**Educational Administration: Theory and Practice**
2024, 30(4), 5609-5616
ISSN: 2148-2403
https://kuey.net/

**Research Article**

# Protool: User Frıendly Tool For Hıgh Throughput Mass Spectrometry Data Analysıs

Ashwini Yerlekar[1*], Hemantkumar Tukar[2], Kalyani Pendke[3], Alok Chauhan[4], Yogesh Narekar[5], Ravindra Rasekar[6], Ashish Nanotkar[7]

[1,2,3,4,5,6,7]S.B. Jain Institute of Technology,Management and Research, Nagpur. [1]Email: ashwini.yerlekar@gmail.com, [2]Email: turkar2930@gmail.com, [3]Email: pendke@gmail.com, [4]Email: alokchauhan.work@gmail.com, [5]Email: yogesh.narekar@gmail.com, [6]Email: rasekar.rav@gmail.com, [7]Email: ashishnanotkar5915@gmail.com

**Citation**: Ashwini Yerlekar et al (2024), Protool: User Frıendly Tool For Hıgh Throughput Mass Spectrometry Data Analysıs, *Educational Administration: Theory and Practice*, *30*(4), 5609-5616
Doi: 10.53555/kuey.v30i4.2257

**ARTICLE INFO**

**ABSTRACT**

Because of its intricate nature, protein structure prediction has become more significant in the field of life sciences. Studying the function of proteins in certain environments and the molecular relationships found in living systems both need an understanding of protein-protein interactions. A single organism's protein expression can fluctuate dramatically within its body, between stages of life, and in response to distinct environmental factors. huge amounts of data are evolved using proteomics procedures, which make it possible to examine a huge number of proteins in order to gain a more thorough understanding of biological systems. To evaluate the actual associations between proteins in living things, huge scope proteomics advances are hence required. The m(mass) to e(charge) proportion of particles is estimated utilizing a method utilized in mass spectrometry.
It's a developing method for characterizing proteins. Additional LC detectors can enhance the sensitivity and specificity of a mass spectrometer. Fluid chromatography is a division strategy that can be used to separate a wide range of organic substances, including peptides, proteins, and tiny molecular metabolites, as opposed to gas chromatography. Data storage, Mass Spectrum visualization, Data Normalization, Peak Detection, and Alignment are the different processes in High Throughput Mass Spectrometry information examination. The goal of this project is to create software for mass spectrometry data analysis. The proteome discovery pipeline has been utilized by the program to predict proteins. There are many tools at one's disposal, but for those working in the biological sciences, a graphical user interface can facilitate quick access to data

**Keywords:** Protein-Protein interaction, Mass Spectrometry, mass -to-charge ratio, Complex, LC (Liquid chromatography) detectors.

## 1 Introduction

Bioinformatics is a modern technique to protein structure prediction and sequence analysis. Creating tools for data analysis in addition to storing and retrieving data is the main challenge facing bioinformatics.Additionally, a number of technologies are available for recognizing, visualizing, and characterizing proteins and their structures. he mass-to-charge proportion of particles can be estimated quantitatively utilizing a mass spectrometer. The two fundamental methods for ionizing entire proteins are electrospray ionization (ESI), which converts test proteins into particles, and framework helped laser desorption/ionization (MALDI), which ionizes sample proteins using a laser prior to injection. Understanding the principles and outcomes of the mass spectrometer is crucial because it is a vital instrument in proteomics. A mass spectrometer may conduct all of the mass spectrometry's procedures on a single sample. They prefer the modularity method, which consists of creating the spectrum, breaking the ion, and figuring out the range for a single ion. It can therefore disassemble complex molecules piece by piece till their structure is known.
Several kinds of data are generated by mass spectrometry. Graphic representation, or the mass spectrum, is the most widely used visualization tool.A mass chromatogram works well for displaying several kinds of mass spectrometry data. Total ion current (TIC), selected reaction monitoring (SRM), and selected ion monitoring (SIM) are a few of the several chromatogram kinds. One helpful instrument for seeing various kinds of mass

