RESEARCH ARTICLE OPEN ACCESS

A Review on Mass Spectrometry: Its tools and data analysis for proteomics

Ms. Ashwini Yerlekar*, Dr. M. M Kshirsagar**, Ms. Priyanka Dudhe***
*(Department of Computer Technology, YCCE, Nagpur, Maharashtra
** (Department of Computer Technology, YCCE ,Nagpur, Maharashtra
*** (Department of Computer Technology, YCCE ,Nagpur, Maharashtra

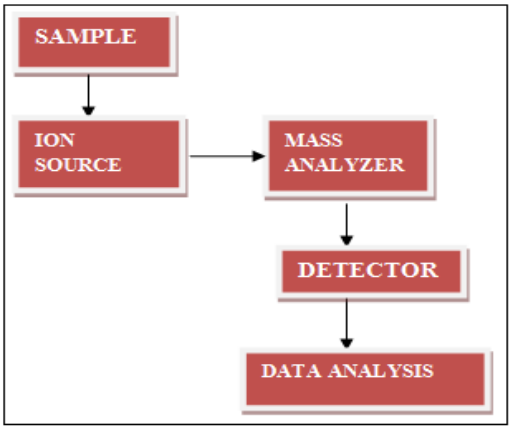
ABSTRACT-
With advent of 'The Human Genome Project' large-scale proteomics has rapidly come to dominate the post genomic age. The Protein Prediction is hard because of its complex structure. Data Analysis is the main issue while predicting the protein. The main problem in proteomics is to estimate the several proteins available in cell structure or tissue sample. Therefore, large scale proteomics technologies are required to measure physical connection of proteins in living organisms. Mass Spectrometry uses the technique to measure mass-to-charge ratio of ion. It's an evolving technique for characterization of proteins. A Mass Spectrometer can be more sensitive and specific, also complement with other LC detectors. Liquid Chromatography, unlike gas chromatography is a separation technique which helps to separate wide range of organic compounds from small molecular metabolites to peptides and proteins. This paper addresses the study of data analysis using mass Spectrometry. It also includes the study of various methods of Mass Spectrometry data analysis, the tools and various applications of Mass Spectrometry.
Keywords **–** Data Analysis, Mass Spectrometry, MS-tools, LC(liquid Chromatograohy)

I. INTRODUCTION

Protein structure prediction is a crucial area of life sciences. Every protein has its sequence, a primary structure; the helices and sheets; a secondary structure, the fold of the protein a tertiary structure, and multimeric formation of its polypeptide subunits, the quaternary structure, Protein structure has been validated for the past several years by chemical and physical methods. The dawn of protein sequencing began in 1950s, upon complete sequencing of insulin and then, ribonuclease.

Mass Spectrometer is a quantitative tool used to measure the mass-to-charge ratio of ions. The two primary methods for ionization of whole proteins are electrospray ionization (ESI-which turns the sample proteins into ions) and matrix-assisted laser desorption/ionization (MALDI- It uses laser to ionize the sample proteins and then push the proteins into the analyzer to produce a Mass Spectrum.). A typical Mass Spectrometer consists of three parts: a mass analyzer, a detector and an ion source. The ion source produces ions from the sample. The Mass Analyzer separates ions with different mass-to-charge ratios, these different ions are detected by detector. Finally, the mass spectrum is generated after all the data have been collected.

FIG 1: SCHEME GRAPH OF Mass Spectrometer



II. HISTORY OF MASS SPECTROMETRY

THE HISTORY OF MASS SPECTROMETRY HAS ITS GROWTH LIES IN PHYSICAL AND CHEMICAL RESEARCH REGARDING THE PROPERTIES OF MATTER. THE RESEARCH CARRIED OUT ON GAS DISCHARGES IN THE MID 19TH CENTURY LED TO THE FINDINGS OF ANODE AND CATHODE RAYS, WHICH RESULTS INTO POSITIVE IONS AND NEGATIVE IONS. IN THE EARLY ERA, MANY PROFESSIONAL SCIENTISTS WERE CONCENTRATED AROUND WITH ELECTRICITY FLOWING THROUGH VACUUM TUBES. THEY WERE FLOODING THE ELECTRONS FROM CATHODE (-VE) TO

