Global Journal of Computer and Engineering Technology (GJCET)

Vol. 3, No. 2, July, 2023 ISSN: 2767-1933

A PRACTICAL APPROACH ON METHODS, ALGORITHMS AND TOOLS FOR COMPUTATIONAL PROTEOMICS

Rajendra Soni, Dr. C.V. Raman University, India (sonirajendra.soni@gmail.com)
Amrita Verma, Dr. C.V. Raman University, India (amrita85024@gmail.com)

ABSTRACT

Proteomics is a scientific discipline that focuses on the comprehensive investigation of proteins, embracing the diverse interests of researchers and medical professionals on their expression and physical characteristics. The discipline of proteomics is experiencing significant growth, as seen by the frequent publication of publications on advancements in technology and scientific investigations.Utilizing these novel instruments. This review largely emphasizes the quantitative aspect of protein expression and its significance.The computational infrastructure required for conducting extensive protein identification and posttranslational analysis.Adjustments. The main focus is placed on the integration of liquid chromatography–mass spectrometry.The utilization of spectrometry techniques, namely liquid chromatography-mass spectrometry (LC-MS) procedures, along with the corresponding tandem mass spectrometry (LC-MS/MS) approaches. Tandem mass spectrometry (MS/MS)Spectrometry, specifically tandem mass spectrometry (MS/MS), encompasses an additional analytical step performed within the instrument subsequent to molecule dissociation.The purpose of the event is to gather structural information, which may include but is not limited to sequence information.This review largely centers on the examination of in vitro digested proteins, namely those referred to as bottom-up or.The topic of interest is shotgun proteomics. Prior to delving into the topic at hand, a concise analysis of recent advancements in instrumentation is presented.The process of selectively increasing or decreasing the concentration of proteins, known as affinity enrichment and depletion, will be discussed. Additionally, an examination of the primary methods employed in this context, namely those that do not involve labelling, will be presented.of proteomics research. This approach involves incorporating stable isotopes into proteins during synthesis, allowing for accurate quantification of protein expression levels.The process of profiling a substantial quantity of proteins. Subsequently, an extensive discourse ensues on the diverse computational methodologies.LC-MS/MS data is commonly employed for the identification of peptides and proteins. This review article thereafter incorporates a concise discourse.This paper discusses the application of liquid chromatography-mass spectrometry (LC-MS) techniques for the determination of three-dimensional structures. The paper finishes with a dedicated section on this topic.This discussion focuses on the application of statistics and data mining techniques in the field of proteomics, with particular emphasis on the appropriate power analysis for clinical investigations.In order to mitigate the issue of over-fitting, it is crucial to employ strategies that are effective in the context of big data sets.

Keywords: Quantitative proteomics, LC–MS, Liquid chromatography–mass spectrometry, Protein identification, Affinity enrichment, Affinity depletion.

1. INTRODUCTION

The objective of this paper is to provide an overview of current advancements in the field of proteomics, with implications for both clinical and fundamental research. The area of proteomics encompasses a wide range of methodologies and is characterized by its diversity, attracting scholars from other disciplines. Individuals who lack expertise in the specific domain may encounter challenges while attempting to comprehend and engage with the subject matter. The field of proteomics has the potential to be utilized in various domains within the realm of biology. Furthermore, the field of medicine is also of great significance. In the context of mutagenesis, the technology has the potential to be utilized. The primary objective was to seek the translation of genetic variations into phenotypic alterations. The investigation focuses on the examination of amino acid sequences and the consequences of genetic changes. Regarding the field of biology, one area of focus involves the study of subsequent biological phenomena, such as the over-expression of proteins or alterations in genetic makeup. In the context of post-translational changes, such as phosphorylation, Alternatively, the process of methylation might also be considered. This analysis, however, centers its attention on the technological aspect. The focus of current research in the field of proteomics has shifted towards broader developments rather than individualized studies. Utilizing the technologies. The authors have endeavoured to structure and categorize subjects and

themes. Recent advancements will be presented in a manner that is easily comprehensible to the audience. Dispensing with any semblance of comprehensiveness. The field of proteins and protein chemistry, which encompasses the study of their properties and characteristics, has its origins in earlier scientific investigations. The field of proteomics has a history that spans more than a century .Gavin et.al (2006). The discipline of proteomics is now seeing significant growth and development. There are several scientific periodicals that are specifically focused on this subject matter, including The field of proteomics and its clinical applications are discussed in the Journal of Proteomics. Two academic journals that focus on proteome research are the Journal of Proteome Research and Molecular and Cellular Proteomics. In addition, scholarly articles that utilize and evaluate proteomics methodologies are also included. This research has been published in various other academic journals. A recent search was conducted on the PubMed database. A search query on the topic of "proteomics" resulted in the retrieval of more than 20,000 citations. Numerous monographs Numerous scholarly works have been dedicated to the topic, with certain publications exploring the incorporation of integration within the context. The integration of proteomics, genomics, and bioinformatics has been widely explored in scientific research. In this context, a comprehensive investigation was conducted to explore the relationship between these fields. The online retailer Amazon.com displayed a wide selection of meticulously designed scholarly publications. Regarding the topic at hand. Given the aforementioned context, the primary objective of the current review is inherently the individual exhibits a humble demean or. Initially, we shall endeavor to demarcate prominent subjects within the subject area. A few of those subjects will be addressed with minimal elaboration.

The inclusion of their mention and a limited number of references to more specialized review articles and significant manuscripts. Several subjects will be addressed. in a more comprehensive manner, yet without any content that could be deemed the analysis conducted was comprehensive and thorough. Therefore, we acknowledge the limitations of this review. We extend our apologies for the numerous deliberate and inadvertent exclusions present in the text. To facilitate the organization of subjects and appropriately delimit the extent of this discussion, in this review, we shall commence by partitioning the domain of proteomics into the realm of quantitative analysis. The field of expression proteomics encompasses the study of protein concentrations, including the concentrations of their modifications, along with the related peptide molecules. In addition to protein identification, this study also encompasses inquiries pertaining to the three-dimensional aspects of proteins. The structural analysis of a molecule can be determined by many techniques such as mass spectrometry, affinity purifications, and other methods. The examination of protein-protein interactions, with recognition of their significance. There exists a certain degree of overlap among these categories Nesvizhskiiet al. (2003). Our attention will be directed on Regarding the field of quantitative proteomics and its corresponding protein identification techniques, with further remarks and citations pertaining to instrumentation and three-dimensional (3D) technology. The analysis of structure by the utilization of mass spectrometry, affinity enrichment, and other pertinent techniques. The field of study encompassing statistics and data miningFrank and Pevzner (2005). In relation to the investigation of protein-protein interactions, In the context of interpersonal exchanges, we shall now proceed to reference a selection of noteworthy accounts, which encompass the identification and characterization of protein complexes and molecular machineries by the application of mass spectrometryMcCormack et al. (1997). The utilization of several approaches in molecular biology, including yeast-two-hybrid, has been documented in the literatureBrückner et al.(2009). The study conducted a series of experiments Keller et al.(2002). explore the field of quantitative proteomics in greater depthScholten et al. (2006). The field can be divided into distinct categories depending on mass spectrometric-based data Suprapto et al. (2007). In the context of liquid or gel separations, and measurements conducted using proteinCravatt et al. (2008). The utilization of affinity arrays and multiplexed immunoassays is observed in several academic research studiesBodenmiller B.et al. (2007). increasingly prevalent in the field of biomedical research and diagnostics. The prevalence of these technologies is expanding and their functionalities are growingMa et al. (2003). There are several noteworthy papers that discuss the applications of protein affinity arrays and multiplexed techniques. Immunological assays encompass a range of techniques, as indicated by Makarov(2000). Tanford and Reynolds (2001)This review, meanwhile, centers its attention on mass spectrometric-based methodologies. This study focuses on the development of assays for quantitative proteomics, with particular emphasis on the integration of multiple techniques Becker et al. (2007). Regarding the subject matter of high-performance liquid chromatography (HPLC), Mass spectrometry, sometimes referred to as LC-MS, along with its accompanying tandem techniques. Mass spectrometry is a widely used technique for the identification of peptides and proteins. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is also used for the same.

Mass spectrometric measurements can be broadly categorised based on whether the final measurements are conducted on intact or degraded proteins. Measurements on intact proteins can be classified based on whether they investigate the sequence structure and/or post-translational determinations (referred to as top-down measurements) or whether they only aim to determine the accurate molecular weight or distribution of weights

in a mixture. While top-down measures are becoming more capable, they are still limited in the field of proteomics. Typically, only a small number of proteins in complicated mixtures may be examined, and these proteins must have high relative concentration and be of moderate size. Additionally, the use of Fourier-transform mass spectrometer (FTMS) instruments, which are often quite costly, is necessary.
Conversely, the primary focus of this review is the examination of proteins by in vitro enzymatic digestion, sometimes known as bottom-up or shotgun.Radulovic et al. (2004) which accounts for the majority of research in this field. Decreasing the size of polypeptides employed in analysis enhances the capacity to measure and recognise a substantial number of proteins, and the reduced mass enables precise resolution and efficient utilisation of a diverse range of instruments. It should be noted that the process of separating, enriching, and depleting intact proteins before digestion can be included as a step in this bottom-up proteomic strategy. It is worth mentioning that bottom-up proteomics is becoming more capable of dealing with larger enzyme fragments by utilising a protease that cuts seldom.

2. METHODOLOGY AND THE WORK

Our focus in this paper is to highlight the Quantitative methods, in quantitative methods We can probably trace the origins of quantitative proteomics back to 1- and 2-dimensional gel electrophoresis. The power of this method was greatly enhanced by mass spectrometry's ability to identify proteins from stained spots. This was initially achieved through so-called peptide fingerprinting with MALDI; moving forward, the method has continued to evolve and grow in capability with a greater variety of general and functional group-specific gel stains, gel preparation methods, and image analysis software. However, the capacity of liquid-based technologies to detect and trace more proteins has recently surpassed that of 2D gels. Nevertheless, 2D gels exhibit global behaviour that is especially useful for studying post-translational modification.
However, the development of quanti-fying for quantitative LC-MS profiling . Hertz et al.(1971) and related identification is the subject of this review. There was clearly no quantification approach that could compete with 2D gels until LC-MS of proteins started to emerge in the 1990s. Due to the complexity of peptide mixtures and the fact that each peptide has its own unique relative sensitivity factor (RSF), as well as the fact that experts in the field were aware of matrix effects—in which the RSF of a molecule can change depending on the presence or absence of other molecules—the problem of how to quantify proteomics using LC-MS remained unsolved for quite some time Shilov et al. (2007). Additionally, most of what we know about metabolic pathways comes from mass spectrometry's lengthy history of using isotope-modified molecules for quantification, which dates back to the 1930s Liu et al.(2004) . Then, in the late 90s, scientists came up with an isotope labelling strategy for differential expression. They used a technique called isotope coded affinity tags (ICAT)Syka et al. (2004), which involved labelling two samples separately with a heavy and light isotope, mixing them so they would go through identical chromatography and mass spectrometry, and then comparing the ratios of the measured heavy and light labelled peptides. The iTRAQ (isobaric tag for relative and absolute quanti-fication) method has been used to refine this methodology Fasolo and Snyder (2009). As more and more chemistries were investigated, isotope labelling of reactants with peptides using stable heavy or light isotopes quickly became a popular practise Mulderet al. (2009). The relatively straightforward 18O method is one of several isotope labelling chemistries developed by various research groups Elias et al. (2004). In this approach, one sample is digested in H218O and the other in a standard aqueous solution to achieve differential labelling at the carboxyl terminusBarr et al.(1996).
The use of isotope-labelled amino acids as a food source for cell lines or tiny organisms is another method of labelling. Even for relatively small numbers of animals, the "stable isotope labelling by amino acids in cell culture" (SILAC) technique Eng et al. (1994) may be used to study species as big as mice, although the cost can add up quickly.
Shortly after isotopic labelling gained traction as a quantification tool, methods claiming to be label-free emerged. One method relied on normalised peptide signal intensities Hu et al. (2007) , while another, known as "spectral counting," used identification frequency Baggerly et al. (2008). The community that had grown fond of isotopic labelling was first resistant to these approaches, but when the validity of label-free procedures was repeatedly proven, they eventually gained widespread acceptance and began to be used Biemann et al. (1966). The concept of keeping matrix effects to a consistent degree through the use of significant chromatographic separation and comparisons of comparable types of samples was crucial to overcome them Clauser et al (1999). As an example, 2007 public research conducted by the Association of Biomolecular Resource Facilities (ABRF) compared several isotope labelling methods with those that did not use either Although intensity-based label-free quantification (MS1 or MS-only) is the more powerful of the two label-free quantification methods, it is currently out of reach for some smaller laboratories due to the complex software and meticulous, repeatable sample preparation procedures it requires Hinkelmann and Kempthorne (1994) . The intensity-based label-free method eliminates the need to redundantly identify peptides in every sample by first decoupling proton and identification and then connecting the two sets of data in silico using precise m/z and retention time. On the

other hand, small laboratories often prefer the spectral counting label-free method because it is the easiest to implement. However, this method has significant uncertainties when there are few identifications per protein, like 5 or fewer, and it might be better described as semiquantitative to qualitative in such cases. Regrettably, the majority of proteins in complex mixtures can only be partially identified using 2D gels and blinded mixture analyses. The label-free method was found to be successful, but the spectral counting approach has limitations due to the need to extensively analyse each sample using MS/MS and the small number of peptides or single peptides that were identified.

In theory, maintaining consistent sample preparation is not hard; rather, it is a question of paying close attention to standardising procedures with suitable assay testing. Important components include educating staff, creating and following standard operating procedures (SOPs), and assessing outcomes using quality control samples. An increasing number of software programmes are being developed to provide intensity-based label-free quantification.

To test preparation, LC-MS analysis, and software tracking performance, it is advised that laboratories utilise a pooled complex mixture containing spikes of varied quantities of exogenous proteins when they first apply the intensity-based label-free technique. Validated findings can also be obtained by regularly using intermittent quality control samples.

At present, 20,000 molecular ions may be differently quantified in a single 1-hour LC-MS run (excluding monoisotopic molecular ions) using intensity-based label-free quantification. This approach may also be applied in two-dimensional chromatography setups, where each fraction from the first stage of chromatography can generate this quantity of monitored molecular ions. For technical replicates (re-analysis of the same sample by LC-MS), seasoned labs can achieve average or median coefficients of variation (CVs) well under 10%. For analysis of different aliquots from a pooled sample, which includes independent chemical processing and LC-MS, the CVs are usually less than 10%. Research using human clinical samples often finds average within-cohort CVs of 25% or less, as was recently reported in a diagnostic/prognostic investigation of brain cancer utilising cerebral spinal fluid Breuker et al. (2008). This is expected due to biological diversity. Depending on the specimen type and preparation methods, a single 1-hour LC-MS run may typically detect 200-800 proteins with a false-discovery rate of 1% worldwide (see to the section below on protein identification for further information). Proteins in blood plasma will normally be less abundant in a 1-hour LC run compared to proteins in a comparable tissue sample because of the masking effect of very abundant proteins. Although the number of proteins monitored and quantified does not simply increase linearly with the number of first-state chromatographic fractions, it does so significantly when more fractions are added to a two-dimensional LC front-end study.

In the section on instruction, MRM was stated as a quantitative technique for validation or general selected focused analysis. For better quantification, accuracy, and quality control, isotopically labelled peptides are the go-to internal standards.

2.1 Software Tools for 3D Protein Structure

This software allows users to import a Protein Data Bank file and view the protein's structure in three dimensions. The structure of a protein is the sole necessary condition for understanding its function. The biochemical function of a protein is thus determined by its structure. Defining a protein's structure in terms of three-dimensional perspectives has proven effective in identifying and categorising both known and unknown protein structures. Helping researchers and biologists deduce a protein's function from its structure is the primary goal. It also gives you the option to see the 3D protein structure in several colour schemes and display formats, such as Space Fill and Ball & Stick.

