Database Creator for Protein/Peptide Mass Analysis, DC-PPMA: A novel standalone computational tool for simplifying the analysis of MS/MS data to identify protein/polypeptide sequences by different proteomic approaches

Pandi Boomathi Pandeswari¹, Isaac Emerson¹, and Varatharajan Sabareesh¹

¹Vellore Institute of Technology

November 30, 2022

Abstract

Rationale: Proteomic studies typically involve use of different types of softwares for annotating experimental tandem mass spectrometric data (MS/MS) and thereby simplify the process of peptide and protein identification. For such annotations, these softwares calculate the m/z values of the peptide/protein precursor and fragment ions, for which a database of protein sequences must be provided as input file. The calculated m/z values are stored as another database, which the user usually cannot view. 'Database Creator for Protein/Peptide Mass Analysis' (DC-PPMA) is a novel standalone software that can create custom databases and the user can view the custom database containing the calculated m/z values of precursor and fragment ions. Methods: Python language was used for implementation and the graphical user interface was built with Page/Tcl, making this tool more user-friendly and easier to analyze. DC-PPMA is freely available at https://vit.ac.in/PPMA/. Results: DC-PPMA contains three modules. Protein/peptide sequences as per user's choice can be entered as input to the first module for creating custom database. In the second module, m/z values must be queried-in, which are searched within the custom database to identify protein/peptide sequences. The third module is suited for peptide mass fingerprinting, for which data arising from both ESI and MALDI MS can be utilized. Conclusions: Mass spectral data acquired from any proteomic approach: bottomup, middle-down and top-down can be interrogated with DC-PPMA. A major facet of DC-PPMA is that the user can 'view' the custom database containing the m/z values of the precursor ions (e.g., proteolytic peptides) and the respective fragment ions (e.g., b & v ions), prior to the database search. The feature of 'viewing' the custom database cannot only be helpful for better understanding the search engine processes; but also, for 'designing multiple reaction monitoring (MRM) methods'. Post-translational modifications and protein isoforms too can be analyzed.

Introduction

Mass spectrometry (MS) is an indispensable tool in proteomics. Due to the high-throughput nature, loads of mass spectral data are generated in any typical proteomics experiment. Therefore, manual interpretation of mass spectral data becomes time-consuming and cumbersome. Consequently, several softwares, including web applications, standalone tools using various algorithms were developed with the key purpose to annotate the mass spectrometric data, thereby simplifying the efforts devoted to data analysis and interpretation [1-12]. Thus far, many software programs have been developed and widely used for the well-established Bottom-up Proteomic (BUP) approach [13, 14]. Similarly, softwares have also been developed for the Topdown proteomics (TDP) (https://www.topdownproteomics.org/ resources/software/). For the approaches involved in middle-down proteomics (MDP), only a few softwares such as YADA, XDIA, isoScale, and Histone coder (https://middle-down.github.io/Software/) are available especially for histone and antibody characterization [15-17]. In all these available softwares, protein sequence database is imperative, which must be entered as an input for identifying proteins. The protein sequences in a database are then used to calculate the m/z values of precursor ions and peptide fragment ions. These calculated m/z values are actually saved or stored in the form of another database, which is subsequently used to annotate the spectra resulting from tandem mass spectrometry (MS/MS) and eventually leading to identify proteolytic peptides and/or proteins. Therefore, at the end of the database search process, the user views only the 'matched hits' in the output, viz., the agreement between the experimental MS/MS spectra and the relevant database entries. This is the typical way of functioning of several proteomic softwares for protein identification. In all these cases, the user cannot view the database containing the m/z values of precursor ions and fragment ions, prior to database search. In other words, the user is aware of the protein sequence database that he/she enters as an input file, whereas the user cannot 'view' and hence, is oblivious of the database comprising m/z values of the precursor ions and the fragment ions that has been generated using the protein sequences, before the database search process. Thus, the user does not know, what is happening with the 'sequence database' that he/she uploads in the search engine.

And, since it is important that the choice of 'optimal database' is critical for more reliable protein identification from MS/MS [18], we decided to develop a new standalone software tool called 'Database Creator for Protein/Peptide Mass Analysis, (DC-PPMA)', wherein the user can 'view' the database containing the calculated m/z values of precursor ions and fragment ions, before the process of database search . So, the user is aware of the 'custom' database of m/z values of precursor and fragment ions that he/she will be using subsequently for MS/MS based search and for further analysis.

In DC-PPMA, the 'database' can be created and tailored according to the proteomic approach that a user follows. Further, DC-PPMA can be used for analysing PTMs, isoforms and also user-defined (custom/new) modifications of targeted peptides/proteins. Furthermore, DC-PPMA is suited for analysing sequences of intact peptides, e.g., natural product polypeptides or synthetic peptides, whose sequences can be entered in an input file. With respect to MD proteomic analysis, two features have been included in DC-PPMA: (i) specialized enzymes used for the MDP are given in the python dictionary and (ii) 'mass range' filter is provided for creating databases containing longer proteolytic/truncated peptides. Additionally, TDP analysis can be performed in DC-PPMA by creating database containing multiply charged ions of intact protein sequences, for which no protease need to be selected. So, DC-PPMA is applicable for any proteomic approach, be it MDP, BUP or TDP. Thus, altogether DC-PPMA can be utilized for the identification and characterization of sequences: (i) derived from transcriptomic data, (ii) targeted proteins of user's interest, (iii) peptide(s) of any length and (iv) custom modified peptides/proteins. So, it can be used not only for mass spectral data analysis for proteomics but also for peptidomics. The detailed workflow of DC-PPMA containing three modules is shown in (**Figure 1**)

Method

DC-PPMA was developed by python code and the graphical user interface (GUI) was built using page/Tcl.

