ENHANCEMENT AND VALIDATION OF CURRENT HUMAN GENOME ANNOTATION VIA NOVEL PROTEOGENOMICS ALGORITHMS

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

DOCTOR OF PHILOSOPHY

in Molecular Biology and Genetics

by
Canan HAS

December 2016
İZMİR

We approve the thesis of **Canan Has**

Assoc. Prof. Dr. Jens ALLMER
Supervisor

Prof. Dr. Anne FRARY
Committee Member

Assist. Prof. Dr. Yavuz OKTAY
Committee Member

Prof. Dr. Talat YALÇIN
Committee Member

Prof. Dr. Şermin GENÇ
Committee Member

27 December 2016

Prof. Dr. Volkan SEYRANTEPE
Head of the Department of Molecular
Biology and Genetics

Prof. Dr. Aysun SOFUOĞLU
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGEMENTS

There are many people who contributed a lot into my life and helped me to be more mature by time. Some of them made me so happy, some of them broke my heart badly. First of all I would like to thank all of them from the bottom of my heart for these experiences in my life, especially in my PhD studies.

Dr. Jens Allmer, we had tough times, good times. We shared success and failure. I have been changed a lot as a person and a researcher during my education under your supervision. I am so grateful for the chance to work in your lab.

Dr. Anne Frary, you are one of the role models who inspired me with your humanity at first then with your scientist identity. It was an honor to meet you, to work with you. Dr. Sami Doğanlar, I feel more confident in respect to my career upon your encouragement. Thank you very much for all of your support.

I would like to thank Dr. Talat Yalçın, Dr. Yavuz Oktay, Dr. Şermin Genç for being my thesis committee member and more importantly for their valuable comments regarding my studies and my future career plans. I owe special thanks to Dr. Ayten Nalbant, Dr. Bünyamin Akgül, and Dr. Çağlar Karakaya. All of them were always supportive in the milestones of my career; getting research assistantship position, graduation, planning my post-graduation, etc. As a molecular biologist, it was a big challenge for me to pursue a career in bioinformatics. IYTE Computer Engineering Department faculties and Dr. Bilge Karaçalı believed in my passion on bioinformatics and accepted me to the informatics classes. I will always appreciate that.

During Ph.D., I had an unforgettable experience in Berlin with the support of DAAD. I am so happy to meet AGABI group at Freie Universitaet led by Dr. Knut Reinert and to have a chance to work with OpenMS team. I am also grateful to KNIME Data Analytics Berlin Team and Konstanz Team for having me as an intern. They allowed me to learn insights of KNIME and I became a fan of data science in general, after my time there.

I would like to thank all of my colleagues at IYTE MBG Department for their kindness and support during my journey and to colleagues at IYTE Proteomics Facility who taught me the wetlab part of proteomics. I should also state my undergrad students. During my six years of TA experience, they gave me one of the biggest pleasures of my life: being able to transfer my knowledge to others.

Aslı Kartal: My fellow. I am blessed to have you in my life. You are one of the people who witnessed my life journey. As in the poem of Gülten Akın: “Gün küçük dağların ardından ve yolumuz var daha. Herşey olgunlaşır, çürüyüp dökülür zincir. En güzeli, yol yürüyüş öğretir. Dostum, eskimeyen arkadaşım.”

Funda Akıncıoğlu, Özge Yoluk: unforgettable supports of my life. No words would be sufficient enough to express the importance of your existences in my life. With the hope of having you during my lifetime.

I should also state my thanks to current and former IYTE-Bioinformatics Lab members. Each of you has a contribution to my life. I learnt a lot from each of you. Each of you kept my back in hard times. Each of you did something to make me smile. I always felt your support behind me. I will always keep your friendships in my heart.

IYTE introduced several nice people to my life, however, two of them are really precious. Özgün Öykü Özdemir, Deniz Keman: Always in my heart and in my thoughts. In the tough times; you are wisperhing the song in Tangled “I ‘ve got a dream”.. I wish our friendship will continue with more memories.

Who would say that I would have a chance to work with my old friends from Istanbul University, currently members of Frary/Doganlar Lab. Elvan Başkurt Öztürk, Süleyman Öztürk and İbrahim Çelik. Guys, you rock the plant world!

The credit also goes to Oliver Christoph; my boyfriend, one of my best friends, wisdom “wait" sound of my life and one of the coolest people I have ever met.

To my beloved family: Çok şanslıyım ki sizler benim ailemsiniz. Has ve Peker ailelerinin üyesi olmak, en çok sizin tarafınızdan sevilmek, en çok sizi sevmek, evlat-abla-kardeş-yeğen-kuzen-arkadaş bir sürü şey olmamız benim bu hayattaki en büyük teşekkürüm. Bu tez ve tüm başarılar bizimdir, hepimizin hayali, hepimizin çabasıdır. Sizleri çok seviyorum.

This project was supported by TUBITAK project grant number 114Z177.

This thesis and everything I did and will do in my life is dedicated to;

My grandparents,

Şerife-Ali PEKER and Saime-Hasan HAS

whom I owe so much than words can explain.

My parents Azize-Erdal Has

whom encouraged me for the academic career and being first discoverers of my potential.

My big family

and

The seven-years-old girl who started everything by being so much proud of herself when she became the first student whose apple turned to dark red within first few weeks of the 1st class as the symbol of learning how to read and write

ABSTRACT

ENHANCEMENT AND VALIDATION OF CURRENT HUMAN GENOME ANNOTATION VIA NOVEL PROTEOGENOMICS ALGORITHMS

Proteogenomics includes the transfer of knowledge from proteomics to genomics and vice versa. To have high confidence in the information transferred it is essential that it be based on experimental results. Genomics is currently fueled by high throughput techniques involving next generation sequencing. Proteomics is based on mass spectrometry (MS) which is also a high throughput approach. Both fields are generating a wealth of data which needs to be correlated and annotated to generate knowledge.

Publicly available human blood plasma mass spectrometric data exist for samples in data repositories such as PeptideAtlas, PRIDE. We acquired high-quality collections from this data and stored it in a custom database developed by us. First, we aimed to amend this data by employing a proteogenomic pipeline PGMiner developed in this study against a custom sequence database which includes all predicted alternative open reading frames as well as the six-frame translation of the human genome and exosome. Then, we correlated the existing annotations with the available mass spectrometric measurements. The human genome in tandem with currently available genome annotations from HAVANA and ENSEMBL enabled us to validate and enhance current gene annotations.

ÖZET

VAR OLAN İNSAN GENOM ANOTASYONUNUN YENİ PROTEOGENOMİK YÖNTEMLER İLE DOĞRULANMASI VE GELİŞTİRİLMESİ

Proteogenomik protemikten genomik alanına veya genomikten proteomik alanına bilginin transferini içerir. İki alanda bilgi üretmek için kimliklendirilmesi ve ilişkilendirilmesi gereken büyük sayıda veri ortaya koyar. Genomik çalışmalarla üretilen verilerin kimliklendirilmesi amaçlanır ve bu kimliklendirmede yüksek güvenilirlik elde etmek için deneysel tekniklerle translasyon düzeyinde doğrulama yapılması şarttır. Genomik yeni nesil dizileme yöntemini içeren yüksek-ölçekli yöntemlerle elde edilirken, proteomik verileri yine yüksek-ölçekli veri üreten bir yöntem olan kütle spektrometreden elde edilir.

PeptideAtlas, PRIDE gibi çeşitli veri bankalarında açık kaynak insan kan plazma dokusuna ait kütle spektrometre verisi mevcuttur. Bu veriler arasından elde edilecek yüksek kaliteli koleksiyonlar geliştireceğimiz veritabanında depolanmıştır. Bu proje kapsamında ilk gerçekleştirilen amaç spektral verileri bu çalışma kapsamında geliştirilen PGMiner akış algoritması kullanarak insan genomunun 6-çerçeve translasyonu, eksozom ve tüm tahmin edilmiş alternatif açık okuma çerçevelerini kapsayan veritabanlarına karşı aranmış ve spektral verilerin hangi peptitlere ait olduğunu anlamlandırılmıştır. Daha sonra var olan gen ve protein anotasyonları ile ilk aşamada peptit tanımlaması yapılan kütle spektrometre ölçümleri ilişkilendirilmiştir. HAVANA ve ENSEMBL’dan elde edilen var olan genom anotasyonları ile mevcut gen anotasyonları doğrulanmış ve geliştirilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
PSM	Peptide-Spectrum Match
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
FDR	False-Discovery Rate
CID	Collision Induced Dissociation
HCD	High-energy Collision Induced Dissociation
ETD	Electron Transfer Dissociation
altORF	Alternative Open Reading Frame
NCBI	National Center for Biotechnology Information
GTF	General Transfer Format
GFF	Generic Feature Format
SNP	Single Nucleotide Polymorphism

CHAPTER 1

MASS SPECTROMETRY-BASED PROTEOMICS AND PROTEOGENOMICS

1.1.Introduction

The proteome of an organism is the entirety of proteins and their isoforms that can be generated from the underlying genome. Proteomics deals with the identification, sequencing, quantification, and other issues related to the proteome. In proteomics, mass spectrometry has become the tool of choice for many areas of proteomics such as peptide identification and sequencing. Recent advances in mass spectrometry-based proteomics have resulted in *n* increasing amount of freely available mass spectrometric data in public databases. The general focus of the data in these MS-databases is protein identification from known protein databases. With the aid of rapid improvements in NGS technology, custom sequence databases can be built by using six- or three-frame translated DNA or RNA sequences. Additionally, available protein sequences predicted gene models and their derivatives such as alternative spliced forms, exon-exon junction peptides, alternative translation products and single-nucleotide polymorphic sequence variants can be used as databases for computational proteomics search.

Searching MS/MS spectra against custom sequence databases generated using genomic and transcriptomic data, and their variants enable refinement of gene models and provide quantitative information on protein expression. This research field is so-called proteogenomics. Within proteogenomics proteomics findings feed annotation of the genomic data while findings in genomics generate possible proteins that can be composed a large search space in proteomics.

1.2.Mass Spectrometry-Based Proteomics

The goal of proteomics is to identify and characterize proteins in a given sample to clarify their sequences, structures, functions, interactions, and subcellular localizations (Aebersold and Mann 2003). In recent years, mass spectrometry (MS) has become the tool of choice, enabling high throughput measurement of peptides and proteins with much accuracy and sensitivity (McHugh and Arthur 2008). In a typical MS-based proteomics experiment, peptides are obtained via enzymatic cleavage from proteins isolated from the sample to be analyzed. The resulting peptides are subjected to liquid chromatography coupled mass spectrometer (LC-MS) (Derrick and Patterson 2001). Here, peptide molecules are ionized and subjected to gas phase; hence ionized peptides are separated according to their mass to charge (m/z) ratios. This first step produces information on the molecular weights of the peptides and is called as MS level 1. However, the molecular weight is not sufficient enough to identify a peptide, therefore proteins and in turn a protein. Therefore, selected ions are subjected to a fragmentation process followed by the second stage of MS. This process is called tandem mass spectra (MS/MS, MS level 2). During fragmentation, peptide bonds between amino acids and/or side-chain residues of amino acids are broken. As a result, different fragment ion types (Domon and Aebersold 2006) are produced as seen peaks having m/z ratios and intensities. The mass difference between peaks in a tandem mass spectrum is used to infer the peptide sequence. Automated mass spectrometry procedures produce high-throughput data meaning that in one run of the machine millions of spectra can be obtained. In addition to peptide-protein identification, mass spectrometry is used to quantify proteins under different conditions of the proteome. These issues lead the development of computational methods or pipelines to perform analysis of tandem mass spectra (Käll and Vitek 2011). In the next section, these analysis strategies will be explained.

1.2.1. Data Analysis Strategies

The peptide identification strategy via computational methods described here can be classified into two main approaches, database search, and de novo sequencing. While database search algorithms search MS/MS spectra against an amino acid database or spectral libraries that were identified previously assigned to peptides, de novo sequencing aims to find peptide sequences by utilizing only the information provided in a tandem mass spectrum without additional sequence information.

1.2.1.1. Database Search

Database search algorithms are based on a general procedure as illustrated in Figure 1. This procedure requires an amino acid sequence database and set of MS/MS spectra in a special file format depending on the algorithm. The amino acid sequence database can be originated from six- frame translation of genomic database or three-frame translation of transcriptome data or protein database. The database is *in silico* digested into peptides according to the enzyme which is used in the wet-lab experiment of MS data preparation process. Most of the available algorithms provide a list of enzymes; however, trypsin is the enzyme which is often used in enzymatic cleavage of protein sequences into peptides. Among generated peptide candidates, ones that match to experimental peptide mass within a user-defined error tolerance are selected for further process. The peptide mass error is often called as precursor mass tolerance (PM). For each candidate theoretical MS/MS spectrum is generated according to user-defined or algorithm pre-defined ion types and user-defined post-translational modifications. The correlation between experimental MS/MS spectrum and theoretical spectrum of each candidate peptide is calculated according to the scoring function of the algorithm. Some of the available algorithms employ multiple scoring functions.

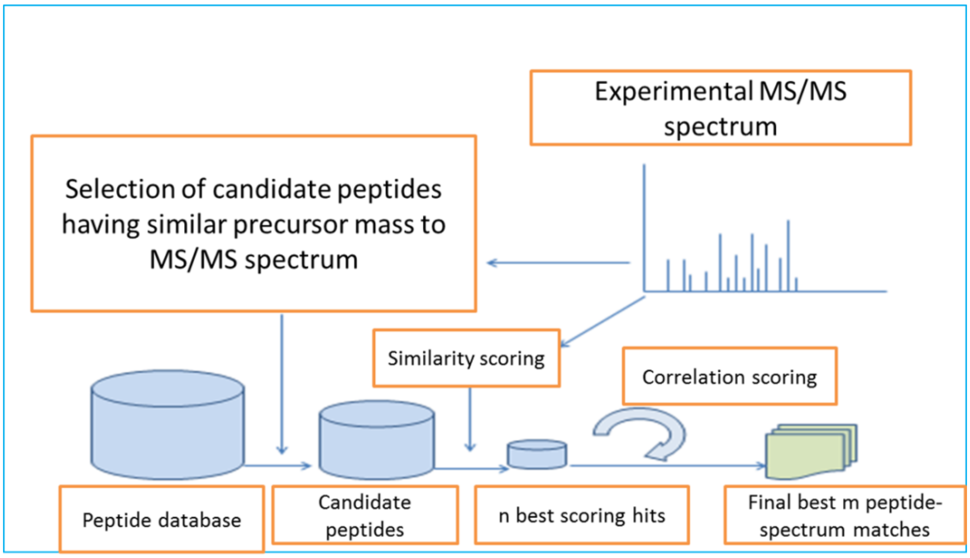

Experimental MS/MS spectrum
Selection of candidate peptides having similar precursor mass to MS/MS spectrum
Similarity scoring
Correlation scoring
Peptide database
Candidate peptides
n best scoring hits
Final best m peptide-spectrum matches

Figure 1. The general concept of database search approach in computational mass spectrometry was illustrated.

There are many commercial and open-source database search algorithms available. MSGF+ (Kim et al. 2010), OMSSA (Geer et al. 2004), Myrimatch (Tabb et al. 2007), Inspect (Tanner et al. 2005), MSAmanda (Dorfer et al. 2014), X!Tandem (Craig and Beavis 2004) are available open-source algorithms. On the other hand, Mascot (Perkins et al. 1999), Sequest (Eng et al. 1994) and PEAKSDB (Zhang et al. 2012) are commercially available products.

1.2.1.2. De novo Sequencing

Many de novo sequencing tools (Allmer 2011; Hoopmann and Moritz 2013) with different algorithmic principles have been published so far. First attempts for de novo sequencing of MS/MS spectra was performed by generation of all possible peptide sequence and evaluation of them by allowing precursor and fragment mass error of the MS instrument. Instead of generating all sequences, optimization approaches like genetic algorithm or ant colony optimization (Has et al. 2012) can be used to explore the vast search space. Many advanced de novo sequencing approaches have been established, but most current solutions employ spectrum graphs (Taylor and Johnson 2001; Frank and Pevzner 2005).

1.2.1.3. Scoring Peptide-Spectrum Matches

The quality of peptide assignment is defined by a scoring function both in de novo sequencing and database search. Many algorithms provide multiple scoring systems with a default score. Best *n* peptide-spectrum matches (PSMs) per spectrum are sorted by default scoring system and reported. However, it is not guaranteed that best n hits are in all correct. Different scoring systems of an algorithm might rank best n hits in different order. Therefore even manual validation of peptide assignments becomes error-prone. The scoring scheme varies between algorithms, therefore, do not present a common-sense to interpret and integrate the results. Therefore, algorithm independent general statistical assessment methodologies to assign significance to PSMs have been developed.

In the general concept of reporting a statistical assessment of PSMs relies on evaluation based on empirical thresholds. The idea behind that is to differentiate PSMs whose scores are greater and equal than a user-defined threshold from PSMs whose scores are less than the threshold. The PSMs above the threshold score are considered as true-positive (TP) while others are regarded as false-positive (FP) hits. The sensitivity of discrimination in TP and FP hits is dependent on how the threshold is set well. High thresholds cause involving more FP hits while low thresholds result in loss of more TPs. For each PSM hit, besides scores assigned by database search algorithm, a comparable and consistent statistical score is assigned at the end of statistical assessment analysis. There have been many statistical validation methodologies published so far (Choi and Nesvizhskii 2008), however, in this section, only a few methodologies will be explained in the chronological order.

The first statistical measure used in computational mass spectrometry-based proteomics is the computation of p-value. p-value definition, in general, is the probability of an event which occurs due to chance. The calculation of p-value is done by calculating the ratio of PSMs with a score above a threshold to all false-positive PSMs. This ratio is called as a false positive rate (FPR) and takes into account variance and sample size. Therefore, a low p-value of a PSM addresses the low probability of being incorrect. As an example, 0.05 p-value threshold indicates that a PSM is FP with 5% chance. Although low p-value scores are considered as significant, the real significance is dependent on the number of PSMs. Therefore, multiple testing needed to validate the significance of the threshold to fitness level. There are many techniques to perform multiple trials, and the simplest and earliest well-known technique is Bonferroni correction (Napierala 2012).

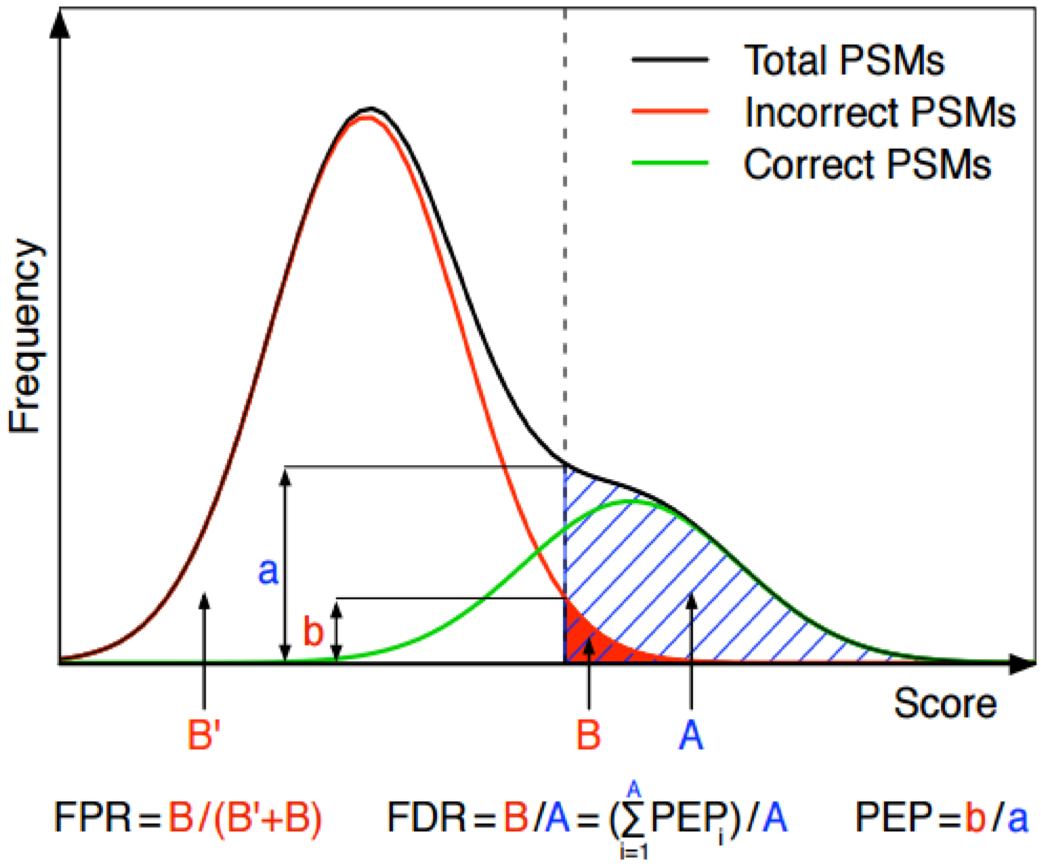


Figure 2. Score distribution (black) of correct (green) and incorrect (red) PSMs. PSMs that are the above-defined threshold (dashed line) point all accepted PSMs (shaded blue region= A) and incorrect PSMs according to the threshold are found as overlapping (red region=B). False positive rate (FPR) and false discovery rate (FDR) formulation derivations, as well as posterior error probability (PEP), are shown. (Source: Brosch et al. 2011)

Bonferroni correction does not provide a balance between false positive and false negative (FN) hits. Therefore false-negative hits might be considered as FP. To overcome this limitation of Bonferroni correction, false discovery rate (FDR) (Benjamini and Hochberg, 1995) method has been developed. FDR is described as the expected proportion of incorrect PSMs among all PSMs above pre-defined score threshold (Figure 2). An example can be given as following: in a set of 100 PSMs with a score above threshold *x*, 10 PSMs are incorrect, thus calculated FDR would be 1%. The FDR threshold, for instance, 1% as in given an example, can be adapted to a different value depending on the goal and perquisites of the experiment. To calculate FDR level, q-value (Käll et al. 2008) is computed. q-value is a local score which defines the significance of an individual PSM as called as minimum FDR threshold. q-value is sensitive to database size, database search algorithm settings, instrument settings so on. Thus the same PSM would differ in q-value under two different search sets. q-value lead to computation of global FDR as representative of all PSMs therefore, it does not reflect a significance measure of single PSM. By Kall et al. (2008) posterior error probability or in other words local-FDR (PEP) measure has been developed to compute PSM specific score (Figure 2). The fact that the sum of the PEPs greater than set threshold score divided by the number of PSMs is also alternative FDR calculation (Keller et al. 2002).

In FDR and PEP calculation, the accepted strategy in general is target-decoy search (Elias and Gygi 2007). Target-decoy approach is defined as usage of decoy version of the target database in addition to standard target database search. Decoy database can be prepared as reversed, random or shuffled version of target database (Reidegeld et al. 2008). Decoy database can be concatenated with target database, or separate search can be conducted (Elias and Gygi 2007). It is expected that PSMs hit to decoy sequences produce a score threshold which can be later used to estimate the number of FP hits. This enables the calculation of the FDR by simply counting the number of the decoy and target PSMs that meet the chosen acceptance criteria. The dependency of some candidates in decoy database; database size impact; unrealistic assumption of normal score distribution in target-decoy databases as limitations of the target-decoy approach based FDR were addressed by Gupta et al. (2011).

Another statistical method is PeptideProphet (Keller et al. 2002) which is a machine learning method based on features defined to discriminate correct-incorrect PMSs and the score of PeptideProphet is based on Sequest scoring systems such as X-corr, deltaCn. Unlike from FDR, PeptideProphet does not accept target and decoy score distributions as same.

A Recent methodology based on target-decoy database usage is named as Percolator (Käll et al. 2007), which has been adapted to many database search algorithms such as Mascot, OMSSA, and MSGF. Percolator requires a training set of correct and incorrect PSM scores and uses semi-supervised learning methodology to discriminate FP and TP hits.

1.3. Genome Annotation

Different from proteomics, genomics is a well-established field with the ability to determine the complete genome of an organism. Next generation sequencing techniques enable cheaper and faster high-throughput genome sequencing of many organisms, leading to data inflation, making data storage and analysis the current bottleneck in genomics (Anderson and Schrijver 2010). The main and final goal of sequencing is to determine genomic elements such as protein coding regions, non-coding regions including longRNAs, miRNAs, alternative splicing products, alternative translation products, regulatory sequences. Therefore, the concept that is called "Genome annotation" has been emerged to provide genome-wide insights into these DNA elements. Genome annotation can be classified into two sub-concepts: structural annotation and functional annotation. Structural annotation aims to identify genes, exon/intron structures, regulatory elements, repeats and variants regarding their localization manner. On the other hand, functional annotation seeks to elucidate the biological function of these items and how they are being expressed. Both annotation types are challenging due to the complexity of genomes, different transcription, translation mechanisms among the various organisms, and structural differences of elements for instance length and nucleotide content effect. Genome annotation in prokaryotes is considered less complex than eukaryotic genome annotation process. Moreover, from simple to higher eukaryotes due to differences in genome structure and organization, setting up a general annotation process is not feasible. Therefore, there are many attempts have been done so far to achieve accurate annotation of the genome. Accurate genome annotation leads better understanding of roles of proteins in biological processes. How structural changes such as variations, the existence of repeats or changes in post-transcriptional and translation mechanisms in DNA affect structures and functions of proteins and later phenotype of the organism through proteins can be enlightened. Genome annotation is achieved either via manual in other words expert-curated process or automated processes (Frishman 2007). In the following section, these two approaches will be explained.

1.3.1. Expert Curation

Manual genome annotation is based on decisions made by human experts. The curated gene models are created by experts in the respective fields using a variety of computational methods but mostly relying on published experimental results. Although manual annotations are considered more trustworthy since experts generate them, they are prohibitively time-consuming. According to a study by Collins et al. (2003), the manual annotation of the human chromosome 22 (1% of human genome) took more than a year for six expert bioinformaticians. The cost of manual annotation and the required high-level expert knowledge have directed researchers to pursue automated annotations techniques. Therefore, this led to the rise of the computationally automated process to perform genome annotation.

1.3.2. Computational Annotation

Automated genome annotation includes the utilization of sequence homology to transfer annotation from known sequences to new ones. In addition to that, application of *ab initio* gene prediction can be done during automated genome annotation. Although there are improvements in well-known genome annotation tools such as HAVANA, National Center for Biotechnology Information (NCBI) and ENSEMBL (Flicek et al. 2012), the accuracy can only reach up to 80% (Harrow et al. 2009). Nonetheless, ENCODE project aims to analyze functional regions of genomes by using both experimental and computational techniques. According to the study of Guigo et al. (2006), only 3.2% of predicted exons could be confirmed by experimental techniques.

