

PERCEPTRON *v 1.0.0.0*

A Next Generation Top-Down Proteoform
Identification and Characterization Platform

USER MANUAL

Biomedical Informatics Research Laboratory, Department of Biology

Lahore University of Management Sciences

<http://birl.lums.edu.pk/>

1 Table of Contents

1	Table of Contents	2
2	Table of Figures	3
3	Introduction to PERCEPTRON	4
3.1.	About PERCEPTRON	4
3.2.	Features	4
4	Hardware and Software.....	6
4.1.	Hardware	6
4.2.	Software	6
4.3.	Testing.....	6
5	Video Tutorials	7
6	Getting Started with PERCEPTRON.....	7
7	GUI description:.....	9
7.1	Window 1: PERCEPTRON Tool for Top-down Proteomics.....	9
7.2	Window 2: Protein Search Query.....	10
7.3	Window 3: Summary and Detailed Results View.....	15
8	Search.....	16
8.1	File Formats.....	16
8.1.1	Raw to mzXML File Format Conversion	16
8.1.2	Raw to mzML File Format Conversion	16
8.1.3	MzXML to MGF File Format Conversion	17
8.1.4	MGF to Flat Text File Format Conversion	17
8.2	Parameters	18
8.3	Databases.....	18

8.4	Modes	18
9	References	20

2 Table of Figures

Figure 1.	PERCEPTRON Homepage	7
Figure 2.	PERCEPTRON Login options	8
Figure 3.	PERCEPTRON - Overview of User Interface	9
Figure 4.	PERCEPTRON - Overview of Basic Parameters	10
Figure 5.	PERCEPTRON - Overview of Experimental Parameters	11
Figure 6.	PERCEPTRON - Overview of De Novo Sequencing Parameters	12
Figure 7.	PERCEPTRON - Overview of Protein Modifications Parameters	13
Figure 8.	PERCEPTRON - Overview of Scoring Component Weight	14
Figure 9.	Summary Results window showing candidate proteins	15
Figure 10.	Detailed Results window showing candidate proteins	15
Figure 11.	Conversion of raw to mzXML	16
Figure 12.	Conversion of raw to mzML	17
Figure 13.	Conversion of mzXML to MGF	17
Figure 14.	Load Default Parameters	18
Figure 15.	Selecting Protein Database	18

3 Introduction to PERCEPTRON

This chapter introduces the user to the PERCEPTRON application and describes its basic features.

3.1. About PERCEPTRON

PERCEPTRON is a freely available web-based proteoform identification pipeline for Top-Down Proteomics (TDP). Top-down proteomics is an emerging experimental protocol for analysis of intact proteoforms. PERCEPTRON search pipeline brings together algorithms for: (i) intact mass tuning, (ii) *de novo* peptide sequence tag extraction, (iii) *in silico* spectral comparison, (iv) identification of post-translational modifications as well as truncated proteins, and (v) a novel composite scoring scheme for candidate protein scoring. PERCEPTRON achieves high performance by leveraging NVIDIA GPU technology coupled with Microsoft ASP.NET and ANGULAR frameworks. The search results obtained include a list of proteins, their scores and details on the matching information. This information can be visualized as well as downloaded. Overall, PERCEPTRON is aimed at filling the crucial void of open-source and open-architecture protein identification software for TDP data, employing state-of-the-art algorithms.

3.2. Features

The salient features of the pipeline are summarized below:

- **Graphical User Interface** - A set of rich and intuitive graphical user interface has been developed for setting up the search parameters as well as for integrating the main components of the engine.
- **Whole Protein Molecular Weight Estimation** - The protein identification begins with the tuning of precursor protein's monoisotopic MW (MS1) as guided by its fragmentation spectra (MS2). Relative abundances and mass/charge (m/z) ratios are used to calculate the consensus MW which is then employed in the search and scoring process.
- **Peptide Sequence Tag Extractor** - Peptide sequence tag ladders (PST) are extracted from the spectra by enumerating successive peaks having MW differences equal to an amino acid and within the user specified mass tolerance. Protein database is then filtered for proteins reporting these PSTs. The length of PST ladders, cumulative mass off-sets and relative abundances are used in calculating the PST scores.
- ***In silico* fragmentation** – *In silico* fragments of candidate proteins are generated by the user selected fragmentation techniques. *In vitro* and *in silico* spectral comparisons are performed and scored.

- **Post-translational Modification (PTM) Search** - Support for predicting typical PTMs has been provided in the tool. Users can select and search variable and fixed PTMs of their choice along with blind-PTMs by simply selecting them from the GUI.
- **Multifactorial Composite Scoring System** - A multifactorial candidate protein scoring scheme incorporating the aforementioned algorithms has been developed. User customization of the parameters and weights in the scoring function is admitted via a GUI.
- **Single and Batch Search** – PERCEPTRON provides support for search in single as well as batch modes. Towards an automated processing of multiple spectral data files, a batch processing mode allows for the selection of multiple files from the folder by clicking the attach file button. The experimental spectra, search parameters and results are automatically stored in the project directory for further processing and visualization.

4 Hardware and Software

4.1. Hardware

For deployment, PERCEPTRON requires GPU that supports CUDA TOOLKIT 7.0

4.2. Software

PERCEPTRON requires:

- Windows Server 2012 R2
- Visual Studio 2013
- Angular 1.7.4
- Node.js 8.11.1
- SQL Server Management Studio 17.6
- CUDA TOOLKIT 7.0
- CUDAFy.NET.1.29.5576.13786
- Microsoft Windows Server v6.2
- .NET Framework 4.5

4.3. Testing

PERCEPTRON has been deployed and tested on Dell Power Edge R730, 2 x Intel Xeon E5-2620, 160 GB RAM (16GBx10), NVIDIA Tesla K40C (2880 Cores). Following Windows are compatible with PERCEPTRON:

- Windows 8.1
- Windows 7
- Win Server 2012 R2
- Win Server 2008 R2

5 Video Tutorials

Video tutorials have been provided for: (i) using PERCEPTRON, (ii) search in single and batch modes, (iii) performing search for MGF and text files, and (iv) interpreting results. The videos are available as a playlist at: https://www.youtube.com/playlist?list=PLaNVq-kFOOn0Z_7b-iL59M_CeV06JxEXmA

6 Getting Started with PERCEPTRON

PERCEPTRON manual, samples and issues database is freely available (under the MIT open license) at (<https://perceptron.lums.edu.pk/>)

To log in, click on the link provided above.

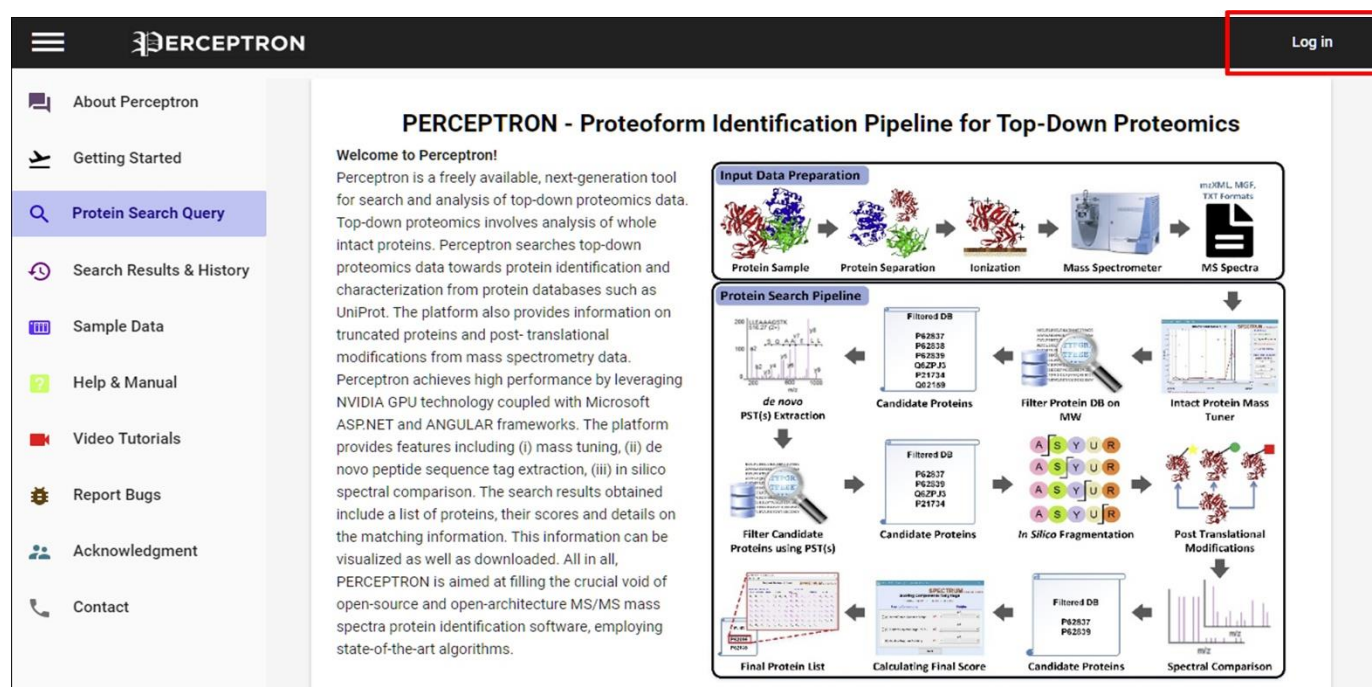


Figure 1. PERCEPTRON Homepage

Click the 'log in' button on the top right. A window will appear. Enter details to proceed.

