MaSS-Simulator: A highly configurable MS/MS simulator for generating test datasets for big data algorithms

Muaaz Gul Awan[1] and Fahad Saeed [1,*]
[1]*Department of Computer Science, Western Michigan University, MI, USA.*

Abstract

Mass Spectrometry (MS) based proteomics has become an essential tool in the study of proteins. With the advent of modern MS machines huge amounts of data is being generated which can only be processed by novel algorithmic tools. However, in the absence of data benchmarks and ground truth datasets algorithmic integrity testing and reproducibility is a challenging problem. To this end, we present MaSS-Simulator, which is an easy to use simulator and can be configured to simulate MS/MS datasets for a wide variety of conditions with known ground truths. MaSS-Simulator offers many configuration options to allow the user a great degree of control over the test datasets which can enable rigorous and large-scale testing of any proteomics algorithm. We assessed MaSS-Simulator by comparing its performance against 8,031 experimentally generated spectra for which we had high confidence peptide matches available. Our results showed that MaSS-Simulator generated spectra matched closely with real-experimental spectra and had a relative-error distribution centered around 25%. In contrast the theoretical spectra for same peptides had relative-error distribution centered around 150%. MaSS-Simulator will enable developers to specifically highlight the capabilities of their algorithms and provide a strong proof of any pitfalls they might face. . Source code, executables and a user manual for MaSS-Simulator can be downloaded from https://github.com/pcdslab/MaSS-Simulator

Mass-Simulator

High performance liquid chromatography (HPLC) combined with tandem mass spectrometry has revolutionized the study of proteins. It has become an essential part of systems biology studies [1], drug discovery research [2], detection and determination of phenotypes of cancer [3], toxicology studies [4] and evolutionary biology [5]. A usual mass spectrometry (MS) based proteomics pipeline consists of breakdown of unknown proteins into smaller chains known as peptides and proceeds by separating them using high-performance liquid chromatography (HPLC). Separated peptides are then transferred to a Mass Spectrometer to obtain MS1 spectra [6]. In the fragmentation process each unknown peptide is broken down into several types of ions to yield an MS/MS spectrum. Each ion in MS/MS spectrum is represented by its mass-to-charge ratio and corresponding intensity to represent its relative abundance. Collision Induced Dissociation (CID), Higher Energy Collisional Dissociation (HCD) [7], Electron Capture Dissociation (ECD), and Electron Transfer Dissociation (ETD) are common dissociation strategies [8] [6].

The data generated by the Mass Spectrometers is processed using an algorithmic pipeline [9]. Usefulness of MS based proteomics relies on the accuracy of this pipeline. These algorithms

*Corresponding Author: fahad.saeed@wmich.edu

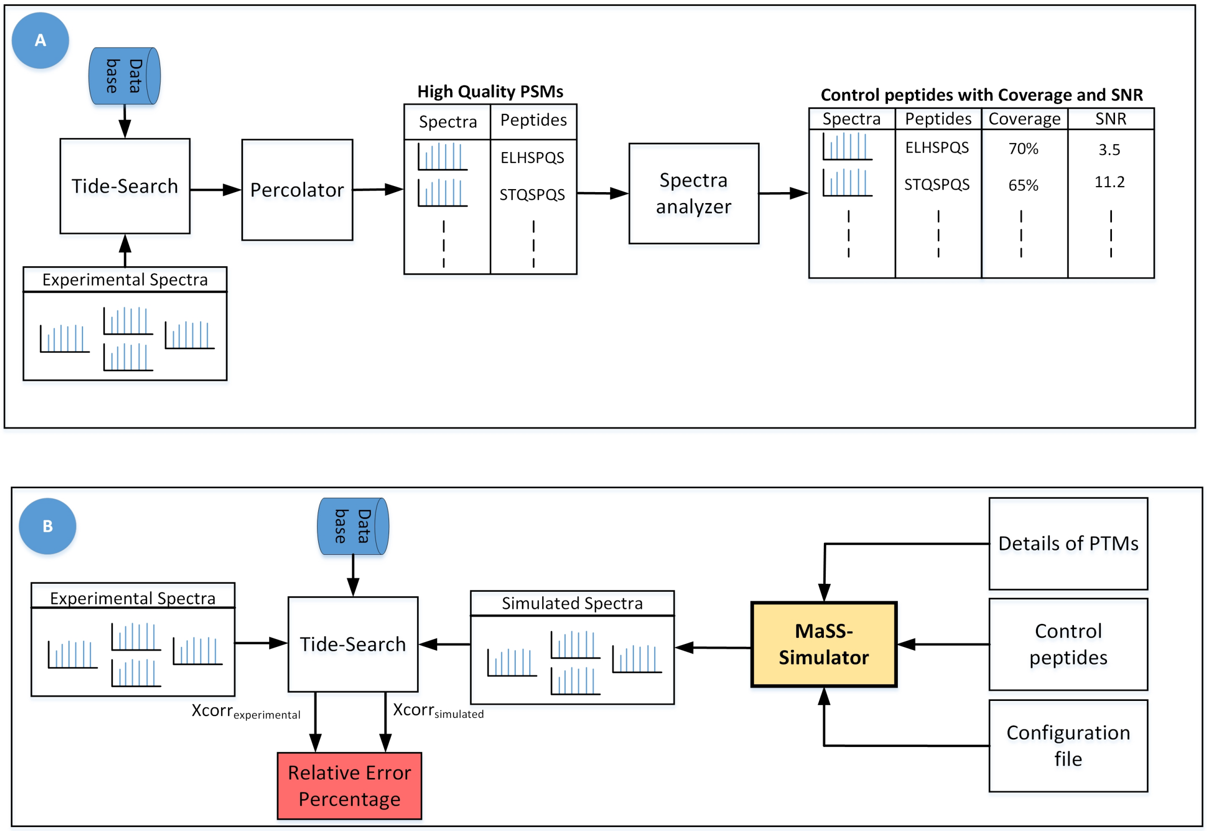

Figure 1: Figure A) shows the workflow used for obtaining experimental spectra with high confidence PSMs along with their Coverage and SNR values. Figure B) shows the workflow for generation and assessment of simulated spectra. To determine the relative error percentage for theoretical spectra, we replaced simulated spectra with theoretical spectra in this workflow.

[10] were either designed for MS/MS spectra generated by a specific ion-dissociation strategy or have only been tested on very limited sets of data. Comparing and assessing the performance of these large number of algorithms is a challenging problem due to the lack of systematic data generation where the parameters of benchmarks are in control of the method developer [11]. Due to the lack of controlled integrity testing, it becomes difficult to tell that which algorithm will function better for a particular type of dataset thus highly limiting the reliability of such softwares.

Generating experimental spectra is a costly process with many parameters not in one's control. One way of obtaining MS data sets in which all the parameters are in control of the method developer is with the help of simulators. Existing simulation techniques enable generating simple theoretical spectra [12] with a lot of parameters not in user's control. MSSimulator [11] can be used to generate a control dataset with desired properties for an LC-MS experiment but offers a very limited control over the simulation of MS/MS spectra. Another simulator with a similar name i.e. MS-Simulator [13] consists of a trained model to generate theoretical spectra with accurate y-ion intensities. These simulators either offer very limited control over the parameters or use pre-trained models specific for a particular instrument or ion-dissociation strategy to simulate MS/MS spectra.

To the best of our knowledge there does not exist a simulator for MS/MS data which will allow careful exploration of the space of the parameters associated with MS/MS data. Such exploration will allow one to identify bottlenecks, strengths and weaknesses in the proposed algorithms for MS based proteomics. Previously such simulators have been used successfully for generation of next generation sequencing data [14].

In this paper we introduce MaSS-Simulator, which offers many configurable options including the selection of ion-series, Ion Generation Probabilities, immonium ions, type and amount of noise, adjustable ion intensities and ability to simulate static and variable modifications of all types. By correctly configuring this simulator with simple configuration text file control datasets with desired properties and ground truth peptides can be obtained and used for assessment of proteomics algorithms.

We have tested MaSS-Simulator for two different dissociation strategies and compared simulated spectra against experimentally generated spectra. Our results have shown that MaSS-Simulator can generate spectra which are very close to experimental spectra regardless of dissociation technique.

Fragmentation process which leads to the generation of MS2 spectra is highly dependent upon the ionization technique, instrument and other factors [15]. For instance, the type of ions present in a spectrum and peptide coverage are dependent upon the type of dissociation strategy [8]. To give user a complete control over the ion fragmentation we introduce a feature of *Ion Generation Probability* (IGP). IGP value for each ion determines the likelihood that a given ion will be generated in the simulation. For instance, if the IGP value of b-ions is set to 40%, then the probability that each b-ion will be generated is 0.4. Using the ion generation probabilities peptide coverage can be controlled. Hence by correctly selecting the ion series and their corresponding IGP values, any dissociation strategy can be simulated. Immonium ions [10] may be formed for some ion dissociation techniques which are helpful in detecting certain amino acids [10]. MaSS-Simulator can be configured to generate these ions with a given IGP value.

Ion intensities depict the relative abundance of the ions. A lot of effort has been made to predict the ion intensities theoretically but the developed models have been trained only for a handful of experimental conditions [13]. For our purposes we used average of relative intensity values as default settings e.g. average intensity of y ions is usually two times that of b ions [16]. Intensity values for each ion series can also be adjusted by the user from the configuration file.

For large scale testing of peptide search engines an elaborate set of spectra with a range of Post Translational Modifications may be required. To help with this, MaSS-Simulator provides the option of simulating any static and variable Post Translational Modification (PTM). All desired types of modifications can be listed in the *modifications.ptm* file by following a simple to understand form. Details of this format can be found in the user manual. For our experiments we tested both static Carbamidomethyl: (C+57.021) and variable modifications (Phosphorylation: STY + 79.966, Deamidation: NQ+0.984 and Oxidation: M+15.995). In most MS2 spectra only about 5-10% of the peaks are useful for peptide deduction and the remaining data is usually noise [9] [17]. Nature and amount of noise in spectra can vary greatly with the experimental conditions. MaSS-Simulator gives an option to add *random* noise peaks in the spectra that can either be *uniformly* distributed across each spectrum or follow a *Gaussian* distribution with the possibility of including a user defined noise model. Intensity values for noise peaks can also be configured as either fixed or randomly generated within a user defined range.

To control the amount of noise we use *Sound to Noise Ratio* (SNR) factor. User can specify a desired SNR value to control the amount of noise to be added to spectra. SNR is given by:

$$SNR = \frac{n(y) + n(b)}{n(N)} * 100$$

Where $n(y)$ is the number of y ions, $n(b)$ is the number of b-ions while $n(N)$ is the number of noise peaks to be added. Generated spectra are output in the form of an MS2 file which can be conveniently converted to any other desired format using software like proteowizard [18].

To assess the spectra generated by MaSS Simulator, we shortlisted 8,031 control peptides for which we had high confidence experimental spectra available. Detailed process of generation of experimental data has been discussed in Supplementary Materials and the work-flow for shortlisting the high-confidence experimental spectra and their corresponding peptides (which will be used as control dataset) has been visualized in Fig. 1 (A). The list of control peptides along with a configuration file containing the SNR and coverage values and a *modification.ptm* was given as input to the simulator as shown in Fig. 1 (B). In the configuration file we used peptide coverage values for control data-set as the IGP values for b and y ion series and the SNR values to control the amount of noise. Other parameters in configuration file for this experiment can be found in the Table 1 of Supplementary Materials. At the output we obtained simulated spectra for each control peptide.

Ideally the simulated spectra should closely match the corresponding experimentally generated spectra. To assess the similarity between the two sets of spectra we use the work-flow given in Fig. 1 (B). The idea is to compare both the experimental and simulated spectra using a proven method/score. For this purpose, we use the xcorr scores obtained from Tide database search software [19] which gives a measure of how closely the spectrum under consideration matches the theoretical spectrum of a particular peptide. We consider the xcorr value for our experimentally obtained spectra to be a gold standard. Consider $xcorr_{exp}$ to be the xcorr score of an experimental spectrum and $xcorr_{sim}$ be the xcorr score attained by a simulated spectrum which has the same target peptide. Smaller the difference between these two scores, more similar the two spectra are. So using the following equation we can compute a relative error percentage, a smaller error means two spectra match closely.

$$RE = \frac{|xcorr_{exp} - xcorr_{sim}|}{xcorr_{exp}} * 100$$

Following the above discussed method, the simulated and the experimental spectra are searched against a rat proteome using Tide [19] algorithm which outputs the list of peptide spectral matches (PSMs) and xcorr scores for each set of input spectra. We consider the PSMs from both sets which have the same target peptide and use their xcorr values to compute a relative error percentage using the above equation. The same procedure is repeated by replacing the simulated spectra with simple theoretical spectra and relative error percentage is computed using the above equation by replacing $xcorr_{sim}$ with $xcorr_{theo}$ which represents the xcorr score for theoretical spectra.

Boxplots were used to compare the relative error distributions for simulated and theoretical spectra as shown in Fig. 1 through 4 of Supplementary Materials. It can be observed that the simulated spectra match the xcorr scores of experimental spectra much more closely than the theoretical spectra. Majority of simulated spectra have an error percentage of 25% which is extremely small compared to the large error percentage for theoretical spectra. Further, it can be observed that the error remains small consistently regardless of the type of ion-dissociation technique used or if the peptides were modified or not.

We also demonstrate the usability of MaSS-Simulator by assessing the performance of peptide database search engine Tide. Details of this study can be found in Supplementary Materials. Our experiments using controlled sets of simulated spectra show that Tide performs poorly for spectra with low coverage or low SNR. Such a study would not have been possible without MaSS-Simulator.

