Editorial: Special Section on High-Performance Computational Biology

Srinivas Aluru, *Senior Member, IEEE*, Nancy M. Amato, and David A. Bader, *Senior Member, IEEE Computer Society*

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 section, we hope to provide a resource for *IEEE Transactions on Parallel and Distributed Systems* (*TPDS*) 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 within the last year, such as those from 454 Life Sciences Inc., have significantly accelerated sequencing capabilities. Using 454 sequencing systems, it is possible to sequence as many as 200,000 short DNA fragments in a 4 hour experiment for a few 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 maize and sorghum. 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, diesease/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 section, we showcase some of the recent work in high-performance computational biology. In addition to the open call for papers, authors whose work was published in the 2005 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 section 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 section.

The first paper in this special issue is on a scalable implementation of the widely used BLAST search program for homology detection between a query sequence and a database of known sequences. In "ScalaBLAST: A Scalable Implementation of BLAST for High-Performance Data-Intensive Bioinformatics Analysis," Christopher Oehmen and Jarek Nieplocha report on ScalaBLAST, a high-performance sequence alignment program they developed to enable applications that require thousands to millions of queries to be performed simultaneously. Such queries are used in applications such as multiple genome/proteome comparisons, and in finding genes in newly sequenced genomes. By using a combination of techniques, including target database distribution, exploiting multilevel parallelism, parallel I/Os and latency hiding, the authors achieve a scalable implementation of this ubiquitous search program.

- *S. Aluru is with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011. E-mail: aluru@iastate.edu.*
- *N.M. Amato is with the Department of Computer Science, Texas A&M University, College Station, TX 77843-3112. E-mail: amato@cs.tamu.edu.*
- *D.A. Bader is with the College of Computing, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: bader@cc.gatech.edu.*

For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org.

1045-9219/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

Certain design patterns repeatedly occur in algorithms for several computational biology applications. For example, two-dimensional dynamic programming is a frequently used technique in computational biology. Moreover, many algorithms exhibit similar data dependencies: an element in the dynamic programming table depending on the three elements with row number or column number or both reduced by one, or an element in the dynamic programming table depending on all previous elements in the same row and all previous elements in the same column, etc. In the paper titled "Parallel Pattern-Based Systems for Computational Biology: A Case Study," Weiguo Liu and Bertil Schmidt develop parallel pattern-based prototypes for computational biology to enable rapid development of HPC applications that rely on certain target design patterns. They target two classes of applications that rely on dynamic programming algorithms or genetic algorithms. They demonstrate that such an approach can achieve good scaling on PC clusters and grid environments.

The next paper in the special section revisits the classic dot plot technique for comparison of biological sequences. The dot plot is a simple comparison matrix that compares every sliding window of a small fixed length k in one sequence with every sliding window of length k in the other sequence. In "High-Performance Direct Pairwise Comparison of Large Genomic Sequences," Christopher Mueller, Mehmet Dalkilic, and Andrew Lumsdaine bring up many surprising issues that arise in developing a high-performance, parallel implementation of the seemingly simple dot plot algorithm. They exploit parallelism at many levels, including the vector processing units and SIMD parallelism within modern microprocessors, multiple processors within a single node, and coarse-grained parallelism across multiple nodes in a cluster with the goal of enabling direct pairwise comparison between large genomes.

Molecular analysis-based drug discovery is an expensive and time consuming process. The first step in this process is the screening of hundreds of thousands of potential candidate molecules for features such as specific binding sites or affinity to bind to a specific protein. This can be formulated as a frequent subgraph mining problem in a lattice of subgraph relationships between subgraphs of candidate molecules. The explosive size of such lattices and the complexity of subgraph isomorphism testing algorithms make this a computationally demanding exercise. In the paper titled "Dynamic Load Balancing for the Distributed Mining of Molecular Structures," Giuseppe Di Fatta and Michael R. Berthold present methods for distributed mining of molecular structures as candidates for drug discovery. They propose methodologies for dynamic partitioning of the highly irregular search space and present a new load balancing technique for distributed mining. They present a framework suitable for loosely coupled, heterogenous systems such as grids and demonstrate the effectiveness of their approach on an HIV-screening data set.

Predicting the structure of a protein from its amino acid sequence is considered a "holy grail" problem in computational biology. In the paper "Predictor@Home: A Protein Structure Prediction Supercomputer Based on Global Computing," Michela Taufer, Chahm An, Andre Kerstens, and Charles L. Brooks III harness the power of volunteered computing resources on a global scale to carry out more accurate protein structure prediction. To assess progress in structure prediction capabilities, the community organizes a Critical Assessment of Structure Prediction (CASP) competition during even years where experimentally determined structures of proteins are compared against computationally predicted structures. The authors utilize Internet connected global computing resources for both protein conformational search and sampling, and structure refinement. They report exploiting more than 380 years of compute time during CASP6 competition which allowed them to increase sampling capacity by one to two orders of magnitude and resulted in more accurate structure prediction.