spectrometry information is a three-layered map. Mass-to-charge (m/z) is shown on the x-hub, power is shown on the y-hub, and an extra boundary is shown on the z-pivot [2].
Database search and de novo search are the two main categories into which peptide identification techniques are divided. The main hunt is finished against an information base that has each amino corrosive, while the second deduces peptide groupings without utilizing hereditary information. These days, database searches are thought to yield higher-quality results for the majority of applications and are more dependable. The de novo search could become more and more interesting as instrument precision advances [2].
A conventional mass spectrometer comprises of three sections: a mass analyzer, an indicator, and a particle source. The particle source produces particles from the example. The mass analyzer isolates particles with fluctuating mass-to-charge proportions, and the locator distinguishes these one of a kind particles. The mass spectrum is ultimately produced once all the data has been gathered. Fig. 1[2] displays a system display for a mass spectrometer.

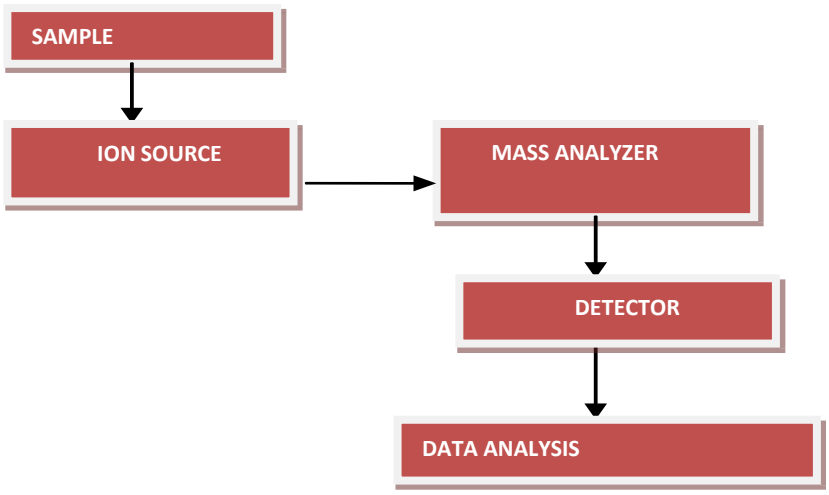


**Fig. 1.** System Dısplay of Mass Spectrometer

## 2 Lıterature Revıew

Numerous approaches have been put out to achieve the aforementioned goal for data analysis. This covers the subsequent methods: The process for finding proteins is described in [1]. Proteomics techniques facilitate the analysis of large protein datasets, providing a deeper understanding of biological processes. The technology of liquid chromatography-mass spectrometry is widely used in high-throughput proteomics to acquire data. With the help of this system, proteomic researchers may easily analyze data on several proteomics technology platforms and handle a wide variety of file types with an intuitive web-based data analysis package.
It may be developed further because of its expandability and versatility. The core techniques for proteomics are mass spectrometry-based protein identification methods, which are crucial for combining and quantitatively evaluating the large datasets produced by biological investigations, which are usually carried out in duplicates. The search algorithm is a widely used technique for peptide and protein structuring using data from 2-D LC ESI MS2 (two-dimensional liquid chromatography-electrospray ionization tandem mass spectrometry). This paper outlines the proteomic analysis workflow by looking at replicate datasets from a single experimental paradigm and using both parametric and non-parametric statistics to generate a list of known proteins along with their corresponding probabilities and notable changes in protein expression. [5].
The necessity to understand mass spectra necessitates the creation of new peak-labelling procedures for protein and peptide labels to be associated with peaks, as well as methods for accurately labelling ion species. The projected protein properties included in the model, such as mass weight, expected concentration, and amino acid sequence, are routinely calculated using this computational technique. This data is integrated into a new probabilistic score that is presented [6].
Antibodies from the same patient's plasma and red blood cell hemolysate react. employing a proteomics technique to identify those hemolysate. Antigens and then create a database with those antigens can aid in illness disorders' diagnosis, prognosis, and treatment [7]. Modern proteomic instruments allow for comprehensive, high-throughput proteome function detection, identification, and research. Advancements in protein separation and labeling methodologies have broadened the scope of protein identification to include even the rarest proteins. These new methodologies generate a profusion of data, which creates new obstacles for data processing and interpretation. The present state of proteomic techniques and their wide range of applications are described in this review, along with a detailed analysis of their benefits and challenges. Possible ways to get over these technological challenges are also mentioned [11].

