International Journal of Therapeutic Applications
ISSN 2320-138X

CPG COUNTER: AN *IN SILICO* APPROACH TO DETERMINE THE POSSIBLE DNA METHYLATION SITES IN HUMAN GENES

S.K. Maity and N. Roy Chattopadhyay*

Department of Biotechnology, Haldia Institute of Technology, I.C.A.R.E. Complex, H.I.T. Campus, Hatiberia, P.O: HIT, Dist: Midnapore (E), West Bengal, India. PIN: 721657.

ABSTRACT

The studies involving regulations of gene expressions need rapid and precise analyses of millions of nucleotide sequences. In silico approaches are fast and accurate, and these save both time and money as these can screen nucleotide sequences to be used in costly molecular biology techniques. Mammalian gene silencing by DNA methylation in 5´CpG3´ dinucleotides play a very crucial role by affecting the remodeling of the underlying chromatin. Therefore the presence of these dinucleotides is studied widely in aberrant expressions of gene(s) seen in abnormal development and in various diseases including cancers. Statistical analyses need exact counts of these nucleotides, which counts may affect the silencing processes. The present report offers a simple in silico approach to determine the counts of these dinucleotides with great precisions and rapid performances involving a countless of nucleotide sequences, without complicated web-based softwares. The modifications of the approach may be used for other sequential patterns also.

Keywords: CpG count, in silico, DNA methylation.

INTRODUCTION

Regulation of gene expression is of immense importance in all living organisms.[1] In metazoans, three basic mechanisms are found to regulate the expressions of genes which ultimately decide for the fate of a cell.[1,2] Any aberration in this regulation may lead to abnormal growth and/or reproduction and/or may give birth to disease(s).[1-6] Mutations can change the genetics of a gene to produce altered expression of that target gene while epigenetic modifications do not affect the nucleotide sequence of the target gene, still producing altered expression of that.[6] The third mechanism targets the production of protein from mRNA i.e. the translation step of the central dogma specified by Crick.[7,8] Micro RNAs, or miRs, can form imperfect matches with the 3'UTRs of mRNAs to stop protein production. This endogenous post-transcriptional regulation occurs in the cytoplasm by the mature miRs which are small non-coding RNAs.[7,9]

Most of the time, mutations are repaired inside the cell where these have occurred.[10] It is reported widely that epigenetic modifications are involved in the regulation of repair genes and also in the genes of miRs.[11-16] Thus, epigenetic modifications gained a lot of interests from the scientists all over the world. Various abnormal developments and diseases are attributed to the aberrant expressions of genes caused by one or more type(s) of epigenetic modifications. In humans, diseases like cancer and hereditary diseases are studied widely in search of these types of regulations.[17-19] Histone modifications and DNA methylations are the key types of epigenetic regulations seen to play the most important role in humans. Methylations, phosphorylations, acetylations, and ubiquitylations of histones affect chromatin remodeling and thus, the expression of a gene present in the region with modified histones. DNA methylations, more specifically the methylations of 5´CpG3´ sites in a DNA, are related to histone modifications and also regulate chromosome remodeling.[20-25] It is not known conclusively whether CpG methylation occurs first and then induce histone modification(s) or histones play the primary role for epigenetic regulations, or these two are not dependent n each other at all.[26-31]

DNA methylations in 5´CpG3´ sites have gained much interest due to a few reasons. First, histones are present throughout all the genetic material of a cell and they are arranged in a specific regular manner. The universal presence of histones along the DNA molecules offers less variation to regulate genes variably.[21-22,24,25] Secondly, it is reported that

***Corresponding Author :**
mailnabanita@gmail.com

Table 1: A few genes tested with the *CpG COUNTER* program.

Sl. No.	Gene Symbol	Accession No.	Gene Size	CpG found manually	CpG found in CpG COUNTER
1	**EGFR**	OTTHUMG00000023661	193808	2574	2620
2	**TCL6**	OTTHUMG00000149979	30536	332	334
3	**FGFR1**	OTTHUMG00000147366	58704	1159	1174
4	**BCL2**	OTTHUMG00000132791	197983	2294	2329
5	**BRAF**	OTTHUMG00000157457	200822	1674	1702
6	**MYC**	OTTHUMG00000128475	7195	333	336
7	**MDM2**	OTTHUMG00000142827	38459	515	524
8	**WNT1**	OTTHUMG00000170403	4262	323	327
9	**ELK 1**	OTTHUMG00000021452	16284	296	302
10	**JUN**	OTTHUMG00000008376	4740	272	276
11	**ABL1**	OTTHUMG00000020813	174930	2714	2764
12	**BAX**	OTTHUMG00000160476	8184	244	249
13	**CCND1**	OTTHUMG00000167877	14588	663	673
14	**PIM1**	OTTHUMG00000016426	6424	269	272

CpG methylations may affect histone modifications to alter gene expressions.[20,23,28,29] The third reason is, the presence of 5´CpG3´ sites in DNA molecules create patterns of sequences made of these dinucleotides and thus, may make a pattern having more such dinucleotides more prone to methylation while another patterns are less prone to the same, as they contain less of these nucleotides. It should be noted that the sequence of these dinucleotides is important for methylation.[32] Though it is not clear whether the exact number of these dinucleotides play any deterministic role for the methylations of cytosine residues in the underlying DNA, it may be inferred that the chance of methylation increases as the number of 5´CpG3´ sites increase.

In silico analyses of nucleotide sequences save time and money for the researchers in the fields of biology, particularly in molecular biology which needs to deal with the nucleotide sequences in almost all experiments. Analyses of CpG methylations are necessary for the studies involving gene expressions and their mechanisms. These are also very important for the studies involving cancers and the aberrant expressions of various genes in these diseases.[33-38] The present report offers a method to determine the exact number of the 5´CpG3´ dinucleotides in any DNA sequence got from any source. Using the similar programs, the other possible dinucleotides can also be found in any DNA region. By changing the number and sequence of the nucleotides, the program can be modified to determine the repeat elements and/or any particular sequence(s) in a DNA or RNA. This method will quickly produce results that may be used for statistical analyses and may generate data which can be used for wet lab-based molecular biology techniques.

MATERIALS AND METHODS

Software supports

The study needs the software support of TURBO C compiler ver. 3.0and above.

Sequences taken during the study

For checking the proper functioning of the program 14 genes were taken (Table 1). Full sequences of these genes are downloaded from the HGNC database (http://www.genenames.org/) and the *vega* formats are used in the study (http://vega.sanger.ac.uk/Homo_sapiens/). The numbers of CpG dinucleotides are calculated both by word (by pressing Ctrl F and then highlighting the searched dinucleotides) and by the program.

RESULTS AND DISCUSSIONS

The program: CpG COUNTER

The following program is to be run to calculate the number of 5´CpG3´ dinucleotides in any nucleotide sequence. Each line of the program is assigned a reference number (e.g. Line no.1, Line no.2, etc.) so that any modification(s) in that line may be referred later on.

```
Line no.1       #include<iostream.h>
Line no.2       #include<fstream.h>
Line no.3       #include<conio.h>
Line no.4       #include<stdlib.h>
Line no.5       void main ()
Line no.6       {
Line no.7       clrscr();
Line no.8       ifstream fin;
Line no.9       char ch,ch2,choice;
Line no.10      X:
Line no.11      cout<<”ENTER SEQUENCE THEN
    PRESS ‘Y’\t”;
Line no.12      cin>>choice ;
Line no.13      if(choice==’Y’)
Line no.14      {
Line no.15      fin.open(“gene.txt”);
    // “gene,txt” is a text file in the same directory
    as ‘.obj’,&’.bak’ files.
Line no.16      int c=0,i,j ;
Line no.17      char k[10000];
Line no.18      for(j=0;j<250; j++)
Line no.19      {
Line no.20      for(i=0;i<10000; i++)
Line no.21      {
Line no.22      L:
Line no.23      fin.get(ch);
Line no.24      if((ch==’ ‘) || (ch==’\n’))
Line no.25      goto L;
Line no.26      k[i]=ch;
Line no.27      }
Line no.28      for(i=0; k[i]!=’\0’; i++)
Line no.29      {
Line no.30      if((k[i]==’C’)&&(k[i+1]==’G’))
Line no.31      c++;
Line no.32      }
Line no.33      }
Line no.34      cout<<”\n\nTotal number of CpG
    :\t”<<c;
Line no.35      fin.close();
Line no.36      cout<<”\n\t\tMORE?(Y/N)”;
Line no.37      Y:
Line no.38      cin>>ch2;
Line no.39      if(ch2==’N’)
Line no.40      exit(0);
Line no.41      else if(ch2==’Y’)
Line no.42      goto X;
Line no.43      else
Line no.44      cout<<”\n\tWRONG
    CHOICE(Y/N)\t”;
Line no.45      goto Y;
Line no.46      }
Line no.47      else
Line no.48      cout<<”\n\tEXIT?(Y/N)\t”;
Line no.49      Z:
Line no.50      cin>>ch2;
Line no.51      if(ch2==’Y’)
Line no.52      exit(0);
Line no.53      else if(ch2==’N’)
Line no.54      goto X;
Line no.55      else
Line no.56      cout<<”WRONG CHOICE(Y/N)” ;
Line no.57      goto Z;
Line no.58      }
```

