Search Feasibility in Distributed MS-Proteomics Big Data

Umair Mohammad
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
umohamma@fiu.edu

Fahad Saeed
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
fsaeed@fiu.edu

Abstract

Making large-scale Mass Spectrometry (MS) data FAIR (Findable, Accessible, Interoperable, Reusable) and democratizing access for the omics research community requires advance access and reuse mechanisms. In this work, we proposed a novel distributed data access infrastructure and developed a simulation test-bed to show the feasibility of this solution. In contrast to existing centralized approaches, participating nodes are relied upon to execute the search algorithm and search based on the ***comparison of raw spectra*** is supported as opposed to simple meta-data based searches. Simulation results using networking, stochastic modelling, and queuing theory, illustrated that search times were reduced by up-to 600 times for up-to a total of fifty billion spectra. Proteomics is vital because of the importance proteins to life and their role in state-of-the-art medicine such as custom drug delivery and cancer treatment. MS-based proteomics involves the fragmentation of proteins into peptide ions to generate raw MS spectra. Traditionally, scientists have relied on meta-data based searches of centralized repositories followed by complex database searches and protein sequencing. Though useful, this technique may result in missed datasets because of poor meta-data or sheer amount of effort and computational time needed. Recently, direct raw spectra search has been proposed with the development of centralized tools such as PeptideAtlas. However, PeptideAtlas hosts $\tilde{1}$3,000 spectra whereas systems supporting billions of spectra are needed. Let us assume users can submit one or more query spectra for search to a central controller. In the proposed novel distributed paradigm, the controller will forward the queries to several nodes hosting a total of multiple MS/MS datasets, where each of the nodes will run the search algorithm against against each spectrum in their local MS/MS dataset, and send the results as URLs/pointers and associated scores back to the controller. The controller will then collate the results and transmit them back to the users. To simulate the system performance, we focused on the distributed process between the controller and the participating nodes. We modeled the the nodes using computational devices present in typical research labs, communication links as the average achievable by combined fiber/Ethernet links, and data loads based on typical storage sizes of spectra and URLs. By running Monte Carlo simulations, we were able to obtain the response time to a single query for various scenarios and assuming an $M/M/1$ queue, we simulated the time degradation due to multiple requests by compounding over the number of requests with a load degradation factor. Testing results for fifty billion spectra indicated that using 500 distributed nodes can provide search results in 10s and 2000 nodes in 5s, a reduction by 100 and 200 times, respectively, compared to a centralized approach which requires 1000s. Considering typical capabilities of modern day servers and computers, a load factor of 0.001% was tested and indicated that the system provided constant time performance up-to 10k concurrent queries. Lastly, accounting for communication link degradation demonstrated that a trade-off can be achieved between performance and number of nodes. Therefore, it is worth investigating the implementation of a distributed big-data access infrastructure for proteomics.

CCS Concepts: • **Computing methodologies** → **Model development and analysis**; • **Applied computing** → **Systems biology**.

Keywords: distributed infrastructure, networked database, proteomics, modelling and simulation, mass spectrometry, spectral search, big omics data

ACM Reference Format:
Umair Mohammad and Fahad Saeed. 2021. Search Feasibility in Distributed MS-Proteomics Big 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, 1 page. https://doi.org/10.1145/3459930.3470855

BCB '21, August 1–4, 2021, Gainesville, FL, USA

ACM ISBN 978-1-4503-8450-6/21/08…$15.00
https://doi.org/10.1145/3459930.3470855

Acknowledgments

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM134384.