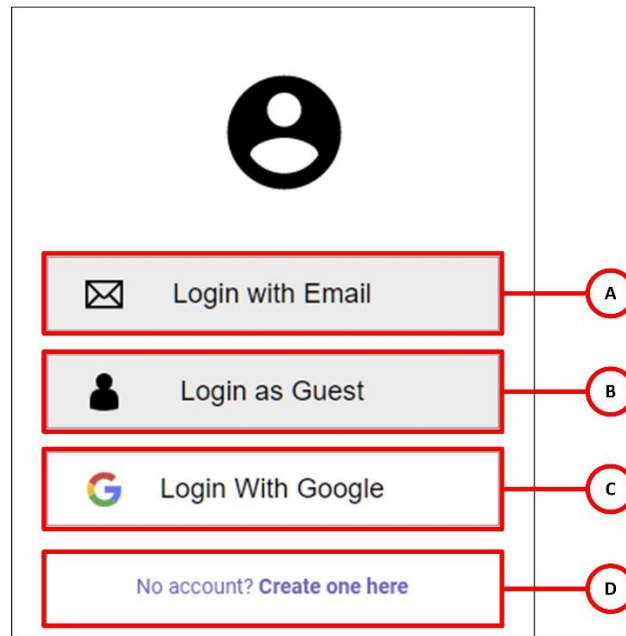


Figure 2. PERCEPTRON Login options

- A. User can login with their email account
- B. Enables user to login as a guest, perform search in PERCEPTRON but the results will not be saved
- C. User can only login via their g-mail account
- D. If the user does not have any account, they can make a new one here in order to login.

7 GUI description:

This chapter presents the interface overview for user facilitation.

7.1 Window 1: PERCEPTON Tool for Top-down Proteomics

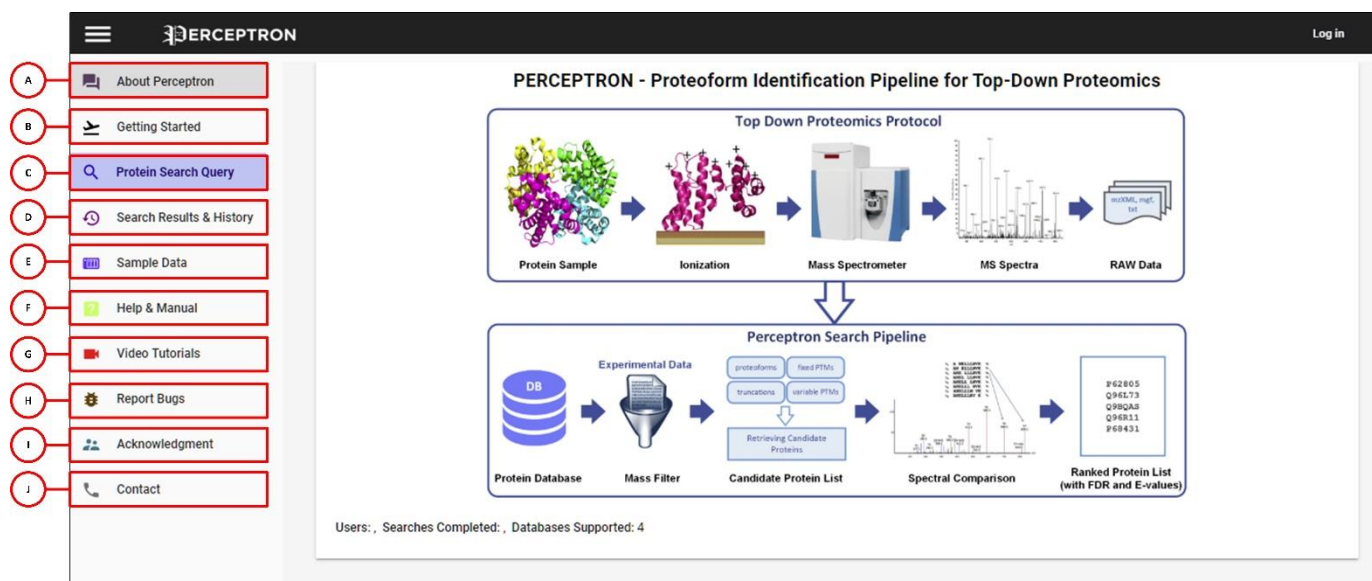


Figure 3. PERCEPTON - Overview of User Interface

- A. About PERCEPTON: A next-generation top-down proteoform search and identification platform
- B. Getting Started: Quick guide to proteoform search and identification using PERCEPTON
- C. Protein Search Query: Job submission - Search top-down proteomics data files to identify and characterize proteoforms
- D. Search Results & History: View search results and history
- E. Sample Data: Sample top-down mass spectrometry data for use with PERCEPTON
- F. Help & Manual: Get assistance with using PERCEPTON and download manual
- G. Video Tutorials: View step-by-step video tutorials demonstrating usage of PERCEPTON
- H. Report Bugs: Report problems and issues here
- I. Acknowledgement: PERCEPTON project team members
- J. Contact: Contact us for further information

In order to start protein search, click on the 'Protein Search Query' tab and the following window will appear:

7.2 Window 2: Protein Search Query

Basic Parameters:

The screenshot shows the 'Submit Protein Identification Query' form in the PERCEPTRON application. The form is divided into several sections. On the left is a sidebar with navigation links: About Perceptron, Getting Started, Protein Search Query (highlighted), Search Results & History, Sample Data, Help & Manual, Video Tutorials, Report Bugs, Acknowledgment, and Contact. The main form area is titled 'Submit Protein Identification Query' and contains the following sections:

- Basic Parameters** (with a link to 'Load Default Parameters'):
 - Protein Search Title**: A text input field (labeled A).
 - Intact Protein Mass**: A text input field (labeled B).
 - Upload Mass-Spectrometry Data File(s)**: A file upload button with a '0 Files Selected' indicator (labeled C).
 - Select Database**: A dropdown menu (labeled D).
 - Enter your E-mail Address ***: A text input field (labeled E).
 - Recieve Top**: A dropdown menu (labeled F).
 - Results**: A button (labeled F).
- Set Experimental Parameters**: A collapsed section.
- Set De Novo Sequencing Parameters**: A collapsed section.
- Set Protein Modifications Parameters**: A collapsed section.
- Set Scoring Components Weight**: A collapsed section.

At the bottom of the form are 'Reset' and 'Submit' buttons. The footer contains 'Contact Us' and '(c) 2019 Biomedical Informatics Research Laboratory (BIRL)'.

Figure 4. PERCEPTRON - Overview of Basic Parameters

- A. In order to start the protein search, user has to enter the 'Protein Search Title' first (for example: MyProject)
- B. User can enter intact protein mass (MS1)
- C. Browse and upload protein database (.fasta)
- D. Browse and upload experimental data (.mzXML/ .MGF/ .txt) for Single mode; Peak-list files for Batch mode
- E. Enter your email address. You will be notified about the compilation of your results via email
- F. Select the number of candidate protein hits to be received in results

Set Experimental Parameters:

The screenshot displays the 'PERCEPTRON' web application interface. The main section is titled 'Submit Protein Identification Query'. Below this, there are 'Basic Parameters' and a 'Set Experimental Parameters' section. The 'Set Experimental Parameters' section contains several input fields and checkboxes, each highlighted with a red box and a corresponding letter (A-I) in a red circle. The parameters are: Mass Mode (radio buttons for MH+ and M(Neutral)), Intact Mass Tolerance (text input with a unit dropdown), Peptide Tolerance (text input with a unit dropdown), Select Fragmentation Type (dropdown menu), Filter Database using MS1 (checkbox), Handle Truncated Proteoforms (checkbox), Tune Intact Protein Mass (checkbox), Neutral Mass Loss (text input with a unit dropdown), and Select Corresponding Special Ions (dropdown menu). The interface also includes a sidebar with navigation links, a top header with the PERCEPTRON logo and user status, and a bottom section with 'Reset' and 'Submit' buttons.

Figure 5. PERCEPTRON - Overview of Experimental Parameters

- A. Select Mass Mode. MS data can only be provided in either m/z form with $z = 1$ or neutral masses
- B. Set the tolerance value for Protein Mass and select its unit
- C. Select the tolerance value for Peptide and select its unit
- D. Select the 'Fragmentation type' from drop down menu
- E. User can filter database by checking the option 'Filter Database using MS1'
- F. Check 'Handle Truncated Proteoforms' to allow search for truncated proteoforms
- G. Provide the value of Neutral loss, if any
- H. Choose the corresponding special ions for the type of fragmentation selected (i.e. a', b', y', z'', a*, b*, y*, z' ions)
- I. Check the option 'Tune Intact Protein Mass' to allow for tuning of MS1 using MS2 data

Set De Novo Sequencing Parameters:

The screenshot displays the PERCEPTRON web interface for submitting a protein identification query. The left sidebar contains navigation links: About Perceptron, Getting Started, Protein Search Query (selected), Search Results & History, Sample Data, Help & Manual, Video Tutorials, Report Bugs, Acknowledgment, and Contact. The main content area is titled 'Submit Protein Identification Query' and includes a 'Basic Parameters' section with fields for 'Protein Search Title', 'Intact Protein Mass', 'Da', 'Upload Mass-Spectrometry Data File(s)', '0 Files Selected', 'Select Database', 'Enter your E-mail Address *', 'Recieve Top', and 'Results'. Below this is the 'Set Experimental Parameters' section, which is expanded to show 'Set De Novo Sequencing Parameters'. This section contains five parameters: 'Enable PST Filtering' (checkbox, labeled A), 'Tolerance For Each Hop' (dropdown menu, labeled B), 'Minimum Tag Length' (dropdown menu, labeled C), 'Maximum Tag Length' (dropdown menu, labeled D), and 'Set the overall mass error tolerance for the whole PST' (text input field, labeled E). Below these are 'Set Protein Modifications Parameters' and 'Set Scoring Components Weight' sections. At the bottom right are 'Reset' and 'Submit' buttons.

Figure 6. PERCEPTRON - Overview of De Novo Sequencing Parameters

- A. Check 'Enable PST Filtering' to filter PSTs
- B. Set the 'Tolerance for each Hop'
- C. Tags will be filtered above the minimum length of PST selected from the drop down menu by the user
- D. Tags will be filtered below the maximum length of PST selected from the drop down menu by the user
- E. Overall mass error tolerance shows error margin for the whole PST

Set Protein Modifications Parameters:

The screenshot shows the PERCEPTRON web interface. The main section is titled 'Submit Protein Identification Query'. Under 'Basic Parameters', there are fields for 'Protein Search Title', 'Intact Protein Mass', 'Da', 'Upload Mass-Spectrometry Data File(s)', 'Select Database', 'Enter your E-mail Address *', 'Recieve Top', and 'Results'. Below this are expandable sections: 'Set Experimental Parameters', 'Set De Novo Sequencing Parameters', and 'Set Protein Modifications Parameters'. The 'Set Protein Modifications Parameters' section is expanded and contains several sub-sections: 'Terminal Modification' (A), 'Blind-PTM Search' (B), 'PTM Tolerance' (C), 'Methionine Chemical Modifications' (D), 'Variable Modifications' (E), 'Fixed Modifications' (F), and 'Cysteine Chemical Modifications' (G). A list of modifications is provided on the left, and buttons for 'Reset' and 'Submit' are at the bottom.

Figure 7. PERCEPTRON - Overview of Protein Modifications Parameters

- A. Allows user to select specified terminal modifications. PERCEPTRON handles four cases: 1) None – No modification, 2) NME – N terminal methionine excision, 3) NME_ACETYLTATION – N terminal acetylation with initiator methionine removed, and 4) M_ACETYLTATION – N terminal methionine acetylation
- B. Allows user to perform Blind-PTM search and find unknown modifications
- C. Set the tolerance value for Post Translational Modification (PTM)
- D. Allows the user to select instrument specific modification on Methionine
- E. User can opt for required Variable ‘Post translation Modifications’ from the list of modifications
- F. Allows the user to select instrument specific modification on Cysteine
- G. Similarly, various ‘Fixed Modifications’ are also selected from the list

Set Scoring Components Weight:

The screenshot displays the PERCEPTRON web application interface. The sidebar on the left contains navigation links: About Perceptron, Getting Started, Protein Search Query (highlighted), Search Results & History, Sample Data, Help & Manual, Video Tutorials, Report Bugs, Acknowledgment, and Contact. The main content area is titled 'Submit Protein Identification Query'. It includes a 'Basic Parameters' section with fields for 'Protein Search Title', 'Intact Protein Mass' (Da), 'Upload Mass-Spectrometry Data File(s)' (0 Files Selected), 'Select Database', 'Enter your E-mail Address *', 'Recieve Top' (dropdown), and 'Results'. Below this are expandable sections for 'Set Experimental Parameters', 'Set De Novo Sequencing Parameters', 'Set Protein Modifications Parameters', and 'Set Scoring Components Weight'. The 'Set Scoring Components Weight' section is expanded and highlighted with a red box, showing three sliders for 'Intact Protein Mass Score Weightage (%)', 'Peptide Sequence Tags Score Weightage (%)', and 'Spectral Comparisons Score Weightage (%)'. A red circle with the letter 'A' points to the 'Spectral Comparisons Score Weightage (%)' slider. At the bottom right are 'Reset' and 'Submit' buttons.

Figure 8. PERCEPTRON - Overview of Scoring Component Weight

- A. Set the respective weights of the Scoring Components from by shifting the slider left or right accordingly

7.3 Window 3: Summary and Detailed Results View

The screenshot shows the 'Summary Results View' window in the PERCEPTRON application. The left sidebar contains navigation links: About Perceptron, Getting Started, Protein Search Query (selected), Search Results & History, Sample Data, Help & Manual, Video Tutorials, Report Bugs, Acknowledgment, and Contact. The main content area displays a table of candidate proteins. Below the table, there is a red asterisk note and a footer with contact information.

Protein Rank	Protein Name	Protein ID	Molecular Weight	Terminal Modification	No. of Modification(s)	Protein Score
1	Protein 0	Q5TBE3	11359.7861	no	2	0.3159
2	Protein 1	P62805	11473.4114	no	2	0.0340
3	Protein 2	Q99525	11115.1934	no	2	0.0274
4	Protein 3	A6NFZ4	11363.5022	no	2	0.0240
5	Protein 4	Q9C005	11355.8325	no	2	0.0227
6	Protein 5	Q9P1G2	11483.6480	no	2	0.0204
7	Protein 6	Q5T752	11334.5254	no	2	0.0183
8	Protein 7	Q9BTM9	11486.0357	no	2	0.0141
9	Protein 8	Q5T292	11431.1326	no	2	0.0136
10	Protein 9	P09341	11407.2531	no	2	0.0129
11	Protein 10	P19875	11494.3653	no	2	0.0129

* Click on the table row for corresponding detailed results.