CpG counts using the CpG COUNTER

The CpG dinucleotide counts got from the manual effort in word file (by Ctrl F) and those got using the CpG COUNTER program written above show differences in their numbers for all the genes tested (Table 1). DNA methylation studies in humans are very important for both the development and diseases. As discussed earlier, the CpG dinucleotide sites play a crucial role for the silencing of genes, and any aberration(s) may generate disease(s) even like a cancer. The analyses of these dinucleotides are required for the expressions of genes and in silico approach saves both time and money. Therefore the exact counts for these dinucleotides are indeed needed and are very useful for these purposes, and also for statistical analyses. The present report offers a simple program to determine these dinucleotides in any nucleotide sequence which can be used for the gene expression analyses.

The program may be modified to get results for other sequence patterns (Table 2; Supplementary material). For example, line no.30 may be modified as if((k[i]==’C’)&&(k[i+1]==’A’)) to get the number of 5´CpA3´ dinucleotides. All the varieties of dinucleotides may be calculated in such a way. The same line (line no.30) may be modified as if((k[i]==’A’)&&(k[i+1]==’T’)&&(k[i+2]==G)&&(k[i+3]==’C’)) to get the number of 5´ATGC3´ repeats. All the varieties of repeats with two or more nucleotides may be calculated in such a way. So the same line (line no.30) may be modified in such a way that the start codons and stop codons can also be found. A modification of the basic program can also be used to highlight the regions with a particular sequence(s).

The in silico analyses of biological sequences (both protein and nucleotide sequences) have gained its importance in last decades. Though various on-line softwares and databases are available for these purposes,[33-38] and most of them are free for the users, they have their specific requirements if they are to be used properly. The present report offers a very simple program written in the language C which itself is simple and useful. Any researcher can use this program and its modifications for his/her purpose(s) and will not require specific formats or internet, at least for the

Table 2: A few modifications of the basic program.