This programme is designed to work exclusively with PDB format files. There is a cap on the total number of atoms and bonds that can be found in a protein. In other words,

Only five thousand atoms or bonds can be used with this instrument. Only researchers and biologists working on diseases and drug discovery will find this tool valuable.

Using a Protein Data Bank file as a starting point, it may display the three-dimensional structure of a protein. From a product standpoint, it also lets consumers examine a protein's function based on its structure.

The following is described by this product:

· Interface with the user
• Interface between software and devices
• Front end of the system
• Connector for hardware
• Recollection
• Management

2.2 Tool Features

You can choose a PDB file from your local disc or the Internet using this tool. The following features are available in the product:

1. A 3D representation of the protein structure is generated.
2. The 3D protein structure is visualised using various visualisation techniques.
3. The 3D protein structure that has been accessed can be freely rotated.
4. Using a unique colour, it identifies each amino acid inside the structure.
5. Individual atoms inside the structure are assigned a unique hue.
6. It assigns a distinct colour to each protein chain in the structure.

3. CONCLUSION

Following in genomics' footsteps, proteomics has experienced fast expansion and maturation. Proteomics has advanced methods for large-scale peptide and protein identification as well as differential expression (profiling) in response to the quantitative findings demanded by scientists. Instrumentation, three-dimensional structure determination, and analysis of post-translational modifications using affinity capture and advanced tandem mass spectrometry methods are just a few examples of the many areas that continue to see intense effort and quick advancements in the field. The field of proteomics has also embraced data mining and statistical tools that are common in other scientific fields.

Because proteins play such a pivotal role in life, recent advances in quantitative proteomics provide hope for the fast growth of our understanding in areas vital to the betterment of human health and other areas connected to biology.

REFERENCES

Barr J.R., Maggio V.L., Patterson Jr.D.G., Cooper G.R., Henderson L.O., Turner W.E., Smith S.J., Hannon W.H., Needham L.L., Sampson E.J.(1996), Isotope dilution—mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I, Clin. Chem. 42 , 1676–1682.

Baggerly K.A., Coombes K.R., Neeley E.S.(2008), Run batch effects potentially com- promise the usefulness of genomic signatures for ovarian cancer, J. Clin. Oncol. 26 , 1186–1187.

Becker C.H., Kumar P., Jones T., Lin H.(2007), Nonparametric mass calibration using hundreds of internal calibrants, Anal. Chem. 79 , 1702–1707.

Biemann K., Cone C., Webster B.R., Arsenault G.P.(1966), Determination of the amino acid sequence in oligopeptides by computer interpretation of their high- resolution mass spectra, J. Am. Chem. Soc. 88 , 5598–5606.

Bodenmiller B., Mueller L.N., Mueller M., Domon B., Aebersold R.(2007), Reproducible isolation of distinct, overlapping segments of the phosphoproteome, Nat. Meth. 4 , 231–237.

Brückner A., Polge C., Lentze N., Auerbach D., SchlattnerU.(2009) , Yeast Two-Hybrid: a powerful tool for systems biology, Int. J. Mol. Sci. 10 , 2763–2788.

Breuker K., Jin M., Han X., Jiang H., McLafferty F.W.(2008), Top-down identification and characterization of biomolecules by mass spectrometry, J. Am. Soc. Mass Spectrom. 19 ,1045–1053.

Clauser K.R., Baker P., Burlingame A.L.(1999), Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching, Anal. Chem. 71 , 2871–2882.

Cravatt B.F., Wright A.T., KozarichJ.W.(2008), Activity-based protein profiling: from enzyme chemistry to proteomic chemistry, Annu. Rev. Biochem. 77 ,383–414.

Elias J.E., Gibbons F.D., King O.D., Roth F.P., Gygi S.P.(2004), Intensity-based protein identification by machine learning from a library of tandem mass spectra, Nat. Biotechnol. 22, 214–219.

Eng J., McCormack A.L., Yates J.R.(1994), An approach to correlate tandem mass spec- tra data of peptides with amino acid sequences in a protein database, J. Am. Soc. Mass Spectrom. 5 , 976–989.

Fasolo J., Snyder M.(2009), Protein microarrays, Meth. Mol. Biol. 548 , 209–222.

Frank A., Pevzner P.,(2005). PepNovo: de novo peptide sequencing via probabilistic network modeling, Anal. Chem. 77 , 964–973.

Gavin A.C., Aloy P., Grandi P., Krause R., Boesche M., Marzioch M., Rau C., Jensen L.J., Bastuck S., Dumpelfeld B., Edelmann A., Heurtier M.A., Hoffman V., Hoefert C., Klein K., Hudak M., Michon A.M., Schelder M., Schirle M., Remor M., Rudi T., Hooper S., Bauer A., Bouwmeester T., Casari G., Drewes G., Neubauer G., Rick J.M., Kuster B., Bork P., Russell R.B., Superti-Furga G. (2006), Proteome survey reveals modularity of the yeast cell machinery, Nature 440 ,631–636.

Hertz H.S., Hites R.A., Biemann K.(1971). Identification of mass spectra by computer- searching a file of known spectra, Anal. Chem. 43 , 681–691.

Hinkelmann K., Kempthorne O.(1994), Design and Analysis of Experiments: Intro- duction to Experimental Design, John Wiley and Sons.

Hu J., He X., Baggerly K.A., Coombes K.R., Hennessy B.T., Mills G.B.(2007), Non- parametric quantification of protein lysate arrays, Bioinformatics 23 , 1986–1994.

Keller A., Nesvizhskii A.I., Kolker E., Aebersold R.(2002), Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. 74 , 5383–5392.

Liu H., Sadygov R.G., Yates III J.R.(2004) , A model for random sampling and estimation of relative protein abundance in shotgun proteomics, Anal. Chem. 76 , 4193–4201.

Makarov A.(2000), Electrostatic axially harmonic orbital trapping: a high- performance technique of mass analysis, Anal. Chem. 72 , 1156–1162.

Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G.(2003), PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry, Rapid Commun. Mass Spectrom. 17 , 2337–2342.

McCormack A.L., Schieltz D.M., Goode B., Yang S., Barnes G., Drubin D., Yates III J.R.(1997), Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level, Anal. Chem. 69 ,767–776.

Mulder J., Bjorling E., Jonasson K., Wernerus H., Hober S., Hokfelt T., UhlenM.(2009), Tissue profiling of the mammalian central nervous system using human antibody-based proteomics, Mol. Cell Proteomics 8 , 1612–1622.

Nesvizhskii A.I., Keller A., Kolker E., Aebersold R. (2003), A statistical model for identifying proteins by tandem mass spectrometry, Anal. Chem. 75 , 4646–4658.

Radulovic D., Jelveh S., Ryu S., Hamilton T.G., Foss E., Mao Y., Emili A.(2004), Infor- matics platform for global proteomic profiling and biomarker discovery using liquid chromatography–tandem mass spectrometry, Mol. Cell Proteomics 3 , 984–997.

Scholten A., Visser N.F., van den Heuvel R.H., Heck A.J. (2006), Analysis of protein–protein interaction surfaces using a combination of efficient lysine acetylation and nanoLC–MALDI-MS/MS applied to the E9:Im9 bacteriotoxin–immunity protein complex, J. Am. Soc. Mass Spectrom. 17 , 983–994.

Shilov I.V., Seymour S.L., Patel A.A., Loboda A., Tang W.H., Keating S.P., Hunter C.L., Nuwaysir L.M., Schaeffer D.A.(2007), The Paragon Algorithm, a next gen-eration search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra, Mol. Cell Pro-teomics 6 , 1638–1655.

Suprapto A., Karni-Schmidt O., Williams R., Chait B.T., Sali A., Rout M.P.(2007), The molecular architecture of the nuclear pore complex, Nature 450 , 695–701.

Syka J.E., Coon J.J., Schroeder M.J., Shabanowitz J., Hunt D.F.(2004), Peptide and pro- tein sequence analysis by electron transfer dissociation mass spectrometry, Proc. Natl. Acad. Sci. U.S.A. 101 , 9528–9533.

Tanford C., Reynolds J.A.(2001), Nature's Robots: A History of Proteins, Oxford University Press, Oxford, New York.