

Alhudaj: CpG islands Detection Tool in Mammalian Genome Using C++

Osamah Alrouwab ^{1*}, Awatef Allafi², Hayat AboSheta², Yasmeeen Tantoush², Dheba Mansour³
and Mahmoud Gargotti ⁴

Department of Biochemistry, Faculty of Medicine, Sabratha University, Sabratha- Libya¹,

Faculty of Biotechnology, ALjafaraUniversity, Alsahla-Libya²,

Department of Zoology, Faculty Sciences, Aljafra University, Almamura -Libya³,

Department of Microbiology, Faculty of Medicine, University of Zawia, Zawia -Libya⁴



Abstract – One of the unique combinations in the mammalian genome, that revolutionized concepts in the fields of genetics and molecular pathology is what is termed the CpG islands. However, the accurate and rapid determination of CpG islands for DNA sequences remains experimentally and computationally challenging. The main goal of this project is to design an offline, cross-platform CpG islands detection tool. The Algorithm implemented in this study was the traditional sliding window algorithm by using the C++ programming language. Three datasets were used for evaluating the performance of the application. The ANK1 gene, SPTB gene, and RET gene sequence files were obtained from NCBI. In this study, the highest CGIs were reported in ANK1 (ankyrin 1) Gene which scored 13 successive islands whereas the lowest score was reported in RET (ret proto-oncogene) Gene which shows only 6 islands. Generally, the program fulfills the boundary limits as expected. We strongly recommend for further work, the implementation of other algorithms in addition to the sliding window algorithm such as the Hidden Markov Model (HMM).

Keywords – ANK1 gene, CpG islands, Hidden Markov Model (HMM), RET gene, sliding window algorithm.

I. INTRODUCTION

Inquisitively, the genome of mammalian species is characterized by the existence of multifarious nucleotides organizations [1]. These unique nucleotides arrangements may take diverse forms, such as palindrome sequence, variable number tandem repeats (VNTRs), and short tandem repeats (STRs). Among these exclusive provisions, the so-called CpG islands (CGIs). The CpG islands, which have been extensively studied and provided noteworthy enlightenment for many mysterious events from both clinical and genetic perspectives [2], [3]. The CpG islands are a nucleotide duo, C and G, which were identified consecutively along a strand of DNA. It is recognized that the occurrence of CpG islands is extremely rare in utmost DNA sequences, because of different biochemical stresses [4]–[6]. However, these islands present in particular segments of the genome, and span a few hundred to a few thousand nucleotides long. The recurrence of CpG dinucleotides, where a cytosine at the 5' position is followed by guanine at the 3' position, is much less than anticipated, which is an outcome of cytosine and guanine quantities, ~20% for each, both equal $0.2 \times 0.2 = 0.4$ [7]–[9]. The potential explanation for this discrepancy in nucleotide ratio is owing to the presence of DNA in the methylated state in the genomes of mammals [10]. Moreover, 5-methylcytosine instantly transforms into thymine, resulting in the decay of the proportion of CpG islands in the whole genome [11]. Interestingly, all CpG Islands (CGIs) are not methylated but are

strictly distinguished by their high G+C content, a minor fraction of CpG Islands were silenced via the deamination [12], [13]. Although after four decades of studies, no fully reliable algebraic algorithm has been developed that can categorize all existing CGI genera in the mammalian genome [14]. Gardiner-Garden and Frommer were the pioneers of this field and successfully built the first theoretical criteria in 1987, their approach was based on a ratio of observed (Obs) to expected (Exp) CpG dinucleotides of more than 0.6, and a GC content of more than 50% with a fixed window length of 200 bp [15], [16]. The weakness of the Gardiner-Garden and Frommer method is the inability to discern the presence of CGIs in certain elements of DNA, which have a high proportion of GC content with a sequence length of 300 b.p. The Alu element in primate genomes is an example of such false-positive predicted CGIs [17], [18]. In addition, some premature CGIs of less than 600 lengths give false positive results. However, it can be established that a significant number of CpG islands are hypermethylated in the printed genes [19]. Certain CpG islands in unprinted districts are now known to be methylated in normal cells, which frequently accompany tissue-specific gene expression models [20]. There is a substantial proportion of methylated CpG islands have also been observed in cancerous cells [19]. The overwhelming majority of the CpG island is located in areas of serious importance like the promoters or start regions of a variety of genes. The CPG islands situated upstream of the transcript starting point are crucial in gene expression regulation and cell differentiation [21]. They are predominantly reported in the 5' region in many mammalian constitutive genes. Perpetually intersecting with or within a thousand bases of the promoter region. The recognition of the promoters of novel genes may be employing CpG islands with a resolution of 2 KB, which seems to be advantageous for small range-scale sequence annotation. Habitually, Many molecular biologists use visual inspection of CpG islands to identify genes [22]. In Biomolecular practice identifying CpG islands is extremely vital. Regardless of the availability of innumerable CpG islands detection tools, all of them were web-based and required to be connected to the internet. Thus, the main objective of this paper is to provide an offline, cross-platform, and reliable CpG detection tool for routine use of Biomolecular practice.

II. MATERIAL AND METHODS

2.1 The datasets

The datasets used for evaluating the performance of the application, have been retrieved from online, publicly accessible databases, the Entrez Gene databases from The National Center for Biotechnology Information (NCBI; <https://www.ncbi.nlm.nih.gov>), during April 12, 2021. Three gene sequence files were used, namely, (a) the SPTB gene (accession ID: **NC_000014.9**), the chromosomal location of this gene 14q23. 3. Spectran proteins, as well as ankyrin, play a role in the organization and stability of the cell membrane. The protein coded by this locus affects the instability of the erythrocyte membranes, and mutations of this gene have been associated with type 2 globulocytosis, hereditary elliptocytosis, and neonatal hemolytic anemia [provided by RefSeq, Nov 2009], (b) the ANK1 gene (accession ID: **NC_000008.11**), the chromosomal location of this gene 8p11.21. This gene encodes ankyrin proteins, which bind integral membrane proteins to the basement actin-spectrin cytoskeleton and perform an essential role in cell proliferation, activation, motility, contact, and the preservation of specialized membrane domains. Several ankyrin isoforms with different affinities for diverse target proteins are expressed in a tissue-specific, developmentally regulated manner [provided by RefSeq, Dec 2008], (c) the RET gene (accession ID: **NC_000010.11**), the chromosomal location of this gene 10q11.21. This locus encodes a transmembrane receptor and a member of the tyrosine-protein kinase family of proteins. Ligand binding such as glial cell-line derived neurotrophic factor (GDNF) and other encoded receptor-related proteins promote receptor dimerization and activation of downstream signaling pathways associated with cell growth, migration, differentiation, and survival [provided by RefSeq, Sep 2017].

2.2 Implementations

The benchmarks have all been achieved on Intel (R) Core (TM) i5-3470 CPU; 3.20GHz and 16 GB DDR3 RAM. The operating system used for the benchmarks was Microsoft Windows 10 64-bit. The application is written in a c++ programming language using Qt5 framework (<https://www.qt.io/product/framework>) and compiled by GCC GNU Compiler (<https://gcc.gnu.org/>).

2.3 Design

The algorithm implemented was the sliding window algorithm as described by Gardiner-Garden and Frommer in 1987 [15]. Each CpG island should be longer than 200 bp, constitute more than 50% G+C, and have a proportion of CpG frequency to the commodity of the C and G frequencies greater than 0.6 (Figure 1). The number of CG combinations (N_{CG}) in each successive 200 bp window should be at least seven.

