IMPORTANCE OF DATABASE NORMALIZATION FOR RELIABLE PROTEIN IDENTIFICATION IN MASS SPECTROMETRY-BASED PROTEOMICS

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Biotechnology

by
Mehmet Direnç MUNGAN

December 2016
İZMİR

We approve the thesis of **Mehmet Direnç MUNGAN**

Examining Committee Members:

Assoc. Prof. Dr. Jens ALLMER
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Prof. Dr. Talat YALÇIN
Department of Chemistry, İzmir Institute of Technology

Prof. Dr. Anne FRARY
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Assoc. Prof. Dr. Çağlar H. KARAKAYA
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Asst. Prof. Dr. Yavuz OKTAY
İzmir Biomedicine and Genome Center, Dokuz Eylül University

27 December 2016

Assoc. Prof. Dr. Jens ALLMER
Supervisor, Department of Molecular Biology and Genetics
İzmir Institute of Technology

Prof. Dr. Talat YALÇIN
Cosupervisor, Department of Chemistry
İzmir Institute of Technology

Assoc. Prof. Dr. Engin ÖZÇİVİCİ
Head of the Department of Biotechnology and Bioengineering

Prof. Dr. Bilge KARAÇALI
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Jens Allmer, who accepted me as a part of his lab and taught me the ways of the bioinformatics.

I am grateful to my co-supervisor Dr. Talat Yalçın, who taught me the chemical part of the proteomics and actual spectrometric process.

I thank my committee members, Dr. Anne Frary, Dr. Yavuz Oktay and Dr. Çağlar H. Karakaya for accepting to be in my thesis defence committee and for their valuable criticisms.

I also thank TÜBİTAK for their support from project 114Z177.

Finally, i want to thank my family and my friends as they have always pointed me in the right direction at the times that i thought i was lost.

ABSTRACT

IMPORTANCE OF DATABASE NORMALIZATION FOR RELIABLE PROTEIN IDENTIFICATION IN MASS SPECTROMETRY-BASED PROTEOMICS

One of the revolutionary steps towards proteomics, was introducing mass spectrometry to protein inference analysis. Its powerful aspects such as speed, and accuracy towards identifying and quantifying proteins have made it the first choice to obtain high-throughput data. Due to development of a variety of fragmentation techniques, mass spectrometry-based analysis even made it possible to acquire knowledge about single polymorphisms and modifications of amino acids of a peptide.

Although this technology provides enormous amounts of data, identification of the proteins is still a hard challenge to overcome due to the shortcomings of computational methods. Herein a novel methodology is offered to better analyze mass spectrometry data and overcome the deficiency of protein identification algorithms in terms of speed and accuracy.

When the spectral data is acquired from an organism by mass spectrometry, database search algorithms are used for protein identification if the protein sequences of the organism are known. These algorithms compare the experimental data from mass spectrometry analysis to theoretical data gathered from known databases of organism to try and find the best match by ranking the PSMs via scoring functions.

Since the databases can be too large to search and multiple databases with different sizes can contain the peptides of experimental data, database search algorithms may fail to produce fair, fast or complete results.

In this work a methodology is presented to overcome unfair scoring of peptides in different size databases and enable database search algorithms to utilize relatively big sized entries such as human chromosome six frame translations. In terms of speed and accuracy the method is found to be better than some of the existing methods.

ÖZET

KÜTLE SPEKTROMETRİ TABANLI PROTEOMİK ÇALIŞMALARINDAKİ GÜVENİLİR PROTEİN TANIMLANMASINDA VERİTABANI NORMALİZASYONUNUN ÖNEMİ

Protein tanımlaması çalışmalarında kütle spektrometrinin kullanılması proteomik alanındaki devrim niteligindeki adımlardan biri oldu. Protein nicelik ve nitelik belirlemelerindeki doğruluk ve hızlı olması gibi özellikleriyle, yüksek-işleme veri alımında kullanılmak üzere ilk seçim haline geldi. Farklı fragmentasyon yöntemlerinin geliştirilmesiyle, kütle spektrometri tabanlı analizler, bir peptiddeki tekli polimorfizmleri ve amino asitlerdeki modifikasyonlarla ilgili bilgi edinilmesini bile mümkün kıldı.

Bu teknolojinin muazzam ölçülerde veri üretmesine rağmen, protein tanımlama çalışmaları, hesaplamalı metodların eksikliklerinden dolayı, aşılması güç bir hedef halinde. Bu çalışmada, protein tanımlama algoritmalarının protein belirlemedeki eksikliklerinin üstesinden gelmek ve kütle spektrometri verilerini hız ve doğruluk yönlerinden daha iyi analiz etmek için orjinal bir algoritma önerilmiştir.

Bir organizmadan kütle spektrometri aracılığıyla spektral veri elde edildiğinde, eğer organizmanın protein sekansları bilinmekteyse, protein tanımlaması için veritabanı arama algoritmaları kullanılır. Bu algoritmalar, peptid-spektrum eşleşmelerindeki en iyi eşleşmeyi skorlama fonksiyonlarına göre bulmak için, kütle spektrometri analizlerinden alınan deneysel verileri, organizmaya ait veritabanlarından elde edilen teorik verilerle karşılaştırır.

Deneysel verilerin karşılığı olan peptidler farklı boyutlardaki veritabanlarında dağınık halde olabileceğinden ve veritabanlarının arama yapılmak için fazla büyük olabileceklerinden dolayı, veritabanı arama algoritmaları adaletli, hızlı veya eksiksiz sonuçlar çıkarmakta başarısız olabilmektedir.

Bu çalışmada farklı boyutlardaki veritabanlarında peptidlerin adaletsiz skorlamalara tabi tutulmasının üstesinden gelmek ve insan kromozomlarının 6 çerçeve translasyonları gibi göreceli olarak büyük boyutlardaki protein sekanslarının, veritabanı arama algoritmaları tarafından işlenmesini sağlamak amacıyla bir metodoloji sunuldu. Metodun, hız ve doğruluk payı pencerelerinden bakıldığında hali hazırda kullanılan çeşitli metotlardan daha iyi olduğu bulundu.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. Proteomics

Proteins consist of amino-acid sequences translated by the instructions encrypted within the genome (Schorlemmer et al., 2012) and are crucial to cells for sustaining their existences. They are the building blocks of any living organism that invoke certain dysfunctions if their structural conformation is in any way deformed (Uversky and Dunker, 2010). One of the main interests of the life sciences is to infer protein sequences and discover their structural properties (Domon and Aebersold, 2006). Development of different approaches has been made to achieve such objectives however, the aim remains elusive (Nesvizhskii et al., 2007).

Total protein content of an organism, proteome, is constantly being illuminated by the discipline of proteomics by means of proteins' interaction with other molecules, their duties, quantitations, post translational modifications and their expressions (Allmer, 2012).

In the early years of peptide sequencing, Edman degradation method was used to generate sequences by controlled separation of single amino acids from the N-terminus of a peptide (Edman, 1950). Great deal of expertise was needed to control such method and it generally failed to identify long sequences (Steen and Mann, 2004). In addition, Edman degradation would take too much time to complete a sequence in a high-throughput manner (Shadforth et al., 2005) and couldn't identify the peptide if it was acetylated at its amino acid terminus (Barton et al., 2009). In early 1990s the mass spectrometry method was begun to widely used for sequencing peptides (Wilm et al., 1996) which was significantly more rapid and could identify the peptide by the means of mass to charge ratios belonged to amino acids (Tyers and Mann, 2003).

1.2. Mass Spectrometry

The analytes of a sample are moved through fields which use electric or electromagnetism therefore, have to be ionized (Eidhammer et al., 2008b). Ionized compounds are analysed and assigned a mass-to-charge ratio (m/z) each, by a technique called mass spectrometry (Aebersold and Mann, 2003). Three critical parts form a mass spectrometer. Initially molecules are converted into gas-phase ions (generally they are protonated which is addition of a proton (Mann et al., 2001)) by an ionization source. After conversion, a mass analyser measures the m/z value belongs to each ion, then comes the part for an ion detector to detect the ions at a certain time (Figure 1.1) (Yates, 2000).

After the innovation of techniques used for ionization processes such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) ms-based proteomics became an ultimate tool for researchers (Aebersold and Mann, 2003)

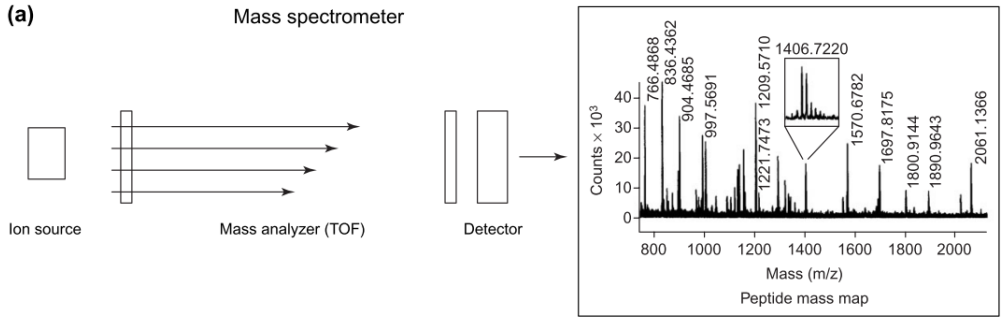

Figure 1.1. A mass spectrometer performing the first stage of mass spectrometry. This stage requires three devices: an ionization source, mass analyser and an ion detector. In this system the ions pass through the mass analyser from ion source to detector which takes a certain amount of time (time-of-flight) that will be used to calculate ions m/z value. (Source: (Yates, 2000))

To identify proteins using mass spectrometry, there are two methods being used traditionally (Pappin et al., 1993). One is called peptide mass fingerprinting (PMF) and the other tandem mass spectrometry (MS/MS) to further inform researcher about the sequential information (Henzel et al., 1993).