Computational genome annotation workflows are based on two broad categories: *ab initio* based and experimental evidence gene predictions. Experimental evidence based methods utilize cDNA (Imanishi et al. 2004), EST (Parkinson and Blaxter 2009). The bottleneck with the usage of cDNA and EST libraries to determine genes is related to low sequence coverage since an only certain portion of genes is transcribed in the cell. In addition to that, sequencing errors, contaminant sequences during the experimental step, truncated cDNAs, the existence of single nucleotide polymorphism (SNP) decrease the quality of EST and cDNA sequences (Nagaraj et al.

2006) and thus might lead to wrong sequence alignments. Besides cDNA and EST sequences, sequence homology-based experimental evidence among evolutionarily related organisms are also used to predict genes. It has been known that exonic regions undergo sequence mutation with slow rate to preserve the functionality of proteins through organisms (Parra 2003). In case usage of homology-based methods species-specific genes are be skipped (Knowles and McLysaght 2009).

The second approach in gene prediction is called *ab initio* gene prediction which utilizes information hidden in DNA sequence. The information which such predictors use is gene signaling sequences such as regulatory regions, promoters, enhancers, and silencers, GC content, statistical features such as length, position. GENSCAN (Burge and Karlin 1997), AUGUSTUS (Stanke et al. 2006) are widely known *ab initio* approach based gene predictors. Although *ab initio* methods are faster than experimental evidence based methods, they are error-prone in compare to experimentally based methods due to low sensitivity and specificity. These algorithms are usually used in case lack of experimental data of organism interest. In addition to that, these algorithms are also used to generate all possible candidates and then validate these candidates on protein level to increase accuracy.

Genomic, EST, cDNA, protein sequences, as well as gene predictions, are deposited in public repositories. ENSEMBL is the one of the first comprehensive repository established to provide this information and high-quality genome annotation of many organisms (Aken et al. 2016). ENSEMBL exploits experimental data such as cDNA and protein data to annotate the genome. In addition to that GENSCAN predictions are also available in ENSEMBL. EST sequences are not included in genome annotation due to high sequencing errors lowering quality. ENSEMBL includes many organisms from vertebrates, plants to bacteria (Kersey et al. 2010). ENSEMBL enables users to analyze data via PERL scripting and SQL querying. It also provides a web-interface which researchers can display existing gene models at all levels as well as custom data loaded by the user. While ENSEMBL provides experimental and computational genome annotation service, HAVANA group at the Welcome Trust Sanger Institute provides manual genome annotation on clone based manner using cDNAs/ESTs and protein sequence data with sequence homology support if any. *HAVANA* also provides *ab initio* gene predictions by GENSCAN and AUGUSTUS. Another manually done genome annotation repository associated to HAVANA is the Vertebrate Genome Annotation (VEGA) database (Wilming and Harrow 2012).

Currently, HAVANA annotated transcripts are also merged into ENSEMBL repository (Aken et al. 2017) and ENSEMBL -HAVANA consensus models are also presented per gene.

1.3.3. Experimental Confirmation

The 40-50% accuracy of computational methods in genome annotation shows that experimental data must confirm results of computational techniques. Early attempts in this aspect include using of cDNA/EST libraries to predict and confirm gene models. However, it does not seem to be possible to determine all transcripts at the moment due to number of cases such as intron-spliced variants (Allmer et al. 2004), alternative translational event (Kochetov 2008). In addition, the underlying sequence assemblies and the mappings of sequences to their corresponding assemblies are not precise so that errors are also transferred during annotation especially when relying on homology. Besides that, it is hard to determine whether a predicted genomic region is protein coding or not since not all RNA transcripts are translated into proteins (Ansong et al. 2008) and ESTs cover not all genes. Therefore, utilizing high-throughput experimental techniques to confirm gene annotation is currently gaining importance in genome annotation strategies (Yates 2000).

1.4. Proteogenomics

Tandem mass spectrometry is one of the state-of-the-art tools, which allows direct, sensitive and high-throughput measurement of proteins, and thus verification of gene models at the translational level. The proteogenomics field emerged from this interplay of genomics, transcriptomics, and proteomics. In this strategy, gene prediction techniques have also exploited tandem mass spectrometric data (Yates et al. 1995). The general workflow of proteogenomic studies starts with the identification of peptides by running database search algorithm(s) on protein database and/or translated genome and transcriptome as well as splice variants, pseudogenes, long-non-coding regions and gene predictions. Identified peptides are then compared with gene models according to genomic locations of peptides and gene models to which the peptides were mapped.

First studies in proteogenomics encompassed confirmation of regions annotated to be protein coding and correction of wrong gene models (Renuse et al. 2011). In addition to that, by performing searches against translated genome or transcriptome, determination of new genes or transcripts has been achieved (Castellana and Bafna 2010). Also, gene annotation of unsequenced or partially sequenced (Jaffe et al. 2004; Castellana et al. 2010) eukaryotic (Choudhary et al. 2001; Collins et al. 2003; Desiere et al. 2005; Fermin et al. 2006; Merchant et al. 2007; Bitton et al. 2010; Brosch et al. 2011; Helmy et al. 2011) and prokaryotic (Himmelreich et al. 1997; Shmatkov et al. 1999; Dandekar et al. 2000; Wang et al. 2009; Wilkins et al. 2009; Venter et al. 2011) organisms can be achieved. A study done by Merrihew et al. (2008) showed the benefits of proteogenomic analysis on the *Caenorhabditis elegans* genome annotation by detection of various errors in 151 gene model and proposing 429 new gene models. In another study, Castellana et al. (2008) reported that 13% of the Arabidopsis thaliana genome annotation was erroneous, and 659 gene models were corrected with the help of protein data. Also, many research labs have been working on detection of biomarkers against diseases (Helmy et al. 2010; Alfaro et al. 2014), environmental response of biological systems aclarification of metabolic pathways such as bioenergetics pathways by using proteogenomic studies (Allmer et al. 2006; Baerenfaller 2008; Wilkins et al. 2009; Tanca et al. 2013). The utilization of proteogenomics can also lead to the confirmation of transcripts which are a product of alternative open reading frames (Castellana et al. 2008; Ning and Nesvizhskii 2010). As shown by de Souza et al. (de Souza et al. 2011) proteogenomics studies can focus on SNPs. In addition to that proteogenomics, studies can be focused on particular problems such as facilitating gene annotation of unsequenced organisms (Yilmaz et al. 2016) and newly sequenced organisms (Gupta et al. 2008). Recent studies in proteogenomics also specialized to meta proteogenomics which involve studies of multiple microbial species found in soil, water, and host environments. Meta proteogenomics which bridges metaproteomics of microbial communities with metagenomics to investigate changes among microbial species in various environments (Seifert et al. 2013). Improvements in mass spectrometry technology and next generation sequencing, development of more specialized sample preparation protocols lead to the production of the tremendous amount of data with high quality. Therefore, the demand for more sophisticated bioinformatics methods at proteomics and genomics side becomes apparent. Therefore, the efforts should be spent to develop more accurate and fast data processing, identification, annotation, and visualization bioinformatics tools by utilizing experimental evidence.

1.5.Thesis Outline

This thesis study is composed of two main objectives. The first objective is to develop fully automated, user-friendly, modular and sustainable proteogenomics pipeline addressing some of the limitations in the current status. A standard proteogenomics bioinformatics pipeline is built of data processing, assignment of peptides to MS/MS spectral data, confidence assignment to found PSMs, mapping of peptides back to genomic data and assessment of gene annotations by comparing genomic locations of the peptide to existing locations and visualization of outcomes. In Chapter 2, limitations at peptide identification step related to database size which effect proteogenomics outcomes were addressed. Proteomics involves the identification of proteins from complex mixtures which is performed using mass spectrometry (MS) followed by computational data analysis. MS/MS spectra can either be sequenced de novo if no sequence is available for the proteins in the mixture or by using database search algorithms when the genomic, transcriptomic or protein sequences are available. The usage of database search algorithms comes along with some limitations regarding database size. Some of the currently used database search tools cannot utilize large databases like the non-redundant protein database or even as small databases as the human chromosome one. Therefore, especially in proteogenomics studies, large databases need to be run independently in smaller chunks, but results from a database having various sequence redundancy and different sizes cannot easily be compared. A new methodology was introduced in Chapter 2 providing proper integration of results from databases different in size and sequence redundancy by equalization of databases to overcome these problems,. In Chapter 3, an accurate and fast exact string matching algorithm, Wu-Manber based peptide mapping tool, Peppig was introduced. Peptide mapping is the crucial step of proteogenomics which finds genomic locations of peptides to enable comparison of locations of peptides against existing genomic features. Database search algorithms do not provide genomic start and end locations of identified peptides. Moreover, consensus peptide identification performed due to its higher confidence is a lack of location information. Therefore, an external tool is needed. Existing peptide mapping tools employing exact string matching algorithms including Peppig were tested against all possible peptide mapping scenarios. In addition to that runtime comparison with increasing, query and database size was performed. Finally, the importance of peptide mapping was shown with a use case study on TCGA Breast cancer data. By using these two tools, with additional novel features, a proteogenomic pipeline, PGMiner was introduced in Chapter 5. PGMiner covers all main steps of proteogenomics and novel features including automated data retrieval from data repositories, equalization of databases, employment multiple database search algorithms, consensus PSM computation, FDR computation, peptide mapping via Peppig, alternative open reading frame prediction, gene annotation comparison. In Chapter 6, PGMiner was tested with human blood plasma MS datasets retrieved from PRIDE and PeptideAtlas repositories. Validation of genes and gene structures, i.e. exon, CDS, as well as correction of existing annotations were shown. A set of novel genes and novel isoforms as a product of alternative translation process were proposed with experimental MS data to compensate currently lacking further experimental data.

CHAPTER 2

IMPACT OF DATABASE SIZE AND DATABASE SEARCH ALGORITHM PARAMETERS ON PEPTIDE IDENTIFICATION

2.1. Introduction

In proteogenomic studies, database search algorithms play an important role to identify peptides to determine novel proteins and their variants. The accuracy and efficiency of peptide identification step via database search algorithms are highly affecting the following steps of proteogenomics studies. As described by Helmy et al. (2012) database search based peptide identification step has bottlenecks in proteogenomics research. The first issue is related to the database features. Databases used in proteogenomics studies can be grouped into three classes. Upon availability of genome and protein database, MS/MS spectra are searched against these databases and if genome annotation is available mapped peptides are compared against the genome annotation. In case genome is newly sequenced and there is no sufficient protein database and genome annotation, translation of genome or transcriptome (RNA-seq) is used as a database. In addition to these database sources, *ab initio* gene predictions performed by gene prediction algorithms such as AUGUSTUS, GENSCAN are also involved into search space. For unsequenced or partially sequenced organisms, genome annotations are often lacking and protein databases are incomplete. Therefore in proteogenomics of such organisms, homology-based database search coupled with de novo sequencing are getting prominent. In both cases, database size and redundancy are at large-scale different than ones used in typical bottom-up proteomics studies. As mentioned in Chapter 1, proteogenomics aims to confirm and improve genome annotations by known and novel peptides. For this reason, genomic and transcriptomic sequences are used to build databases in proteogenomics workflows. Six-frame translation of genomic sequences and three-frame translation of RNA-seq data yield an enormous amount of sequences and often in many proteogenomics studies both data sources are used together with other database sources pseudogenes, cDNA/ EST libraries etc. It has been shown that database search time is highly dependent on database size and MS/MS data size in a linear manner (Edwards 2007). This requires not only a huge demand for computational power but also problems regarding some false negative hits due to non-existing protein sequences, data integration on database and algorithm levels. It has been shown by us and other groups that some of the database search tools used in proteogenomic pipelines cannot utilize databases in large size or including large size database elements. In Table 1, eight database search algorithms were tested to determine minimum database element size in megabytes that can be processed by the algorithm. Among these algorithms, Myrimatch, MSGF+, and OMSSA have no restriction on database element size. However, X!Tandem, Inspect, MSAmanda, pFind, and PEAKS algorithms cannot handle a database element having size 1MB, 5MB, 10MB, 24MB and 260MB respectively.

Table 1. Database search algorithms X!Tandem, Inspect, MSAmanda, pFind, PEAKS, MSGF+, Myrimatch and OMSSA and their maximum size limit on handling FASTA elements in MB.

Algorithm	**FASTA Section Size Limit (MB)**
X!Tandem	1MB
Inspect	5MB
MSAmanda	10MB
pFind	24MB
PEAKS	260MB
Myrimatch	No limitation
MSGF+	No limitation
OMSSA	No limitation

Six-frame translation of human genome yields approximately 6~ GB database. Involving other databases and considering multi miscleavage number inflates a number of candidate peptides by 10 to 100 fold according to Zhou et al. (2010). This leads to increase in the runtime of database search algorithms. To show the relationship between database size and runtime of database search algorithms, we searched randomly selected 5000 spectra against 1MB, 10 MB, 50 MB, 100 MB, 250 MB, 1000 MB databases by MSGF+, X!Tandem, MS Amanda, pFind, OMSSA, Myrimatch, PEAKS, Inspect on 64 GB Ubuntu desktop computer. In Figure 3, log transformed runtimes in MB/s of each algorithm on increasing size databases were shown. Runtimes of Inspect and MS Amanda could not be measured for database size after 10 MB. On the other hand, it has been demonstrated that there is a logarithmic increase in runtimes when database size gets larger. To overcome this problem, either powerful computer needs to be used, or databases need to be chunked.

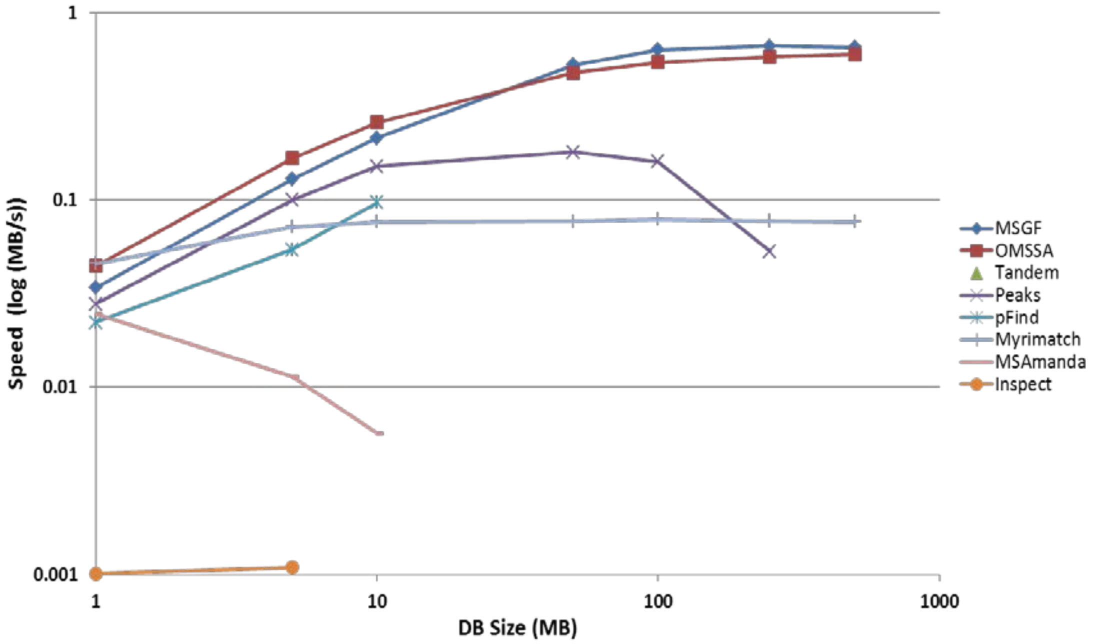


Figure 3. Speed comparison of database search algorithms OMSSA, X!Tandem, MSGF+, PEAKS, pFind, Myrimatch, MSAmanda, Inspect on increasing database size from 1 MB to 1000 MB with fixed number of spectra size as 5000.

Instead of using large size whole genome databases, there are some database pre-processing methods have been developed such as exon-graph, exon-skipping, exon-junction databases, filtering candidates by biochemical properties such as isoelectric point to reduce search space in proteogenomic studies. Without a reduction in the database, six-frame of translation of chromosomes can be searched individually (Bitton et al. 2010). Instead of running chromosomes individually but without reducing the search space into exon level, Fermin et al. (2006) created a library of potential ORFs out of six-frame translated human genome. Branca et al. (2014) presented a method, HiRIEF, based on high-resolution isoelectronic focusing to achieve search space reduction by generating an enzymatic peptidome for the given genome. Utilizing RNA-Seq data also provides knowledge of expressed genes, thus eliminates the complexity related to splicing and masking event occur in the genomic search. Translation of longer assembled RNA-Seq fragments reduces the search space in compare to genome translation. However, it gets along with shortcomings such as the existence of multiple isoforms derived from same reads and difficulty in assembly of multiple RNA-Seq datasets. Woo et al. (2013) introduced a new method named which constructs a splice graph based database from RNA-Seq reads including splicing and mutation events while increasing sensitivity and compression. While these methodologies facilitate runtime of a database search, proper integration of multi-database results remains a problem. As explain in Chapter 1, target-decoy based FDR is not only used a statistical assessment methodology but also as an integration method to find final TP hits. The assumption here that same score distribution will be found in both target and decoy database hits. However, TP hits can be available in decoy database by chance, or isobaric peptides of expected TP hits can be found in decoy database. These two cases might violate the same score distribution assumption of target-decoy approach. Of the 11065 spectra were run against *Escherichia coli*, yeast, pig, mouse and human protein databases that are in size increasing order and their decoy versions by MSGF+, X!Tandem and OMSSA algorithms to test the effect of database size on a number of identifications and score distribution on database search algorithms, In this run, multiple precursor mass tolerance and fragment mass tolerance setting-pairs were used, and optimum results were found for 1.4Da precursor mass tolerance and 0.3Da fragment mass tolerance settings. As shown in Figure 4, some of the predictions OMSSA depends heavily on database size meaning that smaller databases yield a higher number of predictions. In addition to that in X!Tandem showed a slight difference in a number of predictions through increasing database size trend. MSGF+ does not show any significant change. When the scores of target and decoy hits were compared, it was observed that in small databases score distribution between target and decoy version is significant. Therefore some of the retained TP hits will be higher. However, with the increase in size, score distribution between target and decoy databases overlap indicating the higher chance of finding true peptides or isobaric versions of them in decoy database by chance. When small target database to large target database was compared, a significant difference was observed which leads us to consider hits obtained in the small database would surpass hits obtained in a large database when results of multiple databases are integrated to find final hit list. These findings confirm the hypothesis presented by Gupta et al. (2009) that target-decoy approach is not compatible with all database search algorithms to assign statistical significance and to integrate multiple database results.

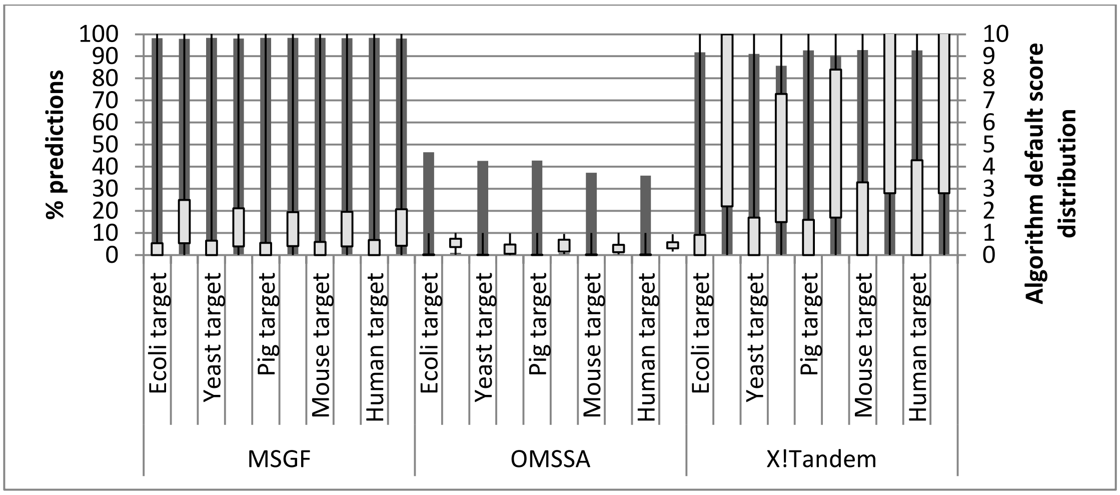


Figure 4. A number of identifications (black bars; left vertical axis) and distribution of E-value scores of algorithms (box-whisker plots, right vertical axis) for MSGF+, OMSSA, and X!Tandem. Results are ordered according to ascending order of database sizes. Score distributions are limited to 10 maximum. In this analysis fragment tolerance and precursor mass tolerance were adjusted to 0.3Da and 1.4 Da respectively.

The second factor that affects the accuracy of database search step is related to developments in mass spectrometry technology (Eng et al. 2011). Database search algorithms require users to set parameters related to the accuracy of mass measurement and sample preparation such as enzymatic cleavage. Since the assignment of peptides to spectra is solely dependent on these parameters, it is important to evaluate algorithms regarding parameter sets to prove which one is better performing in compare others on specific datasets. So far some studies investigated the performance of these algorithms (Kapp et al. 2005; Shadforth et al. 2005). However, few issues are arising in the evaluation step. The first problem is a lack of ground-truth benchmark datasets from known peptides. The second issue is to measure samples in various conditions to assess perturbation effect of analysis. The definition of ground-truth benchmark dataset was recently reported by Allmer et al. (2012). Availability of few benchmark datasets (Keller et al. 2002) and properties of these datasets become a limiting factor in this case. In the literature analysis of specific biological samples (Tabb et al. 2008), complex synthetic samples (Marx et al. 2013) were also measured to evaluate analysis workflows. Although these datasets simulate the real case scenario, peptides available in the sample are unknown due to miscleavage and post-translational modifications or chemical reactions during measurement. In these datasets, peptide-spectrum assignments are based on another algorithm. Therefore assignments cannot be categorized truly due to lack of confidence. In addition to that, the tool used for PSM assignment to evaluate other tools is not the gold standard algorithm. The paper described benchmark dataset as known peptide samples are measured by a set of mass spectrometers, fragmentation methods and measurement settings and correct sequence annotations need to be specified. It is also reported that six measures need to be owned by benchmark datasets. These metrics can be listed as relevance, solvability, scalability, accessibility and independence. In the case of availability of a benchmark dataset, parameters which influence the perturbations can be measured as different combinations of them can be generated.

Precursor mass tolerance and fragment mass tolerance are these parameters which rely heavily on mass analyzer of MS instrument. These tolerances change a number of candidates and scoring of candidate PSMs to find the best PSM. Database search engines compare only the candidate peptides that have similar masses to measured precursor mass within precursor mass tolerance allowed by the mass analyzer. Moreover, the activation method used to fragment peptides into fragment ions also determines types of expected ions. For instance, a-, b- and y- ions are expected in collision-induced dissociation (CID) and higher collision-induced dissociation (HCD) fragmentation techniques, while c- and z- ions are highly expected in ETD fragmentation. Thus, database search algorithms must consider these ion types as well as precursor mass and fragment mass tolerances determined according to the MS instrument and fragmentation method to generate theoretical spectra which are compared against experimental spectra.

In this section, the impact of these two factors on database search step in proteogenomics pipelines will be discussed by using synthetic peptide generated benchmark spectra datasets. A new methodology, called database equalizer, will be introduced to overcome size limitation and integration of different size and redundancy. Algorithm parameter adjustment on example benchmark datasets done via in-house genetic algorithm, which is not part of this thesis, will be explained.

2.2. Methodology

2.2.1. Sample Preparation and MS Analysis

The total of 45 peptide sequences was derived from five different proteins which are cytochrome c (ACN: P00004), bovine serum albumin (ACN: P02769), oval albumin (ACN: P01012), myoglobin (ACN: P68082) and lysozyme c (ACN: P61626). Peptide sequences were synthesized (GL Biochem Ltd, Shanghai, China). Peptide samples for direct-syringe pump measurements were dissolved with 25% ACN in up H2O to reach 1mM. The 1mM stocks were diluted 5 fold under 10% ACN+ 0.1% formic acid. The peptides which contain cysteine and methionine were used in alkalized and non-alkalized form before preparation of stocks. Following that five mix samples were prepared to be analyzed by liquid chromatography coupled tandem mass spectrometers (LC-MS/MS). Each mix was prepared from peptides originated from same protein and in each mix 10 ul from 1mM stock for each involved peptide was added.

The direct-syringe measurements and LC-MS/MS measurements were carried out at ETH Functional Genomics Center (FGC) Zurich Proteomics Facility. Thermo Scientific QExactive instrument with high energy collision dissociation (HCD) fragmentation, Thermo Scientific FUSION instrument with ETch, electron transfer dissociation (ETD) and HCD fragmentations, Thermo Scientific LTQ Orbitrap Velos instrument with collision-induced dissociation (CID) and HCD fragmentation, AB Sciex 5600 TripleTOF instrument at ETH FGC Zurich were employed for direct-syringe pump measurements and LC-MS/MS to simulate real experiment case. A number of spectra generated in this benchmark dataset was shown in Figure 5. Activation time and energy for direct injection and gradient for LC coupled measurement were changed during measurement to achieve varying quality spectral dataset. For CID fragmentation, activation energy was varied as 15eV, 20eV, 30 eV and 45 eV. For HCD fragmentation, activation energy was set as 18eV, 25eV, 35eV and 45eV. In addition isolation width was set to 2, first m/z was set to 50.

