W510–W515 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* *Published online 17 May 2021*
https://doi.org/10.1093/nar/gkab368

PERCEPTRON: an open-source GPU-accelerated proteoform identification pipeline for top-down proteomics

Muhammad Farhan Khalid[1,†], Kanzal Iman[1,†], Amna Ghafoor[1,†], Mujtaba Saboor[1], Ahsan Ali[1], Urwa Muaz[1], Abdul Rehman Basharat[1], Taha Tahir[1], Muhammad Abubakar[1], Momina Amer Akhter[1], Waqar Nabi[2], Wim Vanderbauwhede[2], Fayyaz Ahmad[3], Bilal Wajid[4,5,6] and Safee Ullah Chaudhary[1,*]

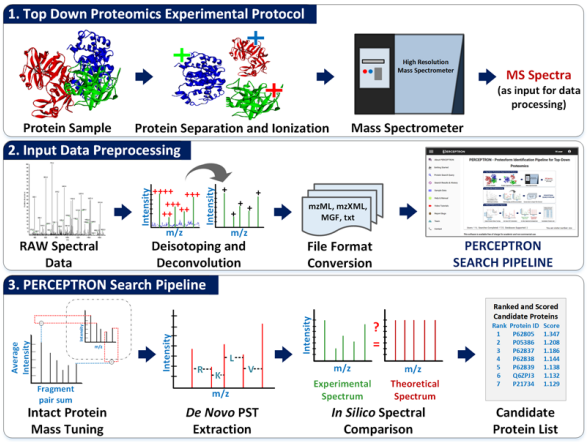
[1]Biomedical Informatics Research Laboratory, Department of Biology, Lahore University of Management Sciences, Lahore, Pakistan, [2]School of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK, [3]Department of Statistics, University of Gujrat, Gujrat, Pakistan, [4]Department of Electrical Engineering, University of Engineering and Technology, Lahore, Pakistan, [5]Department of Computer Science, University of Management and Technology, Lahore, Pakistan and [6]Division of Research and Development, Sabz-Qalam, Lahore, Pakistan

Received March 11, 2021; Revised April 10, 2021; Editorial Decision April 22, 2021; Accepted April 25, 2021

ABSTRACT

PERCEPTRON **is a next-generation freely available web-based proteoform identification and characterization platform for top-down proteomics (TDP). PERCEPTRON search pipeline brings together algorithms for (i) intact protein mass tuning, (ii)** ***de novo*** **sequence tags-based filtering, (iii) characterization of terminal as well as post-translational modifications, (iv) identification of truncated proteoforms, (v)** ***in silico*** **spectral comparison, and (vi) weight-based candidate protein scoring. High-throughput performance is achieved through the execution of optimized code via multiple threads in parallel, on graphics processing units (GPUs) using NVidia Compute Unified Device Architecture (CUDA) framework. An intuitive graphical web interface allows for setting up of search parameters as well as for visualization of results. The accuracy and performance of the tool have been validated on several TDP datasets and against available TDP software. Specifically, results obtained from searching two published TDP datasets demonstrate that PERCEPTRON outperforms all other tools by up to 135% in terms of reported proteins and 10-fold in terms of runtime. In conclusion, the proposed tool significantly enhances the state-of-the-art in TDP search software and is publicly available at https://perceptron.lums.edu.pk. Users can also create in-house deployments of the tool by building code available on the GitHub repository (http://github.com/BIRL/Perceptron).**

GRAPHICAL ABSTRACT



INTRODUCTION

Top-down proteomics (TDP) is an emerging experimental protocol for the analysis of intact proteoforms (1,2). High-resolution mass analyzers coupled with soft ionization techniques have substantially contributed to the growth and impact of TDP (3,4). In particular, the technique has the potential to enhance proteoform identification, improve characterization of post-translational modifications (PTMs) as well as quantification of potential disease biomarkers (4–6). However, the complexity of high-resolution TDP

*To whom correspondence should be addressed. Tel: +92 42 35608352; Fax: +92 42 3572 1673; Email: safee.ullah.chaudhary@gmail.com
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/49/W1/W510/6276909 by University of Glasgow user on 06 July 2021

datasets (7) necessitates high-performance search and analysis tools. Contemporary TDP tools include: MS-Align+ (8), pTop (9), TopPIC (10), MSPathFinder (11) and ProSight PTM (12). The tools, however, are limited in their availability, search accuracy as well as run-time performance (8–11). Furthermore, limited features for non-commercial users (https://assets.thermofisher.com/TFS-Assets/CMD/manuals/Man-XCALI-97801-ProSightPC-User-ManXCALI97801-EN.pdf) impede wider employment of these tools while the general lack of open source code (12,13) hinders their community-based software development (14).