PMF is being widely used because when united with a database, it can speed up the identification process and identify the proteins relatively well (McHugh and Arthur, 2008). The m/z values gathered by the experimental spectra as shown in Figure 1.2, are searched through the reference database to align theoretical mass of an expected peptide (e.g. tryptic) to experimental mass(Kilby, 2007a).

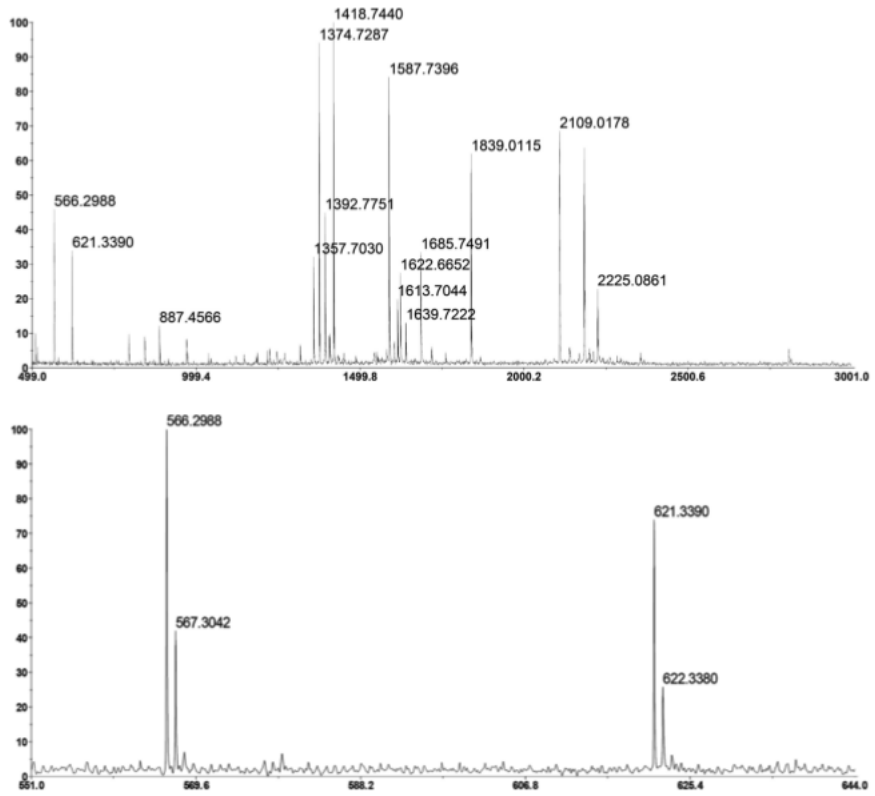

Figure 1.2. An example of the use of PMF. Digestion of matrix metalloproteinase 2 protein with trypsin, results into peptides with matching theoretical masses from the database with isotopic differences. The difference of isotopic peaks are 1 which indicates that single charge is carried by the peptides (Source: (Eidhammer et al., 2008a))

.

Although PMF is being widely used for identification, the experimental spectrum may fail to match any theoretical mass from database, due to unexpected causes of mass differences in a peptide like post-translational modifications, erronous databases or single nucleotide polymorphisms (SNPs) (McHugh and Arthur, 2008).

In MS/MS analysis, tandem mass analyzers are used to measure and select the peptide of interests m/z range and then, after a fragmentation step, calculate the m/z of fragment ions, respectively 1.3 (Nesvizhskii and Aebersold, 2005). Different fragmentation techniques have emerged but in the end, the aim is to procure smaller pieces (generally referred to as ”daughter” or ”fragment” ions) of the selected peptide from the first MS (generally referred to as ”precursor” or ”parent” ions) (Nesvizhskii and Aebersold, 2004).

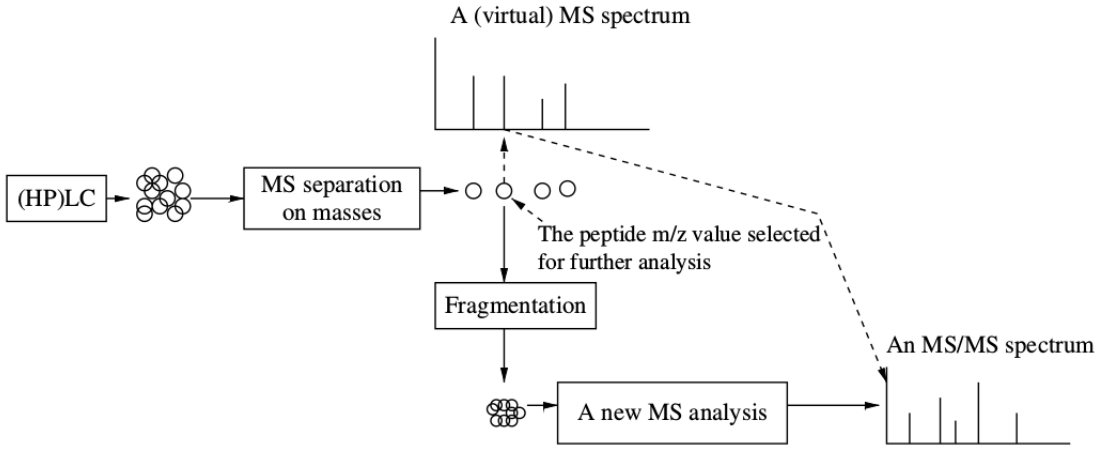

Figure 1.3. MS/MS spectrum obtain workflow. Peptide of interests m/z value is selected after first stage MS for further analysis, which then goes under a fragmentation step to be analyzed again with a new MS run. Constructing MS spectrum isn't needed for such goal. (Source: (Eidhammer et al., 2008c))

.

Information provided by tandem MS data is far more detailed and useful to identify proteins than PMF (Küster et al., 2001) and can be used to determine the even single amino acid variants 1.4, post translational modifications or other structural characteristics since they all cause masses to differ (Domon and Aebersold, 2006).

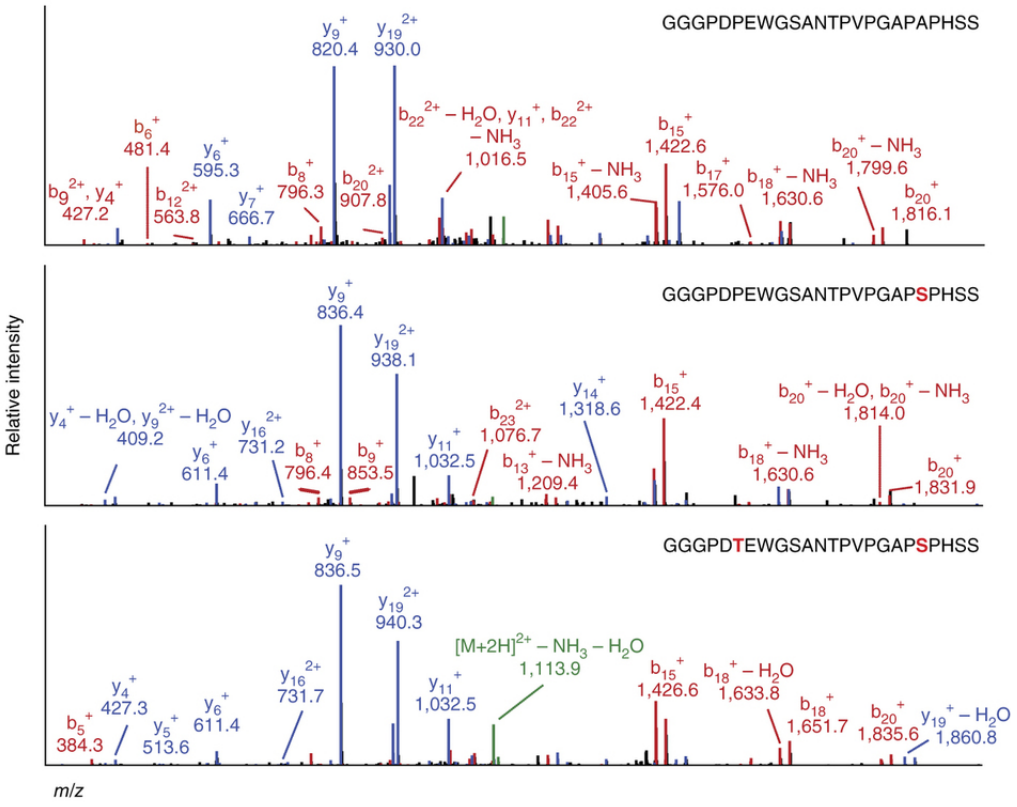

Figure 1.4. Tandem ms result of isotopically recoded glycopeptides. B ions are shown in red and y ions are shown in blue as fragmentation spectra. Single amino acid changes are shown in bold red font. (Source:(Woo et al., 2015))

.

1.3. Bottom-up and Top-down Proteomics

There are currently two approaches that are used to perform proteomic analysis: top-down and bottom-up proteomics (Chait, 2006). To identify proteins , gather detailed information about their sequential information and posttranslational modifications, bottom-up approach is widely used (Aebersold and Mann, 2003).

In bottom-up workflow, proteins are digested with an enzyme in to peptides then ionized by an ionization source for a mass spectrometer to analyse the sample followed by fragmentation if second MS analysis will be used (Chait and Kent, 1992) (Figure 1.5, top panel).

When exercising top-down proteomic approach, researcher must keep the protein intact when introducing the sample to the mass spectrometer where the sample will be converted into the gas phase ions and fragmentated, yielding proteins and its fragments masses (Chait, 2006). Although it seems nice to get both of these informations from a sample, it is hard to analyze large proteins by top-down proteomics as Han et al. managed to set the limit around 200 kD (Han et al., 2006).

