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 Pandeswari[1], Isaac Emerson[1], and Varatharajan Sabareesh[1]

[1]Vellore 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: bottom-up, 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 & y 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 Top-down 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: 1,2-cyclohexanedione (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/zvalues 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 ofm/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

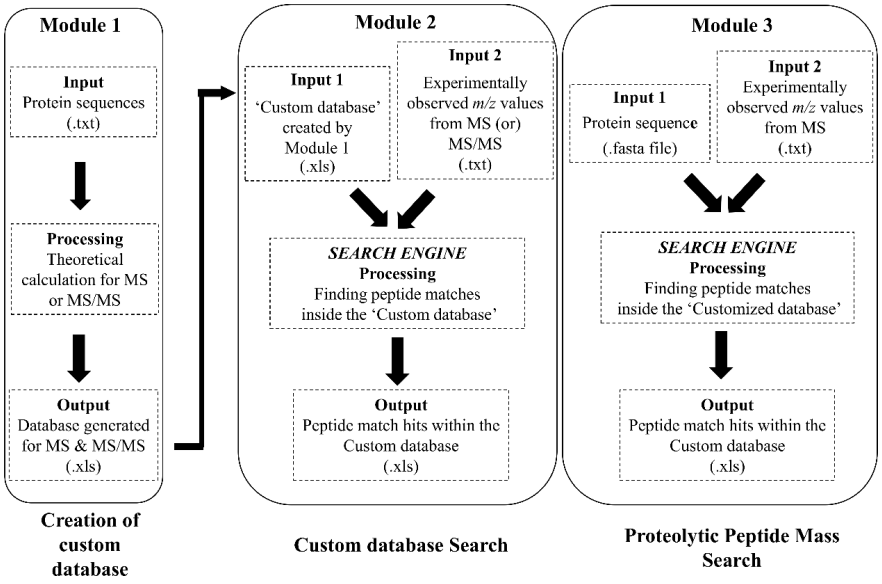


Figure 1. Pipeline for DC- PPMA (Schematic Illustration)

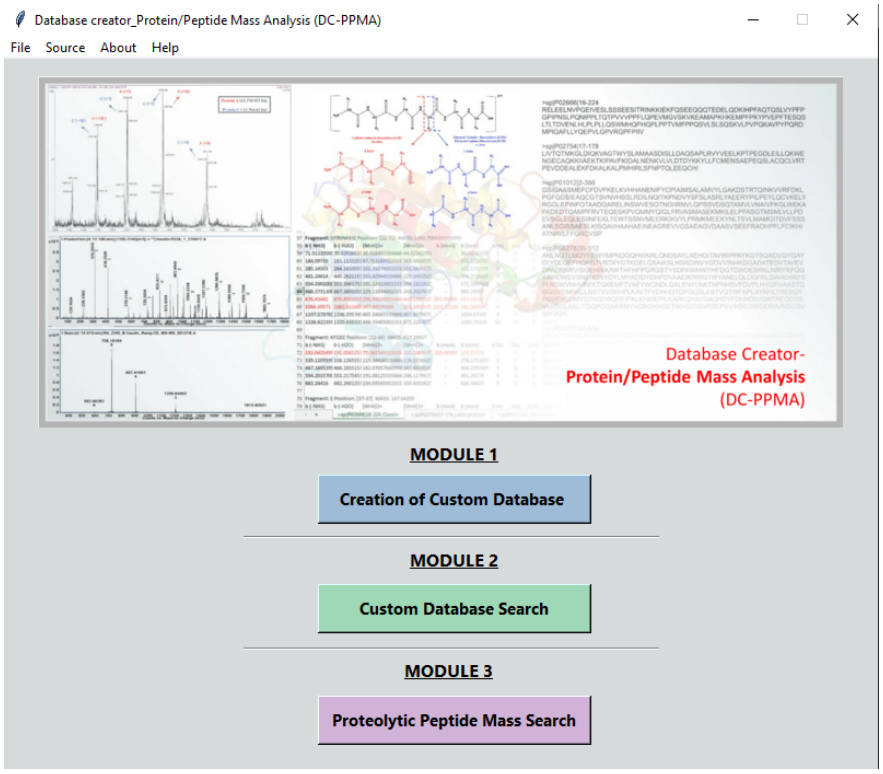
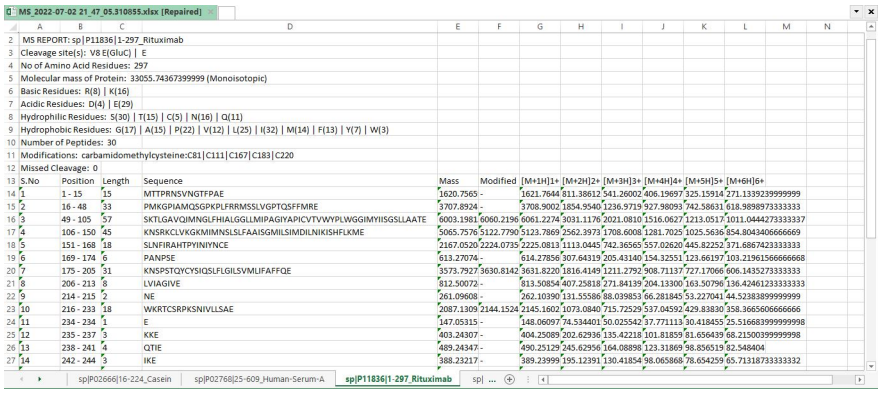


Figure 2. Screenshot of DC-PPMA's homepage. Window for each module can be opened by clicking the respective button given in the homepage.

(a)



MS_2022-07-02 21_47_05.310855.xlsx [Repaired]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
2	MS REPORT: sp|P11836|1-297_Rituximab													
3	Cleavage site(s): V8 E(GluC) | E													
4	No of Amino Acid Residues: 297													
5	Molecular mass of Protein: 33055.74367399999 (Monoisotopic)													
6	Basic Residues: R(8) | K(16)													
7	Acidic Residues: D(4) | E(29)													
8	Hydrophilic Residues: S(30) | T(15) | C(5) | N(16) | Q(11)													
9	Hydrophobic Residues: G(17) | A(15) | P(22) | V(12) | L(25) | I(32) | M(14) | F(13) | Y(7) | W(3)													
10	Number of Peptides: 30													
11	Modifications: carbamidomethylcysteine:C81|C111|C167|C183|C220													
12	Missed Cleavage: 0													
13	S.No	Position	Length	Sequence	Mass	Modified	[M+1H]1+	[M+2H]2+	[M+3H]3+	[M+4H]4+	[M+5H]5+	[M+6H]6+		
14	1	1 - 15	15	MTTPRNSVNGTFPAE	1620.7565	-	1621.7644	811.38612	541.26002	406.19697	325.15914	271.1339239999999		
15	2	16 - 48	33	PMKGPIAMQSGPKPLFRRMSSLVGPTQSFFMRE	3707.8924	-	3708.9002	1854.9540	1236.9719	927.98093	742.58631	618.9898973333333		
16	3	49 - 105	57	SKTLGAVQIMNGLFHIALGGLLMIPAGIYAPICVTVWYPLWGGIMYIISGSLLAATE	6003.1981	6060.2196	6061.2274	3031.1176	2021.0810	1516.0627	1213.0517	1011.0444273333337		
17	4	106 - 150	45	KNSRKCLVKGKMIMNSLSLFAAISGMILSIMIDILNIKISHFLKME	5065.7576	5122.7790	5123.7869	2562.3973	1708.6008	1281.7025	1025.5636	854.8043406666669		
18	5	151 - 168	18	SLNFIRAHTPYINIYNCE	2167.0520	2224.0735	2225.0813	1113.0445	742.36565	557.02620	445.82252	371.6867423333333		
19	6	169 - 174	6	PANPSE	613.27074	-	614.27856	307.64319	205.43140	154.32551	123.66197	103.21961566666668		
20	7	175 - 205	31	KNSPSTQYCYSIQSLFLGILSVMLIFAFFQE	3573.7927	3630.8142	3631.8220	1816.4149	1211.2792	908.71137	727.17066	606.1435273333333		
21	8	206 - 213	8	LVIAGIVE	812.50072	-	813.50854	407.25818	271.84139	204.13300	163.50796	136.4246123333333		
22	9	214 - 215	2	NE	261.09608	-	262.10390	131.55586	88.039853	66.281845	53.227041	44.52383899999999		
23	10	216 - 233	18	WKRTCSRPKSNIVLLSAE	2087.1309	2144.1524	2145.1602	1073.0840	715.72529	537.04592	429.83830	358.3665606666666		
24	11	234 - 234	1	E	147.05315	-	148.06097	74.534401	50.025542	37.771113	30.418455	25.51668399999998		
25	12	235 - 237	3	KKE	403.24307	-	404.25089	202.62936	135.42218	101.81859	81.656439	68.21500399999998		
26	13	238 - 241	4	QTIE	489.24347	-	490.25129	245.62956	164.08898	123.31869	98.856519	82.548404		
27	14	242 - 244	3	IKE	388.23217	-	389.23999	195.12391	130.41854	98.065868	78.654259	65.71318733333332		

sp|P02666|16-224_Casein | sp|P02768|25-609_Human-Serum-A | sp|P11836|1-297_Rituximab | sp| ...