## 2 Various Steps Of Mass Spectrometry Data Analysis

A substantial volume of data is generated in proteomics, necessitating the utilization of numerous tools capable of processing Mass Spectrometry data in diverse formats. The primary objective of this approach is to create a user-friendly interface application tailored for individuals in the life sciences field. However, employing R language for data processing can often prove laborious and burdensome, particularly due to the requirement of installing various packages for statistical analysis. Given the large volume of generated data, it is imperative to ensure proper analysis and storage procedures. The tool aims to facilitate easy access to data for life sciences professionals by employing visualization, normalization (central tendency), peak identification (isPeak function), and peak alignment techniques. Subsequently, the processed data will be transferred to readily available tools for further processing.
Upon receiving the MS data, it undergoes a series of sequential steps as outlined below [1]:

### 3.1 Data Storage and Security

On the PDP platform, investigators can save, process, analyze, and visualize MALDI-TOF ,LC-MS and LC-MALDI-TOF datasets. Datasets ranging in size from a few kilobytes to several terabytes are supported by the system. The workflow document that specifies the logical operations allowed on the raw or processed data is adjustable and is used by the pipeline [3]. The scale of biological data sets necessitates the implementation of a storage solution that offers scalable and redundant data storage. The vast amount of data generated by the mass spectrometry equipment necessitates storage on separate storage units.

### 3.2 Data transformation

Helps to standardized proteomics data. A variety of file formats, including.txt,.mgf,.mzXML,.cdf,.D, and.wiff, are available for the data. After the data is initially retrieved, it must either be converted into a particular format or a tool that can handle many data formats must be created. Today, several software programs convert raw instrument data into mzXML or mzData forms. mzXML and mzData are now commonly used formats. But not all data from mass spectrometers can be converted (such as Agilent MSD TOF data). The ability of Protool to handle data formats like.txt,.mzXML, and.mgf is a significant feature.

### 3.3 Spectral visualization

With its many data visualization features, it allows the user to work with data [21]. For both intermediate data and instrument data formats, Protool offers a variety of data visualization features. By selecting the chosen ion chromatogram, for example, the user can access raw instrument data. Researchers may now directly verify data uploaded to the pipeline thanks to this.

### 3.4 Data normalization

Data normalization enables multiple-experiment analysis. Prior to comparing samples, the data must be normalized. Normalization is an effort to filter general fluctuations in peak intensity caused by experimental errors quantitatively. [24]. The data are normalized by using the central tendency approach. Data bias and noise could be the causes of the relative changes in the data. At this point, the baseline error is eliminated. The data was offered in three formats: processed, baseline, and raw. The central tendency normalization process centers peptide abundance ratios around a mean, median, or other predetermined constant to mitigate the impacts of autonomous systematic bias. A few preprocessing processes, such as binning and calculating the median and total of the available data, are part of the normalization technique [29].

### 3.5 Peak alignment

Ideally, the indistinct peptide or metabolite present on a comparative logical instrument will yield a comparative sign. For instance, the support term and sub-nuclear heap of a peptide chose using a LC-MS system should be unsurprising across various models. Nevertheless, due to contrasts in tests, this most likely won't turn out true to form. Top game plan engages experts to find tops from the very molecule that are accessible in different models, out of the enormous quantities of zeniths procured during assessment.
**isPeak** is the function which is utilized for peak alignment and detection. Plotting the peaks after removing the baseline is made easier by the function. Using **isPeak**, peaks can be found after the baseline has been eliminated. First, the moving average of the k nearest neighbors is used to smooth a range. Enhancing peaks and eliminating phony peaks are two benefits of leveling. Due to the need for some short and wide peaks for analysis, the smoothing is not done precisely. Additionally, accuracy in peak location is required.

Mascot performs additional functions like pattern recognition, protein identification, characterization, and quantization using mass spectrometry data. Matrix Science has developed a tool called Mascot. A plain text (ASCII) file containing peak list data is called a Mascot data file. Different instrument data systems' peak list formats correspond exactly to these requirements. The following formats are automatically recognized by Mascot: Bruker (.XML), mzData (.XML).

