Real-time peptide identification from high-throughput mass-spectrometry data

SUMESH KUMAR and FAHAD SAEED[*], Florida International University, USA

Peptide deduction remains one of the most challenging research problems in the large-scale study of proteomes using high-throughput Mass Spectrometers. The identification of large number of proteins from complex biological samples can be carried out in two steps: 1) tryptic digestion of protein sample to isolate constituent peptides, and then generating MS/MS data using high-thought put mass spectrometers; 2) Once the data is generated various method such as database-search tools are used to compare mass-spectrometry data against a repository of known peptides. Advances in the MS instrumentation now allow generation of high-resolution data in massive volume and velocity making traditional MS based algorithms a bottleneck in the overall workflows [4]. New generation of state-of-the-art database search tools are now capable of producing high-quality matches with impressively low FDR; however, the search time usually takes somewhere between a few weeks to a few months depending on the size of database and search parameters. To accelerate the overall search times, several studies have been proposed which target this computational bottleneck by exploiting specialized hardware architectures including HPC compute clusters and GPUs [2], [1]. Even with these accelerated pipelines the dream of realizing a true real-time processing and deduction of peptides from MS data is a far from realization. One bottleneck preventing the design of true real-time processing of MS based data is the cost of communication of the data required for the existing workflows [3] i.e. moving the data from storage to computational nodes and across hierarchies of system memory, dominates the overall search process in MS data analysis. Therefore, techniques which can minimize the communication cost by enabling the computational search process to execute near the source of data-generation are highly desirable. In particular, specialized computer architecture designed by utilizing FPGAs to process high-resolution MS data as soon as it is generated by a mass-spectrometer can alleviate the latency involved in data storage and movement. FPGA based designs can exploit the inherent data-parallelism and minimize communication overhead by using a custom pipeline design aimed at reducing the number of main memory accesses. In this paper, we propose to design, and develop an FPGA based hardware accelerator. Our design consists of asynchronous parallel processing elements which implement efficient dataflow operations by using configurable data-caching, contention aware bus-arbiter, and double buffering. Our results have shown that we are able to achieve 600x reduction in average number of DRAM accesses and an average of 24x speed-up in the overall computation compared with a CPU. These results were obtained by processing publicly available MS data, whereas real-time performance can be achieved if the search operations are moved close to the source of data generation. In this regard, a streaming network-based hardware accelerator can greatly enhance the scale of proteomics which reads raw data directly from the mass-spectrometer to process the MS data in real-time in a streaming fashion and produce peptides deductions.

CCS Concepts: • **Reconfigurable hardware** → **Accelerator design**; • **Mass-spectrometry** → **Protein identification**.

ACM Reference Format:
Sumesh Kumar and Fahad Saeed. 2021. Real-time peptide identification from high-throughput mass-spectrometry data. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21), August 1–4, 2021, Gainesville, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3459930.3470856

[*] *This research was supported by the NIGMS of the National Institutes of Health (NIH) under award number: R01GM134384.*

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

BCB '21, August 1–4, 2021, Gainesville, FL, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8450-6/21/08.
https://doi.org/10.1145/3459930.3470856

REFERENCES

[1] Chuang Li, Kenli Li, Tao Chen, Yunping Zhu, and Qiang He. 2019. SW-Tandem: a highly efficient tool for large-scale peptide identification with parallel spectrum dot product on Sunway TaihuLight. *Bioinformatics* 35, 19 (2019), 3861–3863.

[2] Jeffrey A Milloy, Brendan K Faherty, and Scott A Gerber. 2012. Tempest: GPU-CPU computing for high-throughput database spectral matching. *Journal of proteome research* 11, 7 (2012), 3581–3591.

[3] Fahad Saeed. 2020. Communication Lower-Bounds for Distributed-Memory Computations for Mass Spectrometry based Omics Data. arXiv:2009.14123 [cs.DC] unpublished.

[4] Muhammad Usman Tariq, Muhammad Haseeb, Mohammed Aledhari, Rehma Razzak, Reza M Parizi, and Fahad Saeed. 2020. Methods for Proteogenomics Data Analysis, Challenges, and Scalability Bottlenecks: A Survey. *IEEE Access* (2020).