(b)

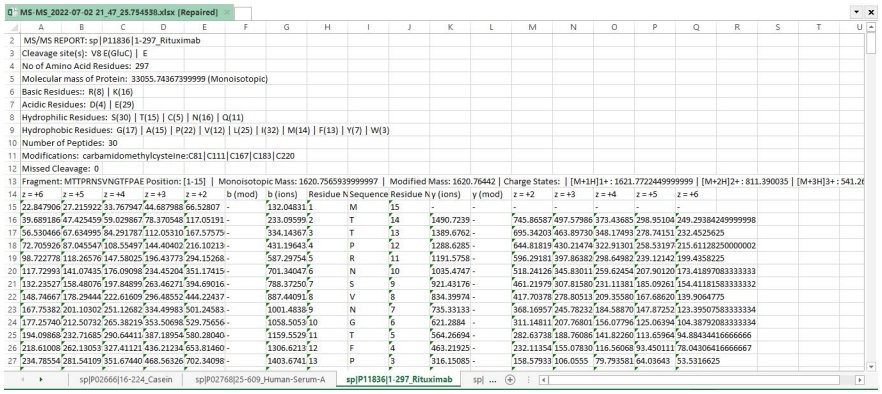
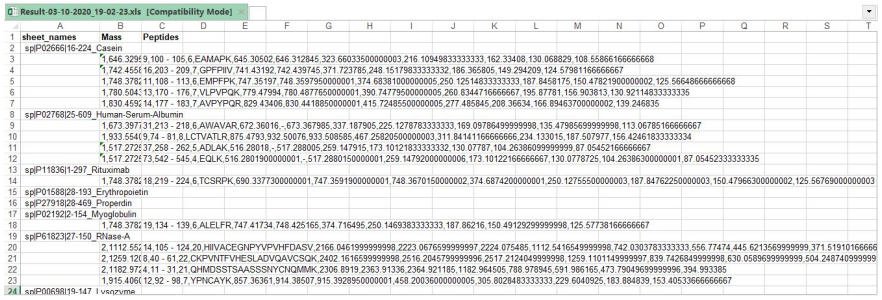


Figure 3. Screenshots of excel files obtained as outputs from Module 1: (a) MS database and (b) MS/MS database.

(a)



(b)

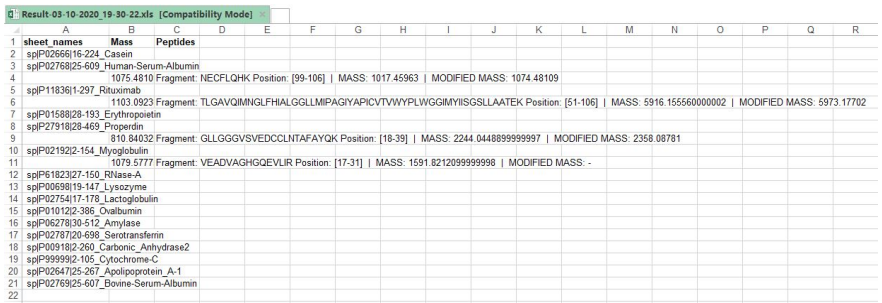


Figure 4. Screenshots of excel files obtained as outputs from Module 2 showing matched hits from the list of protein sequences entered as input: (a) MS based peptide hits;

(b) MS/MS based peptide hits

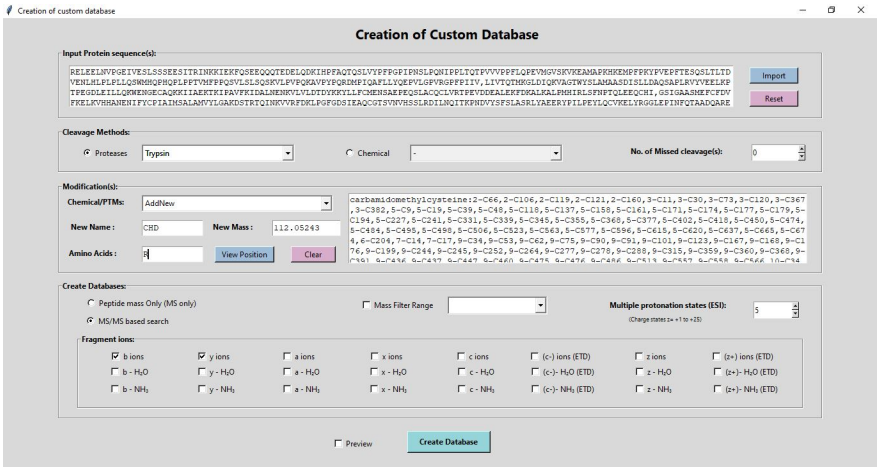


Figure 5. Screenshot of GUI of Module 1: Incorporation of new modification, i.e., CHD modification of Arginine (R) residues is highlighted.