In this work, we present PERCEPTRON, a web-based TDP search engine that derives its core search pipeline from SPECTRUM (15) which is a state-of-the-art MATLAB® toolbox for proteoform identification from TDP spectral data. PERCEPTRON allows for intact protein mass tuning, peptide sequence tags (PSTs)-based filtering, and characterization of post-translationally modified as well as truncated proteoforms. Proteoform scores for each candidate protein are computed by a user-defined weighted combination of algorithm-specific scoring components. PERCEPTRON achieves high-throughput performance by leveraging general-purpose graphics processing units (GPGPU) (https://www.amd.com/en). NVIDIA's® CUDA toolkit has been used for implementing selected search routines on graphics processing units (GPUs). Additionally, a dynamic hardware selection module has also been implemented to optimize CPU-GPU compute load. The proposed platform can be accessed through a set of intuitive web forms that facilitate setting up of search parameters and assist the user in visualizing search results. PERCEPTRON is available for use at https://perceptron.lums.edu.pk; its users can also set up in-house deployments by downloading its source code and database schema from the GitHub repository (https://github.com/BIRL/PERCEPTRON). Taken together, PERCEPTRON is set to serve the proteomics community in the analysis of complex TDP spectra in large-scale whole proteome analysis studies.

MATERIALS AND METHODS

Software architecture and design

Multi-layered Model-View-Controller (MVC) software architecture (https://dotnet.microsoft.com/apps/aspnet/mvc) was employed to implement PERCEPTRON. Angular 7.0 (https://angular.io/) was used to develop an interactive web-based front-end while the middleware comprises two loosely coupled components that include PERCEPTRON RESTful .Net Web API (https://dotnet.microsoft.com/apps/aspnet/apis) and PERCEPTRON Core Service. The back-end database is defined by a *Protein Repository Service* powered by Microsoft® SQL Server Management Studio 17.6 (Figure 1). PERCEPTRON API is consumed by the front-end through REST protocol. Https calls by the front-end initiate a search job. A shared relational database (PERCEPTRON DB) facilitates communication between PERCEPTRON API and PERCEPTRON Core. PERCEPTRON DB also serves to store job parameters and search results. Search operations are stored in PERCEPTRON Core and can be utilized through PERCEPTRON Web API. Each of these components has been developed using C# 7.3 and .NET framework 4.8 (https://docs.microsoft.com/en-us/aspnet/mvc/overview/getting-started/introduction/getting-started), as a standalone module for enhanced maintainability. To execute GPU device kernels on NVidia GPUs, an unmanaged CUDA library in C++ has been implemented. To load-balance, PERCEPTRON uses a heterogeneous mix of both the main processor (CPU) and GPUs for computation. To perform protein search, PERCEPTRON Core retrieves candidate proteins from the *Protein Repository*, which is an optimized fast data-retrieval application to process protein databases.

Input data formats

PERCEPTRON currently supports standardized mass spectrometry-based proteomics data formats, including (i) plain text files comprising of mass to charge ratios (m/z) and relative intensities, (ii) eXtensible Markup Language (XML) files with m/z and relative abundances (mzXML) (16), (iii) Mascot Generic Format (MGF) (17) and (iv) Mass Spectrometry Markup Language (mzML) (18,19) data formats (see Supplementary Table S3, Supported File Formats). Raw data files from mass spectrometry analysis need to be converted into either of these formats before search. Additionally, a customized file reader for MGF file format has also been provided for seamless consumption of MGF files in PERCEPTRON.

