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/

Last Modified: 20-12-2019

1 Table of Contents

1 Table of Contents 2
2 Table of Figures 4
3 Introduction to PERCEPTRON 5
3.1. About PERCEPTRON 5
3.2. Features 5
4 Hardware and Software 7
4.1. Hardware 7
4.2. Software 7
4.3. Testing 7
5 Video Tutorials 8
6 Getting Started with PERCEPTRON 8
7 GUI description: 10
7.1 Window 1: PERCEPTRON Tool for Top-down Proteomics 10
7.2 Window 2: Protein Search Query 11
7.3 Window 3: Summary and Detailed Results View 16
8 Search 17
8.1 File Formats 17
8.1.1 Raw to mzXML File Format Conversion 17
8.1.2 Raw to mzML File Format Conversion 17
8.1.3 MzXML to MGF File Format Conversion 18
8.1.4 MGF to Flat Text File Format Conversion 18
8.2 Parameters 19
8.3 Databases 19

8.4 Modes 19

9 References 21

2 Table of Figures

Figure 1. PERCEPTRON Homepage and Log in button 8
Figure 2. PERCEPTRON Login options 9
Figure 3. PERCEPTRON - Overview of User Interface 10
Figure 4. PERCEPTRON - Overview of Basic Parameters 11
Figure 5. PERCEPTRON - Overview of Experimental Parameters 12
Figure 6. PERCEPTRON - Overview of De Novo Sequencing Parameters 13
Figure 7. PERCEPTRON - Overview of Protein Modifications Parameters 14
Figure 8. PERCEPTRON - Overview of Scoring Component Weight 15
Figure 9. Summary Results window showing candidate proteins 16
Figure 10. Detailed Results window showing candidate proteins 16
Figure 11. Conversion of raw to mzXML 17
Figure 12. Conversion of raw to mzML 18
Figure 13. Conversion of mzXML to MGF 18
Figure 14. Load Default Parameters 19
Figure 15. Selecting Protein Database 19

3 Introduction to PERCEPTRON

This chapter introduces the user to the PERCEPTRON platform along with a description of its 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 (GUI)** - 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 in-house deployment of PERCEPTRON, the hardware requirement includes a graphical processing unit (GPU) that supports CUDA TOOLKIT 7.0.

4.2. Software

Compilation and software build of PERCEPTRON requires the following software:

- 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

The freely available public deployment of PERCEPTRON is hosted on a Dell Power Edge R730, 2 x Intel Xeon E5-2620, 160 GB RAM (16GBx10) and an NVIDIA Tesla K40C (2880 Cores). Following Windows versions are compatible with PERCEPTRON:

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

5 Video Tutorials

Several video tutorials have been developed for employment of PERCEPTRON in protein search. These tutorials are available as a playlist at: https://www.youtube.com/playlist?list=PLaNVq-kFOn0Z_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 initiate use of PERCEPTRON, the user has to register and log in by clicking on the link provided above.

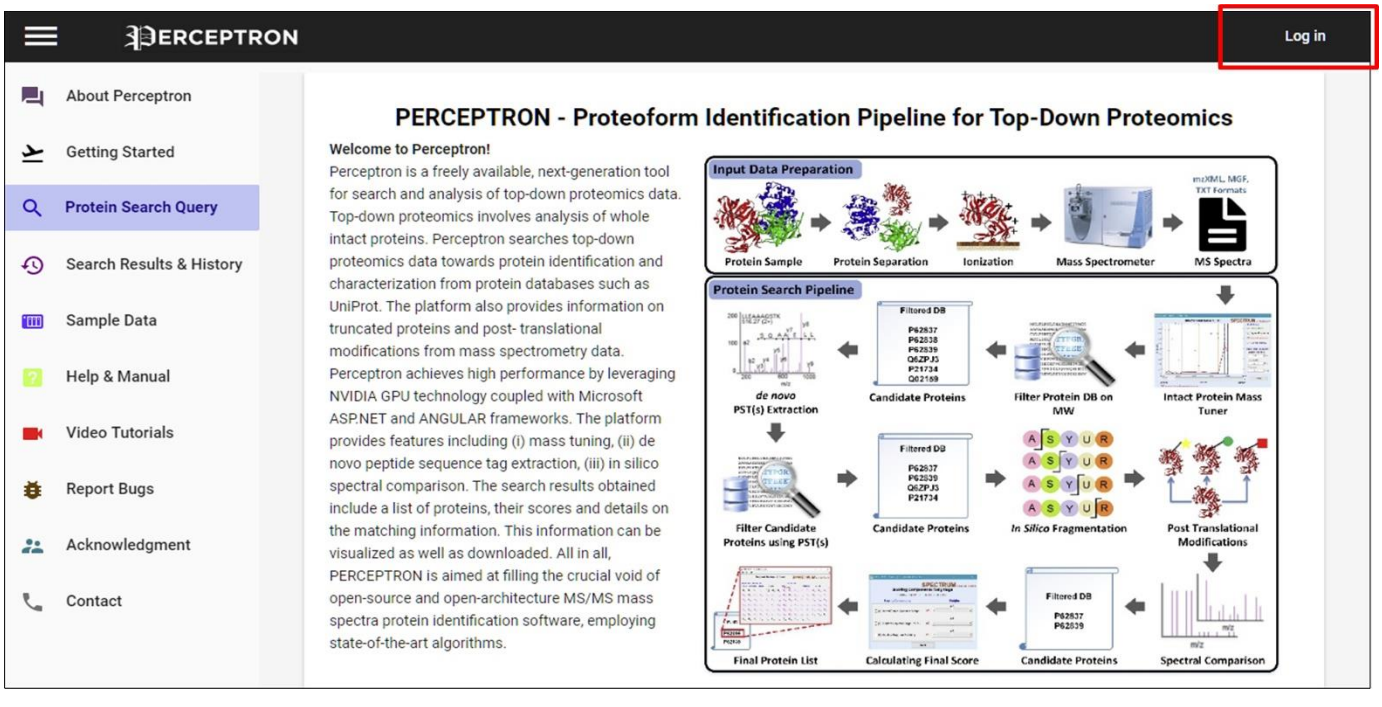

Figure 1. PERCEPTRON Homepage and Log in button

Click the ‘log in’ button on the top right. A window will appear. Enter user credentials to proceed.

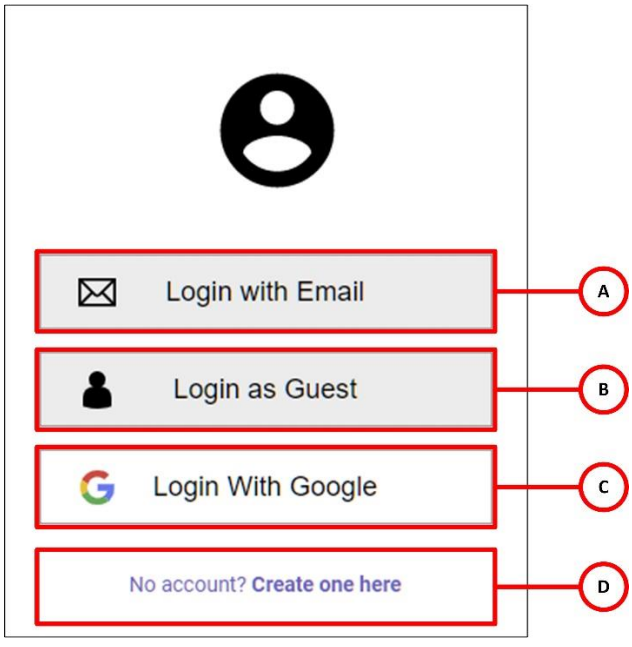

Figure 2. PERCEPTRON Login options

A. User can login with an existing email account
B. Enables user to login as a guest, perform search in PERCEPTRON. Note that results will not be saved in this case.
C. User can also use their Google account to login.
D. Otherwise, users can create a new local account in order to login.

7 GUI description:

This chapter presents the interface overview for user facilitation.

7.1 Window 1: PERCEPTRON Tool for Top-down Proteomics

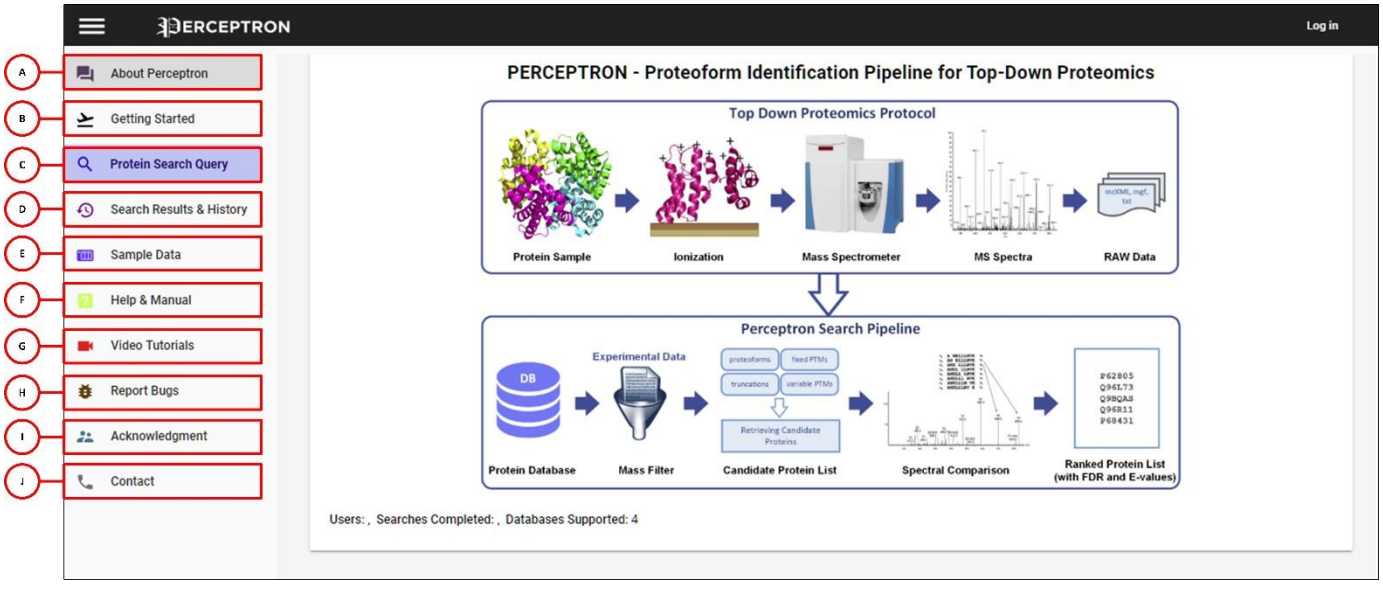

Figure 3. PERCEPTRON - Overview of User Interface

A. About PERCEPTRON: A next-generation top-down proteoform search and identification platform

B. Getting Started: Quick guide to proteoform search and identification using PERCEPTRON

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 PERCEPTRON

F. Help & Manual: Get assistance with using PERCEPTRON and download manual

G. Video Tutorials: View step-by-step video tutorials demonstrating usage of PERCEPTRON

H. Report Bugs: Report problems and issues here

I. Acknowledgement: PERCEPTRON 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:

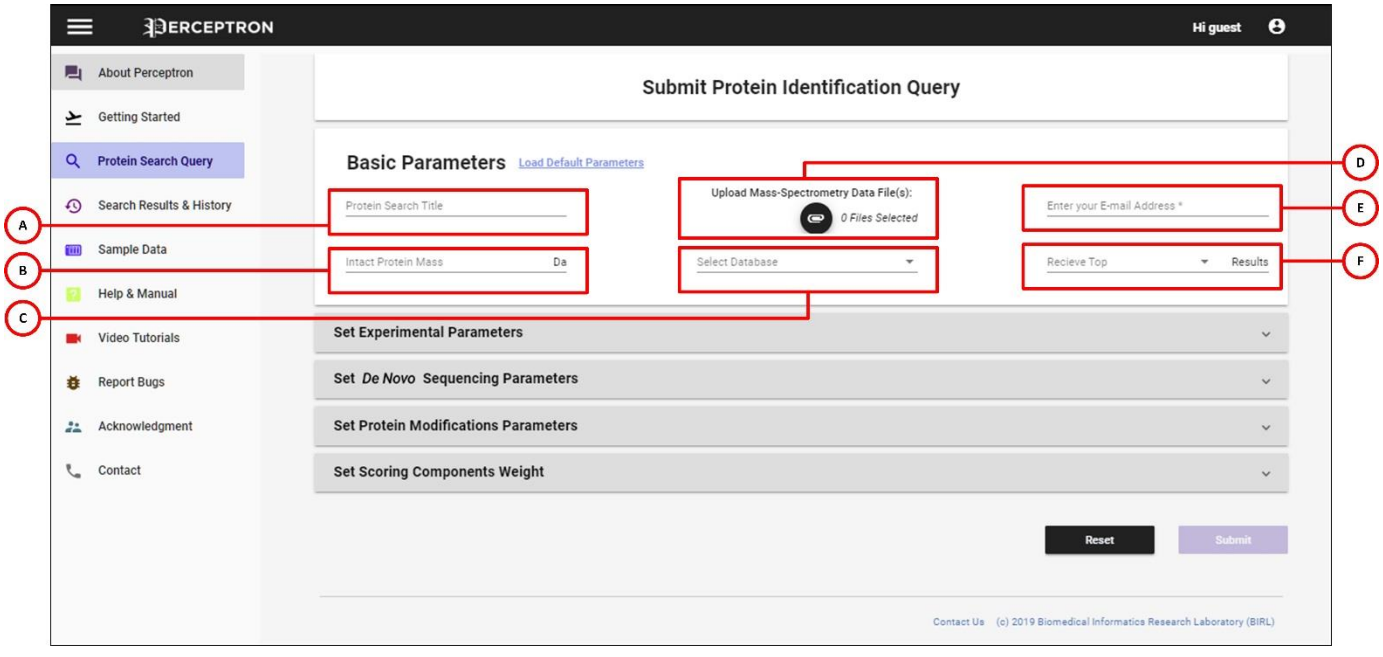

Figure 4. PERCEPTRON - Overview of Basic Parameters

A. In order to start protein search, user must enter the 'Protein Search Title' (for example: "MyProject")
B. Enter intact protein mass (MS1)
C. Select protein database
D. Browse and upload experimental data (.mzXML/ .MGF/ .txt) for Single mode; Peak-list files for Batch mode
E. Enter the email address at which the link for downloading results will be sent. Note that users are notified about the availability of search results upon completion of the job.
F. Provide the number of candidate protein hits to be received in results

Set Experimental Parameters:

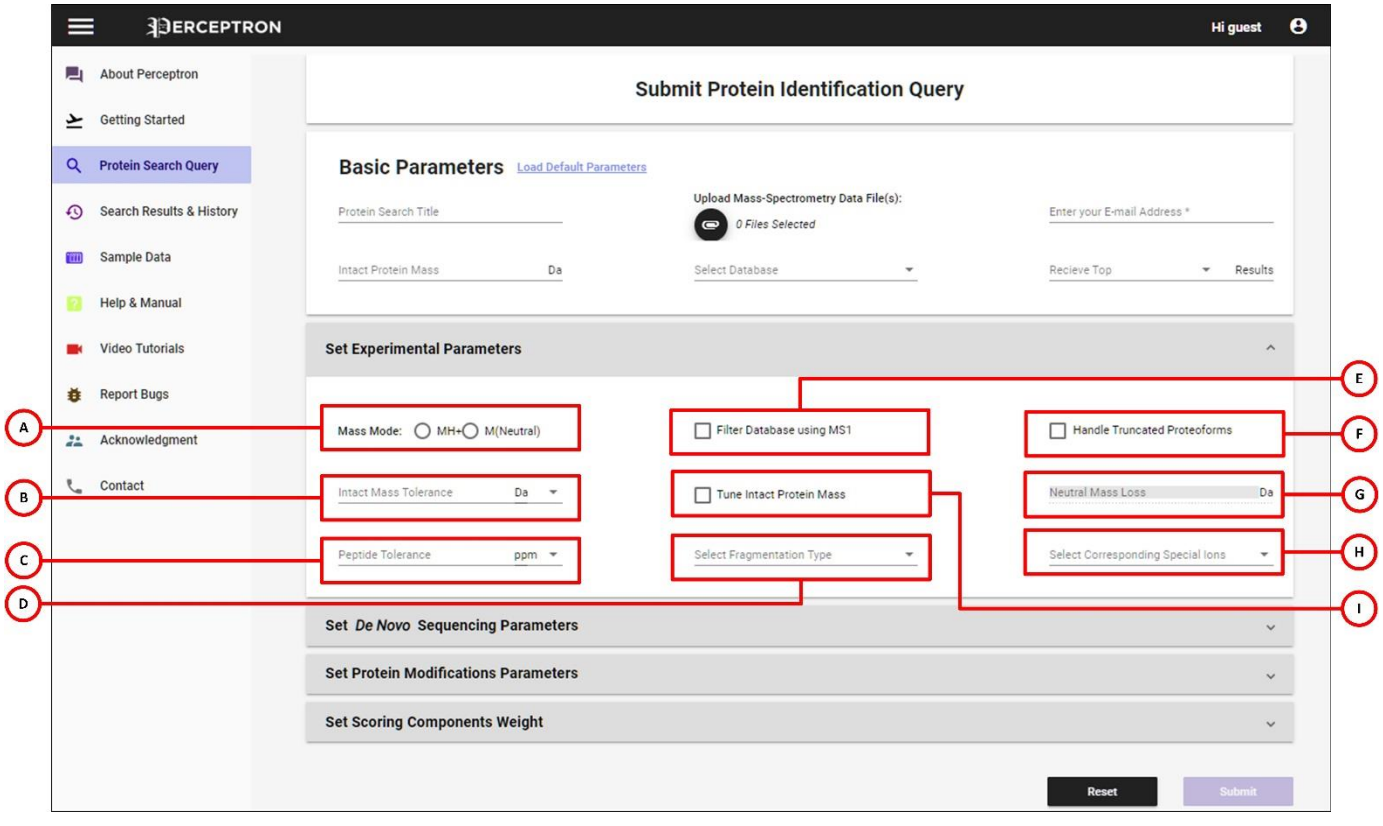

Figure 5. PERCEPTRON - Overview of Experimental Parameters

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:

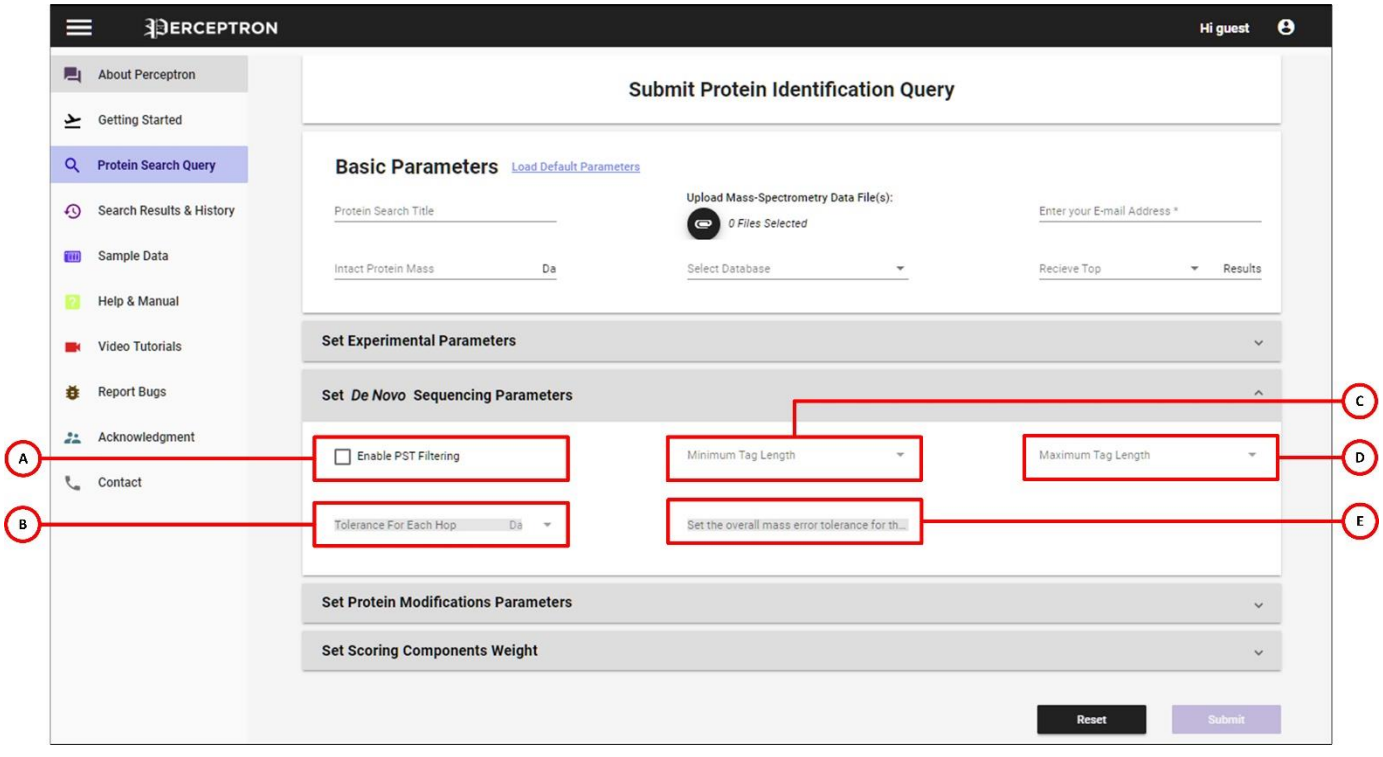

Figure 6. PERCEPTRON - Overview of De Novo Sequencing Parameters

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:

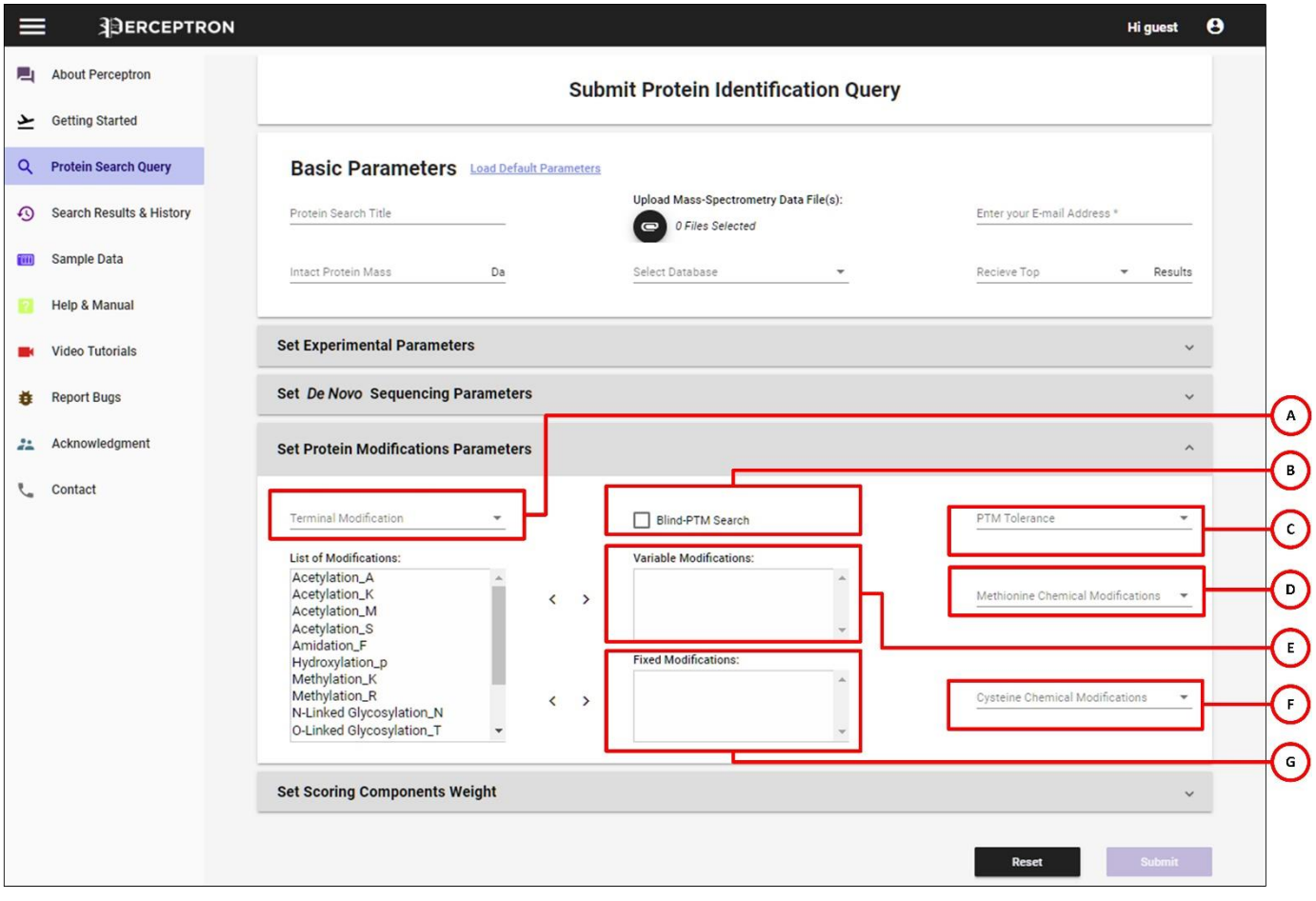

Figure 7. PERCEPTRON - Overview of Protein Modifications Parameters

A. Allows the user to select specified terminal modifications. PERCEPTRON handles four cases: 1) None – No modification, 2) NME – N terminal methionine excision, 3) NME_ACETYLATION – N terminal acetylation with initiator methionine removed, and 4) M_ACETYLATION – N terminal methionine acetylation
B. Select whether to perform Blind-PTM search and unknown modifications
C. Set the tolerance value for Post Translational Modification (PTM) site selection
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:

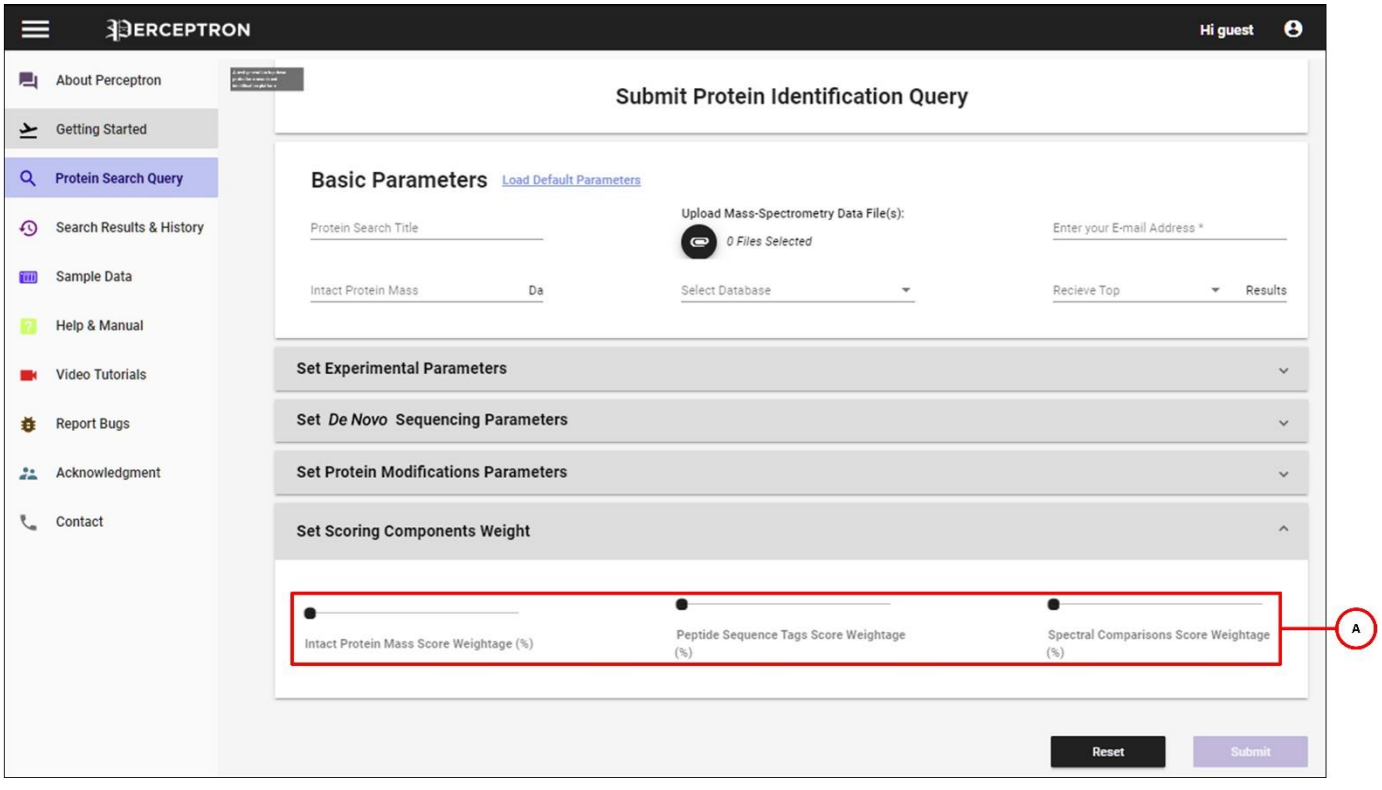

Figure 8. PERCEPTRON - Overview of Scoring Component Weight

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

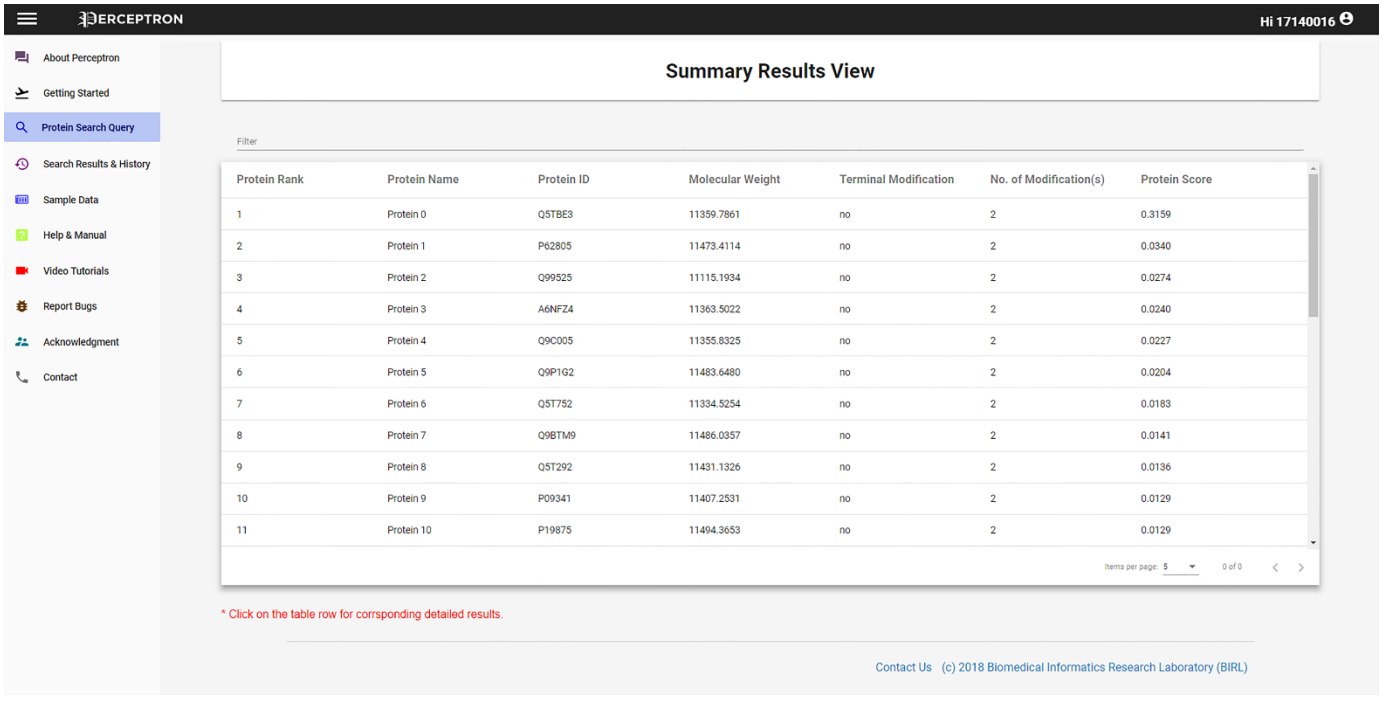

Figure 9. Summary Results window showing candidate proteins

Proteins found after the search are reported along with their protein ID, molecular weight and score. Users can 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.

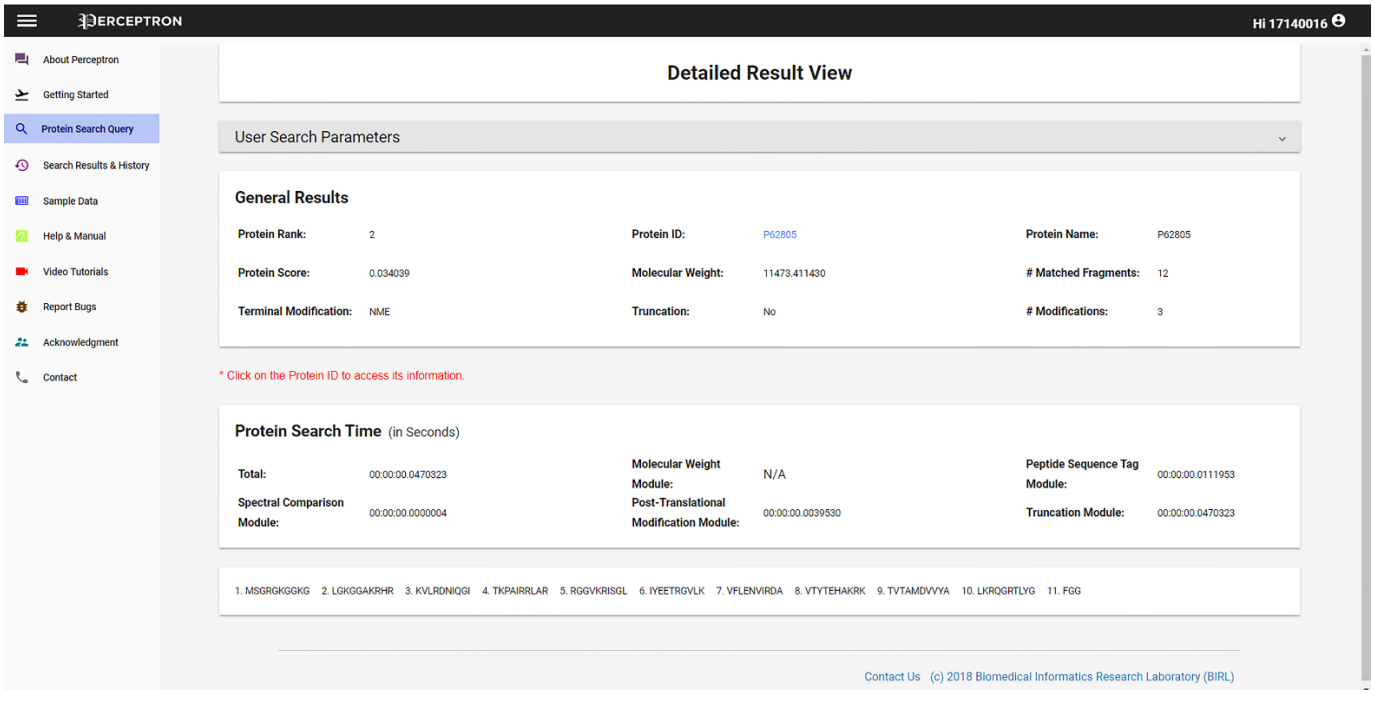

Figure 10. Detailed Results window showing candidate proteins

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-Convert[5].

