

Computational Biology

Advisory Editors

Gordon Crippen, University of Michigan, Ann Arbor, MI, USA
Joseph Felsenstein, University of Washington, Seattle, WA, USA
Dan Gusfield, University of California, Davis, CA, USA
Sorin Istrail, Brown University, Providence, RI, USA
Thomas Lengauer, Max Planck Institute for Computer Science, Saarbrücken, Germany
Marcella McClure, Montana State University, Bozeman, MT, USA
Martin Nowak, Harvard University, Cambridge, MA, USA
David Sankoff, University of Ottawa, Ottawa, ON, Canada
Ron Shamir, Tel Aviv University, Tel Aviv, Israel
Mike Steel, University of Canterbury, Christchurch, New Zealand
Gary Stormo, Washington University in St. Louis, St. Louis, MO, USA
Simon Tavaré, University of Cambridge, Cambridge, UK
Tandy Warnow, University of Illinois at Urbana-Champaign, Urbana, IL, USA
Lonnie Welch, Ohio University, Athens, OH, USA

Editors-in-Chief

Andreas Dress, CAS-MPG Partner Institute for Computational Biology, Shanghai, China
Michal Linial, Hebrew University of Jerusalem, Jerusalem, Israel
Olga Troyanskaya, Princeton University, Princeton, NJ, USA
Martin Vingron, Max Planck Institute for Molecular Genetics, Berlin, Germany

Editorial Board

Robert Giegerich, University of Bielefeld, Bielefeld, Germany
Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Gene Myers, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
Pavel Pevzner, University of California, San Diego, CA, USA

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

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>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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.

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

1	Need for High-Performance Computing for MS-Based Omics Data Analysis	1
	Fahad Saeed and Muhammad Haseeb	
	References	4
2	Introduction to Mass Spectrometry Data	7
	Fahad Saeed and Muhammad Haseeb	
2.1	Proteomics	7
2.1.1	Mass Spectrometry-Based Proteomics	7
2.1.2	MS/MS Data Pre-processing	10
2.1.3	Peptide Identification	10
2.2	Proteogenomics	12
	References	14
3	Existing HPC Methods and the Communication Lower Bounds for Distributed-Memory Computations for Mass Spectrometry-Based Omics Data	21
	Fahad Saeed and Muhammad Haseeb	
3.1	Introduction	21
3.2	Communication Model	23
3.2.1	Sequential Computer	24
3.2.2	Parallel Computer	24
3.3	MS Database Proteomics, Proteogenomics, and Meta-Proteomics Search	24
3.3.1	Generalized Parallel Computing Strategy	25
3.4	Communication Lower Bounds	26
3.5	Meta-Analysis of Results of Current HPC Methods	29
3.6	Discussions	32
3.7	Conclusions	33
	References	34

4	High-Performance Computing Strategy Using Distributed-Memory Supercomputers	37
	Fahad Saeed and Muhammad Haseeb	
4.1	Introduction	37
4.1.1	Background	38
4.1.2	Problem Statement	38
4.2	The HiCOPS Framework	39
4.2.1	Database Indexing	39
4.2.2	Experimental Data Pre-processing	41
4.2.3	Parallel Database Peptide Search	41
4.2.4	Assembling the Local Results	42
4.3	Optimizations	42
4.3.1	Task Scheduling	42
4.3.2	Communication Optimization	43
4.4	Results	44
4.4.1	Experimental Settings	44
4.4.2	Correctness Analysis	45
4.4.3	Speed Comparison	46
4.4.4	Performance Evaluation	47
4.5	Discussion	50
	References	55
5	Fast Spectral Pre-processing for Big MS Data	57
	Fahad Saeed and Muhammad Haseeb	
5.1	A Review of Spectral Pre-processing Methods	57
5.1.1	Spectral Denoising Algorithms	58
5.1.2	Spectral Quality Assessment Algorithms	59
5.1.3	Separation of b-y Ions	59
5.2	MS-REDUCE: An Ultra-Fast Data Reduction Algorithm for Big MS Data	60
5.2.1	Spectral Classification	61
5.2.2	Spectral Quantization	63
5.2.3	Weighted Random Sampling	64
5.3	Performance Evaluation of MS-REDUCE	66
5.3.1	Time Complexity	67
5.3.2	Experimental Verification of the Complexity Analysis	67
5.3.3	Speed Comparison	68
5.3.4	Comparing MS-REDUCE with Other Denoising Methods	69
5.3.5	Quality Assessment	69
5.3.6	Comparison with Random Sampling of Peaks	70
5.3.7	Comparison with Conventional Algorithms	71
	References	74