Contact Us (c) 2018 Biomedical Informatics Research Laboratory (BIRL)

Figure 9. Summary Results window showing candidate proteins

Resultant proteins along with their protein ID, molecular weight and protein score according to the uploaded and selected data are represented in the list. Click on any protein to see the 'Detailed Result View' of the protein. Click on the 'protein ID' to go to the detailed UniProt view of the protein.

The screenshot shows the 'Detailed Result View' window for Protein P62805. The left sidebar is the same as in Figure 9. The main content area displays 'User Search Parameters' and 'General Results'. Below this, there is a 'Protein Search Time' section and a list of peptide sequences.

User Search Parameters

General Results

Protein Rank:	2	Protein ID:	P62805	Protein Name:	P62805
Protein Score:	0.034039	Molecular Weight:	11473.411430	# Matched Fragments:	12
Terminal Modification:	NH2	Truncation:	No	# Modifications:	3

* Click on the Protein ID to access its information.

Protein Search Time (in Seconds)

Total:	00:00:00.0470323	Molecular Weight Module:	N/A	Peptide Sequence Tag Module:	00:00:00.0111953
Spectral Comparison Module:	00:00:00.0000004	Post-Translational Modification Module:	00:00:00.0039530	Truncation Module:	00:00:00.0470323

1. MSGRGKGK 2. LGKGAKRHR 3. KVLRDNIQI 4. TKPAIRRLR 5. RGVKRISGL 6. IYEETRGVLK 7. VFLENVIRDA 8. VTYTEHAKRK 9. TVTAMDVVYA 10. LKRGKRTLYG 11. FGG

Contact Us (c) 2018 Biomedical Informatics Research Laboratory (BIRL)

Figure 10. Detailed Results window showing candidate proteins

8 Search

8.1 File Formats

PERCEPTRON provides support for plain text files (data in columns containing mass to charge ratios (m/z) and relative intensities), eXtensible Markup Language (XML) files with m/z and relative abundances (mzXML)¹, Mass Spectrometry Markup Language (mzML)^{2,3} and Mascot Generic Format (MGF)⁴ data formats in both single and batch file processing modes.

8.1.1 Raw to mzXML File Format Conversion

User can convert raw data files to mzXML file format by using MS-Convert⁵.

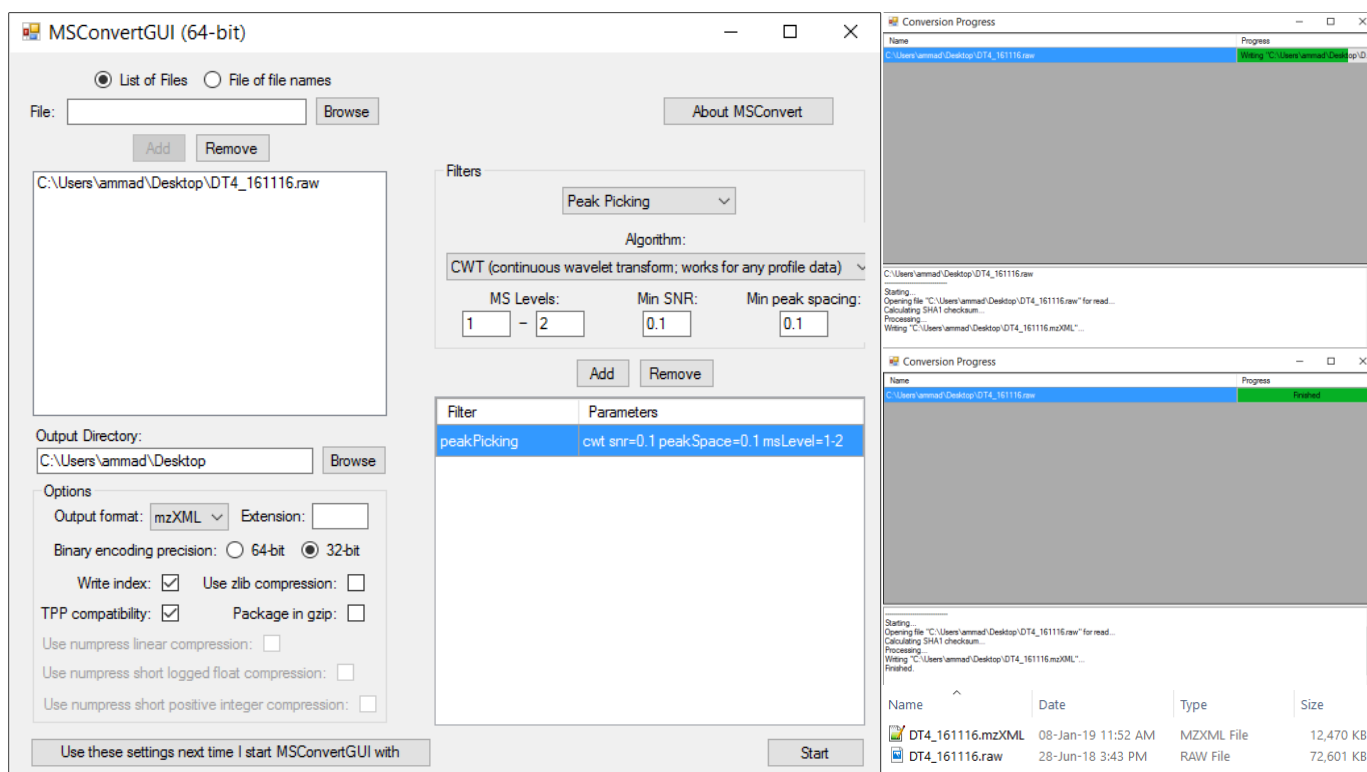


Figure 11. Conversion of raw to mzXML

8.1.2 Raw to mzML File Format Conversion

Raw data files can be converted to mzML file format by using MS-Convert⁵.

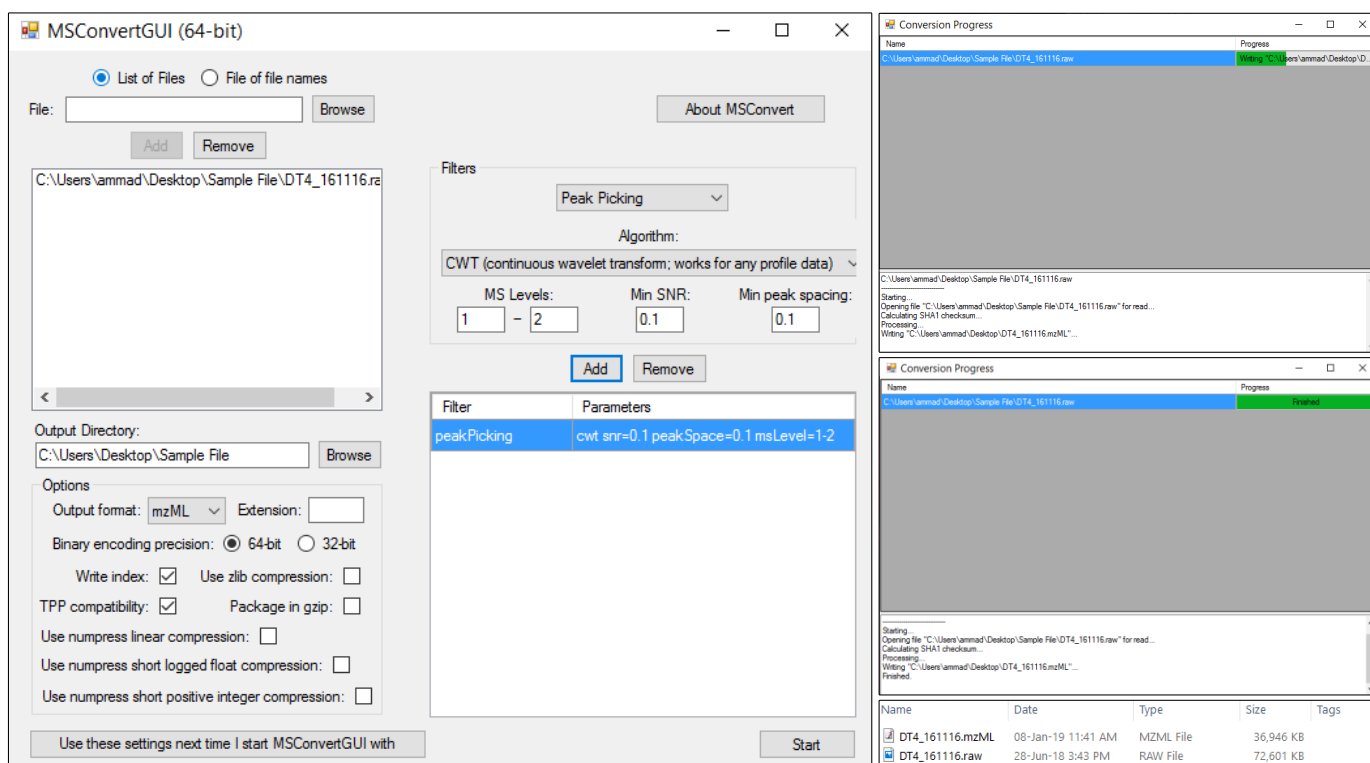


Figure 12. Conversion of raw to mzML

8.1.3 MzXML to MGF File Format Conversion

User can convert mzXML files to MGF using MS-Decov⁶.

