Published in final edited form as:

Biochem Biophys Res Commun. 2009 May 15; 382(4): 643–645. doi:10.1016/j.bbrc.2009.03.076.

CpG islands: algorithms and applications in methylation studies

Zhongming Zhao[a,b,c,*] and **Leng Han**[a]

a *Department of Psychiatry and Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA*

b *Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA*

c *Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA*

Abstract

Methylation occurs frequently at 5′-cytosine of the CpG dinucleotides in vertebrate genomes; however, this epigenetic feature is rarely observed in CpG islands (CGIs) or CpG clusters in the promoter regions of genes. Aberrant methylation of the promoter-associated CGIs might influence gene expression and cause carcinogenesis. Because of the functional importance, multiple algorithms have been available for identifying CGIs in a genome or a sequence. They can be categorized into the traditional algorithms (e.g., Gardiner-Garden and Frommer (1987), Takai and Jones (2002), and CpGPRoD (2002)) or statistical property based algorithms (CpGcluster (2006) and CG cluster (2007)). We reviewed the features of these algorithms and evaluated their performance on identifying functional CGIs using genome-wide methylation data. Moreover, identification of CGIs is an initial step in many recent studies for predicting methylation status as well as in the design of methylation detection platforms. We reviewed the benchmarks and features used in these studies.

Keywords

CpG island; CpG cluster; CG clusters; Methylation; Epigenetics; Promoter; Prediction algorithm

CpG dinucleotides have been commonly observed to be only ~20-25% as what expected in most sequenced mammalian genomes [1,2]. Such a great deficit is attributed to the hypermutability of methylated CpGs to TpGs/CpAs [3]. It has been estimated ~80% of CpGs are methylated in mammalian genomes [4]. In contrast, CpGs in GC-rich regions such as CpG clusters and CGIs are usually unmethylated, which is an important feature in the promoter regions of genes and for the regulation of gene expression [4]. For example, hypermethylation of promoter-associated CGIs in tumor suppressor genes were found to cause carcinogenesis [5]. Although most promoter-associated CGIs remain unmethylated [6], recent studies revealed a sizable fraction of CGIs might be fully methylated in normal cells [6-9].

* Address correspondence to: Zhongming Zhao, PhD, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, PO Box 980126, Richmond, VA 23298-0126, USA, Phone: (804) 828-8129, FAX: (804) 828-1471, Email: E-mail: zzhao@vcu.edu.

Multiple algorithms for identifying CGIs

Multiple algorithms [10-14] have been developed for CGI identification and have been applied in numerous studies. They could be categorized into two groups: the traditional algorithms that are based on three sequence parameters (length, GC content, and ratio of the observed over the expected CpGs (Obs_{CpG}/Exp_{CpG})) and the algorithms based on statistical property in a sequence without imposing the three criteria in the traditional algorithms.

Among the traditional algorithms, the first was proposed by Gardiner-Garden and Frommer in 1987 [10] and its criteria are: length > 200 bp, GC content > 50%, and $Obs_{CpG}/Exp_{CpG} > 0.60$. Many repeat elements such as *Alu* repeats that are abundant in the genomes also satisfy these criteria. To avoid such problem, only the non-repeat portion of the genome sequences has been applied for searching CGIs in a genome [15]. This algorithm was later revised by Takai and Jones, who used more stringent criteria: length ≥ 500 bp, GC content ≥ 55%, and $Obs_{CpG}/Exp_{CpG} \geq 0.65$ [11], and by Ponger and Mouchiroud [12], who used length > 500 bp, GC content > 50%, and $Obs_{CpG}/Exp_{CpG} > 0.60$. The stringent criteria in Takai and Jones' algorithm largely solved the repeat problem [16]. A summary of the features is shown in Table 1.

Recently, two algorithms, CpGcluster and CG clusters, were developed based on statistical property of a sequence without imposing a base compositional *a priori* assumption. CpGcluster detects clusters of CpGs by statistical significance based on the physical distance between neighboring CpGs on a chromosome assuming that the distance distributions between neighboring CpGs differ in CGIs from bulk DNA sequences [13]. CG clusters are defined as CG-dense fragments detected based on empirical species-specific benchmarks, for example, a minimum of 27 CpG dinucleotides in a DNA sequence fragment of no more than 531 bp in length in the human genome [14]. The details of these benchmarks are shown in Table 1.

