Review of CpG Island Recognition Algorithms

Yanning Cai¹*, Qian Dong¹ and Anlan Li²

¹ College of Information Science and Technology, Jinan University, Guangzhou, China
² School of Accountancy, Xin Hua College of Sun Yat-Sen University, Guangzhou, China

*Corresponding author email: caiyanningmitl@sina.com

Abstract. DNA methylation is a common epigenetic process, which exists in many eukaryotes, such as animals and plants. It has the function of regulating gene expression and affecting the stability of genetic material. The unmethylated CpG dinucleotide, known as CGI, generally presents a local aggregation pattern in the genome. Abnormal methylation of CGI is often associated with cancer and abnormal growth. Therefore, it is of great significance to accurately identify CGI and predict its methylation state. This paper summarizes the current mainstream CGI computational recognition methods, analyzes their advantages and disadvantages, and discusses the future direction of CGI recognition algorithm research.

Keywords: CpG island recognition; Biological information; Epigenetics.

1. Introduction

DNA methylation is an important epigenetic process and widely exists in eukaryotes such as animals and plants. Under the action of DNA methyltransferase (DNMT), the methyl group was synthesized on the 5th carbon atom of cytosine in 5'-cpg-3', forming DNA methylation. In recent years, with the development and popularization of genome-wide methylation sequencing technology, the biological significance of DNA methylation gradually revealed. Studies have shown that DNA methylation is widely found in various elements of the genome, including promoters, gene nomenclature, enhancers, silencers and transposons, which can affect the stability of genetic material, gene expression, chromosome activity, transposon silencing, histone modification and other biological processes.

In the human genome, 70% - 80% of CpG dinucleotide is methylated and distributed in all parts of the genome. On the contrary, the demethylated CpG shows a local aggregation pattern, which is called CpG island (CGI)[1]. CGI is not easy to be methylated in the normal biological process, and its abnormal methylated state often affects gene expression, which is related to epigenetic phenomena such as cancer and abnormal plant growth[2][3]. Therefore, it is of great biological significance to recognize CGI in genome and predict its methylation status. In addition, because of the strong correlation between CGI and promoter region, CGI is also considered to be a genomic marker that can effectively predict promoter elements[4].

At present, the methods used to identify CGI are mainly divided into experimental detection method and computational recognition method[5]. The experimental methods include sodium bisulfite (BS SEQ) and immunoprecipitation reaction (medipseq). These methods use different experimental methods to separate methylated and unmethylated cytosine, and then carry out high-throughput sequencing or chip detection on the separated DNA sequence. Its advantage is high accuracy and no ambiguity, but its disadvantage is high cost. In recent years, a variety of CGI computational recognition methods have been proposed. These methods identify CGI in a given DNA sequence by...
establishing a computational model of CGI. CGI methylation pattern analysis has become an important part of bioinformatics. This paper reviews the current mainstream CGI computational recognition methods, focusing on the analysis and comparison of computational models and applicable species.

2. CGI Calculation Identification Method Classification

Although CGI has important biological significance, it is still lack of strict definition. Gardiner garden and Frommer first carried out large-scale computational analysis of CGI based on the sequencing data of vertebrates in 1987[6], and put forward three quantitative conditions of CGI: length greater than 200bp, GC content higher than 50%, ratio of observed CpG to expected CpG (ObsCpG / ExpCpG) is not less than 0.6 (referred to herein as GGF definition). These conditions have a profound impact on the subsequent research of CpG island recognition algorithm. In fact, most of the current mainstream CGI detection algorithms integrate one or more of these three factors.

According to the combination of these factors, CGI detection algorithm can be divided into two categories: rule-based method and statistical feature-based method. The rule-based method filters and discovers CGI regions by using artificial thresholds, which has the advantages of clear knowledge representation and easy understanding. The disadvantage is that it depends on expert knowledge and has limited generalization ability. For example, the recognition algorithm based on a single threshold is difficult to distinguish CGI from ALUs elements in the human genome, because ALUs elements are generally 280bp long, and their GC content and ObsCpG / ExpCpG ratio are also high, which meet the quantitative conditions of CGI; CGI recognition method based on statistical characteristics is widely used in machine learning. The CGI recognition model is constructed by mining the characteristics of CGI fragments different from other DNA sequences, and then the CGI fragments in the new sequences are predicted and recognized by using the model. The feature of this method is its strong generalization ability. According to the types of machine learning models used, CGI recognition methods can be further divided into two categories: supervised learning and unsupervised learning, which are analyzed in turn below.

3. Typical CGI Recognition Method Based on Rules

The rule-based CGI recognition method is also known as sliding window method, which is to scan and filter the DNA sequence according to the pre-set threshold. The DNA fragments that meet the threshold are considered as CGI. These algorithms are represented by GGF, CpGIS, CpGProD, etc[7][8]. After analyzing human chromosome 21 and 22, Takai and Jones put forward stricter standards based on Gardiner garden and Frommer methods[9]: GC ≥ 55%, length ≥ 500bp and ObsCpG / ExpCpG ≥ 0.65. The experimental results show that the new threshold can effectively filter most ALUs and unknown sequences, while the number of CGI at the 5’ end of gene is only slightly reduced.

CpGProD still uses threshold based sliding window method, which is characterized by focusing on recognizing CGI associated with promoter region and TSS in mammalian genome. The advantage of sliding window method is that the rules are intuitionistic and easy to understand. The disadvantage of sliding window method is that the recognition accuracy is easily affected by the threshold value and sliding window size, and it is highly dependent on expert knowledge. Therefore, with the development of sequencing technology and the reduction of sequencing cost, more and more methylated DNA sequence data are available, which makes the prediction method based on statistical analysis and statistical characteristics gradually become the mainstream of CGI recognition algorithm.

4. Typical Unsupervised CGI Identification Method

Unsupervised recognition method is a CGI recognition algorithm based on unsupervised machine learning model. This method usually consists of two steps: ① clustering the adjacent CpG sites to form CpG clusters; ② screening CpG clusters according to GC content, statistical significance, distribution differences and other conditions, and the clusters that meet the conditions are considered as CpG islands.
CpGCluster algorithm is a typical unsupervised CGI recognition algorithm[10]. Based on the hypothesis that CpG distribution in CGI is more compact than that in non CGI, CpGCluster first uses clustering algorithm to aggregate adjacent CpG sites to form CpG clusters, and then selects CGI according to statistical significance. The clustering process is similar to density based clustering: scan the genome from the 5-terminal to the 3-terminal, if the distance between two consecutive CpG loci is less than the threshold, then either merge them into a new CpG cluster, or merge them into an existing CpG cluster.

CpGCluster algorithm has two main parameters: the shortest distance between adjacent CpG sites to form CpG clusters, and the statistically significant value of CpG clusters. Only CpG clusters beyond this threshold can be identified as CpG islands. Since the shortest length of CGI is no longer limited, CpGCluster can often find very short CGI. Therefore, the shortest distance threshold parameters of CpG clusters play a decisive role. Experiments on the human genome show that the median value of the distance between all adjacent CpGs in the genome can be selected as the threshold value, which is also the default setting recommended by the algorithm.

ClusterPSO combines CpGCluster algorithm with Gardiner garden and Frommer (GGF) definitions[11]. Firstly, CpG clusters are generated by CpGCluster algorithm, and then each CpG cluster extends 200 bp to the upstream and downstream respectively, forming a candidate CGI according to the definition of GGF. At the same time, the GGF definition is transformed into the objective function of particle swarm optimization (PSO) algorithm, and the CGI region division which can maximize the objective function value is found by searching.

