Chapter 8.9
PheGee@Home: A Grid-Based Tool for Comparative Genomics

Bertil Schmidt
Nanyang Technological University, Singapore

Chen Chen
Nanyang Technological University, Singapore

Weiguo Liu
Nanyang Technological University, Singapore

Wayne P. Mitchell
Experimental Therapeutics Centre, Singapore

ABSTRACT

In this chapter we present PheGee@Home, a grid-based comparative genomics tool that nominates candidate genes responsible for a given phenotype. A phenotype is the physical manifestation of the interplay of genetic, epigenetic and environmental factors. Our tool is designed to facilitate the discovery and prioritization of candidate genes controlling or contributing to the genetically determined portion of a specified phenotype. However, in order to make reliable nominations of candidate genes from sequence data, several genome-size sequence datasets are required. This makes the approach impractical on traditional computer architectures leading to prohibitively long runtimes. Therefore, we use a computational architecture based on a desktop grid environment and commodity graphics hardware to significantly accelerate PheGee. We validate this approach by showing the deployment and evaluation on a grid testbed for the comparison of microbial genomes.

INTRODUCTION

High-throughput techniques for DNA sequencing have led to an enormous growth in the amount of publicly available genomic data. As of February 2008, 716 complete genome sequences are available and another 2,756 genome-sequencing projects are in progress (www.genomesonline.org). As the sequences of more and more genomes become available, we have reached a critical mass where, instead of focusing on a subset of

DOI: 10.4018/978-1-4666-0879-5.ch8.9

sequences, we can use entire genome data sets to derive global inferences and metadata. Comparative genomics refers to the study of relationships between the genomes of different species or strains. It is currently being used for ortholog detection (Itoh, Goto, Akutsu & Kanehisa, 2005) bacterial pharmacogenomics (Fraser, et al., 2000), clustering of similar protein sequences (Itoh, Akutsu & Kanehisa, 2004), etc. Unfortunately, comparative genomics applications are highly computationally intensive tasks due to the large sequence data sets involved and typically take a few months to complete. These runtime requirements are likely to become even more severe due to the rapid growth in the size of genomic databases.

The objectives of this chapter are therefore two-fold:

1. The presentation of a new comparative genomics tool called PheGee (*Phenotype Genotype Explorer*). PheGee nominates candidate genes responsible for a certain phenotype π given genomic sequence datasets of phenotype positive ($\pi+$) and phenotype negative ($\pi-$) species.
2. The proposition of a hybrid computational grid platform to accelerate PheGee.

The proposed hybrid grid architecture efficiently combines desktop grid computing with GPGPUs (General-Purpose computation on Graphics Processing Units). The driving force and motivation behind this architecture is the price/performance ratio. Using desktop grids as in the volunteer computing approach is currently one of the most efficient and simple ways to gain supercomputer power for a reasonable price. Installing in addition massively parallel processor add-on boards such as modern computer graphics cards within each desktop can further improve the cost/performance ratio significantly. We show how this architecture can be used to accelerate PheGee efficiently. Moreover, the proposed grid approach is flexible and is therefore applicable to a variety of compute-intensive genomics applications.

BACKGROUND

Biological Motivation

PheGee exploits the simple observation: Given a set of species, A, which evince a particular phenotype, π, and a second set of species, B, which do not evince π, then in general genes responsible for π will be present in A and absent from B. Given incomplete knowledge, in the universe of sequenced organisms there will be a third set, C, generally the largest of the three, comprising those organisms undefined in terms of π. To cast this in symbolic terms, if a particular process is the outcome of three genes $x \rightarrow y \rightarrow z$, then in general homologs of x, y, and z will be found in A but not in B. Those genes found in all members of A and not found in B, are therefore the best candidates for genes responsible for the process resulting from x, y, z. However, a perfect concordance between phenotype and genotype in sets of organisms is not guaranteed. Apart from epigenetic and environmental factors that can shape phenotype, the following genetic phenomena might disrupt the relationship.

- *Gene Replacement* (Achtmann & Wagner, 2008): Gene y can be lost in some members of A, say A^*. In A^* the function of y may have been replaced by some other gene, w. Thus, although the phenotype is preserved in all members of A, and although y is a gene that contributes to the A phenotype, y will nevertheless not map to every member of A. In particular, y is not found in A^*.
- *Horizontal Gene Transfer* (Achtmann & Wagner, 2008): Genes x, y, or z may be adopted by a new species through the process of "horizontal gene transfer". In this scenario some members of B, say B^*, will