Results

Software Description

The homepage showing the layout of DC-PPMA containing three modules (pipelines) is shown in (Figure 2). In 'Creation of Custom-Database' (Module 1) the protein sequences of interest can be given in a text (.txt) file which is editable. The customized database will be generated as an output and saved as an excel file. Specifically, the protein/peptide sequence database (or list of protein/peptide sequences) that is entered as input into Module 1 is converted into another database containing the m/z values of precursor ions and/or fragment ions. The role of Module 2, 'Custom Database Search' is to identify the peptide/protein hits for the queried m/z values within the database created by the Module 1. Therefore, Module 1 and Module 2 are interconnected to perform peptide search. Therefore, Modules 1 and 2 should be used together for MS/MS based search and MS/MS data analysis. The Module 3 functions independently for peptide mass fingerprinting (PMF), whereby 'proteolytic peptide mass search' will be performed for m/z values against the protein sequence (fasta file) of a particular biological species. Usually PMF is done using MALDI mass spectrometric data that would typically contain m/z values of singly protonated molecular ions of

proteolytic peptides. However, the Module 3 of DC-PPMA can handle even the conventional ESI mass spectrometric data, which would typically contain m/z values corresponding to multiply protonated ionic species of proteolytic peptides (depending on the length, amino acid composition and sequence). Therefore, the output from Module 3 can be useful to expedite the analysis of ESI-MS based PMF also, in addition to the MALDI-MS based PMF.

In DC-PPMA, the algorithm of peptide search for MS and MS/MS has been designed in such a way that the both the m/z values as well as their respective charge state that are queried in 'Module 2' should match with the values in the custom database that is obtained as output of Module 1. For MS database search, minimum of four queried m/z values should match with the MS database created by the Module 1. In order to perform MS/MS search, minimum of six queried m/z values have to match with a single (proteolytic) peptide corresponding to a protein in the custom MS/MS database that is created from the Module 1. Therefore, peptide search can be done for both MS and MS/MS data, in order to interpret the experimentally observed m/z values both manually as well as by using the Module 2. The observed m/z values and their respective charge states (either from MS or from MS/MS) can be given as input in the form of .txt file. Additionally, error width options are provided, which needs to be appropriately chosen, depending on the mass resolution of the spectrometer used for data acquisitions. The error width option also can be useful to decrease the false positives in the output.

The performance of DC-PPMA was examined using randomly chosen 25 model protein sequences. Among them experimental mass spectral data of eight model proteins under two different conditions: (i) standard trypsin digestion and (ii) trypsin digestion after arginine modification by two different reagents: 1,2cyclohexanedione (CHD) and phenylglyoxal, were considered. Firstly, in the Module 1, the selected 25 protein sequences were entered as input in the form of a .txt file. Trypsin was chosen as the protease (enzyme) and carbamidomethylation was chosen in the modification tab of Module 1 window. For these input parameters, MS and MS/MS databases were created and saved as excel files (**Figure 3**). Subsequently, the observed m/z values from the experiments done on Agilent 6545 LC-MS Q-ToF were queried in the Modules 2 and 3. All the matched tryptic peptides are shown in the excel file output, which was verified by manual interpretation (**Figure 4**). Similarly, the custom modification option was tested by manually entering the molecular mass of CHD (112 Da) available in the Module 1 window (**Figure 5**), which is known to specifically modify arginine residues in proteins [19-21] and the respective custom databases for both MS and MS/MS were created. Those CHD modified peptides that matched with queried m/z values are shown in the output, which have also been confirmed manually. This proves the utility of DC-PPMA for targeted MS-based studies on proteins.

Highlights of DC-PPMA

(i) A major facet of DC-PPMA is that the user can'view' the database (in the form of an excel file) that they will be using further for MS/MS based search or data analysis, viz., prior to database search, the user can know, what are the m/z values of precursor ions (of proteolytic/truncated peptide sequences) and what are the m/z values of fragment ions that will be involved in the search engine process. To the best of our knowledge, this particular facet is not available in any proprietary or online tool that are utilized for proteomic or peptidomic investigations. By viewing/knowing the database containing m/z values of precursor ions and fragment ions, it is possible to know better about 'true negatives and false positives' and the user obtains a better understanding about the process of matching between the experimental MS/MS data and the theoretically calculated values present in the database. Therefore, if DC-PPMA is used to construct appropriate 'decoy database' by choosing suitable protein sequences, then he/she can obtain better comprehension about false discovery rate (FDR). Thus, viewing and analysis of decoy and target databases can be very helpful in more reliable annotation of MS/MS spectra towards proper protein identification.

(ii) Another notable feature of DC-PPMA is that the Module 1 calculates the m/z values for 'singly as well as multiply charged (protonated)' precursor ions (viz., proteolytic peptides) and also calculates the m/z values for 'singly and multiply charged fragment ions: a-, b-, c-, x-, y- and z- ions'. Consequently, in the 'custom' MS/MS database (created from Module 1), the m/z values of each fragment ion (e.g., b,

y, c, z ions) and their respective 'charge state' are generated and shown in the excel sheets, which can be anticipated from MALDI and ESI MS/MS experiments. Therefore, DC-PPMA can be useful for 'manual spectral annotations and interpretations' for proteomic researchers and/or protein/peptide chemists, who use both ESI and MALDI MS/MS.

(iii) Additionally, DC-PPMA can be of immense utility for 'targeted analyses', particularly for multiple reaction monitoring (**MRM**) based experiments, wherein it is essential to 'design suitable channels' that should encompass the m/z values of precursor and pertinent fragment ions. In this context, the output from Module 1 can be used for 'designing **MRM method'**, because the custom database built by Module 1 would consist of m/z values of singly as well as multiply protonated precursor and relevant fragment ions. Thus, DC-PPMA can be helpful for quantitative studies also.

(iv) Furthermore, peptide sequences, either single or multiple sequences can be uploaded in DC-PPMA. This particular feature can be useful for peptidomic investigations and also for *de novo*sequencing exercises or assignments. Consequently, DC-PPMA can indeed prove to be worth for *de novo* sequencing of polypeptides and proteins as well. Additionally, it can be used for discovery based proteomic/peptidomic analysis also, if/when the transcriptome and/or the genome of their sample is also known.

(v) In addition, the custom databases generated by DC-PPMA can also be utilized further for planning about probability-based scoring algorithms or scoring schemes, so as to identify more peptides and proteins in a reliable fashion.

