生物物理学报　第二十二卷　第五期　二〇〇六年十月
ACTA BIOPHYSICA SINICA　Vol.22 No.5　Oct. 2006

人类全基因组范围的 CpG 岛的预测与分析

庄海滨[1]，　朱景德[2]，　刘湘军[1]
（1. 清华大学医学院，生物信息学教育部重点实验室，清华大学生物科学与技术系，北京 100084；
2. 上海交通大学肿瘤研究所，癌基因及相关基因国家重点实验室，上海 200032）

摘要：CpG 岛的甲基化是表观遗传中基因表达调控的重要机制。虽然目前已存在几个从 DNA 序列判别 CpG 岛的标准，但如何在标准中选择合适的参数仍是研究的焦点。文章通过分析比较两种经典 CpG 岛判定标准与三种预测方法，提出了改进的 CpG 岛预测方法——CpGI Seeker。应用该预测方法，结合判定标准中的三个基本参数组合出的 13 组组合参数，在人类全基因组范围内进行了 CpG 岛预测，并统计分析了 CpG 岛的重复序列组成以及相对于基因转录起始位点的位置分布情况。分析结果表明 CpGI Seeker 具有更精确判定 CpG 岛的特性；同时还提示，随着判定标准严格性的增加，CpG 岛的重复序列含量降低，与基因转录起始位点的相关性提高。将 CpG 岛最小尺寸为 500 bp、GC 含量为 60%、CpG 出现率达到 0.65 的组合参数作为标准，是目前预测 CpG 岛的最佳方式。

关键词：CpG 岛；甲基化；表观遗传学
中图分类号：Q75

0 引　　言

CpG 二核苷酸的胞嘧啶 5- 甲基修饰是哺乳动物 DNA 上几近唯一的一种共价修饰。人类和小鼠分别有 55.9%和 46.9%的基因与 CpG 岛有着密切的关联[1]。大量的实验证明启动子区的 CpG 岛在基因的沉默[2]、X 染色体的失活[3]、基因印记[4]等中起着极其重要的作用。特别是在肿瘤分子诊断中，启动子 CpG 岛局部去甲基化状态是原癌基因的转录高度活化状态的必要前提之一，这使得 CpG 甲基化的研究在肿瘤的早期诊断中处于举足轻重的地位。20 多年来，CpG 岛的识别研究也在不断地进行中，并已经成为表观遗传学中研究甲基化的一个重点。

CpG 岛的标准最早是由 Gardiner-Garden 和 Frommer[5]于 1987 年提出的。他们认为 CpG 岛是 GC 含量达到 50%，CpG 二核苷酸的出现率（观测值与期望值的比率）达到 0.6 且长度至少为 200 bp 的 DNA 序列。后来 Takai 等[6]又对 CpG 岛的标准做了部分的修正，认为 GC 含量达到 55%、CpG 二核苷酸的出现率达到 0.65 且长度至少为 500 bp 的 DNA 序列更趋近于分布在基因的 5' 端区域。许多预测 CpG 岛的方法以及软件也相继而出，诸如 CpGi130[7]、CpGProD[8]、CpGIE[9]等。这些软件都有各自不同的检索策略，预测的 CpG 岛也不尽相同。在诸多的判定参数以及不同的预测方法中，哪种参数组合及方法易于更为准确地发现 CpG 岛与真实存在的甲基化，是困惑研究者们的主要问题。恰当的 CpG 岛判定标准、合适的预测方法研究以及全基因组 CpG 岛的预测与分析，是目前表观遗传学迫在眉睫的工作。

本文分析已有的两种经典的 CpG 岛判定标准，整理出 CpG 岛的 3 个基本参数：CpG 岛尺寸、GC 含量和 CpG 出现率。通过分析比较已有的 3 种 CpG 岛预测方法，结合各个方法的优势开发出一种新算法 CpGI Seeker。CpGI Seeker 显示了比其它三种方法更加精确的 CpG 岛预测特性。接着将 3 个基本参数组合成 13 组组合参数，并运用 CpGI Seeker，在人类全基因组范围内进行了 CpG 岛的预测。针对 13 组不同的预测结果，本文还对 CpG 岛的重复序列以及相对于基因转录起始位点的位置

收稿日期：2006-05-16
基金项目：国家自然科学基金资助项目（90412018, 30570850），973 基础研究项目（2004CB518804），教育部科学技术研究重大项目（03180，104232）、教育部“跨世纪人才培养计划”基金项目，上海市自然科学基金项目（04DZ14006，05DZ19318）
通讯作者：刘湘军，电话：(010)62792997，E-mail：frankliu@mail.tsinghua.edu.cn；朱景德，电话：(021)64224285，Email：jdzhu@sjtu.edu.cn