1 Acknowledgements

This research was supported by the NIGMS of NIH under Award Number R15GM120820. Fahad Saeed was additionally supported by NSF CAREER ACI-1651724 grant.

References

[1] R. Aebersold and M. Mann, "Mass-spectrometric exploration of proteome structure and function," *Nature*, vol. 537, no. 7620, p. 347, 2016.

[2] D. E. Scott, A. R. Bayly, C. Abell, and J. Skidmore, "Small molecules, big targets: drug discovery faces the protein–protein interaction challenge," *Nature Reviews Drug Discovery*, vol. 15, no. 8, p. 533, 2016.

[3] Y. Liu, J. Chen, A. Sethi, Q. K. Li, L. Chen, B. Collins, L. C. Gillet, B. Wollscheid, H. Zhang, and R. Aebersold, "Glycoproteomic analysis of prostate cancer tissues by swath mass spectrometry discovers n-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness," *Molecular & Cellular Proteomics*, vol. 13, no. 7, pp. 1753–1768, 2014.

[4] K. Linnet, "Toxicological screening and quantitation using liquid chromatography," *Time-of-Flight Mass Spectrometry. Journal of Forensic Science and Criminology*, vol. 1, no. 1, 2013.

[5] B. Zhao, T. Pisitkun, J. D. Hoffert, M. A. Knepper, and F. Saeed, "Cphos: a program to calculate and visualize evolutionarily conserved functional phosphorylation sites," *Proteomics*, vol. 12, no. 22, pp. 3299–3303, 2012.

[6] X. Han, A. Aslanian, and J. R. Yates III, "Mass spectrometry for proteomics," *Current opinion in chemical biology*, vol. 12, no. 5, pp. 483–490, 2008.

[7] M. P. Jedrychowski, E. L. Huttlin, W. Haas, M. E. Sowa, R. Rad, and S. P. Gygi, "Evaluation of hcd-and cid-type fragmentation within their respective detection platforms for murine phosphoproteomics," *Molecular & Cellular Proteomics*, vol. 10, no. 12, pp. M111–009 910, 2011.

[8] K. F. Medzihradszky and R. J. Chalkley, "Lessons in de novo peptide sequencing by tandem mass spectrometry," *Mass spectrometry reviews*, vol. 34, no. 1, pp. 43–63, 2015.

[9] M. G. Awan and F. Saeed, "Ms-reduce: an ultrafast technique for reduction of big mass spectrometry data for high-throughput processing," *Bioinformatics*, vol. 32, no. 10, pp. 1518–1526, 2016.

[10] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid communications in mass spectrometry*, vol. 17, no. 20, pp. 2337–2342, 2003.

[11] C. Bielow, S. Aiche, S. Andreotti, and K. Reinert, "Mssimulator: Simulation of mass spectrometry data," *Journal of proteome research*, vol. 10, no. 7, pp. 2922–2929, 2011.

[12] J. K. Eng, A. L. McCormack, and J. R. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 11, 1994.

[13] S. Sun, F. Yang, Q. Yang, H. Zhang, Y. Wang, D. Bu, and B. Ma, "Ms-simulator: predicting y-ion intensities for peptides with two charges based on the intensity ratio of neighboring ions," *Journal of proteome research*, vol. 11, no. 9, pp. 4509–4516, 2012.

[14] W. Huang, L. Li, J. R. Myers, and G. T. Marth, "Art: a next-generation sequencing read simulator," *Bioinformatics*, vol. 28, no. 4, pp. 593–594, 2011.

[15] A. Michalski, E. Damoc, J.-P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, and S. Horning, "Mass spectrometry-based proteomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer," *Molecular & Cellular Proteomics*, vol. 10, no. 9, pp. M111–011 015, 2011.

[16] A. Frank and P. Pevzner, "Pepnovo: de novo peptide sequencing via probabilistic network modeling," *Analytical chemistry*, vol. 77, no. 4, pp. 964–973, 2005.

[17] M. G. Awan and F. Saeed, "An out-of-core gpu based dimensionality reduction algorithm for big mass spectrometry data and its application in bottom-up proteomics," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.* ACM, 2017, pp. 550–555.

[18] R. Adusumilli and P. Mallick, "Data conversion with proteowizard msconvert," *Proteomics: Methods and Protocols*, pp. 339–368, 2017.

[19] B. J. Diament and W. S. Noble, "Faster sequest searching for peptide identification from tandem mass spectra," *Journal of proteome research*, vol. 10, no. 9, pp. 3871–3879, 2011.