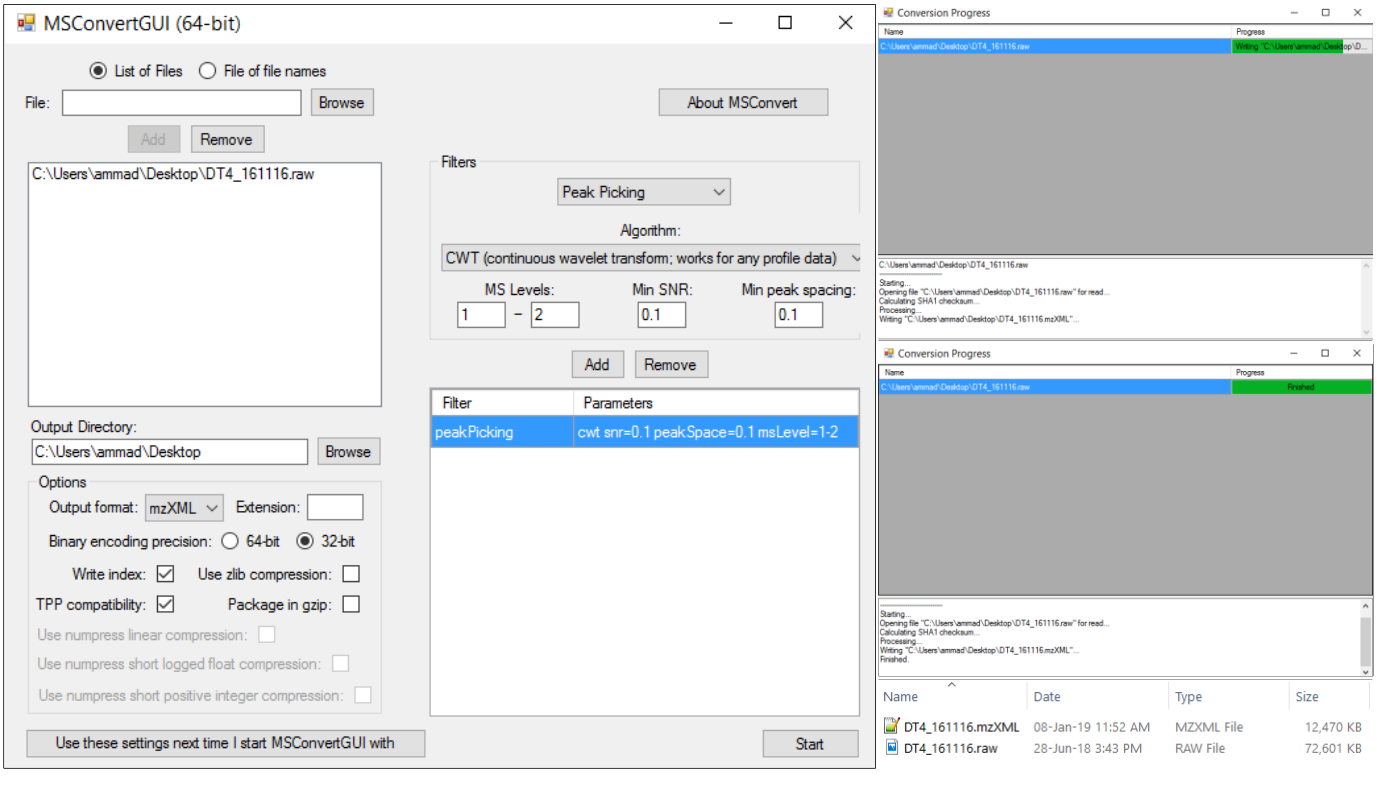

Figure 11. Conversion of raw to mzXML

8.1.2 Raw to mzML File Format Conversion

Raw data files can be converted to mzML file format by using MS-Convert[5].

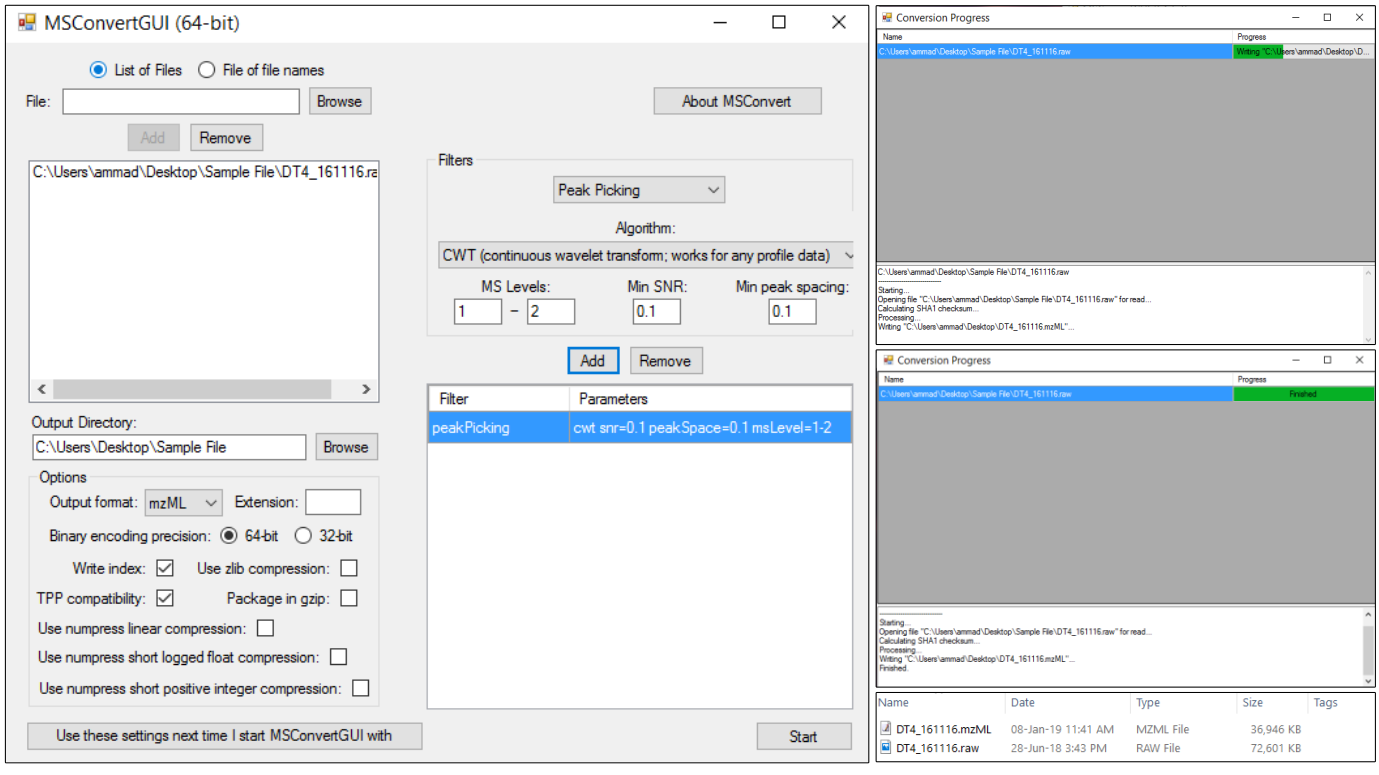

Figure 12. Conversion of raw to mzML

8.1.3 MzXML to MGF File Format Conversion

User can convert mzXML files to MGF using MS-Decov[6].

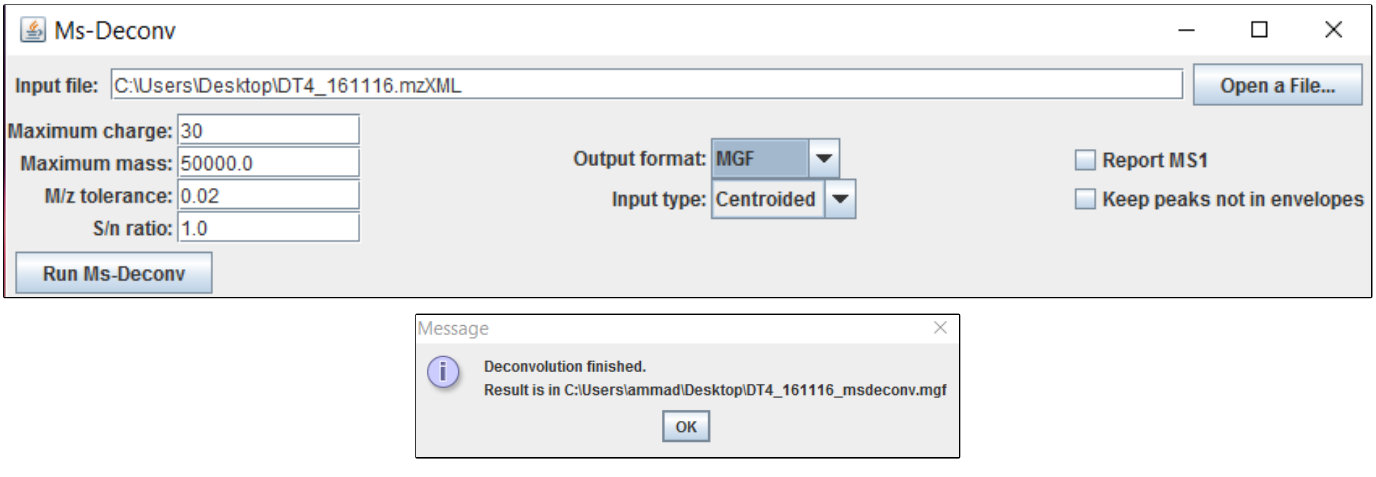

Figure 13. Conversion of mzXML to MGF

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 can employ from the following set of search parameters including: (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.

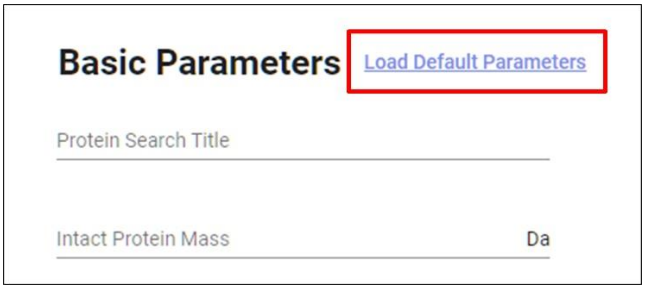

Figure 14. Load Default Parameters

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.

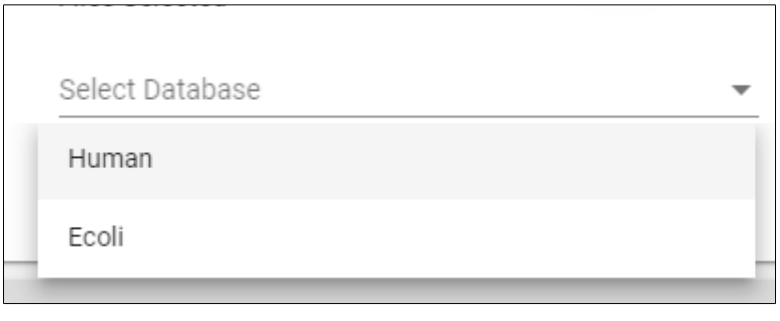

Figure 15. Selecting Protein Database

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.