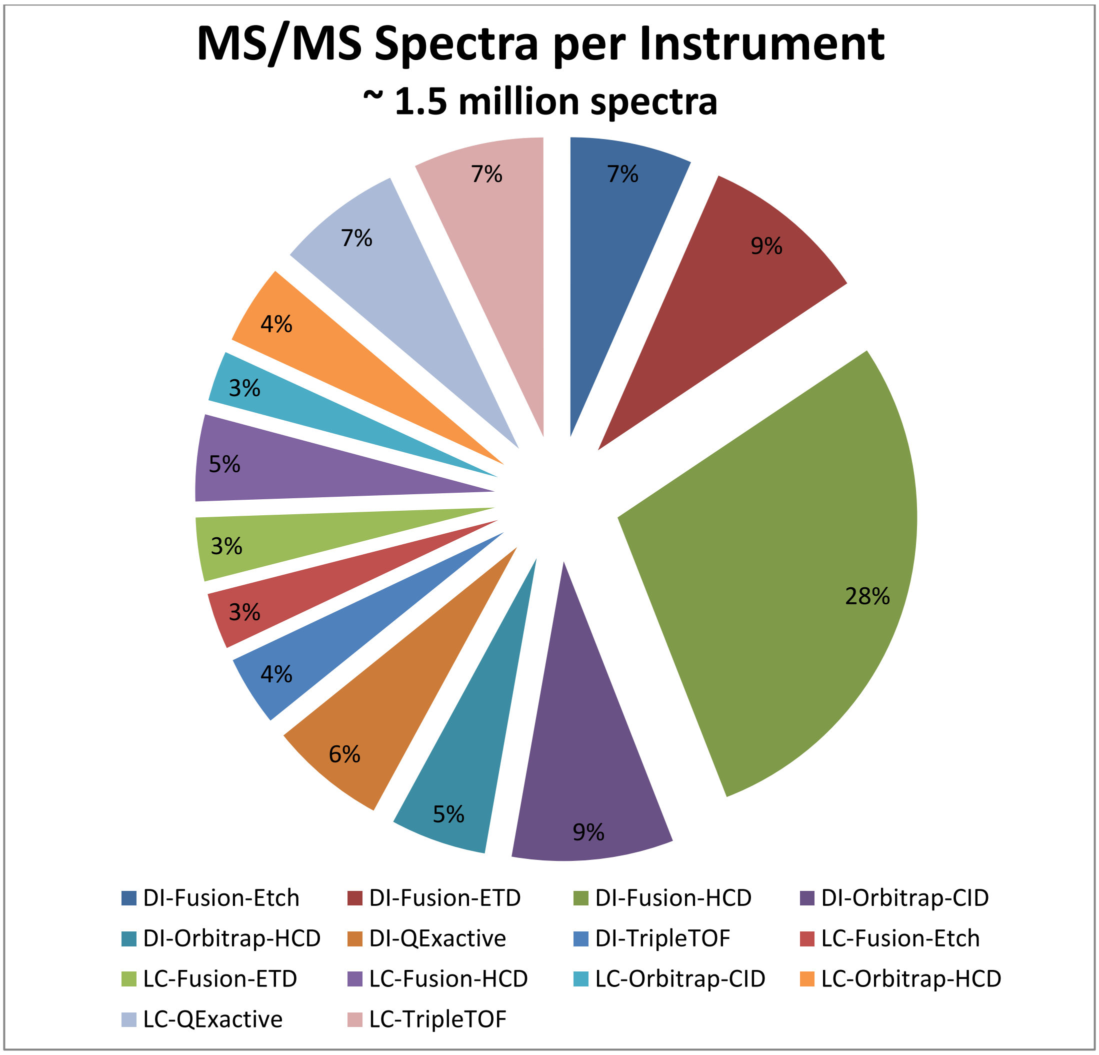


Figure 5. Distribution of benchmark spectra dataset per measurement type-instrument-fragmentation method.

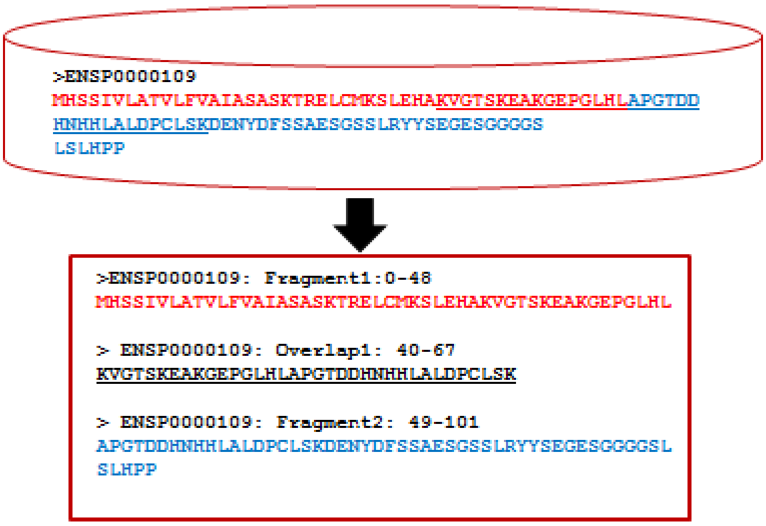
2.2.2. Database Equalizer Algorithm

Database equalizer algorithm was implemented in Java and pseudocode of the algorithm is given in Figure 6. The algorithm is composed of two sections.

```
Input: A number of sequence files $F_1$, $F_2$, ..., $F_n$;  a positive
integer l setting the desired  length of sequence elements;
a positive integer o setting the length of overlapping sequences; a
positive integer m giving the desired number of equalized
files to be produced.
Output: m number of equalized sequence files ($E_1$, $E_2$, ... $E_m$).
Procedure:
get m, l, o
get Files F; n = size(F)
initialize files E[0 .. m]
e = 1
for i = 0 to n
  foreach S in $F_i$ // S represents a sequence
    p=0
    while p < len(S) –l
      E[e++] <- store subsequence(S,p,l)
      p += o
      if e == m
        e = 0
    end while
  end foreach
end for
return E
```

Figure 6. Pseudocode of database equalizer is given.

The first part processes each input database file to generate smaller fragments from each database element. Overlap regions which include the end of the former fragment at o length and the beginning of latter fragment at o length are also generated not to skip fragment cut endings. The illustration Figure 7 shows this process. By default, although fragment length and overlap length are set by users, these sequences are extended until encountering R/K amino acids for protein databases. The reason is that not to lead absence of potential enzyme cleavage products.



```
>ENSP0000109
MHSSIVLATVLFVAIASASKTRELCMKSLEHAKVGTSKEAKGEPGLHLAPGTDD
HNHHLALDPCLSKDENYDFSSAESGSSLRYYSEGESGGGGS
LSLHPP
```

```
>ENSP0000109: Fragment1:0-48
MHSSIVLATVLFVAIASASKTRELCMKSLEHAKVGTSKEAKGEPGLHL

> ENSP0000109: Overlap1: 40-67
KVGTSKEAKGEPGLHLAPGTDDHNHHLALDPCLSK

> ENSP0000109: Fragment2: 49-101
APGTDDHNHHLALDPCLSKDENYDFSSAESGSSLRYYSEGESGGGGSL
SLHPP
```

Figure 7. The first step of database equalizer splits long database elements into fragments up to length *l*. Overlap regions with between any two fragments are generated at length *o*.

In the second part of database equalizer, database elements of different size or origin databases are integrated into same size m number databases. This step is illustrated in Figure 8. With the help of this step, multiple databases become equal regarding redundancy and sequence content. It enables us to compare results from multiple databases.

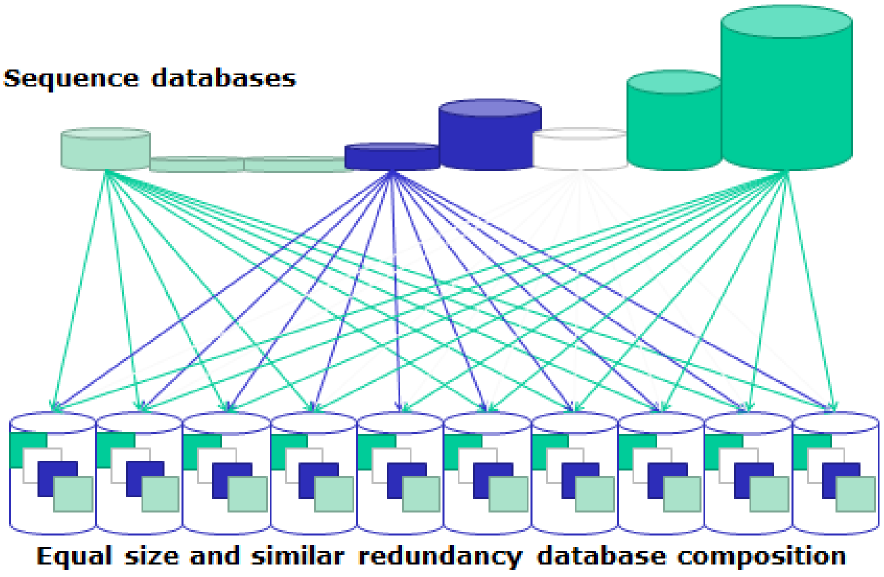


Figure 8. The second step of database equalizer is to merge database elements of different databases into equal size databases.

The hypothesis that scores difference among hits of equalized databases versus different size databases and how a number of candidates effect assigned scores were aimed to be tested. Therefore, six-frame translated human chromosomes 1 and 2 as large, chromosomes 11, 22 as medium size and mitochondria and human protein databases as small size databases were retrieved from ENSEMBL repository. Each database element, i.e. each reading frame in genomic databases and each protein entry in the protein database, were split into 1000 character length fragments and overlap length was set as 100. These databases were integrated into six equal size files. In all database files, proteins of measured peptides were included.

2.2.3. Peptide Identification

The database size effect of all reported scoring functions of eight database search algorithms was compared. Open-source algorithms; Inspect, OMSSA, MSGF+, X!Tandem, Myrimatch, MSAmanda, pFind and commercial algorithm PEAKS were run against six different size databases and six same size databases. Inspect, pFind, X!Tandem, MSAmanda, PEAKS have database element size limitation. Thus, six different size databases were only split into fragments to overcome this limitation. Algorithm settings wereused with default values according to vendor recommendations.

To show whether how scoring functions of algorithms are sensitive to a number of candidate peptides in given database, we also removed competing candidate peptides of expected peptides, which are seen among best 10 hits. After the initial run, best incorrect hits were replaced to A(*n*), i.e. *n* times Alanine, sequences through in each sequence database. By removal of incorrect hits, we aimed to put correct and expected peptides into the first rank. Competitive peptide removal step was repeated two times.

2.2.4. Database Search Algorithm Parameter Adjustment

Parameters of database search algorithms need to be adjusted according to MS instrument and fragmentation model which were used during measurement. In the study, by using the in-house genetic algorithm, X!Tandem and OMSSA algorithms were tested with different parameter settings to influence of algorithm parameters on the number of correct identifications of each algorithm on tested datasets. MSGF+ does not allow parameter optimization since error tolerances as main parameters are set through MS instrument and fragmentation model. Myrimatch, MSAmanda, and Inspect were not tested since the runtime of these algorithms takes so long which is a limiting factor for multiple generations in the genetic algorithm. pFind and PEAKS do not provide executables but only GUI implementation, therefore it was not possible to iterate with multiple settings generated by the genetic algorithm.

The main working principle of the genetic algorithm can be summarized as following. The algorithm starts with default settings represented as an initial chromosome. According to population size set as 100 and number of generations as 20, the random population is generated. For each individual meaning that each parameter setting serial, each algorithm is run on selected spectra against the database. The result of each individual is collected. Afterward, the selection algorithm, crossover, and mutation are respectively applied to the individual. By checking the termination condition, the decision of termination or continuation of iteration is decided. After termination, best individuals are determined. In this study, at max 1500 difficult spectra for LC-Fusion ETD, LC-TripleTOF, Direct Infusion QExactive, and LC-LTQ CID dataset respectively were used. The prominent feature of these spectra is that the datasets were solved by at least three algorithms except OMSSA and X!Tandem algorithms.

2.3. Results and Discussion

DeltaScore ($S_{database1}$-$S_{database2}$) of human chromosome 1 and mitochondrial chromosome and two equal size merge databases were computed to assess the dependency of scoring functions of different database search algorithms to size, In addition to that delta score obtained between initial run and the run after removal of incorrect but candidates as the effect of candidates were plotted as well to show the influence of a number of candidates on score change. For each algorithm, these two delta scores were computed for all reported scoring functions by the algorithm. Here, results of E-value scores of X!Tandem, MSGF+, and OMSSA algorithms were shown as these three algorithms were intended to be used in the last study presented in Chapter 5.

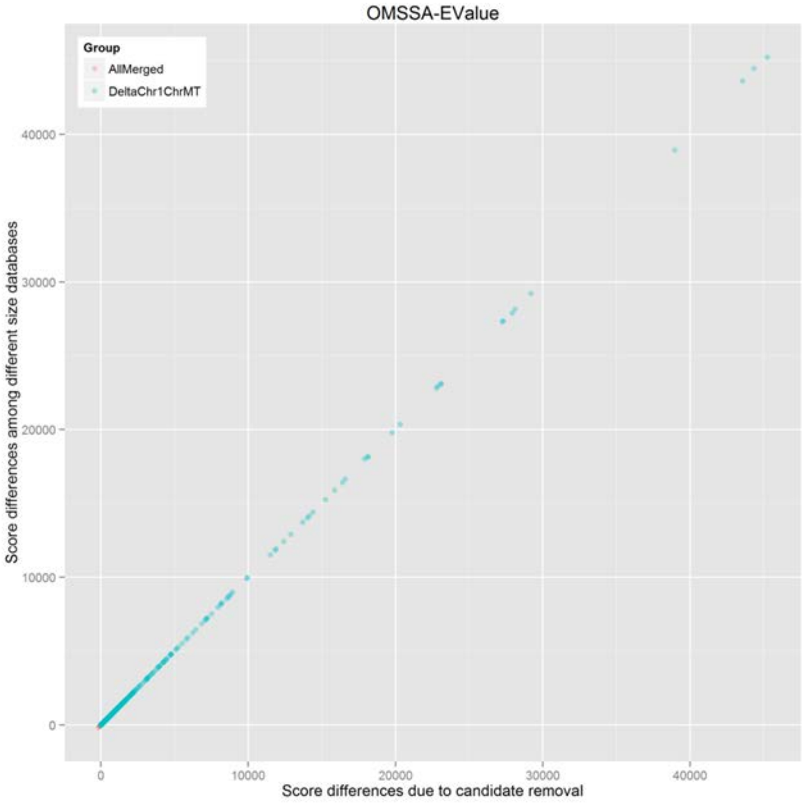


Figure 9. Delta OMSSA E-value score difference between large and small size databases (human chromosome 1 versus human chromosome mitochondria) (blue) and two equal size databases (all merged) (red) were shown in the y-axis. Delta score difference after competitive candidate removal for large-small and equal size databases was shown in the x-axis.

As shown in Figure 9, OMSSA default score type, E-value showed no score difference between split-merge databases (red dots). However, delta difference between human chromosome 1 and the human mitochondrial chromosome is scattered from zero to 30000 (blue dots). This supports the fact that a number of candidates have a big impact on OMSSA E-value scoring function. As stated in the manuscript of OMSSA, E-value described as a score which is the expected number of random hits from a pool of candidates assigned to a spectrum. Here it is expected that random hits should have an equal or better than the actual hit. Therefore, a number of candidates are an important determinant of the value of E-value. However, as shown by the linear line, candidate removal has no significant effect on the E-value scoring of OMSSA. Although the exact numbers of possible candidates in human chromosome 1 and mitochondrial chromosome for expected peptides are not computed, it is seen from the figure that a number of candidates in chromosome 1 is higher than mitochondrial chromosome.

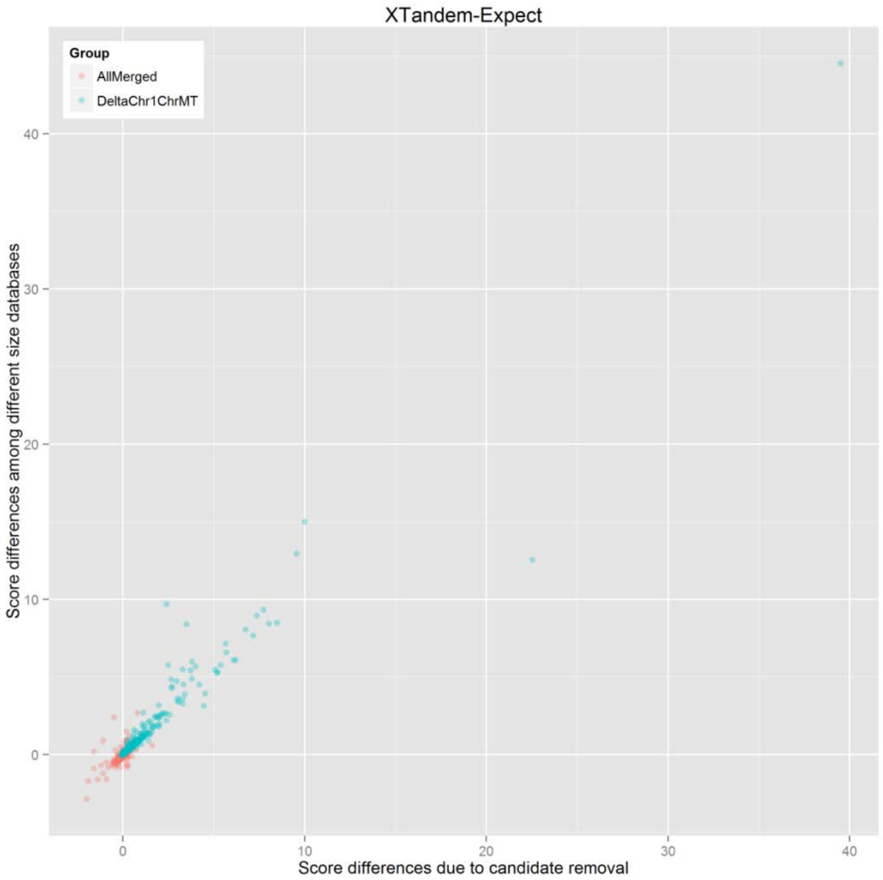


Figure 10. Delta X!Tandem E-value score difference between large and small size databases (human chromosome 1 versus human chromosome mitochondria) (blue) and two equal size databases (all merged) (red) was shown in the y-axis. Delta score difference after competitive candidate removal for large-small and equal size databases was shown in the x-axis.

X!Tandem algorithm returns two scoring types: Expect value (E-value) and hyper score. The first scoring assessment is based on a calculation of dot product by using ion intensities and the number of matching ions according to a comparison of the theoretical spectrum generated from candidate peptide and experimental spectrum. According to the frequency distribution of dot products, survival function is defined. It is a function for discrete stochastic score probability distribution. Here it is defined that the correct matches have greater value than random matches. With this survival function, E-value is computed to indicate the number of PSMs which are expected to have better scores than random matches. Therefore, a number of sequences is a key factor which influences E-value scoring type. On the other hand, hyper score as the other scoring scheme of X!Tandem, is a factorial based on hypergeometric distribution for matches of product ions. The log transformation of the score is returned as hyperscore. The results of X!Tandem matches are ranked according to E-value. Therefore, the comparison of E-values for the given spectra for PSMs obtained from chromosome 1 and mitochondrial chromosome and two equal size and redundancy databases were used. As shown in Figure 10, delta E-values computed from results obtained from different size and redundancy having chromosomes (blue dots) were scattered highly up to 10. On the other hand, delta E-values computed from the equal size and redundancy having databases (red dots) also showed variance, in fact, it is limited between zero to +/- 2.

The last algorithm in this study is MSGF+. MSGF+ algorithm scores peptide spectrum matches based on three main scoring functions; E-value, DeNovo score and MSGF score. E-value scoring is inferred to evaluate the statistical significance of every individual PSM. The computation of these scores is also based on dot product scoring as in X!Tandem. Since E-value is the suggested scoring function by developers of MSGF+ for FDR computing, the impact of candidate number, database size and redundancy were computed for E-value. In Figure 11, candidate removal and database size effects were not observed (red dot). However, candidate removal and database size difference were found as significant for different size databases on MSGF+ E-value.

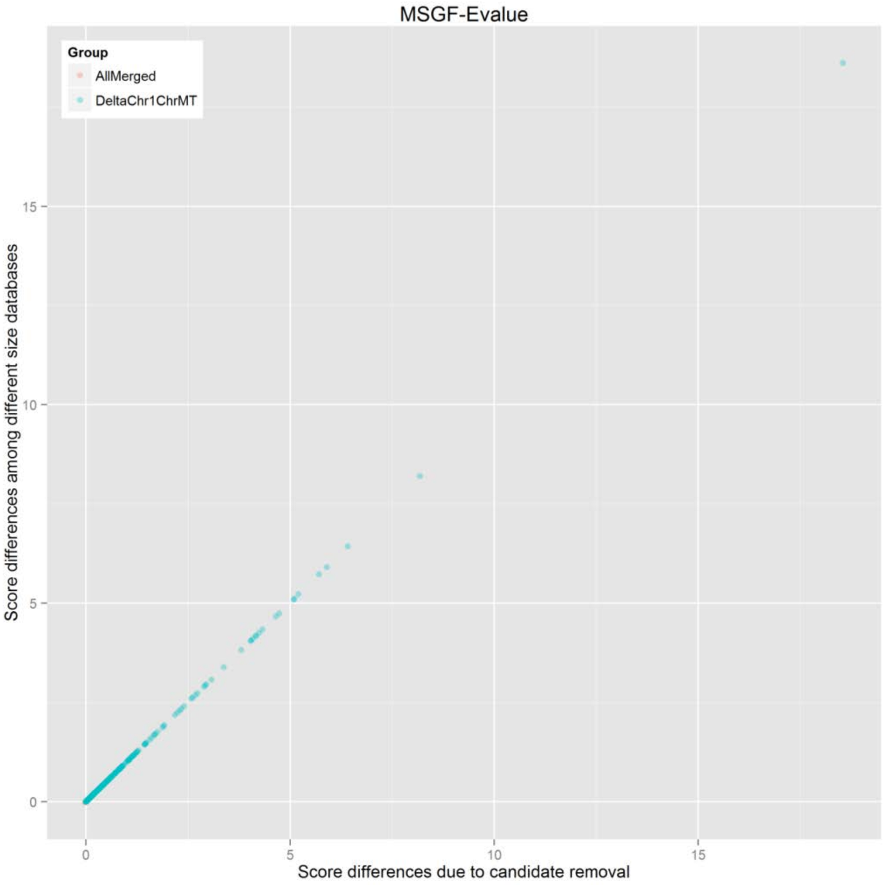
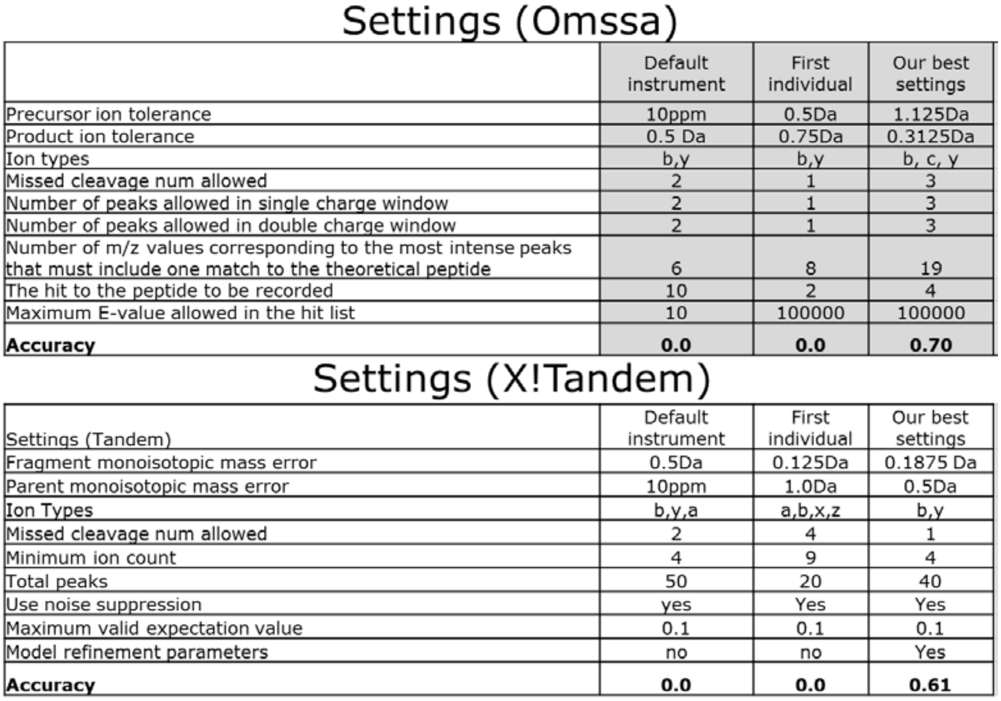


Figure 11. Delta MSGF E-value score difference between large and small size databases (human chromosome 1 versus human chromosome mitochondria) (blue) and two equal size databases (all merged) (red) were shown in y-axis. Delta score difference after competitive candidate removal for large-small and equal size databases were shown in x-axis.

As a conclusion results obtained for OMSSA, X!Tandem, MSGF+ pointed that the database equalizer algorithm usage will not only enable users to speed up the search by parallelization, but also the score difference occurring due to different size will be discarded. The advantage of this will be on the data integration step as in the case of proteogenomics studies.

Another impact factor on database search algorithms is parameter settings before search process. Earlier attempts tested few combinations (Quandt et al. 2014) of parameters in proteomics workflows to determine optimum settings for various MS instrument-fragmentation method based spectra datasets. However, these studies were limited due to a number of combinations used. In this study, OMSSA and X!Tandem algorithms were fed with multiple various combinations more than 1000 by using a genetic algorithm. The main settings which were adjusted for both algorithms were precursor mass tolerance (0 to 4Da), fragment mass tolerance (0 to 2Da), ion types (-a, -b, -c, -x, -y, -z), number of miscleavages (1 to 4 with one interval). In addition to that, minimum ion count (from 1 to 12 with one interval), the total number of peaks (from 0 to 50 with 10 interval), usage of noise suppression (yes or no), model refinement parameters (yes or no) were changed for X!Tandem. For OMSSA, number of peaks allowed in single charge window (from 1 to 4 with one interval), number of peaks allowed in double charge window (from 1 to 4 with one interval), number of m/z values corresponding to the most intense peaks that must include one match to the theoretical peptide (from 1 to 20 with one interval), the hit to the peptide to be recorded (from 1 to 10 with one interval), maximum E-value allowed in the hit list (from 10 to 100000 with 10 interval)settings were changed. Due to limited computational sources, only limited number of spectra up 1500 were used and two algorithms were used. Each algorithm was set to default settings for the first run.

For both algorithms, parameter optimization showed a significant impact on benchmark dataset in terms of accuracy. By considering the dataset which is not solvable with default algorithm settings, it was observed that 0.0 accuracy was increased to 70% for OMSSA algorithm and 61% for X!Tandem algorithm on the LC-LTQ-CID dataset.



Settings (Omssa)

	Default instrument	First individual	Our best settings
Precursor ion tolerance	10ppm	0.5Da	1.125Da
Product ion tolerance	0.5 Da	0.75Da	0.3125Da
Ion types	b,y	b,y	b, c, y
Missed cleavage num allowed	2	1	3
Number of peaks allowed in single charge window	2	1	3
Number of peaks allowed in double charge window	2	1	3
Number of m/z values corresponding to the most intense peaks that must include one match to the theoretical peptide	6	8	19
The hit to the peptide to be recorded	10	2	4
Maximum E-value allowed in the hit list	10	100000	100000
Accuracy	**0.0**	**0.0**	**0.70**

Settings (X!Tandem)

Settings (Tandem)	Default instrument	First individual	Our best settings
Fragment monoisotopic mass error	0.5Da	0.125Da	0.1875 Da
Parent monoisotopic mass error	10ppm	1.0Da	0.5Da
Ion Types	b,y,a	a,b,x,z	b,y
Missed cleavage num allowed	2	4	1
Minimum ion count	4	9	4
Total peaks	50	20	40
Use noise suppression	yes	Yes	Yes
Maximum valid expectation value	0.1	0.1	0.1
Model refinement parameters	no	no	Yes
Accuracy	**0.0**	**0.0**	**0.61**

Figure 12. Comparison of OMSSA and X!Tandem algorithm parameters against defaults, first individual and best settings after genetic algorithm optimization on the LC-LTQ-CID dataset.

In Figure 13 and Figure 14, X!Tandem and OMSSA accuracy results on direct infusion Orbitrap-HCD dataset, LC-TripleTOF dataset and direct infusion Fusion-ETch dataset after parameter optimization via performing genetic algorithm have been given respectively. X!Tandem results show that there is a direct correlation between LC and direct infusion measurement types and accuracy obtained via parameter optimization. The reason here is that during peptide measurement fragmentation efficiency was forced for fluctuation to get the varying quality of spectra. Nevertheless, the same peptide were measured under different settings. Hence spectral annotation was known. On the other hand, varying fragmentation efficiency might have affected some of the peptides in the mixture. Therefore wrong annotations due to missing ions could have been obtained. While a significant increase was observed in direct infusion Orbitrap HCD dataset and direct infusion Fusion-ETch dataset, LC-TripleTOF dataset did not yield a significant accuracy through generations.

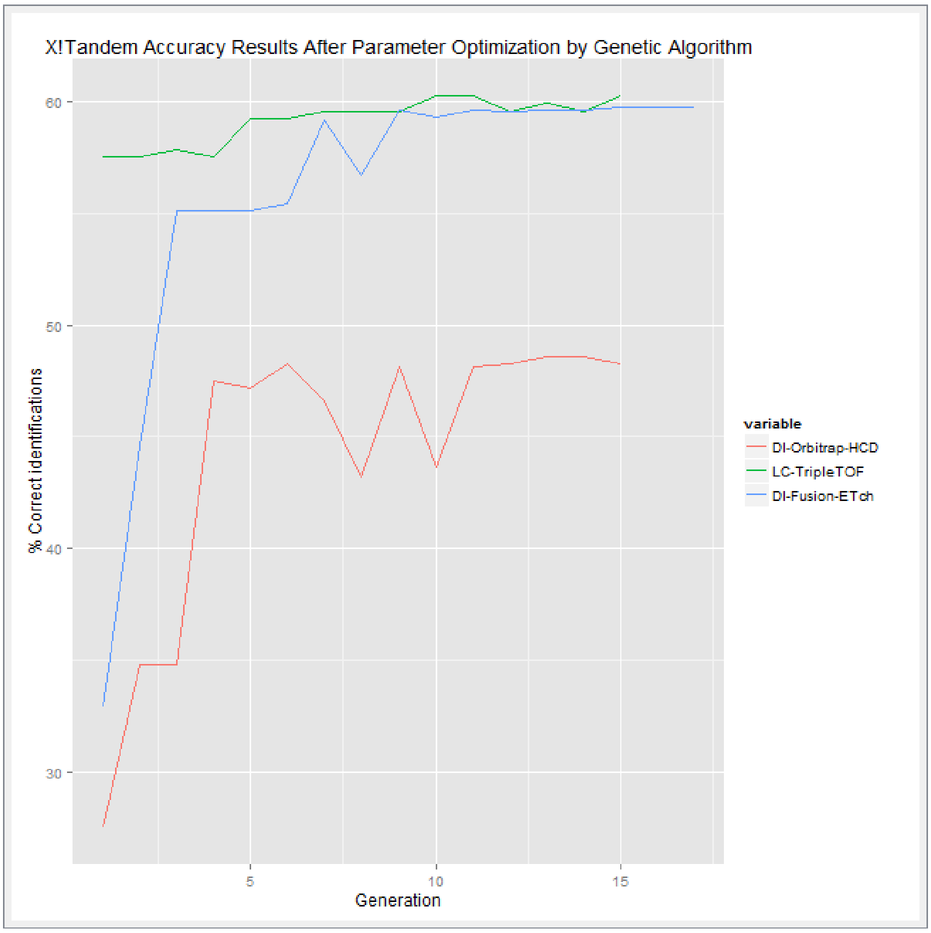


Figure 13. X!Tandem algorithm executed on selected spectra for each instrument-fragmentation method benchmark datasets. The spectra used for this comparison was at max 1500 spectra and are labeled as hard dataset since they are solved at least three algorithms excluding X!Tandem algorithm. In the plot, only three significant datasets were shown as direct infusion-HCD, LC-TripleTOF and direct infusion- ETch datasets.

OMSSA algorithm resulted in major changes in accuracy for all datasets by generations. This tells us that the parameters of OMSSA have the higher impact on scoring functions therefore accuracy. The accuracy increased approximately 25-30% in all datasets. When the results were inspected individually, it was shown that precursor mass tolerance, fragment mass tolerance, ion types, the number of singly and doubly charged peaks in the windows, the number of highly intense peaks and the E-value threshold are the settings which are mostly affecting the accuracy.

This pilot study underpins the importance of training of algorithms according to the MS dataset in use during development and usage. For the different instrument-fragmentation method, different sub datasets with different quality and difficulty level need to be executed on database search algorithms even on de novo sequencing algorithms.

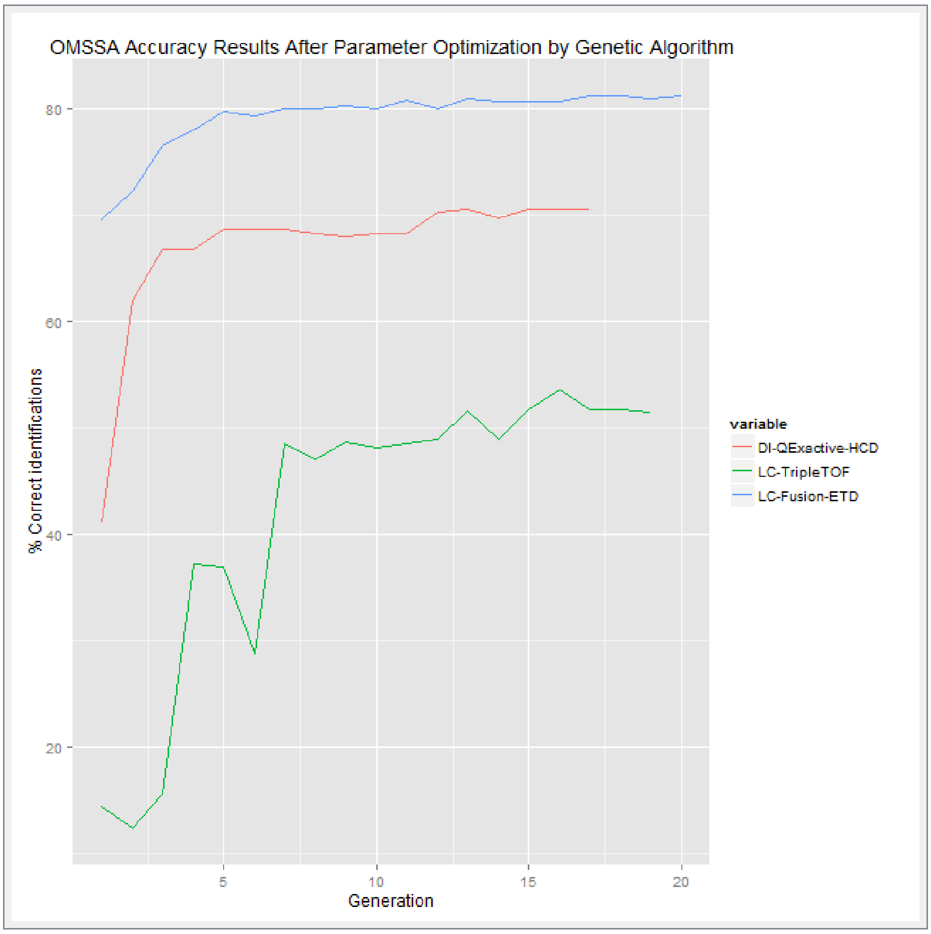


Figure 14. OMSSA algorithm executed on selected spectra for each instrument-fragmentation method benchmark datasets. The spectra used for this comparison was at max 1500 spectra and are labeled as hard dataset since they are solved at least three algorithms excluding OMSSA algorithm. In the plot, only three significant datasets were shown as direct infusion-HCD, LC-TripleTOF and direct infusion- ETch datasets.

CHAPTER 3

ACCURATE PROTEOGENOMIC PEPTIDE MAPPING

3.1. Introduction

One task in this area is to map proteomic data back to the underlying genome. Database search algorithms establish a PSM, but may not report all locations of the assigned peptide in the underlying database (Table 2). Although some database search tools provide all peptide mappings for the PSMs they establish, they fail to reach a large intersection with other database search tools which require the use of multiple database search tools to generate consensus peptide identifications followed by peptide mapping. All peptide mappings to all relevant databases need to be known in proteogenomic studies to enable proper gene model assessment. Therefore, assigning genomic locations of identified peptides is a necessary step, and it has been implemented within proteogenomics pipelines and three standalone tools are available. Peptide mapping in this manner is a non-heuristic process and affords complete correctness. Errors affect the following steps in proteogenomics pipelines such as proposing new gene models with respect to peptide mappings.

Table 2. A non-comprehensive list of database search algorithms with features regarding PSM report. All of them must report PSMs, but there is no requirement for detecting all their occurrences in our opinion. Additionally, genomic databases and thus genomic locations are rarely used, and translation to genomic locations via the use of annotation files is complicated and not performed by database search tools. Four questions define whether and how database search tools report PSMs with multiple locations in a database.

Algorithm	Are all proteins found?	Are all peptide locations in proteins found?	Do parameter settings influence search outcomes?	Does result in output format effect report peptide localization?
OMSSA	Yes	Yes	Yes[1]	No
Masswiz	No	No location information	No	No
Myrimatch	Yes	No location information	No	No
Tandem	Yes	Yes	No	Yes
MSGF+	Yes	No	No	Yes[2]

[1]In case of changes in parameters such as number of returned hits and E-value threshold; OMSSA returns all start & end locations of a peptide in all matching amino acid sequences in the given database (changes may lead to a large increase in runtime).

[2]MSGF+ is intended to return all start & end locations of found peptides in respect to all matching amino acid sequences in the database. However semi-tryptic rule setting or initial scoring system of MSGF+ would affect the reported locations as well as reported amino acid sequences. The scoring system considers flanking amino acids of queried peptides. Therefore for a peptide which is found in two locations of protein A, location 1 is be reported while location 2 is not.

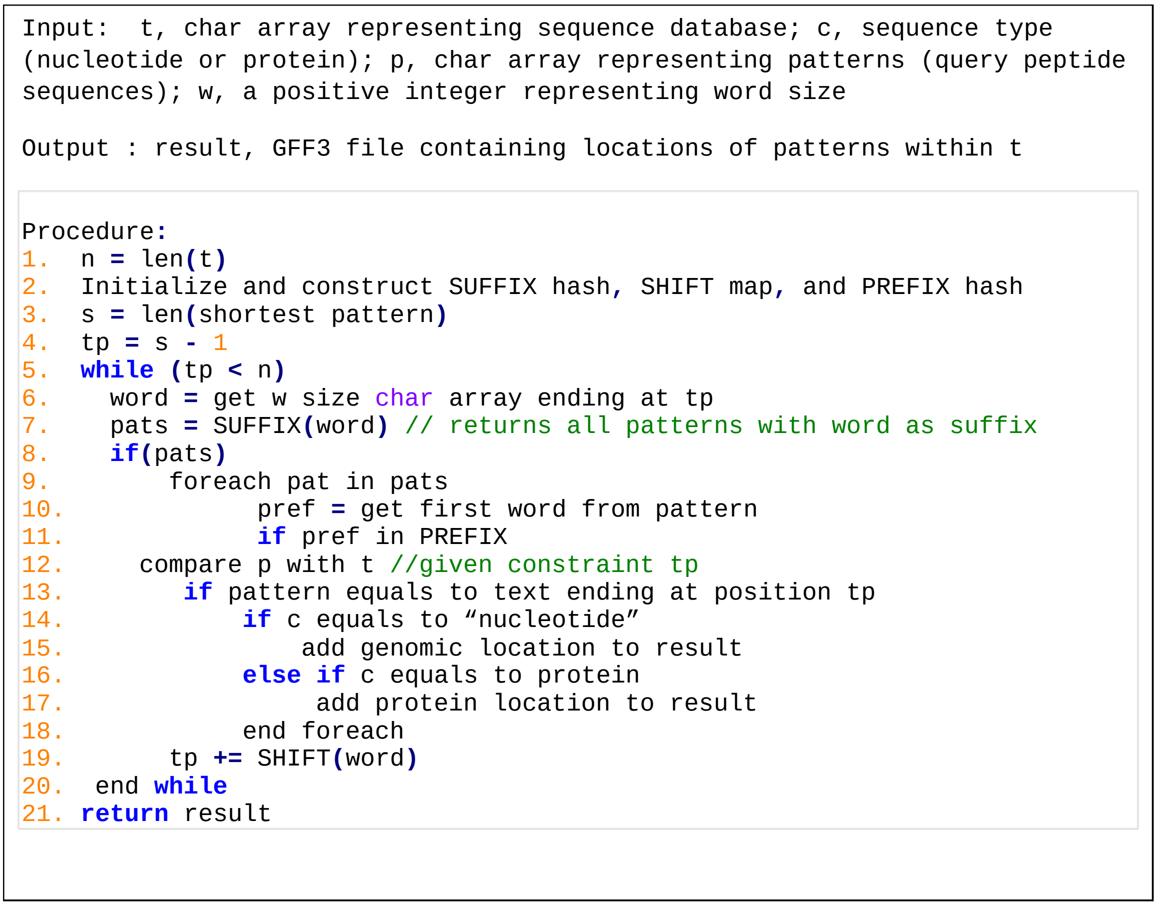
For this reason, software which maps the identified peptides to all their locations within sequence databases (e.g.: genome, transcriptome, proteome) is needed (Menschaert 2015). Furthermore, database search tools in general search only one database at a time and cannot report multiple occurrences when they are spread over multiple databases. Two standalone tools for the purpose of peptide mapping have been published. In addition to these standalone tools, several proteogenomic pipelines (Kumar et al. 2013; Risk et al. 2013; Jagtap et al. 2014; Nagaraj et al. 2015; Has and Allmer 2016) contain peptide mapping as a step and we are aware that there are many more in-house scripts performing this function. The proteogenomics mapping tool (PGM) offers an implementation of the Aho-Corasick algorithm (1975) for mapping peptides to genomes with some additional functionality such as finding expressed open reading frames (Sanders et al. 2011). PGx (Askenazi et al. 2015) has recently been proposed as an alternative method to PGM, performing 2-step indexing to achieve a faster mapping of peptides to protein database and returning corresponding genomic locations of mapped peptides according to given BED file. The problem of peptide mapping can also be solved by non-targeted algorithms like BLAST (Altschul et al. 1990) and BLAT (Kent 2002). Moreover, Allmer et al. (2016) published a set of exact-pattern matching tools which can be used for peptide mapping purpose.

As part of this thesis, a new exact peptide mapping implementation inspired by the Wu-Manber algorithm, called Peppig, has been implemented. Exact pattern matching should lead to correct algorithms which must solve all instances of the peptide mapping problem properly. The correctness of results were ensured by establishing a comprehensive test-set consisting of 245 cases. These test cases address all possible mapping scenarios such as overlapping peptides, multiple occurrences, and tandem repeating peptides in the sequence database. Peptide sequences were designed as tryptic peptides, and they do not contain any unknown characters. PGM, Peppig and PGx were tested on these cases and runtime comparison between Peppig and PGM and one heuristic algorithm BLAT were measured to show a significant difference in the completion of the mapping process. While Peppig was able to solve all test-cases correctly, PGM and PGx only solved 79% and 100% of cases respectively without counting systematic errors. All test-case are available through jlab.iyte.edu.tr/software so that developers can test their in-house tools. Peppig is provided as a standalone tool for proteogenomic peptide mapping through http://jlab.iyte.edu.tr/software.

3.2. Methodology

3.2.1. Peppig Implementation

Peppig algorithm is a modified and fast version of the Wu-Manber algorithm and implemented in Java. The fundamental idea of the Wu-Manber algorithm is to utilize hashing strategy on prefix and suffixes to shift the words according to that. Wu-Manber algorithm is composed of two main phases: preprocessing and scanning. In preprocessing step, smallest length having patterns is found. Afterward, SHIFT table, HASH table, and PREFIX tables are generated. SHIFT table contains maximum shift when a mismatch is encountered. In HASH table holds values which are used when SHIFT table has zero value. PREFIX table, on the other hand, contains matchings of first B characters of patterns. The existence of PREFIX table speeds up the scanning step. In scanning step, first *n* characters of the text are taken from the array, which is called window. Last B characters of the window is compared to the pattern, and according to the shift value in SHIFT table, the window is shifted. In case the shift value is zero, the corresponding value in the HASH table is used. When there are multiple values in the HASH table, PREFIX table is used, and prefix of the pattern is compared. When a prefix match is encountered, then the whole pattern is compared against text by one –character at a time. The pseudocode of Wu-Manber implementation is given in Figure 15. Hashing strategy in Wu-Manber algorithm brings an advantage to less space and less number of comparisons due to large shifts.



```
Input:  t, char array representing sequence database; c, sequence type
(nucleotide or protein); p, char array representing patterns (query peptide
sequences); w, a positive integer representing word size

Output : result, GFF3 file containing locations of patterns within t

Procedure:
n = len(t)
Initialize and construct SUFFIX hash, SHIFT map, and PREFIX hash
s = len(shortest pattern)
tp = s - 1
while (tp < n)
  word = get w size char array ending at tp
  pats = SUFFIX(word) // returns all patterns with word as suffix
  if(pats)
      foreach pat in pats
            pref = get first word from pattern
            if pref in PREFIX
    compare p with t //given constraint tp
       if pattern equals to text ending at position tp
           if c equals to “nucleotide”
               add genomic location to result
           else if c equals to protein
                add protein location to result
           end foreach
      tp += SHIFT(word)
 end while
return result
```

Figure 15. Pseudocode of Peppig is given.

Peppig takes a list of peptide queries and sequence database in FASTA format. Peppig accepts nucleotide and amino acid databases. Via automatic sequence type detection, Peppig determines the database type and translates sequences into six- or three frames. The translation of all frames is done at the same time by reading the sequence at once which accelerates this process. When queried peptides are found in the database, genomic locations of these peptides are computed if database source is nucleotide. Mapping results are returned in GFF format.

3.2.2. Test Scenarios

3.2.2.1. Implementations and Inputs

In order to assure algorithm correctness, the underlying assumptions need to be tested ensuring that input is correctly transformed into the expected output. For testing peptide mapping or exact pattern matching in general, seven broad test scenarios were devised. Peppig and PGM (Release date: Stand-alone jar executable 2011-03-29) are able to translate the given nucleotide sequence database. However PGx (Release date: 2016-03-29) has no such feature. Hence, we used six-frame translated sequence to involve PGx into comparison.

The algorithms were tested against a database with a single element and one with multiple ones. In this respect, as the single database element example nucleotide sequence of human (*Homo sapiens*) pleckstrin homology like domain family member B was used. In multiple database element test case, in addition to human pleckstrin homology like domain family member B, those of homologous sequences in gorilla (*Gorrilla gorrila*), rhesus macaque (*Macaca mulatta*) were included. Query peptides were selected as tryptic peptides that do not contain any common amino acid symbols or "*" symbolizing stop codon. In order to test the effect of foreign symbols existence to query finding in database files, we created two versions of database files; 1) stop codons were symbolized as "*" 2) stop codons were translated to "F", phenylalanine.

3.2.2.1.1. Test Scenario One

Tests in this scenario evaluate whether peptides can be retrieved when they are at the beginning, in the middle, or at the end of a sequence for the six possible translation reading frames of a nucleotide sequence. While this may seem trite, indexing or other transformation of query set or sequence database may introduce errors.

3.2.2.1.2. Test Scenario Two

Pattern matching may retrieve the first match, only; but in proteogenomic peptide mapping, it is essential to detect all matches. Therefore, this scenario tests whether the query peptides are matched to all expected locations in the sequence and whether the locations are correctly reported since shifts in locations can easily occur due to wrong assumptions or calculations. Six queries were created for this scenario. The first one is expected to be found at the beginning and at the end of first-forward reading frame of the six frame translation of the sequence. The second query is expected to be found at the beginning and at the end of the second-forward reading frame of the six frame translation of the sequence. The third query is expected to be found at the beginning and at the end of the third-forward reading frame of the six frame translation of the sequence. The fourth, fifth and sixth queries were expected to be found at the beginning and at the end of the first-reverse reading frame, second-reverse reading frame and third-reverse reading frame respectively.

3.2.2.1.3. Test Scenario Three

Either the queries and/or the sequence database may be transformed for more efficient pattern matching and, therefore, it is important to test whether that was done correctly. To inquire this, queries which contain shared prefixes, infixes, or suffixes were constructed. Three different peptides of varying length, ranging from 10-25 amino acids, which contain shared prefixes, infixes, and suffixes, were selected as queries.

3.2.2.1.4. Test Scenario Four

Tests in this scenario check whether tandem sequence repeats which can occur for low-complex peptide sequences are appropriately mapped to the sequence multiple times.

3.2.2.1.5. Test Scenario Five

Due to miscleavages, peptides might overlap. This scenario checks whether overlapping tryptic peptides can be mapped properly by all tools. Single and four amino acid longer overlapping regions were created.

3.2.2.1.6. Test Scenario Six

Proteogenomic peptide mapping may involve tens of thousands of peptides and it is important to check whether the algorithms support this use-case. The scalability of the implementation is tested with varying number of queries: 50, 100, 500, 1000, 5000, 10k, 50k, and 200k respectively. These peptide queries were distinct randomly generated tryptic peptides. For proteogenomic peptide mapping, we would expect less than 10000 peptide queries in current real world application, but checking whether methods scale well for future implementation when more peptides can be retrieved from MS analysis, seems necessary. Due to the construction of the query set we can only guarantee that at least one match must exist but cannot exclude multiple matches for some of the generated queries.

3.2.2.1.7. Test Scenario Seven

It is often the case that large sequence databases are used in proteogenomic studies. Therefore, it is crucial that peptide mapping algorithms are able to handle large sequence files to search for peptides. In this test scenario, four databases with sizes 100MB, 500MB, 1GB, 5GB, respectively, were used to test whether algorithms scale well in respect to database size.

3.3. Results and Discussion

Functionality and correctness of an algorithm need to be tested against benchmark standards to instill trust in its implementation. Proper benchmark datasets cover the envisioned input and expected wrong input with their associated expected outputs. A test framework is useful, but not always easy to setup when algorithms work in different software environments, for example.

We developed a benchmark dataset encompassing 7 different scenarios with a total of 245 tests. The three peptide available standalone mapping algorithms, PGM, PGx, and Peppig were benchmarked on the developed test cases manually, and the results are presented on a per scenario basis in the following.

Scenario one tests whether the genomic locations of matches are properly reported (Table 3). Peppig solves all test cases correctly which was confirmed by manually validating them in the Artemis genome browser (Rutherford et al. 2000). PGM, on the other hand, had 1 shift in 36% of scenario 1 cases. 33% of cases were not returned by PGM due to implementation error.

Table 3. Test outcomes for scenario one. Matches as a beginning, middle and end locations are tested for the three peptide mapping tools against the expectation provided. Scenario one tests the outcomes for the three available algorithms and presents quantification of detected errors.

		Exactly correct hits	1-shift hits	1> shift hits	Missing hits	Total
Single Element - Star DB	Peppig	17				17
	PGM	4	8		5	
	PGx		15	2		
Multi Element - Star DB	Peppig	34				34
	PGM	14	15		5	
	PGx		29	5		
Single Element – Non-star DB	Peppig	20				20
	PGM	4	6		10	
	PGx		18	2		
Multi-Element – Non-star DB	Peppig	33				33
	PGM	10	9		14	
	PGx		29	4		

PGx found all matches, but the locations of the matches in the databases were incorrect and were shifted by one (88%) or more (12%) bases. The large part of that problem could be due to a different reporting paradigm (shifted by 1 base) which could be accounted for in the downstream analysis, but for larger shifts (mostly in reading frame 2) we expect it to be an implementation error or preparation of BED file.

Scenario two tests whether all matches of a query that locates at the beginning and at the end of the sequence databases are correctly returned. In this test case, the expectation is that an algorithm must scan the complete database starting from the first character to the last character. According to Table 4 PGM reports only the first occurrence of the queries on all four databases, thus failed solving of 41% cases. Thus, it was concluded that it is a bug that not all locations are reported. Peppig and PGx, however, returned all peptides in all locations. However, we observed 1 shift in hits found by PGx.

Table 4. Test outcomes for scenario two. Matches as front and end positions of same queries are tested for the three peptide mapping tools against the expectation provided.

		Exactly correct hits	1-shift hits	1> shift hits	Missing hits	Total
Single Element Star DB	Peppig	4	0	0	0	4
	PGM	2	0	0	2	
	PGx	0	4	0	0	
Multi Element - Star DB	Peppig	4	0	0	0	4
	PGM	2	0	0	2	
	PGx	0	4	0	0	
Single Element - Non-star DB	Peppig	6	0	0	0	6
	PGM	3	0	0	3	
	PGx	0	6	0	0	
Multi Element - Non-star DB	Peppig	10	0	0	0	10
	PGM	7	0	0	3	
	PGx	0	10	0	0	

As shown in Table 5 we assume that some problems might occur during the query tree construction leading to errors in 1 out of 11 test cases. We ran into similar problems with PGM when creating larger query sets with 1000 or more queries randomly extracted from the human chromosome 1. PGx, on the other hand, was able to return peptides according to gene locations given in BED file, but as in other test cases, all locations had 1 index shift. Peppig solves all tests correctly.

Table 5. Results for test cases in scenario three. Peptides having partial similarities (infix, suffix, and prefix) among each other were queried with PGM, Peppig, and PGx and the outcomes compared to the expected results.

		Exactly correct hits	1-shift hits	1> shift hits	Missing hits	Total
Single Element Star DB	Peppig	11/11	0	0	0	11
	PGM	10/11	0	0	1/11	
	PGx	0	11/11	0	0	
Multi Element Star DB	Peppig	11/11	0	0	0	11
	PGM	10/11	0	0	1/11	
	PGx	0	11/11	0	0	
Single Element Non-star DB	Peppig	11/11	0	0	0	11
	PGM	10/11	0	0	1/11	
	PGx	0	11/11	0	0	
Multi Element Non-star DB	Peppig	11/11	0	0	0	11
	PGM	10/11	0	0	1/11	
	PGx	0	11/11	0	0	

Peppig successfully maps queries of tandem repeated amino acid as well as a query of larger sequence randomly located in databases multiple times as tested in scenario four (Table 6). PGM return all expected locations correctly on star symbol containing single and multiple element containing databases. However, in no star symbol containing single and multiple databases, 34% of total cases were returned with one shift. As in other scenarios, PGx returned all expected peptides with single index shift in all databases.

Table 6. Results for the test cases from scenario four. Tandem repeats and multiple occurrences of large peptides needed to be mapped to matches which overlap within the database sequence.