PERCEPTRON web API

PERCEPTRON web application programming interface (API) accepts hypertext transport protocol secure (https) calls from PERCEPTRON front-end. The API consists of three controllers, including (i) User Controller, (ii) Home Controller and (iii) Search Controller. These controllers handle requests and call upon the respective models for further action. The first controller processes user credentials and information in concert with the PERCEPTRON user database (DB) model. A home controller provides for API housekeeping and platform testing. The search controller takes protein search parameters from the front-end and passes them to the PERCEPTRON DB model for storage. PERCEPTRON API interacts with the PERCEPTRON CORE via PERCEPTRON DB. PERCEPTRON users issue a request by submitting protein search query through the front-end. This triggers PERCEPTRON API, which organizes the job information and passes it to the PERCEPTRON DB through the search controller. PERCEPTRON API has a bidirectional data flow architecture; PERCEPTRON API returns processed jobs stored in PERCEPTRON DB to the front-end.

Since PERCEPTRON RESTful web API is https-based, it provides data encryption to secure calls. PERCEPTRON API is developed using .NET Framework 4.8 and hosted on a Dell Power Edge R730, 2× Intel Xeon E5-2620, 160GB RAM (16GBx10), and an NVIDIA Tesla K40C (2880 Cores). Additionally, PERCEPTRON API has been made accessible through a software development kit (SDK) which enables its integration into other TDP platforms.

Downloaded from https://academic.oup.com/nar/article/49/W1/W510/6276909 by University of Glasgow user on 06 July 2021

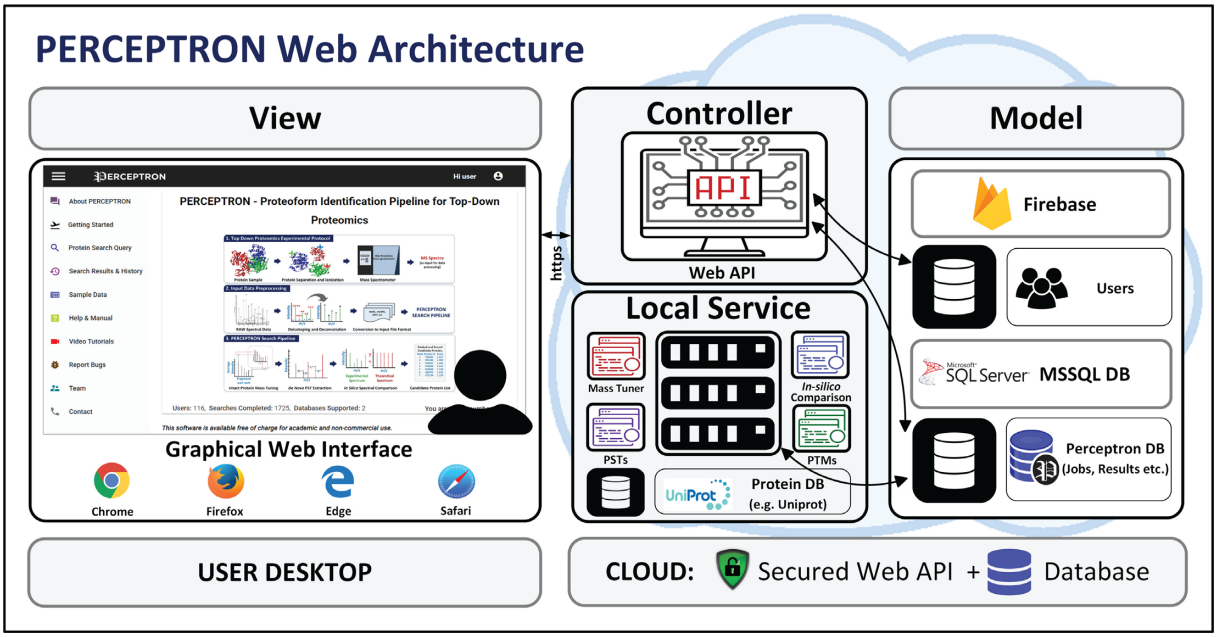


Figure 1. PERCEPTRON software architecture. PERCEPTRON web architecture is based on Model View Controller (MVC). Users can access this web service by signing up or logging in as a guest via PERCEPTRON graphical user interface (GUI) which is compatible with multiple web browsers. Secured PERCEPTRON web application program interface (API) acquires data from the front-end including (i) user credentials and (ii) parameters for protein search query, using https protocol. This information is stored in the PERCEPTRON database (DB). PERCEPTRON Core, which consists of the algorithmic pipeline, executes search using data stored in PERCEPTRON DB and PERCEPTRON Protein Repository. Processed results for each job are then moved to PERCEPTRON DB where they are stored. API returns these results to the front-end from where users can access them using PERCEPTRON GUI.