ANODE (+VE) WHICH EUGEN GOLDSTEIN COINED AS "CATHODE RAYS". HE ALSO FOUND THAT THE RAYS IN CATHODE ARE VISIBLE IF SLITS ARE CUT. THE BACKGROUND GAS PRESENT IN CATHODE IS RESPONSIBLE FOR THE COLOUR OF THESE RAYS. THE COLOURS MAY BE HYDROGEN ROSE, AIR yellow and neon red. With advanced capabilities in the parting of these positive ions, it helps for the discovery of isotopes of the elements which are stable. The stable isotopes of Neon are: ^{20}Ne (neon with 10 neutrons and 10protons) and ^{22}Ne (neon with 12 neutrons 10 protons). Manhattan Project exclusively uses Mass spectrometers for the fission of isotopes of uranium which is essential to manufacture the atomic bomb.

III. VARIOUS TOOLS AVAILABLE FOR PROTEIN PREDICTION USING MASS SPECTROMETRY

Peptide identification algorithms fall into two broad classes: database search and de novo search. The first search takes place against a database containing all amino acid whereas the latter deduce peptide sequences without knowledge of genomic data. At present, database search is more reliable and considered to produce higher quality results for most uses. With advancement in instrument precision, the de novo search may become increasingly captivating.

A. Database search algorithms

i. SEQUEST:

SEQUEST [3] is a patented tandem mass spectrometry data analysis program. It was developed by John Yates and Jimmy Eng in 1994. The algorithm used by this program is covered by several US and European software copyrights. SEQUEST identifies collections of tandem mass spectra to peptide sequences that have been evolved from databases of protein sequences. SEQUEST, like many engines, learns each tandem mass spectrum one by one. The software evaluates protein sequences from a database to calculate the list of peptides. The peptide's intact mass is known from the mass spectrum, and SEQUEST uses this information to find the group of candidate peptides sequences that could be compared to the spectrum by including only those which are equally near the mass of the observed peptide ion.

ii. Mascot:

Mascot [4] is a proprietary identification program readily available from Matrix Science. Instead of using the cross correlation method, it carries out mass spectrometry data analysis through statistical comparisons of matches between projected and observed peptide fragments. In addition to the identification features, support for peptide quantization methods is provided, as per version 2.2.

iii. PEAKS DB:

PEAKS DB [5] is a proprietary database search engine, run in cohesion with de novo sequencing to automatically formalize the search results, allowing for a higher number of searched sequences for the available false search rate. In addition to providing an independent database search, results can be integrated as part of the software's multi-engine (Sequest, Mascot, X!Tandem, OMSSA, PEAKS DB) consensus reporting tool, in Chorus. The tool also provides a list of sequences identified exclusively by de novo sequencing.

B. MS/MS peptide quantification

i. OpenMS / TOPP

OpenMS [13] is a software C++ library for LC-MS/MS data management and analysis; it offers an infrastructure for the development of softwares related to mass spectrometry. It is free software available under the 2-clause BSD license (previously under the LGPL).TOPP - The Opens Proteomics Pipeline - is a set of small applications that can be sequenced to create analysis pipelines tailored for a particular problem. TOPP is developed using the structures and algorithms provided by OpenMS.

ii. MaxQuant

MaxQuant is a proprietary software for numerical proteomics developed by Jorgen Cox and others at the Max Planck Institute of Biochemistry in Martinsried, Germany.[14] The software is released as freeware under the "MaxQuant" Freeware Software License Agreement" and written in C#. The software provides its own search engine called Andromeda, and also accepts the analysis of label free and SILAC based proteomics experiments.

IV. DATA ANALYSIS

Mass spectrometry data analysis particular to the type of experiment producing the data. Fundamental divisions of data are general in understanding any data. Many mass spectrometers work in either electron mode or positive ion mode. Sometimes it is very necessary to know whether the obtained ions are negatively or positively charged. It is often essential in determining the neutral mass but it also symbolizes something about the nature of the molecules.

