Microsoft® Word™ Macro for Analysis of Cytosine Methylation by the Bisulfite Deamination Reaction

BioTechniques 30:116-120 (January 2001)

Rakesh Singal and Sidney R. Grimes

Overton Brooks Veterans
Administration Medical Center
and LSU Health Sciences
Center, Shreveport, LA, USA

ABSTRACT

Cytosine methylation at CpG dinucleotides is an important control mechanism in development, differentiation, and neoplasia. Bisulfite genomic sequencing and its modifications have been developed to examine methylation at these CpG dinucleotides. To use these methods, one has to (i) manually convert the sequence to that produced by bisulfite conversion and PCR amplification, taking into account that cytosine residues at CpG dinucleotides may or may not be converted depending on their methylation status, (ii) identify relevant restriction sites that may be used for methylation analysis, and (iii) conduct similar steps with the other DNA strand since the two strands of DNA are no longer complementary after bisulfite conversion. To automate these steps, we have developed a macro that can be used with Microsoft® Word™. This macro (i) converts genomic sequence to modified sequence that would result after bisulfite treatment facilitating primer design for bisulfite genomic sequencing and methylation-sensitive PCR assay and (ii) identifies restriction sites that are preserved in bisulfite-converted and PCR-amplified product only if cytosine residues at relevant CpG dinucleotides are methylated (and thereby not converted to uracil) in the genomic DNA.

INTRODUCTION

Cytosine methylation at CpG dinucleotides has increasingly been recognized as an important control mechanism in development, differentiation, and neoplasia. Frommer et al. (1) introduced a procedure based on bisulfite-induced oxidative deamination of genomic DNA under conditions in which cytosine is converted to uracil and 5-methyl cytosine (5-mC) remains unchanged. The bisulfite-treated target sequence is then amplified by PCR using strand-specific primers. Upon sequencing of the amplified DNA, all uracil and thymine residues become detectable as thymine, and only 5-mC residues amplify as cytosines. This method is presently the method of choice for the detailed analysis of 5-mC in any given genomic target sequence (6,7,10), and it has been used in our laboratory to examine the methylation status of an avian embryonic globin gene and of the histone H1t promoter (6,9). A number of rapid methods to detect 5mC have been developed based on the bisulfite deamination reaction in combination with PCR amplification. These are suitable for examining limited numbers of CpG dinucleotides and include PCR amplification with primers specific for methylated or unmethylated alleles (3) or identification of restriction sites in PCR-amplified product (4,10). However, to be able to use these methods, one must conduct a time-consuming manual analysis of the sequence of interest. There is a need for a computer program or macro to automate this process.

Programs such as MethTools have been written to help design and analyze DNA using the bisulfite genomic sequencing technique (2). However, the programs are written in PERL 5 or C++. To run the programs, you must have a computer with a UNIX® operating system. Investigators that may not have easy access to this operating system most likely do have access to Microsoft® Word™ and other Microsoft Office® applications. Visual Basic® for Applications (VBA) programming language is included with these applications. Therefore, there is a need for a macro that can be used with Microsoft Word to help in DNA analysis for the bisulfite genomic sequencing technique and its modifications. In this paper, we present such a macro.

MATERIALS AND METHODS

Microsoft VBA supplied with Microsoft Word for Office 2000 was used to prepare this macro. The Macro was tested with Microsoft Word for Office 97 running under Windows 95 and Word for Office 2000 running under Windows 98 on an IBM-compatible computer and with Microsoft Word 98 for the Macintosh on a Power Macintosh G3 running OS 8.1. To run this macro, you need a copy of Microsoft Word running on your computer, and you need to type in or paste in the DNA sequence to be analyzed.

You may obtain the source code for the macro as an e-mail attachment from the authors. To incorporate the macro

into Word for Office 2000 on an IBM-compatible computer, you should open Microsoft Word, click the Tools Macro menu item, and click Visual Basic Editor. In the Visual Basic Editor, click File and then click Import File. Click on the macro file name and click the Open button to import the macro. The file is supplied with the extension ".bas" but it may also be supplied as a text file with the extension ".txt". Both file types can be used. To incorporate the macro into Word on the Macintosh, open the macro text file as a Word file. Click Edit, click Select All, and then click Edit Copy. Click Tools, click Macro, click Macros, and type "Bisulfite" in the Macros Name box. Click Create, click Paste to insert the macro, and click Save Normal. Click File Close and Return to Word. The macro named Bisulfite is then available to analyze your DNA sequence.

A macro button can be added to the menu bar to shorten this process to two mouse clicks. To add the button on the PC, click the Tools Customize menu item. Click the Commands tab and click tools in the left window. Scroll down and hold down the left mouse button on the macro button (a filled triangle pointing to the right). Drag the symbol to the menu bar beside Help. Right click on the symbol, click on the "image and text" choice, and close the Customize window. You can click on the new Macros menu item to display the macro list and then click Bisulfite.

VBA that is built into Word can be used to modify this macro or to write additional applications. Repetitive tasks such as Edit (Find and Replace) that we use to study bisulfite-modified DNA sequences, as described here, can be automated by writing macros using VBA. Therefore, the tools for the use and modification of this macro are available at little expense.

RESULTS AND DISCUSSION

Manual Procedure

The macro presented in this paper has been used to analyze CpG methylation of several genomic DNA sequences including the histone H1t promoter (9). To use this method, one must

manually convert the sequence of interest to the sequence after bisulfite conversion and PCR amplification, taking into account that cytosine residues at CpG dinucleotides may or may not be converted depending on their methylation status. The sequence is then analyzed for restriction enzyme cut sites to find relevant restriction sites that may be used for methylation analysis (4).

Examination of a genomic DNA sequence for the presence of CpG dinucleotides and conversion of all other Cs to Ts within the sequence is done on a word processor or by special software. The process usually takes several steps using a word processor and is required to aid in selection of PCR primers for the forward (top) DNA strand and for the reverse complement (bottom) strand. We typically copy a selection of a genomic sequence in GenBank® format directly from a file found by searching the National Center for Biotechnology Information (NCBI) Web site (<http://www.ncbi.nlm.nih.gov/>). This format has numbers at the beginning of each line of sequence, and the sequence is listed in groups of ten bases, resulting in a sequence with imbedded numbers, imbedded spaces, and a paragraph mark at the end of each line. You should not include annotated text because the macro will retain every g, a, t, and c within the annotation. However, you could easily modify the macro to strip all annotated text.

For this paper, we selected the promoter region (nucleotides 2101–2520) from the rat histone H1t genomic sequence file with accession number M28409, as shown in Figure 1. We used the Edit Copy menu item in Microsoft Internet Explorer® to copy the selected sequence from the NCBI Web site. We opened a new document in Word and used the Edit Paste menu item to paste the selected sequence into the document.

We used the Edit Find menu item to find all numbers (digits) in the sequence. The Find tab was selected, and the special characters "^#" were typed into the Find what: box. The Replace tab was selected, and the Replace with: box was left empty. Pressing the Replace All button deleted all digits (^#). The same procedure was used to replace paragraph marks (^p). To delete spaces,

the space bar was pressed one time with the cursor in the Find what: box, and nothing was entered into the Replace with: box. The font was changed to 10-point Courier New (a nonproportional font) by selecting all text, clicking the Format Font menu item, and selecting the Font type. The text was converted to lowercase by clicking the Format Change Case menu item and selecting lowercase.

We used Find and Replace to convert every CG dinucleotide to x. We then changed each remaining c to t and finally changed every x back to **CG** that was bold and underlined. If you do not remove spaces and paragraph marks, then you must Find and Replace not only adjacent Cs and Gs but also all CGs separated by one or more spaces or paragraph marks. By converting the text to lowercase, the bold and underlined **CG** dinucleotides in the final converted sequence are easier to see. Finally, we saved the modified sequence. The sequence can be loaded or pasted into a program for restriction enzyme analysis.

Restriction analysis of the modified sequence was performed with programs available on the Internet or with the Vector NTI™ 5 Suite of programs on our laboratory computer. We identified restriction sites with CpG dinucleotides in their recognition sequence. These sites would only be preserved in bisulfite-converted and PCR-amplified product if cytosine residues at relevant CpG dinucleotides are methylated (and thereby not converted to uracil) in the genomic DNA.

This rather tedious process is repeated for the reverse complement DNA strand because the two strands are no longer complementary after bisulfite conversion. The reverse complement of the original unmodified sequence must be typed in manually or produced by another software application such as Vector NTI or a Word macro published in BioTechniques (5). A software program for selecting appropriate PCR primers for amplification and sequencing can be used on the converted forward and reverse DNA strands.

Automated Procedure

The macro is installed in Word as described in Materials and Methods. Type

a DNA sequence or paste a sequence into a new document and save the document into a folder on the hard disk. In this paper, we used the promoter region containing nucleotides 2101–2520 from the rat histone H1t genomic sequence shown in Figure 1. Click the Tools Macro menu item, select Macros, and select Bisulfite from the macro list.