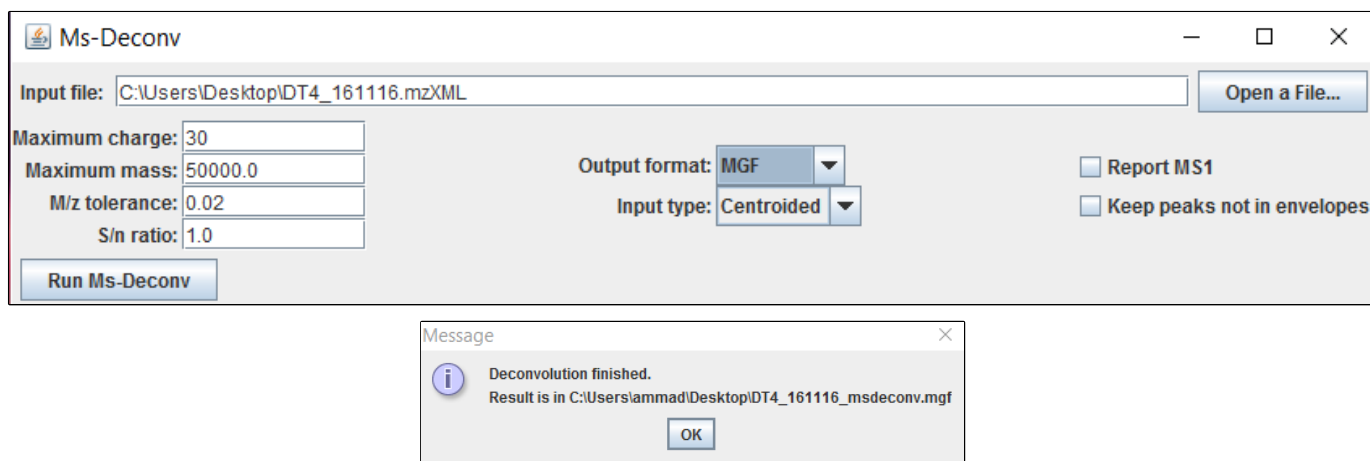


Figure 13. Conversion of mzXML to MGF

8.1.4 MGF to Flat Text File Format Conversion

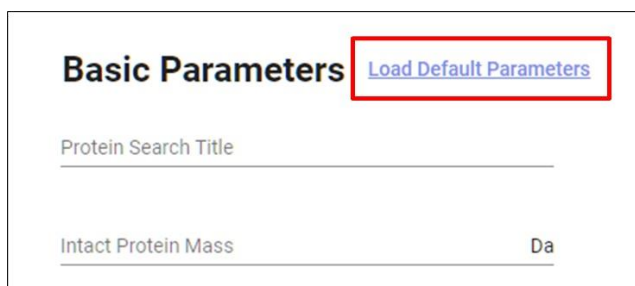
PERCEPTRON converts MGF files to flat text (peak list) using built-in custom file reader.

8.2 Parameters

PERCEPTRON works when all parameters are set. Two kind of parameters can be used by the user which includes: (i) Default Parameters, and (ii) Selected parameters.

How to load default Parameters?

To submit the job using default parameters, select 'Load Default Parameters' option in front of Basic Parameters.

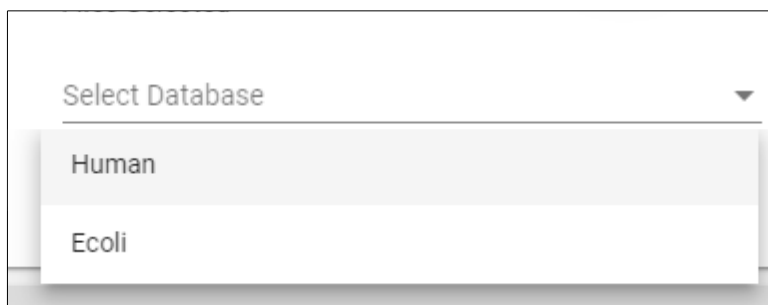


The image shows a web form titled "Basic Parameters". To the right of the title is a blue hyperlink "Load Default Parameters" which is enclosed in a red rectangular box. Below the title, there are two input fields: "Protein Search Title" and "Intact Protein Mass". The "Intact Protein Mass" field has a unit "Da" to its right.

Figure 14. Load Default Parameters

8.3 Databases

SwissProt database is included in PERCEPTRON by default. User can take any protein sequence from other databases (such as Uniprot) in .fasta format.



The image shows a dropdown menu with the label "Select Database". The menu is open, displaying two options: "Human" and "Ecoli".

Figure 15. Selecting Protein Database

8.4 Modes

The search modes are auto-selected based on the number of files that the user inputs. If one file is given as input, PERCEPTRON runs single search mode whereas if multiple files are given as input, the mode switches to batch mode.

- (i) Single Search Mode
- (ii) Batch Mode

Batch mode takes more processing time as it deals with larger data. The experimental spectra, search parameters and results are automatically stored in the project directory for further processing and visualization.

9 References

1. Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 2004;22(11):1459–1466.
2. Turewicz M, Deutsch EW. Spectra, chromatograms, Metadata: mzML-the standard data format for mass spectrometer output. In: *Data mining in proteomics*. Springer; 2011. p 179–203.
3. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD. mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011;10(1):R110. 000133.
4. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20(18):3551–3567.
5. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 2012;30(10):918.
6. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA. Deconvolution and database search of complex tandem mass spectra of intact proteins a combinatorial approach. *Mol Cell Proteomics* 2010;9(12):2772–2782.