A Cloud Infrastructure for Optimization of a Massive Parallel Sequencing Workflow

Olivier Terzo, Lorenzo Mossucca
Istituto Superiore Mario Boella (ISMB)
Torino, Italy
Email: (terzo,mossucca)@ismb.it

Andrea Acquaviva, Francesco Abate, Rosalba Provenzano
Politecnico di Torino
Torino, Italy
Email: (andrea.acquaviva,francesco.abate)@polito.it
rosalba.provenzano@studenti.polito.it

***Abstract*—Massive Parallel Sequencing is a term used to describe several revolutionary approaches to DNA sequencing, the so-called Next Generation Sequencing technologies. These technologies generate millions of short sequence fragments in a single run and can be used to measure levels of gene expression and to identify novel splice variants of genes allowing more accurate analysis. The proposed solution provides novelty on two fields, firstly an optimization of the read mapping algorithm has been designed, in order to parallelize processes, secondly an implementation of an architecture that consists of a Grid platform, composed of physical nodes, a Virtual platform, composed of virtual nodes set up on demand, and a scheduler that allows to integrate the two platforms.**

***Keywords*-grid computing; cloud computing; virtual environment; next generation sequencing; hybrid architecture; massive parallel sequencing.**

I. Introduction

Massive Parallel Sequencing is a term used to describe several revolutionary approaches to DNA sequencing, called Next Generation Sequencing (NGS) technologies, these allow to draw from a single experiment a larger amount of data sequence with the previous technology known as Sanger Sequencing [1]. NGS platform aims to obtain from the molecules of DNA/RNA of smaller fragments, called *read*, which are sequenced in parallel thus reducing the processing time. Aberrant mutations in the RNA transcription, as chimeric transcripts, are on the base of various forms of disease and NGS proved to be extremely helpful in making the detection of these events more accurate and reliable. However, even if from biological point of view NGS technology leads to new exciting perspectives spreading an incredible amount of data, on the other hand it raised new challenges in the development of tools and computing infrastructures. An NGS machine produces millions of reads in a single run that must be successively elaborated and analyzed. TopHat [3] is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. It is built on Bowtie [2], an ultrafast short read mapping program. The purpose is to offer to biologist an friendly infrastructure to conduct their research and to respond to the ever evolving needs of NGS users. Biologists are already using the Amazon Elastic Cloud Computing infrastructure for their research but in some contexts, it is preferable to use a number of instances of a tailored Virtual Machine than submitting jobs to the own existing infrastructure. The paper is organized as follows: Sect. 2 motivation is discussed. Sect. 3 shows architecture design. Sect. 4 is related to performances. The last Sect. draws conclusion and future work.

II. Motivation

NGS data sample consists of millions of reads, and in a classic situation, with only one workstation available, time needed to obtain the output increases significantly. Alignment is a process in which each mapping reference is made to read independently from the other reads, and this means that can perform a parallel analysis of the data. Although alignment is a very basic operation, computational effort is very high that is due to the great number of data involved in the process. This scenario recalls for the need of developing computing infrastructures presenting high performances CPU capability and memory availability.

III. Infrastructure Description

During preliminary phase of reverse engineering of TopHat, blocks of transactions have been identified that were executed sequentially. We have divided TopHat flow into three steps, that are executed by Bowtie: Step (a): left and right segments mapped with Human Genome (HG19); Step (b): segments mapped with HG19; Step (c): segments mapped with segment juncs. At the end of each step, there is a join phase of the results previously obtained. For Steps (a) and (c), since files involved in the elaboration are significant, a common repository has been created that contains the temporary folder used by TopHat. Network File System (NFS) has been used to configure the common repository (CR), that allows computers to share files and folders over a network. Step (b) instead involves small files, these can be performed on a Grid, both physical and virtual, because transfer time is neglected.

1) System Overview: The architecture in Figure 1, called VirtualBio, allows to manage RNA data, prepared by a distributed version of TopHat, is composed of three main components: a Master Node (MN), a set of Physical Worker Nodes that establish the Grid environment while a set of Virtual Worker Nodes that establish the virtualized environment. The MN contains the database, where information

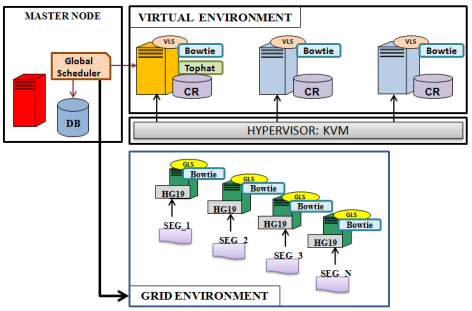

Figure 1. Architecture Overview.

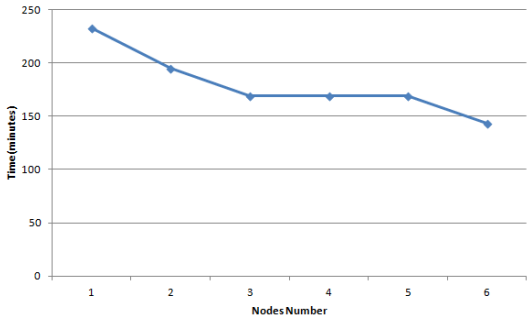

Figure 2. TopHat Execution Time for a Single Sample.

about nodes belonging to the infrastructure have stored, node status, workflow for each analysis and system monitoring. Both environments are configured with the middleware Globus Toolkit, since it allows obtaining a reliable information technology infrastructure that enables the integrated, collaborative use of computers and networks supporting the development of applications for high performance distributed computing environments. The Grid environment consists of machines with high computing power using own machines or belonging to different virtual organizations. Virtualized environment of private Cloud configured with KVM, allows use of pre-installed images, which contain all software and libraries needed for analysis that allows to set up easily machines when you need them, and once used close instance.

2) Schedulers: Three job schedulers have been developed: Grid Local Scheduler (GLS), Virtual Local Scheduler (VLS) and Global Scheduler (GL). GLS has been developed for Step (b), is active on physical machines, it aligns segments with respect to the HG19 through Bowtie. Since transfer of input file can be neglected, Worker Nodes do not need to be in the same subnet, but may also belong to different virtual organization, so system can have greater scalability and can use machines powerful performance. VLS is a scheduler active on virtual machines. Its purpose is to draw up Steps (a) and (c) through Bowtie but Step (a) allows alignment with respect to the HG19 and Step (c) allows alignment with respect to the segment juncs previously constituted by TopHat. Since the considerable size of the files involved in these 2 steps, VLS works directly in the temporary folder that is located in the RC, allowing to avoid wasting time due to the data transfer. GL is located on MN, it checks nodes available in the Grid infrastructure, when not enough nodes are available, it provides to set up virtual machine and distributes input file to the Worker Nodes, then it waits to receive all output files to proceed with the next step of the flow.

IV. Performance Considerations

Input in Step (a), consists of two files and have size about 3 GB each, in Steps (b) and (c), consists of twelve files about 500 MB each instead support files (HG19 and segment juncs) have a size of about 4 GB each. In the original version of TopHat, for elaboration of a single sample, RAM required is about 8 GB and at least 60 GB of free space hard disk. Figure 2 depicts the processing time of entire flow of TopHat for a single sample when nodes number increases. Elaboration with only a node corresponds to original version of TopHat, it means in sequential version. We want to focus the attention on elaboration time when 3/4/5 nodes are available, we obtained no gain of time because each node has more than one segment to process.

V. Conclusion and Future Work

VirtualBio is a tool for NGS analysis, especially for the alignment phase through TopHat and Bowtie. The solution covered both the field of infrastructure and the optimization software. Infrastructure is based on Grid and Virtual environment, using a common repository and a couple of job scheduler. TopHat algorithm has been optimized making parallel independent sections that were sequential. The system allows to reduces the elaboration time already for a single sample. Future work includes an improvement of scheduling policies, balancing number of jobs and resources, this study also opens to a scenario multi samples, allowing to elaborate more sample simultaneously.

References

[1] Sanger F., Nicklen S. and Coulson A.R., *DNA sequencing with chain-terminating inhibitors*, Proc. Natl. Acad. Sci. USA, 1977, pp. 5463-5467

[2] Langmead B., Trapnell C., Pop M. and Salzberg Steven L. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology 10:R25

[3] Trapnell C., Pachter L. and Salzberg Steven L. *TopHat: discovering splice junctions with RNA-Seq*, Bioinformatics (2009), Vol. 25, pp. 1105-1111

[4] Berman F., Fox G. and Hey A.J.G. *Grid Computing Making the Global Infrastructure a Reality*, Wiley, 2005, pp.171-198