PERCEPTRON core

PERCEPTRON Core, termed *local service*, is built using SQL Server Management Studio (version 17.6). It forms the business logic layer and is central to PERCEPTRON working. PERCEPTRON's algorithmic pipeline is a part of this local service. The function of PERCEPTRON CORE is to take job query information stored in the PERCEPTRON DB model, process it, and send it back to the PERCEPTRON DB. For processing the job, PERCEPTRON Core communicates with PERCEPTRON Protein Repository. The processed job results are stored in PERCEPTRON DB. These stored search results are available to be fetched by PERCEPTRON API upon call by the front-end from the PERCEPTRON user.

PERCEPTRON repository

UniProt database has been made available for search in PERCEPTRON. With local deployments, users can add any other database as and when required. To optimize runtime, data is retrieved from .FASTA formatted database. Upon job completion, search results from PERCEPTRON Core are stored in PERCEPTRON DB. Finally, PERCEPTRON API returns these results to the user via the front-end.

RESULTS AND DISCUSSION

Workflow of PERCEPTRON pipeline

PERCEPTRON's proteoform identification workflow has been derived from SPECTRUM (15) which comprises of three salient modules: (a) intact protein mass tuner for tuning precursor whole protein mass (MS1), (b) *de novo* sequencing, and peptide sequence tag (PST) filter to generate PST ladders within user-defined range of lengths, and (c) *in silico* spectral comparator to compare theoretical spectra with the experimental data (Figure 2). Tuned mass obtained in the first step is used to filter user-selected protein database within a user-specified intact protein mass tolerance. The candidate proteins are given a preliminary score based on the user-defined weight for intact protein mass score. Next, peptide sequence tags (PSTs) obtained from MS2 data through *de novo* sequencing are employed to further shortlist candidate proteins. Theoretical spectra are then generated for each candidate protein using the user-specified fragmentation technique followed by the comparison of theoretical and experimental spectra and re-scoring. The overall protein ranking is based on the weighted scores from each algorithmic component. Unknown post-translational modifications (PTMs) are characterized through a blind PTM search module. Additional support for searching terminal, chemical, fixed, and variable modifications along with terminal and single-sided truncations has also been provided.

Salient features

PERCEPTRON derives its algorithmic pipeline from SPECTRUM toolbox (15) and improves upon the usability and performance of the MATLAB-based toolbox (15), as well as other contemporary tools (12,13). PERCEPTRON offers an enhanced availability of algorithms implemented in SPECTRUM through a web-application that does not require a MATLAB (https://www.mathworks.com/products/

Downloaded from https://academic.oup.com/nar/article/49/W1/W510/6276909 by University of Glasgow user on 06 July 2021

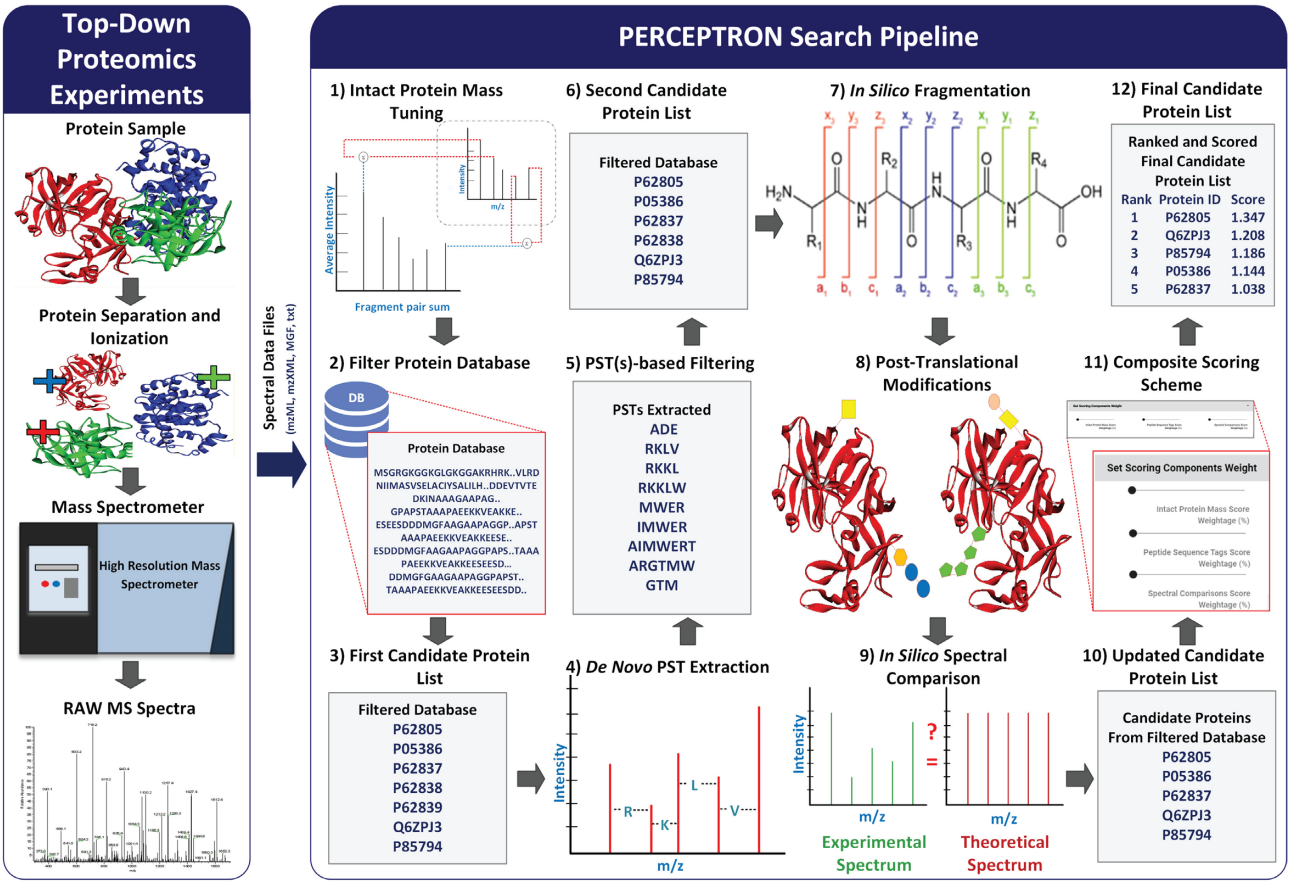


Figure 2. PERCEPTRON Workflow. PERCEPTRON pipeline initiates with the input of MS-based top-down proteomics data in specific input file formats. As the first step, mass-based filtering of user-specified protein database is performed. The resulting candidate protein list is further filtered using sequence-derived information via peptide sequence tag (PST) extractor. Next, the updated candidate protein list is subjected to *in silico* fragmentation. Theoretical spectra obtained using user-specified fragmentation methods are then compared with experimental spectra. The resulting filtered candidate protein list is ranked in light of user-defined weight-based scoring of each algorithm. The final candidate protein list contains ranked proteoforms at 1% false discovery rate (FDR).

matlab.html) license. A conceptual outline of the methodology employed in implementing GPU-based algorithms has been provided in Supplementary Methods A1 – GPU-based Algorithms. The high runtimes of MATLAB-dependent toolbox (15) have been overcome by using graphics processing unit (GPU) acceleration in PERCEPTRON. The source code of PERCEPTRON is publicly available for development as well as creating in-house deployments. The web-based pipeline removes the need for users to compile code besides seamlessly updating the underlying algorithms. Moreover, software updates and bug fixes are also provided seamlessly thus, posing no requirement for users to update the tool. PERCEPTRON's protein repository is automatically updated which offers a continued update of protein databases without any intervention of the user. To expand the interoperability of PERCEPTRON with other applications, an application programming interface (API) has been made accessible for use without the need for graphical user interface (GUI) using a software development kit (SDK).