When you click Run, the macro produces two document files named forward.doc and reverse.doc and saves these two document files to the hard disk in the same folder used to save the original sequence. The two documents, forward.doc and reverse.doc, can be seen by clicking the Windows Menu item and clicking on the document name.

The forward.doc summary page is shown in Figure 2. The summary is composed of four parts. The top part entitled UNMODIFIED FORWARD SEQUENCE contains the original sequence with CG dinucleotides emphasized. Visual inspection of the rat Histone H1t proximal promoter in Figure 2 reveals a cluster of CG dinucleotides.

```

2101 tttttttttt gaatttgatt gaaaatcaag ggcgtttccc actagtctcg acatgcagga
2161 acatttgatc ataaagggtg aaaagtcagg tggaatccaa gttcccgtag gtcacctgtg
2221 tcataacctg agcgattctg cagagagcac acatgcaagt ttctactttt tgggggaaaa
2281 gaaacacaaa tgcccccttc ccagggggcg gggaggcgcc tagggatgca ccaatcacag
2341 cgcgcacctg totatataag gccccccccc ggacccctc tctactccag cgtgttctgc
2401 tctcctgtac tttttttttt aattgggtgt agttcttgac cagctcttga ctatgtcgga
2461 aacggctcct gcagcctcaa gtactcttgt tccagctcct gtagagaaac ctgcaactaa
  
```

Figure 1. A selection of DNA sequence in GenBank format was copied directly from a file found at the NCBI Web site. The file selected for this figure is the rat testis-specific histone H1t gene with accession no. M28409. The sequence in FASTA format will also work. The sequence was pasted to a new Microsoft Word document page as shown.

Analysis of larger genomic DNA fragments containing the H1t gene shows that CG dinucleotides further upstream are in lower abundance (data not shown).

The second section entitled MODIFIED FORWARD STRAND (UNMETHYLATED) contains the sequence that would be obtained after bisulfite treatment and PCR amplification if all CpG dinucleotides were unmethylated. Each c is converted to t if the genomic DNA is unmethylated and positions of

the original CpG dinucleotides are shown as bold, underlined **TG**.

The third section entitled MODIFIED FORWARD STRAND (METHYLATED) contains the sequence that would be obtained after bisulfite treatment and PCR amplification if all CpG dinucleotides were methylated. Note by comparing this sequence to the unmodified sequence that all cs are converted to ts if the c is not present as a CpG dinucleotide. This sequence is the one that must be used for PCR primer design

and analysis by software that identified restriction enzyme cut sites.

The fourth section entitled RESTRICTION SITES OF FORWARD SEQUENCE is identical to the third section, but the sequence has been marked to identify restriction enzymes that have CpG dinucleotides in their recognition sequence. These are sites that are preserved in bisulfite-converted and PCR-amplified product if cytosine residues at relevant CpG dinucleotides are methylated in the genomic DNA.

The reverse.doc summary page is not shown but is composed of the same four sections. Note that the bisulfite-modified and PCR-amplified reverse complement strand is no longer complementary to the modified complementary strand because of the conversion of cs to ts in each sequence. Therefore, to amplify this strand of bisulfite-treated genomic DNA by PCR, you must prepare a primer pair that is different from the primer pair used for the amplification of the forward strand. The restriction enzyme recognition sequences within the PCR product from the modified reverse complement strand are also different from those in the forward strand. Therefore, it is necessary to perform the same sequence analysis for the reverse complement strand.

Use of the Microsoft Word macro presented in this paper provides an inexpensive way to rapidly analyze genomic DNA fragments in preparation for bisulfite genomic sequencing. Details of the bisulfite genomic sequencing method, including the selection of forward and reverse PCR primer pairs for each strand of DNA, have been published (1,6,7,9). To select primer pairs and to design the bisulfite genomic DNA experiment, you must manually alter the DNA to the way it will appear after bisulfite treatment and PCR amplification.