Line no.	Modification	Result(s)
Line no. 30	if((k[i]==’C’)&&(k[i+1]==’A’))	Number of 5´CpA3´ dinucleotides
Line no. 30	if((k[i]==’A’)&&(k[i+1]==’T’))	Number of 5´ApT3´ dinucleotides
Line no. 30	if((k[i]==’G’)&&(k[i+1]==’A’))	Number of 5´GpA3´ dinucleotides
Line no. 30	if((k[i]==’T’)&&(k[i+1]==’T’))	Number of possible thymine dimers
Line no. 30	if((k[i]==’A’)&&(k[i+1]==’U’)&&(k[i+2]==’G’))	Number of start codons in mRNA
Line no. 30	if((k[i]==’U’)&&(k[i+1]==’A’)&&(k[i+2]==’G’))	Number of stop codon ‘UAG’ in mRNA
Line no. 30	If(((k[i]==’U’)&&(k[i+1]==’A’)&&(k[i+2]==’G’))|| ((k[i]==’U’)&&(k[i+1]==’A’)&&(k[i+2]==’G’))|| ((k[i]==’U’)&&(k[i+1]==’A’)&&(k[i+2]==’G’)))	Number of stop codons (all three) in mRNA
Line no. 30	if((k[i]==’any nucleotide’)&&(k[i+1]==’ any nucleotide ’)&&(k[i+2]== ’any nucleotide‘))	Number of any three nucleotide repeats
Line no. 30	if((k[i]==’ any nucleotide’)&&(k[i+1]==’ any nucleotide’)&&(k[i+2]==’ any nucleotide‘)&&(k[i+3]==’ any nucleotide’)&&(k[i+4]==’ any nucleotide’)…)	Number of any repeat with any length
Line no.17	char k[any number may be written in digits];	Nucleotide sequence of any length may be analyzed
Line no.18	for(j=0;j< any number may be written in digits; j++)	
Line no.20	Should be changed accordingly with Line no.17.	

analysis of the sequences in relation to number and/or specific pattern(s).

CONCLUSION

The present provides a very simple program to determine the exact numbers of CpG dinucleotides which play the major role in mammalian gene silencing by DNA methylation. The error-free results and the repeated uses of the program in a single window will help the study of many genes with great precisions and fast performances. Complicated on-line softwares and web-based practices may sometimes limit their uses and the user can not modify the service as per his/her requirement(s). The modifications of this program may be used for finding other sequences also and may be utilized as desired by the user for his/her specific purpose(s).

REFERENCES

1. Gilbert S F, *Developmental Biology* (8th ed.), (Sinauer Associates, Sunderland, USA), 2006.
2. Rudel D & Sommer R J, The evolution of developmental mechanisms, *Developmental Biology.*, 264 (2003) 15.
3. Bruneau B G, The developmental genetics of congenital heart disease, *Nature*, 451 (2008) 943.
4. Chazenbalk G, Chen Y_H, Heneidi S, Lee J_M, Pall M, Chen Y_D, Azziz R, Abnormal expression of genes involved in inflammation, lipid metabolism, and Wnt signaling in the adipose tissue of polycystic ovary syndrome, *J Clin Endocrinol Metab*., 97 (2012) E765.
5. Joenje H, & Patel K J, The emerging genetic and molecular basis of Fanconi anaemia, *Nat Rev Genet.*, 2 (2001) 446.
6. Sadikovic B, Al-Romaih K, Squire J A, & Zielenska M, Cause and Consequences of Genetic and Epigenetic Alterations in Human Cancer, *Current Genomics.*, 9 (2008) 394.
7. Visone R & Croce C M, MiRNAs and Cancer, *The American Journal of Pathology*, 174 (2009) 1131.
8. Crick F H C, On Protein Synthesis, *Periodical of The Symposia of the Society for Experimental Biology 12*, (1958) 138.
9. Filipowicz W, Bhattacharyya S N, & Sonenberg N, Mechanisms of posttranscriptional regulation by microRNAs: are the answers in sight?, *Nat Rev Genet*., 9 (2008) 102.
10. Wood R D, Mitchell M, Sgouros J, Lindahl T, Human DNA Repair Genes, *Science*, 291(2001) 1284.
11. Wheeler J M D, Epigenetics, mismatch repair genes and colorectal cancer, *Ann R Coll Surg Engl.*, 87(2005) 15.

