Parallel Computing 42 (2015) 1–3

Contents lists available at ScienceDirect

Parallel Computing

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

Guest editorial

Parallelism in bioinformatics: A view from different parallelism-based technologies

In Bioinformatics, we can find a variety of problems which are affected by huge processing times and memory/storage consumption, due to the large size of biological data sets and the inherent complexity of biological problems. In fact, Bioinformatics is one of the most exciting research areas in which parallelism finds application. Successful examples are mpiBLAST, RAxML-HPC or ClustalW-MPI, among many others. In conclusion, Bioinformatics allows and encourages the application of many different parallelism-based technologies: multicore computing, cluster computing, cloud computing, hardware accelerators as GPUs, FPGAs, etc.

This special issue brings together high-quality state-of-the-art contributions about parallelism in bioinformatics, from different points of view, that is, from the use of different parallelism-based technologies. This special issue collects the best papers, accepted and presented in PBio 2013 (International Workshop on Parallelism in Bioinformatics, and part of EuroMPI 2013). These papers have been considerably extended and improved by the authors from their original versions.

At present, the application of parallelism in bioinformatics is a very popular research topic. As an example of the current interest in this field, it is worth mentioning that for this special issue we have managed a total of 29 high-quality submissions from many different countries: Germany, France, Italy, Spain, Russia, Portugal, China, Egypt, Saudi Arabia, Puerto Rico, etc. All the papers included in this special issue were reviewed by at least four expert reviewers. Furthermore, all the papers in the special issue received a minimum of two review rounds, where some manuscripts received up to five review rounds. Finally, 7 papers of high quality in emerging research areas were accepted for inclusion in the special issue (acceptance rate = 7/29 = 24.14%). In conclusion, we think these papers, with authors from around the world, bring us an international sampling of significant work.

The domains and topics covered in these 7 papers are timely and important, and the authors have done an excellent job of presenting the material. We are sure this special issue will be very useful for all the readers who are engaged in the many issues surrounding the application of parallelism in bioinformatics.

The title of our first paper is "Combined Hardware-Software Multi-Parallel Prefiltering on the Convey HC-1 for Fast Homology Detection", by Michael Bromberger, Fabian Nowak, and Wolfgang Karl. Protein databases used in research are huge and still grow at a fast pace. Many comparisons need to be done when searching similar (homologous) sequences for a given query sequence in these databases. To further reduce the runtime, bioinformatics tools use two-level prefiltering to reduce the number of time-consuming comparisons. Still, prefiltering is very time-consuming. Highly parallel architectures and huge bandwidth are required for processing and transferring the massive amounts of data. In this article, the authors present an approach exploiting the reconfigurable, hybrid computer architecture Convey HC-1 for migrating the most time-consuming part. The Convey HC-1 with four FPGAs and high memory bandwidth of up to 76.8 GB/s serves as the platform of choice. Limited by FPGA size only, the authors present a design that calculates four first-level prefiltering scores per FPGA concurrently, i.e. 16 calculations in total. This score calculation for the query profile against database sequences is done by a modified Smith–Waterman scheme that is internally parallelized 128 times in contrast to the original Streaming "Single Instruction Multiple Data (SIMD)" Extensions (SSE)-supported implementation where only 16-fold parallelism can be exploited. They tightly integrated the FPGA-based coprocessor into the hybrid computing system by employing task-parallelism for the two-level prefiltering. Despite much lower clock rates, the FPGAs outperform SSE-based execution for the calculation of the prefiltering scores by a factor of 7.9.

The second paper, "High Performance Computing Improvements on Bioinformatics Consistency-Based Multiple Sequence Alignment Tools" by Miquel Orobitg, Fernando Guirado, Fernando Cores, Jordi Llados, and Cedric Notredame, is focused on Multiple Sequence Alignment (MSA), an essential task for a wide range of applications in bioinformatics. With the growth of sequencing data, features such as performance and the capacity to align larger datasets, are gaining strength. To achieve these new requirements, without affecting accuracy, the use of high-performance computing (HPC) resources and tech-