Operation

Users can log in via an existing email account or as a guest. Using an email account to login PERCEPTRON allows users to perform search and results history is saved. However, logging in as a guest only enables protein search while results are not saved. Additionally, users also have an option to create a local account to proceed with protein search.

Adding search parameters and submitting a job

PERCEPTRON web-form allows users to upload top-down proteomics data and set search parameters for proteoform identification. Users can search from a single data file or a batch of multiple files. To initiate search, the user first enters basic parameters, including (i) project title to retrieve results, (ii) experiment data files in specified file formats, (iii) protein database for protein search, (iv) FDR cut-off, (v) email address to receive results, and (vi) number of top hits to be displayed in the search history. Next, before proceeding with search, the user must specify a set of four search parameters: (i) Experimental Parameters, (ii) *De Novo* Sequencing Parameters, (iii) Protein Modification Parameters and (iv) Scoring Components Weight.

Experimental parameters. To add experimental details, the user must select mass mode and fragmentation method that was employed during the experiment. In the current deployment, PERCEPTRON supports nine types of fragmenta-

Downloaded from https://academic.oup.com/nar/article/49/W1/W510/6276909 by University of Glasgow user on 06 July 2021

tion methods. Intact protein mass (MS1) tuning feature allows users to derive tuned MS1 by using the fragment ion data (MS2). The selected protein database can, therefore, be filtered using either experimental MS1 or tuned MS1 within the precursor ion mass tolerance provided by the user. Moreover, the user may specify other experimental details including neutral losses or the type of corresponding special fragment ions to add to the search process. Defining a peptide tolerance allows the filtering of these fragment ions. Additionally, example data files have been provided which can be searched in PERCEPTRON by loading default parameters.

De novo sequencing parameters. PERCEPTRON users can enable sequence-based filtering of protein databases using peptide sequence tags (PSTs). The length of a PST-ladder may range from 1 to 10 amino acids as specified by the user. Additionally, the user must also specify PST tolerance to allow for sequence-based search.

Protein modifications parameters. PERCEPTRON provides search support for: (i) terminal modifications, (ii) chemical modifications, (iii) fixed and variable post-translational modifications (PTMs), (iv) blind PTMs, and (v) truncated proteoforms. Users can select and incorporate one or more of these modifications into the search process.

Setting scoring components weight. PERCEPTRON's composite scoring scheme has been derived from SPECTRUM (15), which allows for weight-based scoring of three factors, including (i) intact protein score, (ii) peptide sequence tags score, and (iii) *in silico* spectral comparison score. Users can assign weights to each of these components to compute an overall protein score.

PERCEPTRON validates proteoform search results by computing false discovery rate (FDR) (see Supplementary Methods A2 – False Discovery Rate (FDR) Estimation Process). To estimate statistical significance of each candidate protein, PERCEPTRON computes E-Value thresholded at 1E−10 for stringency and high confidence.

Once the search parameters are added, users can initiate the proteoform search process by submitting the job. At any step, the user can reset search parameters to clear the web-form and add new search parameters. Once the search is completed, a link with the search results is emailed to the user.

Search results and visualization

PERCEPTRON users can view search results on the website besides downloading results along with search parameters. The summary of search results includes identified proteoforms ranked by protein score. Each Protein ID carries information about protein's molecular weight and the number and type of modifications (terminal modifications, post-translational modifications and truncations). Users can also visualize detailed search results using 'Detailed Visualization' option in 'Detailed Protein Hit View'.

