CpG Analyzer, a Windows-based utility program for investigation of DNA methylation

Yi-Hua Xu, Herbert T. Manoharan, and Henry C. Pitot
University of Wisconsin-Madison, Madison, WI, USA

BioTechniques 39:656-662 (November 2005)
doi 10.2144/000112053

There is substantial evidence that changes in DNA methylation occur during preneoplasia, including both global changes in DNA methylation and changes in CpG dinucleotide methylation sites in specific genes (1). Aberrant DNA methylation within CpG islands is one of the earliest and more common alterations in human malignancies (2). Cytosine methylation in CpG dinucleotides has been observed to be an important control mechanism in development and differentiation (3). The bisulfite genomic sequencing technique (4) has found wide acceptance for the generation of DNA methylation status maps with single-base resolution. This method is based on the selective deamination (induced by bisulfite treatment) of cytosine to uracil, while 5-methylcytosine residues remain unchanged. This bisulfite-modified DNA sequence is amplified by PCR and then sequenced. The uracils in the sequence are detected as thymines on PCR amplification and complement with adenines on formation of the double strand. Methylation status is obtained by the comparison of bisulfite sequence PCR products with the computer-generated bisulfite-modified sequences. Knowledge of the CpG distribution within the sequence, identifying each CpG location, and generating bisulfite-modified sequences are essential in the use of this method, while the ability to highlight CpGs in the sequence text will greatly speed up the process of sequence comparison.

Some programs are available to simplify the process (see the MethDB links web page at 195.83.84.240/links.html). MethTools is a program available only on the UNIX® operating system,

thus excluding Microsoft® Windows users from employing it (5). CpG Island Searcher is a web site that has a simple user interface to identify CpG islands from submitted sequences (6). However, this program does not supply the detailed CpG location information and CpG-highlighted text that are important for a DNA methylation map with single base resolution.

On the Microsoft Windows platform, Anbazhagan et al. used Microsoft Excel® to identify and mark CpG islands (7). CpGs can also be analyzed and highlighted by searching for "cg" in a sequence using most word processing software. However, since the sequence text commonly used is in GenBank® flat file format and consists of line numbers, spaces, and line breaks, the searching process must be performed after these extraneous characters have been eliminated. This involves repeated use of the find and delete

commands. Manual CpG analysis and sequence conversion should be avoided, since not only are these time-consuming, but mistakes are readily introduced. Singal et al. (8,9) used Microsoft Word® macros to simplify the repetitive tasks such as removing numbers, spaces, and line breaks, highlighting CpGs, and generating the bisulfite-modified sequences. However, this method does not record or generate CpG location data. Since the valuable numbers and spaces that indicate the location of CpGs are eliminated from the highlighted text, this makes it difficult to find the exact locations of the highlighted CpGs, which is important for manual inspection later,

Figure 1. User interface of CpG Analyzer. Screenshot taken following analysis of the rat insulin-like growth factor II (IGF II) gene (GenBank accession no. X17012). The whole process can be simply finished by clicking a set of command buttons in sequence (Paste Sequence, CpG Analysis, and Highlight CpG). In the central portion of the window, two RichTextBoxes hold the tested sequence (upper) and resulting highlighted sequence (lower). There are two small panels that give the user the highlighting and content options. As shown in this figure, since the user plans to export the highlighted text to Word and print it on a black and white printer, a font color of gray and a highlight color of red was chosen. Thus, the highlight text will stand out both on the computer screen and in the paper printout. Note that the GenBank flat file format has been preserved. The CpG Distribution Configuration in Sequence panel is located above the RichTextBox controls. A graphic CpG data output is shown here. On the plot, CpG-rich regions can clearly be seen. Several conventional TextBox controls are used to show the CpG analysis results and the current cursor position (indicated by the arrow). The Grid control in the CpG Position Information panel (left side of the window) displays the detailed CpG location information. The user can view all data in detail by moving the scroll bar. Command buttons labeled with Copy CpG Data (Word) and Copy CpG Data (Excel) are used for CpG location data output, depending on the destination software the user has chosen. The data exported into Excel has two columns labeled with CpG_ID and CpG Position, which list all the CpG data found in the sequence. This type of data can be used later for preparing a DNA methylation status map. The data exported into Word is formatted as a table, showing CpG number and location, with four in each line. A ProgressBar control at the top is used to show the process in progress.

during the sequence comparison phases.

