 甲基化的高通量检测方法和相关数据库, 然后重点综述了 DNA 甲基化的预测研究和 CpG 岛不易被甲基化的机制探索, 紧接着简要阐述了 DNA 甲基化同癌症以及其他表观遗传学现象之间的关系, 最后一部分是作者基于 DNA 甲基化的工作体会对于今后研究方向的展望。

1 DNA 甲基化的高通量实验检测方法

DNA 甲基化的检测大体分为两个步骤: a. 待检测样品的前期处理; b. 目标序列的定位和甲基化状态的量化。待检测样品的前期处理方法主要包

括三大类: a. 亚硫酸氢钠(sodium bisulfite)法, 亚硫酸氢钠把非甲基化的胞嘧啶转化为尿嘧啶, 而保持甲基化的胞嘧啶不变, 以此实现两类的分离, 此方法因其准确率高而被誉为检测 DNA 甲基化的“金方法”; b. 限制性内切酶(restriction enzyme)法, 限制性内切酶可以切割非甲基化的位点, 而甲基化以后的识别位点将被保留; c. 利用特定抗体对甲基化的胞嘧啶进行免疫沉淀反应(immunoprecipitation)。这三种方法的主要特点见表 1。前期处理实现了甲基化和非甲基化的胞嘧啶分离以后, 可以对分离出的甲基化 DNA 片段进行基因芯片杂交或者大规模深度测序, 实现高通量的检测(图 1)。以上两个步骤的不同组合产生了多种高通量的实验检测技术, 如亚硫酸氢钠+测序^[17, 18]、亚硫酸氢钠+芯片技术^[19, 20]、酶切+芯片技术^[21]、甲基化抗体+芯片技术^[22]等。目前很多实验室利用以上技术对不同组织中的 DNA 甲基化状态进行了检测, 具体情况见表 2。

Table 1 The characteristics of different pretreatment methods in detection techniques for DNA methylation

表 1 DNA 甲基化实验检测技术各种前期处理方法的特点

处理方法	分辨率	优点	缺点
亚硫酸氢钠	高(单个碱基)	准确, 可应用于任何待检测样品	需要将所有非甲基化的胞嘧啶进行转换
酶切处理	中等	方法相对简单	只能针对内切酶能识别的位点
免疫沉淀	中等	任意序列背景下的甲基化胞嘧啶都能实现免疫沉淀	对发生甲基化的胞嘧啶的密度要求较高

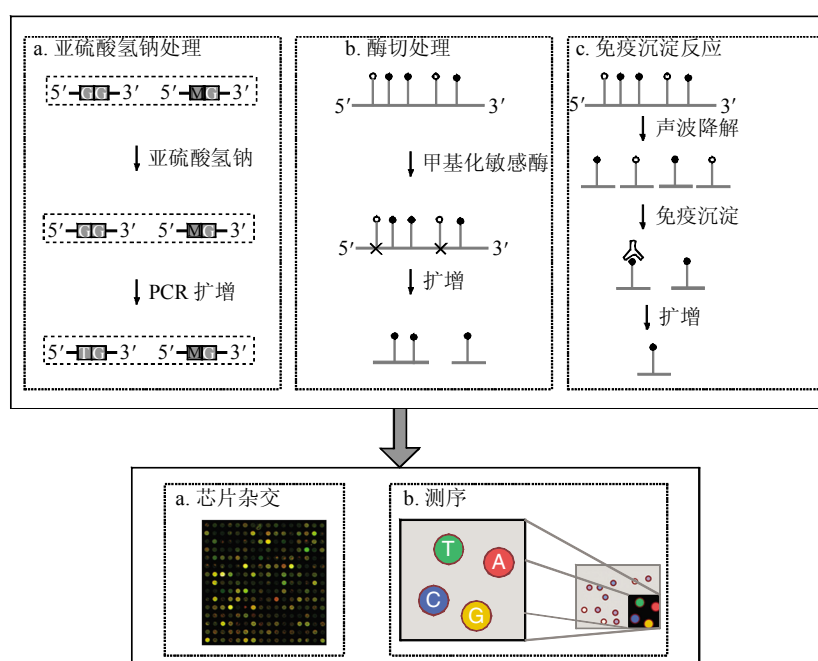


Fig. 1 Sketch of the high-throughput detection techniques for DNA methylation

图 1 高通量检测 DNA 甲基化技术的示意图

Table 2 The published results of human DNA methylation with high-throughput detection techniques
表 2 已发表的大规模检测人的 DNA 甲基化状态的研究结果

