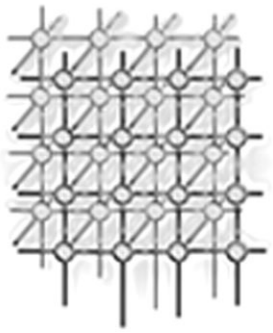
CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2004; **16**:817–821 (DOI: 10.1002/cpe.865)

Special Issue: High Performance Computational Biology



Ever since the structure of DNA was discovered in 1953, biology has been steadily changing from being 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 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 due to 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 are scaling 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 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. In order 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.

The *IEEE International Workshop on High-Performance Computation Biology* (*HiCOMB*, http://www.hicomb.org) has been established as the premier meeting for showcasing novel and important research results that use parallel and distributed computing techniques to address real biological problems. At this writing, workshop co-organizers Srinivas Aluru and David A. Bader have held two successful meetings: the first (*HiCOMB 2002*) was held in April 2002 in Fort Lauderdale, Florida, and the second (*HiCOMB 2003*) was held in April 2003 in Nice, France. Program Chair Dan C. Marinescu is organizing the third meeting for April 2004 in Santa Fe, New Mexico. These meetings are being held in conjunction with the *International Parallel and Distributed Processing Symposium* (IPDPS, http://www.ipdps.org) in the hope of stimulating interest in the field and attracting other researchers to participate in this field.

Following the *2nd IEEE International Workshop on High-Performance Computation Biology* (*HiCOMB 2003*), and with consultation of the 14 Program Committee members:

- Suchendra Bhandarkar, University of Georgia
- Alok Choudhary, Northwestern University
- David W. Deerfield II, Pittsburgh Supercomputing Center
- Gao Guang, University of Delaware
- Bruce Hendrickson, Sandia National Laboratories
- Joseph JáJá, University of Maryland
- Suraj Kothari, Iowa State University
- Timothy Mattson, Intel Corporation
- John Reynders, Celera
- Joel Saltz, Ohio State University
- Quinn Snell, Brigham Young University
- Stefan Unger, Sun Microsystems
- Geert Wenes, National Center for Genome Resources
- Albert Y. Zomaya, University of Western Australia

several authors were invited to submit extended manuscripts for this special issue. Each manuscript was reviewed by at least three, and in some cases four to five, expert reviewers. We are extremely grateful to all the reviewers who provided thoughtful reviews. Ten manuscripts were selected for publication in this special issue.

Studies in computational molecular biology often start with DNA sequences that require alignment for comparison and further analyses. The first paper 'Sequence alignment on the Cray MTA-2', by Bokhari and Sauer, discusses parallel approaches for exact matching and approximate matching using dynamic programming, for instance, using the DNA of *H. Influenzae* which has 1.8 million bases. Several variants of standard algorithms for DNA sequence alignment have been implemented on the Cray Multithreaded Architecture-2 (MTA-2). This work describes the architecture of the MTA-2 and discusses how its hardware and software enable efficient implementation of parallel algorithms with little or no regard for issues of partitioning, mapping or scheduling. It is shown that, for the dynamic programming algorithm, the use of the MTA's 'Full/Empty' synchronization bits leads to almost perfect speedup for large problems on 1–8 processors. These results illustrate the versatility of the MTA's architecture and demonstrate its potential for providing a high-productivity platform for parallel processing.

Progressive sequence model refinement by means of iterative searches is an effective technique for high-sensitivity database searches and is currently employed in popular tools such as PSI-BLAST and SAM. In the second paper, 'Using hybrid alignment for iterative sequence database searches', the authors Li, Lauria and Bundschuh propose a novel alignment algorithm that offers features expected to improve the sensitivity of such iterative approaches, specifically a well-characterized theory of its statistics even in the presence of position-specific gap costs. The authors demonstrate that the new hybrid alignment algorithm is ready to be used as the alignment core of PSI-BLAST and evaluate the accuracy of two proposed approaches to edge effect correction in short sequence alignment statistics that turns out to be one of the crucial issues in developing a hybrid-alignment based version of PSI-BLAST.

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. The design of primers for DNA Chip experiments is a time consuming process that extensively uses bioinformatics. The selection of the primers, which are immobilized on the DNA chip, requires a complex algorithm. Based on several parameters an optimized set of primers is automatically determined for a given gene sequence. Simmler, Singpiel and Männer present a 'Real-time primer design for DNA Chips'. This paper describes a parallel architecture which performs the optimization of the primer selection on a hardware accelerator. In contrast to the pure software approach, the parallel architecture gains a speedup of factor 500 using a PCI-based hardware accelerator. This approach allows an optimization of a specified primer set in real time.

In 'Development and implementation of a parallel algorithm for the fast design of oligonucleotide probe sets for diagnostic DNA microarrays', Meier, Krause, Kräutner and Bode describe an accurate method for the automatic parallel generation of oligonucleotide probe sets for DNA microarrays. This approach includes a component for high-performance specificity evaluation of designed probes in large data sets. The three main algorithmic components of the method, namely probe preselection, hybridization prediction, and probe selection, are explained in detail. The authors introduce new combinatorial techniques for the efficient selection of probe sets of high differentiation capability even from sequence databases of conserved homologous genes. These techniques include the automatic generation of group specific probes as well as the design of excluding probes. The applicability of their program is pointed out by designing a set of oligonucleotide probes that allow a comprehensive parallel identification and differentiation of several groups of extremophilic prokaryotes by DNA microarray.