To simplify some tasks involved in DNA methylation studies and to combine and expand the favorable features of current software, we have developed a utility program, CpG Analyzer. It is designed to run on the Microsoft Windows platform using Microsoft Visual Basic 6 Professional programming language. Figure 1 shows the CpG Analyzer user interface. The tasks for CpG Analyzer include: (i) performing CpG analysis and displaying the detailed results; (ii) generating bisulfite-modified sequences to identify methylated and unmethylated CpGs for use in PCR primer design and PCR product comparison; and (iii) highlighting CpG sequences for creating CpG maps.

The text of the sequence of interest can be typed in or obtained from search results on the Internet. Usually the text shown is in GenBank flat file format, which has a sequence portion and a notation portion. The sequence portion text can be copied and pasted into CpG Analyzer. Whole genes as well as gene segments can be analyzed. Processing occurs in three major steps: (i) CpG analysis; (ii) bisulfite-modified sequence generation; and (iii) CpG highlighting. This analysis process can be completed in approximately 1 min. When the process is finished, three types of results can be obtained: (i) a graphic plot of CpG distribution information; (ii) a detailed table with the CpG number and location data; and (iii) CpG-highlighted sequence text (see Figure 2, steps 1–7). These data can be viewed in CpG Analyzer. The CpG distribution plot clearly shows the density of CpGs in the sequence, from which CpG-rich regions can easily be identified. The CpG location detail can be viewed in the Grid control. The computer-generated bisulfite-modified sequence can be viewed and used for DNA methylation PCR primer design and later for the comparison of PCR products.

CpG Analyzer is designed to use the Microsoft Office suite as the destination software to export the results through the Windows clipboard using the copy and paste commands. Thus, data can be directly used in publication

and presentation. Graphical CpG distribution data can be copied into Microsoft Word for an image record or Microsoft PowerPoint® for a graphical presentation. The number data of CpG locations can be exported to Excel and Word by clicking the command buttons under Grid control. CpG-highlighted sequence text can be copied into Word (Figure 2, step 8).

To utilize CpG Analyzer most efficiently, the following should be noted: in CpG Analyzer, both compacted sequence and GenBank flat file formatted sequence can be used. However, only the sequence portion of each file should be entered. The use of GenBank flat file format has some advantages. The numbers beginning each line indicate the first base position of this line, and a space is used to separate bases into groups of 10 nucleotides. Without these numbers and spaces, the sequence code becomes a block of characters, making

the determination of CpG position in the sequence extremely difficult. CpG Analyzer is designed to highlight CpGs with the original sequence format unchanged. Special precautions were taken in programming to ensure that extra spaces, numbers, and line breaks within the text were not barriers to CpG analysis (Figure 1).

To clearly highlight CpGs in a DNA sequence, the difference in the font used for CpGs and normal bases should be kept as distinct as possible. It is better to keep the CpG in uppercase, bold, and underlined, while the rest of the text is in lowercase. The user can select different font colors for color output. In order to achieve the best alignment of the sequence, the Courier font setting should be used in the RichTextBox control. Since the output of the highlighted text is in RichText format, Word and WordPad can be used for this purpose, but NotePad cannot be used.

Figure 2. Stepwise process of data management in CpG Analyzer. (step 1) When sequence data are loaded into CpG Analyzer (by clicking Paste Sequence), it is changed automatically into lowercase. (step 2) A copy of the sequence to be analyzed is compacted to remove all numbers, spaces, and line breaks. (step 3) CpG location data are obtained using the string-searching mode. (step 4) Detailed CpG location data are organized and displayed in a Grid control. (step 5) CpG location data are plotted in a PictureBox control. (step 6) A pair of bisulfite-modified strands (methylated or unmethylated) are generated using selective deamination (induced by bisulfite) of cytosine to uracil. The original sequence text format is not changed. (step 7) CpGs are highlighted in the original sequence as well as bisulfite-modified (methylated) sequence text. (step 8) Three types of data (CpG-highlighted text, graphic plot of CpG data, and table of numbered CpG locations) can be exported into the Microsoft Office suite using the Windows copy and paste commands.

