Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 5, No. 2, May 2014), Pages: 1-5
www.jacr.iausari.ac.ir

Improved Alignment of Bisulfite Sequencing Data Using CpG Islands

Nadia Barjaste[1,*], Reza Nadimi[2], Majid Alipour[3]

(1) Department of computer engineering, Islamic Azad University, Babol Branch, Babol, Iran
(2) Department of Computer Science, University of Mazandaran, Babolsar, Iran
(3) Department of Biology, Islamic Azad University, Babol Branch, Babol, Iran

niutster@gmail.com; nadimi@umz.ac.ir; alipourrna@yahoo.com

Received: 2013/12/15; Accepted: 2014/01/19

Abstract

DNA methylation is an important biological process involving in human disease such as cancer Insomia and Diabetes. Bisulfite sequencing (BS-Seq) with next-generation technology is an accurate method for measuring DNA methylation. BS-seq data analysis is a considerable way to recognize methylated cytosines and several tools have been developed to analysis BS-Seq such as BS-Seeker, B-SOLANA, BRAT, BSMAP and etc. In this paper, we propose a novel idea to get more efficiency in the sequencing process, This idea will improve the rate of accuracy in the BSOLANA alignment tool using a new method in the preprocessing step. Our method is based on modification in some regions of DNA strand named CpG islands. CpG islands are significant regions in DNA strand which frequency of methylated cytosines is less than other CpG contexts. We compared our method with previous methods in the preprocessing of the original BSOLANA tool using on HG19 reads. The comparison shows that new method provides more ability to align read sequences in the BSOLANA.

Keywords: *Bisulfite sequencing, DNA methylation, Alignment tools*

1. Introduction

DNA methylation, the addition of a methyl group to a cytosine base is an epigenetic modification in many eukaryotes, playing critical roles in many biological processes including gene and transposon silencing, imprinting, and X chromosome inactivation [1]. Regions of DNA, contain cytosines are major choice to determine methylation regulation. Although most of CpG dinucleotides are methylated along the genome (~60-80% in mammals), there are regions of dense CpGs called CpG islands that usually remain unmethylated [1].

Treatment of genomic DNA with bisulfite is currently the golden standard for measuring DNA methylation at single base resolution, which is followed by low or high-throughput sequencing. Sodium bisulfite converts unmethylated cytosines to uracils, leaving methylated cytosines unchanged. DNA polymerases read uracils as thymines, consequently in PCR amplification unmethylated cytosines appear as thymines whereas methylated cytosines would be unchanged [2]. In the aligning of the bisulfite-treated DNA to the original genome, both of the methylated and

unmethylated cytosines that identified as thymines (T) and cytosines (C) have the same matches (cytosines (C)) at the reference genome.
There are several tools to map bisulfite treated reads to reference DNA. In this article we review several Bisulfite Sequencing (BS)-mapping applications available. We also propose a novel idea to achieve higher accuracy in aligning BS reads.

2. Mapping problem

In the short read mapping problem [3], goal is mapping of sequenced reads to the reference genome. In the bisulfite sequencing problem, reads are not exactly a block of reference genome because unmethyled cytosines are changed to thymines and in the complementary strand, goanines are converted to adenine, therefore mapping of bisulfite converted reads is challenging and complex.

3. Alignment tools

To recognize the methylation state, bisufite treated sequences and unmodified reference must be compared. Several tools have been developed to analysis DNA methylation that all of them are common in mapping process (between unmodified sequences and BS-Seq data) but they differ in preprocessing (before mapping) and post processing (after mapping) to recognize best alignment.
Different matching tools are used for mapping process, they've been applied different algorithms such as Reference hashing, Wildcard matching and BWT algorithm. Bowtie [4] is the fastest and memory-efficient matching tool that uses BWT algorithm that performed in the matching process. There are several BS-Seq tools such as Bismark [5], BS-Seeker [6], B-SOLANA [7] that apply the Bowtie tool in mapping process.
DNA conversion in the BS-Seq is one of the several complexities in mapping process. DNA has two strands, top strand and bottom strand. After sodium bisulfite treating, there will be four distinct strands, top bisulfite treated strand, complementary of top strand, bottom bisulfite treated strand and bottom complementary (figure 1). To accomplish this predicament situation, some of mapping tools reduce alphabet cardinality as a preprocessing. In most of them, cytosine to thymine conversion is usual. According to need, the conversion maybe use in top, bottom strand, both of them, or only at the some regions of DNA.

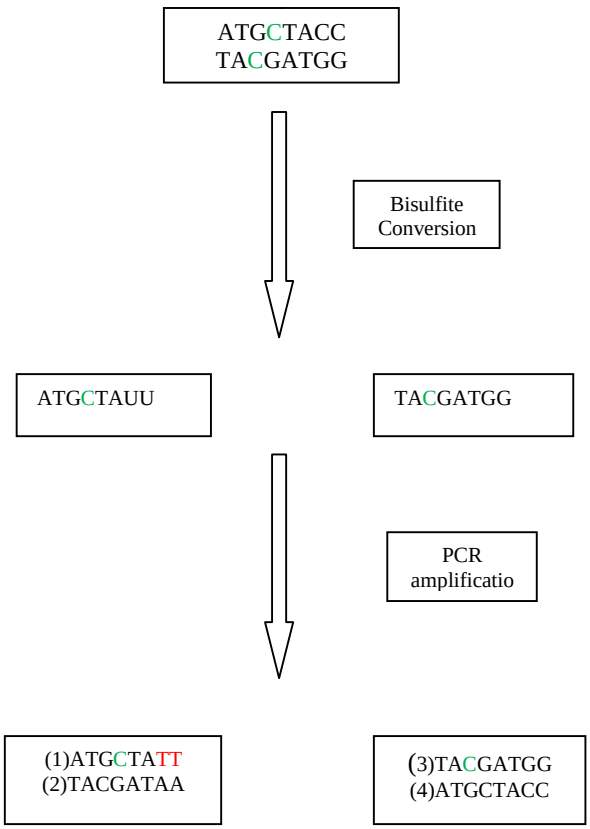

Figure 1:DNA after bisulfite conversion, (1)top bisulfite treated strand, (2)complementary of top strand (3)bottom bisulfite treated strand and (4)bottom complementary, green color determines methylated cytosine and red color determines result of cytosine conversion to thymine

In preprocessing, BSMAP [8] converts all Cs to Ts at the reads, BS-SEEKER, BRAT [9], Bismark convert Cs to Ts in reads and references. C to T conversion or G to A conversion reduce searching space but makes some ambiguous in it. To achieve these ambiguous, after mapping process, count each mismatch to determine the best mapping. BSMAP counts mismatch with utilize of xor operation and bitwise makes between reads and references to select lowest number of mismatches. Alignment tools categorized in two fields: 1.methylation aware tools which consider both cytosines and thymines as potential matches to a genomic cytosines. 2.tools which converts any residual cytosines in Bs-seq and all cytosines in the reference genome.

3.1 Directional and non-directional

To reduce bisulfite treated ambiguity, BS-SEEKER uses directional library. Moreover applies alphabet conversion. In directional library, forward or reverse tags are added appropriate to its directional. The original top or bottom strands can be recognized. Reads with this tags will be discriminates in A/T or C/G conversion in BS-SEEKER policy.

3.2 Characteristics of alignment tools

Several tools have been developed for BS-seq mapping. In practical view, all of them map BS-seq reads to the reference, but they are different in accuracy, speed, flexibility and reporting information. Recent tools are faster than first introduced

tools because they use Bowtie for mapping. Bowtie makes both accuracy and speed in mapping. Bismark and B-SOLANA report additional information such as methylaion analysis in CpG and non-CpG positions. Additional information make ease of analysis about cytosine and thymine in different contexts.

4. New method

According to B-SOLANA, there are two modifications in genome context as preprocessing, in the first one all cytosines are converted to the thymines and in the second one only cytosines in the non-CpG contexts are converted to the thymines. Cytosines in the CpG contexts are methylated more than other cytosines unless in the CpG islands that have high rate of CpG contexts with low rate of methylation. We introduce a new modification method that converts all non-CpG context cytosines and CpG contexts in the CpG-islands

5. Data set

We tested our idea on human genome(hg19) with 250000 reads. To evaluate our idea, BSOLANA's preprocessing method (the second one) was examined. In this conversion, non CpG context were converted to T and in proposed method, CpG islands with CH (H can be A, C, T) were converted to thymines.

6. Results

To assess the efficiency of our method, we compare new method with the original method of BSOLANA. New idea, "conversion in CpG islands and CH (H can be A, T, C) is compared to one efficient preprocessing method, "none CpG context conversion" In bisulfite treated alignment, goal is the mapping of BS-seq data to the reference genome. Consequently, if we can get more alignments, applied method would be efficient. To compare alignment rate of two methods, we aligned reads in two conversions form (preprocessing) with applying Bowtie2 in alignment process. There are three forms of alignments, some reads are aligned to more than one place (more than one time), some are aligned exactly one place (exactly one time) and the other ones are not aligned to any places. Reads with CpG islands and CH conversion obtained 51.07% of alignment rate, the higher rate in comparison with the other method.

Table 1. illustrates results precisely.

Preprocessing Methods	**Exactly 1 time**	**More than 1 time**	**0 time**	**Overall alignment rate**
CpG islands and CH conversion	21.24%	29.83%	48.93%	51.07%
Non CpG context conversion	17.46%	29.70%	52.84%	47.16%

7. Conclusion

Combination of bisulfite treatment and high throughput sequencing is an accurate method to measure DNA methylation. We review some of alignment tools developed for mapping BS-seq data to the reference genome. We presented a novel idea to generate more accurate mapping. Then we judged our method by experiments. The experiments are observed and enlighten us our method achieved to a good efficiency and enough accuracy.

Acknowledgment

We thank Ali Sharifi Zarchi for his helpful comments.
Analysis of data was performed using computing cluster of the Institute for research in fundamental science, Tehran (IPM).

References

[1] J.A Law, S.E. Jacobson, "Establishing, maintaining and modifying DNA methylation patterns in plants and animals," *nature*, 2011.
[2] P.W. Laird, "principles and challenges of gnome wide DNA methylation analysis," *nature*, 2010
[3] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, "Ultrafast and memory efficient alignment of short DNA sequences to the human genome," *Genome Biology*, 2009.
[4] B. Longmead, S.L. Salzberg, "Fast gapped-read alignment with Bowtie2," *Nature Methods*, 2012.
[5]]. F. Krueger, S.R. Andrews, "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications," *bioinformatics*.
[6] P.Y. Chen, S.J. Shawn, M. Pellegrini, "BS Seeker: precise mapping for bisulfite sequencing," *BMC bioinformatics*, 2010.
[7] B. Kreck, G. Marnellos, J. Richter, F. Krueger, R. Siebert, A. Franke, "B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data," *bioinformatics*,2012.
[8] Y. Xi, W. Li, "BSMAP: whole genome bisulfite sequence MAPping program," BMC *bioinformatics*, 2009.
[9] E.Y. Harris, N. Ponts, A. Levchuk, K.L. Roch, S. Lonardi, "BRAT: bisulfite-treated reads analysis tool," *bioinformatics*, 2010.