In addition, Kakumani et al. designed a CGI recognition algorithm based on maximizing signal-to-noise ratio because the probability of connecting nucleotide G after nucleotide C in CGI is often greater than that in non CGI. Gaussian CpG adopted two parameters[12], i.e. Gaussian model and GC content, after forming CpG clusters based on distance clustering Number of CpG islands in CpG clusters[13]. CpG_MI finds CGI by calculating the mutual information of adjacent CpG distances[14]. Therefore, there are two key factors for CGI recognition based on unsupervised learning model: one is the distance calculation method between CpG sites, which determines the morphology of CpG clusters; the other is the CpG clustering screening conditions, which determines the final composition of CGI. On the basis of the rule-based recognition method, this kind of method can integrate more statistical features to describe CGI region, which is more expressive and flexible.

5. Typical Supervised CGI Identification Method
Supervised CGI recognition method refers to the construction of CGI recognition model by using supervised machine learning algorithm. Compared with unsupervised learning, supervised learning needs labeled data, that is, DNA fragments that are known to be CGI and not CGI. The learning algorithm establishes the recognition model based on the annotation data and determines the model parameters. In CGI recognition, there are two main supervised learning algorithms: classification and sequence prediction.

CGI classification refers to the classification of DNA sequences into CGI and non CGI. For example, Bock et al. took CGI in human chromosomes 21 and 22 as training data, and used support vector machine model (SVM) to build a CGI classifier[15]. At the same time, they evaluated many CGI related features, including six aspects: DNA sequence pattern, repetitive distribution pattern, DNA helix structure, transcription factor binding point information, etc. These attributes are then used to predict the epigenetic state of CGI in the whole genome. One of the advantages of using supervised learning model is that it can evaluate and select features, such as combining with typical epigenetic function and genetic status, which can reveal the epigenetic significance and gene expression regulation mechanism of CGI more than the prediction model based on sequence alone.

In recent years, with the improvement of computing performance and the development of big data, deep neural network model is widely used in various machine learning tasks. In CGI recognition, Wang et al. took the interaction between chromosomes as one of the characteristics[16], and based on stack de-noising self encoder, designed a prediction algorithm for the methylation status of CpG sites in human genome, and emphatically verified the effectiveness of the algorithm in the prediction of methylation status of CpG sites in non coding long RNA (lncRNA). The algorithm generates training
data set from methylation sequencing data, and uses one cross validation method to train and verify the effectiveness of the algorithm.

The supervised CGI recognition method based on sequence learning mainly uses sequence learning algorithms such as hidden Markov (HMM) and conditional random field (CRF) to predict DNA sequence. Wu and Spontaneo successively proposed CGI recognition algorithm based on Hidden Markov model; Wu et al. tested on the basis of invertebrate genome and found that the rule-based method was not suitable for invertebrate genome[17][18]; Spontaneo et al. are still testing on the basis of human genome, and simplifies the process of manual CGI inspection by providing a graphical interface[19]; Liu Wei and Chen Ling propose a CGI recognition algorithm based on CRF. Compared with unsupervised learning model, supervised learning method can find more representative CGI features for specific species or tissues, so as to better reveal the methylation pattern and its epigenetic significance. In addition, the first mock exam model is more scalable, for example, it can extend the same model to DNA sequences of different species by using different training datasets.

6. Conclusion
DNA methylation is an important epigenetic process widely existing in all kinds of eukaryotes. The accurate recognition of CpG island and the prediction of its methylation status are helpful for the accurate location of promoters and recognition of genes, as well as for revealing the mechanism of the relationship between methylation and epigenetic phenomena such as cancer and abnormal growth. This paper summarizes the current mainstream CpG island computing and recognition methods, analyzes and compares its mechanism and applicable species. It is found that at present, the research objects of CpG island are mainly vertebrates, and the recognition algorithm of CpG island suitable for other species remains to be studied. Among the three methods, the supervised CGI identification method can better reveal the epigenetic significance of CGI. With the popularization and application of sequencing technology, a large number of methylation sequencing data have been generated, which provides the data basis for the supervised CGI identification model. Therefore, it may also be the main direction of CGI research in the future.