# 4 Experimental Environment

## 4.1 Experimental Setup

The software required are R software (3.0.0), R (D) COM Interface 3.5-1B2_Noncommercial, rscproxy_2.0-5, Visual Studio 2010, Windows OS7. The necessary hardware is System type: 32- or 64-bit; Processor: Intel® CoreTMi3-2370M 2.4 GHZ or above; Installed memory: 4.00 GB or higher.The interfacing is done between C#.Net and R language, where C#.Net is used as front end to design the GUI and R language as back end for the statistical calculations.

The various steps of interfacing are as follows:

- Creating the Windows Form in Visual 2010.
- Add project reference to the R (D) COM library
- Add library references to form code such as STATCONNECTORSRVLib and StatConnectorCommonLib.
- Setting and Initializing Connection
- Invoking R commands and displaying results

## 4.2 Installation of Bioconductor Packages

Bioconductor offers tools for high-throughput genetic data processing and interpretation. With an open development framework, the statistical programming language R is used by the open-source platform Bioconductor. In the R window, by providing the two instructions:

source ("http://bioconductor.org/biocLite.R")

biocLite ()

The many packages that must be installed are listed below:

### readMzXmlData

All mass spectrometry data that is saved in a path-specified mzXML format is read by the software and added to mass spectrum-class objects that are a component of the MALDIquant package.

### RforProteomics:

An Introduction to R offers a beginner's guide to the language and shows how to use it for statistical analysis and graphical depiction. It developed from the previous Notes on R[28].

### The mzR Package

The mzR program provides a uniform interface to numerous mass spectrometry open formats [26]. You can connect to the raw data file using this small bit of code, which is primarily derived from the openMSfile documentation. For mzXML, mzData, or netCDF files, you can also use the adaptable mzR package. Additionally, mzR serves as a foundational instrument for additional packages such as MSnbase [27]. xcms and Target Search provide a higher degree of abstraction for the data.

### The Process library

There are several functions intended for spectrum processing in the PROcess package. These features make it easier to do operations like removing baseline drift, finding peaks, and matching them to a predetermined list of protobiomarkers.