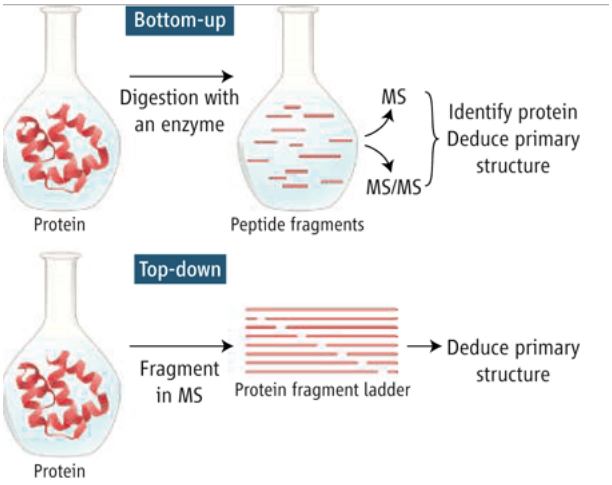

Figure 1.5. Work principals for bottom-up and top-down proteomics. At the top panel, bottom-up method is shown. The analytes are digested then processed by a mass spectrometer whereas at the latter strategy, protein stays intact when ionized and then fragmentated within the mass spectrometer (Source: (Chait, 2006))

.

1.4. Ionization Techniques

As mentioned earlier, in order to measure the molecular weights of an analyte, sample must be ionized and converted into gas phase. Molecules with higher masses are hard to transform into intact gas-phase ions. MS caught biochemists attention with early ionization techniques (Burlingame et al., 1994) but MALDI MS (Whitehouse et al., 1989) and ESI MS (Karas and Hillenkamp, 1988) have made the breakthrough.

Using minimum fragmentation, these ionization methods have the advantage of creating gas-phase ions from larger biomolecules with higher efficiency. Usually, MALDI

combined with TOF mass analyzer (one of the cheapest setups) and ESI combined with ion trap or a Q-TOF mass analyzer are used as instrument setups (Kilby, 2007b).

1.4.1. MALDI

MALDI is used to carry out MS/MS tasks but it is the general choice of ionization source when performing single MS (Eidhammer et al., 2008e). Beside peptides and proteins, it can selected to analyse large, non volatile biomolecules, oligonucleotides, oligosaccharides (Zenobi and Knochenmuss, 1998). MALDI setup is comprised of a matrix, which bears the analyte and a laser source, which energize the matrix compound followed by proton transfer to the analyte (Bökelmann et al., 1995).

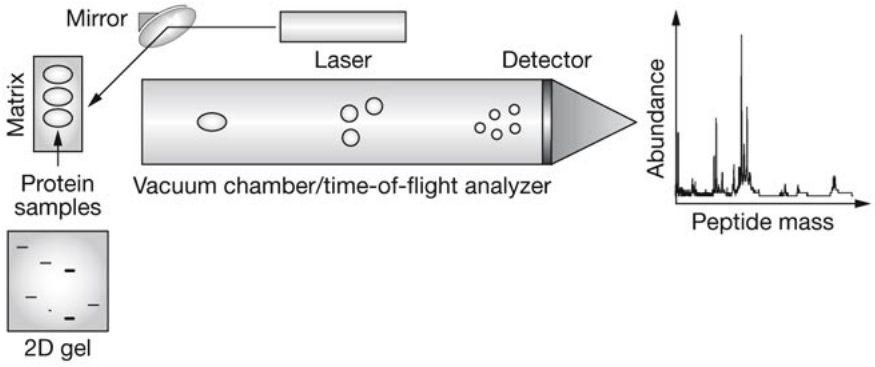

Figure 1.6. MALDI-TOF instrument setup. Here the analyte acquired from 2d gel is coupled with a matrix. Laser beams ionize the molecules which then enter a vacuum chamber and as soon as they hit the detector, according to their retention time, assigned a m/z value. (Source: (Din et al., 2007))

One of the most important factors that affects a MALDI-MS spectra is the matrix quality (Beavis and Chait, 1990). Dried droplet technique is the first and most used method of matrix-sample preparations (Cohen and Chait, 1996). In this technique, applying matrix-analyte mixture comes after the pure surface introduction via rapid evaporation (Vorm et al., 1994). Using volatile compounds like acetone can aid the rapidness of evaporation. For the sake of achieving better MALDI-MS results, different prepatation methods should be used for different samples given that some peptides might not be able to couple well with the used matrix (Kussmann et al., 1997).

1.4.2. ESI

ESI depends on a needle containing electromagnetic field which the sample passes through. Ions are generated by the high electric potential (Patterson and Aebersold, 1995) and the analytes as charged droplets evaporates which leads to higher charge density (Ikonomou et al., 1991). This increase in charge density plays a crucial role in resolving analyte ions into the gas phase, given that dense drops will eventually have a force (Coulomb repulsion force) greater than the surface tension resulting in the creation of smaller drops (Figure 1.7) (Covey et al., 1988).

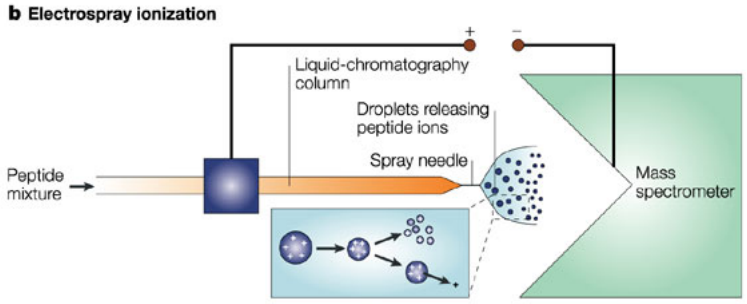

Figure 1.7. ESI workflow (Source: (Steen and Mann, 2004))

In positive ion mode the analyte is sprayed at low pH to encourage positive ion formation.

Positive ionization mode is generally used in MS/MS spectra (Köcher et al., 2003) of analytes to positively charge the ions at low pH (Keshishian et al., 2009) which causes of the appearance of relatively high amounts of anions (Gatlin and Turecek, 1994). For proteins and peptides, ESI effectively produces gas-phase ions (Covey et al., 1988) and different charge states can be present due to the lenght of polypeptide chain (Mirza and Chait, 1994).

1.5. Fragmentation Techniques

To generate structural or sequential knowledge about peptides, various fragmentation techniques have been used for many years (Paizs and Suhai, 2005). In mass spectrometry, peptide fragmentation process relies on imparting energy onto the molecule (Biemann, 1992). Applying energy to the peptide causes breakage of the amide bonds which leads to backbone fragments. Fragment ion type can be a,b,c if the charge is retained on the N-terminal fragment or x,y,z if the charge is retained on the C-terminal fragment (Figure 1.8) (Roepstorff and Fohlman, 1984).

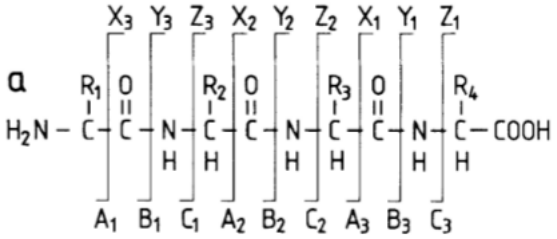

Figure 1.8. Peptide fragmentation scheme leading to different ion types. (Source: (Roepstorff and Fohlman, 1984)).

Different fragmentation techniques have been developed such as CID, HCD, ETD, ECD and EtcD which are useful in specific areas.

1.5.1. Collision Induced Dissociation

Colliding peptides with neutral gas has become the most frequently used fragmentation method in MS/MS. It is commonly referred as collision-induced dissociation (CID) when the ion goes through an activation process by collision and fragmentation (Wells and McLuckey, 2005). Between the carbonyl and amine groups, the peptide backbone amide bonds dissociate by the energy ,built up from the continuous collision of neutral gas and the peptide inside the collision cell, resulting in the product ions and/or neutral losses from the precursor molecule (Figure 1.9). Generated fragment ions are often observed as b and y ions (Molina et al., 2008).

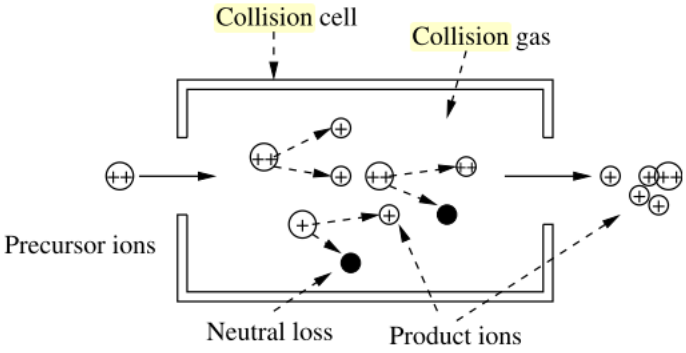

Figure 1.9. The principle of CID. (Source: (Eidhammer et al., 2008d)).

1.5.2. Higher-Energy Collisional Dissociation

Originated from CID a new technique called higher-energy collisional dissociation (HCD) is used specifically by orbitrap mass spectrometer. In HCD ions are trapped first then fragmented later on in a collision cell and returned to C-trap before mass analysis by orbitrap. Like CID technique, it mainly create b and y ions (Olsen et al., 2007).

1.5.3. Electron Capture Dissociation

When free electrons interact with multiply charged peptides, ECD occurs. This electron beam irradiation causes the backbone to fragmentate to mostly c and z fragments (Zubarev et al., 2008). When identifying peptides with modifications, ECD has a great usability (Zubarev et al., 1998). ECD is a nonergodic process given its rapidness of the cleavage being faster than the intramolecular energy randomization (Bakhtiar and Guan, 2005). Disadavantage of ECD however, is that it can only be used with Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry (Zubarev et al., 2008).

1.5.4. Electron Transfer Dissociation

Replacing free electrons with anionic species, ETD carries the mechanism of ECD fragmentation that can be used with non-FT-ICR instruments that can trap ions long enough for the electron transfer to occur. Peptide backbone fragmentation in ETD is extremely similar to ECD resulting in the creation of c and z ions (Syka et al., 2004).