分布进行了统计学的分析。研究结果显示，严格的判定标准有利于预测 CpG 岛，也有利于寻找与转录调控密切相关的甲基化区域。同时将 CpG 岛最小尺寸为 500bp、GC 含量为 60%、CpG 出现率达到 0.65 的组合参数作为标准，是目前预测 CpG 岛的最佳方式。

1 材料和方法

1.1 材料

人类基因组数据版本为 Hg17（2004 年 5 月），来自 UCSC 的 ftp 站点（ftp:// hgdownload.cse.ucsc.edu/goldenPath/hg17/chromosomes/）。人类基因转录起始位点的数据来自 UCSC 基因组生物信息学网站（http://genome.ucsc.edu），共 20646 条 mRNA。重复序列的分析工具 RepeatMasker 3.0 来自 http://www.girinst.org/，重复序列库来自 RepBase（版本：2005 年 5 月 23 日）。

1.2 CpG 岛预测算法——CpGI Seeker

判定 CpG 岛的准则包含 3 个基本参数：CpG 岛最小尺寸、GC 最低含量和 CpG 最低出现率。其中 CpG 出现率是 CpG 二核苷酸的观测个数与期望个数的比率，并且期望个数可以用[C]×[G]/[ACGT]来计算[8]，其中[]表示个数。在固定 3 个基本参数的条件下，可以运用 CpGI Seeker 预测算法进行 CpG 岛的预测，具体步骤如下：

1) 定义搜索窗大小为 CpG 岛最小尺寸。

2) 以 1 nt 为步距，在 DNA 序列上寻找 GC 含量和 CpG 出现率都满足最低要求的搜索窗；最低要求定义为满足 3 个基本参数以及 CpG 期望个数的最低值条件[9]，即 CpG 期望个数（CpG_{exp}）不小于序列长度除以 16。该条件有利于去除数学上的 CpG 岛，即碱基 G 的含量大大超过碱基 C，或者碱基 C 的含量大大超过碱基 G[6]。

3) 找到满足条件的第一个搜索窗后，不断以 1 nt 的跳距移动搜索窗，考察下一个搜索窗直至不满足最低要求。

4) 将满足条件的第一个搜索窗的第一个碱基到最后一个搜索窗的最后一个碱基之间的 DNA 序列作为一个新的大窗口，并考察该大窗口是否满足最低要求，若不满足，则大窗口的 5'、3' 端各减少 1 nt 直至满足条件。

5) 至此，独立的 CpG 岛搜索结束。继续执行 2)～4)，直至找到所有独立的 CpG 岛。

6) 若相邻两个 CpG 岛的距离不超过最小连接距离（200 nt），则尝试连接相邻 CpG 岛：考察从前一个 CpG 岛的第一个碱基到后一个 CpG 岛的最后一个碱基之间的 DNA 序列是否满足 GC 含量和 CpG 出现率的最低要求；若新序列满足条件，则将新序列定义为新的 CpG 岛，进入步骤 7)进行调整。

7) 去除新序列 5' 端和 3' 端所有非 C 或 G 的碱基。

8) 若去除碱基后的序列 GC 最低含量满足条件但 CpG 最低出现率不满足条件，则其 5' 端和 3' 端按照“去除 GC 含量低的一端”的原则依次去除碱基，直到满足基本条件或碰到 CpG 二核苷酸。

9) 若去除碱基后的序列 CpG 最低出现率满足条件但 GC 最低含量不满足条件，则延长其 5' 端和 3' 端直到满足基本条件或达到原来新序列的边界。

10) 若前两种调整都不能找到满足条件的序列，则仍然保留原来的两个 CpG 岛，否则定义为新的 CpG 岛。

1.3 方法比较

将三种公开的 CpG 岛预测工具 CpGIE2.0、CpGi130、CpGProD 与 CpGI Seeker 进行了 CpG 岛预测准确性的比较。CpGIE 来自 http://bioinfo.hku.hk/cpgieintro[9]。CpGi130 的 perl 程序来自 http://ccnt.hsc.usc.edu/cpgislands/[7]。CpGProD 来自 http://pbil.univlyon1.fr/software/cpgprod.html [8]。用来进行 CpG 岛分析的 DNA 序列为 NT_006576，区域：1243289~1290159，总长 46871 bp。该序列来自 NCBI 美国国家生物技术信息中心网站（http://www.ncbi.nlm.nih.gov/）。

2 结 果

2.1 CpGI Seeker 具有的精确识别 CpG 岛的特性

我们将 CpGI Seeker 与 CpGIE2.0、CpGi130、CpGProD 进行了比较。采用的参数为 GC 最低含量 55%，CpG 最小出现率为 0.65，CpG 岛的最小尺寸为 500 nt。具体例子为 DNA 序列 NT_006576，区域：1243289~1290159，总长 46871 bp。