The default setting for highlighted text is CpGs in uppercase and the remaining text in lowercase. If the user prefers to record the nonhighlight text in uppercase, this can be achieved in Word by selecting all the text and using the Change Case command. Since the color and other font styles are still different, CpG highlight effects are maintained.

The first base position information is necessary to number the CpG positions in the sequence. The default setting for this is "1." For whole gene sequence analysis, since the first base is counted as 1, no further setting is required. However, in case a region in the middle or end of the gene is analyzed, the first base should be carefully set. This feature makes it possible for CpG location data generated from the gene segment to be matched with that from the entire gene. Since a maximum of 3000 CpGs in the sequence can be recorded and displayed, for a very large gene, the user should split the sequence into smaller segments for analysis.

Several useful features of the program include: (i) the CpG plotted in the PictureBox is sensitive to the position of the cursor (Figure 1); (ii) by clicking the location of interest on the plot, the text contents of the two RichTextBox controls will scroll to the corresponding location if the sequence is in GenBank flat file format; and (iii) by clicking the CpG_ID of interest in the Grid control, the text contents of the two RichTextBox controls will scroll to the corresponding location.

Management of CpG data requires the computer to run several programs simultaneously, including CpG Analyzer and individual programs inside Microsoft Office such as Word, Excel, or Power Point. To keep these programs running smoothly, an IBM-compatible PC with a high-speed CPU and substantial memory is required. In our laboratory, a Dell® Dimension 4400 computer with an Intel® Pentium® 4 microprocessor (2.0 GHz) and 1000 MB RAM is used, with a Windows XP professional operating system (SP2). Other Windows systems should be compatible. CpG Analyzer is now available from our laboratory upon request (contact us via the Pitot

laboratory web page at mcardle.oncology.wisc.edu/pitot).

ACKNOWLEDGMENTS

The authors express their sincere appreciation to Mrs. Mary Jo Markham and Mrs. Kristen Adler for their expert technical typing of the manuscript. The development of this software was supported in part by grants from the National Cancer Institute of the United States (CA07175, CA22484, and CA45700).

COMPETING INTERESTS STATEMENTS

The authors declare no competing interests.

REFERENCES

1. Pitot, H.C. and D.D. Loeb. 2002. Fundamentals of Oncology, 4th ed., Revised and Expanded. Marcel Dekker, New York.
2. Baylin, S.B., J.G. Herman, J.R. Graff, P.M. Vertino, and J.P. Issa. 1998. Alteration in DNA methylation: a fundamental aspect of neoplasia. *Adv. Cancer Res.* 72:141-196.
3. James, S.J., I.P. Pogribny, M. Pogribna, B.J. Miller, S. Jernigan, and S. Melnyk. 2003. Mechanisms of DNA damage, DNA hypomethylation, and tumor progression in the folate/methyl-deficient rat model of hepatocarcinogenesis. *J. Nutr.* 133(Suppl 1):3740S-3747S.
4. Frommer, M., L.E. McDonald, D.S. Millar, C.M. Collis, F. Watt, G.W. Grigg, P.L. Molloy, and C.L. Paul. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* 89:1827-1831.
5. Grunau, C., R. Schattevoy, N. Mache, and A. Rosenthal. 2000. MethTools—a toolbox to visualize and analyze DNA methylation data. *Nucleic Acids Res.* 28:1053-1058.
6. Takai, D. and P.A. Jones. 2003. The CpG island searcher: a new WWW resource. *In Silico Biol.* 3:235-240.
7. Anbazhagan, R., J.G. Herman, K. Enika, and E. Gabrielson. 2001. Spreadsheet-based program for the analysis of DNA methylation. *BioTechniques* 30:110-114.
8. Singal, R. and S.R. Grimes. 2001. Microsoft Word macro for analysis of cytosine methylation by the bisulfite deamination reaction. *BioTechniques* 30:116-120.
9. Shaw, G. 2000. Useful Microsoft Word Macros for molecular biologists and protein chemists. *BioTechniques* 28:1198-1201.

Received 2 June 2005; accepted 8 September 2005.

Address correspondence to Henry C. Pitot, McArdle Laboratory for Cancer Research, University of Wisconsin-Madison, 1400 University Avenue, Madison, WI 53706-1599, USA. e-mail: pitot@oncology.wisc.edu

To purchase reprints
of this article, contact
Reprints@BioTechniques.com