检测方法	检测区域	样本	参考文献
亚硫酸氢钠+测序	第 6, 20, 22 号染色体	来自 43 个样本的 12 个组织	[8]
酶切+测序	第 21q 号染色体	外周血细胞	[15]
酶切+芯片技术	全基因组	脑组织	[21]
	第 21 和 22 号染色体	脑组织	[23]
甲基化抗体+芯片技术	全基因组	原代肺纤维组织, 结肠癌细胞系	[22]
	全基因组	原代纤维细胞, 精原细胞	[24]
亚硫酸氢钠+芯片技术	371 个基因	乳腺癌细胞系, 结肠癌细胞系, 前列腺癌细胞系	[19]

2 DNA 甲基化的相关数据库

数据库的构建是生物信息学研究的重要内容。目前发表的较大规模的 DNA 甲基化相关数据库有 4 个: MethDB、MethyCancer、PubMeth 和 MethCancer DB, 数据情况见表 3。其中, MethDB 是最早整合文献中 DNA 甲基化数据的数据库, 也是涵盖物种和组织最多的数据库, 其余 3 个数据库都采用了文献挖掘的方法, 收集了文献中已报道的与特定癌症

相关的一些 DNA 甲基化模式。MethDB 和 MethyCancer 还提供了 DNA 甲基化的可视化工具。通过对现有的分散的 DNA 甲基化数据的整理分析, 这些数据库提供了内容独立、记录相对完整、格式比较统一的数据资源, 为后期的深入挖掘提供了极大便利。随着大规模深度测序技术在 DNA 甲基化检测中的应用, 建立相应数据库用以存储和分析相关数据, 开发类似于 CpGViewer^[25]的可视化分析工具将具有重要价值。

Table 3 Databases related to DNA methylation
表 3 DNA 甲基化相关的数据库

数据库	数据情况	提供的服务	参考文献及链接
MethDB	50 个物种, 241 个组织, 共 20 236 条甲基化数据	在线查询	[26]
MethyCancer	含 64 681 个 MethyLoci, 与癌症相关的 7 100 个基因和 57 790 个 SNP	可视化工具	http://www.methdb.de/
		在线查询	[27]
PubMeth	来自 1 000 篇文献, 包括 5 000 条记录	下载	http://methcancer.genomics.org.cn/
		可视化工具	
MethCancer DB	4 720 个记录, 2 199 个基因	在线查询	[28]
		下载	http://www.pubmeth.org/
			[29]
			http://www.methcancerdb.net/methcancerdb/home.seam

3 DNA 甲基化的预测研究

实验手段检测 DNA 甲基化状态的方法虽然比较可靠, 但是其对于人力财力的需求以及检测技术方面的缺陷, 使得目前大规模的 DNA 甲基化的实验数据还比较有限。研究计算预测 DNA 序列甲基化状态的方法显得极为重要。另一方面, 作为实验检测技术的补充, 预测算法能够挖掘出数据中隐藏的重要特征, 为人们进一步认识 DNA 甲基化机理提供重要依据和研究思路。目前发表的对 DNA 甲基化模式进行预测的方法中, 按照预测目标的类型可以分为, 对单个 CpG 双核苷酸的甲基化状态预

测、对 CpG 岛片段以及 CpG 岛的甲基化状态预测。Methylator (<http://bio.dfci.harvard.edu/Methylator/index.html>)是目前唯一的预测单个 CpG 双核苷酸甲基化状态的工具^[30]。该工具基于 MethDB 数据库^[26, 31]中人的 2 839 个 DNA 甲基化模式数据, 以 CpG 位点周围 39 bp 的序列模式为特征, 采用支持向量机的方法得到了 87% 的正确率。该研究还表明, 基因的非翻译区的甲基化程度要高于外显子和内含子。由于单个 CpG 双核苷酸的甲基化状态在细胞生长过程中是动态变化的^[4], 因此 Methylator 的应用相对有限。

2006 年, Rollins 等^[21]发表了人脑组织全基因

组范围的甲基化数据. 基于其得到的 1 948 个甲基化区域和 2 386 个非甲基化区域, 产生了两个预测 CpG 岛片段甲基化状态的工具——HDMFinder^[32]和 MethCGI^[33]. HDMFinder (<http://rulai.cshl.edu/HDMFinder/methylation.htm>) 利用包括 *Alu* 在内的 17 个序列特征能够对长度为 800 bp 的 CpG 岛片段以及 800 bp 的普通 DNA 序列的甲基化状况进行准确预测, 并且还对 22 个常染色体中的甲基化模式进行了预测. 我们开发的 MethCGI 则专门针对 CpG 岛片段 (包括 200 bp、300 bp 以及 400 bp) 进行了预测, 正确率达到了 84%, 提供了在线预测服务 (<http://bioinfo.au.tsinghua.edu.cn/MethCGI/MethCGI.html>). 通过分析特定转录因子结合位点在甲基化和非甲基化的 CpG 岛片段中的分布情况, 我们得到了 4 个在两类中差异最大的转录因子结合位点, 这些结合位点可能为下一步研究 DNA 甲基化的内在机理提供重要依据. 另外, MethCGI 对 MethDB 数据库中包含的来自多个组织的甲基化数据也得到了较好的预测精度, 表明 CpG 岛片段的甲基化模式并没有很强的组织特异性.

CpG 岛的甲基化状态同基因表达的模式密切相关, 因此对于 CpG 岛甲基化状态的预测一直是人们关注的重点. 2003 年, Feltus 等^[34]通过在细胞系中加入过量 DNA 甲基化转移酶(DNMT1), 发现大部分的 CpG 岛仍然保持非甲基化状态, 同时有部分 CpG 岛呈现超甲基化, 采用序列特征对两类的分类正确率可以达到 82%. 在特征选择的过程中, 他们发现两类 CpG 岛在长度、GC 含量、CpG 比例以及和启动子的关系上没有显著性差异. 但是其后对来自体内的 DNA 甲基化数据的研究表明, 甲基化和非甲基化的 CpG 岛在长度、GC 含量以及 CpG 比例上存在显著差异^[33, 35]. Bock 等基于人外周血白细胞中第 22 号染色体得到的 132 个 CpG 岛的甲基化状态数据^[20], 先后得到了两个预测 CpG 岛甲基化状态的分类器. 第一个采用了序列模式, 特定的重复序列以及特殊的 DNA 结构三大类特征^[35], 第二个增加了预测的组蛋白修饰状态和染色质结构特征^[36], 并提供了在线预测工具 Epigraph (<http://epigraph.mpi-inf.mpg.de/WebGRAPH/>). 由于组蛋白修饰以及染色质结构分布是否存在 DNA 序列的偏好性仍是一个有争议的问题, 我们通过增加 Barski 等^[37]实验获得的基因组尺度的组蛋白甲基化数据, 对 CpG 岛的甲基化状态进行了预测^[38]. 该预测模型得到了更高分类正确率的同时, 还发现了

4 种显著影响 CpG 岛甲基化的组蛋白修饰特征 (H3K4me1、H3K4me2、H3K4me3 以及 H3K9me1), 一定程度上量化了部分组蛋白甲基化同 CpG 岛甲基化之间的关系.

4 CpG 岛不易被甲基化的机制探索

虽然人类基因组上有的 CpG 岛处于甲基化状态, 但是大部分 CpG 岛是不易被甲基化的, 这同基因组上约 80% 的 CpG 双核苷酸处于甲基化状态的现象形成鲜明对比. 该现象促使人们思考为什么 CpG 岛不易被甲基化. 早在 1994 年 Brandeis 等^[39]就发现, Sp1 的结合位点在人胚胎干细胞的 *APRT* 基因中能够保护 CpG 岛在重新甲基化的过程中不被甲基化, 后来对人乳腺组织中 *BRCA1* 基因的研究结果同样表明, 启动子区 Sp1 和 CTCF 结合位点突变与否将决定该区域 CpG 岛的甲基化状态^[40]. Turker 等^[41, 42]以 Sp1 在小鼠 *Aprt* 基因中的功能对小鼠体细胞中 DNA 甲基化模式的形成提出了一个动态平衡假说, 认为在 DNA 的重新甲基化过程中, 首先是 B1 之类的重复序列被甲基化, 这些序列被称为甲基化中心, 然后甲基化会从中心往外延伸, Sp1 结合位点通过同 Sp1 的结合能够阻碍甲基化向 CpG 岛区的蔓延, 从而保护 CpG 岛不被甲基化. 后来的研究又发现, 在 Sp1 缺失的情况下 *Aprt* 基因仍然有表达, 即 CpG 岛并没有因 Sp1 的缺失而超甲基化^[43], 另一方面, Sp1 (以及 CTCF) 的结合位点并非存在于所有不易甲基化的 CpG 岛附近, 这也意味着还应该有其他因子参与保护机制的形成.

利用人脑组织基因组尺度的 CpG 岛甲基化数据^[21], 我们系统分析了不易甲基化的 CpG 岛内部以及岛边界序列的序列特征, 发现一些具有锌指结构的转录因子 (包括 Sp1 和 CTCF) 可能的结合位点在 CpG 岛的 400 bp 边界序列中显著富集, 而且其中 80% 的结合位点在人、大鼠和小鼠中是显著保守的^[44]. 基于以上结果, 我们认为, 从小鼠 *Aprt* 基因得到的关于 CpG 岛不易被甲基化的保护机制同样存在于人的基因组中^[44]: *Alu* 序列是重新甲基化过程中的甲基化中心^[45], 在甲基化转移酶的作用下, DNA 甲基化从 *Alu* 序列出发往外延伸, 当 CpG 岛周围富集的转录因子结合位点同具有锌指结构的转录因子结合以后则可以阻止甲基化向 CpG 岛的蔓延, 因此在这种阻止与蔓延的作用达到动态平衡以后, 就形成了比较稳定的甲基化模式. 当细胞所处的环境发生变化时, 这种平衡可能

会被打破, 阻止功能的增强或者侵入能力的提高则可能使 DNA 甲基化往 CpG 岛外或者往内移动, 从而建立新的平衡. 比如, 如果某些顺式元件发生突变, 则本可以结合并阻碍 DNA 甲基化蔓延的特定转录因子可能因为结合能力的减弱使其阻碍功能变

弱甚至消失, 这种情况下, 其附近的 CpG 岛可能会被甲基化^[39, 43]. 因此, 发生癌变的组织中抑癌基因的异常甲基化以及与老化过程相关的甲基化可能是由于这种保护机制的减弱造成的^[45]. 上述推测的保护机制示意图见图 2.