# 5 Experimental Results

## 5.1 Interfacing

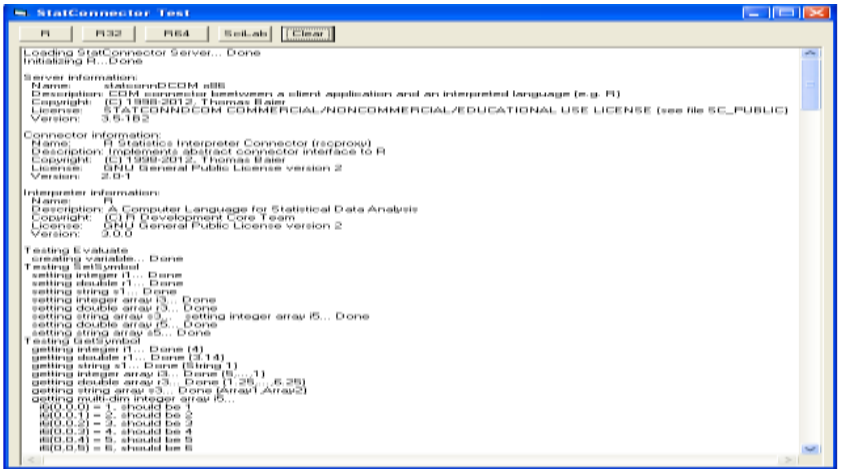


**Fig. 2.** Connection of R with R(D)COM server

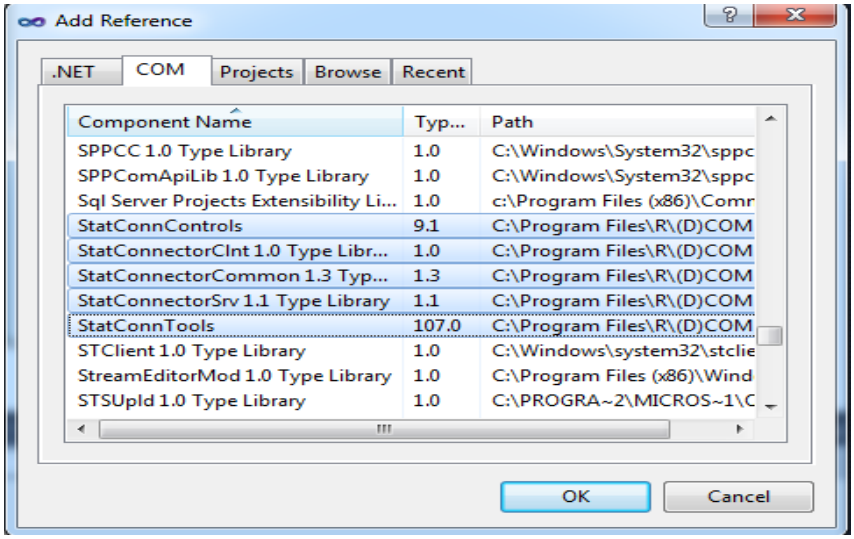


**Fig. 3.** Addition of references for connection of R with C#.Net

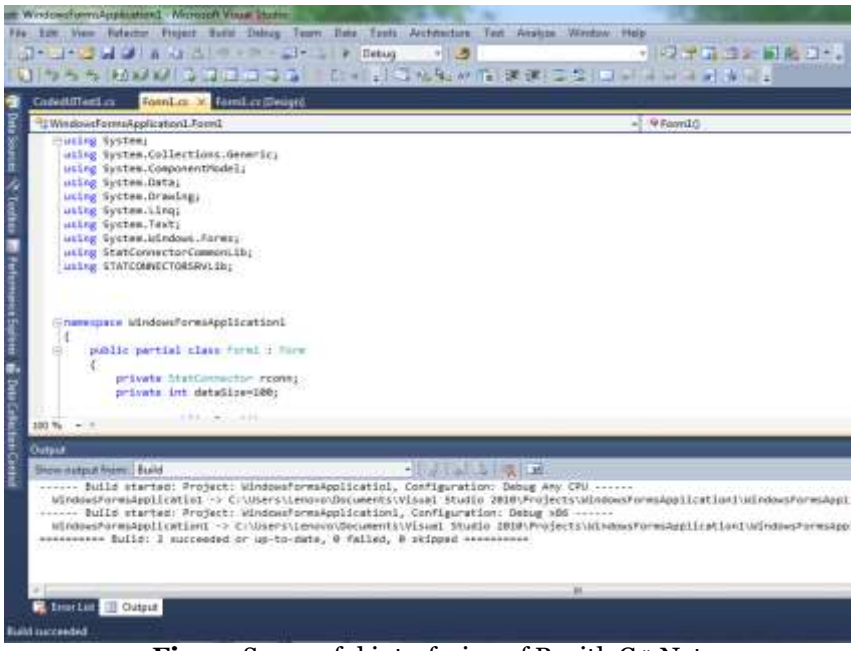

**Fig. 4.** Successful interfacing of R with C#.Net

## 5.2 Spectrum Visualization

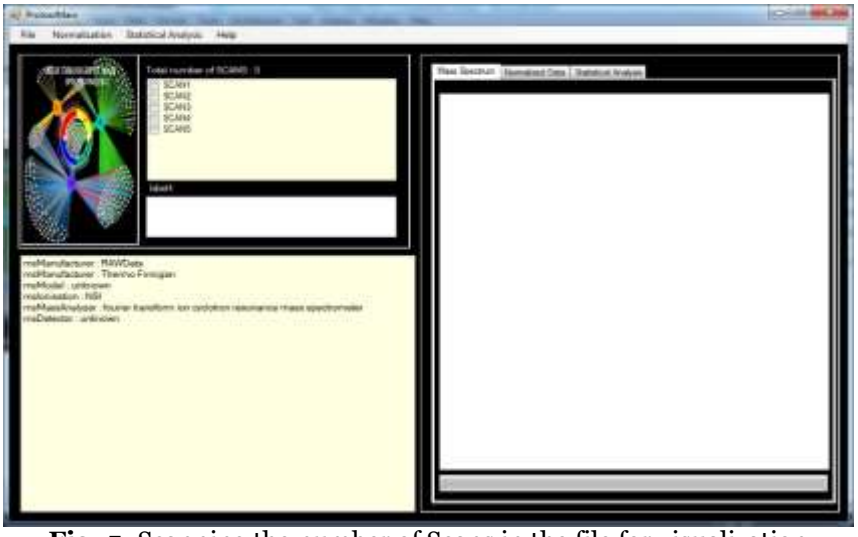

**Fig. 5.** Scanning the number of Scans in the file for visualization

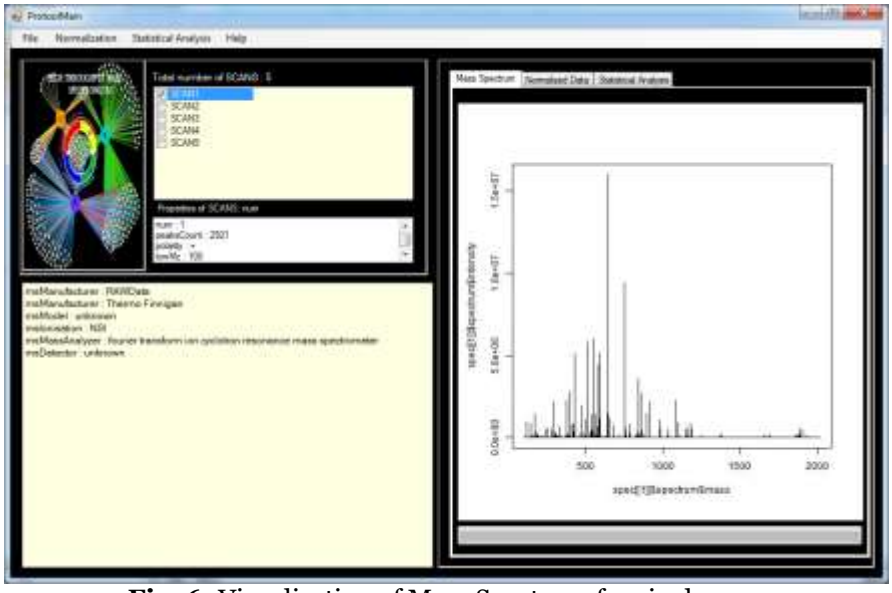

**Fig. 6.** Visualization of Mass Spectrum for single scan

## 5.3 Data Normalization

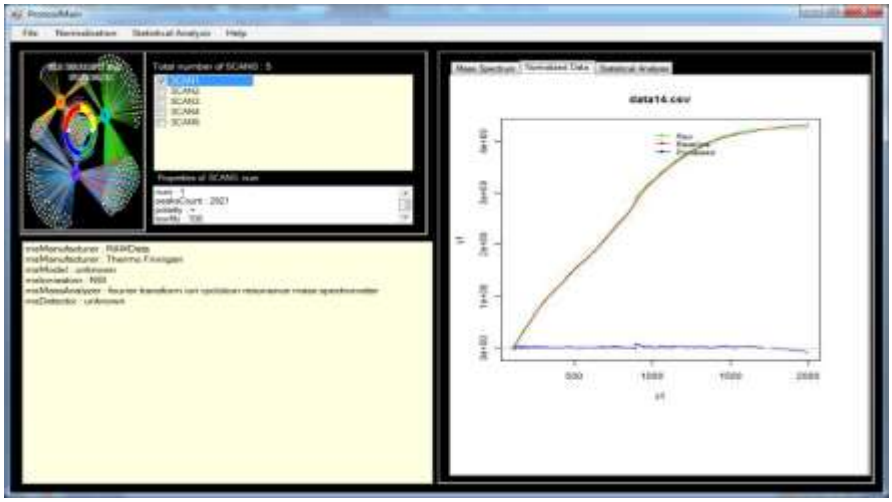

**Fig. 7.** Normalization

## 5.4 Peak detection

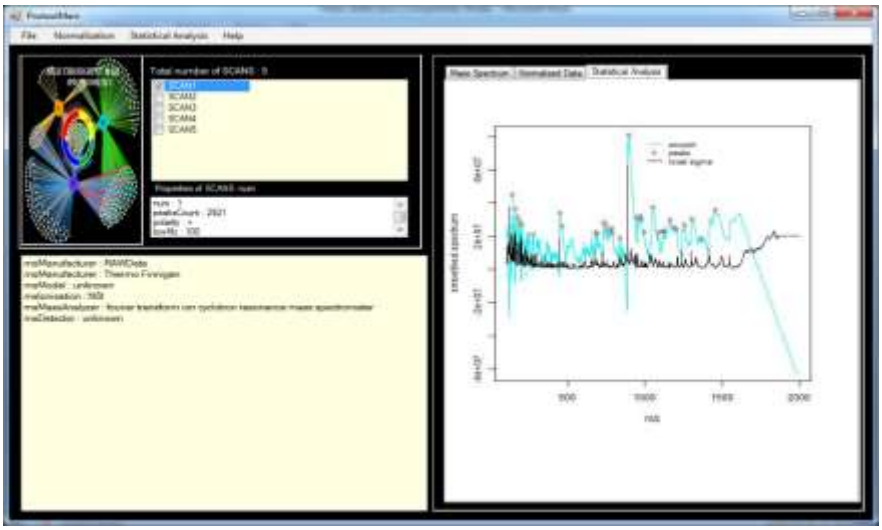

**Fig. 8.** Peak detection

## 6 Conclusion

In the realm of proteomics, there is a vast amount of initiated data. Today, it is imperative that these enormous volumes of data be collected and examined. There are numerous methods and instruments available for protein prediction. One analytical technique for analyzing data that was developed in big quantities is mass spectrometry [27]. The several facets of mass spectrometry have been observed. Tools and databases for mass spectrometry are accessible. People in the bio sciences can utilize the instrument with ease. The resulting output can be entered into Matrix Sciences' easily accessible MASCOT tool. Matrix Sciences licenses the use of a tool called Mascot.
The future scope is that as mass spectrometry tool generates files with various formats, so the data should be accessed from any file format. So, the tool can be further developed for various file formats generated by Mass Spectrometry tool. The data generated is so large, so some remedies can be taken to access this huge amount of data.