Acknowledgments
This work was supported by the Natural Science Foundation of Guangdong Province (Grants No. 2018030310074), Youth Fund Project of Jinan University research and Innovation Fund (Grants No. 21617350), Youth Foundation of Jinan University (No.21617349) and the PhD Foundation of Guangdong Natural Science (No. 2018030310581).

References
[1] Aimée M Deaton, Bird A. Deaton AM, Bird A. CpG islands and the regulation of transcription[J]. Genes & development, 2011, 25(10):1010-1022.
[2] Maekawa H, Ito T, Orita H, et al. Analysis of the methylation of CpG islands in the CDO1, TAC1 and CHFR genes in pancreatic ductal cancer[J]. Oncology letters, 2020, 19(3).
[3] Kausar S, Yang L, Abbas M N, et al. Mitochondrial DNA: A Key Regulator of Anti-Microbial Innate Immunity[J]. Genes, 2020.
[4] Lioznova A V, Khamis A M, Artemov A V, et al. CpG traffic lights are markers of regulatory regions in human genome[J]. BMC Genomics, 2019, 20(1).
[5] Weber M , Davies J J , Wittig D , et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells[J]. Nature Genetics, 2005, 37(8):853-862.
[6] Gardiner-Garden M, Frommer M J. Cpg Islands in Vertebrate Genomes[J]. Journal of Molecular Biology, 1987, 196(2):261-282.
[7] Limin, Jiang, Chongqing, et al. Light CpG: a multi-view CpG sites detection on single-cell whole genome sequence data[J]. Bmc Genomics, 2019.
[8] Ting W, Jinfu N, Larson N B, et al. CpGtools: a python package for DNA methylation analysis[J]. Bioinformatics, 2019.
[9] TAKAI D, JONES P A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22[J].PNAS,2002,99(6):3740-3745.
[10] Yang C H, Chiang Y C, Chuang L Y, et al. A CpGCluster-Teaching–Learning-Based Optimization for Prediction of CpG Islands in the Human Genome[J]. Journal of Computational Biology, 2017; cmb.2016.0178.
[11] YANG C H,LIN Y D,CHIANG Y C, et al.A hybrid approach for CpG island detection in the human genome[J].PLoS ONE,2016,11(1):e0144748.
[12] KAKUMANI R, AHMAD O, DEVABHAKTUNI V. Identification of CpG islands in DNA sequences using statistically optimal null filters[J].EURASIP Journal on Bioinformatics and Systems Biology,2012(1):12.
[13] YU N, GUO X, ZELIKOVSKY A, et al. GaussianCpG:a Gaussian model for detection of CpG island in human genome sequences[J].BMC Genomics,2017,18(Suppl 4):392.
[14] SU J, ZHANG Y, LV J, et al. CpG_MI:a novel approach for identifying functional CpG islands in mammalian genomes[J].Nucleic Acids Research,2010,38(1):e6.
[15] BOCK C,WALTER J,PAULSEN M,et al.CpGisland mapping by epigenome prediction[J].PLoS Computational Biology,2007, 3(6):e110.
[16] WANG Y,LIU T,SHI H, et al. Predicting DNA methylation state of CpGdinucleotide using genome topological features and deep networks[J].Scientific Reports,2016,6:19598.
[17] IRIZARRY R A,WU H,FIEINBERG A P.A species-generalized probabilistic model-based definition of CpG islands[J].Mammalian Genome:Official Journal of the International Mammalian Genome Society,2009,20(9-10):674-680.
[18] WU H, CAFFO B,JAFFEE H A,et al.Redefining CpG islands using hidden Markov models[J].Biostatistics (Oxford,England), 2010,11(3):499-514.
[19] SPONTANEO L, CERCONE N. Correlating CpG islands, motifs, and sequence variants in human chromosome 21[J].BMC Genomics,2011,12(Suppl 2):S10.