Chapter VII
High Performance BLAST
Over the Grid

Vincent Breton
IN2P3, CNRS, Clermont-Ferrand, France

Eddy Caron
Université de Lyon, LIP, CNRS-ENS-Lyon-UCBL-INRIA, France

Frédéric Desprez
INRIA, Université de Lyon, LIP, CNRS-ENS-Lyon-UCBL-INRIA, France

Gaël Le Mahec
CNRS, IN2P3, UBP Clermont-Ferrand, France

ABSTRACT

As grids become more and more attractive for solving complex problems with high computational and storage requirements, bioinformatics starts to be ported on large scale platforms. The BLAST kernel, one of the main cornerstone of high performance genomics, was one the first application ported on such platform. However, if a simple parallelization was enough for the first proof of concept, its use in production platform needed more optimized algorithms. In this chapter, we review existing parallelization and “gridification” approaches as well as related issues such as data management and replication, and a case study using the DIET middleware over the Grid’5000 experimental platform.

INTRODUCTION

Genomics acquiring programs, such as full genomes sequencing projects, are producing larger and larger amounts of data. The analysis of these raw biological data requires very large computing resources. Functional sites and signatures of proteins are very useful for analyzing these
data or for correlating different kinds of existing biological data. These methods are applied, for example, to the identification and characterization of the potential functions of new sequenced proteins, and to the clusterization into protein families of the sequences contained in international databanks. The sites and signatures of proteins can be expressed by using the syntax defined by the PROSITE databank, and written as a “protein regular expression”. Searching one such site in a sequence can be done with the criterion of the identity between the searched and the found patterns. Most of the time, this kind of analysis is quite fast. However, in order to identify non perfectly matching but biologically relevant sites, the user can accept a certain level of error between the searched and the matching patterns. Such an analysis can be very resource consuming. Parallel computing (Zomaya, 2006) and Grids (Krishnan, 2004) are then obvious responses to performance issues.

Genomics involving BLAST (Basic Local Alignment Search Tool) applications (Altschul, Gish, Miller, Myers & Lipman, 1990) (Korf, Yandell, & Bedell, 2003) are the common tool for searching sequences similarities within protein and DNA databases. The BLAST is used to search homologies between amino-acids or nucleotides sequences. It quantifies similarities using some sensibility parameters and a similarity score matrix. Sequences are known as similar if they have enough common characters and not too many differences, which can be substitutions, deletions or insertions. The search for similarities between sequences can be done at a global or local level. At the global level, we search for the best alignment between two sequences permitting large different areas inside the sequences (Needleman, & Wunsch, 1970). At the local level, we search subsequences alignments, not using the subsequences without any alignment for the total similarity score computing (Goad, & Kanehisa, 1982) (Smith, & Waterman, 1981).

However, as the size of available databases increases (now up to hundreds of Gigabytes) (NCBI. Ncbi website. http://www.ncbi.nih.gov/), the search of many sequences is time consuming. Many researches have been done to get faster algorithms. Their goal is to reduce the search scope while looking at high scoring alignments. Heuristics have been produced and among them the BLAST is the most popular. Still, as databases size and number still increase, the use of parallel computing becomes mandatory. Several approaches exist from parallelizing the BLAST kernel itself (and distributing the databases) to splitting the input sequences to different servers and replicating the databases. All these approaches can be combined to get the best performance.

The rest of this chapter is organized as follow. The next section will be devoted to a short presentation of Bioinformatics Grids around the world. Then, in the third section, we present the BLAST kernel and different parallelization levels. In the fourth section, we present some algorithms and heuristics used for the joint replication and scheduling problem which is used for the gridification of the BLAST kernel. The next section presents middleware used for bioinformatics application over the Grid. Finally, before a conclusion, we describe a test case of the deployment of a BLAST application over DIET, a middleware implementing the GridRPC API of the Open Grid Forum (OGF).

**BACKGROUND**

Several projects around the world aim at setting Bioinformatics Grids. These projects are built upon de facto standards like Globus (Globus. http://www.globus.org) or Condor (Frey, Tannenbaum, Livny, Foster, & Tuecke, 2002).

The main challenge for data analysis in life sciences is to offer to the molecular biologists an integrated and up-to-date view of an exponentially growing volume of data in a multiplicity