**Disclosure of Intrests**. The writers have know contending interests to proclaim that are pertinent to the substance of this article.

## References


1. Catherine P. Riley*, Erik S. Gough, Jing He, Shrinivas S. Jandhyala, Brad Kennedy, Seza Orcun, Mourad Ouzzani, Charles Buck, Ali M. Roumani and Xiang Zhang, " The Proteome Discovery Pipeline – A Data Analysis Pipeline for Mass Spectrometry-Based Differential Proteomics Discovery" : The Open Proteomics Journal, vol 3, pg[8-19].(2010)
2. Yerlekar, A., and M. M. Kshirsagar. "A review on mass spectrometry: Technique and tools." J. Eng. Res. Appl 4.4 .17-23. (2014)
3. Bibekanand Mallick, Zhumur Ghosh," BIOINFORMATICS - THE RISING SUN": INDIAN SCIENCE CRUISER" Volume 20 Number.(2006)
4. Ken Pendarvis†, Ranjit Kumar*†, Shane C Burgess, and Bindu Nanduri1," An automated proteomic data analysis workflow for mass spectrometry": BMC Bioinformatics. (2009)
5. Richard Pelikan and Milos Hauskrecht, "Efficient Peak-Labelling Algorithms for Whole-Sample Mass Spectrometry Proteomics": IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 7, NO. 1.(2010)
6. Mahmoud Rafea, HebaZaki Torky,Sultan, "Bioinformatics Data Mining Tool Using Data Collected from Red Blood Cells Hemolysate" : 2nd International Conference on Computer Technology and Development. (2010)
7. Himanshu Grover, Vanathi Gopalakrishnan, "Efficient Processing of Models for Large-scale Shotgun Proteomics Data" : 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing , Collaboratecom 2012 Pittsburgh, PA, United States, October pg[14-17].(2012)
8. Liu Dan, Huang Yuan-yuan, Ma Chen-xiang ," Feature extraction and classification of proteomics data using stationary wavelet transform and naïve Bayes classifier" .(2010)
