Available online at www.sciencedirect.com
ScienceDirect
Future Generation Computer Systems 24 (2008) 752–762
FGCS FUTURE GENERATION COMPUTER SYSTEMS
www.elsevier.com/locate/fgcs

A Bioinfomatics Grid Alignment Toolkit

Maria Mirto*, Sandro Fiore, Italo Epicoco*, Massimo Cafaro, Silvia Mocavero, Euro Blasi, Giovanni Aloisio

University of Salento, Lecce & SPACI Consortium, Italy
Euro Mediterranean Centre for Climate Change, Italy

Received 16 February 2007; received in revised form 6 January 2008; accepted 7 February 2008
Available online 16 February 2008

Abstract

Even though many useful tools for sequence alignment are available, such as BLAST and PSI-BLAST by NCBI and FASTA by the University of Virginia, a key issue regarding sequence databases is their size, growing at an exponential rate. Grid and parallel computing are crucial techniques to maintain and improve the effectiveness of sequence comparison tools, whilst the Web Services approach may guarantee the interoperability among large collections of programs and data. This paper describes BioGAT, Bioinformatics Grid Alignment Toolkit, that offers optimized brokering and a data management system to exploit various bioinformatics alignment tools wrapped as Web Services in a Grid architecture.

Keywords: Bioinformatics; Sequence alignment tools; Data management; Grid computing; Computational Grid; Web services

1. Introduction

A key issue in sequence analysis is the determination of similarity between biological sequences, that is the percentage of sequence matches among nucleotide or protein sequences.

The hypothesis is that similarity relates to functionality: if two sequences are similar, they will have related functionalities. Therefore, the idea is to match sequences of proteins with unknown functions with those of known functions. Only similarity matching makes sense when searching among sequences.

The similarity among sequences is measured through the determination of the optimal alignment of (parts of) sequences, that is the attribution of a score: the higher the score, the higher the similarity between the sequences.

The comparison of two sequences is named *pairwise alignment* (PA).

The sequence alignment is the basis of more complex analysis such as the search for sequence similarities in databases, phylogenetic investigation and the assignment of gene/protein functions.

In bioinformatics, similarity must not be confused with homology. Indeed two or more sequences are homologue if they have a unique ancestor, from which they evolved. It is a qualitative feature that indicates a common phylogenetic origin whereas similarity is a quantitative feature: it measures how the sequence matches on the basis of a comparative approach. Homology implies similarities, but the converse is not true.

In this scenario, another very important approach is the *multiple sequence alignment* (MSA) where many homologue sequences are matched to identify functional domains. Multiple alignment between homologue sequences is more effective with respect to pairwise alignment for studying regions that during the evolution are conserved and that characterize, with good probability, the biological functionality of the sequence. Hence, pairwise and multiple alignment are basic tools in bioinformatics. In particular, there are several applications such as BLAST (Basic Local Alignment Search Tool) [1] and FASTA [2] for PA, and PSI-BLAST [3] for MSA, that offer flexible searching through boatloads of data.

However, these tools are slow if the alignment of many sequences is required.

* Corresponding author at: Centre for Advanced Computational Technologies, University of Salento, Lecce, Italy. Tel.: +39 0832297371; fax: +39 0832297279.
E-mail addresses: maria.mirto@unile.it (M. Mirto), italo.epicoco@unile.it (I. Epicoco).

0167-739X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2008.02.001

Indeed, sequence databases are exploding in size, growing at an exponential rate, currently doubling in about 12 months, exceeding the rate of growth in compute cycles that doubles only every 18 months (Moore's law) [4]. Whenever sequence databases are significantly updated, alignment should be repeated many times to discover new information. It is worth noting here that although clusters are utilized in several laboratories, an increasing computing power is continuously needed to support these operations and provide results in acceptable times.

Thus, the main requirements to exploit the growth in sequence databases without acquiring costly hardware are: (i) the alignment algorithms should not be modified; (ii) avoiding the use of costly multiprocessor machines; (iii) the ability to utilize cheaper hardware such as clusters.