The final paper in the special section is "Adaptive Electrocardiogram Feature Extraction on Distributed Embedded Systems," authored by Roozbeh Jafari, Hyduke Noshadi, and Majid Sarrafzadeh. The authors address the problem of electrocardiogram data analysis taking into account energy consumption constraints common to tiny embedded devices. Potential long-term applications of this research include development of wearable electronics with embedded devices for monitoring patient health.

We hope that the readers will find the papers in this special section 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.

ACKNOWLEDGMENTS

The special section editors wish to thank the authors of all submitted manuscripts, without whom this special section would not have been possible. They also thank the reviewers who provided a thorough evaluation of the submitted manuscripts in a timely manner. The *TPDS* editorial staff has provided assistance throughout the process of bringing out the special section.

Srinivas Aluru is a Professor in the Department of Electrical and Computer Engineering and serves as the chair of the Bioinformatics and Computational Biology graduate program at Iowa State University. He is a member of the Laurence H. Baker Center for Bioinformatics and Biological Statistics, and the Center for Plant Genomics at Iowa State. Previously, he held faculty positions at New Mexico State University and Syracuse University. Dr. Aluru was a recipient of the US National Science Foundation Career award in 1997, an IBM faculty award in 2002, the Iowa State University Young Engineering Faculty Research Award in 2002, and the Warren B. Boast Undergraduate Teaching Award in 2005. He is an IEEE Computer Society distinguished visitor from 2004 to 2006. His research interests include parallel algorithms and applications, bioinformatics and computational biology, and combinatorial scientific computing. He served on numerous program committees and has taken up leadership roles at several conferences and workshops in these areas. His contributions to computational biology are in computational genomics, string algorithms, and parallel methods for solving large-scale problems arising in biology. He cochairs the Annual Workshop on High-Performance Computational Biology (http://www.hicomb.org) and edited a comprehensive handbook on computational molecular biology. He is a member of the ACM, SIAM, Life Sciences Society, and a senior member of the IEEE and the IEEE Computer Society.

Nancy M. Amato received the BS and AB degrees in mathematical sciences and economics, respectively, from Stanford University in 1986, and the MS and PhD degrees in computer science from the University of California, Berkeley and the University of Illinois at Urbana-Champaign in 1988 and 1995, respectively. She was an AT&T Bell Laboratories PhD Scholar and a recipient of a CAREER Award from the US National Science Foundation. She is currently a professor of computer science at Texas A&M University, where she is the codirector of the Parasol Laboratory. Her main areas of research focus are motion planning, computational biology, computational geometry, and parallel and distributed computing. Details about her group's research are available on her Web page: http://parasol.tamu.edu/~amato. She has served on several editorial boards, including the *IEEE Transactions on Robotics and Automation* and the *IEEE Transactions on Parallel and Distributed Systems*. She sits on review panels for the NIH and the US National Science Foundation, and she regularly serves on conference organizing and program committees. She is a member of the Computing Research Association's Committee on the Status of Women in Computing Research (CRA-W) and she codirects the CRA-W's Distributed Mentor Program.

David A. Bader received the PhD degree in 1996 from The University of Maryland, and was awarded a US National Science Foundation (NSF) postdoctoral research associateship in experimental computer science. He is an associate professor in the Computational Science & Engineering Department, a division within the College of Computing, Georgia Institute of Technology. He is an NSF CAREER Award recipient, an investigator on several NSF awards, a distinguished speaker in the IEEE Computer Society Distinguished Visitors Program, and is a member of the IBM PERCS team for the DARPA High Productivity Computing Systems Program. Dr. Bader serves on the steering committees of the IPDPS and HiPC conferences, and was the general cochair for IPDPS (2004-2005), and vice general chair for HiPC (2002-2004). He has chaired several major conference program committees: program chair for HiPC 2005, program vice chair for IPDPS 2006, and program vice chair for ICPP 2006. He has served on numerous conference program committees related to parallel processing and computational science and engineering, is an associate editor for several high impact publications including the *IEEE Transactions on Parallel and Distributed Systems* (*TPDS*), the *ACM Journal of Experimental Algorithmics* (*JEA*), *IEEE DSOnline*, and *Parallel Computing*, is a senior member of the IEEE Computer Society, and a member of the ACM. Dr. Bader has been a pioneer in the field of high-performance computing for problems in bioinformatics and computational genomics. He has cochaired a series of meetings, the IEEE International Workshop on High-Performance Computational Biology (HiCOMB), written several book chapters, and coedited special issues of the *Journal of Parallel and Distributed Computing* (*JPDC*) and the *IEEE TPDS* on high-performance computational biology. He has coauthored more than 75 articles in peer-reviewed journals and conferences, and his main areas of research are in parallel algorithms, combinatorial optimization, and computational biology and genomics.