engineering proceedings

MDPI

Proceeding Paper

A Parallel Tool for the Identification of Differentially Methylated Regions in Genomic Analyses †

Alejandro Fernández-Fraga *, Jorge González-Domínguez and Juan Touriño

Computer Architecture Group, CITIC, Universidade da Coruña, 15071 A Coruña, Spain; jgonzalezd@udc.es (J.G.-D.); juan@udc.es (J.T.)

* Correspondence: a.fernandez3@udc.es

† Presented at 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: Methylation is a chemical process that modifies DNA through the addition of a methyl group to one or several nucleotides. Discovering differentially methylated regions is an important research field in genomics, as it can help to anticipate the risk of suffering from certain diseases. RADMeth is one of the most accurate tools in this field, but it has high computational complexity. In this work, we present a hybrid MPI-OpenMP parallel implementation of RADMeth to accelerate its execution on distributed-memory systems, reaching speedups of up to 189 when running on 256 cores and allowing for its application to large-scale datasets.

Keywords: methylation; whole-genome bisulfite sequencing; high performance computing; MPI; OpenMP

Citation: Fernández-Fraga, A.; González-Domínguez, J.; Touriño, J. A Parallel Tool for the Identification of Differentially Methylated Regions in Genomic Analyses. *Eng. Proc.* **2021**, 7, 44. https://doi.org/10.3390/engproc2021007044

Academic Editors: Joaquim de Moura, Marco A. González, Javier Pereira and Manuel G. Penedo

Published: 21 October 2021

1. Introduction

DNA methylation is a chemical modification of DNA resulting from the addition of a methyl group to a certain nucleotide. This process, which mainly occurs at cytosines within particular regions of the DNA, is associated with different biological functions, and abnormal methylation levels can indicate the presence of certain diseases. For instance, the existence of regions with different methylation levels is a common characteristic for several types of cancer. Therefore, discovering differentially methylated regions is an important research field in genomics, as it can help to anticipate the risk of suffering from some diseases. Nevertheless, the high computational cost associated with this task prevents its application to large-scale datasets.

In this work, we focus on the tool RADMeth [1], since it showed superior accuracy in terms of biological results when compared to several counterparts [2]. RADMeth is a publicly available tool to find individual differentially methylated sites and genomic regions in data from Whole-Genome Bisulfite Sequencing experiments, which is the state-of-the-art technology for obtaining a comprehensive view of DNA methylation. The tool uses beta-binomial regression for high-precision differential methylation analysis over these data, and it can handle medium-size experiments. Despite its accurate biological results, the main drawback of RADMeth is its high computational requirements that prevent its usage on large-size experiments.

The objective of this work is to develop a parallel tool for the identification of differentially methylated regions that provides exactly the same accurate biological results as RADMeth, but employs High Performance Computing (HPC) techniques to accelerate the execution. This goal has been achieved by using the Message Passing Interface (MPI) and OpenMP parallel programming standards to provide support for multiple processes and threads, respectively. This parallel implementation allows for the exploitation of the architecture of distributed-memory systems, such as multicore clusters, and then an analysis of large-size datasets by significantly reducing the execution times associated with them.

2. Parallel Implementation of RADMeth

RADMeth works with datasets that are represented as matrices, where the columns are the samples, the rows are the CpG sites (i.e., regions of the DNA where methylation mainly occurs), and each position in the matrix contains two values: one containing the number of reads associated with a certain CpG site, and another one containing the number of methylated reads. The output dataset is formatted with a row for each CpG site, which contains the p-value from the experiment associated with it. In this work, we propose a hybrid parallel tool based on the MPI message-passing library and OpenMP threads that works with the same data format as RADMeth and guarantees the same exact accurate results, but with significantly reduced execution times.

This has been achieved by applying domain decomposition to the input CpG sites. This means that the CpG sites of the input dataset are divided into blocks so that each block is processed in parallel by a different processing element. Our implementation makes use of this domain decomposition at two levels of parallelism. First, the dataset is statically distributed in blocks of the same size among MPI processes. Besides the execution time gains, this MPI parallelization enables the joint usage of the memory of several nodes. Second, the workload of each process is dynamically distributed among OpenMP threads.

Two performance optimizations have been applied to our parallel tool:

- Parallel data loading and storing, to take advantage of parallel programming not only in the critical phase of the tool but also in the input and output phases. This optimization technique not only improves the performance of the tool on its own, but it also allows the tool to load the entire dataset at once in a distributed and scalable manner, without concern about memory problems, which allows the tool to achieve an even better performance in these phases.
- Dynamic workload distribution. The time to process different CpG sites can present high variability. This may lead to workload imbalances; that is, even though the blocks of data associated with each process have the same size, one block might need more time than another one. This problem is alleviated thanks to the second level of parallelization, where the workload for each MPI process is distributed dynamically among the OpenMP threads launched by it.

3. Results and Conclusions

The experiments for the performance analysis of the parallel version of RADMeth were conducted on 16 nodes of the "Pluton" cluster, a distributed-memory system that is based on Intel Xeon processors and installed at CITIC. Each node is composed of two processors with eight cores each (16 logical threads using HyperThreading) and 64 GB of main memory. The whole system provides a total of 256 cores (512 logical threads with HyperThreading) and 1 TB of memory. In order to keep the reproducibility of the experiments, we have used two representative datasets that are publicly available, Akalin 2012 and Hansen 2014. The Akalin dataset is made of 28,670,426 CpG sites and 2 samples, whereas the Hansen dataset consists of 28,217,449 and 6 samples. In addition to those properties, it is important to mention that the Akalin dataset is a worst-case scenario in terms of workload imbalance, meaning that a high percentage of the workload is associated with a very small contiguous block of CpG sites. A parallel tool with only static data distribution is unable to accelerate the execution of this dataset, proving the dynamic workload distribution to be an outstanding performance optimization, completely compulsory for the tool to scale in critical scenarios.

Table 1 also shows the runtime and speedup (in parentheses) for different number of nodes. Regarding the parallel executions, the same configuration for each node was used, as it proved to be the most efficient one after some preliminary tests: one MPI process with 16 cores each (32 logical threads per MPI process using HyperThreading). The reason is that the dynamic distribution, necessary for high scalability, is implemented at the thread level, so the higher the threads-to-process ratio is, the better the workload balance is, and the better the performance of the parallel tool is. As can be seen in the table, the parallel

version of RADMeth performs well and offers excellent scalability (execution times largely decrease when the number of cores increases). Furthermore, due to the execution time reduction, the parallel tool allows for the analysis of large-scale datasets when this method is applied.

Table 1. Dataset scalability, execution times (in seconds), and speedups (in parentheses).

	Akalin 2012	**Hansen 2014**
Original RADMeth	10,886.90	45,931.20
1 Node	498.65 (21.83)	2424.93 (18.94)
2 Nodes	304.76 (35.72)	1228.00 (37.40)
4 Nodes	326.55 (33.33)	803.70 (57.14)
8 Nodes	215.10 (50.61)	444.67 (103.29)
16 Nodes	121.35 (89.71)	242.97 (189.03)

Author Contributions: Conceptualization, J.G.-D. and J.T.; methodology, A.F.-F., J.G.-D. and J.T.; implementation, A.F.-F.; validation, A.F.-F.; writing—original draft preparation, A.F.-F.; writing—review and editing, J.G.-D. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Innovation of Spain (PID2019-104184RB-I00/AEI/10.13039/501100011033) and by Xunta de Galicia and FEDER funds of the EU (Centro de Investigación de Galicia accreditation 2019-2022, ref. ED431G2019/01; Consolidation Program of Competitive Reference Groups, ED431C 2021/30).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Dolzhenko, E.; Smith, A.D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinform.* **2014**, *15*, 215. [CrossRef] [PubMed]
2. Klein, H.U.; Hebestreit, K. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. *Briefings Bioinform.* **2016**, *17*, 796–807. [CrossRef] [PubMed]