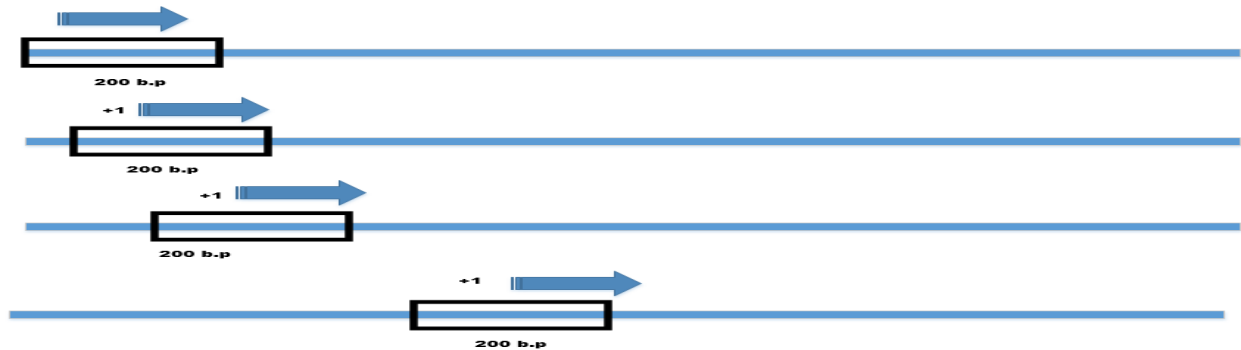


Figure 1: Schematics for the sliding windows algorithm of the CpG Islands.

The search criteria of the algorithm can be abbreviated as follows: (a) setting the windows to 200 bp and evaluating the sequence ($G+C\%$, Obs_{CpG}/Exp_{CpG} , and, N_{CG}) of a CpG island, (b) if the window meets the criteria, print the results and move the window bp at a time, (c) repeat steps a and b until end of the sequence (Figure 1 and 2).

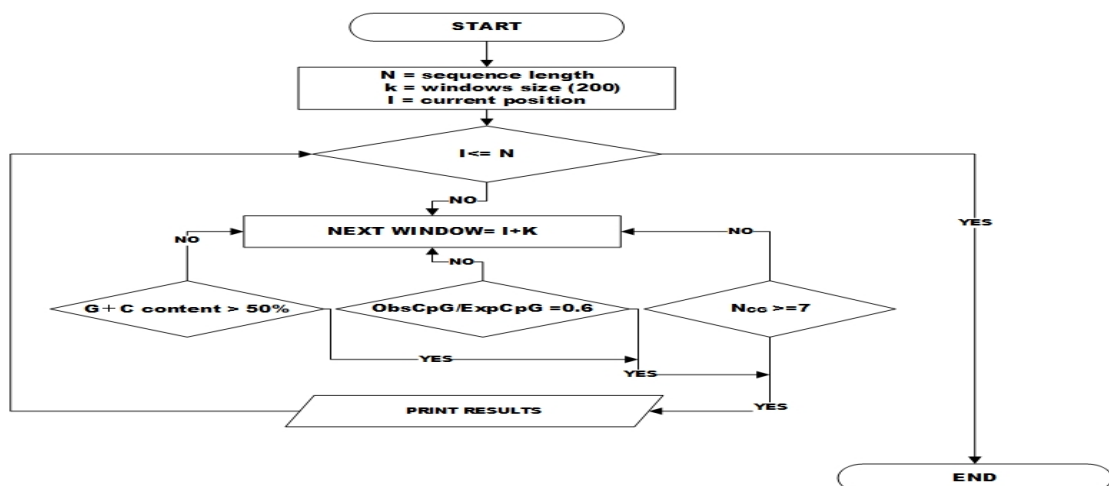


Figure 2: Flowchart for the sliding windows algorithm.

III. RESULTS

The touchstone by which CpG islands were identified was the slide window algorithm as described by Gardiner-Garden and Frommer [15]. All benchmarks were exposed to unified parameters of 200 b.p window size, $Obs_{CpG}/Exp_{CpG} \geq 0.6$, $N_{GC} \geq 7$, and $G+C \geq 50\%$. The CpG islands resulted from analysis of the three selected genes organized to their genomic location into (a) 5' region of a gene CpG island, (b) islands situated in exotic province, (c) unknown scattered islands including the *Alu* repeats.

3.1 CpG Islands in Homo sapiens SPTB (spectrin beta) Gene

The SPTB gene (Gene ID: 6710) on Chromosome 14 encloses 33 exonic regions (Figure 3). We successively extracted 11 CpG islands from various regions, namely, (a) one island on 5` UTR regions that extends 598 b.p upstream to promoter site, (b) three islands harbored at exons 1, 3, and 33 respectively. (c) the rest of the islands designated unknown (Table 1).