		Exactly correct hits	1-shift hits	1> shift hits	Missing hits	Total
Single Element Star DB	Peppig	5	0	0	0	5
	PGM	5	0	0	0	
	PGx	0	5	0	0	
Multi Element Star DB	Peppig	6	0	0	0	6
	PGM	6	0	0	0	
	PGx	0	6	0	0	
Single Element Non-star DB	Peppig	9	0	0	0	9
	PGM	4	5	0	0	
	PGx	0	9	0	0	
Multi Element Non-star DB	Peppig	9	0	0	0	9
	PGM	4	5	0	0	
	PGx	0	9	0	0	

In scenario 5, the ability of mapping overlapping peptides was tested for all algorithms against all databases. In this scenario, 1 and 4 amino acid overlapping sequences were selected to check sequence overlap length effect returned outputs. Peppig and PGM returned an exact number of expected peptides with correct locations. However, PGx failed to return peptides with 4 amino acid overlaps (Table 7), however, after reporting this bug to the developers of PGx, the bug is fixed. PGx returns all expected peptides.

Table 7. Results for the test cases from scenario five. Overlapping peptides needed to be mapped to the database.

		Exactly correct hits	1-shift hits	1> shift hits	Missing hits	Total
Single Element Star DB	Peppig	8	0	0	0	8
	PGM	8	0	0	0	
	PGx	0	7	0	1	
Multi Element Star DB	Peppig	8	0	0	0	8
	PGM	8	0	0	0	
	PGx	0	7	0	1	
Single Element Non-star DB	Peppig	8	0	0	0	8
	PGM	8	0	0	0	
	PGx	0	7	0	1	
Multi Element Non-star DB	Peppig	8	0	0	0	8
	PGM	8	0	0	0	
	PGx	0	7	0	1 (fixed after report)	

In addition to reporting correct results, another concern should be query size and database size that a peptide mapping tool is able to process. First of all, we tested upper limits of given database size for each algorithm in scenario six (Table 8). For given query number 10, four databases with sizes 100MB, 500MB, 1GB and 5 GB were run on 32 GB RAM Windows7 installed workstation. Results showed that Peppig and PGM were able to translate databases up to 5 GB and to search queries. On the other hand, PGx was not able to process databases files larger than 100MB.

Table 8. Results for the test cases in scenario six. Different size databases respectively 100MB, 500MB, 1GB, and 5GB were searched by all algorithms with 10 tryptic peptide queries.

Database size	PGM	Peppig	PGx
100 MB	+	+	+
500 MB	+	+	-
1 GB	+	+	-
5 GB	+	+	-

Upper bound on a number of queries that a tool can map to sequence file was tested in scenario 7. All tools were tested with increasing number of queries from 50 to

200k against 100 MB database. 100MB database was chosen due to being lower database size bound for all tools.

Table 9. Results for the test cases from scenario seven. The ability of mapping increasing query size was checked for PGM, Peppig, and PGx. The upper bound for the number of queries was determined for each of them. Due to unacceptable large runtime, PGM was not tested beyond 1000 queries.

Query size	PGM	Peppig	PGx
50	+	+	+
100	+	+	+
500	5 days runtime	+	+
1000	10 days runtime; 300 GB results	+	+
5000	Not tested	+	+
10000	Not tested	+	+
50000	Not tested	+	-
200000	Not tested	+	-

According to results shown in Table 9 Peppig could finish processing large queries; however, PGM took several days to finish even 500 queries whilst PGx failed extreme query sizes such as 50k and 200k.

In order to establish runtime difference between Peppig, PGM, and BLAT three different sized sequence databases were used. PGx was not included in the comparison since it was designed as three separate Python scripts. Human chromosome 1, 10 and Y are containing 247MB, 134MB, and 57.5MB of bases, respectively, were used as nucleotide sequence databases. Peppig, PGM, and BLAT performances were compared for both translation process and search process. Query sets consisted of 50, 100, 500, 1000, 5000 (normal use case), 10000 and 50000 peptide sequences that were randomly extracted from the six-frame translation of each chromosome. Run times were measured on a workstation with the following specifications: 40 GB RAM Intel Core i7 3.07GHz x. The mint flavor of Ubuntu Linux was installed on the PC, and most competing tasks were disabled during performance measurement. Additionally, where possible, any but the search module was disabled. Note, that searching consumes the largest part of the runtime. For a fair comparison, the measurements were repeated 5 times, and the averages are reported.

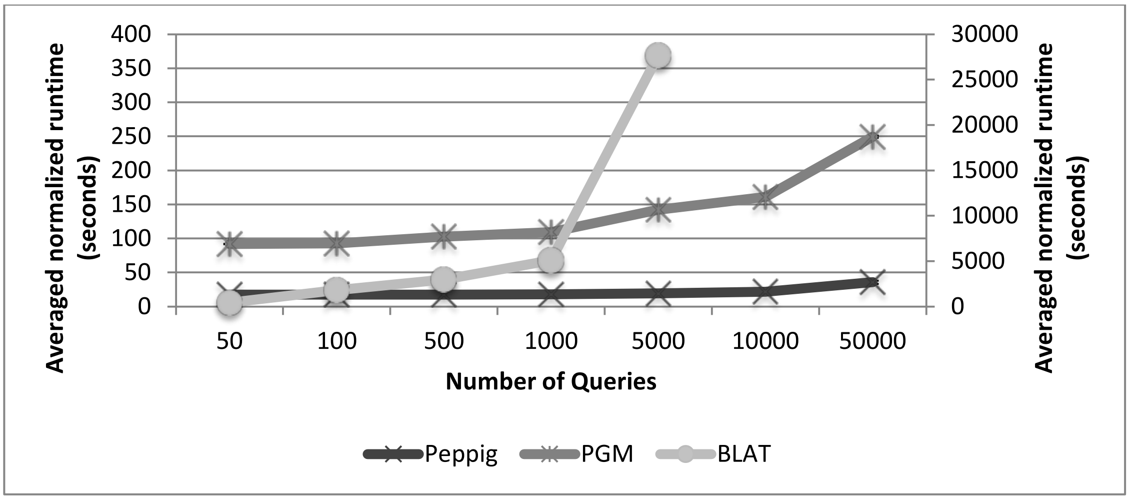


Figure 16. BLAT, Peppig, and PGM were executed on 247MB sized human chromosome 1. BLAT was only tested for 1000 queries due to its long runtime. Runs for each algorithm were repeated 5 times, and run times in seconds were averaged. They are plotted with standard deviations for each peptide query size.

For the largest human chromosome, BLAT shows good performance for few queries, but it quickly degrades, and it becomes evident, that the algorithm was not designed for this kind of task. We chose BLAT over BLAST since BLAT was shown to be faster. Peppig is fastest for normal use cases, and its growth is linear, whereas the growth rates of BLAT and PGM seem exponential.

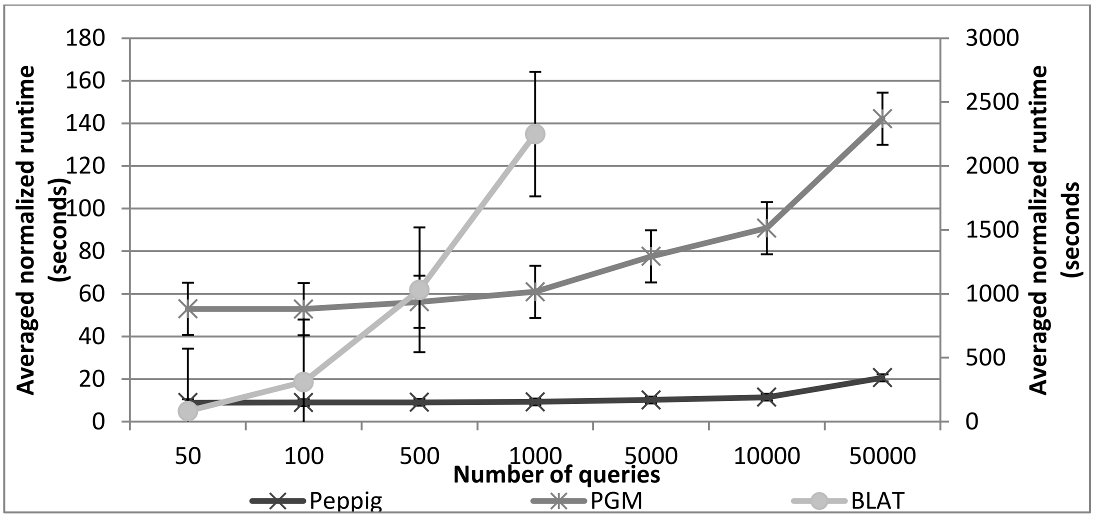


Figure 17. BLAT, Peppig, and PGM were executed on 131MB size human chromosome 10. BLAT was only tested for 1000 queries due to excessive runtime. Runs for each algorithm were repeated 5 times, and run times in seconds were averaged. They are plotted with standard deviations for each peptide query size.

Peppig algorithm is fastest for normal use cases, and its growth is linear, whereas the growth rates of BLAT and PGM seem exponential. Similar results are found for medium (Figure 17) and small chromosomes (Figure 18).

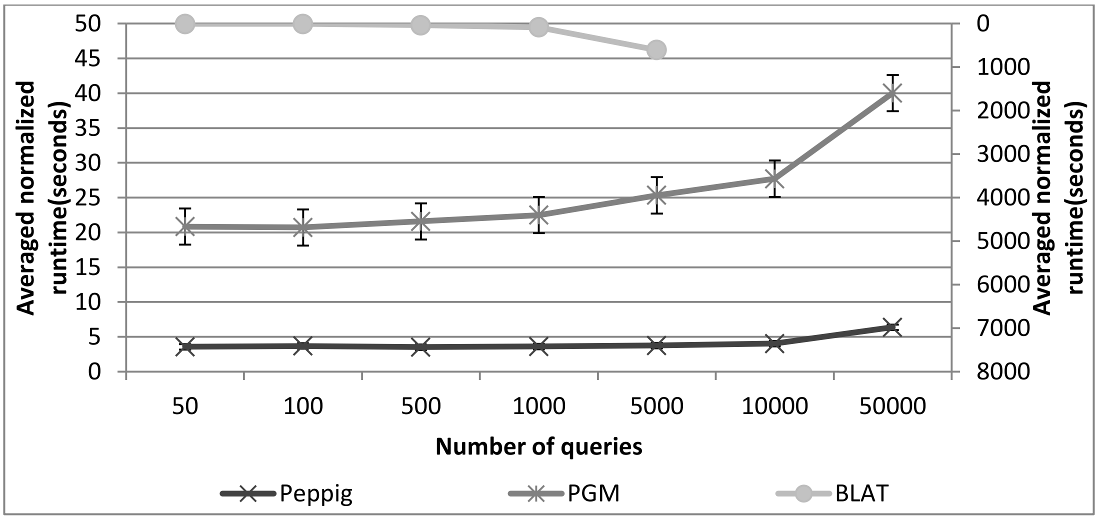


Figure 18. BLAT, Peppig, and PGM were executed on 57.5MB size human chromosome Y. BLAT was only tested for 1000 queries due to long runtime. Runs for each algorithm were repeated 5 times, and run times in seconds were averaged. They are plotted with standard deviations for each peptide query size.

In order to compare runtime performances of Peppig against BLAT and iPiG (Kuhring and Renard 2012) in translated databases, human chromosomes 1, 10 were translated into six frames with following file sizes: 492MB, 268MB respectively. Query files were built from 50, 100, 500, 1000, 5000 peptide sequences due to long run time.

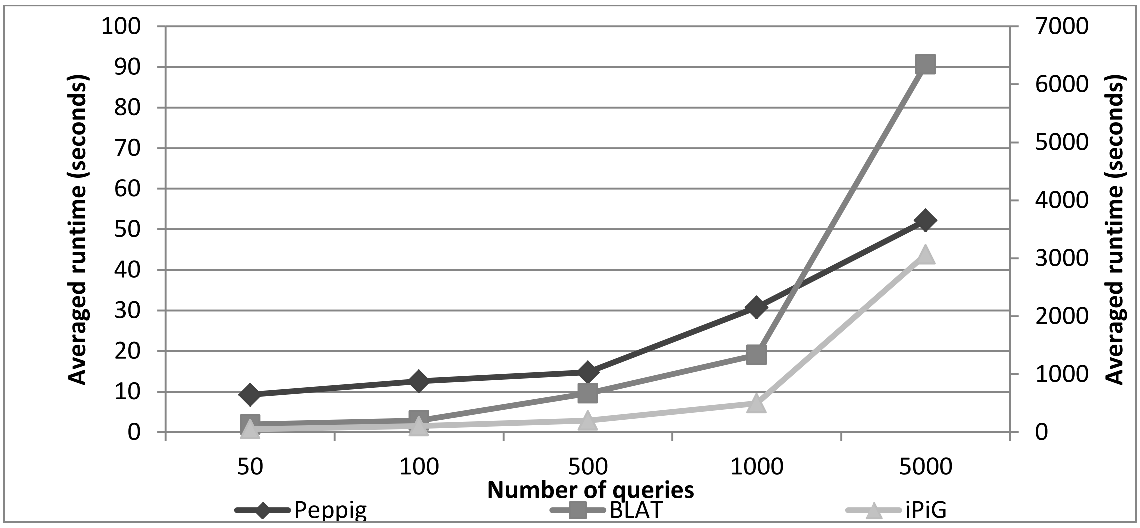


Figure 19. BLAT, Peppig, and iPiG were executed on 492MB sized six-frame translated human chromosome 1. Runs for each algorithm were repeated 5 times, and run times in seconds were averaged. BLAT and iPiG were shown in the second vertical axis. Peppig is on the primary vertical axis.

For the comparison performed on translated databases as shown in Figure 19 and Figure 20, Peppig performed better than BLAT and iPiG in terms of runtime. However, it should be noted that iPiG provides different features than Peppig, for instance, visualization. Moreover, BLAT is intended to perform approximate string matching which cannot be conducted by Peppig.

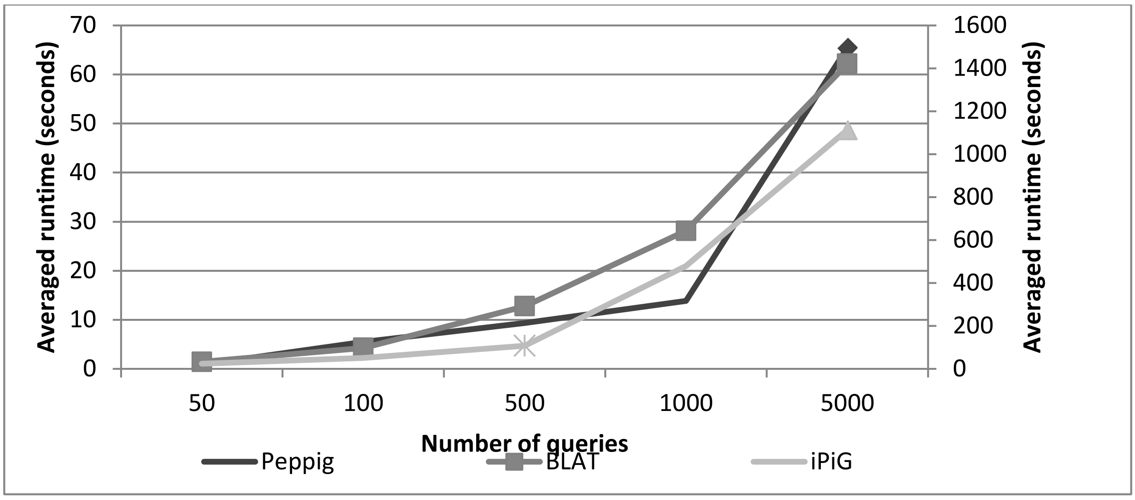


Figure 20. BLAT, Peppig and iPiG were executed on 268MB sized six-frame translated human chromosome 10. Runs for each algorithm were repeated 5 times, and run times in seconds were averaged. BLAT and iPiG were shown in the second vertical axis. Peppig is on the primary vertical axis.

Overall, when the average measurements for all databases including the translation step done by algorithms were taken, compared to PGM (7min/GB), Peppig (~ 70s/GB) is 6 times faster. BLAT is 190, and 30 times slower than Peppig and PGM, respectively.

CHAPTER 4

PGMINER: PROTEGENOMIC PIPELINE ANALYSIS TOOL

4.1. Introduction

In proteogenomic studies, the general workflow is composed of steps as database generation, a database search of mass spectrometric data to identify peptides with a level of confidence, mapping of identified peptides to databases or existing gene models and visualization of the mapping results.

In previous studies, different sequence sources at genomic (Desiere et al. 2005; Fermin et al. 2006; Gallien et al. 2009; Branca et al. 2014) or transcriptomic level (Jagtap et al. 2013; Woo et al. 2013; Zickmann and Renard 2015) have been used as sequence database. These sequences can be listed as a six-frame translation of genomic DNA sequence, three-frame translation of cDNA, six-frame translation of EST, RNA-Seq data. Database generation step took main attention in previously reported methods since size and content of sequence database influence the time needed for peptide identification and accuracy of PSMs (Castellana et al. 2010; Wang et al. 2012; Jagtap et al. 2013).

In peptide identification step of a general proteogenomic workflow, MS/MS spectra are searched against generated sequence databases. Many existing solutions use multiple database search algorithms to assign peptides to spectra. In order to assign a confidence level to PSMs, there are various approaches employed such as False-discovery rate (FDR), Percolator (Wright et al. 2012), second-round search (Jagtap et al. 2013). With the aid of false-positive PSM elimination techniques number of true-positive PSMs are saturated.

In addition to that fast and accurate exact or tolerant mapping of peptides to genomic databases and genomic annotations is a crucial step which influences proposing new gene models and correction/ validation of existing ones. Although database search algorithms return peptide location, not all algorithms return all enzymatic cleavage rule fitting positions are available. In addition to that, when consensus prediction is performed on multiple tool results it is inevitable to perform peptide mapping. In this step, peptides are categorized as exonic, intronic, overlapping or intergenic (Fermin et al. 2006; Khatun et al. 2013).

Another step in proteogenomic studies is protein inference which is described as correlating peptides which are above a certain threshold with protein sequences (Claassen 2012). The peptides can map to several proteins ambiguously. Therefore there are different algorithms to solve this problem (Serang et al. 2010; Zickmann and Renard 2015).

Therefore, confirmed or proposed gene models with mapped peptides need to be formatted in special formats like General Transfer Format (GFF), Browser Extensible Data (BED) format so that they can be visualized on genome browsers such as Ensembl or UCSC. There are also tools developed for this specific purpose (Kuhring and Renard 2012).

Accomplishing all aspects of proteogenomic, there are also platform-based tools which couple all steps into one framework (Fenyö and Menschaert 2015). For this purpose, GenoSuite (Kumar et al. 2013) was developed as an automated proteogenomic pipeline for prokaryotes which fulfills this need. GenoSuite translates given prokaryotic genome and performs a database search on six-frame translated genome by employing multiple algorithms.

In 2014, the BPP (Bacterial Proteogenomic Pipeline) was developed as an alternative to Genosuite (Uszkoreit et al. 2014). The main features of BPP are to enable users to load and use different database search tool results and to visualize pseudo-peptides rising in various experimental conditions.

Peppy (Risk et al. 2013), on the other hand, was the first tool which demonstrated ENCODE Human Tier 1 cell line data analysis as an example for proteogenomic analysis on eukaryotic data. The novelty of algorithm was shown as fast processing in the database generation step. On large genomes, the biggest problem is that computers with moderate RAM cannot handle large databases as eukaryotic organisms typically have. Therefore, Peppy creates first segments from input genome then generates possible peptides from segments in a multithread manner. The MS/MS spectra are searched against generated peptide list via scoring system proposed in Morpheus algorithm (Wenger and Coon 2013). However, Peppy does not map FDR filtered and confidence assigned peptides to annotations with further visualization.

PGTools (Nagaraj et al. 2015) has several features in genomic part (Phase II) and proteomics part (Phase I). In proteomics part, spectra file conversion module, multiple-algorithm supporting database search module, FDR calculation module, protein inference and functional annotation on protein-protein interaction level modules are available. In addition to that PGTools provides visualization of intermediate steps as Venn diagrams to display unique and overlapping peptides, bar charts to display MS/MS search result distribution, TreeMaps to display protein groups, chromosome distribution plot to display annotated peptides on genomic coordinates via visualization module. In genomic part, peptides identified in proteomics part are queried against a sequence database. The sequence database can be established from different genomic or transcriptomic level data sources. ExtractFeature module of this section suggests corrected gene annotations, novel genes, and exons. PGTools was written as command-line and some modules have GUI version as well. The modularity of the program allows users to configure modules and to execute them independently or in assembly to generate different workflows on command-line.

Open source PGalaxy (Jagtap et al. 2014) was designed on Galaxy framework to perform the proteogenomic analysis. The tool provides opportunities sequence assembly as a source to sequence database generation which is used to perform MS/MS analysis. The tool has 4 modules which are listed as peak list/database generation module, database search module which performs two-round database search by using ProteinPilot search tool, data filtering module which maps peptides by using a BLAST-P algorithm to filter unmatched peptides in order to further evaluation of those PSMs. Evaluated peptide-spectrum matches with their genomic locations are presented in GFF which can be integrated into genome browsers.

In this thesis, a new proteogenomic workflow, PGMiner, has been proposed which can be executed in KNIME workflow management platform including

(i) spectral and sequence data acquisition

(ii) six-frame translation of given databases and generation of multiple database chunks to allow parallelization of remaining workflow

(iii) generation of alternative translation products, i.e. alternative open reading frames

(iv) multiple database search algorithm support for peptide identification

(v) decoy database generation and database formatting according to algorithm requirements

(vi) statistical assessment of peptide-spectrum matches by target-decoy based FDR

(vii) consensus peptide assignment by the rank-weight approach

(viii) mapping of peptides to sequence databases

(ix) enzymatic rule filtering of mapped peptides

(x) proteotypic and locotypic peptide determination

(xi) classification of peptides as exonic, intronic, 3’ /5’ overlapping, intergenic to perform an assessment of genome.

(xii) visualization of mapped peptides on the genome.

PGMiner is customizable with further core data mining nodes of KNIME as well as Python and R scripting options available in KNIME or user-developed nodes.

4.2. Methodology

4.2.1. Data Acquisition

Mass spectrometry data are deposited in two main publicly available repositories: PeptideAtlas (Deutsch et al. 2010) available at http://www.peptideatlas.org/ or PRIDE (http://www.ebi.ac.uk/pride/archive/). There are other repositories such as ProteomeXChange (http://proteomexchange.org), ProteomicsDB (https://proteomicsdb.org) which include data from no-longer-existing Tranche repository. These repositories include RAW data of each submitted mass spectrometry data collection/project including additional information regarding sample preparation, measurement details and search results by database search algorithms if any. In some data collections, additional file formats such as such as MGF (Mascot Generic Format), mzML (Deutsch et al. 2010), and mzXML (Pedrioli et al. 2004) are available. These file formats are called as peak file formats which are in a readable format that is converted from binary RAW files. Peptide identification and quantitation tools usually accept spectral inputs in these formats.

In current proteogenomic pipelines, mass spectrometry data files are loaded by the user. To the best of our knowledge, available proteogenomic pipelines do not provide modules for data fetching from mass spectrometry repositories mentioned above. PGMiner allows users to provide as files and it provides two data acquisition nodes to retrieve spectral data from PeptideAtlas and PRIDE repositories. In both modules, the user needs to query keywords such as collection name, organism name or study related words. After query search, all relevant collections with available details are listed. Subfiles related to the collection, for instance, RAW files, aforementioned formatted peak files, search results, are displayed. Selected files are downloaded to pre-set output directory. The current version of PGMiner does not support file conversion from RAW files to other file formats accepted by many algorithms. Therefore, it is suggested to download files in peak file formats. One of widely used mass spectrometry-based proteomics data analysis library, OpenMS (Sturm et al. 2008) is available as community nodes in KNIME. By using file conversion nodes of OpenMS, file formats can be converted to each other.

Sequence data used in proteogenomics studies are genome, transcriptome and protein originated. There are many repositories that deposit sequence information which is widely used in mass spectrometry-based proteomics studies. As mentioned in Chapter 1, ENSEMBL is one of the repositories, which includes data from multiple sources to provide multi-level support for manual annotation by HAVANA and automated predictions made by ENSEMBL predictions.

Protein sequences and their extensive annotations are collected in the UniProt Knowledgebase (UniProtKB). Core data structure of UniProtKB is composed of protein name, accessions and cross-references, sequence information, taxonomic data, related publication citations, biological ontologies, protein classification and indicators of annotation quality. UniProtKB contains two sub-repositories. Unreviewed computationally generated proteins and their functional annotations under UniProtKB/TrEMBL, whilst manually reviewed, and annotated proteins are stored in non-redundant UniProtKB/Swiss-Prot.

Another source for sequence databases is provided by NCBI. Reference sequence database (RefSeq) stores manual genome curation of many organisms generated based on cDNA sequences. From RefSeq, genomic DNA, transcripts, and proteins can be downloaded. NCBI also provides Entrez protein database which is a composition of RefSeq protein, UniProtKB/SwissProt protein database and translated GenBank transcripts.

PGMiner currently supports sequence retrieval from RefSeq and Ensembl repositories. In addition that user can provide pre-downloaded sequence files. However, it must be noted that sequence files must be in FASTA file format. PGMiner does not support of sequence file conversion and resolving ambiguities in provided sequence file.

Another sequence data source in PGMiner is predicted alternative translation products, i.e. altORFs. Hereby, PGMiner amends to enable prediction of alternative start sites for selected gene models when CDS sequences are available. For this, PGMiner mostly follows the linear scanning mechanism where a 40S ribosomal subunit binds to a capped 5'-end of a translation start codon located in an appropriate context (Jackson et al. 2010; Malys and McCarthy 2011; Aitken and Lorsch 2012; Alekhina and Vassilenko 2012). Reinitiation and leaky scanning mechanisms are also considered.

4.2.2. Database Processing

As mentioned previously, genomic, transcriptomic and protein sequence data are used as sequence sources to build custom databases to peptide identification from MS/MS spectra data. Genomic databases are translated into six reading frames. Six-frame translations lead to massive increase in total file size. Nesvizhskii (2014) reported that translation of human results in approximately 3.2 gigabase database. This number is 70 fold more than ENSEMBL protein database which is 50Mb. In addition to genomics databases, RNA-Seq and ribosome profiling technology provide usage of transcriptomic data to improve annotation of the genome (Woo et al. 2013). RNA-Seq reads nevertheless does not provide information about proteins that are translated and splicing event and translated reading frame. Therefore, there have been many efforts have been presented to reduce search space and to increase sensitivity. GenoSuite, Peppy, pGalaxy, ProteoAnnotator, and BPP enable the translation of genome and assembled RNA-Seq data. PGMiner accepts both, genomic and protein sequence files. The database preprocessing node determines the type of sequence as nucleotide or amino acid automatically, and if it is a nucleotide, the node translates nucleotide sequences. The node performs six-frame translation by default; however, in case, it can be set to three-frame prior.

