Chapter 1
Need for High-Performance Computing for MS-Based Omics Data Analysis

Fahad Saeed and Muhammad Haseeb

For the past 30 years, significant efforts were invested for the development of designing and implementing more efficient scoring functions which included highly successful search engines [1–5]. Similar to other domains, numerical algorithms were developed for Mass Spectrometry (MS)-based peptide deduction and are designed and implemented by assuming *number of arithmetic operations* as the sole metric for efficiency. In the last decade, the technological trend of Moore's law has kept making the arithmetic operations faster. As a result, the bottleneck for many algorithms have shifted from computational arithmetic operations efficiency to communication, i.e. communication costs of either moving the data between different levels of memory hierarchy (e.g. RAM, cache) or between different distributed-memory processors connected via a network. There are numerous [6–8] studies that have shown this trend that the cost of moving data exceeds the costs of doing the arithmetic operations [6]. This gap is, and will continue to grow exponentially over time with the introduction of multicore, manycore, and GPU architectures [8–10].

To date, processing of high-throughput MS data is accomplished mostly using serial algorithms. Recent trends in systems biology (e.g. meta-proteomics, non-model proteomics, microbiomes) MS-based experiments point towards the need for larger, better, and faster computational tools. This trend is fed by the MS technology that has rapidly evolved and is now capable of generating increasingly large and complex (multiple species/microbiome/high-dimensional) data sets [11]; leading to high-impact proteomics, meta-proteomics and microbiomes studies directly related to human disease and health. This rapid advances in MS instrumentation must be matched by equally rapid evolution of scalable methods developed for the analysis of these complex data sets. Developing new methods to process MS data is an active area of research [5] but the focus of this research, till date, has been towards improving the efficiency of arithmetic operations. Exclusion of communication costs as a metric, of otherwise highly successful methods, is becoming a severe bottleneck for

F. Saeed and M. Haseeb, *High-Performance Algorithms for Mass Spectrometry-Based Omics*, Computational Biology 34,
https://doi.org/10.1007/978-3-031-01960-9_1

processing of MS data, and now hinders scientific advancements for microbiome, and (meta) proteomics/microbiome research [12] workflows are the most commonly used data processing pipelines that match the high-dimensional noisy MS data (called spectra) to a database of protein sequences. These MS data sets are then processed using databases which may be several times larger than the original proteome (or a combination of multiple proteomes in case of meta-proteomics studies) depending on the search parameters. The data volume can easily reach terabyte level depending on the experiment and search parameters for these workflows.

Comparison across literature indicates decreased scalability of the serial algorithms. The scalability issues for these serial methods are solved using various filtering mechanisms but this has shown increased misidentification of peptides and/or inconsistencies between various search engines [13, 14]. It is not uncommon to have a terabyte scale database against which millions of raw experimental spectra need to be processed for model organisms (e.g. Rattus) and a few dynamic modification as search parameters. Non-model organism is considered the next frontier to accelerate insight into chronic disease in humans [15] requires even larger search spaces which would lead to intractable runtimes for both serial algorithms and traditional HPC methods for MS-based omics.

Increasing size of the spectra and theoretical database search space has led to the development of high-performance computing strategies [16–21] to speed up these search engines. However, similar to serial numerical algorithms, the objective of these HPC methods has been to speed up the arithmetic scoring part of the search engines with little to no efforts to minimize the communication costs. However, within-study direct comparisons of serial versus high-performance computing (HPC) proteomics algorithms have revealed modest gains [22]. Cloud computing-based processing, alone, cannot fill the gap of scalable tools since a non-efficient tool on desktop or cluster is non-efficient on cloud and scaling the cloud resources are costly prohibitive. All these indicate that further formal design and evaluation are warranted for scalable infrastructure for MS-based omics database workflows. High-performance computing algorithms that can leverage multicore, manycore, distributed-memory, and CPU-GPU architectures can make the existing pipelines much more scalable while maintaining a high confidence in peptide identifications. More modern techniques based on deep learning models [23, 24] are known to have scalability problems with increasing size of the training sets and will incur similar bottlenecks when incorporated into regular MS-based omics workflows.

Therefore, there is an urgent need for scalable solutions of more confident peptide identifications without which the integrity, and the confidence in large-scale systems biology studies is not possible, especially for meta-proteomics, proteogenomics, and MS-based microbiome or non-model organisms' studies having a direct impact on personalized nutrition, microbiome research, and cancer therapeutics.

In order to fill this gap, we collectively as a scientific community have to design parallelization techniques, and approaches for high-performance MS omics data analysis. However, in contrast to existing methods, these HPC algorithms must be designed by considering both computational, and communication costs as metrics

for efficiency. One parallel algorithm[1] with provable efficiencies can be designed for distributed-memory architectures, it can be extended for multicore, manycore, and Graphics Processing Units (GPU's), and cloud-platforms. To maximize the impact, effective research works are required to solve the following facets: (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, and (3) Integration of these parallel algorithms to existing workflows using XSEDE Supercomputers, and Amazon Cloud-Computing infrastructure.

Additional capabilities expected from highly scalable HPC methods include new-found abilities to process: (1) large number of potential dynamic PTM's search parameters, (2) Limiting filtering that may lead to omics dark-data, and (3) search against multiple organisms from different taxonomic families for meta-proteomics and microbiome studies. Integration of these HPC methods to existing workflows will enable systems biologist to investigate complex microbiome communities and identify the taxonomic families which is so intimately associated with human health and disease. This novel, substantially different, and scalable computational approach to study complex proteomes (and microbiome) at the proteomics level is expected to allow us to overcome the current scalability limitations of existing workflows. This new class of HPC algorithms, we believe, will open these new horizons for precision nutrition studies, interaction insights for complex microbiome communities and its effects on human gut, and overall mental and physical health.