Features and performance comparison

We have developed two case studies using published datasets (10,20) to test and exhibit PERCEPTRON (see Supplementary Data – Case Study I & II). Case Study I was performed to validate PERCEPTRON for HeLa spectral dataset of Histone H4 protein (20) using parameters given in Supplementary Table S1 – Case Study I – Search Parameters. The detailed results are provided in Supplementary Data S2 – Case Study I – Complete Results. Second case study was performed using *Escherichia coli* dataset (10) towards identifying proteoform count with and without using peptide sequence tags (PSTs). The search parameters used have been provided in Supplementary Table S2 – Case Study II – Search Parameters and detailed results are given in Supplementary Data S4 – Case Study II – Complete Results. We have also compared PERCEPTRON features with other top-down proteomics tools (see Supplementary Table S4 – Features Comparison) besides evaluating runtime performances of PERCEPTRON (CPU and GPU modes) and SPECTRUM, (see Supplementary Table S5 – Performance Comparison). A comparison of search performance for Case Study II has been provided in (Supplementary Table S6 – Search Comparison).

DEPLOYMENT AND AVAILABILITY

PERCEPTRON is freely available for public use along with its source code. The software is deployed on a dual Intel(R) Xeon(R) CPU E7–4830 v3 @ 2.10GHz (48 CPUs) server with 128 GB RAM, and an NVidia Tesla K40 GPU accelerator (https://www.nvidia.com/en-us/location-selector/) with 2880 cores and 12 GB on-board memory operating at 288 GB/sec. The operating system used for the deployment is Microsoft® Windows Server 2012 R2 (https://www.microsoft.com/en-pk/download/details.aspx?id=41703) and the back-end database is Microsoft® SQL Server Management Studio 17.6. The software is available at https://perceptron.lums.edu.pk. The source-code, issues list, release notes, and an updated bug database repository has been made publicly available as a GitHub repository at https://github.com/BIRL/PERCEPTRON.

Obtaining an account

Users can log in with an existing email account or as a guest. Email perceptron@lums.edu.pk for account information.

Future directions

The provision of additional protein databases in PERCEPTRON can facilitate its users for searching organism-specific proteomes. Also, currently, PERCEPTRON does not accommodate amino-acid substitutions and double-sided truncations in protein search. Support for these features will further improve protein characterizing power during spectral search. Furthermore, the implementation of probability-based post-translational modification (PTM) characterization algorithms can significantly enhance the platform. Lastly, provision for relative and absolute protein quantitation can also be a useful feature for PERCEPTRON users.

Downloaded from https://academic.oup.com/nar/article/49/W1/W510/6276909 by University of Glasgow user on 06 July 2021

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: S.U. designed the project and supervised the research; M.F., K.I., A.G., A.R., M.S., U.M., A.A., T.T., M.A., T.M. and S.U carried out the platform development; K.I. carried out the case study and analyses; K.I., W.N., W.V., S.A., S. A. and S.U. wrote the manuscript.

FUNDING

HEC [21-320SRGP/R&D/HEC/2014,20-2269/NRPU/R&D/HEC/12/4792,20-3629/NRPU/R&D/HEC/14/585]; Ignite [SRG-209]; TWAS [RG 14-319 RG/ITC/AS_C]; LUMS [STG-BIO-1008, FIF-BIO-2052, FIF-BIO-0255]. Funding for open access charge: Partially supported by university.
Conflict of interest statement. None declared.