Conclusions

Due to continuous advancements in computational methods, the software and databases used for proteomics are also rapidly evolving. The significance of optimal database for protein identification by MS/MS has been lucidly delineated by Kumar et al. 2017 [18]. Even to devise optimal scoring algorithm for better peptide/protein identification from the data of MS/MS, firstly it is imperative to construct a good database. Consequently, the need to build custom database has become inevitable. The custom databases can be built: (i) according to the individual researcher's specific project and objectives; (ii) for in-house requirements and (iii) to expedite the data analysis and interpretations. These are only a few reasons/purposes (among many), as to, why custom databases are essential for proteomics. Therefore, our objective was to make a standalone computational tool that can ease the process of creating custom database, which can simplify the analysis of MS/MS data of proteome/protein/peptidome. The custom database built with DC-PPMA enables the user to know, what are the 'm/z values' that are involved in the MS and MS/MS database search process, for a given 'protein/peptide sequence database'.

Though some or all of the aforementioned features of DC-PPMA are available in several proprietary software tools that comes along with mass spectrometer, such proprietary tools are either not accessible to everyone or would have limited access. Further, there are online tools which also have the same or very similar functionalities as that of DC-PPMA. However, to the best of our knowledge, we believe that DC-PPMA is unique in that it is a standalone tool, which is freely downloadable. So, DC-PPMA is accessible to everyone. Furthermore, using a particular custom database created by DC-PPMA, both MALDI MS/MS and ESI MS/MS data can be analysed, irrespective of the manufacturer. Another important aspect is that DC-PPMA can be successfully applied for any of the three proteomic approaches, viz., bottom-up, middle-down and top-down. Thus, it can be helpful for both academicians and industrial researchers, particularly for biotechnology industries involved in protein/proteomic/peptidomic investigations. Academicians can consider DC-PPMA as a tool to teach novices and students.

Author contribution

Arnold Emerson implemented the back-end programs. The front-end GUI was designed by Boomathi. Sabareesh conceptualized this project and contributed for the GUI design.

Acknowledgements

Boomathi Pandeswari thanks Council of Scientific and Industrial Research (CSIR), Govt. of India, for Senior Research Fellowship.

References

[1] Orsburn BC. Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. Proteomes. 2021 Mar 23; 9(1):15. doi: 10.3390/proteomes9010015.

[2] Chen C, Hou J, Tanner JJ, Cheng J. Bioinformatics methods for mass spectrometry-based proteomics data analysis. Int J Mol Sci. 2020; 21(8), 2873. https://doi.org/10.3390/ijms21082873

[3] Verheggen K, Ræder H, Berven FS, Martens L, Barsnes H, Vaudel M. Anatomy and evolution of database search engines - a central component of mass spectrometry based proteomic workflows. Mass Spectrom Rev. 2020; 39(3):292-306.

[4] Tsiamis V, Ienasescu HI, Gabrielaitis D, Palmblad M, Schwämmle V, Ison J. One Thousand and One Software for Proteomics: Tales of the Toolmakers of Science. J Proteome Res. 2019;18 (10):3580–5.

[5] Prakash A, Ahmad S, Majumder S, Jenkins C, Orsburn B. Bolt: a New Age Peptide Search Engine for Comprehensive MS/MS Sequencing Through Vast Protein Databases in Minutes. J Am Soc Mass Spectrom. 2019 Nov;30(11):2408-2418. doi: 10.1007/s13361-019-02306-3.

[6] Röst H L, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich H C, Gutenbrunner P, Kenar E, Liang X, Nahnsen S, Nilse L, Pfeuffer J, Rosenberger G, Rurik M, Schmitt U, Veit J, Walzer M, Wojnar D, Wolski WE, Schilling O, Choudhary J S, Malmström L, Aebersold R, Reinert K, Kohlbacher O. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 2016; 13(9):741-8.

[7] Chen T, Zhao J, Ma J, Zhu Y. Web resources for mass spectrometry-based proteomics. Genomics, Proteomics Bioinformatics. 2015;13(1):36 - 39. doi:10.1016/j.gpb.2015.01.004.

[8] Perez-Riverol Y, Wang R, Hermjakob H, Müller M, Vesada V, Vizcaíno J A. Open source libraries and frameworks for mass spectrometry based proteomics: A developer's perspective. Biochim Biophys Acta - Proteins Proteomics 2014;1844 (1 Pt A):63-76.

doi: 10.1016/j.bbapap.2013.02.032. .

[9] Allmer J. Algorithms for the *de novo* sequencing of peptides from tandem mass spectra. Expert Rev Proteomics. 2011; 8:645-657.

[10] Cox J, Neuhauser N, Michalski A, Scheltema R A, Olsen J V, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res 2011; 10(4):1794-1805.

[11] Perkins D N, Pappin D J, Creasy D M, Cottrell J S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999; 20 (18): 3551-3567.

[12] Eng J K, McCormack A L, Yates J R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom 1994; 5:976-989.

[13] Zhang Y, Fonslow B R, Shan B, Baek M C, Yates J R 3rd. Protein analysis by shotgun/bottom-up proteomics. Chem Rev. 2013; 113 (4):2343-94.

[14] Gonzalez-Galarza F F, Lawless C, Hubbard S J, Fan J, Bessant C, Hermjakob H, Jones A R. A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis. OMICS. 2012; 16 (9):431-42.

[15] Carvalho P C, Han X, Xu T, Cociorva D, da Gloria Carvalho M, Barbosa V C, et al. XDIA: Improving on the label-free data-independent analysis. Bioinformatics. 2010; 26 (6):847–848.

[16] Carvalho P C, Xu T, Han X, Cociorva D, Barbosa V C, Yates J R. YADA: A tool for taking the most out of high-resolution spectra. Bioinformatics. 2009; 25 (20):2734–2736.

[17] Pandeswari P B, Sabareesh V. Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. RSC Adv. 2019; 9:313–44.

[18] Kumar D, Yadav A K, Dash D. (2017). Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data. In: Keerthikumar, S., Mathivanan, S. (eds) Proteome Bioinformatics. Methods in Molecular Biology, vol 1549. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-6740-7_3

[19] Pandeswari P B, Sabareesh V. An ESI Q-TOF study to understand the impact of arginine on CID MS/MS characteristics of polypeptides. International Journal of Mass Spectrometry 2021, 459:116453.

[20] Patthy L, Smith E L. Reversible Modification of Arginine Residues. J. Biol. Chem. 1975; 250:557-564.