Gene clustering, the process of grouping related genes in the same cluster, is at the foundation of different genomic studies that aim at analyzing the functions of genes. Microarray technologies have made it possible to measure gene expression levels for thousands of genes simultaneously. For knowledge to be extracted from the datasets generated by these technologies, the datasets have to be presented to a scientist in a meaningful way. Gene clustering methods serve this purpose. In 'A hybrid self organizing maps and particle swarm optimization approach', Xiao, Dow, Eberhart, Ben Miled and Oppelt propose a hybrid clustering approach that is based on self-organizing maps and particle swarm optimization, and implement it on a cluster of workstations. In the proposed algorithms, the rate of convergence is improved by adding a conscience factor to the self-organizing maps algorithm.

Understanding biochemical molecules, known as proteins, and their interactions, is crucial for the discovery of the mechanisms and pathways inside a living cell. Amino acids are the biochemical building blocks of proteins, and a short sequence of amino acids, called a peptide, is often compared

for similarity with a longer sequence representing a proteome. In 'A specialized hardware device for the protein similarity search', authors Marongiu, Palazzari and Rosato present the architecture of PROSIDIS, a special purpose processor designed to search for the occurrence of substrings similar to a given *template string* within a proteome. Similarity is determined by means of a weighting matrix which measures the degree of similarity between any two amino acids. The authors illustrate the advantages derived from designing a special purpose processor to face the protein similarity discovery problem. Some preliminary results are given, reporting the time spent by several conventional computing architectures and by the PROSIDIS processor hosted by a personal computer to solve the same protein analysis problem. The results show that PROSIDIS, implemented on a Xilinx XV1000 FPGA, gives speedup figures ranging from 5.6 up to 55.6.

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. A number of software packages are available for the construction of comprehensive human genetic maps. In 'Parallelization of IBD computation for determining genetic disease maps' Rizk presents a parallelization of the widely used package Genehunter, and uses the Message-Passing Interface (MPI) parallel environment to improve the performance of the function that performs computations of Identity by Descent (IBD) genes of a family.

The exponential growth rate of biological databases has established the need for high performance distributed computing in bioinformatics. Schmidt, Feng, Laud and Santoso, in the paper 'Development of distributed bioinformatics applications with gMP', present the design and use of gMP as a tool for developing distributed bioinformatics applications. gMP is a purely Java-based interface that adds MPI-like message-passing and collective communication to the genomics Research Network Architecture (gRNA). The gRNA is a highly programmable, modular environment specifically designed to invigorate the development of genome-centric tools for life science research. The authors demonstrate the development of a distributed application to detect regulatory elements using correlation with gene expression data. Its implementation with gMP leads to significant runtime savings on the distributed gRNA system.

Biological function of the proteins is determined by their three-dimensional shape, and their shape is determined by their linear sequence. The protein threading problem (PTP) is an extremely important challenge in computational biology. The problem consists of testing whether or not a target sequence query is likely to fold into a 3D template structure core by searching for an alignment which minimizes a suitable score function. In 'Parallel divide and conquer approach for the protein threading problem', Yanev and Andonov propose a network flow like formulation for the protein threading problem and show its equivalence with a generalization of the shortest path problem on a graph with a very particular structure. The underlying Mixed Integer Programming (MIP) model proves to be very appropriate for the PTP–huge real-life instances have been solved much faster by using only the MIP solver CPLEX instead of a known special-purpose branch and bound algorithm. The properties of the MIP model permit a decomposition of the main problem on a large number of subproblems (tasks). The authors show that a branch and cut strategy can be efficiently applied for solving in parallel these tasks, which leads to a significant reduction in the total running time.

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 histories of the group under study and is crucial to understanding 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 the phylogenetic distribution of variation in wild populations. The ambitious NSF *Tree of Life* program is aimed at discovering the complex evolutionary relationships between all known species on the planet. This special issues contains one paper on parallel methods for phylogenetic analysis, 'The AxML program family for maximum likelihood-based phylogenetic tree inference', by Stamatakis and Ludwig. Inference of phylogenetic (evolutionary) trees comprising hundreds or thousands of organisms based on the maximum likelihood criterion is a computationally extremely intensive task. This paper describes the evolution of the `AxML` program family which provides novel algorithmic as well as technical solutions for the maximum likelihood-based inference of huge phylogenetic trees. Algorithmic optimizations and a new tree building algorithm yield running time improvements of a factor greater than 4 compared with `fastDNAml` and *parallel* `fastDNAml`, returning equally good trees at the same time. Various parallel, distributed, and Grid-based implementations of `AxML` give the program the capability to acquire the large amount of required computational resources for the inference of huge high-quality trees.

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 and selected 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.

The parallel and distributed computing community has a lot to offer for solving challenging problems facing modern biology and we strongly believe that such interdisciplinary collaborative research is very rewarding. We hope that this special issue holds out this message for the wider parallel and distributed computing community of researchers and accelerates research in high-performance computational biology.

ACKNOWLEDGEMENTS

The special issue editors wish to 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 a thorough evaluation of the submitted manuscripts. We express our sincere gratitude to Manish Parashar, on the international editorial board for *Concurrency and Computation: Practice and Experience*, for his guidance and support throughout the process of bringing out the special issue, and to Geoffrey Fox in his capacity as the editor of this journal.

DAVID A. BADER
Department of Electrical and Computer Engineering,
University of New Mexico, Albuquerque, NM 87131, U.S.A.

SRINIVAS ALURU
Department of Electrical and Computer Engineering,
Laurence H. Baker Center for Bioinformatics and Biological Statistics,
Iowa State University, U.S.A.