A Hybrid Distributed Framework for SNP Selections

Pengfei Liu[1], Shuai Li[1], Weiying Yi [1] and Kwong Sak Leung[1]
[1]Department of Computer Science and Engeering
The Chinese University of Hong Kong, Hong Kong

Abstract— *With the development of next generation sequencing technology, researchers are able to obtain extremely high-dimensional data. However, only a fraction of the data is related to diseases and the computational time on processing the whole sequences is tremendous. Moreover, using the high-dimensional data directly will greatly reduce the accuracy of the machine learning and data mining algorithms. Single nucleotide polymorphism (SNP) selections is critical for addressing these problems in genome wide association study (GWAS). Typically, it needs days to perform SNP selections even though the relationship between SNPs and diseases is assumed to be linear. More time is needed when the relationship is nonlinear. In order to speed up the SNP selection processes, a CPU-GPU hybrid distributed framework (HDF) specifically for SNP selection algorithms is introduced in this paper. The HDF fully utilizes the computational power of machines. And the interfaces are also provided, which help researchers to extend their SNP selection algorithms into distributed version. The results demonstrate that the acceleration by HDF is about hundreds times on SNP selections with synthetic and real data, compared to single machine.*

Keywords: SNP selection, CPU, GPU, distributed

1. Introduction

Genome-wide association study (GWAS) is an analysis to identify which part of the human genomes are associated to a certain trait. In GWAS, statistical and computational analyses are applied to compare the DNA sequences of the controls (healthy samples) and the cases (patients with the specific genetic disease) in order to identify the related single nucleotide polymorphisms (SNPs) [1][2][3]. With the technology of next generation sequencing, researchers are able to obtain millions of SNPs in DNA sequences. However, only a small fraction of the SNPs are related to diseases, and the rest are irrelevant and regarded as noises. These noises will severely reduce the accuracy and reliability of the GWAS algorithms [4]. Hence, identifying the useful SNPs before analyzing their relationship with diseases is a critical issue in GWAS.

There are about 4 million SNPs in human DNA sequences, and many SNPs work in coordination to manifest a disease [5]. Analyzing such a high dimensional combinatorial relationship increases the computational complexity a lot. To speed up the process of SNP selection, parallel computing is adopted in this work.

Nowadays, most of the computers are equipped with both CPUs and GPUs. In order to maximize the utilization of the computing resources, a CPU-GPU hybrid distributed framework (HDF) specifically for SNP selections is proposed and implemented. The HDF provides interfaces that help researchers extend their SNP selection algorithms into distributed versions. To the authors' best knowledge, the proposed HDF is the first CPU-GPU based distributed system specifically designed for speeding up the SNP selection processes.

The HDF consists of three components:

- **Controller**
  The Controller decomposes the original computational mission from a user into many small tasks, and distributes them to the CPU clients and GPU clients described below. After receiving the progress report from the clients, the Controller merges the progresses and find a new task to distribute, which can be manipulated using the provided interfaces.
- **CPU client**
  The CPU clients use multi-thread architecture to handle the tasks distributed by the Controller and return the results back to the Controller after finishing the task.
- **GPU client**
  The GPU clients use Nvidia CUDA API to process the tasks distributed by the Controller and return the results back to the Controller after finishing the task.

To test the performance of the HDF, we implement an SNP selection algorithm ReliefF on the HDF using the provided interfaces. The experimental results show that the distributed version is about hundreds times faster than the a single thread CPU version.

The rest of the paper is organized as follows. Literature reviews on SNP selection algorithms and distributed systems are introduced in Section 3. Section 4 and 5 are the architecture design and the implementation of the HDF. Experiments are performed in Section 6 and the results are analyzed in Section 7. Section 8 is the conclusion and discussion.

2. Definition of SNP Selections

Each SNP in human genome is either an A-T pair or a C-G pair. For each SNP, the pair with higher probability is called the major allele, otherwise it is called the minor allele.