[21] Takahashi K. The Reaction of phenylglyoxal with arginine residues in proteins. J. Biol. Chem. 1968, 243:6171-6179.

FIGURES



Figure 1. Pipeline for DC- PPMA (Schematic Illustration)



Figure 2. Screenshot of DC-PPMA's homepage. Window for each module can be opened by clicking the respective button given in the homepage.

(a)

1	MS_2022-0	7-02 21	47_05.310	55.xlsx [Repaired] ×											•
	A	В	С	D	E	F	G	н	T	J	K	L	M	N	
	MS REPOR	RT: sp P	11836 1-2	07_Rituximab											
	Cleavage	site(s): '	V8 E(GluC	E											
1	No of Ami	no Acid	Residues:	297											
5	Molecular	mass of	Protein:	33055.74367399999 (Monoisotopic)											
5	Basic Resi	dues: R	(8) K(16)												
	Acidic Res	idues: 0	D(4) E(29												
8	Hydrophil	ic Resid	ues: 5(30)	T(15) C(5) N(16) Q(11)											
	Hydropho	bic Resi	dues: G(1	r) A(15) P(22) V(12) L(25) I(32) M(14) F(13) Y(7) W(3)											
0	Number o	f Peptid	es: 30												
1	Modificati	ions: ca	rbamidom	ethylcysteine:C81 C111 C167 C183 C220											
2	Missed Cl	eavage:	0												
3	S.No	Position	h Length	Sequence	Mass	Modified	[M+1H]1+	[M+2H]2+	[M+3H]3+	[M+4H]4+	[M+5H]5+	[M+6H]6+			
4	1	1 - 15	15	MTTPRNSVNGTFPAE	1620.756	5 -	1621.7644	811.38612	541.26002	406.19697	325.15914	271.133923	9999999		
5	2	16 - 48	33	PMKGPIAMQSGPKPLFRRMSSLVGPTQSFFMRE	3707.892	4 -	3708.9002	1854.9540	1236.9719	927.98093	742.58631	618.989897	3333333		
6	3	49 - 105	57	SKTLGAVQIMNGLFHIALGGLLMIPAGIYAPICVTVWYPLWGGIMYIISGSLLAAT	6003.198	1 6060.2196	6061.2274	3031.1176	2021.0810	1516.0627	1213.0517	1011.04442	73333337		
7	4	106 - 15	0 45	KNSRKCLVKGKMIMNSLSLFAAISGMILSIMDILNIKISHFLKME	5065.757	6 5122.7790	5123.7869	2562.3973	1708.6008	1281.7025	1025.5636	854.804340	6666669		
8	5	151 - 16	8 18	SLNFIRAHTPYINIYNCE	2167.052	0 2224.0735	2225.0813	1113.0445	742.36565	557.02620	445.82252	371.686742	3333333		
9	6	169 - 17	4 6	PANPSE	613.2707	4 -	614.27856	307.64319	205.43140	154.32551	123.66197	103.219615	66666668		
D	7	175 - 20	5 31	KNSPSTQYCYSIQSLFLGILSVMLIFAFFQE	3573.792	7 3630.8142	3631.8220	1816.4149	1211.2792	908.71137	727.17066	606.143527	3333333		
1	8	206 - 21	3 8	LVIAGIVE	812.5007	2 -	813.50854	407.25818	271.84139	204.13300	163.50796	136.424612	33333333		
2	9	214 - 21	5 2	NE	261.0960	8 -	262.10390	131.55586	88.039853	66.281845	53.227041	44.5238389	9999999		
3	10	216 - 23	3 18	WKRTCSRPKSNIVLLSAE	2087.130	9 2144.1524	2145.1602	1073.0840	715.72529	537.04592	429.83830	358.366560	6666666		
4	11	234 - 23	4 1	E	147.0531	5 -	148.06097	74.534401	50.025542	37.771113	30.418455	25.5166839	99999998		
5	12	235 - 23	7 3	KKE	403.2430	7 -	404.25089	202.62936	135.42218	101.81859	81.656439	68.2150039	9999998		
6	13	238 - 24	1 4	QTIE	489.2434	7	490.25129	245.62956	164.08898	123.31869	98.856519	82.548404			
7	14	242 - 24	4 3	IKE	388,2321	7 -	389,23999	195,12391	130,41854	98.065868	78.654259	65 7131873	3333332		

(b)

