J Biosci (2019) 44:143
DOI: 10.1007/s12038-019-9961-8

Review

A review of computational algorithms for CpG islands detection

Rana Adnan Tahir[1,2,†], Da Zheng[1,†], Amina Nazir[1] and Hong Qing[1]*

[1]*Key Laboratory of Molecular Medicine and Biotherapy in the Ministry of Industry and Information Technology, Department of Biology, School of Life Sciences, Beijing Institute of Technology, Beijing, China*

[2]*Department of Biosciences, COMSATS University Islamabad, Sahiwal Campus, Islamabad, Pakistan*

**Corresponding author (Email, hqing@bit.edu.cn)*

[†]*Co-first authors.*

MS received 12 April 2019; accepted 11 September 2019; published online 31 October 2019

CpG islands are generally known as the epigenetic regulatory regions in accordance with histone modifications, methylation, and promoter activity. There is a significant need for the exact mapping of DNA methylation in CpG islands to understand the diverse biological functions. However, the precise identification of CpG islands from the whole genome through experimental and computational approaches is still challenging. Numerous computational methods are being developed to detect the CpG-enriched regions, effectively, to reduce the time and cost of the experiments. Here, we review some of the latest computational CpG detection methods that utilize clustering, patterns and physical-distance like parameters for CpG island detection. The comparative analyses of the methods relying on different principles and parameters allow prioritizing the algorithms for specific CpG associated datasets to achieve higher accuracy and sensitivity. A number of computational tools based on the window, Hidden Markov Model, density and distance-/length-based algorithms are being applied on human or mammalian genomes for accurate CpG detection. Comparative analyses of CpG island detection algorithms facilitate to prefer the method according to the target genome and required parameters to attain higher accuracy, specificity, and performance. There is still a need for efficient computational CpG detection methods with lower false-positive results. This review provides a better understanding about the principles of tools that will assist to prioritize and develop the algorithms for accurate CpG islands detection.

Keywords. Bioinformatics; computational algorithms; CpG island; *CpGcluster*; epigenetics; methylation

1. CpG islands

Cytosine-guanine di-nucleotide (CpG) sites are the DNA sequence regions where the cytosine is followed by the guanine in the linear arrangement of nucleotides in 5' to 3' direction. CpG islands (CGIs) are defined as CpGs clusters within CpG-depleted bulk DNA, as described by Gardiner and Frommer. CpG clusters containing high GC content and CpG percentage close to the expected ratio are recognized as CGIs. The threshold standard was also proposed for CGIs, which turned into a primary criterion in all CGI prediction tools. Gardiner and Frommer described the CGIs having subsequent features including a CpG frequency exceeding 0.6 in observed/expected (O/E), more than 50% of GC content, and the length of the island region greater than 200 bp (Gardiner-Garden and Frommer 1987)

Rigorous modifications were proposed (Takai and Jones 2002) with 0.65 of O/E frequency, 55% of GC contents and a minimum length of 500 bp. The minimum sequence length of the island was amplified to prevent Alu sequences. Alu sequences are described as short repetitive scattered elements of around 280 bp in length comprise high O/E frequency and GC content. Alu elements are involved in the regulation of tissue-specific genes and sometimes alter the expression of genes (Britten 1996). These elements also cause mutations in the human genome (Deininger and Batzer 1999). McClelland and Ivarie (1982) presented a Chi-square test for assigning and identifying the statistically significant CGIs. This technique is considered as a suitable approach according to the CGI definition to identify statistically substantial CpG clusters within CpG-depleted regions (Hackenberg *et al.* 2006).

2. CpG islands' significance

CGIs are the long stretches of DNA (0.5 – 2 kb) with high levels of CpGs and there are about 30,000 CGIs in the human genome. These usually reside near the promoters and the nucleotides remain unmethylated. Conversely, the intragenic CGIs are frequently found as methylated and remain inactive as internal promoters. The mechanisms underlying these contrasting patterns of CGI methylation are poorly understood (Jeziorska *et al.* 2017).

Sometimes, CGIs are abnormally found in other transcriptionally active genes and lead to aberrant methylation in acquired and inherited genomic rearrangements (Tufarelli *et al.* 2003; Jones and Baylin 2007; Ligtenberg *et al.* 2009). Naturally occurring or aberrant transcription within the CGIs are associated with DNA methylation of CpGs (Jeziorska *et al.* 2017). The methylation status of CGIs enables to understand the underlying regulatory roles of methylation in transcription because methylated CpGs are exceptionally associated with transcription.

3. Epigenetics of CpG islands

Epigenetic marks have a momentous role in genetic regulation through DNA methylation, chromatin structure remodeling, histone modification and small non-coding RNAs that do not induce changes in DNA sequences. DNA methylation has a vital role in cell differentiation and development (Jang *et al.* 2017).

3.1 *Methylation of CpGs in cancers*

DNA methylation induces epigenetic variations to the cytosine at CpGs through the addition of a methyl group (Robertson 2005). DNA methylation of CpG plays a significant role in the cellular lineage commitment and memory (Smith and Meissner 2013). It has also been reported that abnormal DNA methylation at particular CpG sites leads to the cancer progression through genome instability (Feinberg and Tycko 2004). DNA methylation anomalies and their underlying mechanisms associated with cancers still need to be addressed, although, variant DNA methylation patterns have been often identified in cancers (Timp and Feinberg 2013). Generally, high levels of methylation are observed in genes and intergenic regions, whereas the short hypomethylated CpG regions frequently interrupt the low-density CpG regions, and usually, methylation does not occur within the CpG-rich promoter sites. Simultaneously, differential methylation for specific cancers between the controlled and other cells is observed in CGIs (Wahlberg *et al.* 2016). DNA methylation patterns among different physiological conditions, development phases, and cell types are of enormous interest to interpret the mammalian gene regulation mechanism. Several analyses have been performed to identify the DNA methylation patterns across different cell types to understand the single CpGs. Numerous advanced studies are ongoing on a large scale to characterize the methylomes, but still, there is an incomplete understanding of normal cell functions, gene expression, disease and methylation (Jones 2012).