As mentioned in Chapter 2, many database search algorithms have limitations on handling large size sequence databases and large database elements in a sequence database. This results in termination of pipelines at peptide identification step. Therefore, reduction of search space by different methods such as extraction of potential open reading frames (Fermin et al. 2006), creation of exon junction graphs (Mo et al. 2008), generation of all possible peptides filtered according to enzymatic rule, length or isoelectric focusing (Branca et al. 2014) have been developed. Current proteogenomics tools allow data preprocessing by overcome size limitation. Peppy generates candidate proteins by digesting six-frame translated genome by stop codons. In silico enzymatic digestion is applied to generated proteins and all theoretical peptides are produced. On the other hand, pGalaxy filters sequence database according to HiRIEF (Branca et al. 2014) technique.

PGMiner uses an approach introduced in Chapter 2 to equalize databases to enable database search tools to perform a database search and to integrate results from multiple sources in a coherent manner. In addition to that equal size, databases will allow parallelization of database search step on distributed computing systems.

4.2.3. Peptide Identification

Proteogenomic pipeline tools involve multiple database search algorithm tools. However, Peppy employs only Morpheus database search algorithm and BPP does not include peptide identification step, but it outsources pre-produced peptide identifications results.

PGMiner is also designed to perform analysis using multiple database search algorithms. OMSSA, X!Tandem, MSGF+ database search algorithms have runner nodes available in PGMiner. These algorithms can be configured via configuration file provided on PGMiner website. In configuration file most of the settings are available. However, for instance for OMSSA, only one output file type is currently supported. Therefore output file setting does not exist. Output directory is set by the user and original output files and .SER file which is accepted input file type by following nodes is produced into that directory.

4.2.4. Scoring Peptide-Spectrum Matches

Statistical score assignment to PSMs for discrimination of FP PSMs from TP PSMs is one of the necessary steps in proteogenomics studies since it influences the accuracy of remaining steps. Peppy, BPP and GenoSuite use a target-decoy approach based FDR while PGTools employs PEP in addition to target-decoy based FDR. PGalaxy performs a two-round database search for reducing the number of FP hits and then it executes ProteinPilot algorithm to determine the confidence level of protein identifications. Details about these methods were given in Chapter 1. PGMiner also allows users to employ target-decoy based FDR although it has been shown that this approach is not compliant to all database search algorithms, for instance, X!Tandem. Currently, PGMiner only supports separate target and decoy database search. Therefore q-values (Käll et al. 2009) are computed for each PSM. The q-value formulation used in this node is:

$$q - value = \left(\frac{FP}{FP + TP}\right)$$

PSMs below a set threshold are considered as FP hit and they are removed from the pool.

$$TP = \ (T_{above\ threshold} - FP\ )$$

4.2.5. Consensus Prediction

Although there is a general algorithmic framework for database search tools, each algorithm employs a different scoring scheme to determine best matching peptide candidates to searched MS/MS spectra. Many factors affect scoring function and therefore accuracy of peptide identification. MS instrument, measurement type, fragmentation methods use in measurement, fragmentation efficiency, properties of peptides, peak intensities can be given as examples which are in factors affecting peptide identification accuracy. Accuracy distribution of eight database search algorithms was previously shown by us on synthetic peptide benchmark dataset that was introduced in Chapter 2.

As shown in Figure 21, PEAKS, pFind, MSAmanda, X!Tandem, Inspect, OMSSA, MSGF+ and Myrimatch were run on direct infusion Orbitrap data set subjected to CID fragmentation against a protein database. While PEAKS, pFind, MSAmanda, and X!Tandem yielded ~50% correct identifications, OMSSA, MSGF+ and Inspect resulted in approximately 43-45% correct identifications. Myrimatch yielded 39% correct predictions. All algorithms except OMSSA and PEAKS resulted in a low number of unidentified spectra. In fact, they yielded high incorrect predictions.

OMSSA and PEAKS returned less incorrect identifications but a high number of unidentified spectra.

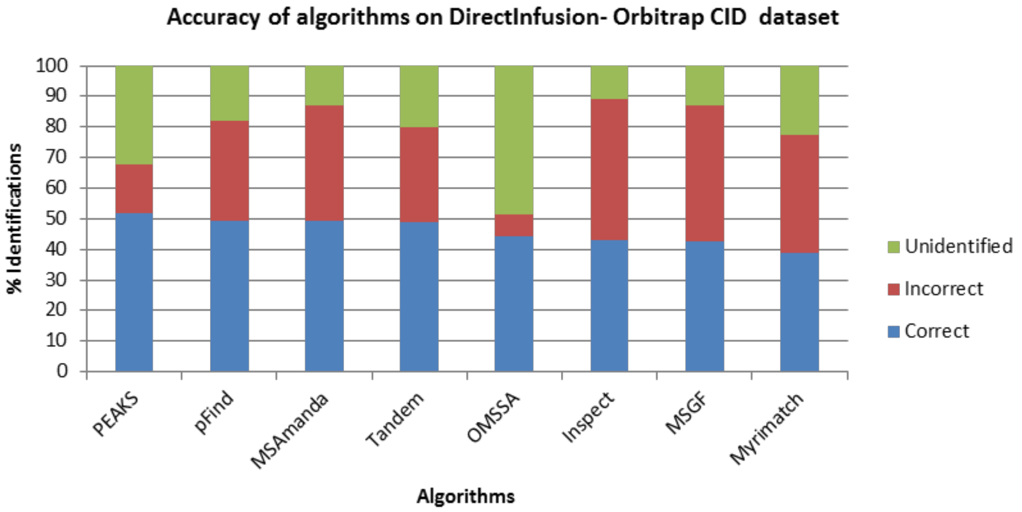


Figure 21. The accuracy of eight database search algorithms; MSAmanda, PEAKS, Inspect, X!Tandem, MSGF+, Myrimatch, OMSSA, and pFind were given on direct infusion Orbitrap CID benchmark dataset.

In Figure 22, accuracy distributions of PEAKS, pFind, MSAmanda, X!Tandem, Inspect, OMSSA, MSGF+, and Myrimatch on LC-Orbitrap CID dataset were shown. In this dataset, some of the peptides in the mixed sample were While PEAKS, pFind, MSAmanda, and X!Tandem yielded ~50% correct identifications, OMSSA, MSGF+ and Inspect resulted in approximately 43-45% correct identifications. Myrimatch yielded 39% correct predictions. All algorithms except OMSSA and PEAKS resulted in a low number of unidentified spectra. In fact, they yielded high incorrect predictions. OMSSA and PEAKS returned less incorrect identifications but a high number of unidentified spectra.

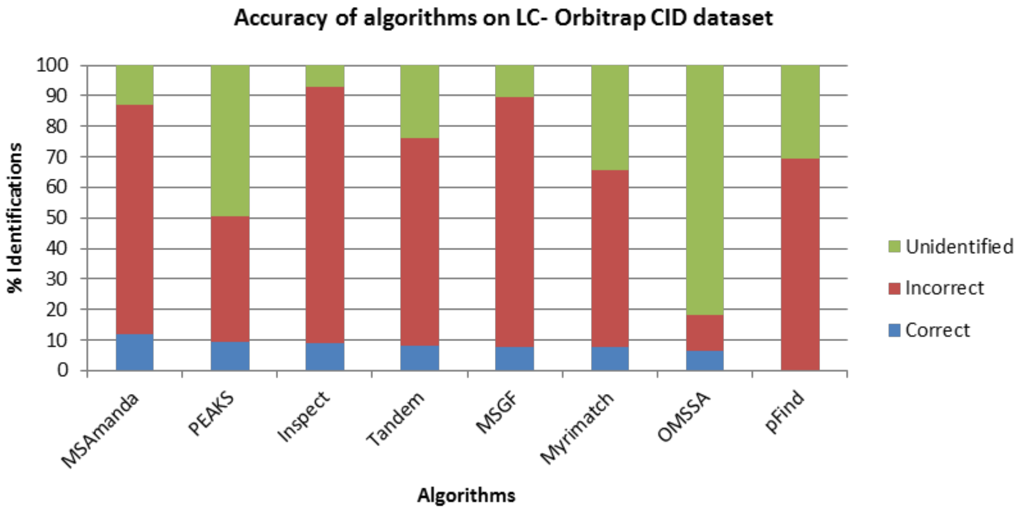


Figure 22. The accuracy of eight database search algorithms; MSAmanda, PEAKS, Inspect, X!Tandem, MSGF+, Myrimatch, OMSSA and pFind were given on LC-Orbitrap CID benchmark dataset.

Overall database search algorithms results were pulled down and shown per MS instrument-fragmentation method dataset in Figure 23. It can be concluded that direct infusion datasets were assigned to expected peptides more than LC datasets by all database search algorithms. However, maximum cumulative accuracy can reach up to 50%. The reason here could be related to varying quality of spectra as well as different criteria are taken into account by database search algorithms.

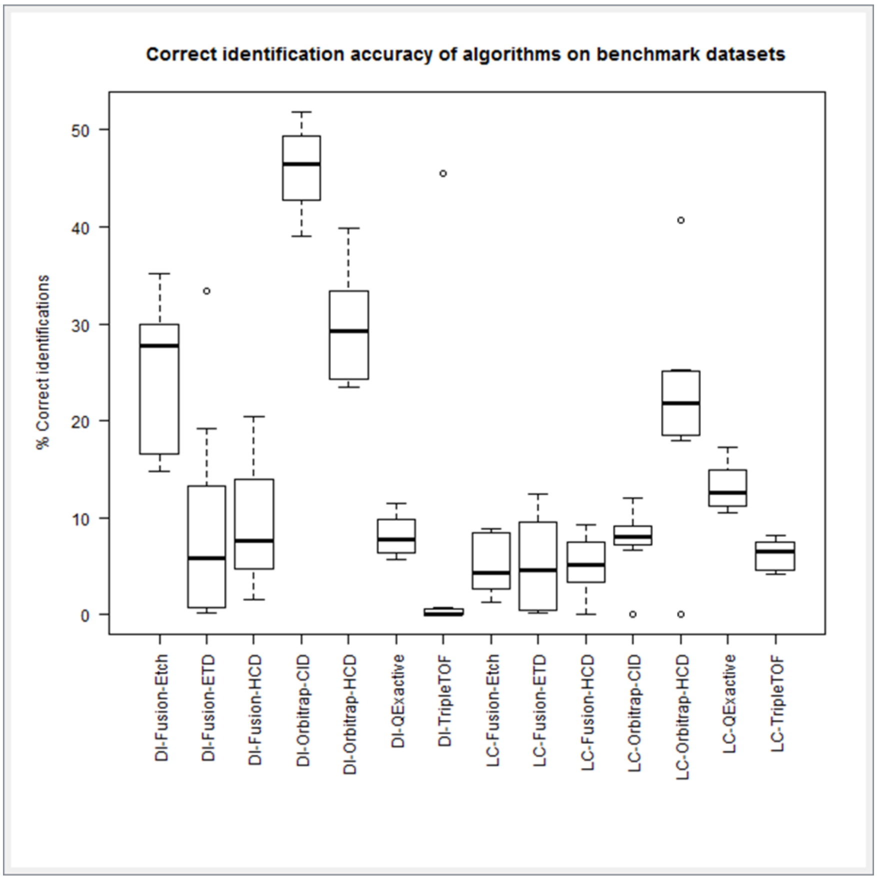


Figure 23. Cumulative % identification distribution of all algorithms per benchmark dataset is shown. According to the results, database search algorithm accuracy results increase in direct infusion dataset in compare to LC dataset. The highest accuracy of database search algorithms reaches at most to 50% accuracy in direct infusion CID dataset. On the other hand, the highest accuracy obtained in LC datasets is around 25%.

As shown previously in the literature (Sultana et al. 2009; Dagda et al. 2010; Nahnsen et al. 2011), consensus identification of multiple tools increases the number of correctly identified spectra. In our datasets, we also tested this observation (Figure 24). Consensus predictions for each dataset were computed according to the first hit based consensus scoring described in Figure 25. In this formulation, for each spectrum, best PSM hit of each algorithm are taken and majority vote peptide is assigned to spectrum.

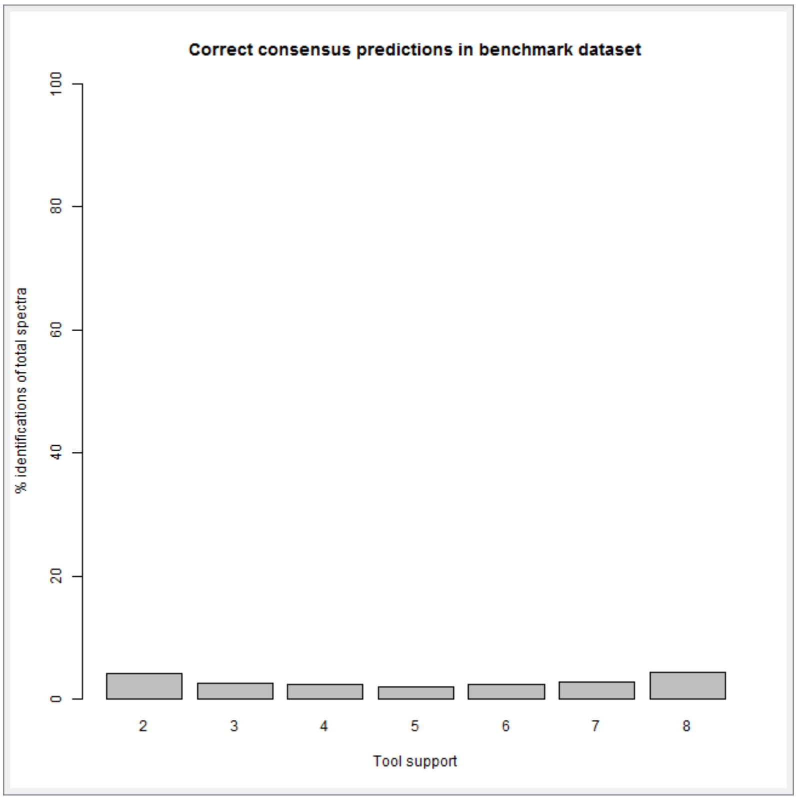
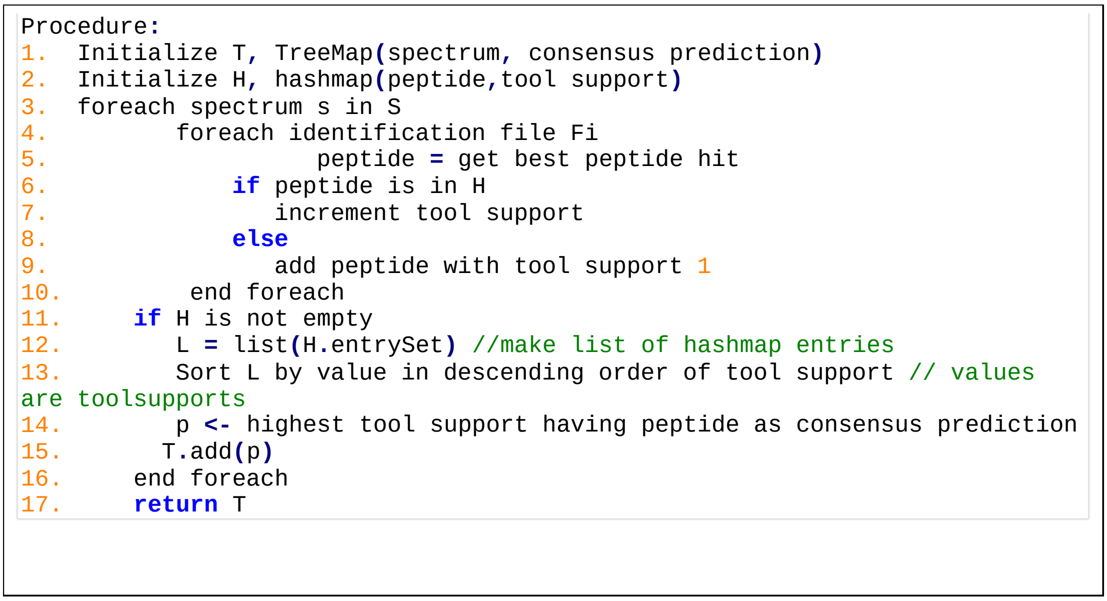


Figure 24. Exclusive consensus predictions from tool support two to tool support eight are given.

Tool support distribution in % correct identifications was shown exclusively in Figure 24. According to the results, we concluded that any tool integration for quality ranging datasets, integration of algorithms' results to form a consensus assignment outperforms all individual algorithm predictions. In direct infusion Orbitrap CID dataset while maximum correct prediction yielding tool reaches to 51%, inclusive best-hit majority vote consensus method reached to 54%. On the other hand, for LC-Orbitrap CID dataset, the best tool gives 12% correct identifications, consensus prediction results in 11% correct predictions with increasing confidence rather than using a single tool.



```
Procedure:
Initialize T, TreeMap(spectrum, consensus prediction)
Initialize H, hashmap(peptide,tool support)
foreach spectrum s in S
       foreach identification file Fi
                 peptide = get best peptide hit
           if peptide is in H
              increment tool support
           else
              add peptide with tool support 1
        end foreach
    if H is not empty
       L = list(H.entrySet) //make list of hashmap entries
       Sort L by value in descending order of tool support // values are toolsupports
       p <- highest tool support having peptide as consensus prediction
      T.add(p)
    end foreach
    return T
```

Figure 25. Pseudocode of best-hit majority vote based consensus prediction calculation is given.

Another consensus prediction calculation approach would be weighting ranks of best n hits of a spectrum for all algorithms. The idea here is some algorithms might assign similar scores to different peptides and report them in a different order. Therefore, we hypothesized that instead of taking a first best hit of each algorithm for a spectrum, best n peptide hits would be scored. The idea of this consensus method is to pool down all peptides, and for each peptide summing up a rank score of each algorithm for the peptide. In case an algorithm has no hit for a peptide than the rank score is penalized as 100. For instance, peptide *x* is found by two algorithms out of three. Algorithm *a* ranks peptide x at first rank; algorithm *b* returns the same peptide at fifth rank. Algorithm *c* does not return the peptide, 100 is assigned as penalty score. Therefore, the cumulative weighted rank score becomes 1+5+100= 106. In this consensus method, the lowest score having peptide is considered as the best candidate hit.

```
Procedure:
for each spectrum s in S
   create hashmap H to score consensus scores of peptides
   for each peptide P in the hit set Atotal(hits)
     consensus score Pcs = 0
     for each algorithm A in hit set Ahits
         if P in the Ahits
              Pcs+= (rank RA)
         else
              Pcs+= (100 as penalty)
      end for
      store Pcs and P in H
   end for
   sort H by value
   P with lowest consensus score is returned as consensus peptide-spectrum match
end for
```

Figure 26. Pseudocode of ranking-weighted consensus prediction calculation is given.

In PGMiner, since three algorithms are suggested to use, three would be the highest score. By default, we considered maximum best 10 hits per algorithm. Therefore the lowest acceptable threshold score is 120 which two algorithms assign peptide to 10th rank while the third algorithm does not include the peptide in best 10 hits. Therefore, we ensure minimum tool support as two. The pseudocode of this algorithm was given in Figure 26.

4.2.6. Peptide Mapping and Genome Annotation Assessment

Mapping peptides to gene annotations is the central intersection of genomics and proteomics in the field of proteogenomics. Most proteogenomics pipelines employ some specialized tool for mapping identified peptides to the underlying genomic database. PGMiner involves Peppig (formerly Lelantos) algorithm described in Chapter 3. As mentioned in Chapter 3, Peppig is a modified WuManber algorithm (Wu and Manber 1994) exact string matching algorithm implementation to achieve unequaled processing time on exponentially growing query numbers. Additionally, enzymatic cleavage rules of the enzyme used during sample preparation are checked following the mapping procedure in order to eliminate potentially false mappings, especially on protein database. Considering the alternative splicing event, some of the peptide mappings might be seen conflicting with the tryptic rule.

As in proteomics studies, in proteogenomic studies listing proteins with unique peptides in a certain probability ratio is important to determine which proteins are unambiguously expressed by a particular gene under different conditions (Nesvizhskii 2014). In this study, a protein inference algorithm was not employed to group proteins unambiguously related to a different level (genomic, transcriptomic) of support. Since the aim of this study is to show usage of peptide mapping tool to point release differences for a database, a coarse definition for "unambiguous" protein has been followed. Therefore proteins and genes which are mapped to more than 2 peptides with the rule of at least one of them are proteotypic are considered as unambiguous identifications.

In the present analysis, peptide identification status was defined either as locotypic or proteotypic according to three fundamental rules; genome-wide location, the number of mapped gene identifiers and support of mapped database origin level such as genomic, transcriptomic and proteomic (Table 10). Peptides which have only one location throughout the genome but can map to multiple genes are named as locotypic regardless of origin support. The reason is that one exon can be shared by multiple genes. Therefore, one peptide can be related to several genes or variants. On the other hand single, genome-wide location in addition to single gene mapping would put a peptide into both the proteotypic and the locotypic category. A genome may contain additional sequences (e.g.: patches) which introduce artificial redundancy. Therefore, a peptide may be mapped to multiple versions of the same gene. In such case, if mapped gene number is one, however, reported locations are multiple; peptide is still named as proteotypic but not locotypic. Often it has been observed that a peptide cannot be mapped to a genomic database. However, mapping result against protein database can be reported (e.g.: due to introns (Allmer et al. 2004)). In case a peptide is a product of splicing and related to only one gene identifier, then the peptide is accepted as proteotypic. Consequently, peptides that are mapped to multiple locations in the genome and have relation to multiple genes cannot be accepted as proteotypic or locotypic. For unambiguous identification of proteins, at least two peptides including at least one proteotypic peptide were required.

Table 10. Peptide classification status is explained according to genome-wide location, a number of genes mapping to peptides, mapped sequence origin. Explanation of each status is also provided.

Case	Genome-wide location	Number of mapped Gene-IDs	Genome/Transcriptomic/Proteomic Origin	Status	Explanation
1	Single	Multi	Multi/Single	Locotypic	One exon can be shared by multiple genes and their transcripts and protein products. Such peptides identify a locus but cannot differentiate among variants.
2	Single	Single	Multi/Single	Proteotypic, Locotypic	The peptide uniquely identifies a single protein without known variants.
3	Multi	Single	Multi/Single	Proteotypic, Not-Locotypic	Genome mapping can be on primary assembly and patch/scaffold. And two different regions might contain same geneID and/or transcript(s) and/or protein(s). Those multiple transcripts and proteins can be predicted and revised (validated), thus have the same sequence.
4	Multi	Multi	Multi/Single	Ambiguous	
5	-	Single	Only transcript and proteome level	Proteotypic	The peptide can be splicing product and cannot be found in a direct genome search. However, in case protein and transcript identifiers direct to the same location, it is proteotypic.

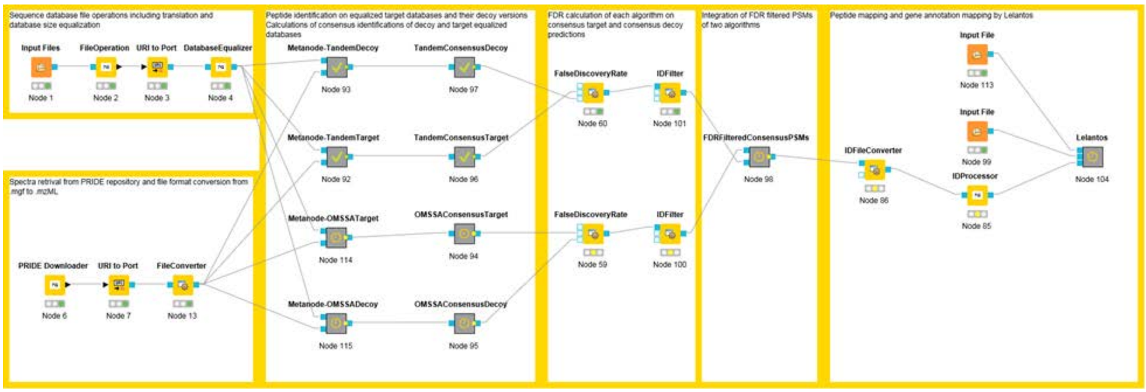
Mapped peptides are intersected with known information such as gene models. Such annotations must be provided in GFF format. Annotations enable PGMiner to categorize peptide identifications into classes confirming known annotations or conflicting with them. This information is one of the most important outcomes of any proteogenomics analysis and PGMiner further provides information whether conflicting peptide identifications are intergenic, intronic, overlapping with exons, etc.

4.2.7. Visualization

pGalaxy and GenoSuite allow the visualization of identified peptides and proteins in their genomic context. PGMiner also lets users view their output GFF on Integrative Genomics Viewer (Thorvaldsdóttir et al. 2013) within the genomic context. In addition to that Artemis Genome Browser (Rutherford et al. 2000) can be used to visualize annotated peptides in their genomic context.

4.3. Discussion

Proteogenomics field gains importance driven with the advent of high-throughput technologies for next-generation sequencing and mass spectrometry-based proteomics. This enables generation of the tremendous amount of data and public share of these data through data repositories. The result of the availability of these data yield depth of information which requires sophisticated data analysis and mining techniques. However, accessibility and operability of these bioinformatics tools lagged behind of production of data by multi-omics technologies. The not only success of the available tools but also merging multiple tools responsible for different analysis steps heighten the challenge in proteogenomics studies. Many integrative pipelines are available to accomplish complete analysis. In this chapter, a new proteogenomic workflow, PGMiner (Figure 27), is presented. A list of comparison for existing features of available tools is given in Table 11.



Sequence database file operations including translation and database size equalization

Spectra retrival from PRIDE repository and file format conversion from .mgf to .mzML

Peptide identification on equalized target databases and their decoy versions Calculations of consensus identifications of decoy and target equalized databases

FDR calculation of each algorithm on consensus target and consensus decoy predictions

Integration of FDR filtered PSMs of two algorithms

Peptide mapping and gene annotation mapping by Lelantos

Figure 27. Screen shot of the workflow designed in KNIME. Gray nodes were called meta nodes which in turn can contain smaller workflows where a specific subtask is carried out (e.g.: Peppig - formerly Lelantos), while single nodes were shown as yellow. As a result of successful execution, nodes were turned to green color.

PGMiner was developed on KNIME Data Analytics platform. In this respect, pGalaxy and PGMiner have a shared property since pGalaxy was drawn up on Galaxy Data Analysis platform. This property becomes prominent and ideal since they are both supported in workflow management systems. Therefore, it is possible to extend these pipelines with further data analysis nodes for special purposes. Here the important thing is to continue accessibility, maintenance, and updates so proteogenomic pipelines can be supported in a time-independent manner. Notable is that these proteogenomic pipelines demand a combination of further omics fields, not only limited to genomics, transcriptomics, and proteomics: Lipidomics, metabolomics, cheminformatics need to be merged to existing pipelines. Such integrations will also require cross-omics visualization techniques which are currently not available in common.

Table 11. Comparison of currently available proteogenomic pipelines is given for basic steps of the proteogenomic workflow.

Pipeline	**Organism**	**Data acquisition**	**Database preprocess**	**Database search algorithms**	**Statistical assessment**	**Peptide mapping**	**Extended features**
GenoSuite (2013)	Prokaryotes	User input	6-ORF translation	OMSSA X!Tandem InsPecT MassWiz	FDR -Peptide level -Protein level	No algorithm mentioned	
Peppy (2013)	Eukaryotes	User input	Generate peptide segments	Morpheus algorithm	FDR	No algorithm mentioned	
BPP (2014)	Prokaryotes	User input	-	Outsource	User dependent	No algorithm mentioned	Proteotypic peptides
ProteoAnnotator (2014)	Prokaryotes Eukaryotes	User input		SearchGUI toolkit	FDR	Against in silico gene annotation	
pGalaxy (2014)	Prokaryotes Eukaryotes	User input		ProteinPilot	Two round searchProteinPilot	Blastp *Ab initio* proteins	
PGTools (2015)	Prokaryotes Eukaryotes	User input		X!Tandem OMSSA MSGF+ Comet	FDR PEP	Blastp *Ab initio* proteins	
PGMiner	Prokaryotes Eukaryotes	-Repository fetch -User input		OMSSA X!Tandem MSGF+	FDR Peptide level	Wu-Manber All databases	Proteotypic peptides AltORFs

CHAPTER 5

ENHANCEMENT OF HUMAN GENOME ANNOTATION BY USING PGMINER PIPELINE

5.1. Introduction

First proteogenomic study on human genome was presented by Choudhary et al. (2001). Since the human genome was not completed yet by that time, it has been shown whether it is possible to examine current annotation with mass spectrometry data. LC-MS/MS collection measured for 22 human proteins were searched against protein database, EST database, and template of International Human Genome Project draft by using Mascot database search algorithm. As a result, they observed that out of 169 spectra 114 spectra were matched to human protein database originated peptides while 11 spectra were hit to bovine trypsin. They could not get any results for 44 spectra. Exon-intron boundary matching peptides, peptides that are missing protein database, peptides missing in nucleotide database, N-terminal peptides were also identified.

Desiere et al. (2005) performed a more detailed proteogenomic study on human genome annotation. In contrast to study by Choudhary et al., peptide-spectrum match quality has been considered. 52 proteome collections composed of proteins obtained from difference cell types such as T cells, B cells, lymphocytes, hepatocytes were searched by SEQUEST database search algorithm (Eng et al. 1994) against human IPI database. Resulting PSMs were recorded by the PeptideProphet algorithm to determine the statistical confidence of hits. As a result 224973 PSMs having p-value score greater than 0. 9 were kept. Among these PSMs, 26840 peptides were defined as proteotypic on the protein level. These peptides were searched against ENSEMBL human protein database by BLAST algorithm. 25754 peptides out of 26840 peptides were found in this database. For unfound 1086 peptides, it has been concluded that the reason could address the existence of single nucleotide polymorphism or novel splicing isoforms. Besides that non-parallel synchronization of IPI database and ENSEMBL human protein database could also lead to unfound peptides. Of the 9747 proteins in ENSEMBL human protein database matched to peptides. In addition to that many peptides matched to many proteins with high scores leading to ambiguous protein identifications. When these proteins were further analyzed, it has been observed that some of those proteins are paralogs having shared domains. Of the 3718 proteins, on the other hand, included at least one proteotypic peptides. Genomic coordinates of matching peptides were calculated and examined. According to the results, 4800 peptides were determined at exon-intron boundaries. Besides, that in this study tissue and disease specific proteins were detected.

The first study targeting human blood plasma proteome was performed by Fermin et al. (2006). In this study, blood plasma and serum proteins were collected from donors belonging to different ethnical and geographical groups such as Africa, Asia. A total of 2,230,502 MS/MS spectra were searched against human genome by X!Tandem database search algorithm. However, the human genome was not used directly; instead, all possible open reading frames were generated out of the human genome. Spectra matching to multiple open reading frames were eliminated, spectra matching to proteotypic peptides were used in following steps. Open reading frames which contain proteotypic peptides and have high scores were analyzed. In this study, only intragenic peptides were analyzed. Peptides were categorized into three main groups as; exonic, exon-intron boundary matching and non-exonic peptides. In order to define exon-intron boundary matching peptides, EST libraries were used in the search and identified ESTs were annotated in UniProtKB database. This study is the first and the only proteogenomic study targeting human plasma proteome so far according to the best of our knowledge. However, in terms of data size and methodology, the study is limited.

In this study, human blood plasma proteome collections available in PeptideAtlas and PRIDE repositories were searched against the whole human genome, CDS collection, protein database and GENSCAN predictions, alternative open reading frames as well as human microbiome collection including bacterial and viral proteins found in human.

5.2. Methodology

5.2.1. Human Genome and Its Annotation

Masked removed all human chromosomes and non-chromosomal and mitochondrial genome, CDS sequences and protein sequences were obtained from ENSEMBL repository as human database build 38 version 86. In addition to that GENSCAN predictions were retrieved from ENSEMBL database. Genome annotation in general transfer format (GTF) was downloaded from ENSEMBL annotation repository. These annotations include both HAVANA and ENSEMBL annotations of all genes, their transcripts, and exons present in these annotations. Alternative translation products were generated via AltORFEv module of PGMiner. Since blood is a tissue in which circulates many proteins including immune system proteins, we assumed that bacteria and virus originated peptides could be found in blood. Therefore, spectra matching to microbiome peptides need to be removed. Human host bacteria and virus originated proteins were retrieved from UniProtKB and Human Microbiome Project databases. All databases were formatted into FASTA file format.

5.2.2. Mass Spectrometric Data (Human Mass Spectra Collection)

The high-resolution instrument measured mass spectrometry data of human blood tissue were retrieved from PeptideAtlas and PRIDE repositories Table 12. These collections were collected from human blood plasma, serum and platelets from different donors having different health conditions and sample preparation protocols were different. All selected collections were measured in LTQ Orbitrap CID. We estimate the currently available data to be approximately 100 GB in size which is roughly equal to 3 million MS/MS spectra. Of these 3 million spectra we estimate 50% to be of useful quality and will be assigned to peptides.

Table 12. Human blood plasma, serum and platelet proteome collections, their descriptions and MS instrument used in this study

Collection	Source	Collection Description
PXD000766	PRIDE	Blood plasma(corona plasma protein)
PXD003666	PRIDE	Blood plasma (MASP-3 protease inhibitor in lectin pathway)
PXD001171	PRIDE	Human serum proteome(hepatocellular carcinoma)
PXD001794	PRIDE	Blood serum(serpin family detection for colorectal adenoma)
PXD002475	PRIDE	Blood serum(HCV discovery)
PXD002762	PRIDE	Blood plasma(biomarker panel for chronic graft versus host disease)
PAe003765	PeptideAtlas	Blood serum proteomics survey
PAe003762	PeptideAtlas	Blood serum proteomics survey

These spectra were obtained in RAW file format and converted to MGF file format by ProteoWizard-(Kessner et al. 2008) MSConvert module. MGF files further split into 2000 spectra having files to decrease runtime of peptide identification step. The scan numbers were renewed starting from zero to 1999. Since algorithms assign index numbers different from scan numbers, spectra having same scan and index numbers were ensured.

5.2.3. PGMiner Pipeline

By using database processing module of PGMiner, all genome origin databases were translated into the six-reading frame. More than 20 X (unknown amino acid) sequences were reduced to 20 X sequences to decrease file size. CDS sequences were translated into the three-reading frame. All FASTA elements were split into fragment by database equalizer module. 50 same size databases were generated.

All spectra were searched against all 50 same size target databases. In this study, we did not perform target-decoy based FDR since we already showed in Chapter 2 that database search algorithms used in this study are not compliant to target-decoy based FDR. In addition to that consensus, prediction implementation ensured high confidence. Therefore, FDR computation was not considered as compulsory. Algorithms were configured to 10 ppm precursor mass tolerance and 0.8 Da fragment mass tolerance by allowing 2 miscleavages. The settings were determined via genetic algorithm briefly described in Chapter 2. Carbamidomethylation of cysteine was set as fixed modification, while oxidation of methionine was set as variable modification.

PGMiner peptide identification modules were executed on Amazon Web Services to save the runtime,. In this configuration, spectra and databases were stored in the S3 storage, and EC2 spot instances were purchased depending on the run and prices. For each spectral file, each database and each algorithm combination an SQS job message was created, one EC2 instance was purchased to execute the job in SQS. EC2 instances were able to parse similar SQS messages and to retrieve spectra and database files from S3 bucket to copy them itself. Every worker instance has the ability to start a job, and each job is indexed with a number so that results of the corresponding number were able to be followed. After completion of each run, in other words, job, Simple Notification Service (SNS) reported the status of the job. In case there is no job to be executed, EC2 instance was able to shut down itself.

For each MGF file, 50 database search results of each algorithm were merged to obtain maximum 50 hits per spectrum. One of the benefits of having multiple files is to collect multiple results from approximately equal size databases. For instance X!Tandem only returns one identification for a database, however in this case at max 50 results for collected for X!Tandem. These 50 PSMs of each spectrum were sorted by the E-value score for MSGF+, OMSSA, and X!Tandem. Maximum best 10 hits were selected per spectrum for each algorithm. Then weighted rank based consensus prediction calculation was performed. All consensus peptides were collected into a FASTA formatted file and mapped against original versions of all sequence databases by Peppig module in PGMiner. Enzymatic cleavage rule was applied to filter mapped peptides according to K/R (lysine-arginine ending rule) to eliminate spurious mappings.

In order to store results, a relational database as shown in Figure 28 was created in MySQL, which integrated all types of data as spectra, PSMs and their all scores, algorithms, peptides, peptide-database mappings associated with the proteomic and genomic context. The reason for creation of this database results in many advantages as following:

- The spectral information with descriptive details regarding the collections is available.
- Algorithm settings are stored in the used databases.
- PSMs identified by each algorithm, spectra and database combinations

can be linked in the database. PSMs are stored with all available scoring schemes returned by algorithms.

- Non-redundant peptides can be reported, by then FASTA files can be generated, mostly identified peptides can be listed.
- Peptide-protein matches are stored. Therefore a number of unique and proteotypic peptides for each protein can be computed easily, and identified proteins can be calculated. The protein-database relationship can be visualized.
- Gene model-peptide associations per chromosome are stored. Therefore gene-centric view of results can be computed. Peptide genomic start and end locations are stored. Peptides which are unique to gene or gene isoforms can be reported efficiently.
- The coverage for chromosomes or gene locus with a proteotypic set of peptides can be computed.

The database connection and data analysis processing codes were written in JAVA by using JAVA SQL connector library. Initially, the database is populated with spectra files, and spectral information is represented in Spectra table and Source table. In the Source table, spectral collection code, MS instrument, data repository information and spectral count are stored. The table is connected to the Spectra table via foreign key as SourceID tuple. In the Spectra table, each spectrum is represented with the file name, charge information, scan identifier, retention time and spectrum peak file. In addition to that identification tool support and consensus, identification tuple is also included as initialized as null.

Database search algorithms are introduced in the database in two tables; DatabaseEngine and EngineSettings. While DatabaseEngine table only stores primary key and name of the database engine, EngineSettings table is connected to DatabaseEngine via foreign key as EngineID. In EngineSettings table, the type/name of the database and algorithm settings are stored. Identified PSMs returned by each database search algorithm with given database and settings to searched spectra are stored in Identifications table. This table is a join table which includes identifiers of Spectra table, Peptide table, and Score table. Score scheme by each algorithm is stored in Score table which is linked to EngineSettings table via a DBEngineSettingsID foreign key. In Peptides table, peptide sequences are stored uniquely. Another information stored regarding peptides is that genomic start and end locations and chromosome information that each peptide is mapped against. In case protein database usage for mapping, Proteins table is created. In Proteins table, the data repository, organism and protein accession number are kept. The connection between peptides and their origin proteins are stored in ProteinPeptideMatches table.

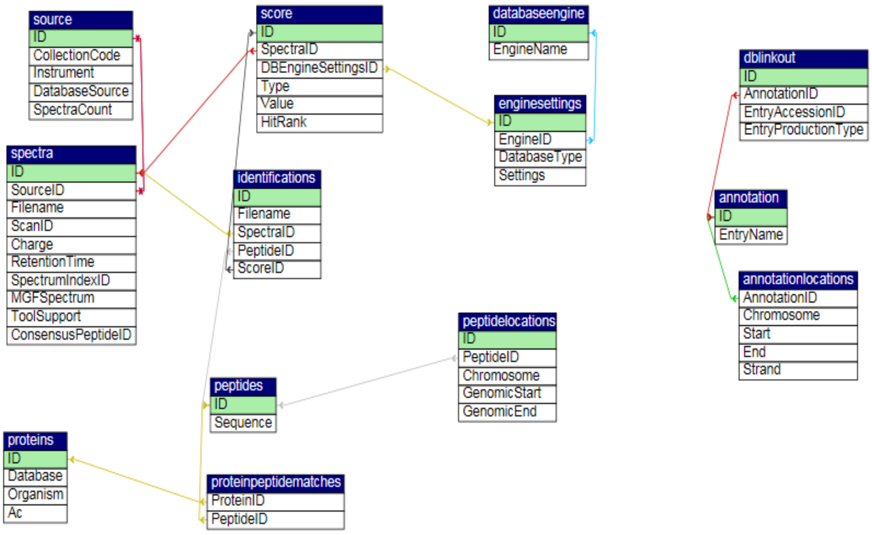


Figure 28. Relational database schema. The database is composed of three main groups as identifications, annotation data and identification mapping to annotations. Notable is that this design is not entirely normalized for optimal performance and specialized usage.

Chromosome annotations including gene, transcript level are represented in the DBLinkOut table. The start and end locations, strand and chromosome information, are stored in AnnotationLocations table. Mapped peptides were compared against these locations, and peptide genomic start and end locations were stored at PeptideLocations table.

5.3. Results and Discussion

By using eight spectral collections reaching to 3 million spectra up to +3 charge were analyzed by MSGF+, X!Tandem and OMSSA algorithms against all human chromosomes, exosome including human host bacteria and viruses, all gene predictions predicted by GENSCAN algorithm, all known human proteins and alternative translation products. Of the 98.84% of spectra were assigned to a peptide at least by one of these algorithms.

MSGF+ identified 155,077 PSMs as discrete while the number of discrete peptides is 476 for OMSSA and 117,749 for X!Tandem. Among these three algorithms, 2,841,667 spectra were identified at least by two algorithms (Figure 29). As mentioned in Chapter 4, prediction rank weighted consensus PSMs was used to compute consensus prediction, which has been considered as having higher confidence than single tool usage. From the spectra identified by at least two tools, 31% of total spectra, 946250 spectra, were found as consensus prediction assigned spectra Figure 29.

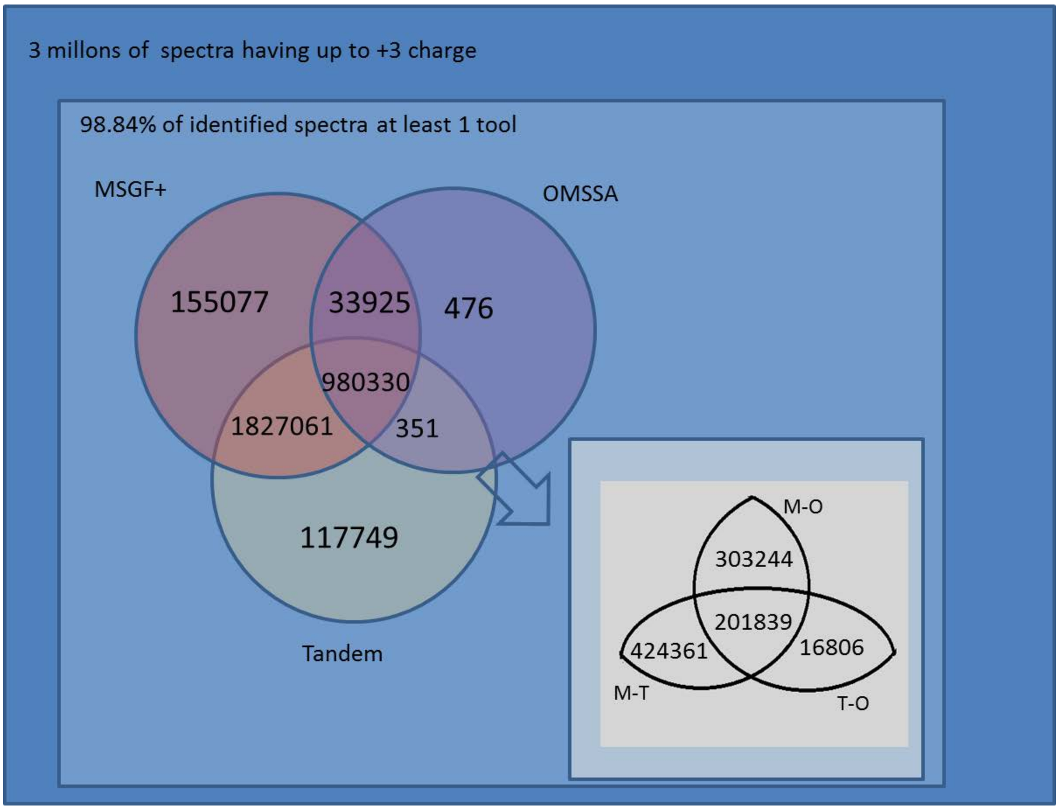

3 millons of spectra having up to +3 charge
98.84% of identified spectra at least 1 tool
MSGF+
OMSSA
155077
33925
476
980330
1827061
351
117749
Tandem
M-O
303244
201839
424361
16806
M-T
T-O

Figure 29. The identification of 3 millions of spectra and rank-weighted consensus identification computed from at least two tool support spectra.

The spectral count is one of the measures which determines the estimate of the relative abundance of peptides/proteins in the sample. Here, the aim is not to compare the number of PSMs in each spectra collection. However, to have an idea of peptide spectral count relation, spectral count distribution was computed for consensus identifications. As shown in Figure 30, the total number of 137,527 unique consensus peptides spectral count distribution was calculated. The highest number of peptides as

131,085, spectral count varies between one to 10 spectral support. The next highest spectral count bin was found as 11-20 spectral support for 2478 peptides. Notable is that on the determination of novel gene models or suggestion of changes in existing models spectral count information for mapping peptides increases significance.

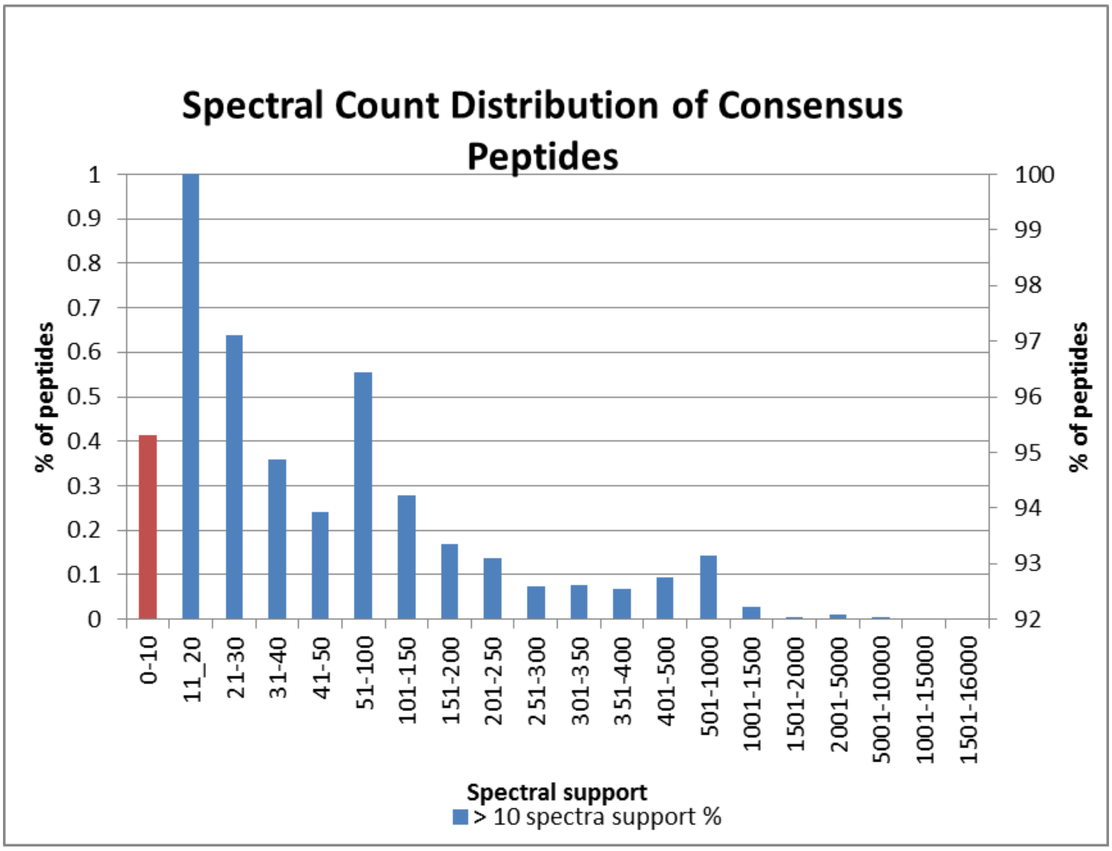


Figure 30. Binned spectral count distribution vs a number of peptide frequency is given. Most of the peptides were found with one to 10 spectral support suggesting significant peptides that can be considered as biomarker candidates or one-hit wonders.

The identified peptides were mapped against used databases individually. The first aim of peptide mapping, as explained in Chapter 3, is to find genomic locations of peptides by performing map against the genomic database to be able to compare genomic regions in terms of gene expression. The second aim is to determine proteins including identified peptides. However, it should be noted here that peptide sequences van be found in regions violating enzymatic cleavage rule. Although non-tryptic peptides or semi-tryptic peptides can occur during cleavage process, in this study only fully tryptic peptides were considered. In Figure 31, the distribution of tryptic mapping peptides in percentage per database was shown. It was investigated that 28 human blood group system genes have been localized in 14 autosomal chromosomes as chromosome 1, 2, 3, 4, 6, 7, 9, 11, 12, 15,17,18,19,22 and two blood group system genes have been found in X chromosome (Lögdberg et al. 2011). In addition to that, there are proteins encoded by other autosomal chromosomes which are circulating in the blood. On the other hand, there are peptides shared by different proteins which are not circulating blood. Therefore, in order to define a complete list of blood-related proteins, further serological, biochemical and molecular assays need to be set.

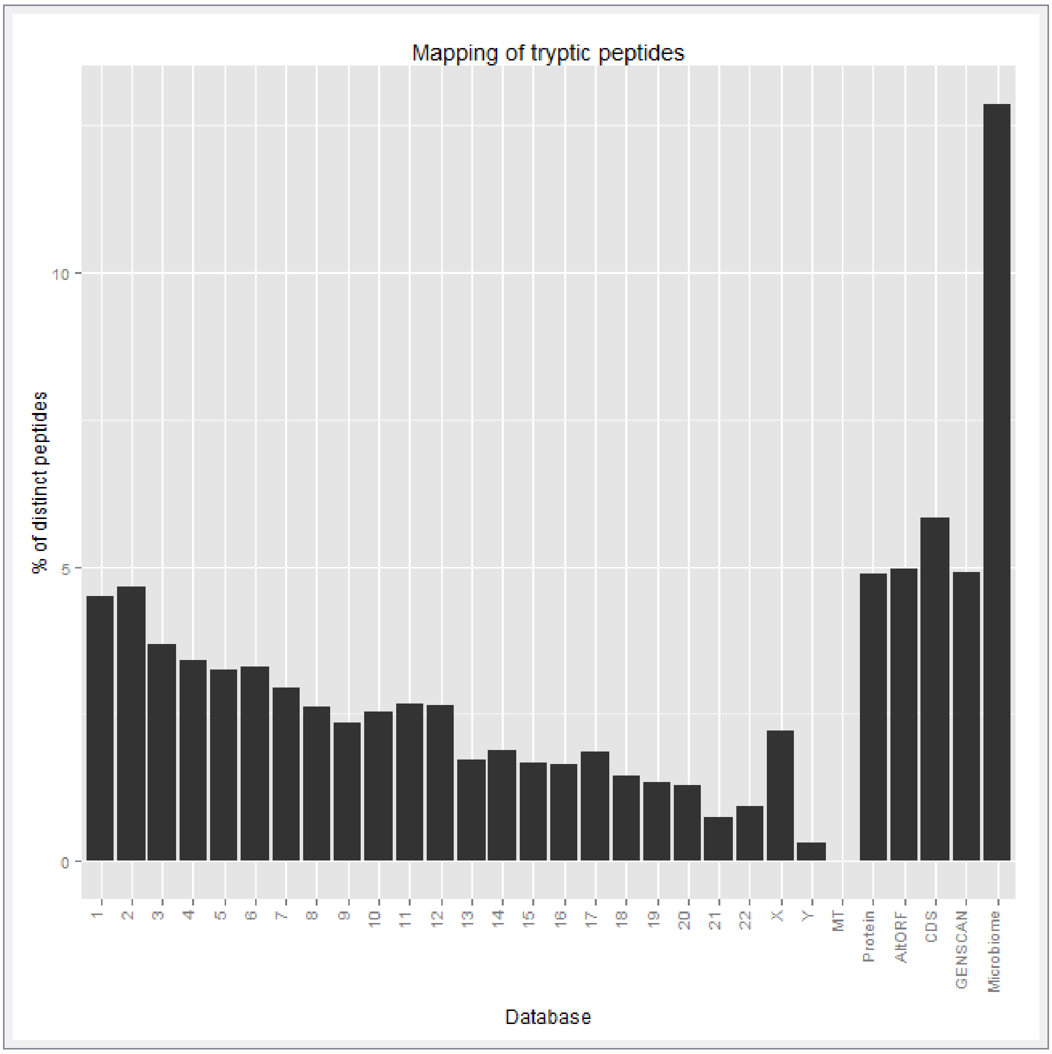


Figure 31. A number of tryptic peptides in percentage per mapping to all databases used in this study. Each peptide is considered only once, so-called distinct. The databases include all human chromosomes, all known ENSEMBL human protein database, alternative translation products (AltORF), three-frame translated coding sequences (CDS), all gene prediction products by GENSCAN, human host bacteria and virus database (Microbiome).

Besides autosomal chromosomes, identified peptides were found in known ENSEMBL human protein database, isoform and proteins encoded via alternative translation process, coding site database, predicted proteins via GENSCAN algorithm and exosomal proteins in the microbiome. The highest percentage is found for microbiome and summed chromosomal databases.

Of the 115,601 consensus peptides, 36565 peptides were shared by all databases (Figure 32). While 61936 peptides were genome specific, 211 peptides were a coding site (CDS) specific, 191 peptides were specific to predicted proteins via GENSCAN, 70 peptides were found only in the database of alternative translation products and 16628 peptides are specific to the microbiome. Peptides which were mapped to protein database were also found in other databases as well. Therefore, there is no protein database specific peptides were found.