1 A	R																			
		C	D	E	F	G	н		1	K	L	M	N	0	Р	Q	R	S	T	L
MS/M	S REPORT: 5	p P11836 1	-297_Rituxi	imab																
Cleava	ge site(s):	V8 E(GluC)	E																	
No of A	Amino Acid	Residues: 2	97																	
Molecu	ular mass of	Protein: 33	8055.743673	999999 (Mo	noisotopi	:)														
Basic R	esidues:: F	(8) K(16)																		
Acidic I	Residues: (D(4) E(29)																		
Hydrop	philic Resid	ues: S(30)	T(15) C(5)) N(16)	Q(11)															
Hydrop	phobic Resi	dues: G(17)	A(15) P	(22) V(12) L(25)	(32) M(14) F(13)	Y(7) W(3	1)											
Numbr	er of Peptid	es: 30																		
Modifi	cations: ca	rbamidomet	thylcystein	e:C81 C11	1 C167 C1	83 C220														
Missed	Cleavage:	0																		
Fragme	ent: MTTPR	NSVNGTFPA	E Position:	[1-15] N	Aonoisoto	oic Mass: 16	20.756593	9999997	Modified	Mass: 1620	.76442 C	harge State	s: [M+1	H]1+:1621.	772244999	9999 [M-	2H]2+:811	.390035	[M+3H]3+	541.2
z = +6	z = +5	z = +4	z = +3	z = +2	b (mod)	b (ions)	Residue I	Sequence	e Residue	Ny (ions)	y (mod)	z = +2	z = +3	z = +4	z = +5	z = +6				
22.847	906 27.2159	22 33.76794	7 44.687988	66.52807	-	132.04831	1	M	15	-	-	-	-	-	-	-				
39.689	186 47.4254	59 59.02986	7 78.370548	117.05191	L -	233.09599	2	т	14	1490.7239	-	745.86587	497.57986	373.43685	298.95104	249.29384	249999998			
56.530	466 67.6349	95 84.29178	7 112.05310	167.57575	5 -	334.14367	3	т	13	1389.6762	-	695.34203	463.89730	348.17493	278.74151	232.45256	25			
72.705	926 87.0455	47 108.5549	7 144.40402	216.10213	b -	431.19643	4	P	12	1288.6285	-	644.81819	430.21474	322.91301	258.53197	215.61128	250000002			
98.722	778 118.265	76 147.5802	5 196.43773	294.15268	<u>k</u> -	587.29754	5	R	11	1191.5758		596.29181	397.86382	298.64982	239.12142	199.43582	25			
117.72	993 141.074	35 176.0909	8 234.45204	351.17419	ş	701.34047	6	N	10	1035.4747	-	518.24126	345.83011	259.62454	207.90120	173.41897	083333333			
132.23	527 158.480	76 197.8489	9 263.46271	394.69016	j	788.37250	7	s	9	921.43176	-	461.21979	307.81580	231.11381	185.09261	154.41181	583333332			
148.74	567 178.294	44 222.6160	9 296.48552	444.2243	7.	887.44091	8	V	8	834.39974	-	417.70378	278.80513	209.35580	167.68620	139.90647	75			
167.75	382 201.103	02 251.1268	2 334.49983	501.24583	k -	1001.4838	9	N	7	735.33133	-	368.16957	245.78232	184.58870	147.87252	123.39507	583333334			
177.25	740 212.507	32 265.3821	9 353.50698	529.75650	ş.	1058.5053	10	G	6	621.2884		311.14811	207.76801	156.07796	125.06394	104.38792	083333334			
194.09	868 232.716	85 290.6441	1 387.18954	580.28040	A-	1159.5529	11	т	5	564.26694		282.63738	188.76086	141.82260	113.65964	94.884344	16666666			
218.61	008 262.130	53 327.4112	1 436.21234	653.81460		1306.6213	12	F	4	463.21925	-	232.11354	155.07830	116.56068	93.450111	78.043064	16666667			
224 70	554 281.541	09 351.6744	468.56326	702.34098	5 -	1403.6741	13	P	3	316.15085	-	158.57933	106.0555	79,793581	64.03643	53.531662	5			

Figure 3. Screenshots of excel files obtained as outputs from Module 1: (a) MS database and (b) MS/MS database.

(a)

A	В	C	D	E	F	G	Н	1	J	K	L	M	N	0	P	Q	R	S	T
sheet_names	Mass	Peptides																	
sp[P02666]16-22	4 Casein																		
	1,646.3	2959,100 - 105	6,EAMAP	K,645.305	502,646.312	2845,323.66	033500000	003,216.10	949833333	333,162.334	08,130.06	8829,108.5	586616666	6668					
	1,742.4	558 16,203 - 20	9,7,GPFP	IV,741.43	192,742.43	9745,371.72	3785,248.	1517983333	33332,186.3	65805,149.	294209,124	4.57981166	666667						
	1,748.3	782 11,108 - 1	3,6,EMPF	PK,747.35	197,748.35	979500000	01,374.683	810000000	05,250.125	1483333333	3,187.8458	3175,150.4	7821900000	0002,125.56	648666666	668			
	1,780.5	04313,170 - 1	6,7,VLPV	PQK,779.4	7994,780.4	877650000	001,390.74	779500000	005,260.83	4471666666	7,195.8778	31,156.903	813,130.92	1148333333	35				
	1,830.4	592 14, 177 - 11	3,7,AVPY	PQR,829.4	43406,830.4	4418850000	001,415.72	485500000	005,277.48	5845,208.36	634,166.8	946370000	0002,139.2	46835					
sp[P02768]25-60	9 Human-S	erum-Albumin																	
	1,673.3	97131,213 - 2	18.6.AWA	AR,672.3	6016,-,673.	367985,337	187905,22	5.1278783	333333,169	.097864999	99998,135.	479856999	99998,113.	067851666	56667				
	1,933.5	54(9,74 - 81,8	LCTVATL	R,875.4793	3,932.5007	5,933.50858	5,467.2582	2050000000	3,311.8414	1166666666	6,234.1330	15,187.507	977,156.42	461833333	334				
	1,517.2	72537,258 - 20	2.5 ADLA	<,516.280	18,-,517.28	8005,259.14	7915,173.	1012183333	33332,130.0	7787,104.2	538609999	9999,87.05	452166666	667					
	1,517.2	72573,542 - 5	15,4,EQLK	516.28019	300000001,	-,517.28801	50000001,:	259.147920	00000006,	173.1012216	66666667,1	30.077872	5,104.2638	630000000	,87.054523	133333335			
sp P11836 1-297	Rituximab																		
	1,748.3	782 18,219 - 23	24,6,TCSR	PK,690.33	773000000	01,747.3591	900000001	,748.3670	150000002,	374.6874200	0000001,25	50.1275550	0000003,18	37.8476225	0000003,15	0.47966300	0000002,128	5.56769000	000000
sp[P01588]28-19	3_Erythrop	pietin																	
sp P27918 28-46	9_Properdir	1																	
sp P02192 2-154	Myoglobul	in																	
	1,748.3	782 19,134 - 13	9,6,ALELF	R,747.417	734,748.425	5165,374.71	6495,250.1	469383333	333,187.86	216,150.491	292999999	998,125.57	738166666	667					
sp P61823 27-15	0_RNase-A																		
	2,1112	552 14, 105 - 12	24,20,HIIVA	CEGNPY	VPVHFDA:	SV,2166.04	519999999	98,2223.06	765999999	97,2224.075	485,1112.5	5416549999	998,742.03	3037833333	33,556.774	74,445.621	3569999999	9,371.5191	01666
	2,1259.	12(8,40 - 61,2	2,CKPVNT	FVHESLA	DVQAVCS	QK,2402.1	516599999	998,2516.2	0457999999	996,2517.21	240499999	98,1259.11	011499999	97,839.742	684999999	8,630.0589	6999999999,	504.248740	09999
	2,1182.	9724,11 - 31,2	1,QHMDS	STSAASS	SNYCNOM	MK,2306.8	919,2363.9	1336,2364.	921185,118	2.964505,78	38.978945,	591.98616	5,473.79049	9699999999	,394.99338	15			
	1,915.4	06(12,92 - 98,	7, YPNCAY	'K,857.363	361,914.385	507,915.392	895000000	1,458.2003	600000000	5,305.80284	83333333.	229.60409	25,183.884	839,153.40	336666666	67			
sn/P00698119-14	7 Lysozym	0																	