4 种 CpG 岛预测算法得到的 CpG 岛如表 1 所示。其中 CpGI Seeker 与 CpGi130 预测得到的 CpG 岛个数为 9 个，而 CpGIE2.0 与 CpGProD 得到的 CpG 岛个数为 8 个。其中 CpGProD 得到的 8

Table 1 Compare of four algorithms predicting CpG islands

Algorithm	Start	End	Length	GC Content (%)	CpG observed/expected
CpGI Seeker	17273	17790	518	61.6	0.65
	19133	19771	639	61.5	0.67
	20165	20747	583	62.1	0.70
	21432	21978	547	56.1	0.65
	25193	25750	558	56.8	0.65
	29259	29758	500	60.4	0.65
	32514	33202	689	58.3	0.65
	35910	37988	2079	69.2	0.69
	39541	42888	3348	69.6	0.71
CpGIE2.0	17286	17791	506	61.3	0.66
	19133	20759	1627	61.4	0.72
	21435	21999	565	56.2	0.65
	25227	25758	532	57.8	0.65
	29259	29768	510	60.4	0.65
	32512	33215	704	58.4	0.65
	35910	37998	2089	69.2	0.68
	39546	42898	3353	69.7	0.71
CpGi130	17274	17790	517	61.7	0.65
	19133	20747	1615	61.4	0.72
	21435	21982	548	56.0	0.65
	25193	25749	557	56.7	0.65
	29259	29758	500	60.4	0.65
	32514	33202	689	58.3	0.65
	35910	37988	2079	69.1	0.69
	39610	40268	659	59.4	0.65
	40427	42888	2462	72.3	0.75
CpGProD	17272	17792	521	61.6	0.65
	19133	20747	1615	61.4	0.72
	21429	21989	561	56.3	0.63
	25159	25784	626	57.2	0.59
	29259	29758	500	60.4	0.65
	32394	33323	930	55.1	0.60
	35910	37988	2079	69.2	0.69
	39541	42888	3348	69.6	0.71

Note: The subject DNA sequence is NT_006576 with the region of 1243289~1290159 and the length of 46871 bp

个 CpG 岛中有 3 个不符合参数的要求。四种算法得出的 CpG 岛在主要位置上都比较一致，但具体在每个 CpG 岛的精确定位却有明显区别。以 CpGI Seeker 预测得到的 CpG 岛为参照，有明显区别的 3 个 CpG 岛如图 1 所示。图 1 显示了 CpGI Seeker 预测得到的第 2 号、第 3 号和第 9 号 CpG 岛，其中 DNA 序列上的 CpG 二核苷酸用竖线标识。在图中可见 CpGI Seeker 识别出的两个 CpG 岛（No.2 和 No.3）在其它三种算法中合并成一个。虽然从 CpG 岛的整体特征上来看，合并后的 DNA 序列仍满足 CpG 基本要求，但实际在序列中的子序列（19780～20031，总长 252 nt）并不满足 CpG 期望个数的最低条件，即 CpG_{exp}×16>length。同时，拆分后的两个 CpG 岛更有利于判断 CpG 甲基化的位置。另外，CpGI Seeker 识别出的一个 CpG 岛（No.9），和算法 CpGIE 以及算法 CpGProD 一致，但算法 CpGi130 却将该 CpG 岛拆成了两个。CpGI Seeker 识别出的这段序列的任何子序列中都显示出了很高的 CpG 出现率，因而该 CpG 岛已经不宜于再次拆分。图 1 也可以清楚地显示出这个结果。综合而言，CpGI Seeker 在 CpG 岛的精确识别方面以及合并分割上体现了比其他三种算法更明显的优势。

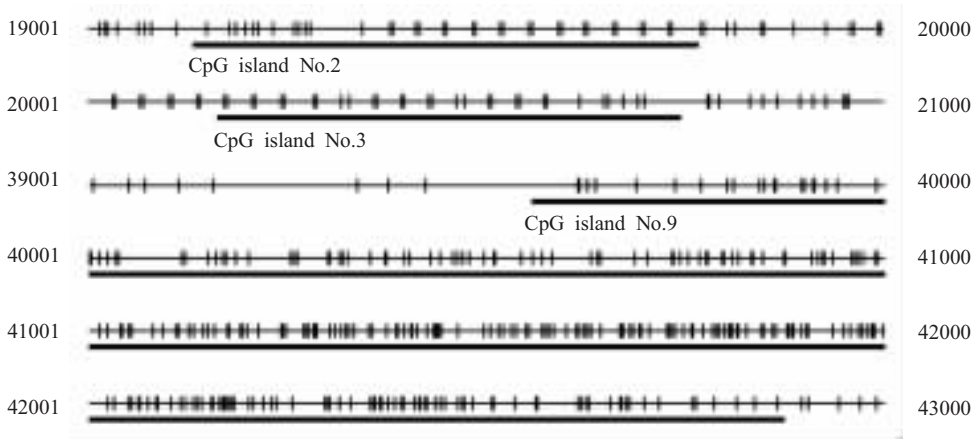

Fig.1 Schematics for parts of predicted CpG islands by CpGI Seeker. The subject DNA sequence was NT_006575, with the region of 1243289~1290159 and the length of 46871 bp. The figure shows the 2nd, 3rd and 9th CpG Islands predicted by CpGI Seeker. CpG dinucleotide was denoted by vertical line

2.2 组合参数选择与全基因组 CpG 岛的预测

Gardiner-Garden 和 Frommer 提出的 CpG 岛标准[5]（GC 含量至少 50%、CpG 最小出现率为 0.60，CpG 岛最小为 200 nt）是经典的预测标准（以下简称标准 1）。而 Takai 和 Jones[6]通过对 21 号、22 号染色体的分析认为 GC 含量 55%、CpG 出现率 0.65 以及 CpG 岛尺寸为 500 nt 形成组合参数而得到的 CpG 岛具有与基因的 5' 端更密切的关联性，也更有利于去除 Alu 重复序列（以下简称标准 2）。为了更全面、更广泛地对 CpG 岛的预测标准进行分析，在 Takai 和 Jones 的基础上，我们取参数 GC 最低含量分别为 50%、55%、60%，CpG 最小出现率为 0.60、0.65、0.70，参数 CpG 岛的最小尺寸为 500 nt，组合成 9 组组合参数；同时为了能够与经典标准进行比较，我们又在 CpG 岛的最小尺寸为 200 nt 的基础上，增加了 GC 含量为 50%，CpG 最小出现率分别为 0.60、0.65、0.70 的 3 组组合参数以及 GC 含量为 55%、CpG 最小出现率为 0.65 的 1 组组合参数，总共形成了 13 组组合参数，详见表 2。在这 13 组组合参数的基础上，我们在人类的全基因组上运用 CpGI Seeker 进行 CpG 岛的预测。

Table 2 13 sets of combined parameters predicting CpG islands

Parameter set number	Size (nt)	GC content	CpG observed/expected
1&	200	50%	0.60
2	200	50%	0.65
3	200	50%	0.70
4	200	55%	0.65
5	500	50%	0.60
6	500	50%	0.65
7	500	50%	0.70
8	500	55%	0.60
9*	500	55%	0.65
10	500	55%	0.70
11	500	60%	0.60
12#	500	60%	0.65
13	500	60%	0.70

Note: The 1st combined parameters (labeled by &) are the criteria of reference[5]. The 9th combined parameters (labeled by *) are the criteria of reference[6]. The 12th combined parameters (labeled by #) are the most appropriate one in current study

13 组不同组合参数预测得到的 CpG 岛统计结果如表 3 所示。3 个参数（岛最小尺寸、GC 最低含量、CpG 出现率）对 CpG 岛个数的显著性 *P* 值分别为 0、0.2053 和 0.13。因而岛最小尺寸对 CpG 岛个数的影响是显著的，而其它两个参数影响并不明显。3 个参数（顺序同上）对 CpG 岛总长的显著性 *P* 值分别为 0.0003、0.0101 和 0.0107。因而 3 个参数对总长都具有显著性作用。预测标准最宽松的 CpG 岛（第 1 组组合参数）在全基因组上有 325 093 个，而最严格的 CpG 岛（第 13 组组合参数）有 23 384，后者只为前者的 7.19%。并且随着预测标准严格性的增加，CpG 岛的个数也在逐步减少。

Table 3　Statistics of CpG islands predicted by 13 sets of combined parameters

Parameter set number	Number of CpG islands	Total length of CpG islands (bp)	Ratio of the length to the whole genome (%)
1	325 093	124 403 922	4.04
2	236 308	93 006 221	3.02
3	197 787	75 907 754	2.47
4	187 724	72 651 919	2.36
5	74 836	72 128 746	2.34
6	55 875	55 402 284	1.80
7	43 789	43 711 931	1.42
8	44 541	49 291 491	1.60
9	36 574	40 811 069	1.33
10	30 143	33 691 996	1.10
11	30 871	35 964 830	1.17
12	26 830	31 122 772	1.01
13	23 384	26 874 098	0.87

从个数上看，第 9～12 组组合参数得到的 CpG 岛个数与人类全基因组的基因个数比较相近。第 13 组组合参数得到的 CpG 岛个数已经远小于人类全基因组的基因个数。从预测 CpG 岛的序列总长来看，预测标准严格性的增加也使得 CpG 岛的序列总长在不断缩小。从整体上看，CpG 岛占人类全基因组总长约 1%～4%，与文献[10]报道的“人类基因组大约有 1%～2%的序列有 CpG 岛组成” 基本相符。第 1～5 组的 CpG 岛总长都大于报道的最高值 2%，第 13 组参数的 CpG 岛总长占全基因组总长的 0.87%，小于文献报道最低值 1%，而第 6～12 组的 CpG 岛总长为全基因组总长的 1%～1.8%，符合文献[10]的报道。

2.3　CpG 岛的重复序列分析与位置分布分析

在 13 组组合参数预测得到 CpG 岛的基础上，我们做了人类重复序列的分析，得到了含有不同百分比重复序列的 CpG 岛的数量占总数的百分比（表 4）。同时，表 5 还提供了 CpG 岛的 3 个基本参数（岛最小尺寸、GC 最低百分比与 CpG 最低出现率）在不考虑交互效应的情况下对重复序列的显著性 *P* 值分析。表 4 显示出前 4 组组合参数与后 9 组组合参数得到的 CpG 岛有较明显的差异：前 4 组得到的 CpG 岛所含有的重复序列百分比比较类似，并且比后 9 组所包含的重复序列高得多，甚至差别非常显著，例如第 4 组组合参数得到的 CpG 岛有 18.09%全部由重复序列组成，但第 5 组参数只有 2.93%的 CpG 岛全部由重复序列组成。表 5 给出的数据说明了造成这一显著差别的主要原因是参数 CpG 岛的最小尺寸从 200 bp 到 500 bp 的改变，即岛最小尺寸在 7 种重复序列含量中的显著性 *P* 值均小于或等于 0.0002。GC 最低百分比对重复序列含量位于 40%～80%的序列也具有较明显的显著性因素（$P<0.01$）。

通过对不同参数组合得到的 CpG 岛做位置分布的分析，得到了属于 6 种不同的基因转录起始位点相关位置分布的 CpG 岛占 CpG 岛总数的百分比（表 6）。同时，表 7 还提供了 CpG 岛的 3 个基本参数在不考虑交互效应的情况下对基因组定位的显著性 *P* 值分析。

Table 4 Repeat content analysis of CpG islands predicted by 13 sets combined parameters

Parameter set number	40%	50%	60%	70%	80%	90%	100%
1	70.06	68.92	67.19	63.06	52.97	34.29	12.09
2	68.62	67.6	66.22	62.4	52.17	33.26	12.16
3	68.2	67.22	65.91	62.16	51.3	32.99	12.31
4	65.82	65.08	64.25	63.1	60.28	50.63	18.09
5	54.88	51.13	42.11	36.4	29.36	18.4	2.93
6	48.89	45.73	37.15	32.02	25.37	16.08	2.9
7	42.55	40.04	32.68	27.49	21.34	13.54	2.63
8	31.91	29.46	25.82	23.96	21.99	16.91	2.62
9	27.62	25.46	22.55	20.92	19.26	14.57	2.57
10	21.15	19.47	17.19	15.87	14.51	10.53	1.82
11	14.63	13.61	12.79	12.3	11.59	10.14	4.27
12	11.92	10.99	10.27	9.81	9.18	8.15	3.83
13	8.51	7.33	6.66	6.07	5.32	4.68	2.45

Note: 7 levels of repeat content of the CpG islands were analyzed in the statistics ranging from 40% to 100% separately. The table shows the percentages (%) of CpG islands under the different levels of repeat content. The percentages were computed by the number having the level of repeat to the total number of CpG islands in 13 different sets of combined parameters

Table 5 The *P*-value for the repeat analysis of the three basic parameters of CpG islands without considering interaction

Basic parameter	40%	50%	60%	70%	80%	90%	100%
Size	0.0002	0.0001	0	0	0	0.0002	0
GC content	0.0002	0.0002	0.0002	0.0005	0.0063	0.1262	0.374
CpG observed/expected	0.1526	0.1567	0.1061	0.0981	0.1456	0.3595	0.5573

Table 6 The association between the distribution of CpG islands and transcription initial sites (TIS) in the 13 sets of combined parameters

Parameter set number	A1	A2	A3	A4	A5	A6
1	6.44	2.23	3.39	2.02	6.48	81.76
2	6.53	2.64	4.56	2.41	6.47	80.03
3	6.60	3.00	5.31	2.79	6.35	78.80
4	6.72	3.03	5.67	2.76	6.63	78.20
5	7.27	2.46	14.42	2.11	6.35	71.25
6	7.34	2.85	18.91	2.44	6.07	66.92
7	7.41	3.32	23.29	2.98	5.98	62.22
8	7.72	2.89	23.87	2.53	6.76	61.68
9	8.01	3.26	28.48	2.89	6.58	56.96
10	8.19	3.73	33.40	3.53	6.64	51.51
11	8.08	3.25	32.73	3.08	7.17	52.78
12	8.33	3.56	37.01	3.44	6.95	48.40
13	8.43	3.98	40.98	4.01	6.87	43.95

Note: The table shows the percentages of CpG islands in different genomic locations. The percentages (%) were computed by the numbers of CpG islands in the given location to the total numbers of CpG islands. The type of A1 to the type of A6 represents 6 different distributions of CpG islands. In type A1, a CpG island ends in the distant upstream of TIS (between upstream 8 kb and upstream 2 kb). In type A2, a CpG island ends in the near upstream of TIS (between upstream 2 kb and TIS). In type A3, a CpG island overlaps the TIS. In type A4, a CpG island starts in the near downstream of TIS (between TIS and downstream 2 kb). In type A5, a CpG island starts in the distant downstream of TIS (between downstream 2 kb and downstream 8 kb). In type A6, a CpG island exists besides the above 5 types of distributions

Table 7 The *P*-value for the genomic location of CpG islands of the three basic parameters of CpG islands without considering interaction

Basic parameter	A1	A2	A3	A4	A5	A6
Size	0	0.00001376	0	0.0178	0.02	0
GC content	0.0003	0.00000007	0.0003	0	0.0001	0.0002
CpG observed/expected	0.0691	0.00000002	0.0212	0	0.0294	0.0103

从处于区域 A6 的 CpG 岛个数百分比中可以看出，除了第 12、13 组组合参数以外，CpG 岛超过一半以上处于非基因转录起始附近区域（±8 kb）。第 5～13 组的组合参数预测得到的 CpG 岛覆盖基因转录起始位点的可能性较大（区域 A3）。第 9 组组合参数得到的 CpG 岛有 28.48%覆盖了基因的转录起始位点，这比文献[6]报道的同样条件得到的 CpG 岛个数高得多（文献报道为 161 个 CpG 岛覆盖了转录起始位点，比率为 14.6%，即 161/ 1101）。第 13 组组合参数显示出该标准的 CpG 岛与基因转录起始位点具有最大的相关性，即 40.98%。从表 7 可以得出除了参数 CpG 最低出现率相对于区域 A1 以外，3 个参数对 CpG 岛的大部分位置分布都起了较显著的作用（$P<0.05$）。

3 讨　论

3.1 已有 CpG 岛判定标准的分析与预测方法改进

标准 1 与标准 2 这两种判定标准的不同主要体现在 GC 含量、CpG 出现率以及 CpG 岛的尺寸这三个基本参数上。其中 CpG 出现率是 CpG 二核苷酸的观测个数与期望个数的比率。CpGIE、CpGi130、CpGProD 等不同的 CpG 岛预测方法虽然都能在同一个 CpG 岛判定标准上进行预测，但预测出的 CpG 岛却各不相同。预测 CpG 岛方法的一种基本模式是：先寻找满足基本参数的第一个搜索窗，然后移动搜索窗，直到不满足条件；再对得到的大窗口进行调整直至大窗口也满足基本条件，并定为一个新的 CpG 岛；最后将找到的所有 CpG 岛进行尽可能多的合并。

CpGi130 在具体的算法实现中采用 200 nt 作为移动搜索窗的跳距，同时合并 CpG 岛时要求两岛距离不超过 100 nt[6]。而 CpGProD 则采用了 1 nt 作为条件，200 nt 作为合并 CpG 岛的最大距离[8]。同时文献[9]还描述了 CpGIE 在合并 CpG 岛时具体的操作细节，包括若连接多个相邻的 CpG 岛后不满足基本参数时，应采用 5'、3' 端各减少 1 nt 的方法直至满足条件，否则保留原来的 CpG 岛。

