SAMasGC: Sequencing Analysis with a Multiagent System and Grid Computing

Roberto González, Carolina Zato, Rocío Benito, María Hernández, Jesús M. Hernández, and Juan F. De Paz

Abstract. Advances in bioinformatics have contributed towards a significant increase in available information. Information analysis requires the use of distributed computing systems to best engage the process of data analysis. This study proposes a multiagent system that incorporates grid technology to facilitate distributed data analysis by dynamically incorporating the roles associated to each specific case study. The system was applied to genetic sequencing data to extract relevant information about insertions, deletions or polymorphisms.

Keywords: Multiagent system, Grid Computing, genetic sequencing, distributed computing, bioinformatics.

1 Introduction

Advances in genetic sequencing have made it possible to carry out massive sequencing and to perform studies on genetic sequencing in multiple patients [2] [3]. One area of medicine in its height of development and fundamental in the application of techniques that facilitate the automatic treatment of data and the extraction of knowledge is genomics. This increase in information has made it necessary to create systems that can perform a distributed analysis of information and be

Roberto González · Carolina Zato · Juan F. De Paz
Department of Computer Science, University of Salamanca
Plaza de la Merced, s/n, 37008, Salamanca, Spain
e-mail: {rgonzalezramos,carol_zato,fcofds}@usal.es

Rocío Benito · María Hernández · Jesús M. Hernández
IBMCC, Cancer Research Center, University of Salamanca-CSIC, Spain
e-mail: {beniroc,jhmr}@usal.es, mahesa2504@hotmail.com

Jesús M. Hernández
Servicio de Hematología, Hospital Universitario de Salamanca, Spain

M.P. Rocha et al. (Eds.): 6th International Conference on PACBB, AISC 154, pp. 209–216.
springerlink.com

adapted to different types of analysis such as with genetic sequencing. The development of distributed applications and parallel computing currently requires the use of complex software and libraries [7], while some, such as MPI and CUDA [8], even use combinations of libraries. It has become necessary to develop systems that facilitate the creation of distributed systems that allow the distributed implementation and execution of different types of analyses+ by applying technologies such as grid computing [6].

The rise in bioinformatics, primarily following the emergence of microarrays and sequencing in particular, has made it possible to generate great volumes of information [5]. With the appearance of expression arrays, specifically BAC arrays and more importantly Exon arrays [5], it became necessary to create systems that would allow the distributed analysis of information to improve the output of algorithms. The use of NGS (next generation sequencing) has noticeably increased the amount of information, which has in turn led to an improvement in output and a reduction in execution time. As a result, it has become necessary to create systems that facilitate the management of distributed systems. These systems must facilitate the creation of algorithms that are executed in a distributed way, which enables the dynamic generation of control flows.

This study proposes the use of multiagent systems [9] capable of distributed execution performed by the integration of grid technology [6]. The multiagent system integrates agents to manage the roles in the case study that is being carried out. Within the context of this study, the proposed system focuses on detecting relevant patterns and mutations, insertions, deletions or polymorphisms, within the sequence data taken from patient samples provided by the Cancer Institute of the University of Salamanca. The analysis of sequencing data requires various types of processes: i) assembly [4] ii) alignment [4] and iii) knowledge extraction [5] in order to analyze sequence data. The Cancer Institute of the University of Salamanca is striving to develop tools to automate the evaluation of data and to facilitate the analysis of information. This proposal is a step forward in this direction and the first step toward the development of a multiagent system.

This article is structured as follows: section 2 reviews the state of the art in genetic sequencing; section 3 presents the proposed architecture and adapts the architecture to the case study; section 4 presents the results and conclusions.

2 Massive Analysis and Sequencing

Sequencing began in the 60s, although it was not until the 80s and the Sanger method [4] that gene and genome sequencing emerged. The sequencing process was a laborious manual process; following the development of automated sequencing in the late 80s the volume of information increased dramatically. The process of separating DNA fragments with automated sequencing was initially performed with gel electrophoresis [1], subsequently replaced by capillary electrophoresis [1], after which pyrosequencing [1] was developed. There are currently various types of NGS with different capabilities in base pairs. Zhang et al. [4] describe the different manufacturers. The length of the fragments of the base pairs can vary according to the sequencing used, from 25 bp to the 500 bp used

with sequencing by the Roche company, which can perform de-novo sequencing [4]. In the near future, the length of sequenced base pairs is expected to increase considerably; in fact, new research in techniques has developed SMRT (single molecule real time) sequencing, which can achieve 10,000 bp, facilitating the processes of new genome assembly and sequencing.