Frequent gene-specific and extensive variations of DNA methylation usually occur in cancer cells as compared to other normal cells (Feinberg and Tycko 2004). Generally, aberrant DNA methylations reported during development comprise global methylation eradication and re-establishment (Mayer *et al.* 2000). The reported quantity of the differentially methylated CpGs between tumor cell types ranges from 0.5% to 20% (de la Rica *et al.* 2013).

3.2 *Methylation states and CpG islands in promoters*

The chromosomal division into clusters with similar methylation states and correlation of these states with epigenetic marks and regulatory sequences still requires attention for deep insights. Significant efforts have been made, including ENCODE (Consortium EP 2012), but there are still substantial gaps in understanding the variations of epigenetic states. CGIs are mostly related to the gene promoters (Saxonov *et al.* 2006), while methylation at CGIs regions is linked with transcriptional repression (Bird 2002).

Advanced technology enabled researchers to directly compute methylation instead of inferred states based on the CpG density (Greally 2013). The increasing geography of important methylation patterns, as well as the relation with CGIs adjacent to promoter regions, is described by genome-wide studies. CGI ‘shores’, defined as a 2 kb of flanking sequence from a CGI, have been reported to be more dynamic than the CGI itself (Irizarry *et al.* 2009). The 2 kb region upstream and downstream of CGIs shores are defined as the ‘shelves’ (Bibikova *et al.* 2011). The ‘open sea’ sites beyond the shores (Sandoval *et al.* 2011), the large DNA methylation regions as ‘valleys’ and ‘canyons’ with low methylation regions have been identified (Hon *et al.* 2012; Xie *et al.* 2013; Jeong *et al.* 2014). The other domains, such as long-range epigenetic silencing’ (LRES) or activation (LREA) and ‘low-methylated regions’ (LMRs) of comparatively high or low methylation are also identified in tumor cells (Bert *et al.* 2013). These domains inevitably depend on the particularized methylation parameters and length. Furthermore, the domains’ relative constancy – for example, canyons and LREAs among conditions and cell types, are not fully described yet (Edgar *et al.* 2014). The understating of dynamic or static behavior of the domains and sites is a crucial step for assigning a function to DNA methylation.

3.3 *Methyl CpG proteins*

Hypermethylation of CGIs usually causes decreased plasticity and gene silencing, while hypomethylation of CGIs, with poor intergenic regions, can modify the DNA methylation landscape during the initiation and development of cancer. It is still challenging to interpret the DNA methylation patterns and their role in epigenome plasticity (Stirzaker *et al.* 2014).

There are different types of proteins involved in the interpretation and modulation of DNA methylation patterns comprising DNA methylation 'editors', 'readers' and 'writers'. CpG methylation patterns are mainly maintained and established by the DNA methylation 'writers' during differentiation and development. These proteins belong to the DNA Methyltransferase (DNMT) family consisting of DNMT1, DNMT3A, and DNMT3B (Okano *et al.* 1998). The DNA methylation 'readers' assist a multilayered regulatory process through the specific binding with methylated CpG di-nucleotides and functions as translators between histone modifications and DNA methylation (Hashimoto *et al.* 2010). These proteins are the members of SET and Ring finger-associated (SRA) domain family and Kaiso family and methyl CpG-binding domain (MBD) proteins (Filion *et al.* 2006). A recently reported DNA methylation group 'editor', that has not been extensively studied, also contains the ten-eleven translocations (TET) protein family. This protein family yields 5-hydroxymethylcytosine by oxidizing the carbon-5 methyl group into hydroxyl residue (Pastor *et al.* 2013). 5-hydroxymethylcytosine promotes the demethylation by converting into unmethylated cytosine in the series of pathways. All the reported types of epigenetic modifiers have a significant role in the interpretation and regulation of the gene expression, chromatin remodeling and DNA methylation (Du *et al.* 2015).

The explosion in large-scale high throughput sequencing data has created an urgent need for fast and efficient computational tools. Sequentially, advanced tools have been developed to analyze the available bulk data. The precise identification of CGIs associated with the epigenetic regulatory mechanisms from mammalian genomes is of great interest and significance. Various computational methods have been proposed for accurate CGI identification based on different algorithms. The evolution of CGI detection methods is depicted in figure 1. In this review, we summarize the recent extensively utilized computational methods to detect CGIs in human genomes. CGIs play a crucial role in chromosome inactivation and nucleosome retention, gene regulation, gene mutation, epigenetic inheritance, and DNA methylation. Various bioassays and computational tools have been developed based on the different strategies for CGIs detection in mammalian genomes. The computational methods, developed in the last ten years, are summarized in this review (table 1).

4. CpG islands detection algorithms

In 1987, *in silico* CGIs prediction was demonstrated for the first time in vertebrates (Gardiner-Garden and Frommer 1987), in consort with CGI definition, and later hitches have been resolved by advanced research. Numerous studies reported that CGIs only occur within the region of gene promoters while Alu repeat regions are eliminated. It also has been seen that CGIs occur in coding and non-coding regions accompanied by gene promoter regions (Meissner *et al.* 2008; Brunner *et al.* 2009).

The development of CGIs computational algorithms greatly accelerated the understanding of genome-wide methylation profiling and helped to determine the CpG-rich regions associated with gene regulation. The CGIs detection algorithms developed in the last decade are mentioned in table 1 along with their summarized and significant details, while these are discussed in detail in the following sections. *In silico* methods for CGI detection are mainly categorized into four classes based on their principal algorithms as window-based, Hidden Markov Model (HMM) based, density-based and distance-/length-based methods being applied in computationally CGI detection (Yu *et al.* 2017).

4.1 *Window-based methods*

Window-based methods (Rice *et al.* 2000; Takai and Jones 2002; Ponger and Mouchiroud 2002) examine the genome by scrolling window and identify the CGIs by canonical statistical standards. An approved algorithm (Takai and Jones 2002) moves the adjustable window for one nucleotide individually to compute the GC content and CpG O/E in the window until it attains the satisfactory CGI content. Afterward, it moves to the subsequent window to calculate the CGI again until the window does not fulfill the selected statistical standards. It again swings back to each nucleotide, until it attains the previous satisfied boundary window. This method is extensively applied due to its effective statistical standards, but a major limitation of this algorithm is the limited window size, which decides the success of prediction. The small window size upsurges the chances of failure to identify the potential CGI and reduces the computing complexity, while larger window size reduces the computing speed and raise the predictive granularity. The low sensitivity is also considered as a drawback of this method for whole CGI prediction where it contains different trivial segments (Yu *et al.* 2017).