REFERENCES

1. Smith,L.M., Kelleher,N.L., Linial,M., Goodlett,D., Langridge-Smith,P., Goo,Y.A., Safford,G., Bonilla,L., Kruppa,G. and Zubarev,R. (2013) Proteoform: a single term describing protein complexity. *Nat. Methods*, **10**, 186.
2. Catherman,A.D., Durbin,K.R., Ahlf,D.R., Early,B.P., Fellers,R.T., Tran,J.C., Thomas,P.M. and Kelleher,N.L. (2013) Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol. Cell. Proteomics*, **12**, 3465–3473.
3. Catherman,A.D., Li,M., Tran,J.C., Durbin,K.R., Compton,P.D., Early,B.P., Thomas,P.M. and Kelleher,N.L. (2013) Top down proteomics of human membrane proteins from enriched mitochondrial fractions. *Anal. Chem.*, **85**, 1880–1888.
4. Ansong,C., Wu,S., Meng,D., Liu,X., Brewer,H.M., Kaiser,B.L.D., Nakayasu,E.S., Cort,J.R., Pevzner,P. and Smith,R.D. (2013) Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 10153–10158.
5. Zhang,J., Guy,M.J., Norman,H.S., Chen,Y.-C., Xu,Q., Dong,X., Guner,H., Wang,S., Kohmoto,T. and Young,K.H. (2011) Top-down quantitative proteomics identified phosphorylation of cardiac troponin I as a candidate biomarker for chronic heart failure. *J. Proteome Res.*, **10**, 4054–4065.
6. Gregorich,Z.R., Peng,Y., Lane,N.M., Wolff,J.J., Wang,S., Guo,W., Guner,H., Doop,J., Hacker,T.A. and Ge,Y. (2015) Comprehensive assessment of chamber-specific and transmural heterogeneity in myofilament protein phosphorylation by top-down mass spectrometry. *J. Mol. Cell. Cardiol.*, **87**, 102–112.
7. Vizcaíno,J.A., Deutsch,E.W., Wang,R., Csordas,A., Reisinger,F., Rios,D., Dianes,J.A., Sun,Z., Farrah,T. and Bandeira,N. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.*, **32**, 223–226.
8. Liu,X., Sirotkin,Y., Shen,Y., Anderson,G., Tsai,Y.S., Ting,Y.S., Goodlett,D.R., Smith,R.D., Bafna,V. and Pevzner,P.A. (2012) Protein identification using top-down spectra. *Mol. Cell. Proteomics*, **11**, doi:10.1074/mcp.M111.008524.
9. Sun,R.-X., Luo,L., Wu,L., Wang,R.-M., Zeng,W.-F., Chi,H., Liu,C. and He,S.-M. (2016) pTop 1.0: a high-accuracy and high-efficiency search engine for intact protein identification. *Anal. Chem.*, **88**, 3082–3090.
10. Kou,Q., Xun,L. and Liu,X. (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, **32**, 3495–3497.
11. Park,J., Piehowski,P.D., Wilkins,C., Zhou,M., Mendoza,J., Fujimoto,G.M., Gibbons,B.C., Shaw,J.B., Shen,Y. and Shukla,A.K. (2017) Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods*, **14**, 909.
12. LeDuc,R.D., Taylor,G.K., Kim,Y.-B., Januszyk,T.E., Bynum,L.H., Sola,J.V, Garavelli,J.S. and Kelleher,N.L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic. Acids. Res.*, **32**, W340–W345.
13. Zamdborg,L., LeDuc,R.D., Glowacz,K.J., Kim,Y.-B., Viswanathan,V., Spaulding,I.T., Early,B.P., Bluhm,E.J., Babai,S. and Kelleher,N.L. (2007) ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.*, **35**, W701–W706.
14. Alexander,N.S. and Palczewski,K. (2017) Crowd sourcing difficult problems in protein science. *Protein Sci.*, **26**, 2118–2125.
15. Basharat,A.R., Iman,K., Khalid,M.F., Anwar,Z., Hussain,R., Kabir,H.G., Tahreem,M., Shahid,A., Humayun,M. and Hayat,H.A. (2019) SPECTRUM–A MATLAB toolbox for proteoform identification from top-down proteomics data. *Sci. Rep.*, **9**, 11267.
16. Pedrioli,P.G.A., Eng,J.K., Hubley,R., Vogelzang,M., Deutsch,E.W., Raught,B., Pratt,B., Nilsson,E., Angeletti,R.H. and Apweiler,R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, **22**, 1459–1466.
17. Perkins,D.N., Pappin,D.J.C., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
18. Turewicz,M. and Deutsch,E.W. (2011) Spectra, chromatograms, metadata: mzML-the standard data format for mass spectrometer output. In: *Data Mining in Proteomics*. Springer, pp. 179–203.
19. Martens,L., Chambers,M., Sturm,M., Kessner,D., Levander,F., Shofstahl,J., Tang,W.H., Römpp,A., Neumann,S. and Pizarro,A.D. (2011) mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics*, **10**, R110. 000133.
20. Frank,A.M., Pesavento,J.J., Mizzen,C.A., Kelleher,N.L. and Pevzner,P.A. (2008) Interpreting top-down mass spectra using spectral alignment. *Anal. Chem.*, **80**, 2499–2505.

Downloaded from https://academic.oup.com/nar/article/49/W1/W510/6276909 by University of Glasgow user on 06 July 2021