The proposed HPC research for MS-based omics data analysis is innovative, in our opinion, because it represents a substantive departure from the status quo. To our knowledge, communication-avoiding parallel algorithms that consider both computation and communication costs to improve the efficiency of these workflows has never been achieved in the context of MS-based omics data analysis. Although rarely employed to date, the incorporation of novel and exciting parallel algorithmic design will make the power of supercomputing, ubiquitous manycore, and GPU-based architectures accessible to large scores of systems biology scientists. Such an accessible HPC framework will serve as the proof-of-concept infrastructure which will *enable bold questions*, and large system biology studies not hindered or hampered by the limiting factors associated with prohibitively long running times and/or non-accessible workflows.

Our hope is that HPC will help in understanding, and studying microbiome MS-based omics in the same way it has made a remarkable difference in our understanding of the cosmos, genomics, and molecular dynamics.

[1] Parallel algorithm and high-performance computing (HPC) method will be used interchangeably in this book.

References

1. Eng JK, Fischer B, Grossmann J, MacCoss MJ (2008) A fast sequest cross correlation algorithm. J Proteome Res 7(10):4598–4602
2. Diament BJ, Noble WS (2011) Faster sequest searching for peptide identification from tandem mass spectra. J Proteome Res 10(9):3871–3879
3. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5(11):976–989
4. McIlwain S, Tamura K, Kertesz-Farkas A, Grant CE, Diament B, Frewen B, Howbert JJ, Hoopmann MR, Kall L, Eng JK et al (2014) Crux: rapid open source protein tandem mass spectrometry analysis. J Proteome Res 13(10):4488–4491
5. 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
6. Ballard G, Demmel J, Holtz O, Schwartz O (2011) Minimizing communication in numerical linear algebra. SIAM J Matrix Anal Appl 32(3):866–901
7. Council NR et al (2005) Getting up to speed: the future of supercomputing. National Academies Press
8. Ballard G, Carson E, Demmel J, Hoemmen M, Knight N, Schwartz O (2014) Communication lower bounds and optimal algorithms for numerical linear algebra. Acta Numerica 23:1
9. Demmel J, Eliahu D, Fox A, Kamil S, Lipshitz B, Schwartz O, Spillinger O (2013) Communication-optimal parallel recursive rectangular matrix multiplication. In: 2013 IEEE 27th international symposium on parallel and distributed processing. IEEE, pp 261–272
10. Solomonik E, Bhatele A, Demmel J (2011) Improving communication performance in dense linear algebra via topology aware collectives. In: SC'11: Proceedings of 2011 international conference for high performance computing, networking, storage and analysis. IEEE, pp 1–11
11. Saito MA, Bertrand EM, Duffy ME, Gaylord DA, Held NA, Hervey WJ, Hettich RL, Jagtap PD, Janech MG, Kinkade DB, Leary DH, McIlvin MR, Moore EK, Morris RM, Neely BA, Nunn BL, Saunders JK, Shepherd AI, Symmonds NI, Walsh DA (2019) Progress and challenges in ocean metaproteomics and proposed best practices for data sharing. J Proteome Res 18(4):1461–1476, pMID: 30702898. http://dx.doi.org/10.1021/acs.jproteome.8b00761
12. Yates III JR (2019) Proteomics of communities: metaproteomics
13. Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, del Toro N, Rurik M, Walzer M, Kohlbacher O, Hermjakob H et al (2016) Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat Methods 13(8):651
14. Tran NH, Zhang X, Xin L, Shan B, Li M (2017) De novo peptide sequencing by deep learning. Proc Natl Acad Sci 114(31):8247–8252
15. Heck M, Neely BA (2020) Proteomics in non-model organisms: a new analytical frontier. J Proteome Res
16. Kulkarni G, Kalyanaraman A, Cannon WR, Baxter D (2009) A scalable parallel approach for peptide identification from large-scale mass spectrometry data. In: 2009 international conference on parallel processing workshops. IEEE, pp 423–430
17. Li C, Li K, Li K, Lin F (2019) Mctandem: an efficient tool for large-scale peptide identification on many integrated core (mic) architecture. BMC Bioinform 20(1):397
18. Sun J, Chen B, Wu F-X (2014) An improved peptide-spectral matching algorithm through distributed search over multiple cores and multiple cpus. Proteome Sci 12(1):18
19. Duncan DT, Craig R, Link AJ (2005) Parallel tandem: a program for parallel processing of tandem mass spectra using pvm or mpi and x! tandem. J Proteome Res 4(5):1842–1847
20. Bjornson RD, Carriero NJ, Colangelo C, Shifman M, Cheung K-H, Miller PL, Williams K (2008) X!! tandem, an improved method for running x! tandem in parallel on collections of commodity computers. J Proteome Res 7(1):293–299

21. Li C, Li K, Chen T, Zhu Y, He Q (2019) Sw-tandem: a highly efficient tool for large-scale peptide sequencing with parallel spectrum dot product on sunway taihulight. Bioinformatics (Oxford, England) 35(19):3861–3863
22. Saeed F, Haseeb M, Iyengar S (2020) Communication lower-bounds for distributed-memory computations for mass spectrometry based omics data. arXiv:2009.14123
23. Qiao R, Tran NH, Li M, Xin L, Shan B, Ghodsi A (2019) Deepnovov2: better de novo peptide sequencing with deep learning. arXiv:1904.08514
24. Chi H, Liu C, Yang H, Zeng W-F, Wu L, Zhou W-J, Niu X-N, Ding Y-H, Zhang Y, Wang R-M et al (2018) Open-pfind enables precise, comprehensive and rapid peptide identification in shotgun proteomics. bioRxiv, 285395