Many CGI prediction algorithms are developed on Gardiner-Garden and Frommer (GGF) principle (Gardiner-Garden and Frommer 1987), for instance, CpGIS (Takai and Jones 2002), CpGProD (Ponger and Mouchiroud 2002), CpGplot (Olson 2002), and particle swarm optimization (PSO) based (Chuang *et al.* 2011) methods. These methods utilize the sliding window approach to scan the genome for CGI detection and cover a wide range of CGI features such as length thresholds, GC content and O/E ratio. Chuang *et al.* (2011) proposed a novel PSO based method that delivers accurate and simple CGI detection accompanied by few parameters and fast convergence. This method employs the sliding-window-based approach in which it sets the window into a multiple of 2,500 bp because of the chromosome sequences. Each window is subjected to PSO for CGI detection; however, DNA sequences scanning is time-consuming through PSO.

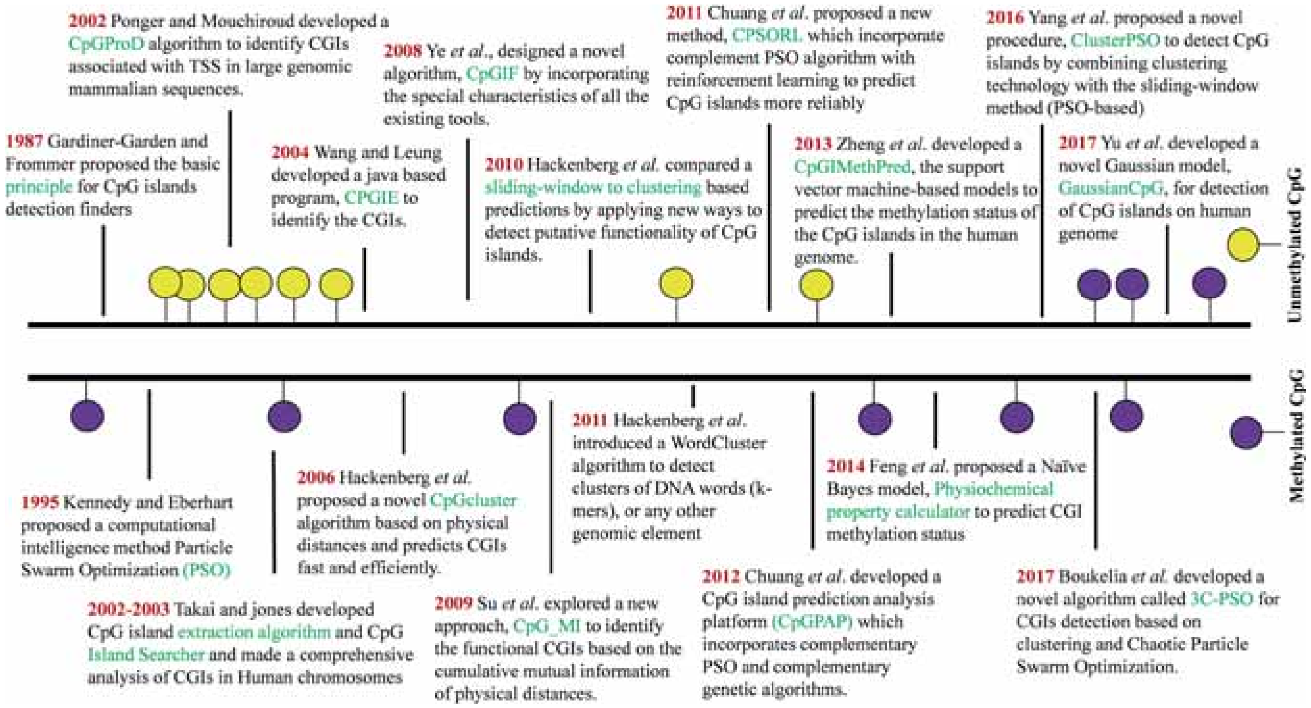
2002 Ponger and Mouchiroud developed a CpGProD algorithm to identify CGIs associated with TSS in large genomic mammalian sequences.
2008 Ye et al., designed a novel algorithm, CpGIF by incorporating the special characteristics of all the existing tools.
2011 Chuang et al. proposed a new method, CPSORL which incorporate complement PSO algorithm with reinforcement learning to predict CpG islands more reliably
2016 Yang et al. proposed a novel procedure, ClusterPSO to detect CpG islands by combining clustering technology with the sliding-window method (PSO-based)
1987 Gardiner-Garden and Frommer proposed the basic principle for CpG islands detection finders
2004 Wang and Leung developed a java based program, CPGIE to identify the CGIs.
2010 Hackenberg et al. compared a sliding-window to clustering based predictions by applying new ways to detect putative functionality of CpG islands.
2013 Zheng et al. developed a CpGIMethPred, the support vector machine-based models to predict the methylation status of the CpG islands in the human genome.
2017 Yu et al. developed a novel Gaussian model, GaussianCpG, for detection of CpG islands on human genome
Unmethylated CpG
Methylated CpG
1995 Kennedy and Eberhart proposed a computational intelligence method Particle Swarm Optimization (PSO)
2006 Hackenberg et al. proposed a novel CpGcluster algorithm based on physical distances and predicts CGIs fast and efficiently.
2011 Hackenberg et al. introduced a WordCluster algorithm to detect clusters of DNA words (k-mers), or any other genomic element
2014 Feng et al. proposed a Naïve Bayes model, Physiochemical property calculator to predict CGI methylation status
2002-2003 Takai and jones developed CpG island extraction algorithm and CpG Island Searcher and made a comprehensive analysis of CGIs in Human chromosomes
2009 Su et al. explored a new approach, CpG_MI to identify the functional CGIs based on the cumulative mutual information of physical distances.
2012 Chuang et al. developed a CpG island prediction analysis platform (CpGPAP) which incorporates complementary PSO and complementary genetic algorithms.
2017 Boukelia et al. developed a novel algorithm called 3C-PSO for CGIs detection based on clustering and Chaotic Particle Swarm Optimization.

Figure 1. The timeline of CGIs detection tools represents the major developments to identify the CpG-rich regions.

