Computational Biology

Endorsed by the *International Society for Computational Biology*, the *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

Fahad Saeed · Muhammad Haseeb

High-Performance Algorithms for Mass Spectrometry-Based Omics

Springer

Fahad Saeed
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, FL, USA

Muhammad Haseeb
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, FL, USA

ISSN 1568-2684 ISSN 2662-2432 (electronic)
Computational Biology
ISBN 978-3-031-01959-3 ISBN 978-3-031-01960-9 (eBook)
https://doi.org/10.1007/978-3-031-01960-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Fahad Saeed dedicates this book to Saba, Haadi, and Emaan.

Preface

To date, the processing of high-throughput Mass Spectrometry (MS) data is primarily accomplished using serial algorithms. Developing new methods to process MS data is an active area of research [1], but there is no single strategy that focuses on *scalability* of MS-based methods [2]. MS is a diverse and versatile technology for high-throughput functional characterization of proteins, small molecules, and metabolites in complex biological mixtures. In the recent years, the technology has rapidly evolved and is now capable of generating increasingly large (multiple terabytes per experiment) [1] and complex (multiple species/microbiome/high-dimensional) data sets [3]. This rapid advances in MS instrumentation must be matched by equally fast and rapid evolution of scalable methods developed for the analysis of these complex data sets. Ideally, the new methods should leverage the rich heterogeneous computational resources available in a ubiquitous fashion in the form of multicore, manycore, CPU-GPU, CPU-FPGA, and IntelPhi architectures. The absence of these high-performance computing algorithms now hinders scientific advancements in MS research [2].

In systems, biology setting workflows (or pipelines) are frequently used which are a sequence of loosely connected computational tasks. Database-search workflows are the most commonly used data processing pipelines which require matching a high-dimensional noisy MS data (called spectra) to a database of protein sequences. The entire workflow is executed using as a script-like structure that executes different algorithms which is then run on a dedicated workstation. The data volume can easily reach terabyte level depending on the experiment and search parameters for these workflows. The currently used state-of-the-art serial and parallel methods are data (and communication cost) oblivious which may not give the best possible performance for these database-search workflows. Currently used state-of-the-art serial algorithms results in unusually long processing times.

Development of parallel computing techniques to deal with this deluge of data has also been limited and has mostly adopted the batch mode (or embarrassingly parallel computing) form of processing. As one might imagine there have been some efforts in developing HPC algorithms that can be used for speeding up the processing. However, most of these efforts have been limited to parallelization of specific algorithms or

workflows without the ability to generalize the end-to-end performance for other existing or new algorithms. In order to take the field of computational proteomics forward and set it afoot with more mature fields such as genomics more concerted HPC efforts. This has resulted in limited speedups (e.g. 30x speedup for 200 cores) and consequently limited usage by proteomics practitioners who might not see the advantage of trivial reduction in processing times.

Our own preliminary study suggests that such workflows when used with data divided among compute nodes in an oblivious manner lead to unbalanced workload and results in hours to weeks of computation with the state-of-the-art software [4]. Therefore, concerted efforts are needed for the development of high-performance computing (HPC) framework for working with large mass spectrometry data sets while benefiting advancements in areas of science including proteomics, proteogenomics, meta-proteomics, and microbiomes. These frameworks must be able to leverage the vast HPC heterogeneous architectures that are ubiquitous in the form of desktops, laptops, clusters, and supercomputers.

The scientific premise of this project is that progress can be gained in developing scalable MS-based omics data analysis tools for non-model organism proteomics/meta-proteomics/proteogenomic will require: (1) Improved data-partitioning strategies allowing minimization of data communication between different levels of memory hierarchy and processing units; (2) Improved parallel algorithms on distributed-memory architectures to address the scalability limitations due to excessive communication costs; (3) Improved parallel algorithms for CPU-GPU/CPU-FPGA architecture to exploit heterogeneity of modern HPC machines; (4) Integration of these parallel algorithms to XSEDE Supercomputers Gateways will make our methods available for large-scale omics system biology studies.

To address these challenges, we will formulate and develop MS-specific parallel computing abstraction that incurs minimal I/O times on the multitude of heterogeneous architectures. Second, to address the diversity of the mass spectrometry user community, we will formulate HPC frameworks that supports scaling down analysis (i.e. working with large data files on relatively inexpensive hardware such as CPU-GPU without fully loading them into memory), as well as scaling up (i.e. ability to execute the workflows on large XSEDE supercomputers and cloud computing infrastructure). By ensuring the compatibility of mass spectrometry-specific data standards, supporting the number of heterogeneous architectures for efficient processing and generating optimized code bases in open-source format will enable the development of scalable analytical methods. Therefore, our framework aims to democratize access to HPC infrastructure for a broader community of life and systems biology scientists, and create a blueprint for a new paradigm for HPC computing for large MS data sets—making HPC the fourth pillar of scientific investigations.

Miami, FL, USA Fahad Saeed
January 2022

References

1. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI (2017) Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat Methods 14(5):513
2. Haseeb M, Afzali F, Saeed F (2019) Lbe: A computational load balancing algorithm for speeding up parallel peptide search in mass-spectrometry based proteomics. In: IEEE International parallel and distributed processing symposium workshops (IPDPSW), IEEE, 2019, pp 191–198
3. Tariq MU, Saeed F (2021) Specollate: Deep cross-modal similarity network for mass spectrometry data based peptide deductions. PloS one 16(10), e0259349
4. Awan MG, Saeed F (2016) Ms-reduce: an ultrafast technique for reduction of big mass spectrometry data for high-throughput processing. Bioinformatics 32(10):1518–1526

Acknowledgements

The research presented was supported by the NIGMS of the National Institutes of Health (NIH) under award number: R01GM134384. The authors were further supported by the NSF under award number: NSF CAREER OAC-1925960. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and/or the NSF. This work used the National Science Foundation (NSF) XSEDE supercomputers through allocations TG-CCR150017 and TG-ASC200004. The authors would also like to acknowledge hardware donation by Intel Altera (DE10-PRO-SX FPGA) and NVIDIA (TITAN Xp GPU).

Contents