Parallel Computing 34 (2008) 613–615

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Guest editorial

High-performance computational biology

1. Editorial

Over the past decade, computational molecular biology has grown into a mature discipline with a well-defined body of core knowledge, and participation from a large and diverse group of researchers. To keep pace with the explosive growth in research in this field, a number of high quality journals and annual conferences have been established. Many universities are actively building academic programs and research centers and groups in computational biology. As a reflection of the maturing of the field, numerous textbooks on computational biology and its various subtopics have been written in recent years, and undergraduate programs are underway. Despite this progress, computational biology continues to be a vibrant discipline with many outstanding research problems and potential for new avenues of investigation for decades to come.

We broadly view high-performance computational biology as the development and application of high-performance computing techniques for extending the reach or scale of investigations in computational biology. A major component of this is the development of parallel and distributed algorithms, and programming environments and systems for aiding biological investigations using high-performance parallel computers, grid computing, and emerging architectures. There is a compelling need for such research given the explosive growth in biological information, the complexity of interactions that underlie many biological processes, and the diversity and interconnectedness of organisms at the molecular level. However, research in high-performance computational biology has not grown as rapidly as computational biology itself. There are subfields of computational biology which have not seen significant influx of ideas from the high-performance computing community. This is perhaps a reflection of the confluence of expertise needed to conduct research in high-performance computational biology, which sets up a barrier to entry for new researchers. Efforts spent in transgressing the barrier are worthwhile given the opportunities for high impact research. By bringing together research in this area as a special issue, we hope to provide a resource for *Parallel Computing* readers interested in this field and aid the entry of new researchers into the field.

The arguments in favor of a sustained effort in high-performance computational biology are stronger than ever. New high-throughput sequencing machines introduced recently, such as those from 454 Life Sciences Inc. and Solexa, have significantly accelerated sequencing capabilities. It is now possible to sequence millions of short DNA fragments in a matter of hours for several thousand dollars. These machines are increasingly being used to sample transcriptomes of many organisms. The sequencing of several complex plant genomes is underway starting with the recently finished maize and sorghum genome sequencing projects. Similar to large-scale genome sequencing projects, comprehensive gene expression profile measurement projects are underway to conduct large-scale microarray experiments on an organism spanning various organs, disease/stress induced states, and developmental stages. Forays into personalized medicine, rational drug design, large-scale systems biology, such as the study of protein–protein interaction networks at the whole organism level, understanding evolutionary relationships and building the tree of life, all require processing vast amounts of data or carrying out highly complex computational tasks.

In this special issue, we showcase some of the recent work in high-performance computational biology. Specifically, authors whose work was published in the 2007 IEEE International Workshop on High-Performance Computational Biology (HiCOMB, http://www.hicomb.org) were solicited to submit extended versions of their papers. Each manuscript submitted to the special issue was subjected to rigorous, independent peer review by three to four reviewers. We are extremely grateful to all the reviewers who agreed and delivered on providing thoughtful reviews within the time constraints imposed for the special issue. Based on the reviewer suggestions and our own reading of the manuscripts, six manuscripts were selected for publication in the special issue.

The first paper in this special issue is "Exploring the viability of the Cell Broadband Engine for bioinformatics applications," by Vipin Sachdeva, Michael Kistler, Evan Speight, and Tzy-Hwa Kathy Tzeng. The authors evaluate the performance

0167-8191/$ - see front matter
doi:10.1016/j.parco.2008.10.001

of bioinformatics algorithms on the Cell Broadband Engine (Cell/B.E.), the heterogeneous multicore chip that IBM has developed for the Sony PlayStation 3. This paper reports on a thorough investigation of three computationally-intensive and popular applications, FASTA, ClustalW, and HMMER, and their suitability for multicore processors such as the Cell/B.E. platform. FASTA performs fast similarity searching where uncharacterized but sequenced "query" genes are scored against vast databases of characterized sequences. ClustalW produces multiple sequence alignments that are needed to organize data to reflect sequence homology, identify conserved sites, and perform phylogenetic analysis. HMMER is a protein sequence analysis tool that employs hidden Markov models for comparisons with sequences. In these cases of popular bioinformatics codes, the authors demonstrate the Cell/B.E.'s suitability for high-performance computation of these bioinformatics workloads.