4.2 *Hidden-Markov-model-based methods*

HMM-based methods (Yoon and Vaidyanathan 2004; Wu *et al.* 2010; Chuang *et al.* 2012; Kakumani *et al.* 2012) calculate the transitive probability between and within the CGIs by applying the statistical transition model. The transition probability is calculated between two consecutive nucleotides during the training state for CGI and non-CGI regions, respectively. The contingency of CG pair in the non-CGI region is much lower than the CpG-rich region. Consequently, the variations between CpG and non-CpG regions are determined by the log-likelihood ratio of the probabilities for every possible sequence (Yoon and Vaidyanathan 2004). However, the efficacy of the HMM-based methods is significantly affected due to the lack of adequate data training and variant patterns. The computing efficiency of the HMM method is not very productive.

The application of HMM was implemented for sequence analysis and later, successfully, for genomes partition (Churchill 1989). Durbin *et al.* (1998) proposed the application of HMMs in CGIs detection. Subsequently, a technique based on HMMs was proposed that permits an extensible approach for CGI detection. The foremost feature of the current approach is that it determines the CGI status as the probability scores. This delivers the flexibility in CGI definition and enables to create CGI lists for other species. The first CGI lists for other species were generated utilizing this approach. CGI lists can also be generated by employing this approach and it significantly increases overlap with currently identified epigenetic marks (Wu *et al.* 2010).

4.3 *Density-based methods*

Density-based methods (Sujuan *et al.* 2008; Elango and Soojin 2011) instinctively determine the density of the CpG sites like window-based methods, which utilize the statistical standards. The percentage of CpG sites in CGI and the full length of CGI is calculated to compute the CGI density. The basic principle of this method is to set the initial seeds for repetitively regulating the density variables and thus increasing the coverage of CpG-rich regions. Initially, the estimated boundaries of CGIs are analyzed by adjusting a loose/low threshold value of the density. Afterward, the strict/high threshold value is utilized to determine the range of the CGI borders, where the sequence meets the density criteria. However, a linear model could probably not define the CpG distribution in CGI while this method heavily depends upon the threshold of the density that denotes linear association of CpG sites and total CpG length, which is considered as a drawback of this method (Yu *et al.* 2017).

Sujuan *et al.* (2008) designed a novel algorithm named CpGIF (CpG Island Finder) by incorporating the distinctive characteristics of all the existing tools to overcome the shortcomings of each method individually. Its algorithm was executed in PERL language including standard gateway interface and UNIX command line. The mathematical CGIs produced by high GC ratio is omitted by employing a density cutoff, which is also implicated in some other tools. Initially, the algorithm detects and records the locations of all the CpG dinucleotides in the sequence from 5' to 3' direction. It attempts to identify all the initial seeds having 0.10

Table 1. Comparison of CGIs detection tool and their characteristics

Software	*GaussianCpG*	*3C-PSO*	*ClusterPSO*	*CpGIMethPred*	*CpGPAP*	*WordCluster*	*CpG_MI*	*CpGIF*	*CpGcluster*
Publication year	2017	2017	2016	2013	2012	2011	2009	2008	2006
URL	–	–	–	*http://users.ece.gatech.edu/~hzheng7/CGIMetPred.zip*	*http://bio.kuas.edu.tw/CpGPAP/*	*http://bioinfo2.ugr.es/wordCluster/wordCluster.php*	*http://202.97.205.78/cpgmi/*	*http://www.usd.edu/~sye/cpgisland/CpGIF.htm*	*http://bioinfo2.ugr.es/CpGcluster/*
Accessibility	Web based	–	–	Standalone	Web based & Standalone	Web-based	Web-based & Standalone	Web-based	Web-based
Input Parameters	Not well defined Chosen by optimizing the biological statistics	GGF Criteria Population size = 300 particles p-value 0.01	GGF criteria p-value <_0.01	GGF criteria	Adjust the CpG parameters	Not Specific Distance models i. Percentile ii. Fixed iii.Chromosomal intersection iv. Genome intersection	GGF criteria	Default Density 0.10 later, density cutoff reduced from 0.9 to 0.5	Default parameters
Performance	Performance Ppv, PC, F1, Acc	high Specificity and Performance	High detection capability Acc, Cc, Pc, SN	High Specificity and accuracy	higher SN and a higher Cc	outperforms other methods based on densities and sliding-window	highest prediction accuracy	Higher correlation and Pc	Higher Ac, SP, and lower FP predictions
Algorithm	Gaussian Model	Hybrid approach	Hybrid approach	Support Vector Machine	Hybrid approach	Distance-based method	Distance-based method	Density-based Method	Distance-based method
Organism	Human Genome	Human Genome	Human Genome	Human Epigenome project	Human chromosomes	Human Genome	Mammalian genomes	Human chromosomes	Human and Mouse
Parameter Evaluation	TP[1], FP[2], FN[3], TN[4], SN[5], SP[6], Acc[7], Mcc, Ppv, Pc and F1 score	Cc, PC, Acc, SP, SN	Acc, SP, SN, Cc, PC	SN, Cc, Acc, SP,	SN, Cc, GC content, coverage O/E ratio	CGIs length, GC content & O/E ratio	length, CpGs No. GC content, O/E ratio	SN, SP, Ppv Pc	Acc, SP, FP SN
Cross Comparison with tools	i. CpGPlot ii. CpGReport iii. CpGProd iv. CpG-Cluster	i. CpGPlot, ii. *CpGcluster* iii. ClusterPSO	i. CpGPlot, ii. *CpGcluster*, iii. CpGProD, iv. CpGIS, v. CPSO vi. PSORL, vii. PSORL,	i. Compared the results with twelve different tissues	i. CpGPlot, ii, CpGProD, iii. CpGIS, iv. *CpGcluster*	i. Densities and sliding-window-based methods	i. CpGProD, ii. CpGIS, iii. *CpGcluster* iv. CpGIF	i. CpGIS, ii. CpGProd, iii. CpGPlot, iv.*CpGcluster*	i. CpGProd ii. CpGIS iii. CpGIE iv. CpGED

Table 1 (continued)

Pros	i. Simplify the complex interaction	i. Higher prediction performance ii. Capability to combine with other algorithms.	i. Accurate prediction ii. Computing efficiency	i. Capable to predict the methylation status ii. utilize unexplored features	i. User-friendly interface ii. graphical visualization, iii. application of associated programs iv. algorithm selection	i. It predicts clusters of DNA words ii. Co-localization with gene annotations	i. Epigenomic modifications ii. Genomic functional elements	i. Fast computing	i. CGIs start and end with a CpGs ii. Integer arithmetic for fast & efficient iii. minimal overlapping of Alu elements
Working	i. Find all CpG sites ii. Cluster these CpGs iii. Apply Gaussian filter iv. Filter clusters v. Collect the filtered clusters vi. Calculate % (G+C) and pick up the clusters	i. Clustering to find the potential CGIs ii. Optimization of CGIs through 3C-PSO	i. Detect the CGIs according to the distance and p-value ii. PSO then predicts the CGIs among candidates	i. Extract and identify the key features that are correlated with DNA methylation patterns and these features are utilized in CGIs predictions.	i. Selection of algorithm; ii. graphic visualization of results iii. application of related tools and dataset	i. Detection of K-mers copies ii. Calculation of distances between copies iii. Detection of clusters iv. Calculation of statistical significance of clusters	i. Locate and record the CpG sites ii. Calculate distances iii. Compute CpG numbers iv. Filter the CpG through extended CpG clustering	i. Scan the genome and record CpG positions ii. Identify all seeds iii. Keep updating the GC contents iv. initial seeds are extended v. Seeds are clustered	i. Clustering involves scanning and identifying all the possible CGIs clusters ii. p-value is assigned to screen the potential CGIs
Reference	Jeziorska *et al.* (2017)	Su *et al.* (2009)	Stirzaker *et al.* (2014)	Turner (2000)	McClelland and Ivarie (1982)	Robertson (2005)	Sandoval *et al.* (2011)	Ponger and Mouchiroud (2002)	Jeong *et al.* (2014)

[1] True Positive

[2] False Positive

[3] False Negative

[4] True Negative

[5] Sensitivity

[6] Specificity

[7] Accuracy

default density and records the number of Cs and Gs in the array. The density cutoff is iteratively decreased from 0.09 to 0.05 by extending initial seeds, and at last, adjacent extended seeds are clustered based on the smaller distance between them. CpGIF takes less time for CGIs prediction that is also an essential feature of the current algorithm (Sujuan *et al.* 2008).

4.4 *Distance-/length-based methods*

Distance-/length-based methods (Hackenberg *et al.* 2006) provide a speedy approach for the prediction of CGIs that assembles data in the context of the distance between CpG sites. This method examines the sequence property between any two inline CpG sites, which also brings criticism on this technique. The same CGI in different situations results in various outcomes, the low predictive sensitivity with irrelevant results due to the sequence composition is also considered as the drawback of this technique (Sujuan *et al.* 2008).

All existing methods were based on the vast parameter space made by the GC content, CpG fraction, and length threshold. The distribution of distance differs in CGIs and bulk DNA between adjacent CpGs due to the high number of CpG dinucleotides at CGIs. Hackenberg *et al.* (2006) developed a novel approach (*CpGcluster*) having the capability to directly determine the CpG clusters based on physical distances. The statistically significant clusters are declared as CGIs after assigning the p-value to each group. The test sequence was retrieved from the experimental CGI library by comparing *CpGcluster* predictions with existing CGI methods. It provides the highest degree of overlap with vertebrate phylogenetic conserved elements and, at the same time, the lowest overlap with Alu retrotransposons.

It has the capability to differentiate functional CGIs in the bulk genome because CGIs overlapping with transcription start site presents maximum statistical importance as compared to other genomic islands. The only integer application in arithmetic operations enables a computationally fast and efficient tool for predicting accurate and statistically significant clusters. The starting and ending with CpG dinucleotide of all the predicted CGIs is another superior feature. *CpGcluster* works only on the distance between adjacent CpGs that lead to low overlap and high specificity with Alu elements while other search parameters in other existing prediction methods are not required as major statistical and search parameters (Hackenberg *et al.* 2006).

WordCluster method simplifies the existing *CpGcluster* algorithm (Hackenberg *et al.* 2006) and is statistically substantial to the clusters of genomic elements. Numerous genetic elements, including conserved non-coding regions, genes, microRNA genes, CpG dinucleotides, and transcriptional factor binding sites are spatially clustered within the human genome. It was developed to locate the clusters of genomic elements by utilizing the set statistical criteria and distance between adjacent copies. The method also determines the CTG/CAG clusters exhibiting the rate of variations between the exterior and interior of clusters. The experiments have shown that the WordCluster method determines the valid clusters of genomic elements and DNA words as it identifies the key clusters of olfactory receptor genes. The application of the current algorithm is also available at a web server that provides additional functions to study the functional insights of overlapped genes by an annotation enrichment program (Hackenberg *et al.* 2011).

The accurate prediction of CGIs accompanied by the epigenetic regulatory function from large genome datasets is of keen interest to researchers. Su *et al.* (2009) developed a novel CpG mutual information (CpG_MI) model based on cumulative mutual information of physical distances between adjacent CpGs for the identification of functional CGIs. The algorithm exhibits the highest prediction accuracy and also explores the new functional CGIs coinciding with gene promoter regions as these are skipped by other approaches. The CpG_MI can also be employed in other mammalian genomes to determine the potential functional CGIs due to their similar cumulative mutual information and CpG di-nucleotide content in six mammalian genomes. It also assists in identifying the associations between epigenomic modifications and genomic functional elements. This model can efficiently scan large genome datasets and identify the potential functional CGIs (Su *et al.* 2009).

5. Other advanced methods

Numerous computational methods have been proposed based on different principles, with some pros and cons. Simultaneously, the aforementioned methods can either achieve higher sensitivity with the loss of specificity or higher specificity with low sensitivity. It also infers that the CGI definition may diverge from the ground truth (Glass *et al.* 2007). Different algorithms have also been proposed based on hybrid approaches and novel methods to attain high specificity and sensitivity.