The human genome is estimated to be about 3,000 million base pairs long and contain around 25,000 genes [10]. Consequently, sequencing genome fragments of 500 bp at a time is costly and requires computational techniques that can join contig fragments to generate the complete genome. Sequencing is not usually applied to just any part of the genome; instead, specific exon sequences corresponding to the DNA code are selected. Exons are the part of the DNA that is represented in the messenger RNA. The regions that are transcribed in the messenger RNA can later be converted into proteins [11], hence the relevance of its analysis and the detection of variants.

3 Proposed SMASasGC Architecture

Bioinformatic data analysis requires different processes and algorithms that vary according to the data recovered. Nevertheless, these different types of analyses all share a common characteristic, namely the high computing cost involved. It has become necessary to develop systems that facilitate the efficient development of distributed systems. In order to analyze distributed data in an efficient manner, grid technology [6] and multiagent systems [9] are integrated to produce a high-performance hybrid architecture. The architecture created for this study contains three separate layers: the coordination layer contains agents assigned to maintain the algorithms specific to the case study; the control layer contains agents responsible for controlling the grid; and the specialization layer contains the agents and the processes specific to the case study.

Figure 1 displays the agents that correspond to the coordination and control layer. These layers are independent of the case study, allowing their functionality to be reused.

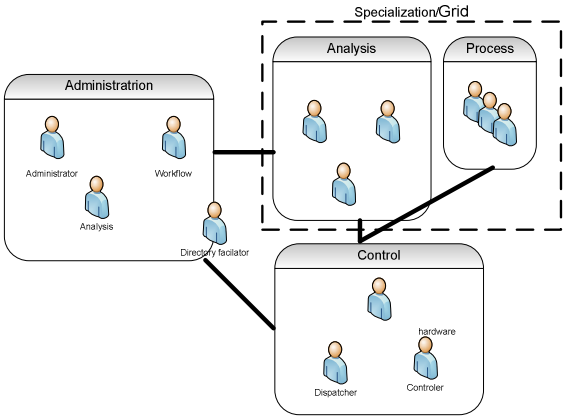

Fig. 1 Architecture for system coordination and control

The coordination layer includes the administrator, analysis, workflow and directory facilitator agents. The administrator agent is in charge of storing and controlling project data for their subsequent analysis. Each project contains information about the flow of analysis and the results obtained by the applied algorithms. Additionally administrator agent includes all associated roles including the status of the project process, launched tasks, and the type of data associated with the different case studies. The analysis agent is in charge of recovering previous flows of analysis and storing new flows of analysis that may be used to recommend flows of execution with similar data. The workflow agent is responsible for creating new flows of execution based on existing algorithms for different types of analyses. Finally, the directory facilitator (DF) stores and administers existing algorithms for each of the possible types of analysis; its functionality is similar to that of a web services DF.

The control layer includes the agents responsible for controlling the state of execution for the grid. The dispatcher, hardware and controller agents are available in this layer. The dispatcher agent is in charge of recovering and distributing tasks between the nodes. The hardware agent controls machine resources available on the grid. The controller agent controls the machine load level to control the state of execution for the machines containing grid.

The specialization layer is composed of agents and processes that are executed in the grid nodes. These processes and agents are specific to each case study and are responsible for defining the hardware needs for their execution, and breaking the tasks down into subtasks that are sent to the dispatcher.

3.1 Sequence Analysis

The process of sequence data analysis varies according to the results that one wants to obtain. It normally requires a process of assembly, alignment and knowledge extraction to automatically process the data. As the architecture must be specific to this end, agents and processes specialized in performing these tasks are required. The agents are responsible for establishing the restrictions and procedures for distributing tasks along the grid nodes according to available resources.

The specialization layer in this case study was composed of the following agents: assembly, alignment and knowledge extraction. Each agent defines the following roles for the purpose of carrying out the task for which they were added to the system: manage available algorithms to execute the task; manage the resources needed to apply each algorithm; determine the preconditions for executing tasks; manage the nodes required to execute the tasks; break the tasks down into subtasks that are subsequently queued in the dispatcher.

Each agent in this stage has various processes that are executed through grid in a distributed manner. The processes can vary according to the algorithm that is selected in the work flow of the project created in the analysis. Thus, the algorithms developed for each stage of the case study are as follows:

3.1.1 Assembly

The assembly process varies according to the size of each reading that is used. In this particular case, we chose to use the algorithm provided by the manufacturer of the sequencer being used. Different assembly algorithms can be seen in [4] and [16]. Roche provides the Newbler assembler [15], which was used in this study.

3.1.2 Alignment

The alignment process consists of establishing the fragment of the reference genome that is most similar to the fragment of the patient being treated. The alignment algorithms are applied to different fields in addition to bioinformatics.

While there are many different ways to carry out the alignment process, performance is ultimately the most important factor. The alignment algorithms used are local, since the sequence to be aligned, or the contigs, is smaller in size than the reference genome. Local alignments are based on the Smith-Waterman algorithm [12]. The alignments can be given in pairs or groups according to the number of fragments that must be analyzed simultaneously. There are currently many alignment algorithms, but the most commonly used are BLAST (Basic Local Alignment Search Tool) [13] and BLAT (BLAST-Like Alignment Tool) [14]. BLAT can perform an alignment faster than BLAST, but it cannot ensure that the final alignment is the best one possible, although performance is greatly improved. Additionally, there are many algorithms that can be found in different review articles such as [4] and [16].

3.1.3 Extraction of Knowledge

The classification algorithms can be divided in: decision trees, decision rules, probabilistic models, fuzzy models, based on functions, ensemble. During the extraction of knowledge phase, different analyses of the variations were performed to detect the following types of alterations: SNP, point mutation not SNP, insertions and deletions. The process of detecting SNP and other point mutations is simple since it only involves searching information in databases that contain the previously published information.

4 Results and Conclusions

Genetic sequencing was applied to a data set taken from patients with leukemia. Specific genes were sequenced from a total of 8 patients, each of whom had approximately 110,000 sequence fragments that corresponded to the regions relevant to this study. The sequenced fragments vary in length for the different patients, as shown in figure 2.

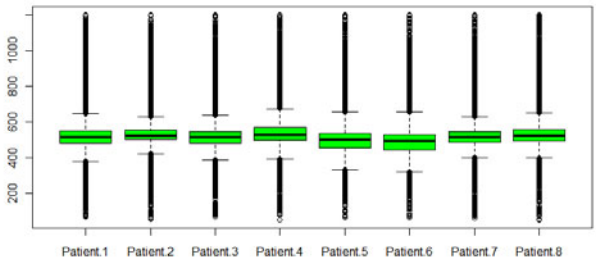

Fig. 2 Box diagram of the lengths of the fragmented segments.

The version of the reference genome used in this study corresponds to HG18; this is because the information used as a reference in selecting the sequence regions was obtained in previous studies using the same version.

The system was tested to validate the generic architecture proposed and to validate the system's capacity for analyzing the proposed case study. The main goal in validating the architecture was to determine the efficiency of the system and its ability to distribute tasks according to the existing nodes. To validate the increase in performance, we selected the Newbler, BLAT. Figure 3 demonstrates the size of the contigs that were assembled once the Newbler algorithm was applied and executed on the grid.

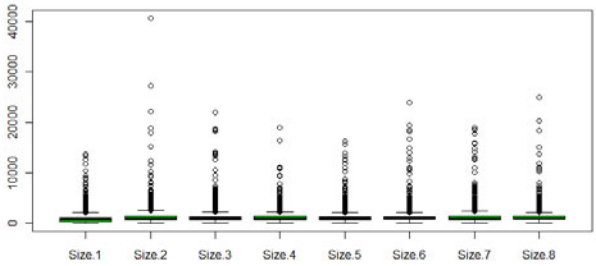

Fig. 3 Box diagram with the length of the assembled contigs

The average sizes for each of the aligned patients are as follows: 1023.134 1356.855 1292.022 1251.492 1239.192 1306.707 1302.321 1362.151.