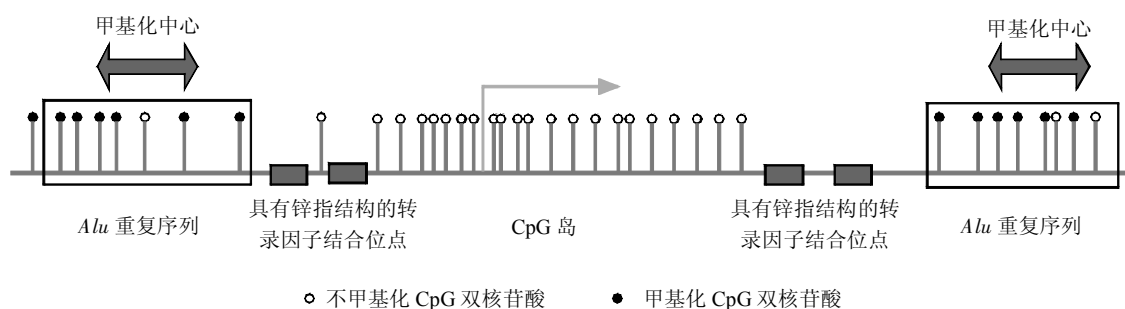


Fig. 2 Sketch of the hypothetical mechanism why CpG islands are resistant to DNA methylation

图 2 CpG 岛不易被甲基化的机制假说示意图

5 DNA 甲基化与癌症的关系

DNA 甲基化同众多疾病的发生与发展密切相关, 比如 II 型糖尿病、自身免疫疾病以及各种癌症. 特别是癌症, 其发生过程中癌细胞基因组整体的欠甲基化和局部区域的超甲基化是其典型特征^[46]. 超甲基化体现在抑癌基因的启动子区被异常甲基化, 整体的欠甲基化同重复序列(如转座子)以及癌基因的启动子区的甲基化程度减小、基因组遗传不稳定性的增加密切相关. 研究特定癌症中超甲基化基因的工作较多, 包括乳腺癌^[47, 48]、结肠癌^[22, 49]、髓细胞性白血病^[50]、肝癌^[51]、涎腺腺样囊性癌^[52]和卵巢癌^[53]等. 其中, Keshet 等^[49]基于结肠癌的数据发现, 结肠癌中超甲基化的基因不仅属于一个特定的功能类型并在染色体上聚集, 而且这些基因中还存在着共同的 motif, 他们推测这类基因可能受到相同转录因子的调节. 目前, 对于癌细胞和正常细胞中 DNA 甲基化模式差异性的认识还存在不同观点. Weber 等^[22]对原代纤维细胞、正常结肠黏膜和结肠癌细胞系中 6 000 个 CpG 岛甲基化模式进行了比较, 认为肿瘤细胞中仅有小部分 CpG 岛的启动子区存在异常甲基化. 但是 Gebhard 等^[50]在髓细胞性白血病中却发现了大量异常超甲基化的基因, 并认为癌细胞中的甲基化模式显著不同于正常细胞. 以上两种结论的差异是来源于细胞系之间或是所采用的技术平台之间的差异, 还有待于对更多癌细胞中的 DNA 甲基化模式进行分析研究. 此外,

Bibikova 等^[19]对结肠癌、乳腺癌、肺癌和前列腺癌细胞系以及正常细胞中 371 个基因中的 1 536 个位点的甲基化状态进行聚类分析, 发现了 55 个可以较好地地区分癌细胞和正常细胞以及区分各种癌细胞类型的位点. 该方法表明, 各种癌症都存在其特异的 DNA 甲基化生物标记.

6 DNA 甲基化与其他表观遗传学现象之间的关系

细胞发育过程中, 各种表观遗传学现象之间不是孤立存在而是密切联系的. DNA 甲基化同组蛋白甲基化共同调控基因表达的现象最早在链孢霉 (*Neurospora crassa*) 中得到证实^[54], 进一步的生化研究结果认为, DNA 甲基化是受组蛋白甲基化调节的^[55]. 而在哺乳动物中, 有研究认为, DNA 甲基化是建立和维持其他表观遗传学现象的基础^[56], 比如 DNA 甲基化位点可以募集诸如组蛋白去乙酰化酶等具有抑制功能的复合物, 同时去除掉该位点附近的组蛋白乙酰化标记^[4]. 也有研究认为, DNA 甲基化是受组蛋白修饰调控的, 比如有报道称组蛋白修饰 H3K9me 能够促进 DNA 甲基化的进程^[57]. 目前, 虽然哺乳动物中存在基因组尺度的 DNA 甲基化和组蛋白修饰数据, 但是还没有研究者针对同一个组织进行两种修饰的系统研究. 2008 年初, 北京生命科学研究所采用覆瓦芯片 (Tiling Array) 技术对水稻第 4 和 10 号染色体中的 DNA 甲基化和组蛋白甲基化进行了系统研究, 发现了 DNA 甲基化

