

## Chapter 4.4

# High-Throughput GRID Computing for Life Sciences

**Giulia De Sario**

*Istituto di Tecnologie Biomediche, CNR, Italy*

**Angelica Tulipano**

*Istituto di Tecnologie Biomediche, CNR, Italy*

**Giacinto Donvito**

*INFN, Sezione di Bari, Italy*

**Giorgio Maggi**

*INFN Bari, Italy & Università e Politecnico di Bari, Italy*

**Andreas Gisel**

*Istituto di Tecnologie Biomediche, CNR, Italy*

### **ABSTRACT**

*The number of fully sequenced genomes increases daily, producing an exponential explosion of the sequence, annotation and metadata databases. Data analysis on a genome-wide level or investigation within a specific data repository has become a data- and calculation-intensive process occupying single computers and even larger computer clusters for month or even years. In most cases such applications can be subdivided into many independent smaller tasks. The smaller tasks are particularly suited to distribution over a computational GRID infrastructure, which drastically reduces the time to reach the final result. In our analysis of gene ontology data and their associations to gene products of any kind of organism in a search to find gene products with similar functionalities, we developed a system to divide the full search into a large number of jobs and to submit these jobs to the GRID infrastructure as long as all jobs are processed successfully, guaranteeing an analysis of the data without missing any information.*

DOI: 10.4018/978-1-4666-0879-5.ch4.4

## INTRODUCTION

Data analysis in bioinformatics—due to the drastically high rate of increase in the sheer volume of data, not only in size but also in diversity—is becoming a very complex and data-intensive procedure occupying large numbers of computer units which often, for a typical user under normal condition, are not available. Sequencing projects all over the world, including high-throughput approaches such as the microarray technology and next-generation sequencing or the large scale mass spectrometry analysis, are responsible for this exponential growth of complex biological data sets. Data analysis within such projects and even more complex projects, such as comparisons and integration processes between such projects, often involves the examination of several big data sets with sizes on the order of hundreds of gigabytes. Fortunately, many of these analyses can be divided into many small tasks, producing the possibility of distributing the workload on an infrastructure such as the computational GRID. However, when the number of jobs necessary to carry on a particular analysis becomes huge, controlling the full production is not a simple enterprise. It is very important to carefully monitor each job, watching the success of each submitted job in order to be able to complete the full analysis without any missing data.

The biological task we are describing in this chapter is the comparison of gene products in a new, non-conventional way to find gene products with similar functionality. Usually gene products are compared by aligning sequences and looking for sequence similarity with the assumption that a high similarity corresponds to similar functionalities (Skolnick and Fetrow 2000). However, the relation sequence-function is not always true and often only small differences in the sequence may result in drastic changes in functionality. Those sequence differences are hardly detectable within a conventional sequence alignment.

Further, several gene products can have similar functionality but the active site or the conformation can be absolutely different, originating from an absolutely different Abstract

*The number of fully sequenced genomes increases daily, producing an exponential explosion of the sequence, annotation and metadata databases. Data analysis on a genome-wide level or investigation within a specific data repository has become a data- and calculation-intensive process occupying single computers and even larger computer clusters for month or even years. In most cases such applications can be subdivided into many independent smaller tasks. The smaller tasks are particularly suited to distribution over a computational GRID infrastructure, which drastically reduces the time to reach the final result. In our analysis of gene ontology data and their associations to gene products of any kind of organism in a search to find gene products with similar functionalities, we developed a system to divide the full search into a large number of jobs and to submit these jobs to the GRID infrastructure as long as all jobs are processed successfully, guaranteeing an analysis of the data without missing any information.*

For this reason we propose a different approach to finding gene products which have similar functionalities. In most sequence databases there are keywords describing the functionality and localisation on a cellular level of the gene product they harbour. Those keywords and other functional annotations in other biological databases were converted to a standardised vocabulary describing the gene products on the level of the molecular functions and biological processes with which they are involved and the cellular components where they are localised, i.e., the gene ontology (GO (Ashburner, Ball et al. 2000)). The GO consortium (GOC) took the lead on this effort, so that we now have a repository of more than 3,6 million gene products described by more than

18 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/high-throughput-grid-computing-life/64517](http://www.igi-global.com/chapter/high-throughput-grid-computing-life/64517)

## Related Content

---

### Securing Real-Time Interactive Applications in Federated Clouds

Michael Boniface, Bassem Nasser, Mike Surrigeand Eduardo Oliveros (2012). *Achieving Real-Time in Distributed Computing: From Grids to Clouds* (pp. 160-173).

[www.irma-international.org/chapter/securing-real-time-interactive-applications/55247](http://www.irma-international.org/chapter/securing-real-time-interactive-applications/55247)

### Harnessing the Cloud for Mobile Social Networking Applications

Juwel Rana, Josef Hallberg, Kåre Synnesand Johan Kristiansson (2012). *Evolving Developments in Grid and Cloud Computing: Advancing Research* (pp. 1-11).

[www.irma-international.org/chapter/harnessing-cloud-mobile-social-networking/61979](http://www.irma-international.org/chapter/harnessing-cloud-mobile-social-networking/61979)

### Architecture Exploration Based on Tasks Partitioning Between Hardware, Software and Locality for a Wireless Vision Sensor Node

Muhammad Imran, Khursheed Khursheed, Abdul Waheed Malik, Naeem Ahmad, Mattias O'Nils, Najeem Lawaland Benny Thörnberg (2012). *International Journal of Distributed Systems and Technologies* (pp. 58-71).

[www.irma-international.org/article/architecture-exploration-based-tasks-partitioning/66057](http://www.irma-international.org/article/architecture-exploration-based-tasks-partitioning/66057)

### Accelerating Training Process in Logistic Regression Model using OpenCL Framework

Hamada M. Zaheraand Ashraf Bahgat El-Sisi (2017). *International Journal of Grid and High Performance Computing* (pp. 34-45).

[www.irma-international.org/article/accelerating-training-process-in-logistic-regression-model-using-opencl-framework/185772](http://www.irma-international.org/article/accelerating-training-process-in-logistic-regression-model-using-opencl-framework/185772)

### Service Quality Model Evaluation

Fotis Aisopos, Magdalini Kardara, Vrettos Moulos, Athanasios Papaoikonomou, Konstantinos Tserpesand Theodora Varvarigou (2012). *Achieving Real-Time in Distributed Computing: From Grids to Clouds* (pp. 59-73).

[www.irma-international.org/chapter/service-quality-model-evaluation/55242](http://www.irma-international.org/chapter/service-quality-model-evaluation/55242)