5.1 *Hybrid algorithms*

Numerous hybrid approaches have been proposed by incorporating the features of window-based, clustering and density-based methods to overcome limitations such as window size, time-consuming and sensitivity. Hackenberg *et al.* (2010) have employed novel techniques to determine the putative functionality of CGIs by correlating the clustering approach, specifically *CpGcluster,* with the sliding-window-based algorithm. It has been observed through the co-localization of genomic regions that *CpGcluster* exhibits less overlap for Alu retrotransposons, at the same time, showing a higher overlap with conserved elements and promoter regions. *CpGcluster* exclusively predicts the CpG islets, which makes it different from another method. The islands predicted by the window-based method can counterfeitly overlap various methylation domains along with the regulated promoters, which designate the incorrect combination of many CGIs into a very long and single island. The predicted length of the island is the main difference between the clustering and sliding-window-based methods. Therefore, *CpGcluster* appears as a better choice for determining the short, but putatively functional, CGIs (Hackenberg *et al.* 2010).

CGIs have been described to understand and clarify gene regulation and local chromatin structures. Yang *et al.* (2016) developed a novel hybrid method by merging the clustering approach with the sliding-window-based method for CGIs detection. Generally, a clustering technique based on the physical distance directly calculates all the probable CpG clusters by neglecting the CpG criteria. The cluster-based algorithms can scan the entire genome for the CpG clusters, whereas the statistically significant CpG are scrutinized through p-value rule. The efficiency of sliding-window-based PSO also needs to be optimized by eliminating the large trivial processing of DNA fragments. Thus, a hybrid method, ClusterPSO was proposed for high accuracy and sensitive prediction of CGIs from the human genome. The clusterPSO prediction involves two main procedures. Initially, *CpGcluster* works by employing clustering technology and distance threshold to identify all the possible CpG candidates from DNA sequences and then validates through p-value. Subsequently, CGIs are predicted, from the scrutinized CpG candidates, through a PSO-based approach with defined CGI criteria. The CGI prediction in the human genome was conducted by ClusterPSO and other eight existing prediction methods to evaluate the efficacy of this method. ClusterPSO exhibits a higher detection capability when compared to all other assessed methods, in terms of accuracy, correlation coefficient, performance coefficient, specificity, and sensitivity. A hybrid approach of PSO and *CpGcluster* significantly reduced the accurate prediction and computing time, respectively (Yang *et al.* 2016).

Generally, CGIs are used to predict the promoter regions, but these could also be utilized as tumor markers when these have abnormal methylation in cancer cells. Boukelia *et al.* (2016) developed a hybrid approach, *i.e.,* 3C-PSO (Clustering and Complementary Chaotic-PSO) to identify and predict the CGIs in the human genome dataset. The clustering method scans the whole genome and identifies the potential CGI clusters, and these clusters are further refined by a complementary chaotic PSO, an optimization technique to find precise CGIs. The hybrid approach effectively overcomes the limitations of each method individually and delivers a high sensitivity detection of CGIs in the human genome. The performance of this method has been assessed by comparing five measures, such as correlation coefficient, performance coefficient, accuracy, specificity, and sensitivity, with other existing CGIs prediction methods. The comparative analysis was performed with CpGPlot, *CpGcluster* and ClusterPSO tools by considering six sequences from NCBI to evaluate the prediction outcomes of 3C-PSO in the human genome. The population size was set to 300 particles in the complementary chaotic PSO, while the p-value was adjusted to 0.01 and the distance threshold was set to 65th position in the clustering method. CGIs were defined for fitness functions by adjusting the parameters such as minimum length to 200 bp, O/E ratio to 0.6, GC content to 0.5 and the gap between nearby islands to 100 bp. This method surpasses the other existing tools and exhibits high specificity and performance (Boukelia *et al.* 2016).

A CpG island prediction analysis platform (CpGPAP) was developed to explore the genome sequences for methylation, medical and biological insights. This method incorporates complementary PSO and complementary genetic algorithms and delivers a user-friendly web-based interface for input sequences. The correlation coefficient and sensitivity of CpGPAP is higher, compared to *CpGcluster*, CpGIS, and CpGProD algorithms, over an entire chromosome. This algorithm possesses three major characteristics as graphical visualization, application of associated programs and datasets, and selection of prediction algorithm. The standalone version has no limitation of input sequence length and the visual display function; therefore, it works as a useful tool to explore and analyze the genomic CGIs. Initially, it works by employing optimization algorithms to predict CGIs and, subsequently optimization parameters are adjusted prior to the input sequences for CGIs association. At last, CGI-related information of GC content, O/E ratio, start and end position and length is predicted and presented in the graphical display. It allows the users to modify or adjust the CpG parameters and visual representation of prediction (Chuang *et al.* 2012).

5.2 *Gaussian model*

In silico approaches are based on the principle of GC content for CGI prediction, but it has been shown that experimentally-validated CGIs diverge from these artificial standards. Differences in GC content, variant patterns, and CGI lengths indicate that the accurate prediction of CGI is not a simple, linear, or statistical task. Therefore, there was an essential need to reveal the underlying mechanisms for accurate predictions of CGI and to develop efficient computational algorithms.

Yu *et al.* (2017) proposed a novel Gaussian model based on the potential energy of each CpG site that satisfies the distribution of Gaussian energy with its primary structure. The Gaussian model was developed to reflect the basics of microscopic connections in the complex human genome. Gaussian CpG is based on the Gaussian model and is designed to identify the CGIs in the human genome. Initially, it scans through the genomic primary structure to investigate the energy distribution for each CpG site, and then adjusts the statistical parameters for the human genome. This method efficiently predicts the CGIs due to adequate sensitivity and specificity in the known human CGI data. The objectives of the Gaussian model were to simplify the microscopical interactions between nucleotides based on Gaussian energy accumulation and distribution. The biological and statistical methods have been optimized for deliberate selection of the Gaussian function parameters. This novel method is validated for both artificial and real data sets by applying pseudo-potential analysis on CGIs (Yu *et al.* 2017).

5.3 *Naïve Bayes algorithm*

The significance of DNA methylation in several important mechanisms like cell differentiation and development has been documented in genome-wide studies of methylomes. A Naïve Bayes model (Feng *et al.* 2014) was developed to computationally predict the methylation status of the CGIs by utilizing the pseudo-tri-nucleotide composition. Naïve Bayes is a statistical algorithm, typically implicated in bioinformatics (Yousef *et al.* 2007; Feng *et al.* 2013b; Feng *et al.* 2013a) and considers the attribute variables as independent from each other given the outcome. Three DNA physicochemical properties were examined in pseudo-tri-nucleotide composition for DNA sequence formulation and three cross-validation methods including jackknife, independent dataset, and sub-sampling tests were used. The proposed model exhibited an 88.2% success rate in determining the CGI methylation status by the jackknife test. It is therefore considered as a valuable method to assess the methylation status of the CGIs.

