Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT

J. Parallel Distrib. Comput. 63 (2003) 671–673

Journal of Parallel and Distributed Computing

http://www.elsevier.com/locate/jpdc

Editorial

Guest Editor's Introduction: Special issue on high-performance computational biology

Ever since the structure of DNA was discovered in 1953, biology has been steadily changing from a descriptive science concerned with the behavior and characteristics of organisms to a mathematical discipline that relates the essential life processes to the underlying biomolecular data. This discovery has stimulated the growth of molecular biology, the study of how biomolecular sequences are related to the functioning of organisms. These developments have brought biology closer to computer science. In many ways, the underlying mechanisms are similar to what we employ in building and programming computers. The characteristics of a life form are coded in its DNA (program), which is processed in each cell (executed) to produce the proteins (outputs) that carry out essential life processes. The field holds immense potential for future discoveries that are unrivaled in significance, such as the design of protein sequences to fold into a specific configuration to efficiently administer drugs and the possibility of treating diseases by altering the genetic code.

The need to discover biomolecular sequences, to relate the sequences to their structure and function, and to understand the sequences through mutual comparison has resulted in a number of interesting problems for algorithm designers and has led to the development of computational molecular biology. The area has attracted many competent researchers, and the field is developing at a rapid pace, as evidenced by the growth of conference meetings and avenues for publication. Algorithms for solving biological problems are often associated with long running times. This arises from various factors: (1) Biological data are obtained by experiments which are prone to errors. The need to deal with errors and uncertainties results in algorithms with high complexity. (2) The data size itself may be large and result in long running times. (3) Many of the problems are shown to be NP-hard, and techniques such as energy minimization and branch and bound are used. As biologists progressed from the study of simple biomolecular data from less complex organisms to the eventual goal of understanding and manipulating entire genomes of complex organisms, the corresponding computational needs were scaled similarly. We believe that effective use of parallel computers is becoming increasingly important for solving meaningful biological problems in reasonable time.

The field of computational molecular biology is replete with applications that require processing of large amounts of data. The basic problem of finding DNA sequences that exhibit homology to a given query sequence requires searching databases containing over tens of billions of nucleotides, and still growing at an exponential rate. The recent assembly of the mouse genome required processing over 33 million fragments of a total size of over 17 billion bases to assemble the genome of size over 3 billion bases. In comparative genomics, two or more genomes of such enormous sizes must be compared to discover common genes and interesting evolutionary relationships among species. To construct trees representing evolutionary relationships among species, algorithms explore a large search space of potential trees. Biomolecular simulations such as protein structure determination require a large number of iterations, making it important to accelerate the runtime per iteration. In these and many other applications, parallel processing can enable the solution of realistic problem instances.

Our intent in bringing out this special issue is twofold: We wish to showcase some of the important research work being carried out in the field of parallel computational biology to the broader parallel and distributed computing community. Through this special issue, we also hope to stimulate interest in the field and attract other researchers to participate in this field.

The call for papers for the special issue generated more than 30 submissions with authors representing 10 countries spanning five continents. Only manuscripts pertaining to the special issue focus were further considered, yielding 27 manuscripts that were each subjected to 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. On the basis of the reviewer suggestions and our own reading of the manuscripts, eight manuscripts were selected for publication in the special issue.

The first paper in this special issue is on implementing a genetic linkage analysis package on parallel computers.

doi:10.1016/S0743-7315(03)00128-X

Genetic linkage analysis is aimed at identifying the approximate locations of disease-causing genes using data from closely related individuals with a high incidence of the disease. In this paper, Gavin C. Conant, Steven J. Plimpton, William Old, Andreas Wagner, Pamela R. Fain, Theresa R. Pacheco, and Grant Heffelfinger present a parallel implementation of the Genehunter package to facilitate multipoint linkage analysis on distributed memory parallel computers.

