Network Modeling Analysis in Health Informatics and Bioinformatics (2021) 10:23
https://doi.org/10.1007/s13721-021-00298-3

SHORT COMMUNICATION

Benchmarking mass spectrometry based proteomics algorithms using a simulated database

Muaaz Gul Awan[1] · Abdullah Gul Awan[2] · Fahad Saeed[3]

Received: 21 September 2020 / Revised: 26 February 2021 / Accepted: 12 March 2021

Abstract
Protein sequencing algorithms process data from a variety of instruments that has been generated under diverse experimental conditions. Currently there is no way to predict the accuracy of an algorithm for a given data set. Most of the published algorithms and associated software has been evaluated on limited number of experimental data sets. However, these performance evaluations do not cover the complete search space the algorithmand the software might encounter in real-world. To this end, we present a database of simulated spectra that can be used to benchmark any spectra to peptide search engine. We demonstrate the usability of this database by bench marking two popular peptide sequencing engines. We show wide variation in the accuracy of peptide deductions and a complete quality profile of a given algorithm can be useful for practitioners and algorithm developers. All benchmarking data is available at https://users.cs.fiu.edu/~fsaeed/Benchmark.html

Keywords Benchmarking · Peptide search algorithms · Proteomics · Mass-spectrometry

1 Introduction

For more than four decades, benchmarks have been used to assess the reproducibility and reliability of hardware (e.g. SPEC computer architecture benchmark) or software (e.g. BAliBASE for sequence alignments). Benchmarks provide a method of comparing the performance of given entity across various possible variables and gives a relative performance by running a number of standard tests. These benchmarks ensure the reproducibility of the software for diverse conditions.

Mass spectrometry (MS) based proteomics (Aebersold and Mann 2003; Iglesias-Gato et al. 2016; Ebhardt et al. 2015; Tsai et al. 2015; PedroM and Bengt 2016; Saeed 2015) has revolutionized the study of system biology, and relies heavily on large number of software tools that automate the process of annotation and assessment of MS/MS spectra (Gul Awan and Saeed 2016; Kong et al. 2017; McIlwain et al. 2014). However, majority of the software tools that are published have been evaluated on a small set of experimental data which represents only a fraction of experimental conditions that would be encountered by the algorithm in real-world. The reliability of these algorithms and software packages then becomes questionable when they are encountered with novel data sets as demonstrated in Sect. 5, and 6. Further, it is up to the proteomics (or proteogeonomics/metaproteomics) practitioners to select a tool which would give the best accuracy for a given data set without using any quantifiable metric. The informed decision generally rests on what software tool the user is more comfortable with *instead* of what software would be best for this specific collected data set.

Similar problems have been encountered in other fields of science. One example most relevant to proteomics is the multiple-alignment problem in genomics analysis. Like peptide deduction, multiple-alignment is challenging because of many-solutions for a given data. In order to standardize the algorithm and software development, researchers came up with different benchmarks that could be used to assess the

✉ Fahad Saeed
fsaeed@fiu.edu

Muaaz Gul Awan
mgawan@lbl.gov

Abdullah Gul Awan
abdullah_gul@live.com

[1] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[2] Al-Khwarizmi Institute of Computer Science (KICS), University of Engineering & Technology (UET), Lahore, Pakistan

[3] School of Computing and Information Sciences, Florida International University, Miami, FL, USA

Published online: 26 March 2021

multiple alignment algorithms. Benchmarks such as BaliBase are a hallmark of alignment algorithm development (Arbelaez et al. May 2011; Zhenqin et al. 2018; Freytag et al. 2018) and have standardize the metrics that can be used for assessment.

In this paper, we present a standard reference database for proteomics as a bench-marking dataset for proteomics algorithms. Proposed database consists of spectra that have been simulated using *MaSS-Simulator* (Gul Awan M and Saeed F (2018)) while carefully adjusting its input parameters to cover a wide range of experimental conditions such as different dissociation strategies and peptide coverage. The benchmark data sets then can be used for assessing the accuracy and sensitivity of the existing and new algorithms, and thereby providing a scale against which all the algorithms can be evaluated.