5.4 *Support vector machine*

In silico prediction of DNA methylation can be utilized to scrutinize the main features associated with methylation patterns and for genome-wide methylation profiling. A CpGIMethPred tool based on support vector machine was proposed (Zheng *et al.* 2013) to determine the methylation status of the CGIs in the human genome under normal conditions. Recognized potential features including histone methylation status, CGI-specific features, distribution patterns of conserved elements and transcription factor binding sites, DNA structure patterns and DNA sequence composition patterns are considered for prediction. Unknown features comprising the histone acetylation status, gene functions, and nucleosome positioning tendencies are also incorporated for additional information.

Kolmogorov-Smirnov (Wang *et al.* 2003), Chi-squared (Turner 2000) and Fisher's exact (Agresti 1992) are statistical tests employed to identify features with significant differences in statistical patterns between negative and positive datasets. The predictive models are particularly trained and validated from the human epigenome project datasets, DNA methylation of CD4 lymphocytes, and afterward, tested by extracting the DNA methylation data from 11 other normal cell types and tissues. The results have demonstrated that i) the predictive methods simplify well to various cell types and tissues, ii) CGI methylation status could be efficiently determined through eight-dimensional feature space, which integrates all types of information, iii) acetylation and methylation data plays a vital role in prediction, while lack of information deteriorates the model performance, and, iv) models attain more accuracy and specificity by integrating the features of histone acetylation, gene functions and nucleosome positioning (Zheng *et al.* 2013).

6. Conclusion and future prospects

DNA methylation is an important heritable event in the epigenetic marks of the genome associated with developmental events and gene regulation. The methylation status of CGIs delivers functional insights to determine the regulatory roles in transcription because methylated CpGs are exceptionally associated with transcription. The computational prediction of methylation can boost the genome-wide methylation profiling and identify the underlying key features of various methylation patterns. There are numerous computational tools available for CGI detection based on physical distances, physiochemical properties, clustering, sliding windows, and patterns. These computational algorithms work on different principles and deliver high accuracy, sensitivity, or specificity, but only for specific datasets and parameters. The comparative analyses of the tools help to identify the features and limitations thus allow prioritizing the CGIs detection algorithms for target epigenomes. It has been suggested that novel algorithms could be developed by incorporating the previous principles, enhancing traditional algorithms, modifying parameters and creative advancements.

There is also a need for advanced and more efficient computational tools for the rapid and accurate detection of the CGIs in whole genomes. Generally, the capability to computationally predict the CGIs and methylation status offers great promise to boost up the discoveries in epigenetics and gene regulations. The accurate identifications of CGIs from high throughput sequencing data through computational approaches could lead to significant breakthroughs in epigenetic mechanisms and associated neurological disorders.

Acknowledgements

We are thankful to Sheikh Arslan Sehgal, University of Chinese Academy of Sciences, Talal Jamil Qazi and Lucienne N. Duru, Beijing Institute of Technology, Beijing, for their kind support and suggestions throughout the manuscript.

References

Agresti A 1992 A survey of exact inference for contingency tables. *Stat. Sci.* **7** 131–153

Bert SA, Robinson MD, Strbenac D, Statham AL, Song JZ, Hulf T, Sutherland RL, Coolen MW, *et al.* 2013 Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* **23** 9–22

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, *et al.* 2011 High density DNA methylation array with single CpG site resolution. *Genomics* **98** 288–295

Bird A 2002 DNA methylation patterns and epigenetic memory. *Genes Dev.* **16** 6–21

Boukelia A, Benmounah Z, Batouche M, Maati B and Nekkache I 2016 A Novel Algorithm for CpG Island Detection in Human Genome Based on Clustering and Chaotic Particle Swarm Optimization; in *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics* Springer pp 70–81

Britten RJ 1996 DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci.* **93** 9374–9377

Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, Neff NF, Anton E, Medina C, *et al.* 2009 Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* **19** 1044–1056

Chuang L-Y, Huang H-C, Lin M-C and Yang C-H 2011 Particle swarm optimization with reinforcement learning for the prediction of CpG islands in the human genome. *PLoS One* **6** e21036

Chuang L-Y, Yang C-H, Lin M-C, Yang C-H 2012 CpGPAP: CpG island predictor analysis platform. *BMC Genet.* **13** 13

Churchill GA 1989 Stochastic models for heterogeneous DNA sequences. *B. Math. Biol.* **51** 79–94

Consortium EP 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57

de la Rica L, Urquiza JM, Gómez-Cabrero D, Islam AB, López-Bigas N, Tegnér J, Toes RE and Ballestar E 2013 Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. *J. Autoimmun.* **41** 6–16

Deininger PL and Batzer MA 1999 Alu repeats and human disease. *Mol. Genet. Metab.* **67** 183–193

Du Q, Luu P-L, Stirzaker C and Clark SJ 2015 Methyl-CpG-binding domain proteins: Readers of the epigenome. *Epigenomics* **7** 1051–1073

Durbin R, Eddy SR, Krogh A and Mitchison G 1998 Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge

Edgar R, Tan PPC, Portales-Casamar E and Pavlidis P 2014 Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. *Epigenet. Chromatin* **7** 28

Elango N and Soojin VY 2011 Functional relevance of CpG island length for regulation of gene expression. *Genetics* **187** 1077–1083

Feinberg AP and Tycko B 2004 The history of cancer epigenetics. *Nat. Rev. Cancer* **4** 143

Feng P-M, Ding H, Chen W and Lin H 2013a Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013

Feng P-M, Lin H and Chen W 2013b Identification of antioxidants from sequence information using Naive Bayes. *Comput. Math. Methods Med.* 2013

Feng P, Chen W and Lin H 2014 Prediction of CpG island methylation status by integrating DNA physicochemical properties. *Genomics* **104** 229–233

Filion GJ, Zhenilo S, Salozhin S, Yamada D, Prokhortchouk E and Defossez P-A 2006 A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol. Cell. Biol.* **26** 169–181