Our extensive comparison of Takai and Jones' algorithm with CpGcluster revealed a high false positive rate in CpGcluster, which largely limits its utility in searching promoter-associated CpG clusters in vertebrate genomes [17]. Here, we further compared Takai and Jones' algorithm with CG clusters developed by Glass *et al.* [14]. We found that their species-specific benchmark, for example, a minimum of 27 CpG dinucleotides in a DNA fragment ≤ 531 bp for the human, was approximate to the three sequence parameters in Takai and Jones' algorithm. For example, when we examined their length distribution in the human genome, only 431 (1.0%) out of 41,487 human CG clusters were shorter than 500 bp, the minimum length in Takai and Jones' algorithm. Compared to CG clusters, Takai and Jones' algorithm had a slightly higher proportion of promoter-associated CGIs (35.0% versus 32.3%, detailed data not shown). Moreover, identification of CG clusters is not as straightforward as that of CGIs in Takai and Jones' algorithm.

Evaluation on methylation status in the promoter-associated CGIs

Performance of a CGI identification algorithm relies on whether its identified CGIs tend to be associated with promoter regions and to be unmethylated or hypomethylated. Recently, we [17] evaluated the performance of Takai and Jones' algorithm and CpGcluster using genome-wide methylation data generated in Weber *et al.* [6]. The evaluation based on sensitivity (SN, the chance of identifying CGIs in hypomethylated promoter regions), specificity (SP, the chance of not identifying CGIs in hypermethylated promoter regions), and correlation coefficient (CC, the extent of agreement between the SN and SP) suggested that Takai and Jones' algorithm had overall better performance. Here, we extended this evaluation to the five major algorithms in Table 1. All the algorithms had high sensitivity but moderate specificity (Table 1). When we compared the overall performance, Takai and Jones' algorithm had the highest correlation coefficient (CC = 0.48), suggesting its best performance on methylation status. CpGProD and CG clusters had similar CC values with Takai and Jones', as both

algorithms utilize similar sequence information (see Table 1 and the above section). More DNA methylation status data at single base resolution, which is now available only in a small genome [18] or some specific regions in mammalian genomes [8,19], will provide further evaluation on the performance of these algorithms.

Application of CGIs in predicting and detecting methylation profiling

DNA methylation studies in genomic regions, chromosomes and genomes have accelerated recently thanks to the rapid advancement of high throughput technologies [20]. Using experimentally verified DNA methylation data, investigators predicted methylation status based on sequence attributes around CGIs such as DNA sequence patterns, repeats, transcription factor binding sites (TFBSs) and predicted DNA structure [21-24]. These studies greatly improved our understanding of the inherent relationship between CGIs, DNA composition (sequence, repeats and structure) and methylation status. Table 2 summarizes the algorithms for identifying CGIs in predicting DNA methylation status with some other genomic factors. For example, Feltus *et al.* (2003) used both Gardiner-Garden and Frommer's and Takai and Jones' algorithms to identify CGIs [21]. Das *et al.* (2006) used Takai and Jones' algorithm [22], while Fang *et al.* (2006) and Bock *et al.* (2006) applied less stringent criteria than Takai and Jones' algorithm: length > 400 bp, GC content > 50%, and $Obs_{CpG}/Exp_{CpG} > 0.60$ [23, 24]. Because sensitivity in identifying hypomethylated CGIs is likely high (Table 1), combining other information such as sequence attributes could increase specificity, as shown in these studies.

CGIs are regions of interest for detecting methylation profiling in large-scale experiments using different platforms. Interestingly, most studies used the annotation of CGIs available in the popular UCSC Genome Browser (http://genome.ucsc.edu/), which is based on Gardiner-Garden and Frommer's algorithm [25-30]. Takai and Jones' algorithm was used in other study [31]. Furthermore, investigators often modified these three sequence parameters in the traditional algorithms for their specific designs (Table 2) [6,8,32,33]. Although definition of CGIs varied in these studies, all of them focused on CpG-rich regions. While a single-base resolution provides the most useful information, Bock *et al.* (2008) suggested that, considering the currently available array-based platforms, it might be sufficient to measure average methylation level in the CpG-rich regions [34]. Importantly, no statistical property based algorithms has been applied in such methylation studies yet.