(b)

	A	В	C	D	E	F	G	н	1	J	K	L	M	N	0	P	Q	R
she	eet_names	Mass	Peptides															
sp	P02666 16-224	Casein																
sp	P02768 25-609	Human-Ser	um-Albumir	1														
		1075.4810	Fragment:	NECFLQH	K Position:	[99-106]	MASS: 1	017.45963	MODIFIE	D MASS:	1074.48109							
sp	P11836 1-297_	Rituximab																
		1103.0923	3 Fragment:	TLGAVQIN	INGLFHIAI	LGGLLMIP	AGIYAPIC	VTVWYPL	VGGIMYIIS	GSLLAAT	EK Position	: [51-106]	MASS: 4	916.155560	000002	MODIFIED	MASS: 597	3.17702
sp	P01588 28-193	Erythropoie	tin															
sp	P27918 28-469	Properdin																
		810.84032	2 Fragment:	GLLGGGV	SVEDCCL	NTAFAYQ	K Position	[18-39]	MASS: 224	4.0448899	999997 N	10DIFIED I	MASS: 235	8.08781				
) sp	P02192 2-154_I	Myoglobulin																
		1079.5771	7 Fragment:	VEADVAG	HGQEVLI	R Position:	[17-31]	MASS: 159	1.82120999	999998	MODIFIED N	ASS: -						
2 sp	P61823 27-150	RNase-A																
sp 3	P00698 19-147	Lysozyme																
sp	P02754 17-178	Lactoglobul	in															
sp	P01012 2-386_0	Ovalbumin																
i sp	P06278 30-512	Amylase																
7 sp	P02787 20-698	Serotransfe	min															
s sp	P00918[2-260_0	Carbonic_An	hydrase2															
e sp	P9999992-105_0	Cytochrome-	C															
) sp	P02647[25-267	Apolipoprot	ein_A-1															
l sp	P02769 25-607	Bovine-Seru	um-Albumin															
2																		

Figure 4. Screenshots of excel files obtained as outputs from Module 2 showing matched hits from the list of protein sequences entered as input: (a) MS based peptide hits;

(b) MS/MS based peptide hits

				Creation c	or custom Data	ibase			
Input Protein seque	nce(s):								
RELEELNVPGEIV VENLHLPLPLLQ TPEGDLEILLQK FKELKVHHANEN	VESLSSSEESITF SWMHQPHQPLPPT KENGECAQKKIIA IFYCPIAIMSALA	RINKKIEKFQSEEQC FVMFPFQSVLSLSQS AEKTKIFAVFKIDAI AMVYLGAKDSTRTQI	QTEDELQDKIHP KVLPVPQKAVPY NENKVLVLDTDY NKVVRFDKLPGF	FAQTQSLVYPFPGPIPN PQRDMPIQAFLLYQEPV KKYLLFCMENSAEPEQS GDSIEAQCGTSVNVHSS	ISLPONIPPLTOTPVVVP VLGPVRGPPPIIV,LIVTO SLACQCLVRTPEVDDEALE SLRDILNQITKPNDVYSF	PFLQPEVMGVSKVKEAMAI 2THKGLDIQKVAGTWYSLJ 2KFDKALKALPHHIRLSFI SLASRLYAEERYPILPEYI	PKHKEMPFPKYPVEPFTE: MMAASDISLLDAQSAPLR NPTQLEEQCHI,GSIGAA LQCVKELYRGGLEPINFQ1	SOSLTLTD VYVEELKP SMEFCFDV TAADQARE	Import Reset
Cleavage Methods:									
Proteases	Trypsin		•	C Chemical		•	No. of Missed cles	avage(s):	0
Modification(s):									
Chemical/PTMs:	AddNew		•	carbamidometh; ,3-C382,5-C9,	ylcysteine:2-C66,2 5-C19,5-C39,5-C48,	-C106,2-C119,2-C12 5-C118,5-C137,5-C1	1,2-C160,3-C11,3-C 58,5-C161,5-C171,5	:30,3-C73,3 -C174,5-C1	-C120,3-C3 77,5-C179,
Chemical/PTMs: New Name :	AddNew	New Mass :	•	carbamidometh; ,3-C382,5-C9; C194,5-C227,5 5-C484,5-C495	ylcysteine:2-C66,2 5-C19,5-C39,5-C48, -C241,5-C331,5-C33 ,5-C498,5-C506,5-C	-C106,2-C119,2-C12 5-C118,5-C137,5-C1 9,5-C345,5-C355,5- 523,5-C563,5-C577,	1,2-C160,3-C11,3-C 58,5-C161,5-C171,5 C368,5-C377,5-C402 5-C596,5-C615,5-C6	30,3-C73,3 -C174,5-C1 ,5-C418,5- 20,5-C637,	-C120,3-C3 77,5-C179, C450,5-C47 5-C665,5-C
Chemical/PTMs: New Name : Amino Acids :	AddNew CHD	New Mass : View Position	112.05243 Clear	carbanidometh ,3-C382,5-C9; C194,5-C227,5 5-C484,5-C495, 4,6-C204,7-C1 76,9-C198,9-C 7301 9-C436 9	ylcysteine:2-C66,2 S-C19,5-C39,5-C48, -C241,5-C331,5-C33 ,5-C498,5-C506,5-C 4,7-C17,9-C34,9-C5 244,9-C245,9-C252, -C437 0-C447 0-C44	-C106,2-C119,2-C12 5-C118,5-C137,5-C1 9,5-C345,5-C355,5- 523,5-C563,5-C577, 3,9-C62,9-C75,9-C9 9-C264,9-C277,9-C2 0 9-C475 9-C476 9-	1,2-C160,3-C11,3-C 58,5-C161,5-C171,5 C368,5-C377,5-C402 5-C596,5-C615,5-C6 0,9-C91,9-C101,9-C 78,9-C288,9-C315,9 C486 9-C513 9-C557	30,3-C73,3 -C174,5-C1 5-C418,5- 20,5-C637, 123,9-C167 -C359,9-C3 0 -C558 4-	-C120,3-C3 77,5-C179, C450,5-C47 5-C665,5-C ,9-C168,9- 60,9-C368, ~566 10-C3
Chemical/PTMs: New Name : Amino Acids : Create Databases:	AddNew CHD R	New Mass : View Position	112.05243 Clear	carbamidometh ,3-C382,5-C9,1 C194,5-C227,5 5-C484,5-C495, 4,6-C204,7-C1 76,9-C199,9-C C301 0-C376 0	yloyateine:2-C66,2 5-C19,5-C39,5-C48, -C241,5-C331,5-C33 ,5-C498,5-C506,5-C 4,7-C17,9-C34,9-C5 244,9-C245,9-C252, _r437 0_r447 0_r44	-C106,2-C119,2-C12 5-C118,5-C137,5-C1 9,5-C345,5-C355,5- 533,5-C565,5-C577, 3,9-C62,9-C75,9-C7 9-C264,9-C277,9-C2 0 0-C475 0-C476 0-	1,2-C160,3-C11,3-C 58,5-C161,5-C171,5 C368,5-C377,5-C402 5-C396,5-C615,5-C6 10,9-C91,9-C101,9-C 178,9-C288,9-C315,9 C488, 0-C513, 0-C557	30,3-C73,3 C174,5-C1 ,5-C418,5- 20,5-C637, 123,9-C167 C359,9-C3 0-C558 9-	-C120,3-C3 77,5-C179, C450,5-C47 5-C665,5-C ,9-C168,9- 60,9-C368, C566 10-C3
Chemical/PTMs: New Name : Amino Acids : Create Databases: C Peptide m	AddNew CHD R	New Mass : View Position	112.05243 Clear	carbamidometh ,3-C382,5-C9,1 C194,5-C227,5 5-C484,5-C495, 4,6-C204,7-C1 76,9-C199,9-C C301 0-C436 0	ylcysteine:2-C66,2 5-C19,5-C39,5-C48, -C241,5-C331,5-C33 ,5-C498,5-C506,5-C 4,7-C17,9-C34,9-C5 244,9-C245,9-C252, _r437 0-r447 0-r44	-C106, 2-C119, 2-C12 5-C118, 5-C137, 5-C13 9, 5-C345, 5-C355, 5- 523, 5-C563, 5-C577, 3, 9-C62, 9-C78, 9-C9 9-C264, 9-C277, 9-C2 0 0-P475 0-P476 0-	11,2-C160,3-C11,3-C 58,5-C161,5-C171,5 C368,5-C377,5-C40 5-C596,5-C615,5-C6 0,9-C91,9-C101,9-C 78,9-C289,8-C315,9 C486 0-C513 0-C557 Multiple protonation st	30, 3-C73, 3 C174, 5-C1 5-C418, 5- 120, 5-C637, 123, 9-C167 C359, 9-C3 1 0-C559, 9-C3 1 0-C559, 9-C3 ates (ESI):	-C120,3-C3 77,5-C179, C450,5-C47 5-C665,5-C ,9-C168,9- 60,9-C368, C566 10-C3
Chemical/PTMs: New Name : Amino Acids : Create Databases: C Peptide m C MS/MS ba	AddNew CHD R ass Only (MS only) sed search	New Mass : View Position	↓ 112.05243 Clear	Carbamidometh , 3-0.382,5-05, 5-0484,5-0227,5- 5-0484,5-049,7-01 76,9-0299,9-0 76,9-0299,9-0 76,9-0299,9-0	ylcysteine:2-C66,2 5-C19,5-C39,5-C48, -C241,5-C33,5-C33,5-C33,5-C33,5-C33,5-C33,5-C33,5-C34,5-C248,5-C248,5-C252, ra17_0-ra17_	-C106,2-C119,2-C12 5-C118,5-C137,5-C1 9,5-C345,5-C355,5 523,5-C543,5-C555,5 523,5-C543,5-C577, 3,5-C24,9-C77,5-C 0,0-C47,5-C77,5-C 0,0-C47,5-C-C47,0-C	11,2-C160, 3-C11, 3-C 59,5-C161,5-C171,5 C368,5-C377,5-C402 5-C596,5-C615,5-C6 0,5-C51,5-C101,5-C 748,6-C515,5-C65,5-C6 (C48,6-C513,6-C557 Multiple protonation st (Carge states 2+10+	30, 3-C73, 3 -C174, 5-C1 , 5-C418, 5- 120, 5-C637, 123, 9-C167 -C359, 9-C3 -C359, 9-C3 -c359, 9-C3 -c559, 9-C379, 9-C37	-C120, 3-C3 77, 5-C179, C450, 5-C47 5-C665, 5-C 9-C168, 9- 60, 9-C368, C566, 10-C3 5 2