Gardiner-Garden M and Frommer M 1987 CpG islands in vertebrate genomes. *J. Mol. Biol.* **196** 261–282

Glass JL, Thompson RF, Khulan B, Figueroa ME, Olivier EN, Oakley EJ, Van Zant G, Bouhassira EE, *et al.* 2007 CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.* **35** 6798–6807

Greally JM 2013 DNA Methylation: Bidding the CpG island goodbye. *Elife* **2** e00593

Hackenberg M, Barturen G, Carpena P, Luque-Escamilla PL, Previti C and Oliver JL 2010 Prediction of CpG-island function: CpG clustering vs. sliding-window methods. *BMC Genomics* **11** 327

Hackenberg M, Carpena P, Bernaola-Galván P, Barturen G, Alganza ÁM and Oliver JL 2011 WordCluster: Detecting clusters of DNA words and genomic elements. *Algorithms Mol. Biol.* **6** 2

Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J and Oliver JL 2006 *CpGcluster*: A distance-based algorithm for CpG-island detection. *BMC Bioinform.* **7** 446

Hashimoto H, Vertino PM and Cheng X 2010 Molecular coupling of DNA methylation and histone methylation. *Epigenomics* **2** 657–669

Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, Valsesia A, Ye Z, *et al.* 2012 Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22** 246–258

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, *et al.* 2009 The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41** 178

Jang HS, Shin WJ, Lee JE and Do JT 2017 CpG and non-CpG methylation in epigenetic gene regulation and brain function. *Genes* **8** 148

Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, Zhang X, Chavez L, *et al.* 2014 Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat. Genet.* **46** 17

Jeziorska DM, Murray RJ, De Gobbi M, Gaentzsch R, Garrick D, Ayyub H, Chen T, Li E, *et al.* 2017 DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc. Natl. Acad. Sci.* **114** E7526–E7535

Jones PA 2012 Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13** 484

Jones PA and Baylin SB 2007 The epigenomics of cancer. *Cell* **128** 683–692

Kakumani R, Ahmad O and Devabhaktuni V 2012 Identification of CpG islands in DNA sequences using statistically optimal null filters. *EURASIP J. Bioinform. Syst. Biol.* **2012** 12

Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TY, Bodmer D, *et al.* 2009 Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3′ exons of TACSTD1. *Nat. Genet.* **41** 112

Mayer W, Niveleau A, Walter J, Fundele R and Haaf T 2000 Embryogenesis: Demethylation of the zygotic paternal genome. *Nature* **403** 501

McClelland M and Ivarie R 1982 Asymmetrical distribution of CpG in an 'average'mammalian gene. *Nucleic Acids Res.* **10** 7865–7877

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, *et al.* 2008 Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454** 766

Okano M, Xie S and Li E 1998 Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat. Genet.* **19** 219

Olson SA 2002 Emboss opens up sequence analysis. *Brief. Bioinform.* **3** 87–91

Pastor WA, Aravind L and Rao A 2013 TETonic shift: Biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Bio.* **14** 341

Ponger L, Mouchiroud D 2002 CpGProD: Identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18** 631–633

Rice P, Longden I and Bleasby A 2000 EMBOSS: The European molecular biology open software suite. Elsevier Current Trends

Robertson KD 2005 DNA methylation and human disease. *Nat. Rev. Genet.* **6** 597

Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M and Esteller M 2011 Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6** 692–702

Saxonov S, Berg P and Brutlag DL 2006 A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci.* **103** 1412–1417

Smith ZD and Meissner A 2013 DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.* **14** 204

Stirzaker C, Taberlay PC, Statham AL and Clark SJ 2014 Mining cancer methylomes: prospects and challenges. *Trends Genet.* **30** 75–84

Su J, Zhang Y, Lv J, Liu H, Tang X, Wang F, Qi Y, Feng Y, *et al.* 2009 CpG_MI: A novel approach for identifying functional CpG islands in mammalian genomes. *Nucleic Acids Res.* **38** e6-e6

Sujuan Y, Asaithambi A and Liu Y 2008 CpGIF: An algorithm for the identification of CpG islands. *Bioinformation* **2** 335

Takai D and Jones PA 2002 Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci.* **99** 3740–3745

Timp W and Feinberg AP 2013 Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* **13** 497

Tufarelli C, Stanley JAS, Garrick D, Sharpe JA, Ayyub H, Wood WG and Higgs DR 2003 Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.* **34** 157

Turner N 2000 Chi-squared test. *J. Clin. Nurs.* **9** 10

Wahlberg P, Lundmark A, Nordlund J, Busche S, Raine A, Tandre K, Rönnblom L, Sinnett D, *et al.* 2016 DNA methylome analysis of acute lymphoblastic leukemia cells reveals stochastic de novo DNA methylation in CpG islands. *Epigenomics* **8** 1367–1387

Wang J, Tsang WW and Marsaglia G 2003 Evaluating Kolmogorov's distribution. *J. Stat. Softw.* **8** 1–4

Wu H, Caffo B, Jaffee HA, Irizarry RA and Feinberg AP 2010 Redefining CpG islands using hidden Markov models. *Biostatistics* **11** 499–514

Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, *et al.* 2013 Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153** 1134–1148

Yang C-H, Lin Y-D, Chiang Y-C and Chuang L-Y 2016 A hybrid approach for CpG island detection in the human genome. *PloS ONE* **11** e0144748

Yoon B-J and Vaidyanathan P 2004 Identification of CpG islands using a bank of IIR lowpass filters [DNA sequence detection]; in *Digital Signal Processing Workshop, 2004,* and *the 3rd IEEE Signal Processing Education Workshop* IEEE pp 315–319

Yousef M, Jung S, Kossenkov AV, Showe LC and Showe MK 2007 Naïve Bayes for microRNA target predictions – machine learning for microRNA targets. *Bioinformatics* **23** 2987–2992

Yu N, Guo X, Zelikovsky A and Pan Y 2017 GaussianCpG: A Gaussian model for detection of CpG island in human genome sequences. *BMC Genomics* **18** 392

Zheng H, Wu H, Li J and Jiang S-W 2013 CpGIMethPred: Computational model for predicting methylation status of CpG islands in human genome. *BMC Med. Genomics* **6** S13

Corresponding editor: BJ Rao