http://dx.doi.org/10.1016/j.parco.2015.02.001
0167-8191/

niques is crucial. In this paper, the authors apply HPC techniques in T-Coffee, one of the more accurate but less scalable MSA tools. They integrate three innovative solutions into T-Coffee: the Balanced Guide Tree to increase the parallelism/performance, the Optimized Library Method with the aim of enhancing the scalability, and the Multiple Tree Alignment, which explores different alignments in parallel to improve the accuracy. The results obtained show that the resulting tool, MTA-TCoffee, is able to improve the scalability in both the execution time and also the number of sequences to be aligned. Furthermore, not only is the alignment accuracy not affected by these improvements, as would be expected, but it improves significantly. Moreover, it is important to highlight that the presented methods are not just restricted to T-Coffee, but may be implemented in any other alignment tools that use similar algorithms (progressive alignment, consistency or guide trees). The results showed that their approach decreases the memory requirements of T-Coffee by 75%, reduces the execution time by 92%, and finally, allows T-Coffee to align more than 2000 sequences while the standard T-Coffee is only able to align 1000 sequences.

Our third paper "Triangulating Molecular Surfaces over a LAN of GPU-Enabled Computers", authored by Sérgio E.D. Dias and Abel J.P. Gomes, presents an OpenMPI-OpenMP-CUDA solution for triangulating molecular surfaces. Computing, triangulating, and rendering molecular surfaces is an important topic in bioinformatics, computational biology, and biochemistry. Standalone GPU-enabled computers are adequate to triangulate and render molecular datasets with some tens of thousands of atoms at most. But, a standalone GPU-enabled computer has a limited capacity to host programmable graphics cards, which in turn have also their constraints in terms of memory space. Thus, in spite of the huge memory space made available and the tremendous processing power of the current GPU-based graphics cards, there remains a scalability problem when it is necessary to triangulate and render big molecules with hundreds of thousands to millions of atoms. In order to overcome this scalability problem the authors use an OpenMPI-OpenMP-CUDA solution that runs on a loosely-coupled GPU cluster over a LAN (Local Area Network). More specifically, they propose a fast, scalable, parallel triangulation algorithm for molecular surfaces that takes advantage of multicore processors of CPUs and GPUs available over a local network, with each CPU core working as the master of a single GPU. In conclusion, one of the main contributions of this paper is that likely introduces the first marching cubes algorithm that triangulates molecular surfaces on CUDA devices over a network of GPU-enabled computers.

The parallelization of proteins docking is addressed in the fourth paper, "Inverse Docking Method for New Proteins Targets Identification: A Parallel Approach" by Romain Vasseur, Stéphanie Baud, Luiz Angelo Steffenel, Xavier Vigouroux, Laurent Martiny, Michaël Krajecki, and Manuel Dauchez. Molecular docking is a widely used computational technique that allows studying structure-based interactions complexes between biological objects at the molecular scale. The purpose of this article is to develop a set of tools that allows performing inverse docking, i.e., to test at a large scale a chemical ligand on a large dataset of proteins, which has several applications on the field of drug research. The authors developed different strategies to parallelize/distribute the docking procedure, as a way to efficiently exploit the computational performance of multi-core and multi-machine (cluster) environments. The experiments conducted to compare these different strategies encourage the search for decomposing strategies since it improves the execution of inverse docking. Their success lies in the optimization of crucial parameters such as the number of 3D generated boxes, their dimensions (chosen depending on the size and shape of the proteins), their overlap, and the number of simulations runs (according to the 3D box volume to explore). Inverse docking methods are yet new approaches and this article likely presents the first tools that allow really performing large-scale inverse "blind" docking on HPC environments. It is important to remind that inverse/reverse docking is one of the main new topics that are animating the molecular docking community.

The fifth paper ("Geometrical Motifs Search in Proteins: A Parallel Approach" by Marco Ferretti and Mirto Musci) focuses on the analysis of the 3D structures of proteins. This is a very important problem in life sciences, since the geometric set-up of proteins has a deep relevance in many biological processes. The complexity of the analysis and the continuous increase in the number of proteins whose 3D structure is known, call for efficient and quick algorithms. Parallel processing is becoming an enabling tool for such research. A key component in the geometric description of a protein is the structural motif, a 3D element which appears in a variety of molecules and is usually made of just a few simpler structures, the secondary structures elements. This paper presents the Cross Motif Search (CMS) and the Complete CMS (CCMS) algorithms, two highly optimized and efficient parallel methods to detect the presence and location of all common motifs of secondary structures in a given protein pair (CMS) or across an arbitrary large dataset of proteins (CCMS). The main difference between their proposal and the state of the art is the innovative focus that CMS puts on the geometric description of the structural motifs, rather than on the topological/biological description employed by competing algorithms. The advantage of a geometrical approach is that it enables to retrieve the exact location of the common substructures in a protein pair. The paper analyzes all possible forms of serial and parallelism optimization of the proposed algorithms, both shared memory and message passing. It introduces a complete parallel implementation of CMS, based on OpenMP, and discusses its scalability on shared-memory architectures. Both small-scale and medium-scale testing shows that the methods produce very interesting results in real applications.

The sixth paper "High Performance Solutions for Big-data GWAS" is authored by Elmar Peise, Diego Fabregat-Traver, and Paolo Bientinesi. In order to associate complex traits with genetic polymorphisms, genome-wide association studies (GWAS) process huge datasets involving tens of thousands of individuals genotyped for millions of polymorphisms. When handling these datasets, which exceed the main memory of contemporary computers, one faces two distinct challenges: (1) millions of polymorphisms and thousands of phenotypes come at the cost of hundreds of gigabytes of data, which can only be kept in secondary storage; (2) the relatedness of the test population is represented by a relationship matrix, which, for large

populations, can only fit in the combined main memory of a distributed architecture. In this paper, by using distributed resources such as Cloud or clusters, the authors address both challenges: the genotype and phenotype data is streamed from secondary storage using the double-buffering technique, while the relationship matrix is kept across the main memory of a distributed memory system. With the help of these solutions, they develop separate algorithms for studies involving only one or a multitude of traits. The results show that these algorithms sustain high-performance (at least one order of magnitude faster than other wide-spread GWAS-codes) and allow the analysis of enormous datasets (in fact, their implementation scales in all problem sizes). Therefore, these algorithms form a viable basis for the challenges posed by the scale of current and future genome-wide association studies.

The spatial transmission of influenza (commonly known as "the flu") is tackled in the seventh paper ("Towards Efficient Large Scale Epidemiological Simulations in EpiGraph" by Gonzalo Martín, David E. Singh, Maria-Cristina Marinescu, and Jesús Carretero). Specifically, the authors focus on understanding the propagation of flu-like infectious outbreaks between geographically distant regions due to the movement of people outside their base location. Understanding the patterns that viruses, such as influenza, follow when they propagate among the population of widely-spread geographic regions is fundamental for an agile response of public health authorities. The authors' approach incorporates geographic location and a transportation model into their existing region-based, closed-world EpiGraph simulator to model a more realistic movement of the virus between different geographic areas. This paper describes the MPI-based implementation of this simulator, including several optimization techniques such as: data partitioning methods that exploit data locality and enable load balance, inter-process communication optimization based on a two-level MPI communicator schema, and a novel approach for mapping processes onto available processing elements based on the temporal distribution of process loads. They present an extensive evaluation of EpiGraph in terms of its ability to simulate large-scale scenarios, as well as from a performance perspective. In particular, they include an experimental evaluation of EpiGraph when simulating 92 urban regions in Spain consisting of 21,320,965 inhabitants. The results show that they can scale their simulations to run efficiently over large areas.

We sincerely hope that you enjoy this special issue. We also have hope that the paper collection as a whole can pleasantly introduce the readers to the composite and challenging arena of the application of parallelism in bioinformatics, and can help in giving a fresh view of several state-of-the-art solutions from diverse domains. Before concluding we want to express our sincere gratitude to some people who have helped us in this challenge. First of all, we would like to thank Prof. Dr. Jeffrey K. Hollingsworth (Editor-in-Chief of Parallel Computing journal) for trusting us, as well as for all his help. This special issue would not have been possible without the assistance of some people from Elsevier who have helped us during its preparation: Iswarya Samikannu, Hilda Xu, Jessey Huyan, Mary Shyla Sivasubramaniyam, and other people; to whom we give many thanks. We extend our sincere thanks to all the authors who submitted papers for this special issue and the many reviewers, whose dedicated efforts made this special issue possible.

Guest editors
Miguel A. Vega-Rodríguez
David L. González-Álvarez
Department of Technologies of Computers and Communications,
University of Extremadura, Escuela Politécnica,
Campus Universitario s/n, 10003 Cáceres, Spain
E-mail addresses: mavega@unex.es (M.A. Vega-Rodríguez), dlga@unex.es (D.L. González-Álvarez)
URLs: http://arco.unex.es/mavega (M.A. Vega-Rodríguez), http://arco.unex.es/dlga (D.L. González-Álvarez)