Grid computing [5] together with parallel alignment tools, wrapped as Web Services [6], are crucial techniques to maintain and improve the effectiveness of sequence comparison tools. Partitioning the load into different jobs for each simulation is a good choice for the alignment applied to a large dataset, because the input data, composed by a set of sequences (generally in FastA format), are compared with biological databases that are made of a set of sequences in various formats (flat files and FastA). Therefore, it is possible to parallelize the execution by splitting both the input dataset and/or the database, sending each data partition to a Grid node, and merging the results. The solution proposed in this paper is based on the possibility to manage the untapped processing power of desktop PCs within an enterprise Grid network to process computationally intensive jobs for scientific applications and in particular for the bioinformatics domain.

It is worth noting here that there are other works in the bioinformatics area that use this approach such as BOINC [7] and related projects, and the Condor Project [8], which pioneered, in the Grid computing field, the idea of using the idle time of organizational workstations to do high-throughput computing. However, those systems are not optimized for sequence alignment based on legacy bioinformatics tools. Moreover, because distributed alignment tasks are not completely independent, a "simple" merge of the results is not enough. Indeed, several parameters must be modified on the basis of the dimension of the database used to search for the sequences to be aligned, and the input sequence.

Finally, the tools are run on biological data banks that must be indexed with specific software before starting the computation, otherwise these tools will not work. Therefore, customized components such as merging, indexing tools and format translation filters are needed to support alignment.

In this paper we describe BioGAT, Bioinformatics Grid Alignment Toolkit, which aims at providing a Grid framework to support parallel alignment tools execution. In particular we have chosen the BLAST, PSI-BLAST and FASTA tools because they represent a "de facto" standard for sequence alignment. Our solution uses a Grid platform based on the Globus Toolkit 4 [9] in which each module is deployed as a Web Service.

The proposed system allows splitting the computation on a biological dataset among several computational nodes. An important advantage is that this approach (fine grained) is different from others, such as the "simple job farming" (coarse grained) currently used by several Grid projects.

BioGAT has been developed as part of the ProGenGrid (Proteomics and Genomics Grid) project [10] jointly with GRelC (Grid Relation Catalog) project [11] at the University of Salento in Lecce, Italy.

The remainder of the paper is organized as follows. Section 2 describes the PA and MSA tools that have been embedded in our system, whereas Section 3 presents the approach to optimize the execution of concurrent alignment tools. Motivations for a Grid Alignment Toolkit are discussed in Section 4 whilst Section 5 discusses the architecture of BioGAT. Section 6 describes a case study related to BLAST and a preliminary test. Section 7 recalls related work and finally Section 8 concludes the paper highlighting future work.

2. Alignment tools

BLAST, or Basic Local Alignment Search Tool [1], is an alignment tool that uses a measure of local similarity to score sequence alignments in order to identify regions of good local alignment.

The basic BLAST algorithm can be implemented in DNA and protein sequence database searches, motif searches, gene identification searches and in the analysis of multiple regions of similarity in long DNA sequences.

Sequences must be input in one of three formats: FastA sequence format, NCBI Accession numbers, or GIs (GenBank Identifiers).

Position Specific Iterative (PSI)-BLAST is a sensitive sequence similarity search tool that uses an iterative searching method and unique scoring scheme to detect weakly related homologues [3].

PSI-BLAST often requires human input in order to prevent unrelated hits from "corrupting" the scoring matrix and incorrectly biasing the search toward false positives. The ability to recognize weak sequence patterns makes PSI-BLAST a critical tool for comprehensively identifying all of the members of a protein family or instances of a functional domain.

However, PSI-BLAST is the best (perhaps the only) freely-available reliable automated system for homologous sequence discovery and its popularity signals are a clear challenge for computational biology to shift up a gear.

The FASTA bioinformatics tool [2], developed by W.R. Pearson at the University of Virginia, provides quick search and local alignment of sequences contained within a specified database. FASTA offers many of the functionalities provided by BLAST. Although BLAST tools are faster, FASTA provides more accurate sequence alignments.

3. Concurrent execution of alignment tools

There are two approaches for the execution of concurrent alignment applications: