SPECTRUM *V 1.0.0.0*

A MATLAB Toolbox for Proteoform Identification from Top-Down Proteomics Data

USER MANUAL

Biomedical Informatics Research Laboratory, Department of Biology

Lahore University of Management Sciences

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

1 Table of Contents

2 Table of Figures

3 Introduction to SPECTRUM

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

3.1. About SPECTRUM

SPECTRUM, a top-down proteomics toolbox based on the popular mathematical computing platform MATLAB, is presented in this work. SPECTRUM provides a richly featured environment for searching and identifying proteoforms from top down proteomics data obtained from high resolution mass spectrometers. The toolbox stands to fill the critical gaps in availability of open source platforms for TDP data analysis as well as the MATLAB® Bioinformatics Toolbox which currently only provides limited spectral data analysis features. The proposed toolbox can be employed by computational proteomics instructors in their educational and training endeavors. SPECTRUM comprises of a front-end GUI and the back-end search "engine'.

3.2. Features

The salient features of the toolbox are summarized below:

- **Graphical User Interface** - A set of rich and intuitive graphical user interfaces 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 toolbox. 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 Processing Mode Search** - Towards an automated batch processing of multiple spectral data files (e.g. peak-lists, MGF and mzXML), a batch processing mode has also been implemented. 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

SPECTRUM requires a minimum 500 MB of disk space and 2 GB RAM for optimal functioning.

4.2. Software

SPECTRUM requires “Windows 7, Windows 8, Windows 10, Windows Server 2012 R2 or Windows Server 2016” and “MATLAB R2015a MATLAB R2015a, MATLAB R2015b, MATLAB R2016a, MATLAB R2016b or MATLAB R2017a” along with MATLAB Bioinformatics Toolbox for functioning.

4.3. Testing

All testing is based upon the following Operating System Releases:

- Windows 7, 64 bit
- Windows 10, 64 bit
- Windows Server 2012 R2, 64 bit
- Windows Server 2016, 64 bit

All testing is based upon the following MATLAB versions:

- MATLAB R2015a
- MATLAB R2015b
- MATLAB R2016a
- MATLAB R2016b
- MATLAB R2017a

.

5 Video Tutorials

User can watch video tutorials to guide in: (i) downloading and execution of SPECTRUM, (ii) search in single and batch modes, (iii) downloading protein databases from UniProt and their incorporation in SPECTRUM. The videos are available as a playlist at: https://www.youtube.com/playlist?list=PLaNVq-kFOn0YH6DpEMlXuxGwz8y7abZfQ

6 Getting Started with SPECTRUM Toolbox

SPECTRUM source code, manual, samples and issues database is freely available (under the MIT open license) at (https://github.com/BIRL/SPECTRUM)

To download, click on 'Download ZIP'.

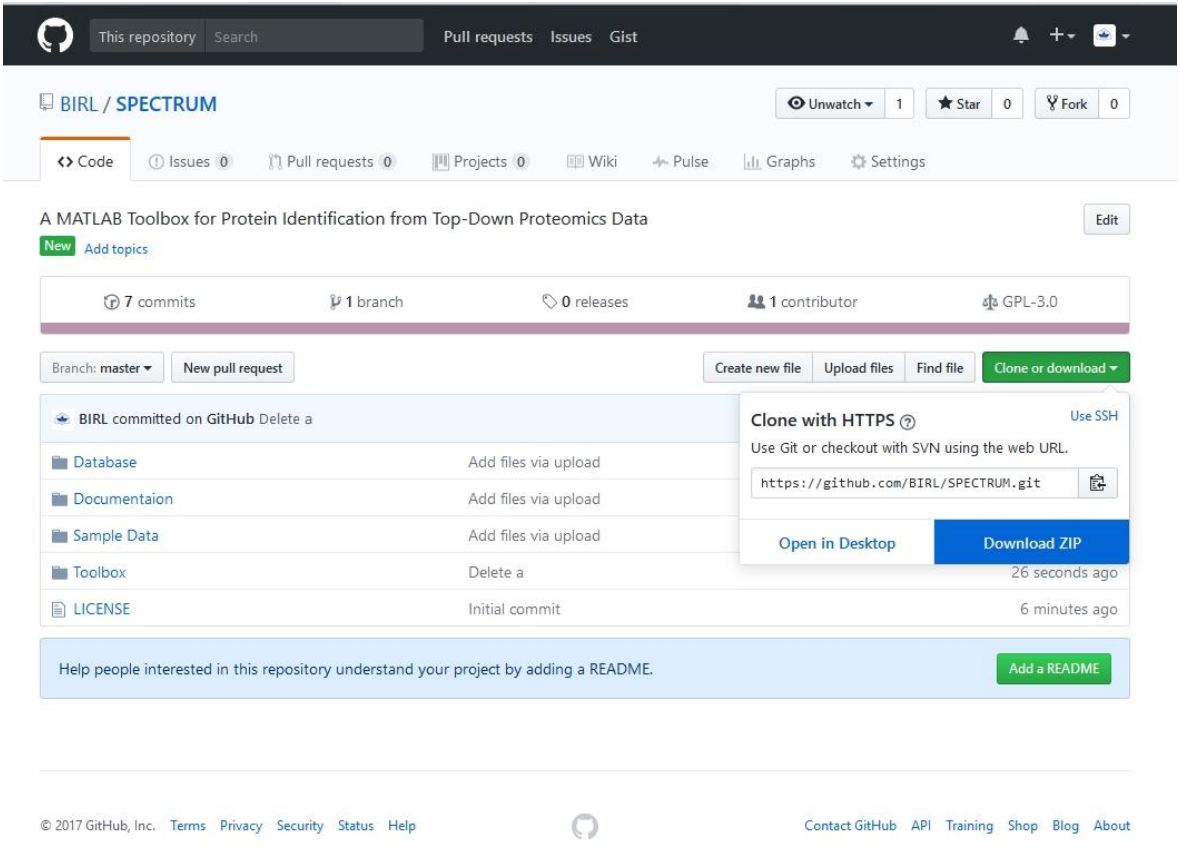

Figure 1. Downloading SPECTRUM

Once downloaded, unzip/extract the files from the folder.

7 GUI description:

This chapter presents the interface overview for user facilitation.