1.5.5. EThcD

One of the latest techniques in fragmentating the peptides is the combination of two: ETD and HCD. In EThcD, ions are first exposed to ETD technique in the ion trap stage.For further fragmentation, resulting ions (both precursors and products) are then moved to the collision cell to be exposed to the HCD technique (Frese et al., 2012). Since its a combination of HCD and ETD, EThcD fragmentation shown to produce b, c, y, and z ions and and better spectral quality than ETD or HCD alone (Frese et al., 2013).

1.6. Mass Spectrometers

1.6.1. Orbitrap Mass Analyzers

Mass spectrometers such as orbitrap, work by the principle of changing ion velocity in an electrostatic field. In orbitrap, ions are confined by the coaxial and outer electrodes, allowing ions to oscillate along the axial electrode. Severance and detection of the ions are accomplished by their different oscillation frequencies, caused by their distinctive m/z values (Figure 1.10) (Makarov, 2000).

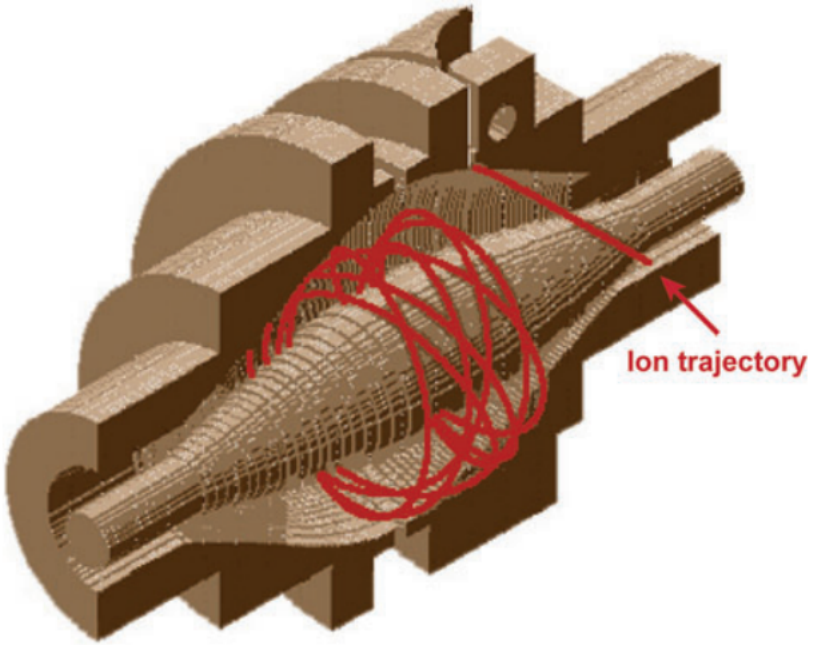

Figure 1.10. Ion trajectory shown by schematic diagram of the Orbitrap. (Source: (Hu et al., 2005)).

1.6.2. Time of Flight Mass Analyzers

Time of flight analyzers depend on the acceleration and deceleration of passing ions peculiar to their m/z values. First, ions are exposed to electrical potential which makes them faster in accordance with their charges. Next comes the field-free drift region where ions move through without any accelerating effect which makes them pierce through the field solely depending on their kinetic energies. Therefore, time taken to pass through the field free region is characteristic to different m/z values of ions (Guilhaus, 1995).

1.7. Computational Methods

In proteomics, one of the crucial tasks that needs to be done meticulously, is the correct identification of the proteins by computational methods. To accomplish such aim, there are two methods that are generally used by the researchers: database search and de novo sequencing.

De novo sequencing is used for finding novel peptides when the sequential information of the proteins origins is ambigious, but if the spectra acquired from MS or MS/MS processes are known to be in a database containing proteome of the sample, database search algorithms are used (Allmer, 2011).

1.7.1. Database Search Algorithms

One of the most frequently used method is database searching when identifying proteins. Commercial (Peaks (Ma et al., 2003)), and free software programs such as X!Tandem (Craig and Beavis, 2003), Omssa (Geer et al., 2004), MS-GF+ (Kim and Pevzner, 2014), PFind (Li et al., 2005), MyriMatch (Tabb et al., 2007), MS Amanda (Dorfer et al., 2014) and InsPecT (Tanner et al., 2005) have been developed in order to tackle the protein identification problem.

General idea of a database search algorithm is digesting the database with an enzyme defined by user (usually trypsin as default), generating theoretical spectra of the

digested peptides, comparing input spectra to the theoretical spectra and scoring accordingly (Xu and Ma, 2006).

1.7.2. False Discovery Rate

When analyzing MS/MS data by using database search algorithms, depending on the quality of spectral data and target database (actual sequences of the organism of interest), a level of ambuiguity arises for peptide-spectrum matches (PSM) (Chen et al., 2005). False discovery rate (FDR) is a commonly used method to identify the false positive results of PSM population (Elias and Gygi, 2007).

For this method, beside of the target database, a decoy database (database that ideally should contain none of the correct PSMs) must be created. Creating a database with the reversed forms of the target sequences provides a simple way to create decoy databases. After acquiring PSM results for both databases, FDR is used to explain false positive identifications by confining the PSM that passed the score criteria for decoy database (Elias and Gygi, 2007).

1.7.3. Aim

Protein identification by database searching methods, relies on the aspects of both spectra and database. Some of the algorithms fail to scan relatively big sizes of database and even if they don't, aspects of the database such as size and redundancy affects the scoring system thus the identification of the protein directly. Herein a normalization method is described in order to enable database search algorithms to effectively analyze spectral data.

CHAPTER 2

MATERIAL AND METHODS

2.1. Spectral Dataset

A total of 45 peptides were synthesized from 5 proteins (cytochrome c (ACN: P00004), bovine serum albumin (ACN: P02769), oval albumin (ACN: P01012), myoglobin (ACN: P68082) and lysozyme c (ACN: P61626)) by GL Biochem Ltd. These peptides were mixed mixture at LC-LTQ Orbitrap XL facility at FCGZ. HCD, ETD, EthcD, CID fragmentations were used to generate MS/MS data by the Orbitrap and TOF mass analyzers. 5000 spectra were arbitrarily chosen to measure speed of algorithms on sequence databases with different total size and database entry size. 4137 spectra were selected according to consensus of results of 8 algorithms run on human protein database with default settings of database search algorithms.

2.2. Split and Merge Method

Given that the scoring systems of database search algorithms react differently depending on the candidate peptide counts of the databases and algorithms themselves may fail to work because of higher sized databases.

In order to integrate different sized databases and enable all tools to utilize them, split and merge (SM) method was used. In split method, all entries in all databases were first split into 1000 amino acid long entries and 100 amino acid long overlap entries between splitting points. Overlap entries must be generated since the continuum of the peptide sequence might be broken in splitting process.Trypsin enzyme was used for protein cleavage so each type of entry was extended until a Lysine (“K”) or an Arginine (“R”) amino acid was found.

After splitting of the raw databases they were all merged into 6 databases which were created by taking entries proportionally from each database and basically merge

them in six files. Merging was done proportional in respect of size so that every merged database had the same size and in respect of entry number from each split file so that every merged database had the similar redundancy with the others.

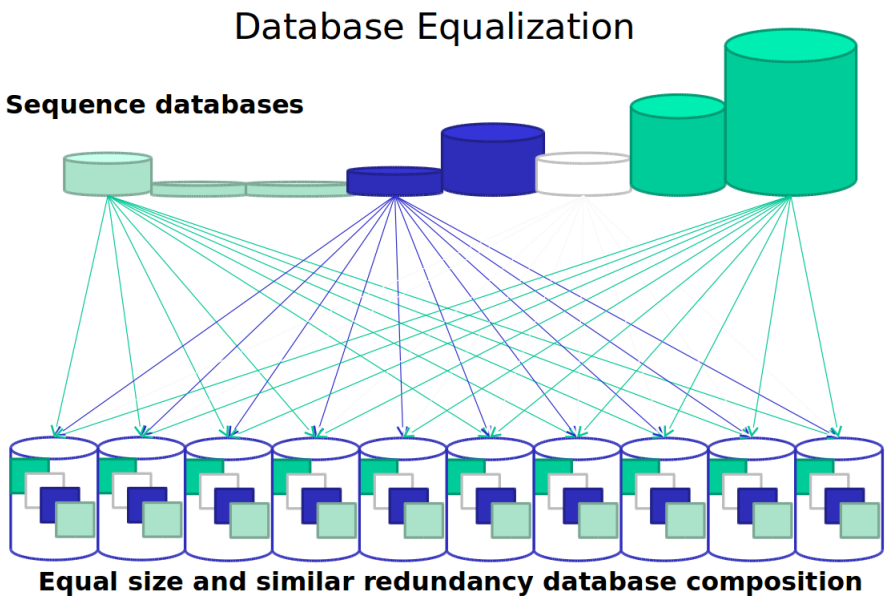

Figure 2.1. The principle of split and merge method.

Split version and split-merge version of human chromosome databases were used in score and accuracy comparisons. In total 6 split files and 6 split-merge files (all files have the db size 275MB) were produced.

2.3. Evaluating the Influence of Contig Size on Database Search Tool Performance

2.3.1. Speed and Limitation Measurements

The main aim of this step is to prove that algorithms are struggling to process large fasta elements and some of them even fail to produce any result or take much longer than split versions of them. In this study, 7 different sized raw (contains one fasta element)

databases and split versions of them were used for speed comparisons and tool limitations (Table :2.1).

Table 2.1. Database source and sizes for speed evaluation.

Contig Composition	Raw Size	Split Size
Human Chr1 + Part of Chr2	500 MB	607 MB
Part of human Chr1	250 MB	305 MB
Part of human Chr1	100 MB	120 MB
Part of human Chr1	50 MB	60 MB
Part of human Chr1	10 MB	12 MB
Part of human Chr1	5 MB	6 MB
Part of human Chr1	1 MB	1.3 MB

After creating the databases required, each tool was run on default settings of the algorithms, using 38MB sized MGF file containing 5000 spectra and all 14 of databases, 5 times for each database size (pFind and Peaks were run only once since they don't have any command line executables). Processes that took more than 5 hours were not evaluated. Inspect, Myrimatch, MSAmanda were run on Ubuntu 15.04, XTandem, OMSSA, MS-GF+, pFind and PEAKS were run on Windows systems respectively.

2.3.2. Accuracy and Score Comparisons

6 Frame translated Human chromosomes 1, 2, 11, 20, MT and Ensembl human protein database were used for accuracy and score comparisons. Unknown amino acids ("X") were removed from all databases and any character sequence that matches our synthetic peptides were replaced with Alanine ("A") characters. Synthetic peptide replacement was done in order to prevent unfair score comparisons between databases due to different peptide candidate counts (Table :2.2).

Table 2.2. Database sources and sizes for accuracy and score comparisons.

Contig Composition	Raw Size	Split Size
Human Chr1	456.2 MB	552.9 MB
Human Chr2	482.4 MB	585.6 MB
Human Chr11	265.5 MB	322.2 MB
Human Chr20	120.5 MB	146.2 MB
Human ChrMT	33.6 KB	40.5 KB
Human Protein Database	35.7 MB	43.5 MB

There are two case tests for score and accuracy comparisons. In one case, the synthetic peptides were distributed equally amongst both split and merge versions of databases as a single entry, considering the spectral data is obtained from multiple sources. After the peptide distribution, split and merge versions of the databases are searched for the spectral data. Based on these algorithm runs, FDR method was used for further data analysis. Another method called Second Run Search (SRS) was also used. In SRS method, the resulting PSM's of the first algorithm search were brought together to form a database entry for each spectrum. Which means that if a spectrum was assigned to peptides in 4 databases, an entry was created formed by these peptides and searched again. This case involves accuracy comparisons for FDR, SRS and SM.

In another case, all of the peptides are put in all databases as a single entry. The aim here was to place the correct identifications to first rank by removal of all competitive peptides and compare score differences. All the peptides that came with a higher score than the correct identification in the previous run were removed from database and search took place again for two competitive removal rounds (cr2). After placing the correct identification to first rank, the score differences between different size databases were shown to demonstrate the effects of database sizes to scoring functions.

The algorithm settings were adjusted as following: precursor mass tolerance 1.5 Da, fragment mass tolerance 0.4 Da, trypsin cleavage with maximum 2 missed cleavage allowance, monoisotopic mass. There were no post-translation modifications set as fixed or variable.

CHAPTER 3

RESULTS

This chapter involves the results of tool limitations and speed comparisons between split and raw databases as well as accuracy and score comparisons between SM, FDR and SRS methods.

3.1. Tool Limitations and Speed Assessment

Selected tools were subjected to limitation test which was done by using raw databases (sequence under one identifier) size up to 500 MB. Only Peaks program demonstrated dependence to spectra size limited to 5000 spectra, the other tool limits were irrelevant to spectral data (Table :3.1).

Table 3.1. Contig size limit for selected database search algorithms.

Algorithm	Raw Size Limit
MSAmanda	10 MB
Inspect	5 MB
XTandem	1 MB
Peaks	260 MB
pFind	24 MB
Myrimatch	No Limitation
OMSSA	No Limitation
MS-GF+	No Limitation

Split versions of these raw databases however, demonstrated none of the limitations above up to 600 MB of databases but caused many algorithms to crash, when spectral data and database size was increased up to 150k and 7.5 GB respectively, proving the need of separating entries when enormous data search is required (Table :3.2).

Table 3.2. Database size: 7.5 GB; 150k spectra PC: 64GB RAM; 8 cores

Database Search Algorithm	Run Report
MsAmanda	Crashes after two weeks with disk full problem (>20GB)
MS-GF+	Crashes after 2 hours due to memory problem (when loading db)
Omssa	Crashes after several hours due to Cthread: WrapperError
pFind	Doesn't end but no errors given.
XTandem	Crashes after a 1.5 weeks at the end of run ("unanticipated cleavage")
Inspect	Crashes within minutes trying to index database
Peaks (commercial version)	Finishes after 1.5 hours (free version may differ)
Myrimatch	Finishes after 11 days

For speed assesment, raw and split sized databases, mentioned in Table 2.1., are searched for 5000 spectra on default settings. As seen in figure 3.1 and figure 3.2 rapidness of the tools are increased at least two folds for all algorithms except MS-GF+ when used split databases.

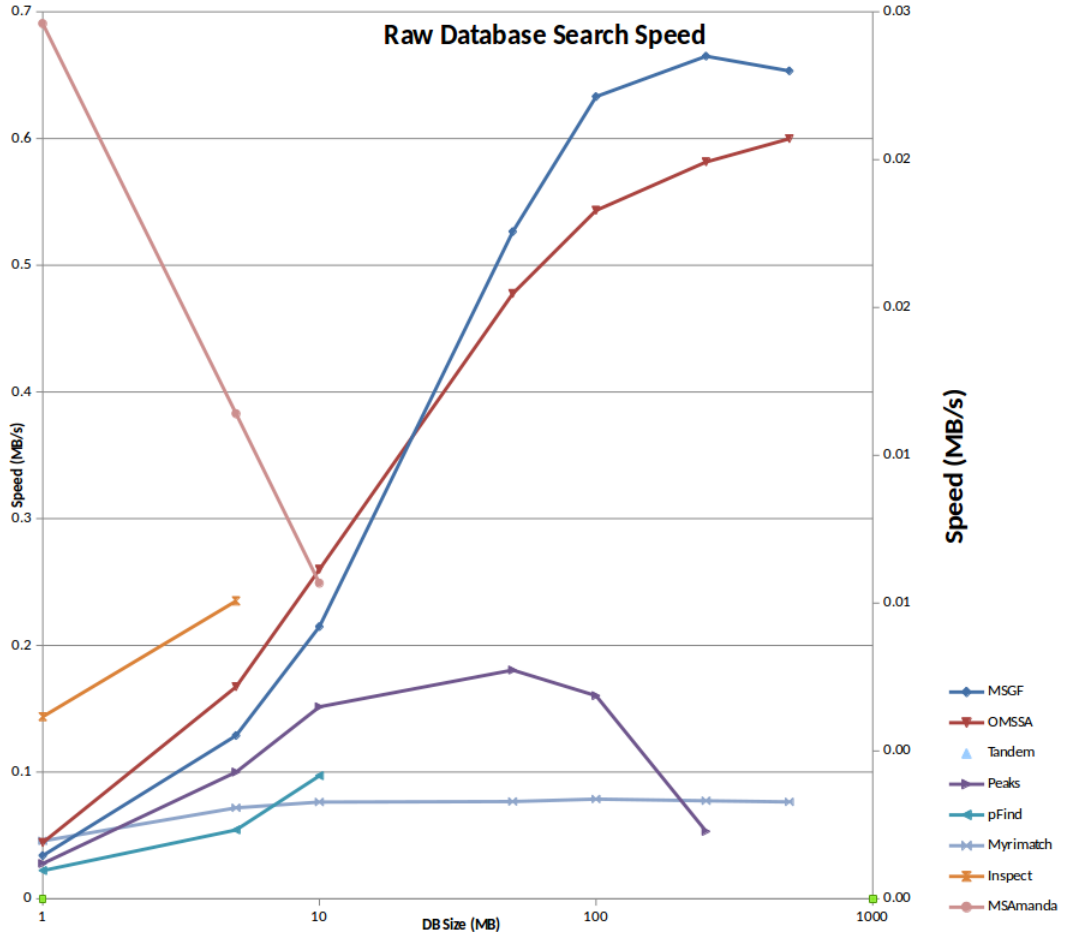

Figure 3.1. Speed assesment of the given algorithms. Using raw databases up to 500 MB, MB of database processed per second is shown in the figure. Given that some of the algorithms (Inspect and MSAmanda) are too slow compared to other algorithms, their speed values are shown in secondary y axis at the right side of the figure. Except for Myrimatch, MS-GF+ and Omssa, tools are limited to certain size of raw databases as shown in table 3.1

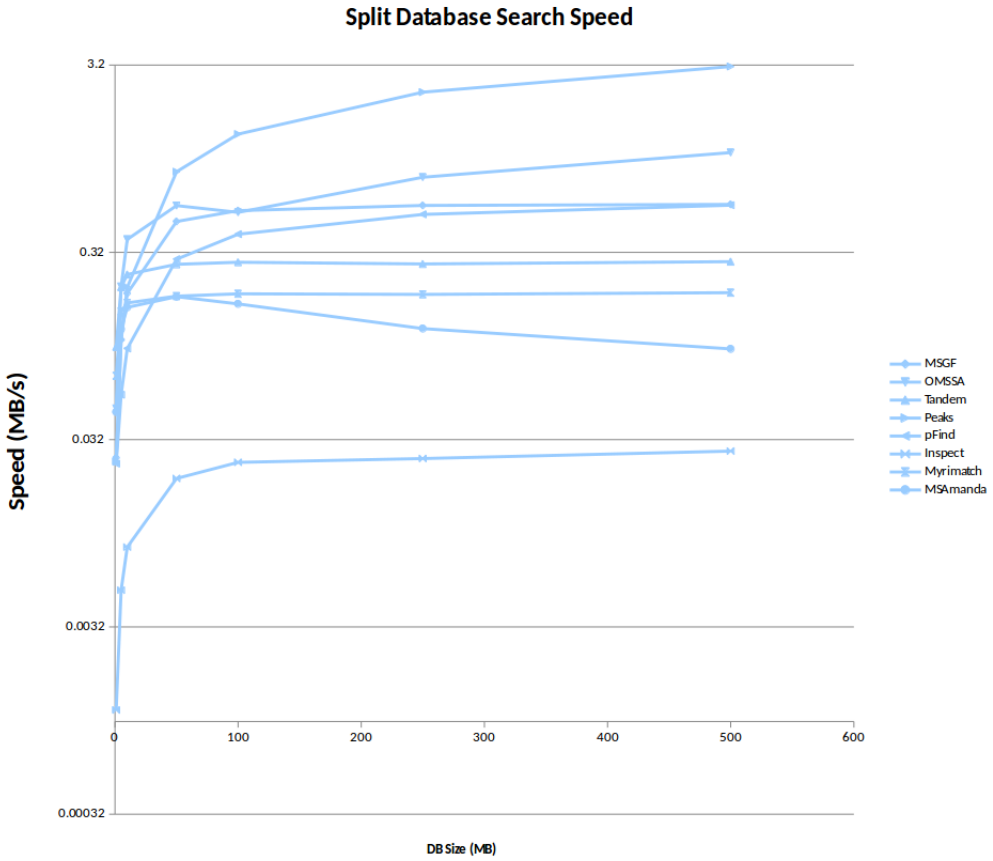

Figure 3.2. Speed assesment of the algorithms searched on split databases. All of the tools could process up to 500 MB of database and their speeds were much higher than its raw versions.

3.2. Accuracy Comparisons of Different Methodologies

For accuracy comparisons, default scoring systems of all algorithms are used to provide correct identification percentages for all methodologies (SM,FDR,SRS). As seen in figure 3.3 SM method towers above FDR and SRS methods for all algorithms. The correct identification percentages are acquired by dividing the correct identification counts in all databases by the spectra count.

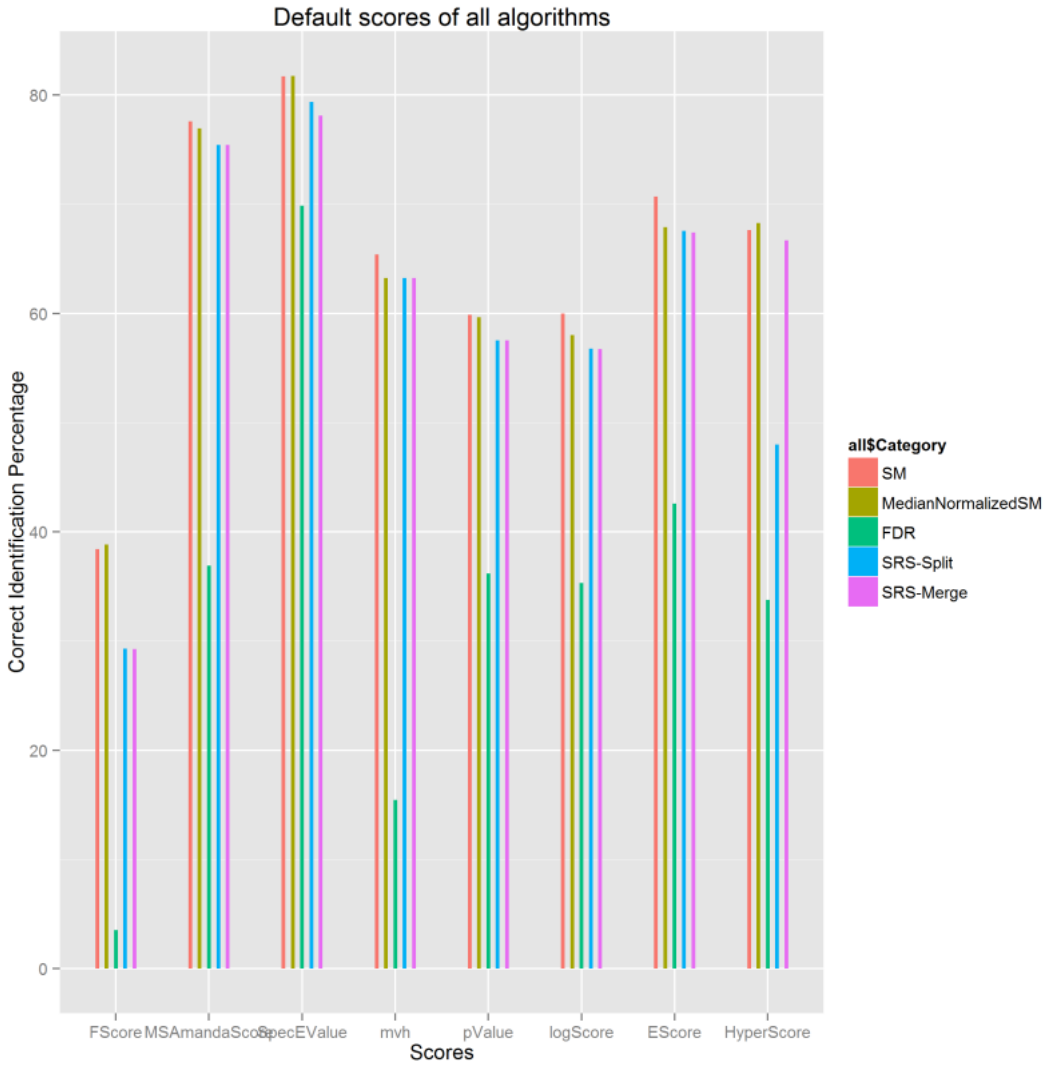

Figure 3.3. Method comparisons of all algorithms. From left to right the scoring systems belongs to; Inspect, MSAmanda, MS-GF+, Myrimatch, Omssa, Peaks, pFind, XTandem.

In respect of different scoring systems of algorithms, an example is given in figure 3.4 which represents accuracy comparisons of methodologies for all scoring systems belongs to Inspect search tool.

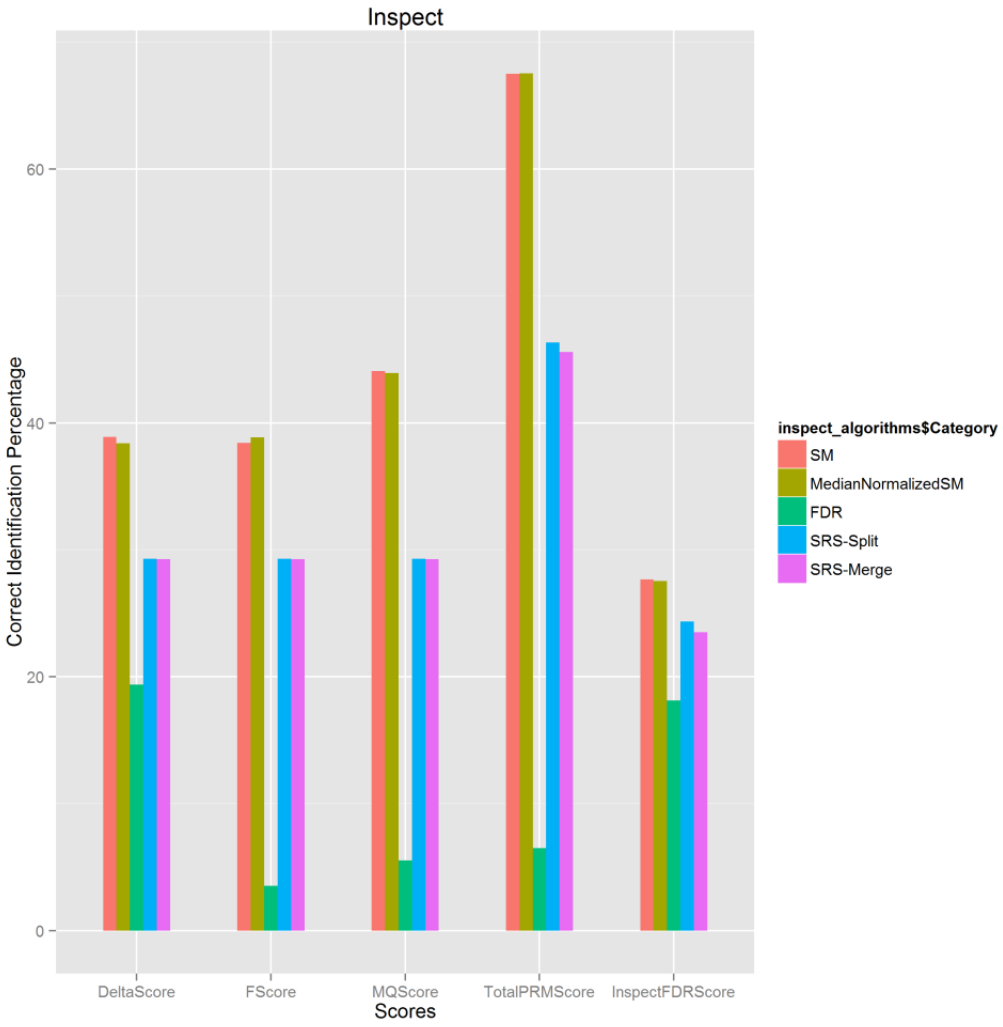

Figure 3.4. Comparison of methodologies for different scoring functions of Inspect algorithm

3.3. Score Comparisons Between Different Sized Databases by Competitive Removal

The main aim of competitive removal (cr) step was to carry all of our correct peptides to first rank in order to compare their scores fairly for all databases. In order to achieve this the wrongly identified peptides that had better scores than the correct peptides were removed from databases.

Two cases were represented in figures 3.5 and 3.6. In figure 3.5 score differences belonging to correctly identified peptides in all SM databases were compared to score differences of all correctly identified peptides of Human Chromosome 1 database and Human Chromosome MT (relatively big and small sized databases).

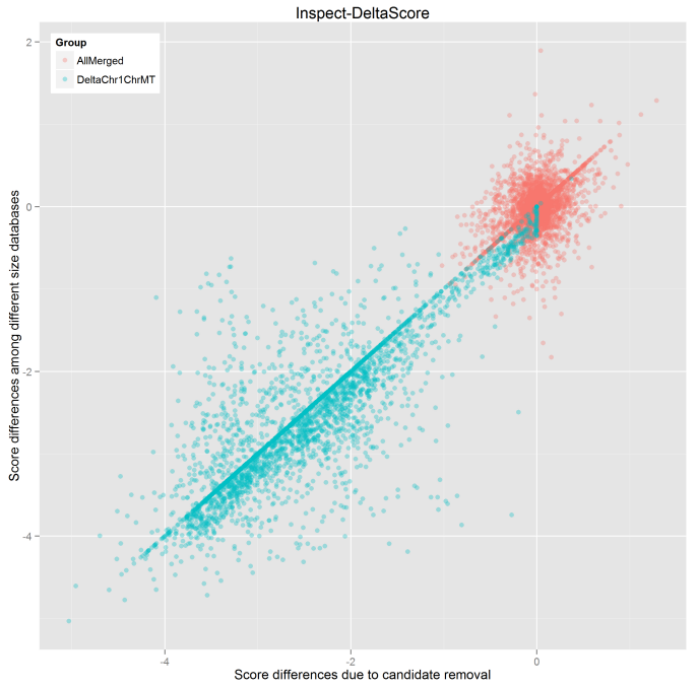

Figure 3.5. Scatter plot of the score differences between merged databases (shown in red dots) and Human Chromosome 1 and Human Chromosome MT. While the score differences are clustered around 0 which indicates that the correctly identified peptides have similar scores in all databases when SM is used, the score differences between the correctly identified peptides in Human Chromosome 1 and Human Chromosome MT varies from 0 to -4 which clearly indicates the size effects of databases towards scoring functions.

In figure 3.6 a medium sized database (Human Chromosome 11) was used to compare the score differences between the raw (Split) database and the same database after the competitive removal step to further prove that the candidate peptide counts effects the scoring functions of database search algorithm.

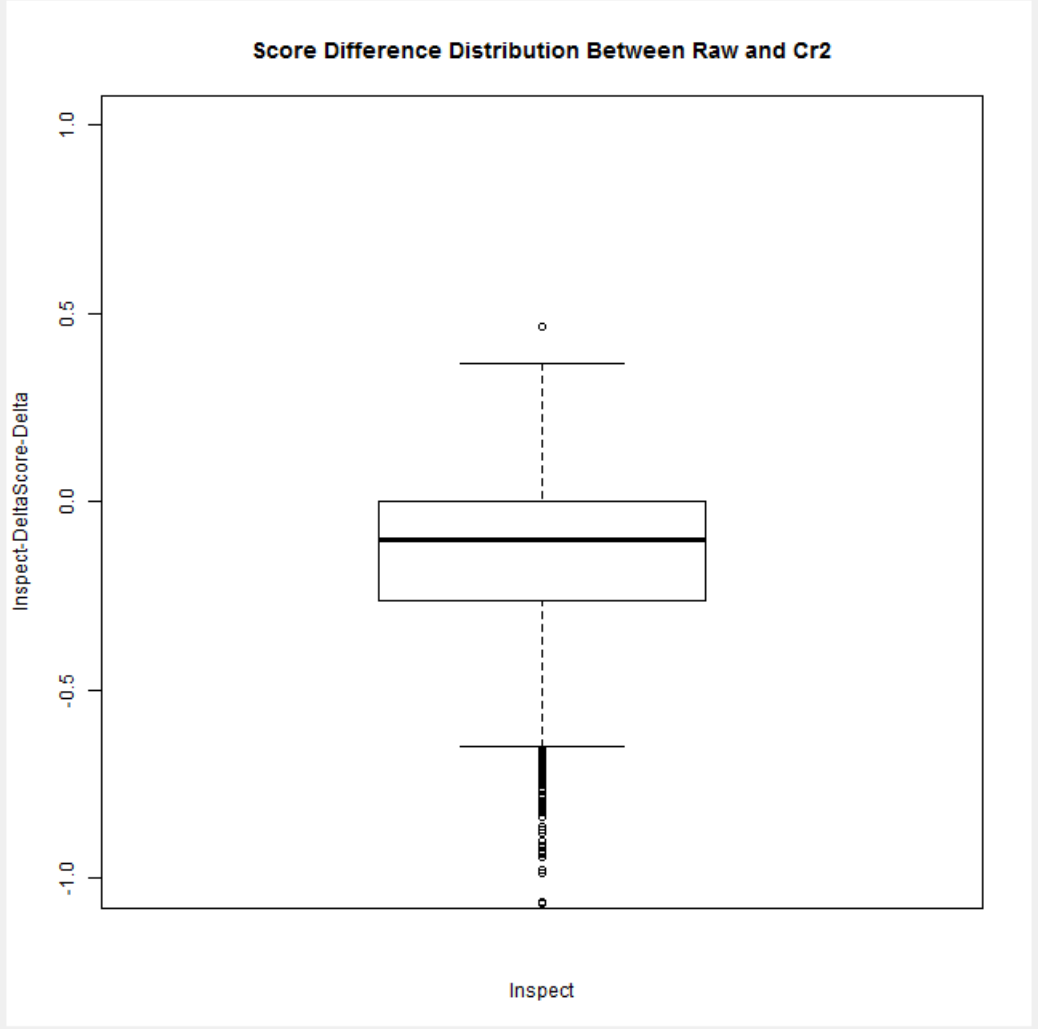

Figure 3.6. Box plot of the score differences between the raw and cr2 induced Human Chromosome 11 database. Since the peptide candidates are abundantly present in raw database, the score assigned to correctly identified peptides are lower.

In figure 3.7 score differences between the peptide sequence matches of split and merge databases and between chromosome 1 and chromosome MT databases were shown. As seen in the figure, the differences are much smaller in split and merged databases which demonstrates the normalization effect.

Score Difference on IQR Before and After Competitive Removal

After First Competitive Removal Round

		All S-M DBs			Chr1 vs MT	
Algorithm	**Score**	**IQR**			**IQR**	
MSGF	Evalue	4.99E-06			0.01139	
pFind	ExpectScore	2.38E-08			2.9E6	
Myrimatch	mvh	0.00			0.00	
OMSSA	EValue	0.000480664			6.7823	
Inspect	DeltaScore	0.287			1.236	
Tandem	ExpectScore	0.001			0.04	
MSAmanda	MSAmandaScore	0.00			0.00	

After Second Competitive Removal Rounds

		All S-M DBs			Chr1 vs MT	
Algorithm	**Score**	**IQR**			**IQR**	
MSGF	Evalue	1.33E-05			0.03349	
pFind	ExpectScore	4.05E-09			0.00036	
Myrimatch	mvh	0.00			0.00	
OMSSA	EValue	0.002265773			16.7425	
Inspect	DeltaScore	0.19175			1.309	
Tandem	ExpectScore	0.001715			0.20376	
MSAmanda	MSAmandaScore	0.00			0.00	

Figure 3.7. Score differences of the peptide sequence matches between the same sized databases and different sized databases. Scoring algorithms depending on the candidate peptide counts are significantly closer to zero in the split and merged databases than raw databases.

CHAPTER 4

CONCLUSION

In proteomics mass spectrometry is the most used technique for identifying proteins. Database search method is one of the main strategies to analyze mass spectrometric data. Due to enormous amount of data provided by mass spectrometers and the database sizes that can haul up to gigabytes, splitting the databases in order to increase speed and accuracy of database search algorithms has become crucial.

Herein a new methodology named Split and Merge is presented to better analyze large and different sized databases. The first step of the method involves splitting the databases to entries by dividing them to desired amino acid sequence length and at the same time, preserve the possible candidate peptides by keeping track of the overlaps between splitting points.

After enabling the database search algorithms to utilize large databases, merging step of the different sized databases to a single, desired sized of multiple databases is done to prevent wrong identifications through scoring deficiencies caused by high\low peptide candidate counts. The merging step allowed database search algorithms to increase the accuracy of the PSMs above commonly used methods such as FDR and SRS when using different sized databases.

CHAPTER 5

FURTHER WORK

Since the database search algorithms have different scoring methods and the accuracy of the peptide identification proccess is dependent on the spectral quality, there are two tasks that need to be done. First the analysis should be done using one workflow, organised by the superior parts of the currently available database search algorithms. To achieve this task separate parts of the tools should be joined into one tool that could use the best part of each algorithm. After creating the algorithm, it must be optimized to proccess spectra that is generated by different types of machines which may be done using genetic algorithms to generate optimal settings for peptide identification.

REFERENCES

Aebersold, R. and M. Mann (2003). Mass spectrometry-based proteomics. *Nature 422*(6928), 198–207.

Allmer, J. (2011). Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert review of proteomics 8*(5), 645–657.

Allmer, J. (2012). A Call for Benchmark Data in Mass Spectrometry-Based Proteomics. *Journal of Integrated OMICS*.

Bakhtiar, R. and Z. Guan (2005). Electron capture dissociation mass spectrometry in characterization of post-translational modifications. *Biochemical and biophysical research communications 334*(1), 1–8.

Barton, S. J., J. C. Whittaker, S. G. Hospital, K. Street, M. Spec, and I. Introduction (2009). REVIEW OF FACTORS THAT INFLUENCE THE ABUNDANCE OF IONS PRODUCED IN A TANDEM MASS SPECTROMETER AND STATISTICAL METHODS FOR DISCOVERING THESE FACTORS. *Mass Spectrometry Reviews*, 177–187.

Beavis, R. C. and B. T. Chait (1990). Rapid, sensitive analysis of protein mixtures by mass spectrometry. *Proceedings of the National Academy of Sciences 87*(17), 6873–6877.

Biemann, K. (1992). Mass spectrometry of peptides and proteins. *Annual review of biochemistry 61*(1), 977–1010.

Bökelmann, V., B. Spengler, and R. Kaufmann (1995). Dynamical parameters of ion ejection and ion formation in matrix-assisted laser desorption/ionization. *Eur. Mass Spectrom 27*, 156–158.

Burlingame, A., R. K. Boyd, and S. J. Gaskell (1994). Mass spectrometry. *Analytical chemistry 66*(12), 634R–683R.

Chait, B. T. (2006). Mass spectrometry: Bottom-up or top-down? *Science 314*(5796), 65–66.

Chait, B. T. and S. B. Kent (1992). Articles. weighing naked proteins: Practical, high-accuracy mass measurement of peptides and proteins. *Science 257*, 1T.

Chen, Y., S. W. Kwon, S. C. Kim, and Y. Zhao (2005). Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *Journal of proteome research 4*(3), 998–1005.

Cohen, S. L. and B. T. Chait (1996). Influence of matrix solution conditions on the maldi-ms analysis of peptides and proteins. *Analytical chemistry 68*(1), 31–37.

Covey, T. R., R. F. Bonner, B. I. Shushan, J. Henion, and R. Boyd (1988). The determination of protein, oligonucleotide and peptide molecular weights by ion-spray mass spectrometry. *Rapid Communications in Mass Spectrometry 2*(11), 249–256.

Craig, R. and R. C. Beavis (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid communications in mass spectrometry 17*(20), 2310–2316.

Din, S., A. M. Lennon, I. D. Arnott, T. Hupp, and J. Satsangi (2007). Technology insight: the application of proteomics in gastrointestinal disease. *Nature Clinical Practice Gastroenterology & Hepatology 4*(7), 372–385.

Domon, B. and R. Aebersold (2006, apr). Mass spectrometry and protein analysis. *Science (New York, N.Y.) 312*(5771), 212–7.

Dorfer, V., P. Pichler, T. Stranzl, J. Stadlmann, T. Taus, S. Winkler, and K. Mechtler (2014). MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research 13*, 3679–3684.

Edman, P. (1950). Method for determination of the amino acid sequence in peptides. *Acta chem. scand 4*(7), 283–293.

Eidhammer, I., K. Flikka, L. Martens, and S. Mikalsen (2008a). *Fundamentals of mass spectrometry*, pp. 74. Wiley.

Eidhammer, I., K. Flikka, L. Martens, and S. Mikalsen (2008b). *Fundamentals of mass spectrometry*, pp. 66. Wiley.

Eidhammer, I., K. Flikka, L. Martens, and S. Mikalsen (2008c). *Tandem MS or MS/MS analysis*, pp. 120. Wiley.

Eidhammer, I., K. Flikka, L. Martens, and S. Mikalsen (2008d). *Tandem MS or MS/MS analysis*, pp. 124. Wiley.

Eidhammer, I., K. Flikka, L. Martens, and S. Mikalsen (2008e). *Tandem MS or MS/MS analysis*, pp. 70. Wiley.

Elias, J. E. and S. P. Gygi (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods 4*(3), 207–214.

Frese, C. K., A. M. Altelaar, H. van den Toorn, D. Nolting, J. Griep-Raming, A. J. Heck, and S. Mohammed (2012). Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Analytical chemistry 84*(22), 9668–9673.

Frese, C. K., H. Zhou, T. Taus, A. M. Altelaar, K. Mechtler, A. J. Heck, and S. Mohammed (2013). Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (ethcd). *Journal of proteome research 12*(3), 1520–1525.

Gatlin, C. L. and F. Turecek (1994). Acidity determination in droplets formed by electrospraying methanol-water solutions. *Analytical Chemistry 66*(5), 712–718.

Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant (2004). Open mass spectrometry search algorithm. *Journal of proteome research 3*(5), 958–964.

Guilhaus, M. (1995). Special feature: Tutorial. principles and instrumentation in time-of-flight mass spectrometry. physical and instrumental concepts. *Journal of Mass Spectrometry 30*(11), 1519–1532.

Han, X., M. Jin, K. Breuker, and F. W. McLafferty (2006). Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science 314*(5796), 109–112.

Henzel, W. J., T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley, and C. Watanabe (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences 90*(11), 5011–5015.

Hu, Q., R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks (2005). The orbitrap: a new mass spectrometer. *Journal of mass spectrometry 40*(4), 430–443.

Ikonomou, M. G., A. T. Blades, and P. Kebarle (1991). Electrospray-ion spray: a comparison of mechanisms and performance. *Analytical Chemistry 63*(18), 1989–1998.

Karas, M. and F. Hillenkamp (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry 60*(20), 2299–2301.

Keshishian, H., T. Addona, M. Burgess, D. Mani, X. Shi, E. Kuhn, M. S. Sabatine, R. E. Gerszten, and S. A. Carr (2009). Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Molecular & cellular proteomics 8*(10), 2339–2349.

Kilby, P. M. (2007a). *Protein Identification by Peptide Mass Fingerprinting*, Volume 21, pp. 61. Wiley Online Library.

Kilby, P. M. (2007b). *Protein Identification by Peptide Mass Fingerprinting*, Volume 21, pp. 5. Wiley Online Library.

Kim, S. and P. A. Pevzner (2014). Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications 5*.

Köcher, T., G. Allmaier, and M. Wilm (2003). Nanoelectrospray-based detection and sequencing of substoichiometric amounts of phosphopeptides in complex mixtures. *Journal of mass spectrometry 38*(2), 131–137.

Kussmann, M., E. Nordhoff, H. Rahbek-Nielsen, S. Haebel, M. Rossel-Larsen, L. Jakobsen, J. Gobom, E. Mirgorodskaya, A. Kroll-Kristensen, L. Palm, et al. (1997). Matrix-assisted laser desorption/ionization mass spectrometry sample preparation techniques designed for various peptide and protein analytes. *Journal of Mass Spectrometry 32*(6), 593–601.

Küster, B., P. Mortensen, J. S. Andersen, and M. Mann (2001). Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics 1*(5), 641–650.

Li, D., Y. Fu, R. Sun, C. X. Ling, Y. Wei, H. Zhou, R. Zeng, Q. Yang, S. He, and W. Gao (2005). pfind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics 21*(13), 3049–3050.

Ma, B., K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie (2003). Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry 17*(20), 2337–2342.

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical chemistry 72*(6), 1156–1162.

Mann, M., R. C. Hendrickson, and A. Pandey (2001, jan). Analysis of proteins and proteomes by mass spectrometry. *Annual review of biochemistry 70*, 437–73.

McHugh, L. and J. W. Arthur (2008). Computational Methods for Protein Identification from Mass Spectrometry Data. *PLoS Computational Biology 4*(2), 12.

Mirza, U. A. and B. T. Chait (1994). Effects of anions on the positive ion electrospray ionization mass spectra of peptides and proteins. *Analytical chemistry 66*(18), 2898–2904.

Molina, H., R. Matthiesen, K. Kandasamy, and A. Pandey (2008). Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Analytical Chemistry 80*(13), 4825–4835.

Nesvizhskii, A. I. and R. Aebersold (2004). Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug discovery today 9*(4), 173–181.

Nesvizhskii, A. I. and R. Aebersold (2005). Interpretation of shotgun proteomic data the protein inference problem. *Molecular & Cellular Proteomics 4*(10), 1419–1440.

Nesvizhskii, A. I., O. Vitek, and R. Aebersold (2007, oct). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods 4*(10), 787–97.

Olsen, J. V., B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann (2007). Higher-energy c-trap dissociation for peptide modification analysis. *Nature methods 4*(9), 709–712.

Paizs, B. and S. Suhai (2005). Fragmentation pathways of protonated peptides. *Mass spectrometry reviews 24*(4), 508–548.

Pappin, D. J., P. Hojrup, and A. J. Bleasby (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current biology 3*(6), 327–332.

Patterson, S. D. and R. Aebersold (1995). Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis 16*(1), 1791–1814.

Roepstorff, P. and J. Fohlman (1984). Letter to the editors. *Biological Mass Spectrometry 11*(11), 601–601.

Schorlemmer, M., J. Abián, C. Sierra, D. de la Cruz, L. Bernacchioni, E. Jaén, A. Perreau de Pinninck, and M. Atencia (2012, jan). P2P proteomics – data sharing for enhanced protein identification. *Automated experimentation 4*(1), 1.

Shadforth, I., D. Crowther, and C. Bessant (2005, nov). Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics 5*(16), 4082–95.

Steen, H. and M. Mann (2004, sep). The ABC's (and XYZ's) of peptide sequencing. *Nature reviews. Molecular cell biology 5*(9), 699–711.

Syka, J. E., J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America 101*(26), 9528–9533.

Tabb, D. L., C. G. Fernando, and M. C. Chambers (2007). Myrimatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research 6*(2), 654–661.

Tanner, S., H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna (2005). Inspect: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical chemistry 77*(14), 4626–4639.

Tyers, M. and M. Mann (2003). From genomics to proteomics. *Nature 422*(6928), 193–197.

Uversky, V. N. and A. K. Dunker (2010). Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1804*(6), 1231–1264.

Vorm, O., P. Roepstorff, and M. Mann (1994). Improved resolution and very high sensitivity in maldi tof of matrix surfaces made by fast evaporation. *Analytical Chemistry 66*(19), 3281–3287.

Wells, J. M. and S. A. McLuckey (2005). Collision-induced dissociation (cid) of peptides and proteins. *Methods in enzymology 402*, 148–185.

Whitehouse, C. M., R. Dreyer, M. Yamashita, and J. Fenn (1989). Electrospray ionization for mass-spectrometry of large biomolecules. *Science 246*(4926), 64–71.

Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, and M. Mann (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature 379*(6564), 466–469.

Woo, C. M., A. T. Iavarone, D. R. Spiciarich, K. K. Palaniappan, and C. R. Bertozzi (2015). Isotope-targeted glycoproteomics (isotag): a mass-independent platform for intact n-and o-glycopeptide discovery and analysis. *Nature methods 12*(6), 561–567.

Xu, C. and B. Ma (2006). Software for computational peptide identification from ms–ms data. *Drug Discovery Today 11*(13), 595–600.

Yates, J. R. (2000). Mass spectrometry from genomics to proteomics. *Outlook 16*(1), 5–8.

Zenobi, R. and R. Knochenmuss (1998). Ion formation in maldi mass spectrometry. *Mass Spectrometry Reviews 17*(5), 337–366.

Zubarev, R. A., N. L. Kelleher, and F. W. McLafferty (1998). Electron capture dissociation of multiply charged protein cations. a nonergodic process. *Journal of the American Chemical Society 120*(13), 3265–3266.

Zubarev, R. A., A. R. Zubarev, and M. M. Savitski (2008). Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet? *Journal of the American Society for Mass Spectrometry 19*(6), 753–761.