Various ion source results in arrays of fragments obtained from the fundamental molecules. The ionization source creates many fragments and mostly single-charged (1-) radicals (odd number of electrons), whereas non-radical quasimolecular ions, an electrospray source usually produces that are

frequently multiply charged. Tandem mass spectrometry produces fragment ions post-source and can change the kind of data obtained by an experiment.

A strongly prepared biological sample will probably contain a specific amount of salt, adducts get formed with the analyte molecules in various analyses Knowledge of the initial sample can provide insight into the component molecules of the sample and their peptides(protein in form of fragments). A sample from this synthesis/manufacturing process will probably consist of impurities chemically related to the target component.

Results can also depend heavily on sample preparation and how it was introduced. Much of the energetics of the desorption/ionization event is controlled by the matrix rather than the laser power. Mass spectrometry can measure sample purity, molecular structure and molar mass. Each of these queries requires a different experimental procedure; so adequate definition of the experimental goal is an early requirement for collecting the proper data and evolving the data.

V. MASS INTERPRETATION OF MASS SPECTRA

Since the specific structure or peptide sequence of a molecule is decrypted through the set of fragment of masses, the combined usage of various techniques helps in the interpretation of mass spectra. The first way of identifying an unknown compound is to compare its experimental mass spectrum against databases of mass spectrum. The software assisted interpretation or manual interpretation of mass spectra must be performed if match in spectral database in not found. Computer simulation of ionization and fragmentation processes occurring in mass spectrometer is the fundamental tool for assigning structure or peptide sequence to a molecule. Such simulation is often supported by a fragmentation library that contains published patterns of known decomposition reactions. Software developed on this idea has been developed for both small molecules and proteins.

VI. ADVANTAGES AND DISADVANTAGES OF MASS SPECTROMETRY

A. Advantages:

1. High Sensitivity– ability to detect very small amounts)
2. High Selectivity– Ability to tell molecules apart in a mixture
3. High Time Resolution
4. Low Cost
5. small sample size
6. fast
7. differentiates isotopes
8. can be combined with GC and LC to run mixtures, or can also be run in tandem for proteins or peptides etc.

B. Disadvantages*:*

1. Doesn't directly give structural information (although we can often figure it out)
2. Needs pure compounds.
3. Difficult with non-volatile compounds.

VII. APPLICATIONS OF MASS SPECTROMETRY

Mass Spectrometry consists of determining the isotopic composition of elements in a molecule, identifying unknown compounds, and determining the structure of a compound by understanding its fragmentation. Other uses include evaluating the amount of a compound in a sample or studying the fundamentals of gas phase in chemistry MS is now in very common use in analytical laboratories that study chemical, physical, or biological properties of a great variety of compounds.

A. Isotope ratio MS: isotope dating and tracking

Mass spectrometry is also used to determine the composition of elements in a sample. Variations in mass among isotopes of an element are very less, and the less abundant isotopes of an element are typically very rare, so a very sensitive instrument is requirement of Mass Spectrometry. These instruments, sometimes allude to as isotope ratio mass spectrometers (IR-MS), it uses a single magnet to bend a beam of ionized particles towards a series of Faraday cups which convert particle impacts to electric current. Isotope ratios are important markers of a variety of processes. The age of materials for example as in carbon dating is determined by isotopes ratios. Marking with isotopes which are stable are also used for protein quantification.

B. Atom probe

An atom probe is an instrument that combines field ion microscopy (FIM) time-of-flight and mass spectrometry to map the location of individual atoms.

C. Pharmacokinetics

Pharmacokinetics is often studied using mass spectrometry because of the complex nature of the matrix (often blood or urine) and the need for high sensitivity to observe low dose and long time. With a triple quadruple mass spectrometer, LC-MS is the most easily accessible instrumentation used; Tandem mass spectrometry is usually used for added probabilty. For quantization of usually a single pharmaceutical in the samples, standard curves and

internal standards are used. The samples represent different time points as a pharmaceutical is administered and then metabolized or cleared from the body. There is currently considerable interest in the use of very high sensitivity mass spectrometry for which is a promising alternative as micro dosing studies, to animal experimentation.