7.1 Window1: SPECTRUM MATLAB Toolbox for Top-down Proteomics

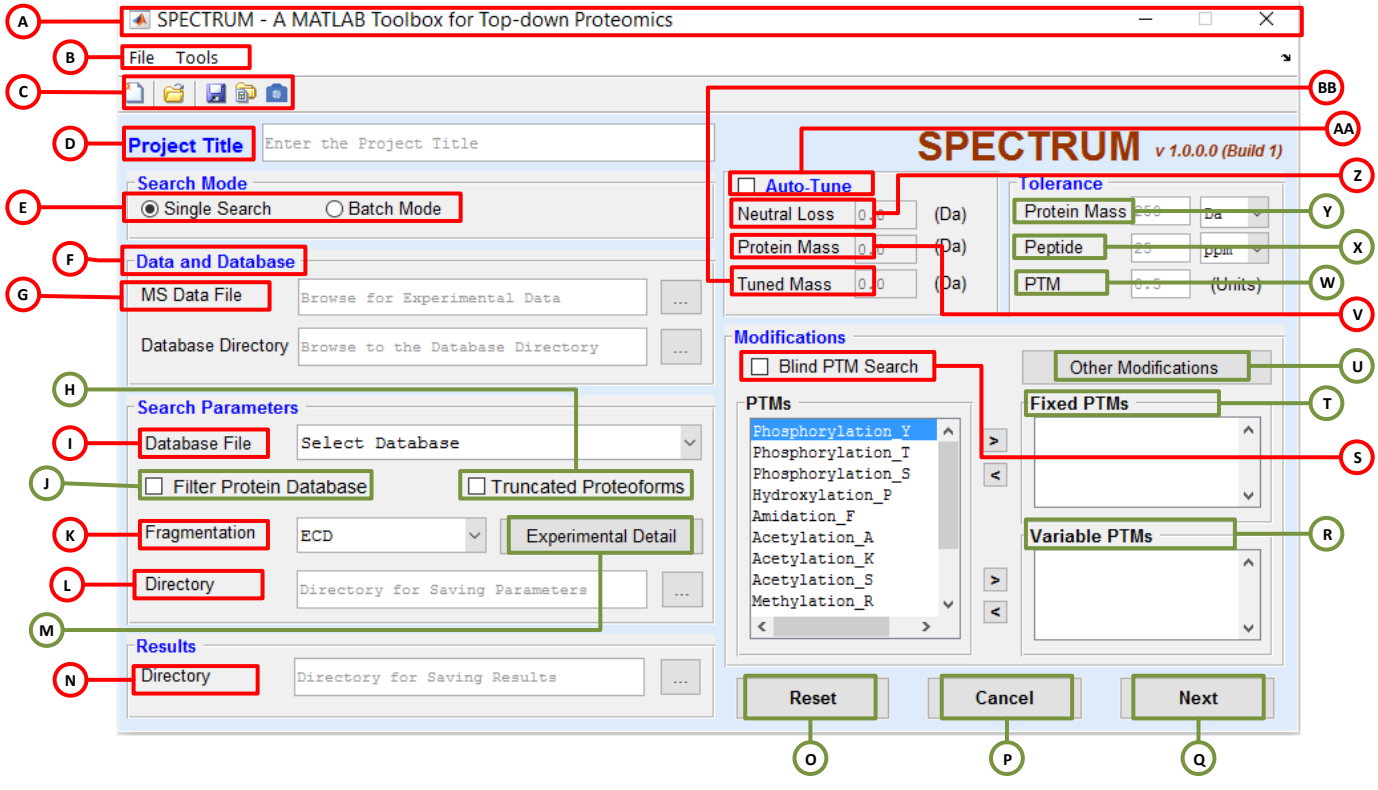

Figure 2. SPECTRUM - Overview of user interface

A. At the top of the SPECTRUM interface is the 'Title bar', which displays the name of the program and the title of the current document.

B. Just below the title bar, there is a 'Menu bar' which lists the heading for each drop-down menu. It allows the user to perform all basic functions (reset, load, save or browse the default/saved parameters). Commands are grouped under each of these menu headings according to function.

C. Under the Menu bar, 'Toolbar' contains shortcuts to some of the most frequently used commands from the menu bar.

D. In order to start a project, user has to enter the 'Project title' first (for My project)

E. Select the search mode according to your requirement (i.e. single search or batch mode). Batch mode option will lead to drop down menu of 'file type'.

F. Browse and upload experimental data (Peak-list file for Single mode; .mzxml/ .mgf/.txt file for Batch mode)

G. Browse and select the protein database

H. User can allow for truncated proteoforms

I. Select Database from the drop down menu

J. User can filter database by checking the option -'Filter DB'

K. Select the 'Fragmentation technique' from drop down menu.

L. User can select a file to store the selected parameters

M. 'Handle Special Ions' option will lead to a new window (Fig. 6) where user can select the fragmentation ions option (i.e. a', b', y', z", a*, b*,y*, z' ions)

N. User can select a file to store the results

O. User can clear the already uploaded data by clicking on 'Reset' option

P. Click on 'Cancel' option to close the window

Q. After filing all the requirements, user can proceed further by clicking on 'Next' option below.

R. User can opt for required Variable 'Post translation Modifications' from the list of modifications

S. Allow user to perform Blind-PTM search and find unknown modifications

T. Similarly, various 'Fixed Modifications' are also selected from the list

U. Allow user to select terminal modifications and Cystine and Methionine chemical modifications

V. After uploading the MS input file, user can see the 'Protein mass' in the box

W. Set the tolerance value for Post Translational Modification (PTM)

X. Set the tolerance value for Peptide

Y. Set the tolerance value for Protein Mass

Z. Provide the value of Neutral loss, if any

AA. Check the option 'Auto-tune' for tuning protein mass and proceed to the next step

BB. Tuned mass can be seen when user proceed to the next step

7.2 Window 2: Select Fragmentation Ions

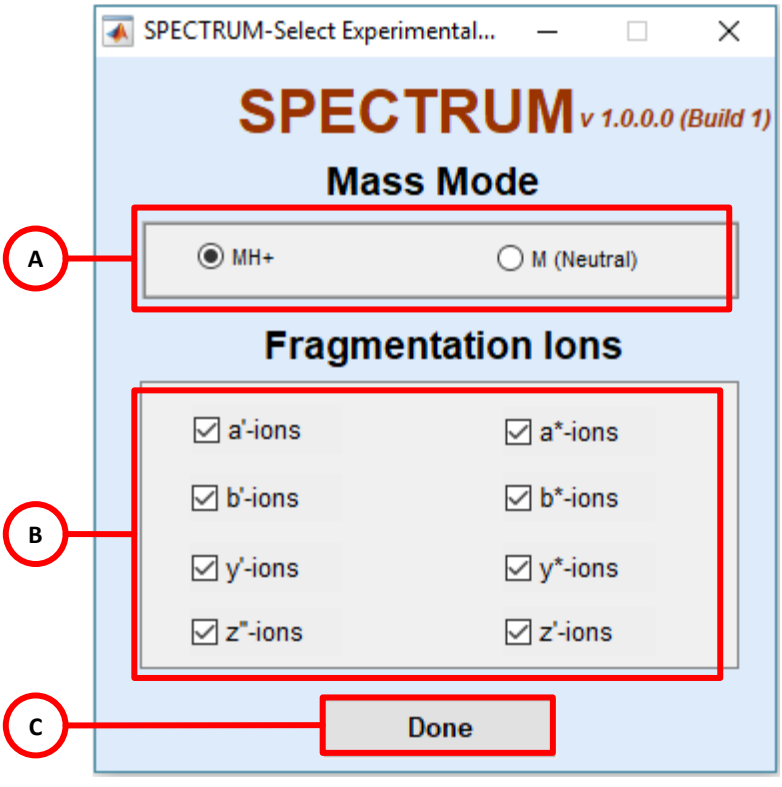

Figure 3. Selecting fragmentation ions

A. Select Mass Mode. MS data can only be provided in either m/z form with z = 1 and neutral masses.

B. 'Handle Special Ions' in the previous window, option leads to a new window where user can select the fragmentation ions option (i.e. a', b', y', z", a*, b*, y*, z' ions)

C. Click on 'Done' to record the selection.

7.3 Window 3: Intact Protein Mass Tuner

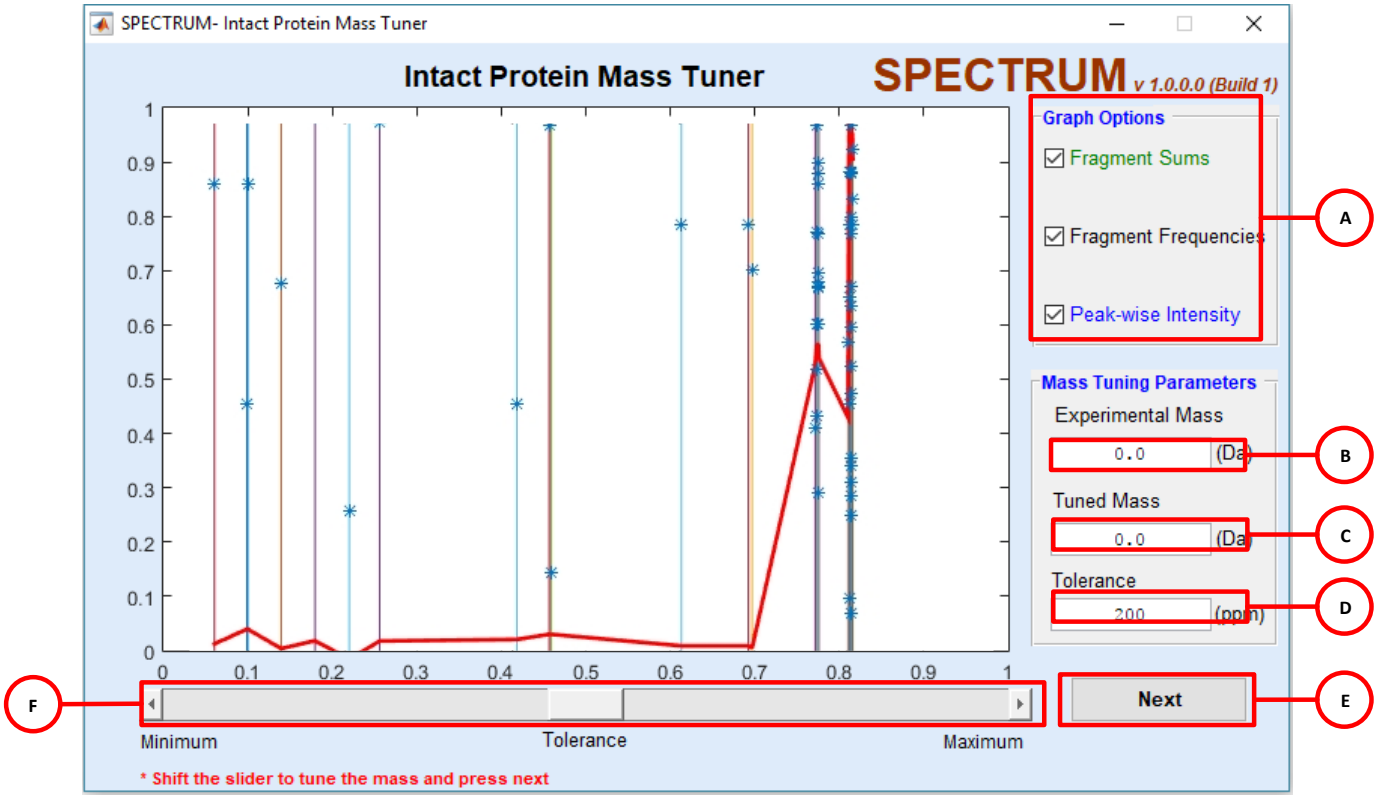

Figure 4. Intact Protein Mass Tuner

A. Graph between 'Frequency' on Y-axis and 'Molecular weight' on X-axis is shown when a used selected 'Auto-tune' option from the previous window. Graph represents Fragment Sums, Fragment Frequencies and Peak-wise Intensity displayed by green, black, red and blue colors respectively. User can select from the list of Attributes to be represented in the graph.