After completing the assembly process, the fragments were aligned using the BLAT algorithm, which obtained PSL output files. The specific files used to predict the alterations are the following: matches, misMatches, repMatches, nCount, qNumInsert, qBaseInsert, tNumInsert, tBaseInsert, qSize, qStart, qEnd, tSize, tStart, tEnd, blockCount, blockSizes. The analyzed alterations are insertions, deletions and polymorphisms, in order to detect these alterations the system used the SNP130 table from UCSC. The final number of alterations in the patient 1 are shown in table 1, the number of unknown alterations (insertions and deletions) is high due to the pathology and the analyzed regions.

Table 1 Variants

Initial Contigs	Contigs with variants	SNPs	Unknown variants
1310	1100	6584	2021

The next step was to analyze the execution time in order to observe the system's scalability according to the number of nodes on which the processes were dropped. The execution time decrease linearly with the number of nodes in the system. The table 2 shows the execution times in seconds for several nodes.

Table 2 Time in seconds as the number of nodes

	1 Node	2 Nodes	3 Nodes
Segmentation	691s	405s	241s
Polymorphism	2340s	1145s	710s

The multiagent system has made it possible to integrate algorithms that can adapt to a specific case study, facilitating the distributed execution of work flows. The system facilitates the integration of algorithms for different case studies and reduces the execution time in an efficient manner, so long as it remains possible to improve performance by separating tasks for their more effective execution in grid technology.

Acknowledgments. This work has been supported by the MICINN TIN 2009-13839-C03-03.

References

[1] Mitchelson, K.R., Hawkes, D.B., Turakulov, R., Men, A.E.: Chapter Overview: Developments in DNA Sequencing. Perspectives in Bioanalysis 2, 3–44 (2007)

[2] Soo, R.A., Wang, L.Z., Ng, S.S., Chong, P.Y., Yong, W.P., Lee, S.C., Liu, J.J., Choo, T.B., Tham, L.S., Lee, H.S., Goh, B.C., Soong, R.: Distribution of gemcitabine pathway genotypes in ethnic Asians and their association with outcome in non-small cell lung cancer patients. Lung Cancer 63(1), 121–127 (2009)

[3] Esteban, F., Royo, J.L., González-Moles, M.A., Gonzalez-Perez, A., Redondo, M., Moreno-Luna, R., Rodríguez-Sola, M., Gonzalez, A., Real, L.M., Ruiz, A., Ramírez-Lorca, R.: CAPN10 alleles modify laryngeal cancer risk in the Spanish population. European Journal of Surgical Oncology (EJSO) 34(1), 94–99 (2008)

[4] Zhang, J., Chiodini, R., Badr, A., Zhang, G.: The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics 38(3), 95–109 (2011)

[5] Corchado, J.M., De Paz, J.F., Rodríguez, S., Bajo, J.: Model of experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine 46(3), 179–200 (2009)

[6] Gregoretti, F., Laccetti, G., Murlib, A., Olivaa, G., Scafuri, U.: MGF: A grid-enabled MPI library. Future Generation Computer Systems 24(2), 158–165 (2008)

[7] Jin, H., Jespersen, D., Mehrotra, P., Biswas, R., Huang, L., Chapman, B.: High performance computing using MPI and OpenMP on multi-core parallel systems. Parallel Computing (in Press)
[8] Rakić, P.S., Milašinović, D.D., Živanov, Ž., Suvajdžin, Z., Nikolić, M., Hajduković, M.: MPI–CUDA parallelization of a finite-strip program for geometric nonlinear analysis: A hybrid approac. Advances in Engineering Software, 42(5), 273–285 (2011)
[9] Wooldridge, M.: Introduction to Mult iAgent Systems. John Wiley & Sons (2002)
[10] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature 431, 931–945 (2004)
[11] Sahua, S.S., Pand, G.: dentification of Protein-Coding Regions DNA Sequences Using A Time-Frequency Filtering Approach. Genomics, Proteomics & Bioinformatics 9(1-2), 45–55 (2011)
[12] Khajeh-Saeed, A., Poole, S., Perot, J.B.: Acceleration of the Smith–Waterman algorithm using single and multiple graphics processors. Journal of Computational Physics 229(11), 4247–4258 (2010)
[13] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990)
[14] Kent, W.J.: BLAT—the BLAST-like alignment tool. Genome. Res. 12, 656–664 (2002)
[15] Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M.: Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380 (2005)
[16] Miller, J.R., Koren, S., Sutton, G.: Assembly algorithms for next-generation sequencing data. Genomics 95(6), 315–327 (2010)