9. Marcus Bantscheff & Markus Schirle & Gavain Sweetman & Jens Rick & Bernhard Kuster, "Quantitative mass spectrometry in proteomics: a critical review": Published online: 1 August 2007# Springer-Verlag. (2007)
10. Kondethimmanahalli Chandramouli and Pei-Yuan Qian. " Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity": published by SAGE.(2013)
11. Robert B. Northrop, Anne N. Connor, CRC Press, "Introduction to Molecular Biology, Genomics and Proteomics for Biomedical Engineers: IEEE ENGINEERING IN MEDICINE AND BIOLOGY MAGAZINE.(2010)
12. M. Tsiknakis, Grangeat, Binz, G. Potamias, F. Lisacek, L. Gerfault, C. Paulus, D. Manakanatas, V. Kritsotakis, M. Perez, D. Plexousakis, S. Kaforou, D. Kafetzopoulos, "Functional specifications of an integrated proteomics information management and analysis platform": Proceedings of the 29th Annual International Conference of the IEEE EMBS Cité Internationale, Lyon, France.(2007)
13. Saeed Amir*, Haihui Wang, Fangtao Sun, "Application of Diffusion based framelet transform to the MS-Based Proteomics Data Preprocessing": Proceedings of 2013 10th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan.(2013)
14. Eleftheria A. Mylona, Michalis A. Savelonas, Member, IEEE, Dimitris Maroulis, Member, IEEE, and Sophia Kossida, "A Computer-Based Technique for
15. Automated Spot Detection in Proteomics Images", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 15, NO. 4. (2011)
16. Sabine Bachmayer, "Preprocessing of Mass Spectrometry Data in the field of Proteomics": Proteomics and Bioinformatics, Masters Degree Program for Bioinformatics, University of Helsinki, Finland.(2008)
17. Laurent Gatto_ and Lisa M. Breckels, "A short tutorial on using pRoloc for spatial proteomics data analysis", Cambridge Center for Proteomics University of Cambridge. (2013)

18. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Nucleic Acids Res.30[PMC free article] [PubMed].(2002)
19. Colantuoni C, Henry G, Zeger S, Pevsner J. Biotechniques. 32:1316–1320. [PubMed] (2002)
20. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Nucleic Acids Res. ;29:2549–2557.[PMC free article] [PubMed],(2001)
21. Liang, C; Smith, JC; Hendrie, Christopher (2003). "A Comparative Study of Peptide Sequencing Software Tools for MS/MS. American Society for Mass Spectrometry".
22. Tureček, František; McLafferty, Fred W; "Interpretation of mass spectra". Sausalito: University Science Books.(1993)
23. Mistrik, R. "A New Concept for the Interpretation of Mass Spectra Based on a Combination of a Fragmentation Mechanism Database and a Computer Expert System". In Ashcroft, A.E., Brenton, G., Monaghan, J.J. (Eds.), Advances in Mass Spectrometry, Elsevier, Amsterdam, vol. 16, pp. 821(2004)
24. Monroe, M.E.; Shaw, J.L.; Daly, D.S.; Adkins, J.N.; Smith, R.D. MASIC: "A software program for fast quantitation and flexible visualization of chromatographic profiles from detected LCMS (/MS) feature"s. Comput. Biol. Chem., 32, 215-7.(2008)
25. Braisted, J.C.; Kuntumalla, S.; Vogel, C.; Marcotte, E.M.; Rodrigues, A.R.; Wang, R.; Huang, S.T.; Ferlanti, E.S.; Saeed, A.I.; Fleischmann, R.D.; Peterson, S.N.; Pieper, R. "The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC- MS/MS proteomics result". BMC Bioinformatics, 9, 529.(2008)
26. Saito, A.; Nagasaki, M.; Oyama, M.; Kozuka-Hata, H.; Semba, K.; Sugano, S.; Yamamoto, T.; Miyano, S. AYUMS: "An algorithm for completely automatic quantitation based on LC-MS/MS proteome data and its application to the analysis of signal transduction". BMC Bioinformatics, 8, 15.(2007)
27. Dudhe, Priyanka S., Manali M. Kshirsagar, and Ashwini S. Yerlekar. "A review on 2D gel electrophoresis: a protein identification technique." Int. J. Comput. Sci. Inf. Technol. 5 , 856-862.(2014)
28. Yerlekar, Ashwini, and Priyanka Dudhe. "A review on study and comparison between 2D gel electrophoresis and mass spectrometry." IOSR J Comput Eng 16 :97e104.(2014)
29. Callister, Stephen J., et al. "Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics." Journal of proteome research 5.2 : 277-286.(2006)