Table 1 Organization of CpG islands in SBTP gene

Island tier	Number of CpG islands
5` regions	1
Exon	3
Unknown	7
Total	11

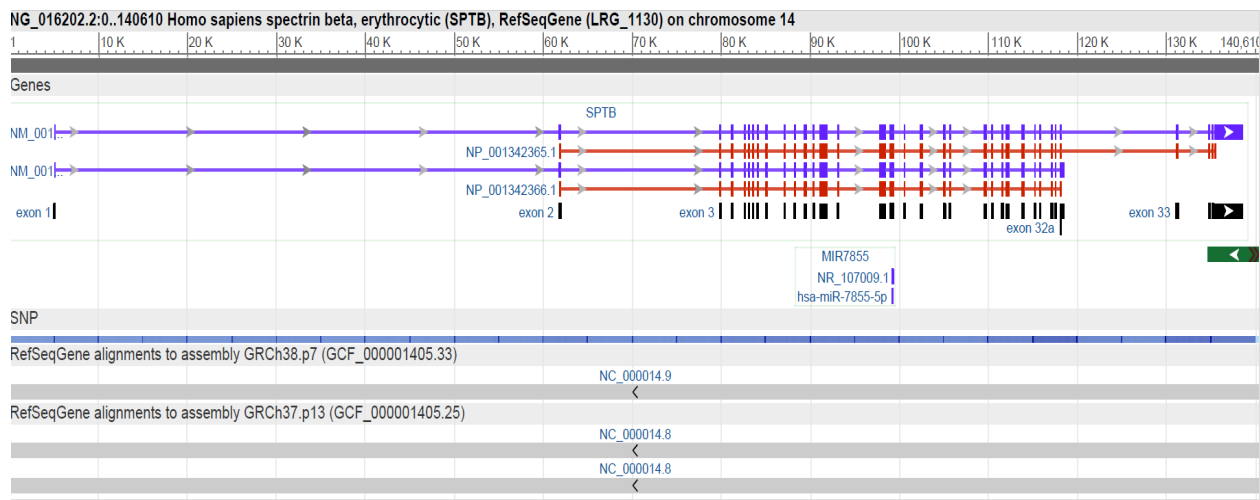


Figure 3: SPTB gene Genomic regions, transcripts, and products.

3.2 CpG Islands in Man ANK1 (ankyrin 1) Gene

Essentially, the ANK1 gene (Gene ID: 286) on Chromosome 8 enfolds 39 exon provinces (Figure 4). The number of CpG islands mined were 13 islands, one island at 5` UTR region and on exons number 4,5,6,7,15,16,25 and 31 correspondingly. Only 4 islands have reported unknown (Table 2).

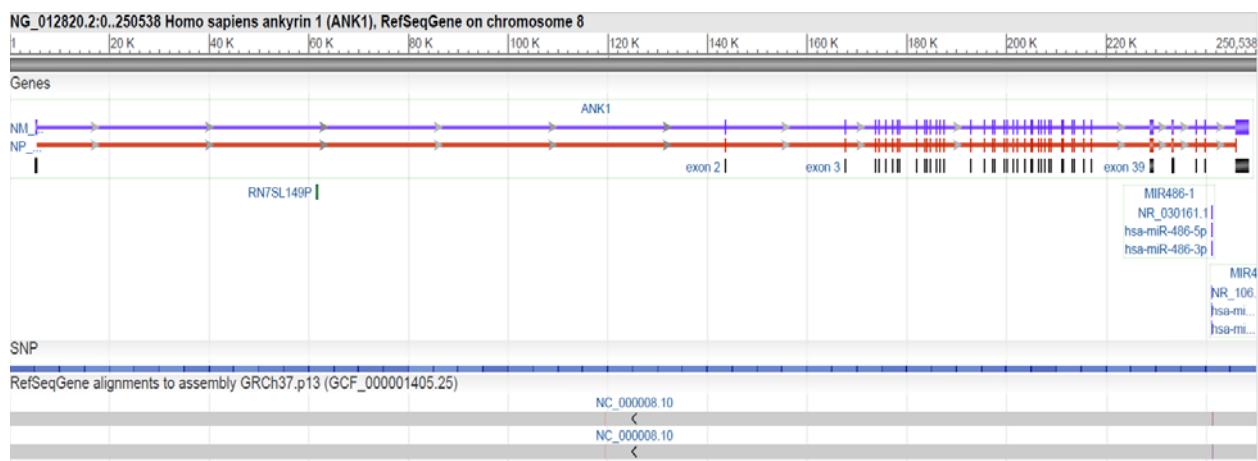


Figure 4: ANK1 gene Genomic regions, transcripts, and products.

Table 2 Organization of CpG islands in ANK1 gene

Island tier	Number of CpG islands
5' regions	1
Exon	8
Unknown	4
Total	13

3.4 CpG Islands in Human RET (ret proto-oncogene) Gene

Principally, the RET gene (Gene ID: 5979) on Chromosome 10 encloses 19 exon sticks (Figure 5). The number of CpG islands excavated were 6 islands, one island at 5' UTR region, and the rest of them considered unknown. No islands have been reported in the exonic region (Table 3).

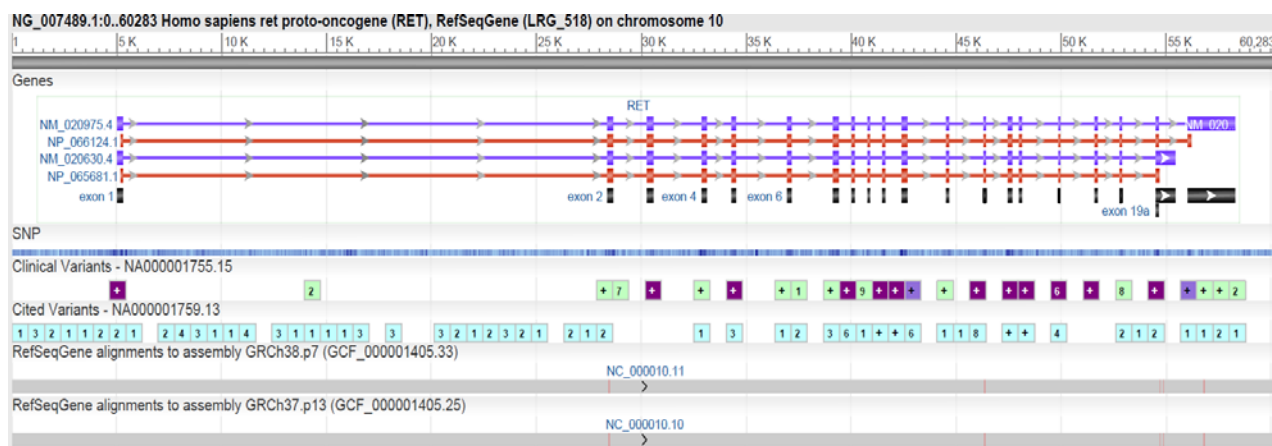


Figure 5: RET gene Genomic regions, transcripts, and products.

IV. DISCUSSION

This paper introduces Alhudaj, across platform offline software for detecting mammalian CpG islands, especially those dwelling the 5'-regions or more precisely the promotor province of constitutive genes. There is indeed a substantial need for the exact mapping of DNA CpG islands to realize the broad biochemical processes. However, accurate determination of CpG islands from the entire genome using computational and experimental methodologies remains challenging. In 1987, Gardiner-Garden and Frommer [15], introduced the first in silico window-based CGIs detection approach. In silico techniques for CGI, identification is primarily classified into four classes depending on their principal algorithms, which are window-based, Hidden Markov Model (HMM) based, density-based, and distance-/length-based methods used in the computational CGI detection. [23], [24].

Kim, Ki-Bong (2010) developed a Window-based CpG Island Search tool implemented in Java and designed for use on any system. The window-based graphical user interface of CpG Islands Detector allows the end-user to more easily utilize this program to locate CpG islands in a genomic DNA sequence. Furthermore, considering CpG islands are frequently present in the 5' regions of vertebrate genes, this approach may be used to identify putative genes in genomic sequences[25]. Chuang, Li-Yeh, et al. (2012) proposed a novel procedure to locate CpG islands by merging clustering technology with the sliding-window approach (PSO-based). Clustering technology is employed to obtain and process the locations of all potential CpG islands, thereby eliminating the need for lengthy and unneeded DNA fragment processing and boosting the effectiveness of sliding window-based particle swarm optimization (PSO) search. ClusterPSO, the suggested method, offers scalable sensitive identification of CpG islands in the human genome. Furthermore, ClusterPSO's detection effectiveness in the human genome is compared to eight CpG island detection techniques. ClusterPSO surpassed all other test techniques in terms of detection efficiency for CpG islands in the human genome, including specificity, sensitivity, accuracy, performance coefficient (PC), and correlation coefficient (CC). Furthermore, the integration of clustering technology and the PSO approach may effectively overcome their disadvantages while retaining their benefits. As a result, clustering technology may be combined with the optimization algorithm approach to improve CpG island discovery [26]. Fan, Zhenxin, et al. (2017) develop a new CGIs search tool, namely CpGIScan (CpG Islands Scan). A CpG island is characterized in this study by three types of parameters: the window length, the frequency of guanine and cytosine (G + C), and the ratio of observed to predicted CpGs. The algorithm in CpGIScan is based on the sliding window method. To reduce the time required to identify CGIs, multithread technology is employed in their program. CpGIScan was compared to existing widely used tools to benchmark its performance [24]. The outcome of our program and online web-based CpG island detection tools the Sequence Manipulation Suite (https://www.bioinformatics.org/sms2/cpg_islands.html) was identical, with only one unworthy difference in calculation of *ObsCpG / ExpCpG*.

V. CHALLENGES

Categorizing CGI according to their type (exonic, 5' UTR, or Alus) was challenging, due to the limitation of the algorithm implemented in this paper. However, we put the focus on a particular task, the detection of CGI. The deamination of CGI type was done visually with the assistance of information provided by the NCBI genomic database.

VI. CONCLUSIONS

In this study, the highest CGIs were reported in ANK1 (ankyrin 1) Gene which scored 13 successive islands whereas the lowest score was reported in RET (ret proto-oncogene) Gene which shows only 6 islands. Generally, the program fulfills the boundary limits as expected. We strongly recommend further work, implement other algorithms in addition to the sliding window algorithm.

References

- [1] P. Takacs, "Philosophical, Historical, and Empirical Investigations into the Concept of Biological Fitness." The Florida State University, 2017.
- [2] E. Balzano, F. Pelliccia, and S. Giunta, "Genome (in) stability at tandem repeats," 2020.
- [3] I. M. Bonapace and F. Macchi, "UHRF1 coordinates DNA methylation and histone post-translational modifications in colon cancer."

- [4] N. Aluru, S. I. Karchner, K. S. Krick, W. Zhu, and J. Liu, "Role of DNA methylation in altered gene expression patterns in adult zebrafish (*Danio rerio*) exposed to 3, 3', 4, 4', 5-pentachlorobiphenyl (PCB 126)," *Environ. epigenetics*, vol. 4, no. 1, p. dvy005, 2018.
- [5] M. L. Brandi, S. K. Agarwal, N. D. Perrier, K. E. Lines, G. D. Valk, and R. V Thakker, "Multiple endocrine neoplasia type 1: latest insights," *Endocr. Rev.*, vol. 42, no. 2, pp. 133–170, 2021.
- [6] V. V Khrustalev, T. A. Khrustaleva, N. Sharma, and R. Giri, "Mutational pressure in Zika virus: local ADAR-editing areas associated with pauses in translation and replication," *Front. Cell. Infect. Microbiol.*, vol. 7, p. 44, 2017.
- [7] A. Unnikrishnan, W. M. Freeman, J. Jackson, J. D. Wren, H. Porter, and A. Richardson, "The role of DNA methylation in epigenetics of aging," *Pharmacol. Ther.*, vol. 195, pp. 172–185, 2019.
- [8] S. Campuzano, M. Pedrero, P. Yáñez-Sedeño, and J. M. Pingarrón, "Advances in electrochemical (bio) sensing targeting epigenetic modifications of nucleic acids," *Electroanalysis*, vol. 31, no. 10, pp. 1816–1832, 2019.
- [9] J. E. DeNizio, B. J. Dow, J. C. Serrano, U. Ghanty, A. C. Drohat, and R. M. Kohli, "TET-TDG Active DNA Demethylation at CpG and Non-CpG Sites," *J. Mol. Biol.*, vol. 433, no. 8, p. 166877, 2021.
- [10] D. S. Dunican, H. K. Mjoseng, L. Duthie, I. M. Flyamer, W. A. Bickmore, and R. R. Meehan, "Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells," *BMC Biol.*, vol. 18, no. 1, pp. 1–21, 2020.
- [11] L. Alsøe et al., "Uracil accumulation and mutagenesis dominated by cytosine deamination in CpG dinucleotides in mice lacking UNG and SMUG1," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, 2017.
- [12] S. Branciamore, Z.-X. Chen, A. D. Riggs, and S. N. Rodin, "CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of HOX and other transcription factors," *Proc. Natl. Acad. Sci.*, vol. 107, no. 35, pp. 15485–15490, 2010.
- [13] V. Vymetalkova, P. Vodicka, S. Vodenkova, S. Alonso, and R. Schneider-Stock, "DNA methylation and chromatin modifiers in colorectal cancer," *Mol. Aspects Med.*, vol. 69, pp. 73–92, 2019.
- [14] R. A. Tahir, D. A. Zheng, A. Nazir, and H. Qing, "A review of computational algorithms for CpG islands detection," *J. Biosci.*, vol. 44, no. 6, pp. 1–11, 2019.
- [15] M. Gardiner-Garden and M. Frommer, "CpG islands in vertebrate genomes," *J. Mol. Biol.*, vol. 196, no. 2, pp. 261–282, 1987.
- [16] K. Kowal, A. Tkaczyk, T. Ząbek, M. Pierzchała, and B. Ślaska, "Comparative analysis of CpG sites and islands distributed in mitochondrial DNA of model organisms," *Animals*, vol. 10, no. 4, p. 665, 2020.
- [17] D. Yalcin, "Sequential, Spatial and Functional Disposition of CpG Islands." The University of Nebraska-Lincoln, 2020.
- [18] M. J. Pajares, C. Palanca-Ballester, R. Urtasun, E. Alemany-Cosme, A. Lahoz, and J. Sandoval, "Methods for analysis of specific DNA methylation status," *Methods*, vol. 187, pp. 3–12, 2021.
- [19] P. A. Jones and S. B. Baylin, "The fundamental role of epigenetic events in cancer," *Nat. Rev. Genet.*, vol. 3, no. 6, pp. 415–428, 2002.
- [20] W. S. Park et al., "Hypermethylation of the RUNX3 gene in hepatocellular carcinoma," *Exp. Mol. Med.*, vol. 37, no. 4, pp. 276–281, 2005.

- [21] A. Bird, "DNA methylation patterns and epigenetic memory," *Genes Dev.*, vol. 16, no. 1, pp. 6–21, 2002.
- [22] E. Ratti and F. Boem, "Junk or Functional DNA? ENCODE and the Function Controversy."
- [23] Y. Yu, T. Jia, and X. Chen, "The 'how' and 'where' of plant micro RNA s," *New Phytol.*, vol. 216, no. 4, pp. 1002–1017, 2017.
- [24] Z. Fan, B. Yue, X. Zhang, L. Du, and Z. Jian, "CpGIScan: an ultrafast tool for CpG islands identification from genome sequence," *Curr. Bioinform.*, vol. 12, no. 2, pp. 181–184, 2017.
- [25] K.-B. Kim, "CpG islands detector: a window-based CpG island search tool," *Genomics Inform.*, vol. 8, no. 1, pp. 58–61, 2010.
- [26] L.-Y. Chuang, C.-H. Yang, M.-C. Lin, and C.-H. Yang, "CpGPAP: CpG island predictor analysis platform," *BMC Genet.*, vol. 13, no. 1, pp. 1–9, 2012.