An observable trait of an individual, called phenotype, is often determined through a complex interaction of multiple genes at different loci on the genome. The paper titled "Least Squares Fit of Genomic Data by Sums of Epistatic Effects" offers a window into exploring this relation. Philip Hanlon, William Andrew Lorenz, and Dave Strenski model the epistatic effects using Additive Epistatic Effect (AEE) functions and present an algorithm for finding the best AEE function that models a given data set. The authors present a detailed analysis of their implementation, exploiting many features of the Cray SV1ex and future Cray X1 architectures.

Microarrays, also known as DNA chips, allow the simultaneous measurement of expression levels of thousands of genes. By allowing us to study the variation in gene expression, microarrays are helpful in identifying genes responsible for genetic diseases. They are also useful in identifying biochemical pathways. Clustering analysis of gene expression data is often the first step in microarray data analysis and allows detection of co-regulated genes. The paper "Clustering Analysis of Microarray Gene Expression Data by Splitting Algorithm," by Ruye Wang, Lucas Sharenbroich, Christopher Hart, Barbara Wold, and Eric Mjolsness, contains a top-down splitting approach to unsupervised clustering of microarray data.

Curiosity about pattern and process in the evolution of diversity has motivated many biologists and natural historians to study phylogenetics. Phylogenetic analysis attempts to reconstruct, from data generated from extant species, the evolutionary history of the group under study. The history is generally represented by a bifurcating (binary) tree, a phylogeny. Phylogenetic analysis is crucial to answering some fundamental open questions in evolution. It also has practical applications such as discovering biochemical pathways unique to target organisms, studying the epidemiology of viruses such as HIV to predict the progression of disease with time in an individual, and developing improved strains of crops based on understanding of phylogenetic distribution of variation in wild populations. The ambitious NSF *Tree of Life* program is aimed at discovering the complex evolutionary relationships between all knows species on the planet. This special issue contains two papers on phylogenetic analysis. Bayesian phylogenetic inference is used to find the phylogenetic tree that best explains a given set of evolutionary relationships and the probability that a given tree is correct with respect to given data. Xizhou Feng, Duncao A. Buell, John R. Rose, and Peter J. Waddell present a parallel algorithm for Bayesian phylogenetic inference using a Markov Chain Monte Carlo (MCMC) method. One of the methods employed in phylogenetic tree construction is to first compute the optimal phylogenetic tree for every combination of four species, called quartets, and then build phylogenetic trees consistent with all the quartets. In the paper titled "Molecular Phylogenetics: Parallelized Parameter Estimation through Quartet Puzzling," Heiko A. Schmidt, Ekkehard Petzold, Martin Vingron, and Arndt von Haeseler present a parallel algorithm for estimating the parameters for evolutionary models to reconstruct phylogenetic trees.

The structures of biochemical molecules such as proteins are crucial to their function, and computational structural biology is an active area of research. Computationally determining the three-dimensional structure of protein molecules directly from their primary sequence information is considered a "Holy Grail" problem in computational molecular biology. We have included three papers that deal with structures in biology. The first paper, titled "A Grid-Aware Approach to Protein Structure Comparison," by Carlo Ferrari, Concettina Guerra, and Giuseppe Zanotti, deals with software for comparison of protein structures. In dealing with protein structures, structure similarity often translates to function similarity. The authors present a distributed algorithm for retrieving proteins matching a given query protein from a structure database. Dan C. Marinescu and Yongchang Ji address the problem of determining the structures of viruses with unknown symmetry in their paper titled "A Computational Framework for the 3D Structure Determination of Viruses with Unknown Symmetry." Viruses are large macromolecules covered by a protein shell, and the binding sites of this protein shell determine which cells can be infected by the viruses. The authors present a computational framework for structure determination using experimental data obtained by cryo-transmission electron microscopy. The final paper in this special issue is titled "Blue Matter, an Application Framework for Molecular Simulation on Blue Gene" and is written by Blake G. Fitch, Robert S. Germain, Mark Mendell, Jed W. Pitera, Michael C. Pitman, Aleksandr Rayshubskiy, Yuk Sham, Frank Suits, William C. Swope, Chris Ward, Yuri Zhestkov, and Ruhong Zhou. Blue Gene is a massively parallel computer being designed by IBM for study of biomolecular phenomena. Although the machine is still being designed, the authors report on an application framework being developed for biomolecular simulation on the Blue Gene and report experimental results using simple performance models.

We hope that the reader will find the papers in this special issue informative and useful. The coverage of various subfields of computational biology represented by the selected papers is of course significantly influenced by the manuscripts that were submitted for the special issue. Hence, the special issue should not be viewed as providing comprehensive coverage of the breadth of research on this topic. It is our hope that the papers published in this special issue provide an insightful sampling of some of the interesting research work carried out by our colleagues combining parallel and distributed computing and computational biology. Some research areas in computational biology not represented by the papers in this special issue but which have attracted significant attention of the parallel processing community include sequence alignments, DNA fragment assembly, clustering of expressed sequence tags (ESTs), and string algorithms motivated by applications in computational biology.

For those of you who are interested in conducting research in this field or who simply wish to stay informed, we conduct an annual workshop on high performance computational biology (HiCOMB; see http://www.hicomb.org for details) in conjunction with the International Parallel and Distributed Processing Symposium. The parallel and distributed computing community has much to offer for solving challenging problems facing modern biology, and we strongly believe that such interdisciplinary collaborative research is very rewarding. We hope that the special issue holds out this message for the wider parallel and distributed computing community of researchers and accelerates research in high-performance computational biology.

Acknowledgments

The special issue editors thank the authors of all submitted manuscripts, without whom this special issue would not have been possible. We also thank the many anonymous reviewers who provided thorough evaluations of the submitted manuscripts despite the short amount of time in which such feedback was requested. We express our sincere gratitude to Sartaj Sahni for his guidance and support throughout the process of bringing out the special issue, in his capacity as the editor in charge of Theory, Algorithms, and Programming for the *Journal of Parallel and Distributed Computing*.

Srinivas Aluru is an associate professor and Associate Chair for Graduate Education in the Department of Electrical and Computer Engineering at Iowa State University. He is affiliated with the Laurence H. Baker Center for Bioinformatics and Biological Statistics, and serves as associate chair for the Bioinformatics and Computational Biology Graduate Program. Earlier, he held faculty positions at New Mexico State University and Syracuse University. He received his B.Tech degree in computer science from the Indian Institute of Technology, Chennai, India, in 1989 and his M.S. and Ph.D. in computer science from Iowa State University in 1991 and 1994, respectively. His research interests include parallel algorithms and applications, bioinformatics and computational biology, and combinatorial scientific computing. He is a recipient of an NSF CAREER Award, an IBM Faculty Award, and a Young Engineering Faculty Research Award from Iowa State University. Dr. Aluru served on program committees and has taken up other leadership roles at several conferences and workshops in the areas of parallel processing, scientific computing, and computational biology. He is a member of ACM, and a senior member of IEEE and IEEE Computer Society. He has co-authored a book and over 40 articles in peer-reviewed journals and conferences.

David A. Bader is an associate professor and Regents' Lecturer in the Department of Electrical and Computer Engineering of the University of New Mexico (UNM). He received his Ph.D. in electrical engineering in 1995 from the University of Maryland and was awarded a National Science Foundation (NSF) Postdoctoral Research Associateship in Experimental Computer Science before joining UNM in 1998. He is an NSF CAREER Award recipient, an investigator on six NSF awards including three ITR 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 is the general co-chair for IPDPS (2004–2005) and vice general chair for HiPC (2002–2003). He has served on numerous conference program committees related to parallel processing and is an associate editor for the *ACM Journal of Experimental Algorithmics* in the area of parallel algorithms, a senior member of the IEEE Computer Society, and a member of the ACM. Dr. Bader has given several keynote talks on high-performance computing for problems in computational genomics. He has co-authored over 38 articles in peer-reviewed journals and conferences, and his main areas of research are in parallel algorithms, combinatorial optimization, and computational biology and genomics.

Srinivas Aluru[a]
[a] *Department of Electrical and Computer Engineering, Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA* and

David A. Bader[b]
[b] *Department of Electrical and Computer Engineering, University of New Mexico, Albuguerfue, NM 87131, USA*