In summary, DNA methylation studies have greatly accelerated during the past three years and are expected to grow even faster due to the rapid advancement of high throughput technologies such as microarray and next-generation sequencing [35]. This review provides useful information and guidance on CGI identification algorithms for gene and methylation prediction, gene feature analysis, and epigenetic and epigenomics studies.

Acknowledgements

We thank Drs. Jacob L. Glass and John M. Greally for kindly providing us their annotations of CG clusters. This work was supported by a NIH grant (LM009598) from the National Library of Medicine, the Thomas F. and Kate Miller Jeffress Memorial Trust Fund, and Institutional Research Grant IRG-73-001-31 from the American Cancer Society.

References

1. Zhao Z, Zhang F. Sequence context analysis in the mouse genome: single nucleotide polymorphisms and CpG island sequences. Genomics 2006;87:68–74. [PubMed: 16316740]
2. Han L, Su B, Li WH, Zhao Z. CpG island density and its correlations with genomic features in mammalian genomes. Genome Biol 2008;9:R79. [PubMed: 18477403]

3. Matsuo K, Clay O, Takahashi T, Silke J, Schaffner W. Evidence for erosion of mouse CpG islands during mammalian evolution. Somat Cell Mol Genet 1993;19:543–555. [PubMed: 8128314]
4. Antequera F. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 2003;60:1647–1658. [PubMed: 14504655]
5. Esteller M. Epigenetics provides a new generation of oncogenes and tumour-suppressor genes. Br J Cancer 2006;94:179–183. [PubMed: 16404435]
6. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 2007;39:457–466. [PubMed: 17334365]
7. Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, Iwasaka T, Mukai T, Sakaki Y, Ito T. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res 2004;14:247–266. [PubMed: 14762061]
8. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 2006;38:1378–1385. [PubMed: 17072317]
9. Illingworth R, Kerr A, Desousa D, Jorgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, et al. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol 2008;6:e22. [PubMed: 18232738]
10. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol 1987;196:261–282. [PubMed: 3656447]
11. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA 2002;99:3740–3745. [PubMed: 11891299]
12. Ponger L, Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 2002;18:631–633. [PubMed: 12016061]
13. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martinez-Aroza J, Oliver JL. CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics 2006;7:446. [PubMed: 17038168]
14. Glass JL, Thompson RF, Khulan B, Figueroa ME, Olivier EN, Oakley EJ, Van Zant G, Bouhassira EE, Melnick A, Golden A, et al. CG dinucleotide clustering is a species-specific property of the genome. Nucleic Acids Res 2007;35:6798–6807. [PubMed: 17932072]
15. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. [PubMed: 11237011]
16. Wang Y, Leung FC. An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 2004;20:1170–1177. [PubMed: 14764558]
17. Han L, Zhao Z. CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? BMC Bioinformatics 2009;10:65. [PubMed: 19232104]
18. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008;452:215–219. [PubMed: 18278030]
19. Farthing CR, Ficz G, Ng RK, Chan CF, Andrews S, Dean W, Hemberger M, Reik W. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. PLoS Genet 2008;4:e1000116. [PubMed: 18584034]
20. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008;24:133–141. [PubMed: 18262675]
21. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. Predicting aberrant CpG island methylation. Proc Natl Acad Sci USA 2003;100:12253–12258. [PubMed: 14519846]
22. Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ. Computational prediction of methylation status in human genomic sequences. Proc Natl Acad Sci USA 2006;103:10713–10716. [PubMed: 16818882]
23. Fang F, Fan S, Zhang X, Zhang MQ. Predicting methylation status of CpG islands in the human brain. Bioinformatics 2006;22:2204–2209. [PubMed: 16837523]

24. Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet 2006;2:e26. [PubMed: 16520826]
25. Khulan B, Thompson RF, Ye K, Fazzari MJ, Suzuki M, Stasiek E, Figueroa ME, Glass JL, Chen Q, Montagna C, et al. Comparative isoschizomer profiling of cytosine methylation: the HELP assay. Genome Res 2006;16:1046–1055. [PubMed: 16809668]
26. Oakes CC, La Salle S, Smiraglia DJ, Robaire B, Trasler JM. A unique configuration of genome-wide DNA methylation patterns in the testis. Proc Natl Acad Sci USA 2007;104:228–233. [PubMed: 17190809]
27. Ehrich M, Turner J, Gibbs P, Lipton L, Giovanneti M, Cantor C, van den Boom D. Cytosine methylation profiling of cancer cell lines. Proc Natl Acad Sci USA 2008;105:4844–4849. [PubMed: 18353987]
28. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 2008;454:766–770. [PubMed: 18600261]
29. Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, Graf S, Tomazou EM, Backdahl L, Johnson N, Herberth M, et al. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Res 2008;18:1518–1529. [PubMed: 18577705]
30. Shann YJ, Cheng C, Chiao CH, Chen DT, Li PH, Hsu MT. Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. Genome Res 2008;18:791–801. [PubMed: 18256232]
31. Ushijima T, Watanabe N, Okochi E, Kaneda A, Sugimura T, Miyamoto K. Fidelity of the methylation pattern and its variation in the genome. Genome Res 2003;13:868–874. [PubMed: 12727906]
32. Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, Segal E, Pikarski E, Young RA, Niveleau A, Cedar H, et al. Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat Genet 2006;38:149–153. [PubMed: 16444255]
33. Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH. Large-scale structure of genomic methylation patterns. Genome Res 2006;16:157–163. [PubMed: 16365381]
34. Bock C, Walter J, Paulsen M, Lengauer T. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. Nucleic Acids Res 2008;36:e55. [PubMed: 18413340]
35. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008;9:465–476. [PubMed: 18463664]

Table 1

Benchmarks for CpG islands (CGIs) identification and evaluation on methylation status

Algorithm	Benchmarks	Evaluation on methylation status[a]		
		SN	S P	CC
Gardiner-Garden and Frommer (1987) [10]	Length > 200 bp, GC > 50%, Obs_{CpG}/Exp_{CpG} > 0.60	0.96	0.38	0.39
Gardiner-Garden and Frommer (1987) in repeat-masked genome [15]	Length > 200 bp, GC > 50%, Obs_{CpG}/Exp_{CpG} > 0.60 in repeat-masked genome	0.95	0.43	0.39
Takai and Jones (2002) [11]	Length ≥ 500 bp, GC ≥ 55%, Obs_{CpG}/Exp_{CpG} ≥ 0.65	0.93	0.62	0.48
CpGPRoD (2002) [12]	Length > 500 bp, GC > 50%, Obs_{CpG}/Exp_{CpG} > 0.60	0.95	0.52	0.46
CpGcluster (2006) [13]	Clusters of CpGs separated by median distance, significance *P*-value $<10^{-5}$	0.96	0.42	0.40
CG clusters (2007) [14]	Human ≥ 27 CpGs in DNA fragment ≤ 531 bp Mouse ≥ 24 CpGs in DNA fragment ≤ 585 bp	0.94	0.54	0.45

[a] Sensitivity $SN=\frac{TP}{TP+FN}$, Specificity $SP=\frac{TN}{TN+FP}$, and correlation coefficient $CC=\frac{TP\times TN-FN\times FP}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}}$ where *TP* (true positives) are hypomethylated promoters containing computationally identified CGIs; *TN* (true negatives) are hypermethylated promoters without detecting CGIs; *FP* (false positives) are hypermethylated promoters containing CGIs; and *FN* (false negatives) are hypomethylated promoters without detecting CGIs.

Table 2

Application of CpG islands (CGIs) in predicting or detecting methylation status

Algorithms	Length (bp)	GC content (%)	Obs_{CpG}/Exp_{CpG}	Sequence attributes	References
CGIs used in computational prediction of methylation status					
Gardiner-Garden and Frommer (1987)	200	50	0.60	DNA patterns	[21]
Takai and Jones (2002)	500	55	0.65	DNA patterns	[21]
				GC content, di- and tri-nucleotide count, *Alu* coverage, hexamers	[22]
				GC content, distribution of *Alu* Y, CpG ratio, TpG content, TFBS	[23]
Specific parameters	400	50	0.60	DNA sequence properties and patterns, repeat frequency and distribution, predicted DNA structure	[24]
CGIs used in experimental detection of methylation status					
Gardiner-Garden and Frommer (1987)	200	50	0.60		[25-30]
Takai and Jones (2002)	500	55	0.65		[31]
Specific parameters	200	65	0.80		[32]
	300	55	0.50		[33]
	400	50	0.60		[8]
	500	55	0.75		[6]