Protein-interaction networks play an important role in understanding the functional and organizational principles of biological processes. Promising computational techniques for key systems biology research problems such as identification of signaling pathways, novel protein function prediction, and the study of disease mechanisms, are all based on graph topological characteristics of these networks. In the next paper in the special issue, "A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms," David A. Bader and Kamesh Madduri develop a multicore parallel algorithm for betweenness centrality and shows its novelty in identifying critical proteins in the Human and Yeast protein-interaction networks. The authors demonstrate a novel contribution that graph-theoretic algorithms from social network analysis may be used to solve important biological queries in mere seconds on multicore processors.

Sadaf R. Alam, Pratul K. Agarwal, and Jeffrey S. Vetter, in their paper "Performance characteristics of biomolecular simulations on high-end systems with multi-core processors," study the performance of high-performance applications that model multi-scale biological processes occurring inside the cell. Specifically, the paper includes a study of biomolecular simulation based on molecular dynamics and a characterization the computation, communication, and memory efficiencies, on the Cray XT supercomputer system. The authors focus on two test biological systems, one models a protein-DNA complex in explicit solvent and counter-ions to allow the system to be charge neutral and the second models cellulose degrading enzyme cellulase complex. Through this work, the authors develop novel strategies such as a memory affinity scheme for optimizing the biomolecular simulation using the multicore processors on each node of the Cray XT as well as the parallelization using MPI across the nodes.

In the paper titled "Biomolecular committor probability calculation enabled by Processing in Network Storage," Paul Brenner, Justin M. Wozniak, Doug Thain, Aaron Striegel, Jeff W. Peng, and Jesus A. Izaguirre, also consider computationally complex and data-intensive atomistic biomolecular simulations, and design a new computational framework based on Processing in Network Storage (PINS). This novel distributed approach overcomes bandwidth, compute, organizational, and security challenges, required to solve meaningful simulation problems. As an example, performance of PINS is reported for the committor probability calculation of a solvated protein domain that requires 500 independent simulations and generates over one million output files.

An ultimate goal in computational biology is to understand biological function and the interrelation between contributing processes. Ribonucleic Acid (RNA) molecules play an important role in areas such as gene function and regulation. In "RNAVLab: A virtual laboratory for studying RNA secondary structures based on grid computing," Michela Taufer, Ming-Ying Leung, Thamar Solorio, Abel Licon, David Mireles, Roberto Araiza, and Kyle L. Johnson, present a virtual laboratory for predicting the folding into secondary structures of RNA molecules. Methods that minimize thermodynamic free energy require significant computation, and the authors demonstrate a framework using grid computing that systematically samples RNA nucleotide segments. The study focuses on the virus family *Nodaviridae* and shows how RNAVLab can reduce the processing time while maintaining the prediction accuracy.

The final paper in the special issue is "Integrating FPGA acceleration into HMMer," authored by Tim Oliver, Leow Yuan Yeow, and Bertil Schmidt. The authors also improve the performance of HMMer, a popular package for biological sequence database search, by using a hardware reconfigurable processor called a Field Programmable Gate Array (FPGA). As reported in this paper, an accelerated FPGA implementation now supports the full Plan7 Viterbi algorithm and gives a competitive price/performance ratio using the Xilinx FPGA. The authors suggest that because of the demonstrated speedups and small form factor, these accelerators could be used in cluster and grid computing to realize high-performance platforms that are capable of solving a wide variety of computational bioinformatics problems.

We hope that the readers will find the papers in this special issue informative and useful. Readers with continued interest in high-performance computational biology research are referred to the Annual Workshop on High-Performance Computational Biology (HiCOMB, see http://www.hicomb.org for details) held in conjunction with the International Parallel and Distributed Processing Symposium.

Acknowledgements

The guest editors wish to thank the authors of all submitted manuscripts, without whom this special issue would not have been possible. They also thank the reviewers who provided a thorough evaluation of the submitted manuscripts in a timely manner. The Parallel Computing editorial staff has provided assistance throughout the process of bringing out the special issue. Finally, we acknowledge the 2007 HiCOMB program co-chairs, Ananth Grama and Shankar Subramaniam, for their efforts in soliciting these works for the special issue.

David A. Bader
College of Computing, Georgia Institute of Technology,
Atlanta, GA 30332, United States
URL: http://www.cc.gatech.edu/~bader

Srinivas Aluru
Department of Electrical and Computer Engineering,
Iowa State University,
Ames, IA 50011, United States
URL: http://www.ee.iastate.edu/~aluru

Available online 15 October 2008