12. Swisher E M, Gonzalez R M, Taniguchi T, Garcia R L, Walsh T, Goff B A & Welcsh P, Methylation and protein expression of DNA repair genes: association with chemotherapy exposure and survival in sporadic ovarian and peritoneal carcinomas, *Molecular Cancer*, 8 (2009) 48.
13. Cuozzo C, Porcellini A, Angrisano T, Morano A, Lee B, Pardo A D, Messina S, Iuliano R, Fusco A, Santillo M R, Muller M T, Chiariotti L, Gottesman M E & Avvedimento E V, DNA Damage, Homology-Directed Repair, and DNA Methylation, *PLoS Genetics*, 3 (2007) 1144.
14. Lujambio A, Calin G A, Villanueva A, Ropero S, Sánchez-Céspedes M, Blanco D, Montuenga L M, Rossi S, Nicoloso M S, Faller W J, Gallagher W M, Eccles S A, Croce C M, & Esteller M, A microRNA DNA methylation signature for human cancer metastasis, *Proc Natl Acad Sci*., 105 (2008) 13556.
15. Lujambio A & Esteller M, How epigenetics can explain human metastasis: A new role for microRNAs, *Cell Cycle*, 8 (2009) 377.
16. Zhang H, Li Y & Lai M, The microRNA network and tumor metastasis, *Oncogene*, 29 (2009) 937.
17. Van der Maarel S M, Epigenetic Mechanisms in Health and Disease, *Ann Rheum Dis*, Suppl III, 67 (2008), iii97.
18. Hirst M & Marra M A, Epigenetics and human disease, *Int. J Biochem. Cell Biol*., 41 (2009) 136.
19. Robertson K D, DNA methylation and human disease, *Nat. Rev. Genet.*, 6 (2005) 597.
20. Kondo Y, Epigenetic Cross-Talk between DNA Methylation and Histone Modifications in Human Cancers, *Yonsei Med Journal*, 50 (2009) 455.
21. Rice J C & Allis C D, Histone methylation versus histone acetylation: new insights into epigenetic regulation, *Current Opinion in Cell Biology*, 13 (2001) 263.
22. Crepaldi L & Riccio A, Chromatin learns to behave, *Epigenetics*, 4 (2009) 23.
23. Wang J, Hevi S, Kurash J K, Lei H, Gay F, Bajko J, Su H, Sun W, Chang H, Xu G, Gaudet F, Li E & Chen T, The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation, *Nat Genet.*, 41 (2009) 125.
24. Jenuwein T & Allis C D, Translating the histone code, *Science*, 293 (2001) 1074.
25. Roth S Y & Allis C D, Histone acetylation and chromatin assembly: a single escort, multiple dances?, *Cell*, 87 (1996) 5.
26. Mikkelsen T S, Ku M, Jaffe D B, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T K, Koche R P, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander E S & Bernstein B E, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature*, 448 (2007) 553.
27. Weber M, Hellmann I, Stadler M B, Ramos L, Pääbo S, Rebhan M & Schübeler D, Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome, *Nat Genet*., 39 (2007) 457.
28. Hotz H R & Peters A H, Protein demethylation required for DNA methylation, *Nat Genet*., 41 (2009) 10.
29. Ooi S K T, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin S P, Allis C D, Xiaodong Cheng X & Bestor T H, DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA, *Nature*, 448 (2007) 714.
30. Tachibana M, Matsumura Y, Fukuda M, Kimura H & Shinkai Y, G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription, *EMBO J*., 27 (2008) 2681.
31. Kondo Y, Shen L, Cheng A S, Ahmed S, Boumber Y, Charo C, Yamochi T, Urano T, Furukawa K, Kwabi-Addo B, Gold D L, Sekido Y, Huang T H & Issa J P, Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation, *Nat Genet*., 40 (2008) 741.
32. Esteller M, Cancer epigenomics: DNA methylomes and histone-modification maps, *Nat Rev Genet.*, 8 (2007) 286.
33. Ashburner, M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat.Genet.*, 25, 25-29.
34. Grunau, C. *et al.* (2001) MethDB – a public database for DNA methylation data. *Nucl. Acids Res.*, **29**, 270–274.
35. Amoreira, C. *et al.* (2003) An improved version of the DNA methylation database (MethDB). *Nucl. Acids Res.*, **31**, 75–77.
36. Huang, D.W. *et al.* (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology,* **8**, R183.1-R183.16.
37. *GENOME-WIDE ASSOCIATION STUDIES. (2008)* http://gwas.nih.gov/. visited latest by December, 2013.
38. Vega Genome Browser, (2003) http://vega.sanger.ac.uk/Homo_sapiens/Search, visited latest by December, 2013.