6	A Easy to Use Generalized Template to Support Development of GPU Algorithms	77
	Fahad Saeed and Muhammad Haseeb	
6.1	GPU Architecture and CUDA	78
6.1.1	CUDA Overview	79
6.1.2	CPU-GPU Computing	79
6.2	Challenges in GPU Algorithm Design	80
6.2.1	Need for Data Parallel Design	80
6.2.2	Data Transfer Bottlenecks	80
6.2.3	Non-coalesced Memory Accesses	81
6.2.4	Warp Divergence	81
6.2.5	Exploiting Coarse Grained and Fine Grained Parallelism	81
6.3	Basic Principles of GPU-DAEMON	81
6.3.1	Simplifying Complex Data Structures	82
6.3.2	Simplifying Complex Computations	83
6.3.3	Efficient Array Management in GPU	83
6.3.4	Exploiting Shared Memory	84
6.3.5	In-Warp Optimizations	84
6.3.6	Result Sifting	85
6.3.7	Post Processing Results	85
6.3.8	Time Complexity Model for GPU-DAEMON	85
	References	86
7	Computational CPU-GPU Template for Pre-processing of Floating-Point MS Data	89
	Fahad Saeed and Muhammad Haseeb	
7.1	Simplifying Complex Data Structures	89
7.2	Efficient Array Management	90
7.2.1	Splitter Selection	90
7.2.2	Bucketing	91
7.3	In-Wrap Optimizations and Exploiting Shared Memory	92
7.4	Time Complexity Model	92
7.5	Performance Evaluation	93
7.5.1	Sorting Using Tagged Approach (STA)	93
7.5.2	Runtime Analysis and Comparisons	94
7.5.3	Data Handling Efficiency	94
	References	97
8	G-MSR: A GPU-Based Dimensionality Reduction Algorithm	99
	Fahad Saeed and Muhammad Haseeb	
8.1	G-MSR Algorithm	99
8.1.1	Simplifying Complex Data Structures	101
8.1.2	Simplifying Complex Computations	101
8.1.3	Efficient Array Management	102
8.1.4	Exploiting Shared Memory	102

8.1.5	In-Warp Optimizations	102
8.1.6	Result Sifting	103
8.1.7	Post Processing Results	103
8.2	Results and Experiments	103
8.2.1	Time Complexity Model	103
8.2.2	Experiment Setup	104
8.2.3	Scalability and Time Analysis	105
8.2.4	Quality Assessment	105
8.2.5	Reductive Proteomics for high-resolution instruments	106
8.2.6	Comparison with Unified Memory	107
	References	110
9	Re-configurable Hardware for Computational Proteomics	111
	Fahad Saeed, Muhammad Haseeb, and Sumesh Kumar	
9.1	Introduction	111
9.1.1	Construction of a Field-Programmable Gate Array	111
9.2	Popular Architectural Configurations Using FPGAs	112
9.2.1	Systolic Array Configuration	113
9.2.2	Parallel Asynchronous PEs Connected to the System Bus	114
9.2.3	Parallel Processors with Communication Interconnect	114
9.3	FPGA Design for Computational Proteomics	116
9.3.1	Architecture Overview	117
9.3.2	Processing Element (PE)	118
9.3.3	Bus-Arbitration Module	119
9.3.4	Binary Search Module	119
9.3.5	Ion-Matching Circuit	120
9.3.6	Experiments and Results	121
9.4	Conclusion	124
10	Machine-Learning and the Future of HPC for MS-Based Omics	125
	Fahad Saeed and Muhammad Haseeb	
10.1	Why HPC is Essential for Machine-Learning Models	126
10.2	Preliminary Data and Findings	127
	References	128
	Glossary	131
	Index	137