Chapter 4.8
Functional Genomics Applications in GRID

Luciano Milanesi
Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche, Italy

Ivan Merelli
Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche, Italy

Gabriele Trombetti
Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche, Italy

Paolo Cozzi
Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche, Italy

Alessandro Orro
Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche, Italy

ABSTRACT

A common ongoing task for Functional Genomics is to compare full organisms' genome with those of related species, to search in huge database for functional annotation of novel sequences and to identify specific patterns of them, such as ESTs, genes, and microRNA. The prediction of these patterns has a relevant computational cost, while public genome archives exceed one billion sequence traces from over 1,000 organisms and this number is increasing rapidly as costs decline, but powerful solution must be enabled in order to perform efficient searches. This means that Functional Genomics applications require significant computational infrastructures, where reusable tools and resources can be accessed. In particular, grid computing seems to fulfill both the computational and data management requirements, even if porting applications on this infrastructure can be difficult. The implementation of a suitable environment for the management of distributed computations can provide reliable advantage, reducing the gap between the requirements of the functional genomic domain and the potential of this technology.

DOI: 10.4018/978-1-4666-0879-5.ch4.8

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

The Human Genome Project was just the first step in understanding humans at the molecular level. In the post genomic era, many questions still remain unanswered, including functions, regulations and interactions of the estimated 30,000 human genes. Moreover, breakthrough technologies such as the High Throughput ones allow, nowadays, the fast generation of enormous amounts of data related to different organisms. This is particularly true for the sequencing, where Roche and Illumina technologies enable the possibility of sequencing many bacterial genomes and even the production of multiple individual references, which allow reconsidering statistically the features of the genetic sequences.

The bottleneck of these studies is now related to the computational scalability of the downstream analysis. Researchers must work through the genome assemble, the prediction of genes and microRNA and the annotation of sequences against the available databases. Then, much work needs to be accomplished to understand the role of the expressed sequence tag (ESTs), which are short sub-sequences of a transcribed spliced nucleotide sequence, and of microRNA, which are single-stranded RNA molecules, which regulate gene expression.

The genome-wide analysis of gene and the related translated proteins, using this functional perspective, is usually referred to as Functional Genomics. In particular, this field of molecular biology attempts to make use of the vast wealth of data produced by genomic projects to describe gene and protein functions, regulations and interactions. In other words, Functional Genomics uses mostly high-throughput techniques to characterize the abundance gene products and clearly it deals with huge quantity of data from which is difficult to turn into real knowledge.

As a consequence, the merging between molecular biology and computer science fields is becoming a key point to mine information from the huge amount of data produced in laboratories. A computer science approach plays a crucial role to deal with the complexity of this scenario and in particular High Performance Computing is suitable to face these challenges. The exploitation of parallel computing and distributed solutions through the porting of bioinformatics applications and developing new systems on computer clusters and grid infrastructure improves performance when addressing genome wide analysis.

In this chapter we are going to discuss an infrastructure we developed to enable the management of biological large scale computations on the EGEE projects grid implementation, which relies on the *gLite* middleware. This infrastructure is useful to manage data on the grid, in terms of database replications and updating and for the submission and monitoring of the jobs. The idea is to provide an example of how a common infrastructure can be used on the top of a grid computing infrastructure to deal with different computation in the context of the Functional Genomics. The background section concerns a general introduction to grid computing technologies and some details related to the specific grid infrastructure employed in this work. In the next section a presentation of the bioinformatics software tested in our framework is provided. The core of the discussion is about the infrastructure developed to deal with large scale challenges, both in terms of computational management and data handling. The last couple of sections present respectively some challenges performed with this infrastructure and the performance figure obtained with a detailed description of the problems that researchers can encounter in using this platform.

BACKGROUND

Grid technology is a very important step forward from the Web, which simply allows the sharing of information over the internet. This paradigm of distributed computing aims to promote the

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/functional-genomics-applications-grid/64521

Related Content

Data Mining in Proteomics Using Grid Computing
Fotis Psomopoulosand Pericles Mitkas (2012). *Grid and Cloud Computing: Concepts, Methodologies, Tools and Applications (pp. 918-940).*
www.irma-international.org/chapter/data-mining-proteomics-using-grid/64522

An Effective Volleyball Trajectory Estimation and Analysis Method With Embedded Graph Convolution
Guanghui Huang (2023). *International Journal of Distributed Systems and Technologies (pp. 1-13).*
www.irma-international.org/article/an-effective-volleyball-trajectory-estimation-and-analysis-method-with-embedded-graph-convolution/317936

A Comparative Study of Range-Free and Range-Based Localization Protocols for Wireless Sensor Network: Using COOJA Simulator
Essa Qasem Shahra, Tarek Rahil Sheltamiand Elhadi M. Shakshuki (2017). *International Journal of Distributed Systems and Technologies (pp. 1-16).*
www.irma-international.org/article/a-comparative-study-of-range-free-and-range-based-localization-protocols-for-wireless-sensor-network/171979

Task-Based Crowd Simulation for Heterogeneous Architectures
Hugo Perez, Benjamin Hernandez, Isaac Rudominand Eduard Ayguade (2016). *Innovative Research and Applications in Next-Generation High Performance Computing (pp. 194-219).*
www.irma-international.org/chapter/task-based-crowd-simulation-for-heterogeneous-architectures/159045

Efficient Resource Allocation Mechanism for Federated Clouds
Chien-Yu Liu, Kuo-Chan Huang, Yi-Hsuan Leeand Kuan-Chou Lai (2015). *International Journal of Grid and High Performance Computing (pp. 74-87).*
www.irma-international.org/article/efficient-resource-allocation-mechanism-for-federated-clouds/141358