Chapter 10
Data Intensive Computing for Bioinformatics

Judy Qiu
Indiana University - Bloomington, USA

Saliya Ekanayake
Indiana University - Bloomington, USA

Jaliya Ekanayake
Indiana University - Bloomington, USA

Stephen Wu
Indiana University - Bloomington, USA

Thilina Gunarathne
Indiana University - Bloomington, USA

Scott Beason
Computer Sciences Corporation, USA

Jong Youl Choi
Indiana University - Bloomington, USA

Geoffrey Fox
Indiana University - Bloomington, USA

Seung-Hee Bae
Indiana University - Bloomington, USA

Mina Rho
Indiana University - Bloomington, USA

Yang Ruan
Indiana University - Bloomington, USA

Haixu Tang
Indiana University - Bloomington, USA

ABSTRACT

Data intensive computing, cloud computing, and multicore computing are converging as frontiers to address massive data problems with hybrid programming models and/or runtimes including MapReduce, MPI, and parallel threading on multicore platforms. A major challenge is to utilize these technologies and large-scale computing resources effectively to advance fundamental science discoveries such as those in Life Sciences. The recently developed next-generation sequencers have enabled large-scale genome sequencing in areas such as environmental sample sequencing leading to metagenomic studies of collections of genes. Metagenomic research is just one of the areas that present a significant computational challenge because of the amount and complexity of data to be processed. This chapter discusses the use of innovative data-mining algorithms and new programming models for several Life Sciences applications. The authors particularly focus on methods that are applicable to large data sets coming from high throughput devices of steadily increasing power. They show results for both clustering and dimension reduction algorithms, and the use of MapReduce on modest size problems. They identify two key areas where further research is essential, and propose to develop new O(NlogN) complexity

DOI: 10.4018/978-1-61520-971-2.ch010

algorithms suitable for the analysis of millions of sequences. They suggest Iterative MapReduce as a promising programming model combining the best features of MapReduce with those of high performance environments such as MPI.

INTRODUCTION

Overview

Data intensive computing, cloud computing, and multicore computing are converging as frontiers to address massive data problems with hybrid programming models and/or runtimes including MapReduce, MPI, and parallel threading on multicore platforms. A major challenge is to utilize these technologies and large scale computing resources effectively to advance fundamental science discoveries such as those in Life Sciences. The recently developed next-generation sequencers have enabled large-scale genome sequencing in areas such as environmental sample sequencing leading to metagenomic studies of collections of genes. Metagenomic research is just one of the areas that present a significant computational challenge because of the amount and complexity of data to be processed.

This chapter builds on research we have performed (Ekanayake, Gunarathne, & Qiu, Cloud Technologies for Bioinformatics Applications, 2010) (Ekanayake J., et al., 2009) (Ekanayake, Pallickara, & Fox, MapReduce for Data Intensive Scientific Analyses, 2008) (Fox, et al., 2009) (Fox, Bae, Ekanayake, Qiu, & Yuan, 2008) (Qiu, et al., 2009) (Qiu & Fox, Data Mining on Multicore Clusters, 2008) (Qiu X., Fox, Yuan, Bae, Chrysanthakopoulos, & Nielsen, 2008) (Twister, 2011) on the use of Dryad (Microsoft's MapReduce) (Isard, Budiu, Yu, Birrell, & Fetterly, 2007) and Hadoop (open source) (Apache Hadoop, 2009) to address problems in several areas, such as particle physics and biology. The latter often have the striking all pairs (or doubly data parallel) structure highlighted by Thain (Moretti, Bui, Hollingsworth, Rich, Flynn, & Thain, 2009). We discuss here, work on new algorithms in "Innovations in Algorithms for Data Intensive Computing" section, and new programming models in "Innovations in Programming Models Using Cloud Technologies" and "Iterative MapReduce with Twister" sections.

We have a robust parallel Dimension Reduction and Deterministic Annealing clustering, and a matching visualization package. We also have parallel implementations of two major dimension reduction algorithms – the SMACOF approach to MDS and Generative Topographic Mapping (GTM) described in "Innovations in Algorithms for Data Intensive Computing" section. MDS is $O(N^2)$ and GTM O(N) but, since GTM requires the points to have (high dimensional) vectors associated with them, only MDS can be applied to most sequences. Also, since simultaneous multiple sequence alignment MSA is impractical for interesting biological datasets, MDS is a better approach to dimension reduction for sequence samples, because it only requires sequences to be independently aligned in pairs to calculate their dissimilarities. On the other hand, GTM is attractive for analyzing high dimension data base records, where well defined vectors are associated with each point – in our case each database record. Distance calculations (Smith-Waterman-Gotoh) MDS and clustering are all $O(N^2)$, and will not properly scale to multi-million sequence problems and hierarchical operations to address this are currently not supported for MDS and clustering except in a clumsy manual fashion. In the final part of "Innovations in Algorithms for Data Intensive Computing" section, we propose a new multiscale (hierarchical) approach to MDS that could reduce complexity from $O(N^2)$ to O(NlogN) using ideas