IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Literature Survey on Framework for Generating MS-MS Spectra for MS Spectrometry Data

Xylia Lamisa Nazareth[1], Shetty Mamatha Gopal[2]

4th Sem M. Tech, Dept of CS&E, Sahyadri College of Engineering and Management, Adyar, Mangalore, India[1]

Assistant Professor, Dept of CS&E, Sahyadri College of Engineering and Management, Adyar, Mangalore, India[2]

Abstract: It is no understatement that mass spectrometry is revolutionizing biological studies. Current mass spectrometers are capable of a tremendous throughput, and can survey thousands of proteins from complex sample mixtures. The most commonly applied mass spectrometry techniques is Tandem-MS. This high-throughput technology generates huge amounts of data. Databases are critical for recording and carefully storing this data. This project considers data acquisition performed by a vendor specific software that generates a file annotating the mass spectra and look up file. The mass spectra data file is given as an input to Hadoop. Here the file is divided into smaller chunks of large file based on the requirements. These smaller chunks are then compared with look up file and required analytics is performed. And the output is given in an interactive visualization environment.

Keywords: Mass Spectrometry, Hadoop, Database, Tandem-MS, Data Acquisition.

I. INTRODUCTION

Current mass spectrometers are capable of a tremendous throughput, and can survey thousands of proteins from complex sample mixtures. This allows for unbiased, global surveys of biological systems and is the reason mass spectrometry has been a consistent aspect of biomarker and discovery based experiments. Combined with the pace of whole genome sequencing, combined applications of proteomics and genomics have allowed for a comprehensive description of an organism's genome and proteome. In the context of human disease, this has allowed for identification of differentially regulated structural variants, annotation of novel coding regions. In proteomics, there are several methods in which the protein composition of a sample is determined. Similar to next-generation

Sequencing technologies, mass spectrometry is currently in a deluge of data. As the throughput and capabilities of mass spectrometers is rapidly increasing, the computational needs have similarly evolved. Broadly, analysis of mass spectrometry data processing as it pertains to proteomics entails three steps: 1) acquisition of mass spectra, 2) assignment of MS2 mass spectrum to peptide sequences, and 3) inference of proteins from the peptide sequences identified.
The first step of data acquisition is performed by vendor specific software that generates a file annotating the mass spectra. Following spectra acquisition, MS2 spectrums are assigned to peptide sequences.
In this method, a user defines what proteins they wish to identify in a sample and the search engine produces theoretical spectra corresponding to the enzymatically cleaved peptides from the user supplied input.Then, the search engine proceeds to match experimental spectrum against theoretical spectra to determine which theoretical spectra best explain the experimental data.

The above methods are primarily concerned with the identification of peptide and protein species. For many applications, such as identifying the proteins present in a sample or discovery of novel genes, a qualitative analysis of mass spectrometry data is sufficient. However, when quantitative information is utilized, powerful analyses are now possible. The utility of quantitative MS data is highly dependent on the accuracy and comprehensiveness of the tools used for analysis. The first step in quantitative proteomic data analysis is to associate spectral information of MS and MS/MS scans with peptide sequences, followed by quantification of the identified peptides.While increasing the number of spectral assignments allows for a greater number of quantitative measurements.Analogous to database search engines, each program used for quantisation has its own method for quantifying data. Thus, data quantified by one program may be missed by another and there may be inherent limitations that are common to several programs.
Big Data is a large complex information asset that outruns our current efficiency processing it. Convoluted reasoning of data may make a slight data appear to be vast. The term ‘Big Data’ describes innovative techniques and technologies to capture, store, distribute, manage and analyse petabyte or larger-sized datasets with high-velocity and different structures. Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

for analytics purposes.Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment.
Hadoop includes a fault tolerant storage system called the Hadoop Distributed FileSystem, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them.The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel.

II. LITERATURE SURVEY

This paper proposes about mass spectrometry and how it has revolutionized proteomics studies in a manner analogous to the impact of next-generation sequencing on genomics and transcriptomics. Several groups have used mass spectrometry to catalogue complete proteomes of unicellular organisms and to explore proteomes of higher organisms, including mouse andhuman. A general limitation of current proteomics methods is their dependence on predefined protein sequence databases for identifying proteins. To overcome this, we also used a comprehensive proteogenomic analysis strategy to identify novel peptides/proteins that are currently not part of annotated protein databases. This approach revealed novel protein-coding genes in the human genome that are missing from current genome annotations in addition to evidence of translation of several annotated pseudogenes as well as non-codingRNAs.[1]
Min-Sik Kim, Jun Zhong,and Akhilesh Pandey proposed in their paper regarding Common errors in mass spectrometry-based analysis of post-translational modifications.Mass spectrometry (MS) is a powerful tool to analyze complex mixtures of proteins in a highthroughput fashion. Proteome analysis has already become a routine task in biomedical research with the emergence of proteomics core facilities in most research institutions. Here, it review the most common errors in MS-based PTM analyses with the goal of adopting strategies that maximize correct interpretation in the context of biological questions that are being addressed. Finally, it provide suggestions that should help mass spectrometrists, bioinformaticians and biologists to perform and interpret MS-based PTM analyses more accurately.MS has rapidly become a method of choice for analysis of complex mixtures of proteins in a high-throughput fashion. PTM of proteins is a common mechanism by which the function of proteins can be precisely and dynamically regulated. Here, they have discussed many aspects of current proteomics pipeline for PTM analysis in biology where we should be cautious. Before we rush to biological conclusions from mass spectrometric data, we should keep the following parameters in mind: (i) the resolution of mass spectrometer being employed; (ii) isobaric PTMs, chemical modifications and amino acids; (iii) enzymes used; (iv) search algorithm; (v) protein databases; (vi) search parameters for PTMs; (vii) site localization algorithm; (viii) co-elution of modified peptides; and (ix) interpretation of quantisation data. Some of the problems that we have listed can be solved, for example, by using different proteases, multiple algorithms, more curated protein databases or optimized searching parameters. Other problems can be solved through improvements in current bioinformatics pipelines such as integration of different PTM databases for better and more complete annotation.[2]

Christopher J. Mitchell, Min-Sik Kim, Chan Hyun Na, and Akhilesh Pandey proposed paper where Quantitative mass spectrometry data necessitates an analytical pipeline that captures the accuracy and comprehensiveness of the experiments. Currently, data analysis is often coupled to specific software packages, which restricts the analysis to a given workflow and precludes a more thorough characterization of the data by other complementary tools. To address this, we have developed PyQuant, a cross-platform mass spectrometry data quantification application that is compatible with existing frameworks and can be used as a stand-alone quantification tool. In addition, PyQuant can perform specialized analyses such as quantifying isotopically behaviour samples where the label has been metabolized into other amino acids and targeted quantification of selected ions independent of spectral assignment. PyQuant is capable of quantifying search results.[3]

This Review provides a general understanding of paper spray-MS, including the methodology and theory associated with a number of different related applications. This method has become a direct sampling/ionization method for mass spectrometric analysis at ambient conditions, and as a result, it has greatly simplified and increased the speed of mass-spectrum analysis. It has now become an increasingly popular and important method for MS. The first part of this review discusses the fundamentals of paper spray. Some modifications are also reviewed, including nib-assisted paper spray, droplet monitoring, high-throughput paper spray, leaf spray, tissue spray and wooden tip spray. The second part focuses on recent applications, including the analysis of DBS, foodstuffs, drugs and oil. These studies show that paper spray-MS has great potential for use as a fast sampling ionization method and for the direct analysis of biological and chemical samples at ambient conditions.PS is now recognized as a fast sampling ionization method for the direct analysis of raw biological and chemical samples using MS. It is noteworthy that the method is operated at ambient conditions; this is very useful in terms of sample preparation prior to MS analysis. By applying a high voltage, analyte ions can be generated from a wet paper or porous substrate and only a small volume of solvent is needed. Matrix

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

effects are minimal, samples obtained by chromatographic separation, extraction or other techniques used to purify samples can all be easily applied. Furthermore, besides chromatograph paper, leaf or tissue can also be used for spray ionization and further applications might emerge.[4]

The emerging glycomics and glycoproteomics projects aim to characterize all forms of glycoproteins in different tissues and organisms. Tandem mass spectrometry (MS/MS) is the key experimental methodology for high-throughput glycan identification and characterization. Fragmentation of glycans from high energy collision-induced dissociation generates ions from glycosidic as well as internal cleavages. The cross-ring ions resulting from internal cleavages provide additional information that is important to reveal the type of linkage between monosaccharides. As a result, they can rarely distinguish from the mass spectra isomeric oligosaccharides, which have the same saccharide composition but different types of sequences, branches or linkages.In this paper, we describe a novel algorithm for glycan characterization using MS/MS. The preliminary performance tests show that our scoring function, which simply counts the number of support peaks, gives encouraging results. However, in most cases, the real structure is scored similarly to several other oligosaccharides. The more complicated the oligosaccharide structure was, the more optimal solutions we got. Furthermore, this simple scoring function prefers linear structure to branching structure, maybe because linear structures often have more hypothetical cross-ring cleavages, and hence have more (random) chances of matching experimental peak. In order to improve scoring function, a probabilistic model is required to describe how likely it is that one fragmentation could happen and one ion could be captured by the MS.[5]

In this study, they used an integrated transcriptomic and proteomic strategy to validate and improve the existing zebrafish genome annotation. They undertook high-resolution mass-spectrometry-based proteomic profiling of 10 adult organs, whole adult fish body, and two developmental stages of zebrafish, in addition to transcriptomic profiling of six organs. Workflow for Manual Genome Annotation—Peptide sequences identified from the alternate database searches were filtered for 1% FDR and compared with the protein database to identify novel peptides. These novel peptide PSMs were further checked via manual inspection for validity of the peptide identification. For functional analysis of novel genes, the protein sequence was obtained from the longest ORF in the frame of the peptides. If the peptide matched to multiple transcripts, the transcript that had the longest ORF in the frame of the peptides was selected. RT-PCR validation was carried out for novel genes, genome assembly error, novel exons, and novel splice events. One unique feature of this study was the additional manual annotation of all novel events, including manual validation of peptide spectrum matches. Accurate annotation of all genomes is obviously highly desirable.[6]

This paper is proposed on behaviourofintegrating transcriptomic and proteomic data for accurate assembly and annotation of Genomes. The primary goal of whole genome sequencing efforts in any new organism is to provide accurate assembly and annotation of all proteincoding genes in the genome. Complementing genome sequence with deep transcriptome and proteome data could enable more accurate assembly and annotation of newly sequenced genomes. This paper describe a systematic approach for an integrated transcriptomic and proteomic data-based reanalysis of genome assembly and annotation using An. Stephensi genome as a proof of principle. This paper demonstrated both the need and the utility for simultaneous large-scale transcriptomic and proteomic analysis as an integral part of whole genome sequencing projects by using recently generated whole genome sequences of 16 Anopheline species. Incorporation of proteomic evidence allowed us to change the designation of over 87 predicted ‘non-coding RNAs’ to conventional mRNAs coded by protein-coding genes. Importantly, extension of the newly corrected genome assemblies and gene models to 15 other newly assembled Anopheline genomes led to the discovery of a large number of apparent discrepancies in assembly and annotation of these genomes. Our data provide a framework for how future genome sequencing efforts should incorporate transcriptomic and proteomic analysis in combination with simultaneous manual curation to achieve near complete assembly and accurate annotation of genomes [7]

F. Cliquet, G. Fertin, I. Rusu, and D. Tessier proposed a solution to deal with the peptide identification problem in the case of unsequenced species by considerably improving Packet Spectral Alignment. Initially it showed how strongly enhanced PSA, a method we developed in, by selecting the most interesting alignment positions and by fixing each parameter, such as the filtering of the peaks or the number of allowed modifications. All of these improvements were validated in terms of (i) quality and (ii) speed in comparison to the reference method, SA, on experimental data. These results confirmed that PSA behaves better than SA and is therefore a good choice for our new framework. We then presented a new method, PSAwEL,that makes it possible to increase first rank identifications in presence of modifications. We have then proposed integrating all of these improvements, together with PSAwEL, within a new peptide identification framework. This new framework was evaluated using Brachypodium experimental data, showing that it led to the improvement of the number of first-ranked peptide identifications. In the future, we plan to add another step to this framework, in order to be able to automatically interpret the detected modifications. Such an interpretation may be possible in most cases by using a databank that references all known modifications.[8]

This proposed paper provides a technique to implement self-tuning in Big Data Analytic systems. Hadoop’s performance out of the box leaves much to be desired, leading to suboptimal use of resource, time and money. This paper introduces Starfish, a self-tuning system for big data analytics. Starfish builds on Hadoop while adapting to user

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

needs and system workloads to provide good performance automatically, without the need for users to understand and manipulate the many tuning knobs in Hadoop. Explores the MADDER properties (i.e Magnetism, Agility, Depth, Data-lifecyle-awareness, Elasticity, and Robustness). The behaviour of a map reduce job is controlled by settings of more than 190 configuration parameters. If the user does not specify the settings, then default values are used. Good settings for these parameters depend on job, data, and cluster characteristics. Starfish's Just In Time Optimizer addresses unique optimization problems to automatically select efficient execution techniques for map reduce jobs.[9]

III. CONCLUSION

The utility of quantitative MS data is highly dependent on the accuracy and comprehensiveness of the tools used for analysis.An emerging issue as well is that the pace of mass spectrometry data acquisition is outstripping many of the existing tools.To address various issues, we are developing a novel big data framework for quantitative analysis of MS data. This framework is a versatile, cross-platform quantisation tool that can be used in conjunction with existing data analysis frameworks or as a quantification node for a minimal, light-weight mass spectrometry data analysis pipeline.

REFERENCES

[1] Min-Sik Kim, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujog,"A draft map of the human proteome",Nature India,2015,doi:10.1038/nature13302.
[2] Min-Sik Kim, Jun Zhong, Akhilesh Pandey,"Common errors in mass spectrometry-basedanalysis of post-translational modifications",Proteomics 2016, 16, 700–714.
[3] Christopher J. Mitchell, Min-Sik Kim‡, Chan Hyun Na, and Akhilesh Pandey,"PyQuant: AVersatile Framework for Analysis of Quantitative Mass Spectrometry Data", TechnologicalInnovation and Resources,2016
[4] "Paper spray-MS for bioanalysis ",Bioanalysis - 6(2):199 - Full Text ,April 21, 2017
[5] Haixu Tang , Yehia Mechref and Milos V. Novotny,"Automated interpretation of MS/MS spectra of oligosaccharides",Bioinformatics,Vol. 21 Suppl. 1 2005.
[6] Dhanashree S. Kelkar, Elayne Provost, Raghothama Chaerkady,Babylakshmi Muthusamy,"Annotation of the Zebrafish Genome through an Integrated Transcriptomic and Proteomic Analysis", Technological Innovation and Resources,2014.
[7] T. S. Keshava Prasad, Ajeet Kumar Mohanty, Manish Kumar, Sreelakshmi K,"Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes",Cold Spring Harbour Laboratory,November 15, 2016.
[8] F. Cliquet, G. Fertin, I. Rusu, and D. Tessier,"Proper alignment of MS/MS spectra from unsequenced species ",Semantic Scholar,April 21, 2017.
[9] H.Herodotou,H.Lim,G.Luo,N.Borisov,L.Dong,F.B.Cetin,and S. Babu. Starfish: A Selftuning System for Big Data Analytics. In CIDR, pages 261–272, 2011