Parallelism-based approaches in computational biology: a view from diverse case studies

Miguel A. Vega-Rodríguez¹ · Sergio Santander-Jiménez²

Published online: 14 July 2017
© Springer Science+Business Media, LLC 2017

Abstract Computational biology allows and encourages the application of many different parallelism-based approaches. This special issue brings together high-quality state-of-the-art contributions about parallelism-based approaches in computational biology, from different points of view or perspectives, that is, from diverse case studies. The special issue collects considerably extended and improved versions of the best papers, accepted and presented in PBio 2016 (4th International Workshop on Parallelism in Bioinformatics, and part of Euro-Par 2016). The domains and topics covered in these 6 papers are timely and important, and the authors have done an excellent job of presenting the material.

Keywords Parallelism · Computational biology · Case studies

1 Introduction

In computational biology, 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, computational biology is among the most exciting research areas in which parallelism finds applica-

tion. Successful examples are mpiBLAST, RAxML-HPC or ClustalW-MPI, among many others. Therefore, computational biology allows and encourages the application of many different parallelism-based approaches: multicore computing, cluster computing, supercomputing, cloud computing, grid computing, green computing, hardware accelerators such as GPUs, FPGAs, etc.

This special issue brings together high-quality state-of-the-art contributions about parallelism-based approaches in computational biology, from different points of view or perspectives, that is, from diverse case studies. This special issue collects the best papers, accepted and presented in [1] (4th International Workshop on Parallelism in Bioinformatics, and part of Euro-Par 2016). These papers have been considerably extended and improved by the authors from their original conference versions.

2 Important data

At present, the application of parallelism in computational biology 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 18 high-quality submissions from different countries, such as USA, Germany, Spain, Iran or Taiwan. All the papers included in this special issue were reviewed by at least five expert reviewers. Furthermore, all the papers in the special issue received a minimum of three review rounds. Finally, 6 papers of high quality in emerging research areas were accepted for inclusion in the special issue (acceptance rate = $6/18 = 33.33\%$). In conclusion, we think these papers bring us an international sampling of significant work.

The domains and topics covered in these six papers are timely and important, and the authors have done an excellent

✉ Miguel A. Vega-Rodríguez
mavega@unex.es

Sergio Santander-Jiménez
sesaji@unex.es

¹ Department of Computer and Communications Technologies,
University of Extremadura, Escuela Politécnica, Campus
Universitario s/n, Cáceres 10003, Spain

² INESC-ID, Instituto Superior Técnico, Universidade de
Lisboa, Lisboa 1000-029, Portugal

job of presenting the material. We are confident that this special issue will be very useful for all the readers who are engaged in the many issues surrounding the application of parallelism in the computational biology domain.

3 Detailed content

The title of our first paper is “Optimization of SAMtools sorting using OpenMP tasks”, by Weeks and Luecke (see [2]). SAMtools is a widely-used genomics application for post-processing high-throughput sequence alignment data. Such sequence alignment data are commonly sorted to make the analysis more efficient. This paper describes a case study on the performance analysis and optimization of SAMtools for sorting large BAM (binary alignment/map) files. OpenMP task parallelism and memory optimization techniques resulted in a speedup of $5.9\times$. Performance improvements to this important tool have the potential to reduce time to solution for a variety of genomics problems facing agriculture, oncology, pathology, and other life sciences.

The second paper, “parallel high-dimensional multi-objective feature selection for EEG classification with dynamic workload balancing on CPU-GPU architectures” by Escobar, Ortega, González, Damas, and Díaz (see [3]), is focused on taking advantage of CPU-GPU heterogeneous architectures to accelerate electroencephalogram (EEG) classification and feature selection problems by evolutionary multi-objective optimization, in the context of brain computing interface (BCI) tasks. In this paper, the authors have used the OpenCL framework to develop parallel master-worker codes implementing an evolutionary multi-objective feature selection procedure in which the individuals of the population are dynamically distributed among the available CPU and GPU cores.

Our third paper “speed and accuracy improvement of higher-order epistasis detection on CUDA-enabled GPUs”, authored by Jünger, Hundt, González-Domínguez, and Schmidt (see [4]), combines a novel preprocessing filter, namely SingleMI, with massively parallel computation on modern GPUs to further accelerate epistasis discovery. The discovery of higher-order epistatic interactions is an important task in the field of genome wide association studies which allows for the identification of complex interaction patterns between multiple genetic markers. These studies are focused on the prediction of specific diseases, therefore, they are very important. The authors have used CUDA for parallelising their software.

The exploration of the trade-offs between data locality and data dispersion in NUMA systems is addressed in the fourth paper, “a performance comparison of data and memory allocation strategies for sequence aligners on NUMA

Architectures” by Lenis and Senar (see [5]). They have performed experiments with several popular sequence alignment tools on two widely available NUMA systems to assess the performance of different memory allocation policies and data partitioning strategies. They conclude that memory interleaving is the memory allocation strategy that provides the best performance when a large number of processors and memory banks are used. In the case of data partitioning, the best results are usually obtained when the number of partitions used is greater, sometimes combined with an interleave policy.

The fifth paper, entitled “two level parallelism and I/O reduction in genome comparisons” by Torreno and Trelles (see [6]), applies two different strategies to accelerate GECKO (a pairwise and multiple genome comparison modular application) while producing the same results. First, a two-level parallel approach parallelising each independent internal pairwise comparison in the first level, and the GECKO modules in the second level. A second approach consists on a complete rewrite of the original code to reduce I/O. Both strategies outperform the original code, which was already faster than equivalent software. Thus, much faster pairwise and multiple genome comparisons can be performed, what is really important with the ever-growing list of available genomes.

The sixth paper “a cloud-based enhanced differential evolution algorithm for parameter estimation problems in computational systems biology” is authored by Teijeiro, Pardo, Penas, González, Banga, and Doallo (see [7]). The authors propose a parallel implementation of an enhanced Differential Evolution algorithm using Spark. The proposal drastically reduces the execution time, by means of including a selected local search and exploiting the available distributed resources. The performance of the proposal has been thoroughly assessed using challenging parameter estimation problems from the domain of computational systems biology. Two different platforms have been used for the evaluation, a local cluster and the Microsoft Azure public cloud. Additionally, it has been also compared with other parallel approaches, another cloud-based solution (a MapReduce implementation) and a traditional HPC solution (an MPI implementation).

4 Conclusion

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 area of the application of parallelism in computational biology, giving a fresh view of several state-of-the-art solutions from diverse perspectives. 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.

C.S. Raghavendra (Editor-in-Chief for Special Issues of the Cluster Computing journal) and Prof. Dr. S. Hariri (Editor-in-Chief of the Cluster Computing journal), for trusting us, as well as for all their help. We also 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.

Acknowledgements This work was partially funded by the AEI (State Research Agency, Spain) and the ERDF (European Regional Development Fund, EU), under the contract TIN2016-76259-P (PROTEIN project). Sergio Santander-Jiménez is supported by the Post-Doctoral Fellowship from FCT (Fundação para a Ciência e a Tecnologia, Portugal) under grant SFRH/BPD/119220/2016.

References

1. PBio: 4th International Workshop on Parallelism in Bioinformatics. <http://arco.unex.es/mavega/pbio/2016/> (2016). Accessed 22 June 2017
2. Weeks, N.T., Luecke, G.R.: Optimization of SAMtools sorting using OpenMP tasks. *Clust. Comput.* (2017). doi:[10.1007/s10586-017-0874-8](https://doi.org/10.1007/s10586-017-0874-8)
3. Escobar, J.J., Ortega, J., González, J., Damas, M., Díaz, A.F.: Parallel high-dimensional multi-objective feature selection for EEG classification with dynamic workload balancing on CPU-GPU architectures. *Clust. Comput.* (2017). doi:[10.1007/s10586-017-0980-7](https://doi.org/10.1007/s10586-017-0980-7)
4. Jünger, D., Hundt, C., González-Domínguez, J., Schmidt, B.: Speed and accuracy improvement of higher-order epistasis detection on CUDA-enabled GPUs. *Clust. Comput.* (2017). doi:[10.1007/s10586-017-0938-9](https://doi.org/10.1007/s10586-017-0938-9)
5. Lenis, J., Senar, M.A.: A performance comparison of data and memory allocation strategies for sequence aligners on NUMA architectures. *Clust. Comput.* (2017). doi:[10.1007/s10586-017-1015-0](https://doi.org/10.1007/s10586-017-1015-0)
6. Torreno, O., Trelles, O.: Two level parallelism and I/O reduction in genome comparisons. *Clust. Comput.* (2017). doi:[10.1007/s10586-017-0873-9](https://doi.org/10.1007/s10586-017-0873-9)
7. Teijeiro, D., Pardo, X.C., Penas, D.R., González, P., Banga, J.R., Doallo, R.: A cloud-based enhanced differential evolution algorithm for parameter estimation problems in computational systems biology. *Clust. Comput.* (2017). doi:[10.1007/s10586-017-0860-1](https://doi.org/10.1007/s10586-017-0860-1)

Miguel A. Vega-Rodríguez received the PhD degree in Computer Engineering from the University of Extremadura, Spain, in 2003. He is currently an Associate Professor (accredited as Full Professor) of computer architecture in the Department of Computer and Communications Technologies, University of Extremadura. He has authored or co-authored more than 620 publications including journal papers (more than 120 JCR-indexed journal papers), book

chapters, and peer-reviewed conference proceedings, for which he got several awards - such as Best Paper Awards in ISDA'11, ICEC'09, and IEA-AIE'08. He has contributed to the organization of several international conferences and workshops, namely as general chair or co-chair. He has edited 10 special issues of international JCR-indexed journals. In addition, he is an editor and a reviewer of diverse international JCR-indexed journals. His main research interests include parallel and distributed computing, bioinformatics, reconfigurable and embedded computing, and evolutionary computation.

Sergio Santander-Jiménez received the Ph.D. degree in Computer Engineering from the University of Extremadura, Spain, in 2016. He is currently a postdoctoral fellow at the R&D Instituto de Engenharia de Sistemas e Computadores (INESC-ID), Instituto Superior Técnico (IST), Universidade de Lisboa (UL), Portugal. He has co-organized several international workshops on high-performance computing, computational intelligence, computational biology and bioinformatics, reviewing articles on these topics for different international JCR-indexed journals. His main research interests include evolutionary and bioinspired computing, multi-objective optimization, parallel and distributed computing, and their applications to real-world biological problems.