2 Existing methods of algorithmic evaluation

Evaluating any algorithm requires ground truth data sets. In case of MS proteomics, these datasets consist of spectra that have been annotated with corresponding peptides. The methods of generating ground-truth datasets consists of annotating acquired MS/MS spectra using any existing peptide search algorithm and then evaluating these matches using statistical analysis tools specifically designed for MS proteomics data. These methods assign each peptide-to-spectra match (PSM) with statistical confidence value, each method may employ a different strategy or a variation of some existing method to determine statistical significance of these matches (Käll et al. 2008; Elias and Gygi 2007; Käll et al. 2007; Shteynberg et al. 2011; Keller et al. 2002). Among these, the most popular and widely accepted metric has been percolator's False Discovery Rate (FDR). The Peptide Spectral Matches (PSMs) obtained from a search algorithm are processed by percolator which assigns each PSM with an FDR value. PSMs with FDR value of less than 1% are generally accepted as ground-truth data. Flow chart of FDR based generation of ground-truth has been shown in Fig. 1. Despite the ingenuity and wide acceptance of this method the ground truth data is far from perfect with wide variation in quality of spectra for a given FDR (Savitski et al. 2015).

To evaluate our argument, we used the proposed reference database to assess the reliability of a popular PSM assessment algorithm which assigns FDR to PSMs. Our results have shown that even with 1% FDR filtering, there are cases when up-to 35% of the PSMs are incorrect.

In this paper, we present a benchmarking database with large number of parameters that are used to simulate this data. Using this database, we demonstrate the shortcomings of existing ground-truth evaluation methods. We follow this discussion with two experimental demonstrations of how benchmarking database can be used to evaluate different types of peptide sequencing algorithms. For the sake of this study we use two algorithms, (1) Tide (Database search) (Diament and Noble 2011) and (2) Novor (Denovo sequencing) Ma (2015).

3 Proposed benchmarking database

The benchmarking reference database has been constructed by considering many possible combinations of six variables which govern the nature of MS/MS spectra that are generated during MS based proteomics experiments. A list of variables and the possible states that they can assume has been shown in Table 1. Real world MS spectra are stochastic and one can argue that six variables are not enough to simulate experimental conditions. Despite this argument, we have previously shown that *correctly* chosen small number of variables can be used to simulate MS/MS spectra that are very close to the real-world spectra with relatively small error percentage (Gul Awan and Saeed 2018). Since these

Table 1 Each parameter can take several possible states

Parameter	Possible states	State 1	State 2	State 3
Peptide length	1, 2	<15	>30 & <51	
Post-translational modifications	1, 2	No PTMS	2 PTMs per peptide	
Peptide coverage	1, 2, 3	10 to 30%	30 to 70%	70 to 100%
Percentage of sound (POS)	1, 2, 3	7 to 10%	3 to 6%	1 to 3%
Companion ions coverage	1, 2, 3	10 to 30%	30 to 70%	70 to 100%
Noise peak intensity	1, 2, 3	30 to 160%	30 to 90%	30 to 35%

The table describes the possible states each parameter can assume and the values it holds while in that state are the variables that can easily be customized using MaSS-Simulator (Gul Awan and Saeed 2018), we chose them as the control variables for the proposed benchmarking database.

By assigning different values to each variable a unique experimental condition can be simulated using MaSS-Simulator. For this database we put together three hundred and twenty-four (324) possible experimental conditions by using variables and their states from Table 1. As a result, we have a very comprehensive benchmarking database covering a wide variety of dissociation strategies, noise content, relative intensities of signal peaks, peptide coverage, peptide length and peptide modifications. For each of the possible experimental conditions we simulate 1000 spectra in a single .ms2 file accompanied with a corresponding ground-truth file. In the table below, each variable can take several states. Using a combination of all possible states we have generated three hundred and twenty-four MS/MS spectra files with their corresponding peptide sequences available. Brief details of all the parameters have been provided in Table 2.

4 Assessment of FDR based method

A popular post-processing algorithm for evaluating PSMs from database search is Percolator from the Crux Toolkit (Käll et al. 2007). Percolator makes use of the Target-Decoy strategy and uses semi-supervised learning approach to assign FDR indicators to each PSM. Existing proteomics software evaluation is primarily done using percolator generated ground-truth spectra and is frequently considered as the gold-standard.

To evaluate how accurately Percolator assigns FDR values to PSMs we used the above discussed benchmarking database. The standard spectra were labelled with peptide sequences using Tide database search algorithm (Diament and Noble 2011) and then post-processed using Percolator. Since we already had the ground-truth for simulated spectra available, we were able to evaluate the accuracy of FDR values assigned to each PSM. We filtered out PSMs which had been assigned an FDR value of 1% or less and then evaluated the accuracy by comparing the assigned peptide against the ground-truth peptides. It can be observed in Fig. 2 that for multiple experimental conditions Percolator is not able to accurately assign the FDR values. Plots in Fig. 2 give a comprehensive analysis of Percolator. Additional plots for remaining benchmark files can be found in Supplementary materials.

Our experiments and the results in the figure also suggest that spectra with low-coverage and peptides with shorter lengths are incorrectly assigned using Percolator.

Table 2 Description of different parameters used to develop the standard proteomics database

Features	Description
Peptide length	Number of amino acids in a peptide
Post-translational modifications	Number of PTMs that can occur in a peptide
Peptide coverage	Peptide coverage provided by the resulting b/y-ions
Percentage of sound (POS)	Percentage of b/y-ions with respect to other peaks
Companion ions coverage	Neutral losses and isotopic ions accompanying each b/y-ion
Noise peak intensity	Intensity of noise peaks relative to the intensity of sound peaks

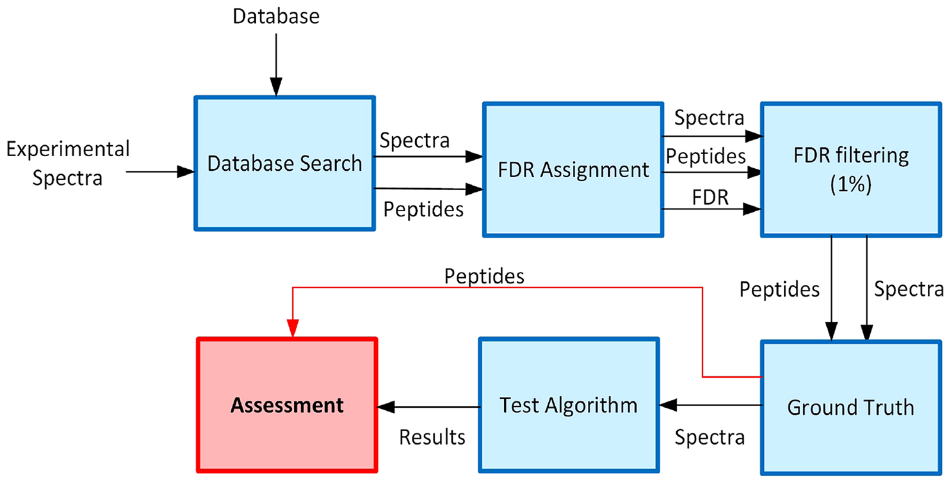
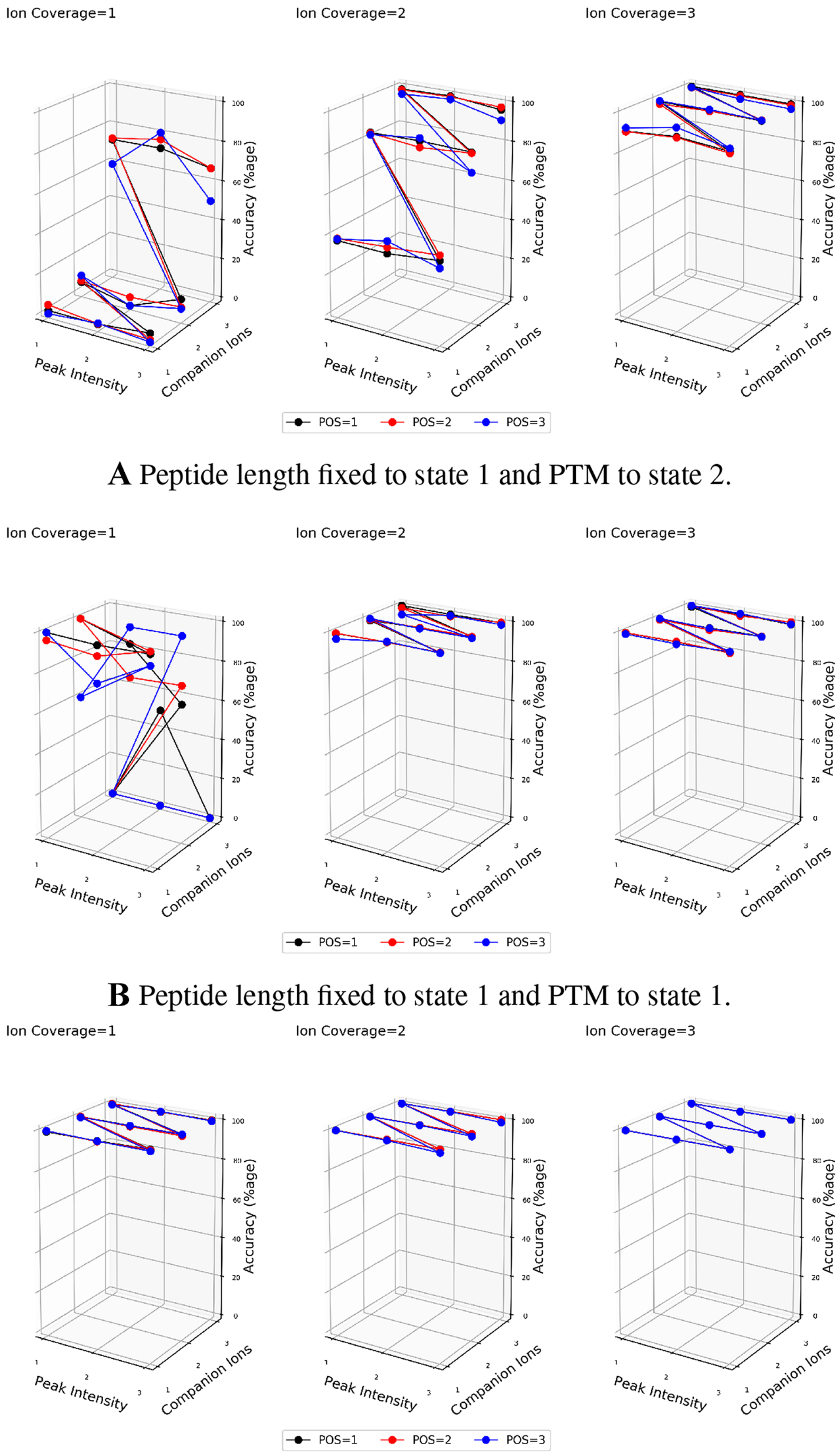


Fig. 1 Conventional work flow for generation ground-truth datasets for proteomics and evaluation of peptide sequencing algorithms

Fig. 2 Plots showing percentage of Peptide correctly filtered using 1% FDR criteria



A Peptide length fixed to state 1 and PTM to state 2.

B Peptide length fixed to state 1 and PTM to state 1.

C Peptide length was fixed to state 2 and PTM to state 1.

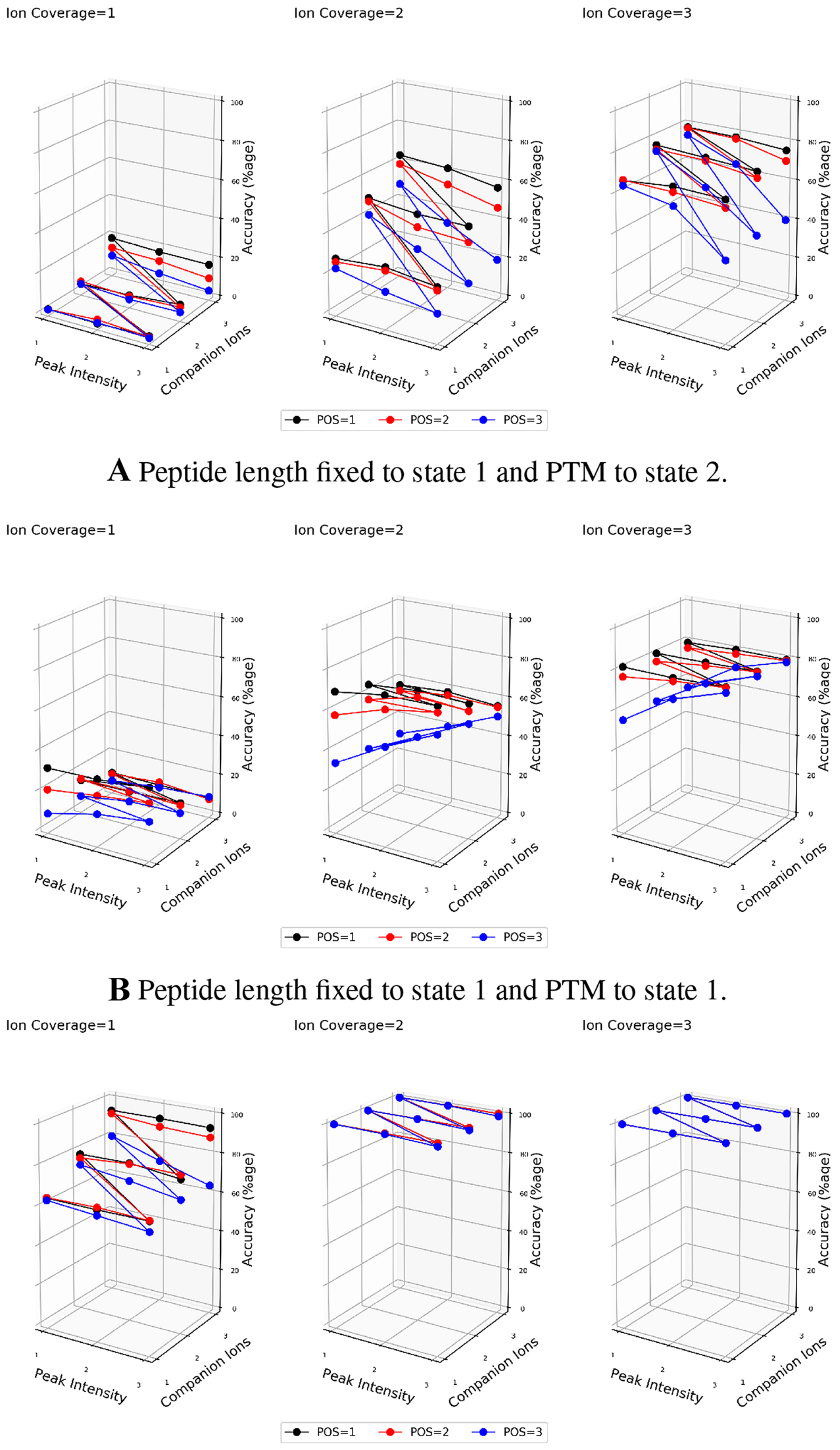


Fig. 3 Plots showing percentage of Peptides correctly identified using Tide database search

A Peptide length fixed to state 1 and PTM to state 2.

B Peptide length fixed to state 1 and PTM to state 1.

C Peptide length was fixed to state 2 and PTM to state 1.

These results support the hypothesis that low FDR PSM's cannot always be assumed to be a reliable ground-truth.

5 Evaluation of database search algorithm

To demonstrate the usability of the benchmarking reference database, we evaluated the performance of Tide database search software to observe its performance over various experimental conditions. Figure 3 shows plots which represent the behavior of Tide-Search engine across all the possible conditions covered by the proposed benchmarking database (additional plots available in supplementary materials). It can be observed that with decreasing peptide coverage the performance of Tide falls drastically and with a combination of low coverage and high noise the number of correct peptides even falls below 10%. In Fig. 3b it can be observed that introduction of two PTMs can also negatively affect the accuracy of Tide. But as shown in Fig. 3c, for longer peptide lengths Tide performs quite well in general but can still give a large number of incorrect matches when peptide coverage is low.

6 Evaluation of denovo sequencing algorithm

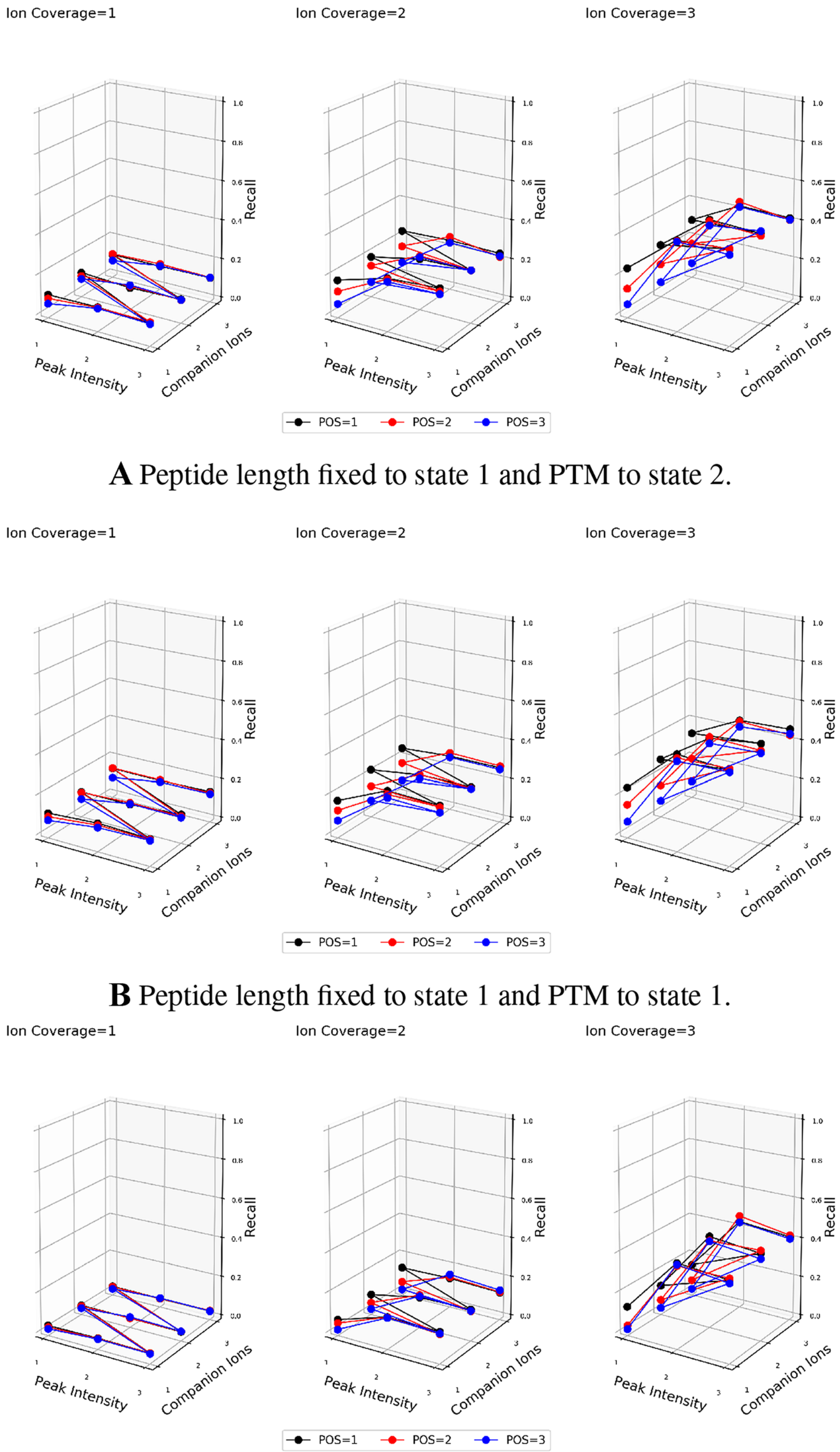
To demonstrate the usability of the benchmarking reference database, we evaluated the performance of *Novor* which is a denovo search software to observe its performance over various experimental conditions. Novor was used with its default settings, each .ms2 file was processed by Novor and the results were then evaluated by measuring recall value for each sequenced peptide and averaging the recall for each file in database (one file contains 1000 simulated spectra).

It can be observed in Fig. 4 that increasing Ion Coverage improves the recall significantly, across all lengths of peptides. Similarly, it can be seen that Companion Ions also have a significant effect on recall and low percentage of companion ions adversely effects Novor's performance. In general, performance of Novor is better for smaller peptides as compared to longer peptides. Results for additional datasets can be found in supplementary materials. It can be noted that the insights about the algorithm's performance, such as the effect of Companion Ions' population and peptide length on overall accuracy of the algorithm are not possible to understand when algorithms are published with handful of experimental datasets. This can be only possible when minor details of the spectra in the dataset are known and understood as was the case for the proposed benchmarking database.

7 Closing remarks

We have introduced a novel benchmarking strategy that can be used for evaluation of the algorithms on a wide variety of experimental conditions. This coverage of experimental conditions on our benchmarking data set will allow developers to report their accuracy of peptide deductions on a uniform scale and metrics. This benchmarking database strategy is the first step towards making peptide deduction algorithms more reliable and predictable in performance, and proteomics results more reproducible. Using this benchmarking database during algorithm development process can provide valuable insights into the design of the algorithms. Such benchmarking, if widely adopted, will help identify pitfalls and steer the algorithmic development process in correct direction. Profiling of these search algorithms will also make their performance more predictable for proteomics practitioners.

Fig. 4 Plots showing recall for Novor algorithm across the spectra from benchmarking database



A Peptide length fixed to state 1 and PTM to state 2.

B Peptide length fixed to state 1 and PTM to state 1.

C Peptide length was fixed to state 2 and PTM to state 1.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13721-021-00298-3.

Acknowledgements Research reported in this paper was supported by NIGMS of the National Institutes of Health under award number: R01GM134384. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Fahad Saeed was further supported by the National Science Foundations (NSF) under the Award Numbers NSF CAREER OAC-1925960. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

Author Contributions MA devised the method and database, wrote the software for generation of the database and wrote the manuscript. AA performed experiments and analyzed the results. FS proposed the initial idea, designed the experiments, supervised the research, and wrote and edited the manuscript.

Data availability The proposed database is available at: https://users.cs.fiu.edu/ fsaeed/Benchmark.html

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Code availability Proposed databases were generated using the MaSS-Simulator software (Gul Awan and Saeed 2018).

References

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422(6928):198

Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell 33(5):898–916

Diament BJ, Noble WS (2011) Faster request searching for peptide identification from tandem mass spectra. J Proteome Res 10(9):3871–3879

Ebhardt HA, Root A, Sander C, Aebersold R (2015) Applications of targeted proteomics in systems biology and translational medicine. Proteomics 15(18):3193–3208

Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4(3):207–214

Freytag S, Tian L, Ingrid L, Ng M, Bahlo M (2018) Comparison of clustering tools in r for medium-sized 10x genomics single-cell RNA-sequencing data. F1000Research 7

Gul Awan M, Saeed F (2016) MS-reduce: an ultrafast technique for reduction of big mass spectrometry data for high-throughput processing. Bioinformatics 32(10):1518–1526

Gul Awan M, Saeed F (2018) Mass-simulator: a highly configurable simulator for generating ms/ms datasets for benchmarking of proteomics algorithms. Proteomics 18(20):1800206

Iglesias-Gato D, Wikström P, Tyanova S, Lavallee C, Thysell E, Carlsson J, Hägglöf C, Cox J, Andrén O, Stattin P et al (2016) The proteome of primary prostate cancer. Eur Urol 69(5):942–952

Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4(11):923

Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7(01):29–34

Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. Anal Chem 74(20):5383–5392

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

Ma B (2015) Novor: real-time peptide de novo sequencing software. J Am Soc Mass Spectrom 26(11):1885–1894

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

PedroM C, Bengt F (2016) Emerging systems biology approaches in nanotoxicology: towards a mechanism-based understanding of nanomaterial hazard and risk. Toxicol Appl Pharmacol 299:101–111

Saeed F (2015) Big data proteogenomics and high performance computing: Challenges and opportunities. In Signal and information processing (GlobalSIP). In: 2015 IEEE Global Conference on. IEEE, pp 141–145

Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M (2015) A scalable approach for protein false discovery rate estimation in large proteomic data sets. Mol Cell Proteom 14(9):2394–2404

Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI (2011) iprophet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteom 10(12):M111-007690

Tsai T-H, Song E, Zhu R, Di Poto C, Wang M, Luo Y, Varghese RS, Tadesse MG, Ziada DH, Desai CS et al (2015) LC-MS/MS-based serum proteomics for identification of candidate biomarkers for hepatocellular carcinoma. Proteomics 15(13):2369–2381

Zhenqin W, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) Moleculenet: a benchmark for molecular machine learning. Chem Sci 9(2):513–530