和组蛋白修饰在基因组不同位置上存在的各种相互关系, 还进一步分析了它们同染色体结构以及基因表达之间的相关性^[58]. 如果将类似的高通量技术同时用以检测哺乳动物中 DNA 甲基化和其他表观遗传因素, 则必将加快我们对于 DNA 甲基化和其他表观遗传因素如何共同调节基因表达的理解和认识.

7 结论和展望

作为直接作用于 DNA 序列的表观遗传修饰, DNA 甲基化在基因表达调控中扮演着重要角色. 虽然 DNA 甲基化是各种表观遗传现象中最早被人们认识也是研究相对较成熟的现象之一, 但是目前对于 DNA 甲基化的机理机制的探索还仍然处于比较初级的阶段. 随着各种高通量检测技术的发展和运用, 各种正常组织以及癌症组织的 DNA 甲基化数据将不断涌现, 这就给从事生物信息研究的科学工作者提供了大量研究数据、提出了大量新的研究课题. 分析利用这些海量的数据无疑需要开发更多更有效的计算方法和工具. 利用来自不同样本不同组织的 DNA 甲基化数据, 可以深层次探讨 DNA 甲基化在人群中的差异性以及组织特异的 DNA 甲基化模式. 虽然保护机制的假说能够一定程度上解释 CpG 岛不易被甲基化的机制, 但是该假说是否全面揭示了其真实机理, 是否还有其他更多甚至更重要的因素参与其中, 这些问题都有待进一步验证. 另外, 结合其他表观遗传现象对人类基因组进行系统研究也是认识基因表达的必然之路. 在医学应用上, 寻找和特定疾病相关的 DNA 甲基化标识, 利用这些标识对疾病进行检测, 并研制使异常甲基化还原的药物也是极具发展前景的方向.

参 考 文 献

- Goldberg A D, Allis C D, Bernstein E. Epigenetics: a landscape takes shape. *Cell*, 2007, **128**(4): 635~638
- Bird A P. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet*, 1987, **3**: 342~347
- Reik W, Santos F, Dean W. Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology*, 2003, **59**(1): 21~32
- Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev*, 2002, **16**(1): 6~21
- Bird A P, Wolffe A P. Methylation-induced repression—belts, braces, and chromatin. *Cell*, 1999, **99**(5): 451~454
- Kim J, Kollhoff A, Bergmann A, *et al.* Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. *Hum Mol Genet*, 2003, **12**(3): 233~245
- Bradbury J. Human epigenome project—up and running. *PLoS Biol*, 2003, **1**(3): E82
- Eckhardt F, Lewin J, Cortese R, *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 2006, **38**(12): 1378~1385
- Ehrlich M, Gama-Sosa M A, Huang L H, *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res*, 1982, **10**(8): 2709~2721
- Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*, 1987, **196**(2): 261~282
- Takai D, Jones P A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci USA*, 2002, **99**(6): 3740~3745
- Wang Y, Leung F C. An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, 2004, **20**(7): 1170~1177
- Larsen F, Gundersen G, Lopez R, *et al.* CpG islands as gene markers in the human genome. *Genomics*, 1992, **13**(4): 1095~1107
- Hansen R S, Stoger R, Wijmenga C, *et al.* Escape from gene silencing in ICF syndrome: evidence for advanced replication time as a major determinant. *Hum Mol Genet*, 2000, **9**(18): 2575~2587
- Yamada Y, Watanabe H, Miura F, *et al.* A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Res*, 2004, **14**(2): 247~266
- Hanahan D, Weinberg R A. The hallmarks of cancer. *Cell*, 2000, **100**(1): 57~70
- Rakyan V K, Hildmann T, Novik K L, *et al.* DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol*, 2004, **2**(12): e405
- Cokus S J, Feng S, Zhang X, *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 2008, **452**(7184): 215~219
- Bibikova M, Lin Z, Zhou L, *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*, 2006, **16**(3): 383~393
- Zhang D, Wang Y, Bai Y, *et al.* A novel method to quantify local CpG methylation density by regional methylation elongation assay on microarray. *BMC Genomics*, 2008, **9**: 59
- Rollins R A, Haghighi F, Edwards J R, *et al.* Large-scale structure of genomic methylation patterns. *Genome Res*, 2006, **16**(2): 157~163
- Weber M, Davies J J, Wittig D, *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*, 2005, **37**(8): 853~862
- Schumacher A, Kapranov P, Kaminsky Z, *et al.* Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res*, 2006, **34**(2): 528~542
- Weber M, Hellmann I, Stadler M B, *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, 2007, **39**(4): 457~466

- 25 Carr I M, Valleley E M, Cordery S F, *et al.* Sequence analysis and editing for bisulphite genomic sequencing projects. *Nucleic Acids Res*, 2007, **35**(10): e79
- 26 Grunau C, Renault E, Rosenthal A, *et al.* MethDB—a public database for DNA methylation data. *Nucleic Acids Res*, 2001, **29**(1): 270~274
- 27 He X, Chang S, Zhang J, *et al.* MethCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res*, 2008, **36** (Database issue): D836~841
- 28 Ongenaert M, Van Neste L, De Meyer T, *et al.* PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res*, 2008, **36**(Database issue): D842~846
- 29 Lauss M, Visne I, Weinhaeusel A, *et al.* MethCancerDB——aberrant DNA methylation in human cancer. *Br J Cancer*, 2008, **98** (4): 816~817
- 30 Bhasin M, Zhang H, Reinherz E L, *et al.* Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett*, 2005, **579**(20): 4302~4308
- 31 Amoreira C, Hindermann W, Grunau C. An improved version of the DNA methylation database (MethDB). *Nucleic Acids Res*, 2003, **31** (1): 75~77
- 32 Das R, Dimitrova N, Xuan Z, *et al.* Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci USA*, 2006, **103**(28): 10713~10716
- 33 Fang F, Fan S, Zhang X, *et al.* Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, 2006, **22**(18): 2204~2209
- 34 Feltus F A, Lee E K, Costello J F, *et al.* Predicting aberrant CpG island methylation. *Proc Natl Acad Sci USA*, 2003, **100** (21): 12253~12258
- 35 Bock C, Paulsen M, Tierling S, *et al.* CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet*, 2006, **2**(3): e26
- 36 Bock C, Walter J, Paulsen M, *et al.* CpG island mapping by epigenome prediction. *PLoS Comput Biol*, 2007, **3**(6): e110
- 37 Barski A, Cuddapah S, Cui K, *et al.* High-resolution profiling of histone methylations in the human genome. *Cell*, 2007, **129** (4): 823~837
- 38 Fan S, Zhang M Q, Zhang X. Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem Biophys Res Commun*, 2008, **374**(3): 559~564
- 39 Brandeis M, Frank D, Keshet I, *et al.* Sp1 elements protect a CpG island from de novo methylation. *Nature*, 1994, **371**(6496): 435~438
- 40 Butcher DT, Mancini-DiNardo DN, Archer TK, *et al.* DNA binding sites for putative methylation boundaries in the unmethylated region of the BRCA1 promoter. *Int J Cancer*, 2004, **111**(5): 669~678
- 41 Turker M S. The establishment and maintenance of DNA methylation patterns in mouse somatic cells. *Semin Cancer Biol*, 1999, **9**(5): 329~337
- 42 Turker M S. Gene silencing in mammalian cells and the spread of DNA methylation. *Oncogene*, 2002, **21**(35): 5388~5393
- 43 Mummaneni P, Yates P, Simpson J, *et al.* The primary function of a redundant Sp1 binding site in the mouse aprt gene promoter is to block epigenetic gene inactivation. *Nucleic Acids Res*, 1998, **26**(22): 5163~5169
- 44 Fan S, Fang F, Zhang X, *et al.* Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome. *PLoS ONE*, 2007, **2**(11): e1184
- 45 Issa J P. Aging, DNA methylation and cancer. *Crit Rev Oncol Hematol*, 1999, **32**(1): 31~43
- 46 Jones P A, Baylin S B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 2002, **3**(6): 415~428
- 47 Ordway J M, Budiman M A, Korshunova Y, *et al.* Identification of novel high-frequency DNA methylation changes in breast cancer. *PLoS ONE*, 2007, **2**(12): e1314
- 48 Piotrowski A, Benetkiewicz M, Menzel U, *et al.* Microarray-based survey of CpG islands identifies concurrent hyper- and hypomethylation patterns in tissues derived from patients with breast cancer. *Genes Chromosomes Cancer*, 2006, **45**(7): 656~667
- 49 Keshet I, Schlesinger Y, Farkash S, *et al.* Evidence for an instructive mechanism of *de novo* methylation in cancer cells. *Nat Genet*, 2006, **38**(2): 149~153
- 50 Gebhard C, Schwarzfischer L, Pham T H, *et al.* Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer Res*, 2006, **66**(12): 6118~6128
- 51 Yu J, Zhang H Y, Ma Z Z, *et al.* Methylation profiling of twenty four genes and the concordant methylation behaviours of nineteen genes that may contribute to hepatocellular carcinogenesis. *Cell Res*, 2003, **13**(5): 319~333
- 52 Zhang C Y, Mao L, Li L, *et al.* Promoter methylation as a common mechanism for inactivating E-cadherin in human salivary gland adenoid cystic carcinoma. *Cancer*, 2007, **110**(1): 87~95
- 53 Yang H J, Liu V W, Wang Y, *et al.* Differential DNA methylation profiles in gynecological cancers and correlation with clinico-pathological data. *BMC Cancer*, 2006, **6**: 212
- 54 Tamaru H, Selker E U. A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. *Nature*, 2001, **414**(6861): 277~283
- 55 Jackson J P, Lindroth A M, Cao X, *et al.* Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*, 2002, **416**(6880): 556~560
- 56 Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, 2003, **33**(Suppl): 245~254
- 57 Lehnertz B, Ueda Y, Derijck A A, *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol*, 2003, **13** (14): 1192~1200
- 58 Li X, Wang X, He K, *et al.* High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell*, 2008, **20**(2): 259~276

Progress of Bioinformatics Study in DNA Methylation*

FAN Shi-Cai^{1,2)}, ZHANG Xue-Gong¹⁾**

⁽¹⁾MOE Key laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China;

⁽²⁾School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

Abstract DNA methylation is an important epigenetic system. It plays many crucial roles in the gene regulation. With the development of the high-throughput detection techniques, the bioinformatics study has been an active hot topic in the research of DNA methylation. The major achievements and progress on the prediction of DNA methylation status, the mechanism that the majority of CpG islands are resistant to DNA methylation, the relationship between DNA methylation and other epigenetics, as well as the association between aberrant DNA methylation and the tumorigenesis were reviewed in this article.

Key words DNA methylation, bioinformatics study, CpG island, prediction, mechanism, epigenetics

DOI: 10.3724/SP.J.1206.2008.00426

*This work was supported by grants from The National Natural Science Foundation of China (30625012, 60721003), The National Basic Research Program of China (2004CB518605) and The Hi-tech Research and Development Program of China (2006AA02Z325).

**Corresponding author.

Tel: 86-10-62794919, E-mail: zhangxg@tsinghua.edu.cn

Received: June 13, 2008 Accepted: August 18, 2008