VBA included with the Microsoft Office applications is an excellent programming tool and can be used to prepare macros for simple repetitive tasks such as converting the forward DNA sequence (top strand) to the reverse complement sequence (complementary strand), as described by Shaw (5), or it can be used in longer macros to combine several tasks into a more complex

FORWARD STRAND:

UNMODIFIED FORWARD SEQUENCE:

tcttctttttgaaatttgattgaaaaatcaaggg**CG**tttccactagttctgacatgcagga
acatttgtcaataaagggtaaaaagtcaggtggaatccaagttcc**CG**tgggtcacctgtg
tcataacctgag**CG**atttctgcagagagcacacatgcaagtttctactttttgggggaaaa
gaaacacaaatgcccttccccagggg**CG**ggggagg**CG**cctagggtgcaccaatcacag
CGCGcctgctctatataaggccccccc**CG**gacccctctctactccag**CG**ctggt**CG**c
tctctgtacttttttttaattgggtgtagttcttgaccagctcttgactatgt**CG**ga
aa**CG**gctcctgcagcctcaagtactcttgttcagctcctgtagagaaacctgcaactaa

MODIFIED FORWARD STRAND (UNMETHYLATED):

ttttttttttgaaatttgattgaaaaatcaaggg**TG**ttttttattagttttgatagtagga
atatttgttaataaagggtaaaaagttaggtggaatttaagtttt**TG**tgggttatttgtg
ttataatttgag**TG**atttctgtagagagtatatatgtaagttttattttttgggggaaaa
gaaatataaatgttttttttttagggg**TG**ggggagg**TG**tttagggatgtattaatatag
TGTGttttgttttatataagggttttttt**TG**gatttttttttatttttag**TG**ttgtt**TG**t
ttttttgtatttttttttaattgggtgtagtttttgattagttttgattatgt**TG**ga
aa**TG**gtttttgtagttttaagtattttttgttttagttttgtagagaaatttgaattaa

MODIFIED FORWARD STRAND (METHYLATED):

ttttttttttgaaatttgattgaaaaatcaaggg**CG**ttttttattagttttgatagtagga
atatttgttaataaagggtaaaaagttaggtggaatttaagtttt**CG**tgggttatttgtg
ttataatttgag**CG**atttctgtagagagtatatatgtaagttttattttttgggggaaaa
gaaatataaatgttttttttttagggg**CG**ggggagg**CG**tttagggatgtattaatatag
CGCGttttgttttatataagggttttttt**CG**gatttttttttatttttag**CG**ttgtt**CG**t
ttttttgtatttttttttaattgggtgtagtttttgattagtttttgattatgt**CG**ga
aa**CG**gtttttgtagttttaagtattttttgttttagttttgtagagaaatttgaattaa

RESTRICTION SITES OF FORWARD SEQUENCE:

ttttttttttgaaatttgattgaaaaatcaaggg**CG**ttttttattagttttgatagtagga
atatttgttaataaagggtaaaaagttaggtggaatttaagtttt**CG**tgggttatttgtg
ttataatttgag**CG**atttctgtagagagtatatatgtaagttttattttttgggggaaaa
gaaatataaatgttttttttttagggg**CG**g (AciI) gggagg**CG**tttagggatgtattaa
ttatag**CG**c (HinfI) (HhaI) G (BstUI) ttttgttttatataagggttttttt**CG**ga
ttttttttttatttttag**CG**ttgtt**CG**ttttttgtatttttttttaattgggtgtagt
ttttgattagttttttgattatgt**CG**gaaa**CG**gtttttgtagttttaagtattttgtttt
agttttgttagagaaatttgaattaa

Figure 2. The Bisulfite macro was run to analyze the sequence shown in Figure 1. Two summary documents are produced named forward.doc and reverse.doc. Each document is composed of four parts: (i) the original genomic DNA sequence with CG dinucleotides shown bold and underlined, (ii) the modified sequence resulting from bisulfite treatment of unmethylated genomic DNA with TG shown in bold and underlined to indicated the positions of CG dinucleotides in the original sequence, (iii) the modified sequence resulting from bisulfite treatment of the genomic DNA with CG shown preserved by 5mC methylation, and (iv) the modified sequence marked with restriction enzymes recognition sites preserved by CpG dinucleotide methylation.

program such as the one described here. Individuals familiar with BASIC programming language should find the Visual Basic Editor relatively easy to use especially with the extensive help menus and examples provided. Those who do not write computer programs should be able to use the macro recorder to automate complex tasks. The code produced can be edited and debugged easily to improve the macro.

The Bisulfite macro described here consolidates several time-consuming manual tasks into a program that can be run in Microsoft Word by clicking the Tools Macros menu item, clicking on macros, and clicking the macro name Bisulfite. The documents produced, forward.doc and reverse.doc, can be used to identify CpG dinucleotide rich regions of genomic DNA. The modified sequences in the second and third sections of each document can be pasted into software applications to help design appropriate forward and reverse PCR primer oligonucleotides. Restriction enzyme recognition sequences marked in the fourth section of each document identify enzymes that can be used to examine the PCR product derived from bisulfite-treated genomic DNA (4). If electrophoretic analysis of these restricted PCR products indicates the presence of the restriction site within the PCR product, then the CpG dinucleotide at that site was preserved by methylation of the cytosine in the form of 5-mC.

Primer Design for Bisulfite Genomic Sequencing versus the Methylation-Specific PCR (MSP) Assay

Genomic DNA is treated with bisulfite under conditions that convert unmethylated cytosine was to uracil, while the 5-mC remains unchanged (1). For the MSP assay (3), the PCR amplification of the modified DNA is then carried out with a set of primers designed to distinguish methylated and unmethylated alleles in the bisulfite-modified DNA, taking advantage of the sequence differences resulting from the bisulfite modification. These primers are designed for genomic sequences containing (i) several cytosines, which allows discrimination between modified and unmodified DNA, and (ii) CpG pairs near the 3' end of the

primers to provide maximal discrimination in PCR between methylated and unmethylated DNA. In addition, primers for methylated DNA are made shorter than those for unmethylated DNA, so that the resultant PCR product from unmethylated and methylated alleles could be easily distinguished on gel electrophoresis (8).

For bisulfite genomic sequencing and restriction analysis, the PCR amplification of the modified DNA is carried out with a set of primers that can amplify both methylated and unmethylated sequences. These primer sequences include cytosines that allow discrimination between modified and unmodified DNA and are devoid of CpG dinucleotides (6,9). Mixed bases (C/T) may also be substituted for cytosine residue in CpG dinucleotides.

It should be obvious that the program described in this paper will be useful for preparing sequence files that help in the design of primers for both bisulfite genomic sequencing and MSP assays. This macro is typically used with DNA sequences that are only a few hundred bases in length, as shown in Figures 1 and 2, but it can be used with much longer genomic sequences to help identify CpG islands and to find regions of DNA rich in CpG dinucleotides. We have tested it on sequences several kilobases in length.

In summary, the features included with this macro should be beneficial to those using the bisulfite genomic sequencing technique and its modifications, as well as to those conducting MSP analysis. Other useful features could be added using VBA. For example, the macro could be modified to produce a graphical representation of the linear DNA sequence that shows the relative positions of CG dinucleotides or a CG dinucleotide density distribution plot. This type of modification can be accomplished using VBA included with Word and Excel®. VBA is readily available for use, and its simplicity and power make it a relatively easy task to accomplish this type of valuable modification.

ACKNOWLEDGMENTS

This work was supported by the

Department of Veterans Affairs (R.S. and S.R.G.) and by National Institutes of Health grant no. HD29381 (S.R.G.). We thank Charles Robinson, D.Sc., P.E., Research Service (151), VA Medical Center, Shreveport, for making time available on his Power Macintosh to test this macro.

REFERENCES

1. 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.
2. 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.
3. Herman, J.G., J.R. Graff, S. Myohanen, B.D. Nelkin, and S.B. Baylin. 1996. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. USA* 93:9821-9826.
4. Sadri, R. and P.J. Hornsby. 1996. Rapid analysis of DNA methylation using new restriction enzyme sites created by bisulfite modification. *Nucleic Acids Res.* 24:5058-5059.
5. Shaw, G. 2000. Useful Microsoft® Word™ macros for molecular biologists and protein chemists. *BioTechniques* 28:1198-1201.
6. Singal, R., R. Ferris, J.A. Little, S.Z. Wang, and G.D. Ginder. 1997. Methylation of the minimal promoter of an embryonic globin gene silences transcription in primary erythroid cells. *Proc. Natl. Acad. Sci. USA* 94:13724-13729.
7. Singal, R. and G.D. Ginder. 1999. DNA methylation. *Blood* 93:4059-4070.
8. Singal, R., Z.J. Tu, J.M. Vanwert, G.D. Ginder, and D.T. Kiang. 2000. Modulation of the connexin26 tumor suppressor gene expression through methylation in human mammary epithelial cell lines. *Anticancer Res.* 20:59-64.
9. Singal, R., J.M. Vanwert, M. Bashambu, S.A. Wolfe, D.C. Wilkerson, and S.R. Grimes. 2000. The testis-specific histone H1t gene is hypermethylated in nongerminal cells. *Biol. Reprod.* 63:1237-1244.
10. Xiong, Z. and P.W. Laird. 1997. COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Res.* 25:2532-2534.

Received 21 July 2000; accepted 2 November 2000.

Address correspondence to:

Dr. Sidney R. Grimes
Medical Research Service (151),
Overton Brooks Veterans Administration
Medical Center
510 E. Stoner Ave., Shreveport, LA
71101-4295, USA.
e-mail: srgrimes@prysm.net