总体而言，CpGIE 方法比 CpGi130 与 CpGProD 更具有识别 CpG 岛的潜力：在文献[9]提到的序列 NT_000874.1 中，CpGi130 找到 13 个 CpG 岛，CpGProD 找到 10 个 CpG 岛，而 CpGIE 则找到 12 个 CpG 岛。通过具体分析 3 种算法找到的 CpG 岛，结果显示 CpGIE 找到的 12 个 CpG 岛包含了 CpGi130 的 13 个 CpG 岛，即 CpGi130 找到的 2 个 CpG 岛被 CpGIE 合并成一个。此外，CpGProD 只丢掉了一个 GC 含量与 CpG 出现率都很接近最低基本参数的 CpG 岛，而另两个预测出来的 CpG 岛（大小分别为 2 305 bp 与 2 544 bp）则在 CpGIE 中分别变成了两个小 CpG 岛（大小为 1 211 bp 与 666 bp，970 bp 与 1 454 bp）。从这些结果来看，CpGProD 评估 CpG 岛的能力与 CpGIE 相当。但 CpGProD 预测得到的 CpG 岛会出现 CpG 最低出现率不满足条件的情况，所以需要在搜索策略上进行加强。而 CpGIE 的合并 CpG 岛的思想值得我们进一步挖掘。

CpGI Seeker 在 CpGIE 预测方法的基础上，结合 CpGProD 与 CpGi130 的思想并做了改进：将搜索窗的跳距改为 1 nt，增加搜索密度；采用 CpGProD 对 CpG 期望个数的计算方式；设定相邻两个 CpG 岛能够进行合并的最大距离为 100 nt；合并 CpG 岛根据实际情况进行新 CpG 岛的调整，并在失败时采用保留原来 CpG 岛的方式。具体算法详见材料与方法。具体的比较结果（图 1、表 1）显示，CpGI Seeker 有较 CpGIE、CpGi130 和 CpGProD 更精确的 CpG 岛识别特性。CpGI Seeker 能够消除预测不正确的 CpG 岛，同时对于 CpG 岛的精确定位也更加仔细。

3.2 参数选择与全基因组 CpG 岛的综合分析

从第 1 组组合参数到第 13 组组合参数，参数的严格性基本上是逐步递增的。随着参数严格性的增加，预测得到的 CpG 岛个数逐渐减少，CpG 岛总长也逐渐降低。结果表明，第 6～12 号组合参数更符合真实的基因组的 CpG 岛情况。

同时随着参数严格性的增加，CpG 岛中所含

重复序列的含量也在逐步降低。3 个基本参数中，岛最小尺寸对重复序列的含量起着最大的显著性作用。GC 最低百分比对部分含重复序列的 CpG 岛也起着显著性作用。重复序列的分析结果还提示，CpG 岛的尺寸变大使得一部分原来长度较短的 CpG 岛无法在严格的标准中成为 CpG 岛，而这一部分长度较短的 CpG 岛则很可能大部分由重复序列组成。第 5～7 组组合参数，用其预测的含 40%～50%重复序列的岛个数比第 8～12 组参数要多很多。而第 5～7 组组合参数与第 8～13 组的主要区别是 GC 含量只需要为 50%，这就使得一部分重复序列由于本身满足 50%的 GC 含量而形成 CpG 岛。虽然部分重复序列可以导致甲基化，但主要由非重复序列构成 CpG 岛更是人们关心的焦点。因而第 8～13 组组合参数更有利于我们预测 CpG 岛。

CpG 岛的位置分布说明了甲基化的广泛分布。甲基化不仅出现在基因转录起始位点附近的区域，在外显子以及基因间区域出现的可能性也很大。但从区域 A3 显示，从第 5 组开始，已经有超过 14%的 CpG 岛出现在基因的转录起始位点，这个比率甚至从第 10 组开始就超过了 32%，在第 13 组参数下得到了 40.98%的最高值。所以把研究的重心集中在覆盖转录起始位点而不是远离转录起始位点的 CpG 岛更有利于集中研究甲基化对基因转录的调控作用。

随着 CpG 岛参数严格性的增加，预测得到的 CpG 岛与基因转录调控的关联性增大，同时基因无关性也随之降低。同样在第 9 组组合参数的条件下，CpGI Seeker 得到的覆盖基因转录起始位点的 CpG 岛比 CpGi130 高得多，也充分说明 CpGI Seeker 算法的 CpG 岛基本判定准则与 CpG 岛搜索策略比 CpGi130 优秀，得到的 CpG 岛也更加精确。

根据 CpG 岛预测的总体统计，从序列组成是否有重复序列以及相对于转录起始位点的位点分布的情况来看，在 CpG 岛预测标准的 3 个基本参数中，CpG 岛的最小尺寸是具有最显著区分力的一个参数。CpG 岛最小尺寸的增大使得预测得到的 CpG 岛个数大大减小，同时 CpG 岛的碎片也大大降低。其次，在同样的变化情况下，GC 最低含量比 CpG 出现率更具有筛选 CpG 岛的特点。3 个基本参数的严格性逐步递增造成了 CpG 岛预测个数逐步减少，同时其序列构成也逐步减少了重复序列的含量。严格性增加还使得预测的 CpG 岛与基因关联性增加，特别是和转录起始位点相关的 CpG 岛的个数得到了增加。因此，更加严格的 CpG 岛预测标准虽然有可能丢失一些真实的可能甲基化的 DNA 片段，但它让我们能够更加集中地研究与转录调控密切相关的 CpG 岛，从而更有利于找到易于甲基化甚至高甲基化的区域。但严格性的增加也同样会导致真实 CpG 岛的丢失。虽然目前暂时还无法得知哪些 CpG 岛丢失了，但是从 CpG 岛的总长情况来看，第 13 组组合参数已经不满足基因组的真实情况了。因而相比之下，在目前的状况下，利用第 12 组组合参数，即最小尺寸为 500 nt、GC 最低含量为 60%、CpG 最低出现率为 0.65，作为判定标准进行 CpG 岛预测更易于发现甲基化的区域。

本文所进行的 CpG 岛全基因组分析主要从 CpG岛个数、长度，以及重复序列含量与基因转录起始位点的相关性这几个方面进行分析。在重复序列含量的分析中，本文并没有具体区分重复序列本身的类别，这也需要在下一步进行深入的分析。现有的分析结果表明，CpG 岛中确实有很大一部分有重复序列组成，因而那些对甲基化有明显帮助的重复序列值得我们进一步研究。CpG 岛的位置与基因转录起始的关系也验证了 CpG 岛对基因转录的重要作用。同时芯片技术的广泛应用，使得甲基化的实验研究能够更广阔地开展。结合具体的实验数据可以更有效地改进现有的 CpG 岛算法，也有助于研究者们发现更多的 CpG 岛区域与甲基化区域，探索甲基化与基因调控的关系。

参考文献：

[1] Antequera F, Bird A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci USA*, 1993,90(24): 11995～11999

[2] Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev*, 2002,16(1):6～21

[3] Panning B, Jaenisch R. RNA and the epigenetic regulation of X chromosome inactivation. *Cell*, 1998,93(3):305~308

[4] Feil R, Khosla S. Genomic imprinting in mammals: an interplay between chromatin and DNA methylation? *Trends Genet*, 1999,15(11):431~435

[5] Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*, 1987,196(2):261~282

[6] Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci USA*, 2002,99(6):3740~3745

[7] Takai D, Jones PA. The CpG island searcher: a new WWW resource. *In Silico Biol*, 2003,3(3):235~240

[8] Ponger L, Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 2002,18(4):631~633

[9] Wang Y, Leung FC. An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, 2004,20(7):1170~1177

[10] Bird AP. CpG-rich islands and the function of DNA methylation. *Nature*, 1986,321(6067):209～213

PREDICTION AND ANALYSIS OF CpG ISLANDS IN THE HUMAN GENOME

ZHUANG Hai-bin[1], ZHU Jing-de[2], LIU Xiang-jun[1]

(*1. Department of Biological Science and Biotechnology, School of Medicine and Ministry of Education Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China; 2. Cancer Epigenetics and Gene Therapy, The State-key Laboratory for oncogenes and Related Gene Shanghai Cancer Institute, Shanghai Jiaotong University, Shanghai 200032, China*)

Abstract: Methylation of CpG Islands is one of the most important mechanisms in the epigenetic regulation of gene expression. Although there exist some criteria for CpG Islands prediction from DNA sequences, the appropriate selection of parameters is still a challenging problem. After comparing two classic criteria of CpG Islands and three prediction algorithms, the authors proposed an improved algorithm, named CpGI Seeker, which showed the better performance than other algorithms. Through using the CpGI Seeker, they did the prediction of CpG Islands in the human genome using 13 combinatorial parameters from 3 basic parameters. Moreover, they analyzed the repeat sequences appeared on the CpG Islands and the location of CpG Islands relative with the transcription initial sites (TISs) of genes. The results demonstrated that more strict criteria of CpG Islands lead to less repeat sequences and more associated with TISs of genes. It is found that the combinatorial parameters of size=500nt, GC%=60% and CpG observed/expected=0.65 is the best criterion for predicting CpG Islands at present.

Key Words: CpG island; Methylation; Epigenetics

This work was supported by grants from The National Natural Science Foundation of China (90412018, 30570850), The National Science Foundation and the National Research Program for Basic Research of China (2004CB518804), the Key Project of Chinese Ministry of Education (03180, 104232), Trans-Century Training Programe Foundation for the Talents by The Ministry of Education and The Shanghai Science Foundation (04DZ14006, 05DZ19318)

Received: May 16, 2006

Corresponding author: LIU Xiang-jun, Tel：+86 (10)62792997，E-mail：frankliu@mail.tsinghua.edu.cn; ZHU Jing-de, Tel：+86 (21) 64224285，E-mail：jdzhu@sjtu.edu.cn