Chemical/PTMs: New Name : Amino Acids : Create Databases: C Peptide m C MS/MS ba Fragment ions:	AddNew CHD R ass Only (MS only) sed search	New Mass : View Position	112.05243 Clear	Carbamidometh , 30.382,5-05, 5-0484,5-0227,5- 5-0484,5-049,7-01 76,9-0299,9-0 76,9-0299,9-0 76,9-0299,9-0	<pre>ylcysteine:2-C66, 2 5-C19, 5-C39, 5-C49, 5-C30, 5-C30, 5-C30, 5-C30, 5-C30, 5-C500, 5-C241, 5-C304, 5-C500, 5-C244, 5-C244, 5-C244, 5-C245, 5-C252,ray7 0_r44, 5-C245,ray7 0_r45, 5-C255,ray7 0_r45, 5-C255,ray7</pre>	-C106,2-C119,2-C12 5-C116,5-C137,5-C1 9,5-C345,5-C355,5 523,5-C563,5-C577, 9-C264,9-C277,9-C2 9-C264,9-C277,9-C2 0 0-C475,9-C476,0-C277,9-C2	11,2-C160, 3-C11, 3-C 56, 5-C161, 5-C111, 5- C560, 5-C377, 5-C402 5-C596, 5-C215, 5-C0 0, 9-C91, 9-C101, 9-C 778, 9-C289, 9-C315, 9 C486 6-C513 6-C53 Multiple protonation at (Charge states - + 10 +	30, 3-C73, 3 -C174, 5-C1 , 5-C410, 5- 120, 5-C637, 123, 9-C167 -C359, 9-C3 -C359, 9-C3 -c-558 6- ates (ESI): -25	-C120, 3-C3 77, 5-C179, C450, 5-C47 5-C665, 5-C 9-C168, 9- 60, 9-C368, C566, 10-C3
Chemica/PTMs: New Name : Amino Acids : Create Databases: C Peptide m G MS/MS ba Fragment ions: Fr b ic	AddNew CHD R ass Only (MS only) sed search	New Mass : View Position	112.05243 Clear aions	Carbanidometh) 3-C382,5-C5); C194,5-C227,5 5-C484,5-C495, 5-C484,5-C495, 1-C484,5-C495, -	<pre>ylcysteine:2-C66, 2 5-C19, 5-C39, 5-C49, 5-C49, c241, 5-C33, 1, 5-C3 y 5-C499, 5-C506, 5-C 244, 9-C249, 9-C249, 9-C2 244, 9-C249, 9-C249, 9-C2 244, 9-C245, 9-C352,</pre>	-C106,2-C119,2-C12 5-C116,5-C137,5-C1 5-C35,5-C353,5-C353,5 523,5-C553,5-C553,5-C577, 5-C264,5-C277,5-C2 - 0775,5-C5 - 0755,5-C5 - 0755,5-C5	11,2-C160, 3-C11, 3-C 569,5-C161, 5-C171, 5 569,5-C171, 5-C172, 5 5-C596, 5-C715, 5-C0 9, 5-C51,9-C101, 9-C 769, 9-C219, 9-C101, 9-C 769, 9-C289, 9-C315, 9 Case 6-C513, 6-C537 Multiple protonation st (Charge states z+ 10 + Zions	(30, 3-C73, 3 (-C174, 5-C1) (, 5-C418, 5-(1) (, 5-C418, 5-(1) (, 5-C418, 5-(1) (-C359, 9-C3) (-C359, 9-C	-C120, 3-C3 77, 5-C179, C450, 5-C47 5-C665, 5-C 60, 9-C368, 9- 60, 9-C368, 10-C3 5 5 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Chemical/PTMs: New Name : Amino Acids : Create Databases: C Peptide m G MS/MS ba Fragment ions: Fragment ions: F b io	AddNew CHD R ass Only (MS only) sed search ons H ₂ O	New Mass : <u>View Position</u> y jons y y H ₂ O	▼ 112.05243 Clear □ a ions □ a - H ₂ O	Carbanidometh ,3-C382,3-C5, C184,5-C227,5 5-C484,5-C484, 4,6-C204,7-C1 76,9-C185,5-C (-15,5-C185,5-C (-15,5-C185,5-C (-15,5-C185,5-C) (-15,5-C)	ylcysteine:2-C66,2 S-C19,S-C39,S-C49,S-C49, C-241,S-C31,S-C33 S-C498,S-C506,S-C 244,S-C24,S-C34,S-C 244,S-C245,S-C34,S- c-447 c-r44 ange	-C106,2-C119,2-C12 -C119,5-C137,5-C1 5,-C139,5-C385,5- 523,5-C385,5-C385,5- 523,5-C38,5-C385,5- 5,-C26,9-C27,5,-C5 -C26,5,5-C3,5-C5 -C26,5,5-C3,5-C5 -C26,5,5-C3,5-C5 -C26,5-C3,5-C5 -C26,5-C3,5-C5 -C26,5-C3,5-C5 -C26,5-C3,5-C5 -C26,5-C3,5-C5 -C26,5-C5	11,2-C160, 3-C11, 3-C 58,5-C161,5-C111, 5-C 586,5-C197,5-C121,5 5-C586,5-C2101,9-C 70,5-C21,5,-C101,9-C 70,5-C216,5,-C101,9-C 70,5-C216,5,-C315,5 Multiple protonation st (Charge states z= +10+ Carge states z= +10+ z ions z z-H ₂ O	(30, 3-C73, 3 (-C174, 5-C1) (, 5-C418, 5- (20, 5-C637, (-C359, 9-C37) (-C359, 9-C37) (-C359, 9-C37) (-C359, 9-C37) (-C359, 9-C37) (-C359, 9-C37) (-C359, 9-C37) (-C359, 9-C37) (-C359) (-C3	-C120, 3-C3 77, 5-C179, C450, 5-C479, 5-C665, 5-C , 9-C168, 9 60, 9-C368, 10- 5 5 5 5 5 5 2 5 2 5 2 5
Chemical/PTMs: New Name : Amino Acids : Create Databases: C Peptide m C MS/MS ba Fragment ions: F bi b - b - b - b - b -	AddNew CHD R ass Only (MS only) sed search H ₅ O NH ₅	New Mass : View Position	▼ 112.05243 Clear □ a ions □ a - H ₂ O □ a - NH ₃	Cashami dometh , 2-032, 3-05, Clash, 3-022, 3-05, d, 2-023, 5-05, d, 2-023, 7-05, d, 2-05, d, 2-05, d	y1cysteine:2-C66,2 5-C18,0-C18,5-C16, -C181,5-C131,5-C13 5-C481,5-C135,5-C16,5-C2 5-C481,5-C16,5-C2 4,5-C18,5-C16,5-C2 -r447 α-r447 α-r44 mage Γ cims Γ c-N60 Γ c-N66		11,2-C140,3-C11,3-C 83,5-C41,5-C17,5-6 645,5-C477,5-6 645,5-C477,5-6 645,5-C477,5-6 645,5-C477,5-C4 945,5-C48,5-C475,5-C47	(30, 3- C73, 3 (-C174, 5- C1 (5 - C140, 5- (20, 5- C637, (123, 9- C167) (-C355, 9- C167) (-C355	

Figure 5. Screenshot of GUI of Module 1: Incorporation of new modification, i.e., CHD modification of Arginine (R) residues is highlighted.