VIII. CONCLUSION

The data originated through Mass Spectrometry is in different format. According to the availability of tools the data is processed through various formats. With change in format the accession of data even change. The various steps of Mass Spectrometry analysis are studied and data is processed for protein prediction. So we have tried to study the mass Spectrometry data analysis, its techniques and applications and advantages and disadvantages.

REFERENCES

[1]. [1] Cheng Lu , Introduction to mass Spectrometry.

[2]. [2]http://www.colorado.edu/ chemistry/ chem5181/ MS1_Intro_SampleInt.pdf

[3]. [3]Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III (1994). "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database". J Am Soc Mass Spectrom 5 (11): 976–989

[4]. [4]Perkins, David N.; Pappin, Darryl J. C.; Creasy, David M.; Cottrell, John S. (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data".Electrophoresis 20 (18): 3551–67

[5]. [5]Liang, C; Smith, JC; Hendrie, Christopher (2003). A Comparative Study of Peptide Sequencing Software Tools for MS/MS. American Society for Mass Spectrometry.

[6]. [6] Colinge, Jacques; Masselot, Alexandre; Giron, Marc; Dessingy, Thierry; Magnin, Jérôme (2003). "OLAV: Towards high-throughput tandem mass spectrometry data identification".Proteomics 3 (8): 1454–63

[7]. [7] "OMSSA ms/ms search engine". Pubchem.ncbi.nlm.nih.gov. Retrieved 2011-09-27

[8]. [8] "RAId MS/MS search engine". QMBP NCBI NLM NIH. Retrieved 2008-01-01

[9]. [9] Bartels, Christian (31 May 1990). "Fast algorithm for peptide sequencing by mass spectroscopy". Biological Mass Spectrometry 19 (6): 363–368.

[10]. [10] Savitski, Mikhail M.; Nielsen, Michael L.; Kjeldsen, Frank; Zubarev, Roman A. (2005). "Proteomics-Grade de Novo Sequencing Approach". Journal of Proteome Research 4 (6): 2348–54

[11]. [11] Ma, Bin; Zhang, Kaizhong; Hendrie, Christopher; Liang, Chengzhi; Li, Ming; Doherty-Kirby, Amanda; Lajoie, Gilles (2003). "PEAKS: powerful software for peptidede novo sequencing by tandem mass spectrometry". Rapid Communications in Mass Spectrometry 17 (20): 2337–42.

[12]. [12] "Lutefisk - de novo MS/MS Sequencing". J. Alex Taylor. Retrieved 2009-11-16.

[13]. [13] Junker, J.; Bielow, C.; Bertsch, A.; Sturm, M.; Reinert, K.; Kohlbacher, O. (2012). "TOPPAS: A Graphical Workflow Editor for the Analysis of High-Throughput Proteomics Data". Journal of Proteome Research 11 (7): 3914–3920.

[14]. [14] Cox, J.; Mann, M. (Dec 2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification". Nat Biotechnol 26 (12): 1367–72.

[15]. [15] Rusconi, F. (2009). "massXpert 2: a cross-platform software environment for polymer chemistry modelling and simulation/analysis of mass spectrometric data". Bioinformatics 25 (20): 2741–2

[16]. [16] V. S. Antonov, V. S. Letokhov, and A. N. Shibanov (1980). "Formation of molecular ions as a result of irradiation of the surface of molecular crystals". Pis'ma Zh. Eksp. Teor. Fiz. 31: 471.JETP Lett.31: 441.

[17]. [17] Brown, R. S.; Lennon, J. J. (1995). "Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer".Anal. Chem. 67 (13): 1998–2003.

[18]. [18] 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, 2010, vol 3, pg[8-19]

[19]. [19] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Ho_, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy,

[20]. L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz,

[21]. J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol, 30(10):918_20, Oct 2012.

[22]. [20] L. Gatto and K. S. Lilley. MSnbase _ an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics.