2012 Third International Conference on Intelligent Systems Modelling and Simulation

Reverse Engineering of TopHat: Splice Junction Mapper for Improving Computational Aspect

Olivier Terzo, Lorenzo Mossucca
Infrastructure and Systems for Advanced Computing (IS4AC)
Istituto Superiore Mario Boella (ISMB)
Via P.C. Boggio 61, Torino, Italy
E-mail: (terzo,mossucca)@ismb.it

Andrea Acquaviva, Francesco Abate,
Elisa Ficarra, Rosalba Provenzano
Department of Control and Computer Engineering
Politecnico di Torino
Corso Duca degli Abruzzi 24, Torino, Italy
E-mail: (andrea.acquaviva,francesco.abate)@polito.it
elisa.ficarra@polito.it, rosalba.provenzano@studenti.polito.it

***Abstract*—TopHat is a fast splice junction mapper for Next Generation Sequencing analysis, a technology for functional genomic research. Next Generation Sequencing technology allows more accurate analysis increasing data to elaborate, this opens to new challenges in terms of development of tools and computational infrastructures. We present a solution that cover aspects both software and hardware, the first one, after a reverse engineering phase, provides an improvement of algorithm of TopHat making it parallelizable, the second aspect is an implementation of an hybrid infrastructure: grid and virtual grid computing. Moreover the system allows to have a multi sample environment and is able to process automatically totally transparent to user.**

***Keywords*-Next Generation Sequencing; Grid Computing; Virtualization; Cloud; TopHat; Scheduler; E-science.**

I. Introduction

Next Generation Sequencing (NGS) technologies has completely revolutionized the functional genomic research leading to an unprecedented availability of biological data [1]. NGS machine produces millions of reads in a single run that must be elaborated and analyzed. The main novelty consists in the possibility of sequencing an entire genome or transcriptome sample with substantially lower costs respect to the previous Sanger sequencing methodology [2]. The reason behind this biotechnological performance increase is the capability of NGS machines to chop the DNA/RNA molecules into small fragments, namely 'reads', that are successively sequenced in parallel with considerable saving in terms of time and economic resources. From biological and technical point of view NGS technology leads to new challenges in terms of development of tools and computational infrastructures [5]. In fact actually biotechnological laboratories are able to produce a huge amount of DNA/RNA sequencing. The first immediate effect is related to capabilities of modern computing infrastructure and tools to be in condition to analyze this huge quantity of data for biological analysis, gene expression profiling, small non coding RNA profiling, novel genes discovery, aberrant transcript event detection [3], the second effect is related to the execution times, in fact by growing volume of data the total process time for NGS elaboration increases dramatically. The reads alignment is a very basic operation that maps the reads on a genome reference in order to reconstruct the original sample sequence and reveal fundamental biological information. The huge number of reads to be mapped and the possibly large dimension of the genome reference itself make the alignment a not trivial operation. Several tools and programs have been recently developed to optimize the alignment phase, in particular, Bowtie [6] has been widely diffused because more efficient with short reads respect to previous solutions, it supports multi-threaded processing and it presents an efficient memory usage. However, the alignment phase can be complicated due to splicing events occurring in the data samples. TopHat [7] is build on top of Bowtie and it is specifically aimed at finding new splice junction and mapping reads on top of them. The importance of finding new junctions are: to improve knowledge of trascriptional biological process and to improve reads mapping probability. Consequently, it performs a more accurate alignment. Therefore, the main challenges in the elaboration of NGS data can be reconducted to the elaboration of million of reads for each single sample. Due to this consideration some specific challenges for news scenarios are reconducted for developing new computing infrastructures by adapting biological tools for a integration in virtualized grid infrastructure with high performances CPU capability and memory availability in a multi user context. However, an accurate analysis on how optimally integrate NGS data on a grid based system is performed. The nature of NGS data makes this aspect not trivial because of the tradeoff between the data transferred on the network and the improvement of flexibility for a multi user context. Study for parallelization of some TopHat workflow analysis in order to find an optimal process scheduling in a flexible computational infrastructure for the main execution steps of TopHat have been explored. The rest of the paper is organized as follows: Section 2 presents authors contribution and motivations, Section 3

978-0-7695-4668-1/12 $26.00 © 2012 IEEE
DOI 10.1109/ISMS.2012.76

IEEE computer society

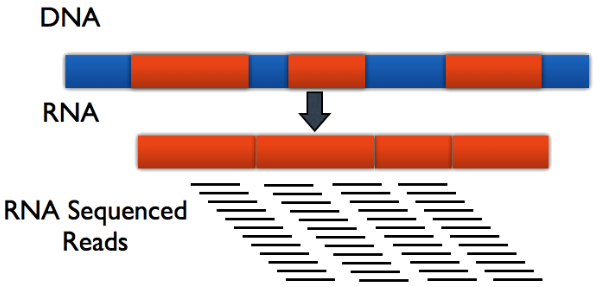

Figure 1. Alignment phase.

gives an overview of the TopHat algorithm and Bowtie tools, Section 4 describes the computing infrastructure and how TopHat was adapted for a parallelization of some mapping reads process in a multi user environment, Section 5 shows performance aspects, Section 6 draws the conclusions and road map for future work.

II. Proposed approach

The contribution presents a new approach for the DNA/RNA analysis based of TopHat in a flexible infrastructure in term of scalability in up and down scaling. The main idea is to create a flexible architecture to process the entire processing chain for Next Generation Sequencing. The solution allows: to share the computational resources, to transfer a great deal of files and to submit several mapping process in parallel mode. The system makes the process regardless of the number of available nodes on the computing infrastructure without any human interaction so that the nodes number is transparent to the user. The infrastructure is an integrated system devoted to handle automatically DNA/RNA samples in multi user context. It is composed of a biological softwares (TopHat, Bowtie), middleware for the infrastructure management, central repository, relational database and specific scheduler for resource and job controller. The software used for the entire processing chain, installed on each worker node, is a modified version of native TopHat tool that is a fast splice junction mapper for RNA-Seq reads. The Bowtie tool aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour (see Figure 1). Each TopHat instance provides sequential and parallel executions due to the alignment phase of Bowtie in a hybrid and flexible computing infrastructure composed by physical workers nodes in a very common standard grid computing architecture and with virtual grid nodes in a cloud computing infrastructure. The integration between the two infrastructures and the execution of multiple instances of TopHat is managed by the scheduler. The data obtained by executing TopHat, are used for further biomedical applications as: detection of new isoform, differential gene expression analysis, detection of aberrant mutation. The Globus Toolkit is used as middleware [9], since it allows obtaining a reliable information technology infrastructure that enables the integrated, collaborative use of computers, networks and databases. Two schedulers are developed: global and local scheduler [10]. The global scheduler, installed on grid master, allows to collect all jobs to elaborate and to distributing them among the worker nodes. Instead, the local scheduler is installed on each worker nodes and its aim is to process data in accordance with the grid master request.

A. Motivation

In NGS technology context the amount of data and the number of samples to be analyzed is growing constantly. It is a positive factor for increasing more accurate studies and results in term of reliable identifications of mutations in aberrant splicing events, fused genes and open new perspective and challenges for adapting tools that are in condition to make pre and post processing in modern infrastructure like virtualized machine, grid and cloud computing. A NGS data sample consists in millions of data and the time needed for the process execution increase dramatically with only one workstation for processing NGS data. The alignment phase is a process which each mapping reference is made in independent way and can be execute in a parallel way on a distributed computing context. The alignment is a very basic operation but actually the alignment tool that we use, TopHat, provides a sequential analyze of each block of reads and in a single user way and consequently in a typical scenarios were more samples need to be analyze the total time for processing data is not acceptable. Sequence alignment is a way of arranging the sequence od DNA/RNA a protein to identify region of similarity. In consideration of this a first challenge is to adapt the tool by making a reverse engineering of the code for a transformation of all possible main and heavy sequential alignment process and a parallelization way in order to obtain an optimization of process time. A second challenge is to adapt the alignment tool for a multi user and multi sample process in a single instance. This new scenario characterizes the capabilities to make splice junction mapping in an flexible and distributed computing infrastructure.

III. Background

A. TopHat Algorithm

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. TopHat is a collaborative effort between the University of Maryland Center for Bioinformatics and Computational Biology and the University of California, Berkeley Departments of Mathematics and Molecular and Cell Biology. TopHat receives as input reads

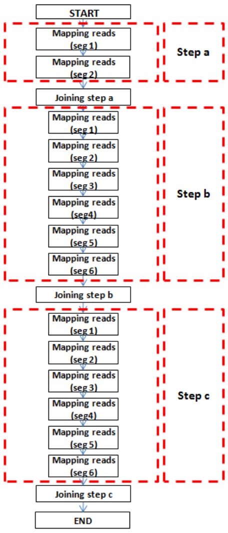

Figure 2. Simple Sequential TopHat Flow.

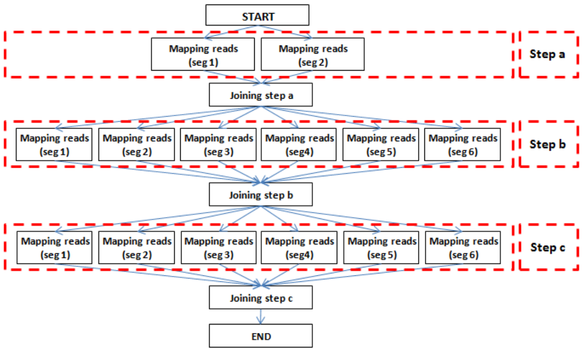

Figure 3. Parallel TopHat Flow.

produced by the Illumina Genome Analyzer, although users have been successful in using TopHat with reads from other technologies [7]. The input sample consists of two files of about 37 million of reads each. The two files are FASTA formatted paired-end reads. Dealing with paired-end reads means that the reads are sequenced by the sequencing machine only on the end of the same DNA/RNA molecule, thus the sequence in the middle part is unknown. Each sequenced end of the same read is also referred as mate. It results in two distinct files, the first one consists in the first mate of the same reads and the second one consists in the opposite mate. TopHat finds junctions by mapping reads to the reference in two phases. In the first phase, the pipeline maps all reads to the reference genome using Bowtie. All reads that do not map to the genome are set aside as 'initially unmapped reads'. Bowtie reports, for each read, one or more alignment containing no more than a few mismatches in the 5'-most s bases of the read. The remaining portion of the read on the 3' end may have additional mismatches, provided that the Phred-quality-weighted Hamming distance is less than a specified threshold.

B. Alignment Tools

The short reads alignment is surely the most common operation in RNA-Seq data analysis. The purpose of the alignment is to map each short read fragment onto a genome reference. From the computational point of view, each short read consists in a sequence of four possible characters corresponding to the DNA bases and the sequence length depends on the sequencing machine adopted for the biological experiment [6]. The main novelty introduced by NGS technology is the capability of sequencing small DNA/RNA fragments in parallel, increasing the throughput and producing very short reads as output. However, this feature makes the computational problem more challenging because of the higher amount of reads produced and the accuracy in the mapping (the shorter sequence length, the higher probability of having multiple matches). For this reason many alignment tools specifically focussed on the alignment of short reads have been recently developed. In the present contribution, we are interested in characterizing the performances of alignment tools on real NGS data. On the wave of this remark, Bowtie has been chosen, a wide diffused alignment program particularly aimed at align short reads. In order to detect the actual limitation of the alignment phase, we consider real NGS data coming from the analysis of Chronic Myeloid Leukemia. In our analysis flow, the HG19 assembly produced in the 2007 is considered as reference genome the last human genome assembly produced to now. In order to increase the computational performances during the read mapping, Bowtie program creates an index of the provided human genome reference. This operation is particularly straightforward from the computational point of view, but it must be performed only one time for the human genome reference and it is independent on the mapping samples. The alignment phase itself is particularly suitable to be parallelized. In fact, each mapping operation is applied to each read independently on the other reads mapping.

IV. VirtualBIO NGS Infrastructure

In a preliminary phase of reverse engineering, studying TopHat, blocks of transactions have been highlighted that were executed sequentially. We have identified three main blocks, that can be executed independently: step a) left and right check reads segments mapping with HG19; step b) left and right mapping segments with HG19; step c) left and right mapping segments with segment juncs. Figure 2 depicts the original version of TopHat, where is clear an sequential approach, instead Figure 3 represents the parallelization of TopHat algorithm implemented in our solution. A feature of these 3 blocks is that they are performed by a external software (Bowtie). For steps (a) and (c), since files involved in the development are significant, we created a common repository that contains the temporary folder used by TopHat. Instead the step (b) uses small files these can be performed on a grid, both physical and virtual, because the transfer times can be neglected. Only difference that the input files are transferred to worker nodes through Globus. These Worker Nodes when the process is terminated, resend the output file to the node that requested execution. This platform aims to be a service that is given to biologists for the NGS analysis but not only, the solution also provides a case study where multiple users require the execution of analysis simultaneously. The architecture, called VirtualBio NGS (see Figure 4), is composed of three main components: a Master Node(MN), a part consists of the Physical Worker Nodes (PWN) that set the grid environment while a part consists of Virtual Worker Nodes (VWN) that set the virtualized environment. The MN is a physical machine with good hardware characteristics, is responsible of Certification Authority, contains the database, where all information about the nodes belonging to the infrastructure, the node status, the flow of the various biological analysis that can be made in the system and system monitoring have stored. 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, networks and databases. The Globus Toolkit is a collection of software components designed to support the development of applications for high performance distributed computing environments, or computational grids. In spite of the success of Grid computing in providing solutions for a number of large scale science fields one of the problems is the scalability of the system. The agents are a solution to provide flexibility and scalability [3], [4]. In fact, the first requirement for the scheduler is that it never requests information on each nodes to take a decision to schedule the jobs. The system agent, installed on each worker node, is used to monitor the availability of each service on the node and, periodically, it sends its status information to the database to Master Node, and if the node has all services active means it is able to execute jobs. The advantage is that the scheduler only queries the database for the pre-selection of the nodes list in condition to receive a job. A specific function checks if the feedback information from all nodes has been sent and, in case of missing status, the node is considered not available for jobs execution.

A. Distributed Platform

The grid environment consists of machines with high computing power. Grid environments are scalable, making them effective for uses where storing large amounts of data are important. The only requirement is to have the necessary software installed for the processing (Bowtie and Globus). On each worker node of the grid environment is installed the Grid Local Scheduler, an essential component for performing biological tests. Virtualized environment also helps to improve infrastructure management, allowing the use of virtual node template to create virtual nodes in a short time, speeding up the integration of new nodes on the grid and, therefore, improving the reactivity and the scalability of the infrastructure. Hypervisor KVM was used for the creation of Full Virtualized machines. By adding virtualization capabilities to a standard Linux kernel, the virtualized environment can benefit from all the ongoing work on the Linux kernel itself. The virtualized environment has pre-installed images, which contain all software (Bowtie and TopHat), local schedulers (VLS, GLS) and support data (HG19). To have images already configured allows to set up easily machines when you need them, and once used the close the instance.

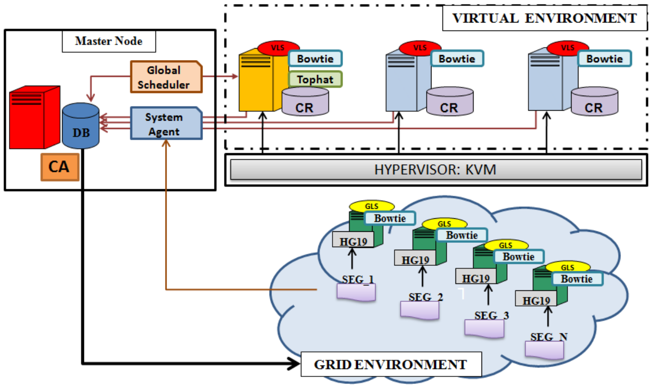

Figure 4. Computing Architecture.

V. Scheduler Approach

The Grid Local Scheduler (GLS) is a scheduler active on physical machines, has been developed for the design phase (b), it aligns the segment with respect to the human genome (HG19) through Bowtie. Since the transfer of the input file is not influential, the worker nodes do not need to be in the same subnet as the master node, but may also belong to different virtual organization, so system can have greater scalability and can use machines powerful performance. The Virtual Local Scheduler (VLS) is a scheduler active on virtual machines. Its purpose is to draw up the steps (a) and (c) of TopHat. As the GLS, the VLS performs the mapping files through Bowtie. The step (a) allows the alignment with respect to the human genome (HG19) and step (c) allows the alignment with respect to the segment juncs previously constituted by TopHat. Since the considerable size of the files involved in these two steps, the VLS works directly on the temporary folder that is located in the common repository, allowing to avoid wasting time due to the transfer

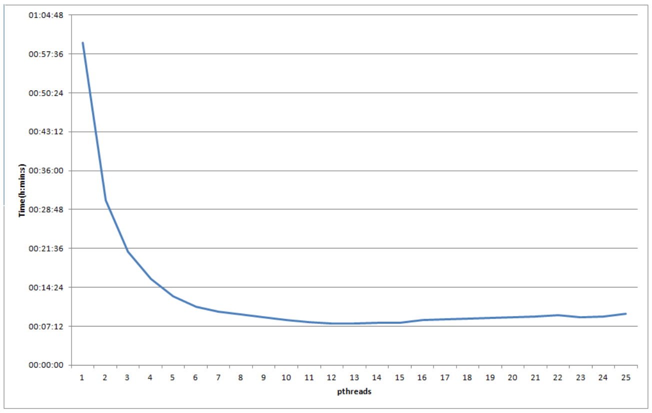

Figure 5. Bowtie Execution Time.

of data. Even in this case the interaction with the database is essential and very frequent, network problems may affect the entire biological analysis.

VI. Experiment

As we explained earlier, during an analysis phase of the algorithm, 3 main blocks have been identified, (a) left and right mate aligned with HG19, (b) segments aligned with HG19, (c) segments aligned with segment juncs. In Figure 5, processing time of a single segment of the variation of the parameter pthread is depicted. The processing time of each segment depends on parameter pthread that is specified in command of Bowtie and refers to the number of parallel processes that can run. Once past this threshold, the trend is no longer regular, this is due to the scheduling allocation of the CPU operating system. This test allowed to have a vision on the processing time will have access to machines with different power, opening to a more accurate scheduling policy adapted to the needs of time of the biologist. The test was run on a machine with 12 CPUs, it is worth noting that in order to gain the maximum time the number of pthread must be equal to the number of CPUs.

The aim of tests performed is to compare execution time for multi samples using the two version of TopHat: the original version that sequential approach and version modified exploiting the distributed environment. For this test phase, we wanted to use an architecture which consists of three machines with four CPUs. In Figure 6, we can notice that already only a sample processed with the version of TopHat Grid, a time savings of 40% is obtained instead increasing the number of samples to be processed, it is worth noting that the percentage of earned time is about 30%, this is due to the jobs queues that are created on the nodes.

VII. Conclusions and Future Works

VirtualBio NGS is a tool for NGS analysis, in particular 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 Grid Platform, using a common repository and a couple of job schedulers. The TopHat algorithm has been optimized making parallel independent sections that were sequential and has been modified for giving a multi user environment were before on the native version of TopHat was for a single user instance. This distributed system allows to reduces the elaboration time for a single sample by at least 40% and about 30% in a multi sample context, using machines in the own virtual organization but not only, this value can change it depends on the power of the machines used. Future works include the improvement of scheduling policies, balancing jobs and resources, this study also opens to a scenario for increasing the capabilities of the scalability of the infrastructure through an integration of Public Cloud as Amazon.

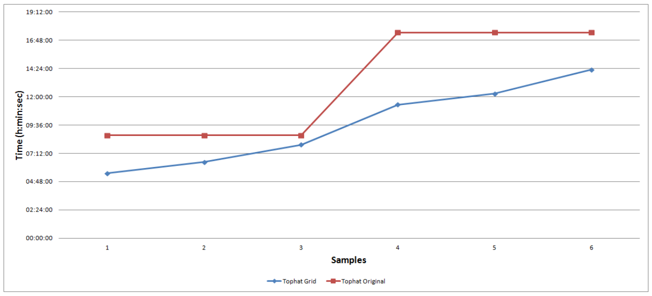

Figure 6. Original TopHat vs TopHat Grid.

References

[1] De Magalhes J.P., Finch C.E., Janssens G., *Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions.*, Ageing Research Reviews, 2010 Jul;9(3):315-23.

[2] Sanger F., Nicklen S., Coulson A.R., *DNA sequencing with chain-terminating inhibitors*, Proc. Natl. Acad. Sci. USA 74 (1977) 54635467.

[3] Kircher M., Kelso J., *High-throughput DNA sequencing concepts and limitations.*, Bioessays. 2010 Jun;32(6):524-36. Review.

[4] Maher C.A., Palanisamy N., Brenner J.C., Cao X., Kalyana-Sundaram S., Luo S, Khrebtukova I., Barrette T.R., Grasso C., Yu J., Lonigro R.J., Schroth G., Kumar-Sinha C., Chinnaiyan Y., *Chimeric transcript discovery by paired-end transcriptome sequencing*, AM. Proc Natl Acad Sci U S A. 2009 Jul 28

[5] Pop M., Salzberg S.L., *Bioinformatics challenges of new sequencing technology*, Trends Genet. 24, 2008

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

[7] Trapnell C., Pachter L., Salzberg S.L. *TopHat: discovering splice junctions with RNA-Seq*. Bioinformatics doi:10.1093/bioinformatics/btp120

[8] Langmead B., Hansen K., Leek J. *Cloud-scale RNA-sequencing differential expression analysis with Myrna* Genome Biology 11:R83

[9] Berman F., Fox G., Hey A.J.G. *Grid Computing Making the Global Infrastructure a Reality*, Wiley, 2005

[10] Kurowski K., Nabrzyski J., Oleksiak A., Weglarz J. *Scheduling jobs on the Grid-Multicriteria approach*, Computational Methods in Science and Technology, 12(2), 123-138, 2006