B. B. 'Experimental mass' shows the mass of respective protein

C. 'Tuned mass' shows represents more accurate and precise protein mass

D. 'Tolerance' shows the value you set by shifting the slider

E. User can tune the mass by shifting the slider below, to left or right and press next

F. To proceed further, press 'Next'

7.4 Window 4: Peptide Sequence Tags (PSTs)

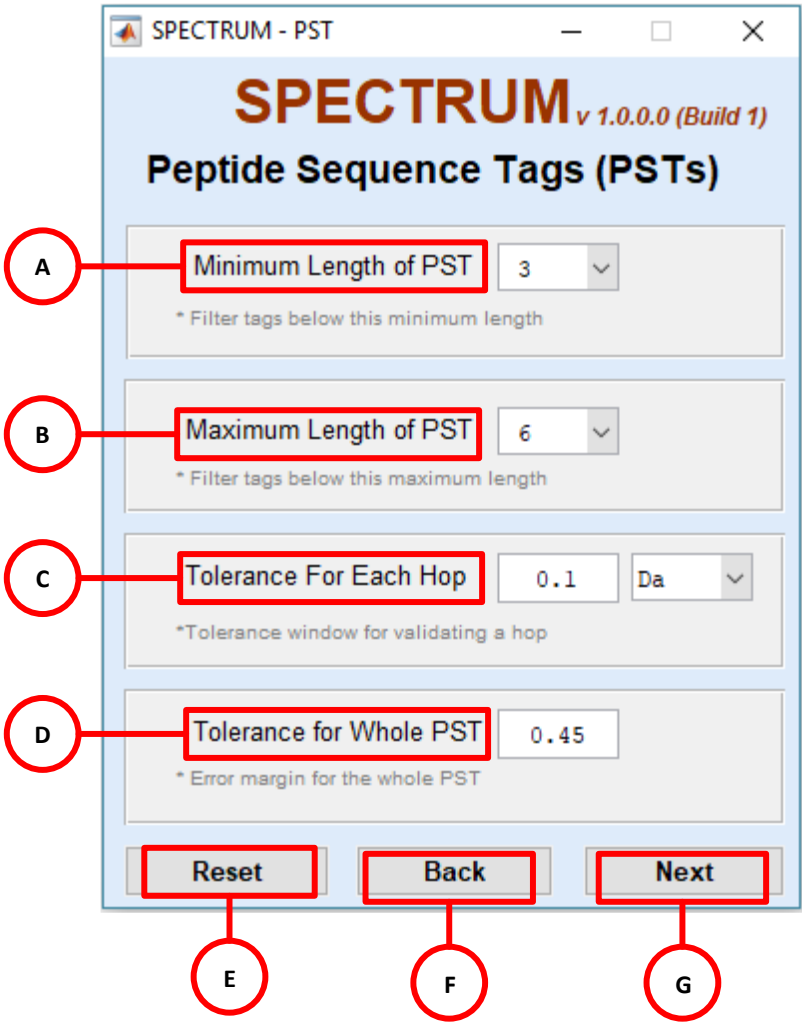

Figure 5. Protein Sequence Tags (PSTs)

A. Tags will be filtered below the minimum length of PST selected from the drop down menu by the user
B. Tags will be filtered below the maximum length of PST selected from the drop down menu by the user
C. Set the 'Tolerance for each Hop'
D. 'Tolerance for Whole PST' shows error margin for the whole PST
E. Click 'Reset' to clear the already selected data
F. Click on 'Cancel' option to close the window
G. Click 'Next' to proceed to the next window

7.5 Window 5: Chemical Modifications

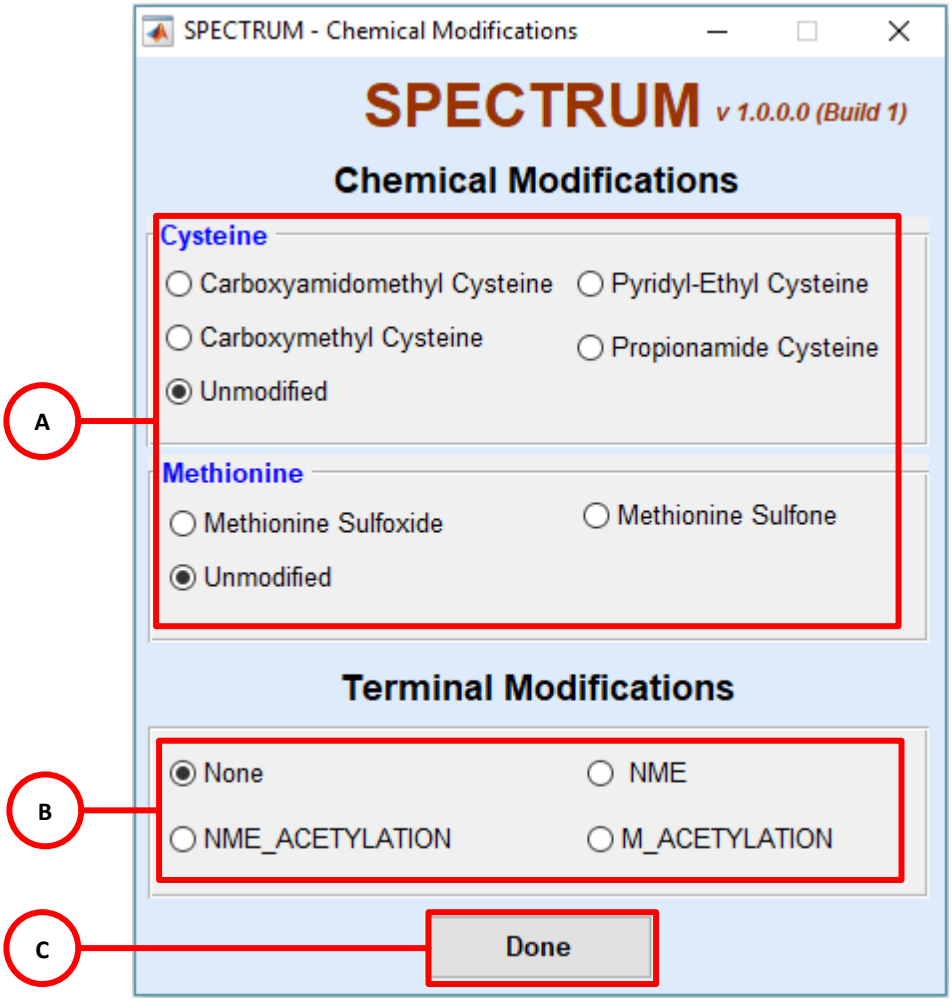

Figure 6. Select instrument specific chemical modification

A. Select instrument specific modification on Cysteine and Methionine.

B. Select Terminal modification. Spectrum 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.

C. To proceed further, press 'Done'

7.6 Window 6: Components Score

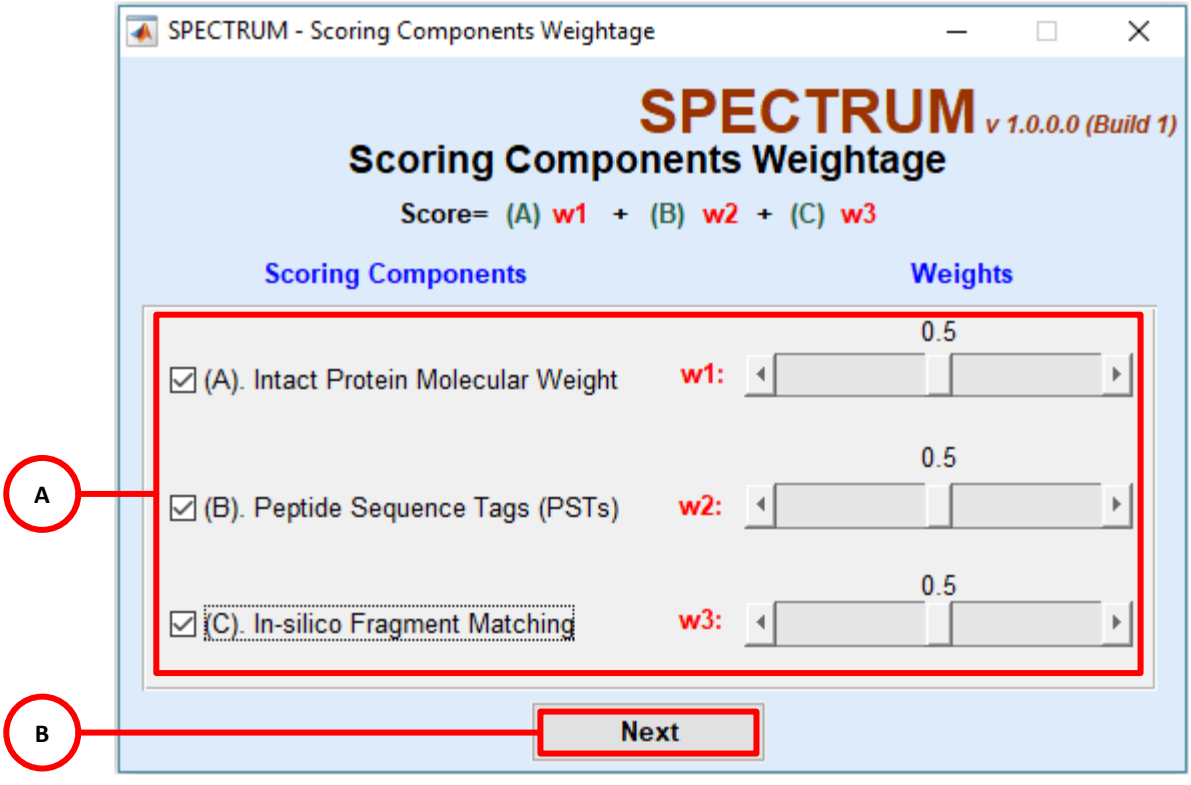

Figure 7. Select weight for each scoring component

A. Check the Scoring Components from the list and set their respective weights by shifting the slider left or right accordingly

B. Click 'Next' to proceed

7.7 Window 7: Candidate proteins

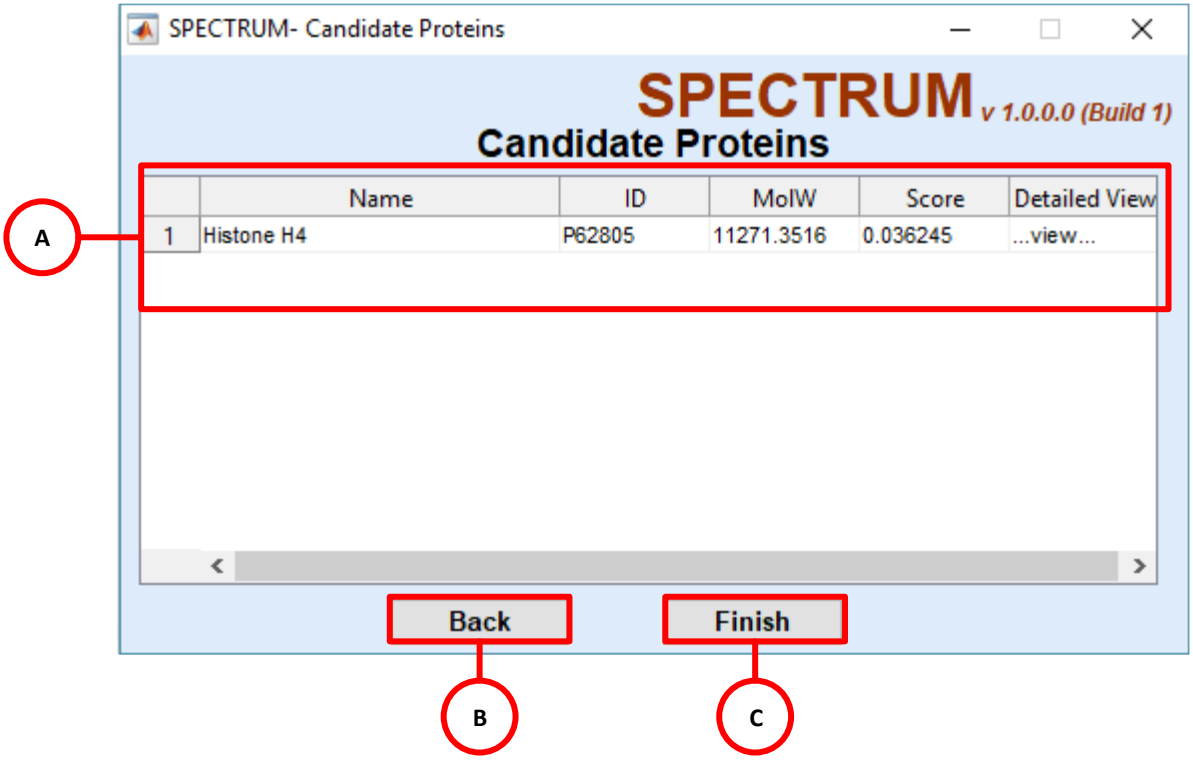

Figure 8. Summary result window showing candidate proteins

A. Name of the resultant protein along with molecular weight, score and detailed view according to the uploaded and selected data is represented under the list of 'Candidate Proteins'.

B. Click 'Back' to revert to the previous window

C. Click 'Finish' to end the process

7.8 Window 8: Detailed Protein View

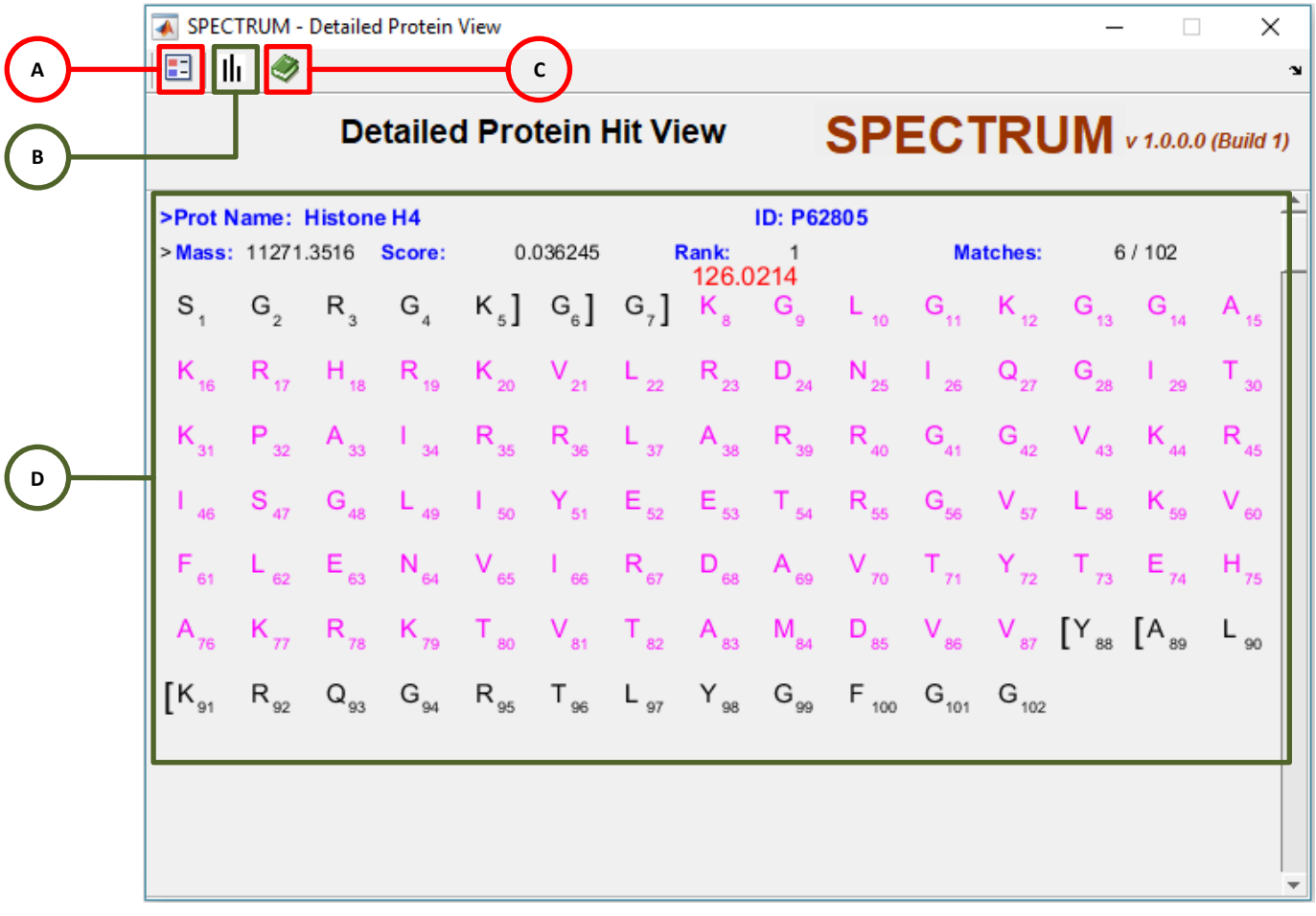

Figure 9. Detailed view of a selected protein

A. Open the window of legends (Figure 10), which might be present in the sequence of candidate proteins.

B. Open the window (Figure 11) containing matched theoretical and experimental N terminal and C terminal fragments.

C. Leads to a window (Figure 12) containing list of full form of Amino acids along with abbreviations and one-letter symbol of protein to facilitate the user.

D. 'Detailed Protein View' window shows proteins name, mass, score, rank, matches and amino acids (along the mass) present in the protein.

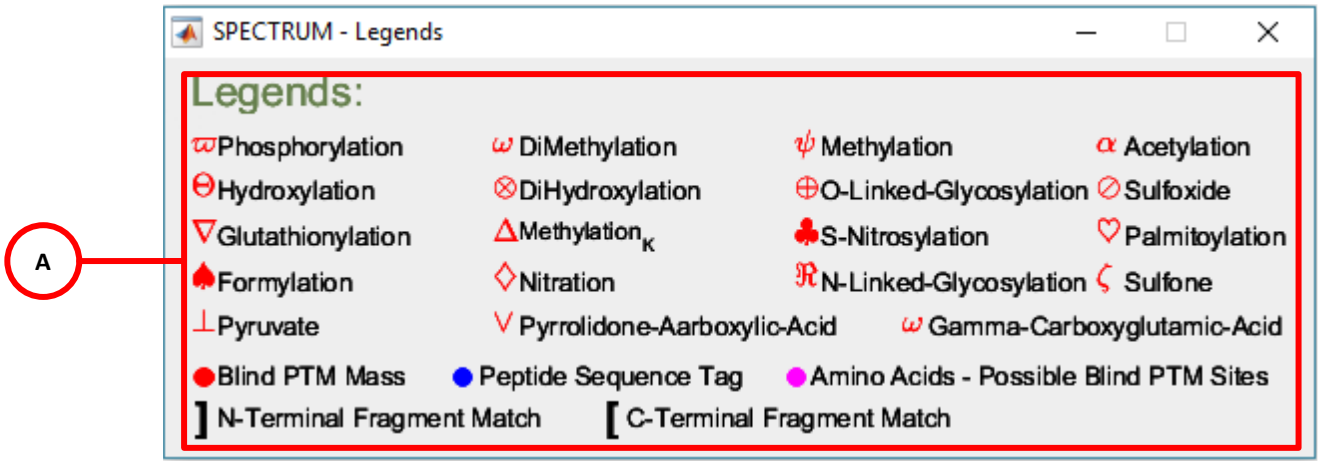

Figure 10. Legend list

A. Legends to help understand the results provided in Figure 10.

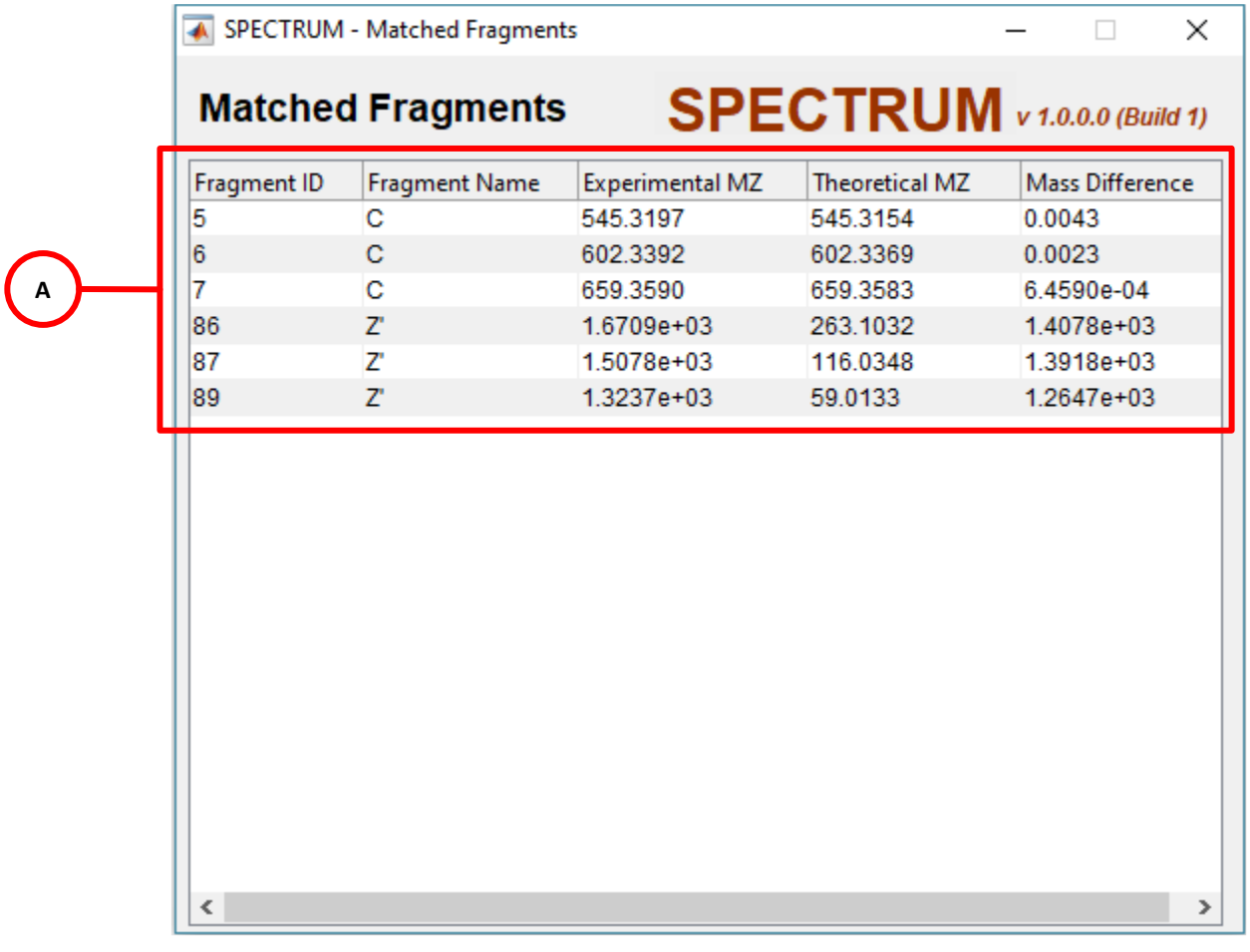

Figure 11. Spectral matches

A. Spectral matches of the protein from Figure 10.

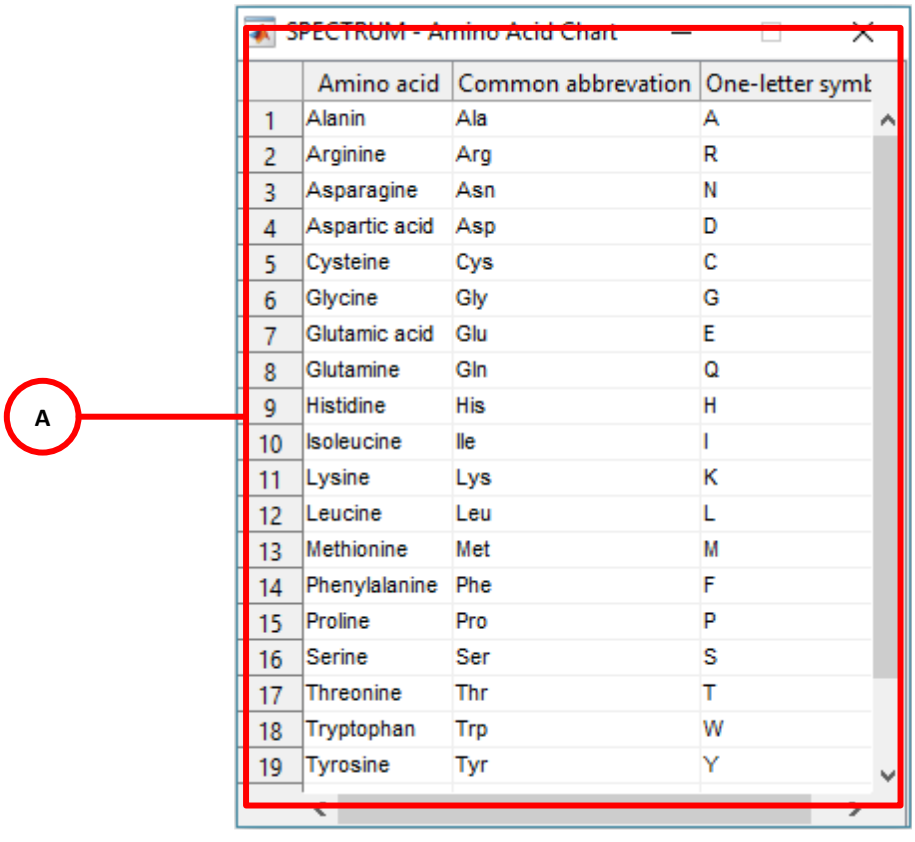

Figure 12. Amino acid chart - Full form and abbreviation of amino acids

A. Amino Acid symbols used in representation of protein sequence in Figure 10.

8 Search

8.1 Parameters:

SPECTRUM works when all parameters are set. Two kind of parameters can be used by the user which includes (i) Default Parameters and (ii) Passed/selected parameters.

How to load default Parameters?

To submit the job using default parameters, go to 'File' tab on the menu bar. Once done, user will be left with different commands. Select 'Load Default Parameters & Settings'.

One shortcut is to simply press the second icon on the toolbar.

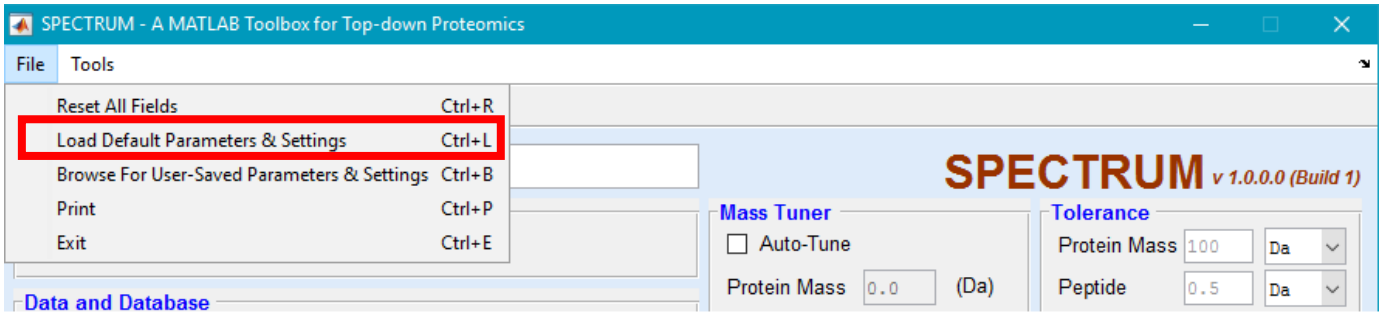

Figure 13. Loading default parameters through menu bar

Figure 14. Loading default parameters through shortcut toolbar

How to save passed/selected Parameters?

Parameters can be passed according to user requirement and can be reused later. For this purpose, user has to save the selected parameters, simply press the highlighted icon below.

Figure 15. Saving user selected parameters through shortcut toolbar

How to load saved parameters?

To browse for the above saved parameters, select the option 'Browse for User-Saved Parameters & Settings' or click on forth icon highlighted below.

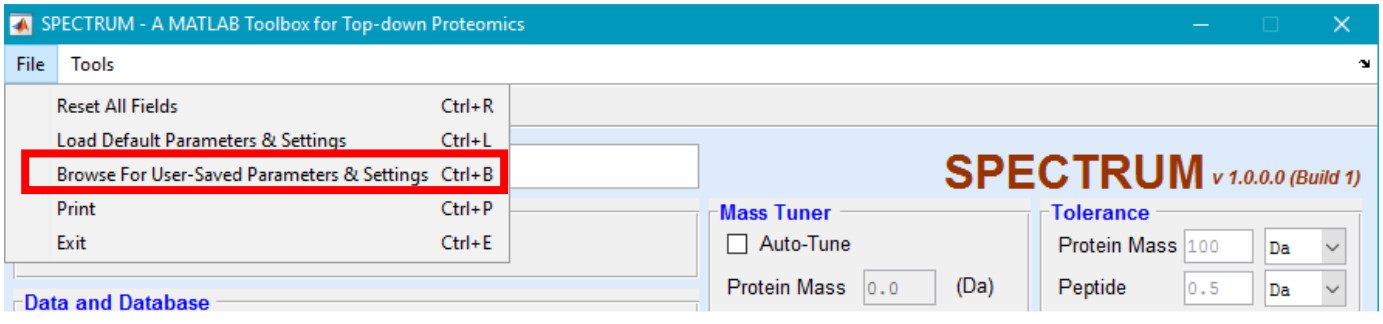

Figure 16. Loading user-saved parameters through menu bar

Figure 17. Loading user-saved parameters through shortcut toolbar

8.2 Databases

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

8.3 Modes

User can select 'Search mode' that best fits the need. SPECTRUM works with two search modes.

(i) Single Search Mode

(ii) Batch Mode

Databases and Peak-list files are uploaded according to the selected mode. Single search mode demands experimental data in .txt file whereas Batch mode supports .txt, .mzXML, .mgf files.

Batch mode takes more processing time as it deals with larger data.

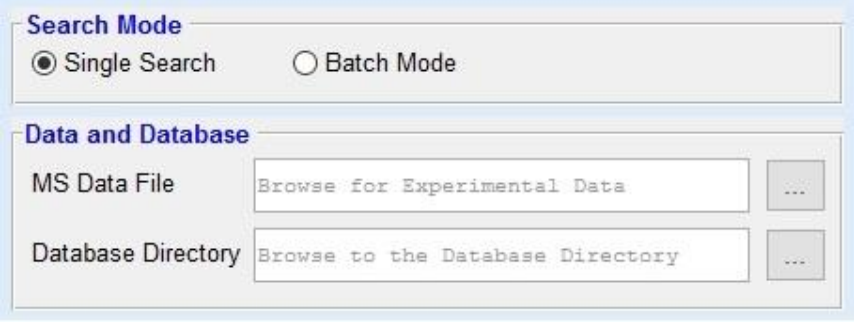

Figure 18. Selecting search mode

The experimental spectra, search parameters and results are automatically stored in the project directory for further processing and visualization.