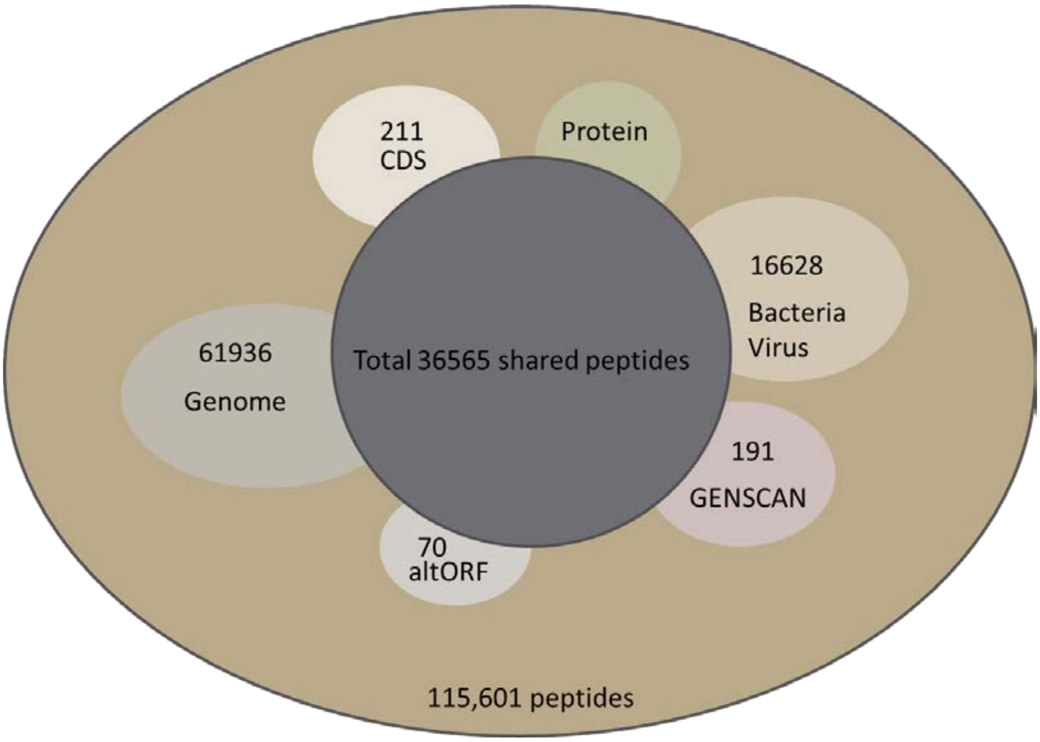
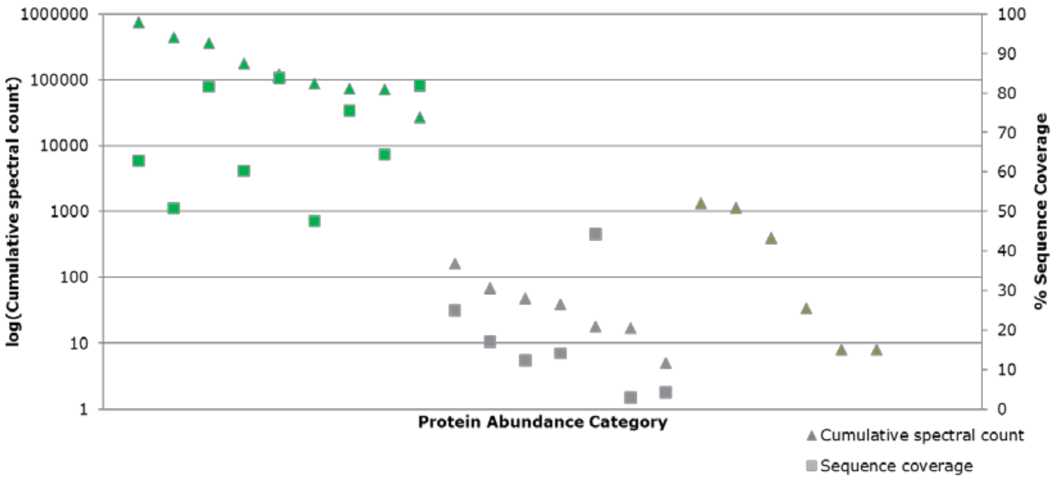


Figure 32. Database specific and shared peptides from the total number of 115, 601 consensus peptides. Among these peptides, 36, 565 were found in all databases; 61,936 peptides were genome specific; 211 peptides were CDS specific; 16,628 of them were bacteria/virus specific; 191 of them were GENSCAN prediction specific and 70 of them were altORF specific. For protein database, there were no protein-database specific peptides found indicating that these peptides were shared in the genome or the CDS as expected.

The blood plasma, serum, and platelet proteome is composed of many proteins with top 22 proteins accounting for 90% of total protein content (Tu et al. 2010; Qian et al. 2008)This limits detection of medium and low abundance proteins which are present in 10 orders of magnitude in terms of protein concentration (Millioni et al. 2011). Human blood is known as a rich source of biomarker proteins which are found in low abundance. Many depletion protocols have been developed for the removal of the abundant proteins such as albumins, complement components, apolipoproteins etc. In order to analyze the found proteins, proteins having at least two peptides were listed with sequence coverage by merging overlapping peptides and spectral count. The protein atlas generated in this study (Figure 33) showed that most abundant blood proteins were not efficiently removed from the samples. In this study, it was observed that albumin, complement component protein types, serpin, fibronectin, keratin, plasminogen were some of most abundantly found proteins according to a high number of total spectral count. On the other hand, low abundant proteins such as RAS oncogene family, bone marrow stromal cell antigen responsible from rheumatoid arthritis, premature ovarian failure 1B protein playing role in early age menopause due to autoimmune disease POF , S100 calcium binding protein A14 responsible for delirium, synaptonemal complex protein playing role in cervical cancer and ret proto-oncogenes were observed with low spectral coverage but high-mid sequence coverage.



High

- Albumin
- Complement component types
- Serpin
- Fibronectin
- Keratin
- Apolipoprotein

Low

- RAS oncogene family
- Bone marrow stromal cell antigen
- Premature ovarian failure, 1B
- S100 calcium binding protein A14
- Synaptonemal complex protein 1
- Ret proto-oncogene

Microbiome

- HIV
- Lachnospiraceae bacterium
- Streptococcus sp.
- Hepatitis B
- Human papillomavirus
- Epstein-Barr virus

Figure 33. Protein atlas of this study is shown in three protein abundance categories as high, low and virus/bacteria proteins.

Apart from human proteins, microbial proteins including viral and bacterial proteins have been detected in the samples. Infectious diseases via virus and bacteria and normal bacterial flora of human body indicate the importance of understanding the host-pathogen interaction and identification and detection of viral, bacterial pathogens in human blood and plasma. The detection of these pathogens has importance for blood supply. In literature, there are protocols based on the microarray, polymerase chain reaction techniques on the genomic and transcriptomic level (Duncan et al. 2015; Kourout et al. 2016). On the other hand, genomic and transcriptomic level detection does not provide insights on the protein level. Therefore, mass spectrometry based proteomics data would enable detection of pathogen proteins in blood with fast detection even real-time detection in the clinic during surgical operations. In this study, HIV, Hepatitis B, human papillomavirus, cancer-causing Epstein-Barr virus, lung infection causing streptococcus sp. microbes were detected via pulling down organism-specific peptides. In total, 545 viruses and bacteria were identified with distinct peptides. On the other hand, it should be noted that these results do not yield 100% accuracy since there was no experimental validation test had been conducted.

Another aspect of this study is to find novel genes or to refine existing gene models based on ab initio gene prediction. GENSCAN algorithm was performed to predict genes based on genomic features which address features of typical protein-coding genes in human genome. In total, 3468 GENSCAN predictions were mapped to peptides. Of the 10 predictions having at least one GENSCAN unique peptides and total greater and equal than two peptides were detected. Six of those cannot be resolved since they have same peptide set. However, four of predictions were unambiguously identified since they contain proteotypic peptides. In Figure 34, GENSCAN6003, immunoglobulin kappa V prediction, was shown. The gene has four mapped peptides with one as proteotypic which is the indicator of unambiguous identification. Three exons were verified by these peptides. However, two exons did not have any mapped peptides. In addition to the existence of proteotypic peptide, another line of evidence was obtained by performing BLAST search against *Mus musculus* protein database. The peptide matching exons were aligned to *M. musculus* homolog gene. HAVANA gene model was not completely available for this gene model.

In order to increase the accuracy of gene prediction tools, predictions verified via peptides need to be used as gold models.

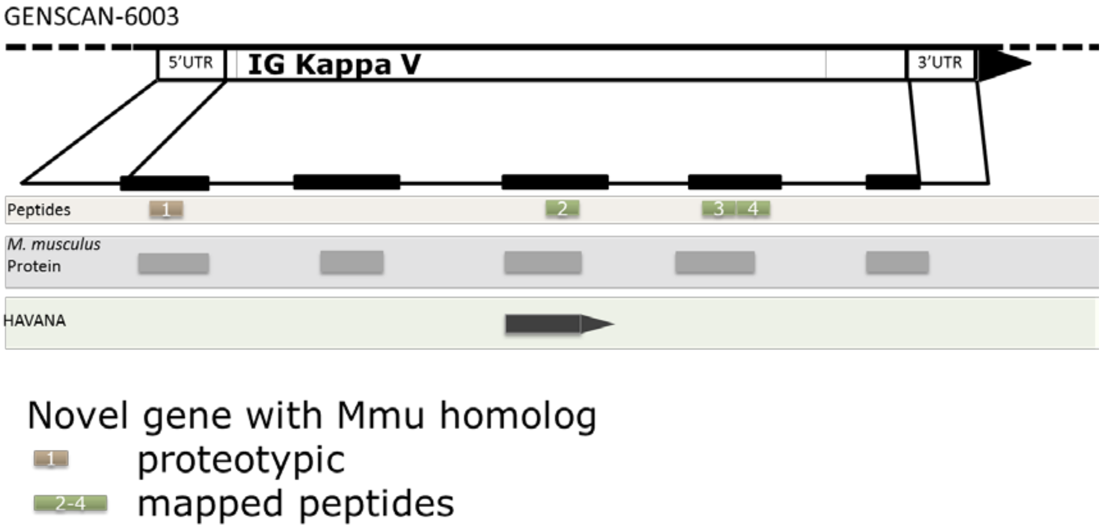


Figure 34. GENSCAN6003 prediction for IG Kappa V with proteotypic peptides and other mapped peptides.

The early findings were suggesting one gene-one transcript-one protein model which is known as central dogma model. However, it was shown that most eukaryotic mRNAs have potential to encode several proteins due to the existence of several translation start sites (Ingolia 2014). These translation initiation sites lead to the expression of alternative open reading frames (altORFs). Ribo-Seq experiments revealed that many eukaryotic organisms including yeast (Ingolia et al. 2009), plants

(Liu et al. 2013), mammalians (Ingolia et al. 2011) have alternative translation initiation sites activated by different translation mechanisms. Previously, novel proteins have been detected via proteomics experiments. Mass spectrometry-based proteomics enables detection of altORF in a more accurate way since altORF products are expected to be shorter than known primary products. Previously, in different organisms, altORFs have been determined via proteogenomic methods. In human blood tissue, it was also shown that in total 1018 altORF proteins in serum and plasma had been detected (Vanderperre et al. 2013). The goal of this study is to detect altORFs in an exhaustive way by involving altORF predictions in the initial search space. These predictions were computed using altORFEv module of PGMiner by considering three major mechanisms; linear scanning mechanism (Kochetov 2008), leaky scanning model (Bazykin and Kochetov 2011; Van Damme et al. 2014), reinitiation model (Kozak 2001). In this study, 16147 altORFs were supported by peptides. In order to increase the accuracy of predictions, altORF distinct peptides, i.e. proteotypic peptides, were considered. Considering that 23 predictions with at least one proteotypic peptide and in total, more than one peptide was reported. However, 21 of those proteins were immunoglobin variants. Therefore it was not possible to distinguish them. Of the four predictions were distinctly confirmed. These protein isoforms are IG_V gene (ENST00000626108.1) via leaky scanning mechanism, myosin heavy chain 7 (ENST00000355349.3) via leaky scanning mechanism, feutin B (ENST00000420570.1) via leaky scanning and prostaglandin D2 synthase (ENST00000471521.5) via leaky scanning products. In Figure 35, an example for IG-V(ENST00000626108.1) was given indicating the peptides in the known protein product and the isoform. It should be noted that peptide MTQSPSSLSAS overlaps with second translation initiation site and the suffix amino acid of this peptide is not R or K. This indicates that this peptide cannot be produced in a fully enzymatic cleavage process. The exons of this gene are confirmed, however, due to the number of spectra analyzed, a high number of peptides could not be found and high sequence coverage could not be reached.

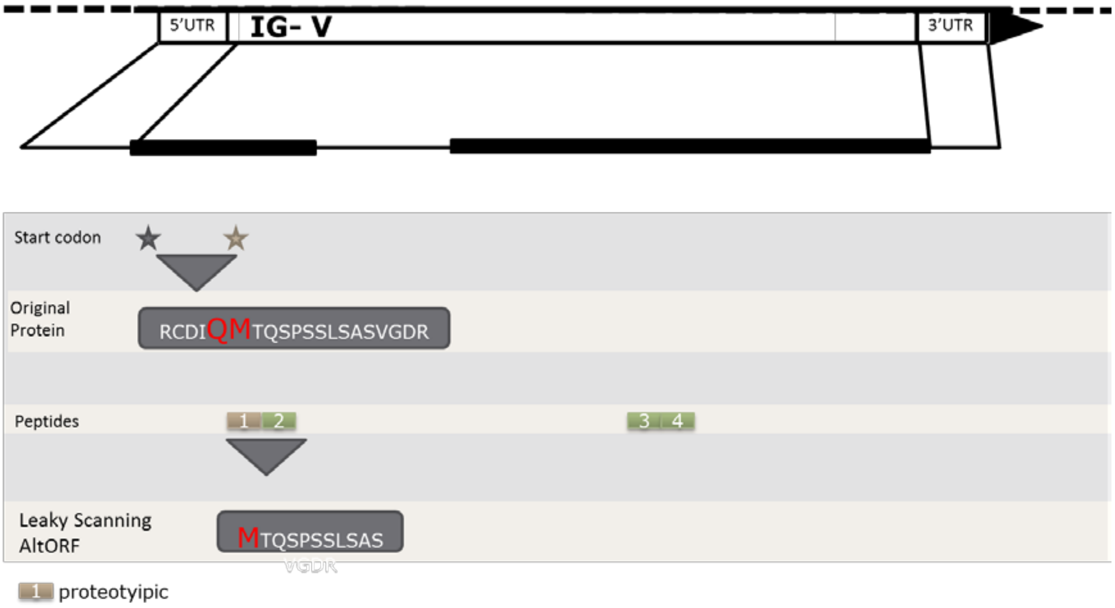


Figure 35. Alternative translation product prediction from IG-V (ENST00000626108.1) via leaky scanning mechanism.

In the next step, HAVANA and ENSEMBL gene models were analyzed individually. For this analysis, matching models, i.e. transcripts, were classified into five categories according to the type of mapping peptides; exonic, intronic, distinct, 3' overlapping, 5' overlapping (Figure 36):

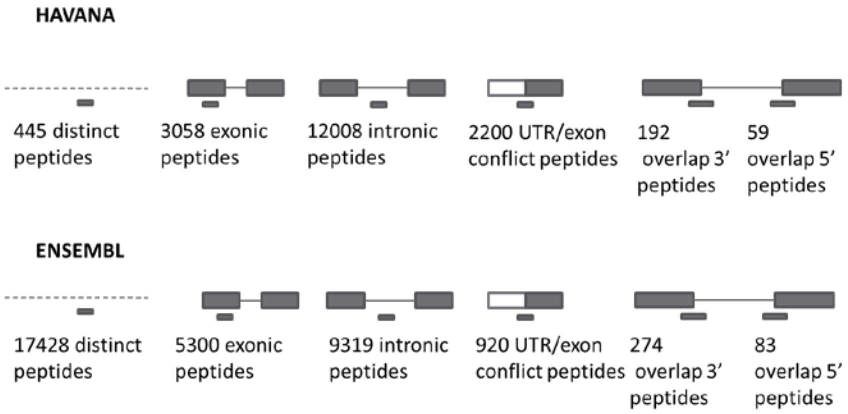


Figure 36. Mapped peptides were categorized as exonic, intronic, UTR/exon conflicting, overlapping and distinct. Here, the number of peptides for each category was given for ENSEMBL and HAVANA models.

The first type consists of confirmed models meaning that models are only having exonic peptides. Of the 230 HAVANA models were confirmed via only exonic peptides, while 267 ENSEMBL models were confirmed with exonic peptides. Of the 2021 HAVANA models also have intronic peptides. In contract, 293 ENSEMBL models were found with intronic peptides. Here these peptides were not further analyzed according to splice site existence around these peptides. However, further analysis of these models individually would lead to correction of models as suggesting new exon regions instead of intronic regions or splice site products that were not detected in this study. ENSEMBL automated models might be considered more accurate in terms of exon-intron prediction than HAVANA manual annotations. On the other hand, as stated earlier, these models were not validated via experiments or inspected further by considering splice models. Of the 2459 HAVANA genes have conflicting peptides meaning that a peptide P is considered as exonic for T1 of gene G, but intronic for T2of the same gene. The number of conflicting gene models increased in ENSEMBL as 4887. In HAVANA models, there is only one gene model was found as having only 3’ exon overlapping peptide, however, in ENSEMBL, this number was found as three. Only 3’ exon overlapping peptide having model indicates exon boundary correction. Peptides matching to intergenic regions also take great interest to determine novel genes. In ENSEMBL, 1454 gene models include distinct peptides as outside of the open reading frames. In HAVANA, 313 gene models include distinct peptides therefore named as distinct models.

In this study, a proteogenomics study on human blood proteome was presented. In total 3 millions of spectra were searched against human genome, known human proteins, coding sites, alternative translation products, gene predictions by GENSCAN algorithm and microbial proteins. It was demonstrated that these MS/MS spectra led to the identification of new gene models, validation, and refinement of existing gene models, protein isoforms, human contaminant proteins.

CHAPTER 6

CONCLUSION

It has been more than a decade that genome and transcriptome sequencing of model organisms such as human, mouse, yeast have yielded a vast amount of data with significant effort. Despite this fact that accurate genome annotation as the final and main aim of sequencing still remains challenging. The reason behind performing genome annotation is to determine protein coding genes and regulatory regions. Current genome annotation methods mainly rely on computational annotations supported by sequence homology and transcriptional evidence such as EST, CCDS regions. Nevertheless, the accuracy of genome annotation is correlated to experimental studies since it requires experimental validation. Sequence homology requires known sequence availability, while transcriptional evidence are not always verified on the protein level. Therefore, accurate experimental data should be collected from proteome data. Mass spectrometry-based proteomics method is state of the art for sequencing and quantification of peptides derived from protein samples. The idea behind the usage of proteomics is to validate translation of genes to confirm genomic regions such as exons. This defines the field proteogenomics, and proteogenomics enables not only validation of translation, but also a refinement of proposed gene annotations and identification of novel genes as well as new protein isoforms as products of alternative translation. There have been many efforts to perform high-throughput analysis of proteomics and genomics data in a sequenced manner to automate data processing, identification, genome mapping and assessment. This thesis is aimed to address some issues of the field proteogenomics.

In Chapter 2, database size and database search algorithm parameters as factors affecting performance, the accuracy of peptide identification process via database search were analyzed. The reliability and sensitivity of following steps in bottom-up and top-down proteomics are dependent on this. In addition to conventional proteomics analysis, this also underpins proteogenomic analysis. Since proteomic data is applied for assessment of genome annotation, low sensitivity and specificity propagate to incorrect annotations.

In Chapter 3, a fast and accurate peptide mapping tool was introduced. Although database search algorithms identify peptides, before annotation of genomic data via proteomics data, a challenge arises through the finding of all occurrences of peptides on the genomic level. Besides, that location of peptides needs to be represented in a format which is recognized by genome browsers. In many proteogenomics studies, in-house scripts are in used, and they are not available publicly. Few tools have been announced to overcome these problems. However, these tools were not assessed with further tests to check whether they fulfill expectations of mapped peptides. Therefore, a new tool based on modified version of the Wu-Manber algorithm was developed, and it was assessed by expected peptide mapping scenarios. These scenarios were established to evaluate the accuracy of two available and the new tool and to determine processing capacity of these tools in terms database size and number of queries.

In Chapter 4, a new proteogenomic annotation pipeline, PGMiner was developed to close the gap between proteomics data analysis and genome annotation assessment. Most of the proteogenomic tools have been presented in GUI version or as a platform plugin like Galaxy. However, these tools come along with installation or modularity problems. The available tools require external input data loading, one or multiple database search tool support and heuristic peptide mapping implementations. In addition to that tools have restrictions on usage of big size databases such as complete human genome translation. PGMiner, however, tackles external data upload, database size restriction issues and provides multiple database search tool usage support with further FDR and consensus prediction implementations. In addition to that exact peptide mapping algorithm with enzymatic cleavage, rule filtering was provided. The tool was developed on KNIME Analytics Platform to provide user-friendly, extendable with custom or readily available data analysis and scripting nodes.

In Chapter 5, PGMiner was applied to human blood plasma proteome MS dataset to demonstrate the usage of the tool. Here, it was shown that MS-based proteomics data facilitated the refinement and validation of existing genome annotation. Besides confirmation of existing models provided from HAVANA manual annotations and ENSEMBL automated pipeline, novel gene models were introduced and alternative translation products were determined. The confidence here was ensured via signature peptides which are known as proteotypic peptides. These peptides are specific to only one location through the genome or spliced form of a transcript. In addition to that homology information was used a complementary line of evidence. Human blood is a source of biomarkers as well as a source for detecting pathogen-originated peptides/proteins. In this study, microbes which are responsible for certain infections such as HIV, hepatitis have been detected. These findings, on the other hand, need to be proven by further experimental studies or checking the donor information in particular samples. As a conclusion, this study demonstrated the importance and the value of exploiting proteomics data for genome annotations as a line of experimental evidence. It would be important to improve the current computational genome annotation pipelines with proteomics data and to bring the knowledge gained from features of proteomics data to newly developed computational annotation tools. Another outcome of this study addressed the issues related to wet lab part of MS experiments. Due to the inefficient removal of highly abundant blood proteins, low abundance proteins in the dynamic range were masked in a high level. In the future, more sophisticated experimental procedures will lead to the production of higher quality MS data to increase the sensitivity of proteogenomics outcomes.

REFERENCES

Aebersold, Ruedi, and Matthias Mann. 2003. “Mass Spectrometry-Based Proteomics.” Nature 422 (6928) (March): 198–207.

Aho Alfred V, and Corasick Margaret J. 1975. “Efficient String Matching: An Aid to Bibliographic Search.” Commun ACM 18: 333–340.

Aitken, Colin Echeverría, and Jon R Lorsch. 2012. “A Mechanistic Overview of Translation Initiation in Eukaryotes.” Nature Structural Molecular Biology 19 (6): 568–576.

Aken, Bronwen L, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, et al. 2017. “Ensembl 2017.” Nucleic Acids Research 45 (D1) (January): D635–D642.

Aken, Bronwen L, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, et al. 2016. “The Ensembl Gene Annotation System.” Database : The Journal of Biological Databases and Curation 2016: 1–19.

Alekhina, Olga M, and Konstantin S Vassilenko. 2012. “Translation Initiation in Eukaryotes: Versatility of the Scanning Model.” Biochemistry Biokhimii͡a 77 (13): 1465–77.

Alfaro, Javier A, Ankit Sinha, Thomas Kislinger, and Paul C Boutros. 2014. “Onco-Proteogenomics: Cancer Proteomics Joins Forces with Genomics.” Nature Methods 11 (11) (November): 1107–13.

Allmer, Jens. 2011. “Algorithms for the de Novo Sequencing of Peptides from Tandem Mass Spectra.” Expert Review of Proteomics 8 (5): 645–657.

Allmer, Jens. 2012. “A Call for Benchmark Data in Mass Spectrometry-Based Proteomics.” Journal of Integrated OMICS 2 (2): 1–5.

Allmer, Jens. 2016. “Exact Pattern Matching: Adapting the Boyer-Moore Algorithm for DNA Searches.” PeerJ Preprints 4:e1758v1.

Allmer, Jens, Christine Markert, Einar J Stauber, and Michael Hippler. 2004. “A New Approach That Allows Identification of Intron-Split Peptides from Mass Spectrometric Data in Genomic Databases.” FEBS Letters 562 (1-3): 202–206.

Allmer, Jens, Bianca Naumann, Christine Markert, Monica Zhang, and Michael Hippler. 2006. “Mass Spectrometric Genomic Data Mining: Novel Insights into Bioenergetic Pathways in *Chlamydomonas reinhardtii*.” Proteomics 6 (23): 6207–6220.

Altschul, Stephen F, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. “Basic Local Alignment Search Tool.” Journal of Molecular Biology 215 (3): 403–10.

Anderson, Matthew W, and Iris Schrijver. 2010. “Next Generation DNA Sequencing and the Future of Genomic Medicine.” Genes 1 (1): 38–69.

Ansong, Charles, Samuel O Purvine, Joshua N Adkins, Mary S Lipton, and Richard D Smith. 2008. “Proteogenomics: Needs and Roles to Be Filled by Proteomics in Genome Annotation.” Briefings in Functional Genomics & Proteomics 7 (1) (January): 50–62.

Askenazi, Manor, Kelly V Ruggles, and David Fenyö. 2015. “PGx: Putting Peptides to BED.” Journal of Proteome Research 15 (3): 795–799.

Baerenfaller, Katja. 2008. “Genome-Scale Proteomics Reveals *Arabidopsis thaliana* Gene Models.” Science 320 (5878): 938-41.

Bazykin, Georgii A, and Alex V Kochetov. 2011. “Alternative Translation Start Sites Are Conserved in Eukaryotic Genomes.” Nucleic Acids Research 39 (2): 567–577.

Bitton, Danny A, Duncan L Smith, Yvonne Connolly, Paul J Scutt, and Crispin J Miller. 2010. “An Integrated Mass-Spectrometry Pipeline Identifies Novel Protein Coding-Regions in the Human Genome.” PloS One 5 (1) (January): e8949.

Branca, Rui M M, Lukas M Orre, Henrik J Johansson, Viktor Granholm, Mikael Huss, Åsa Pérez-Bercoff, Jenny Forshed, Lukas Käll, and Janne Lehtiö. 2014. “HiRIEF LC-MS Enables Deep Proteome Coverage and Unbiased Proteogenomics.” Nature Methods 11 (1) (January): 59–62.

Brosch, Markus, Gary I Saunders, Adam Frankish, Mark O Collins, Lu Yu, James Wright, Ruth Verstraten, et al. 2011. “Shotgun Proteomics Aids Discovery of Novel Protein-Coding Genes , Alternative Splicing , and ‘“ Resurrected ”’ Pseudogenes in the Mouse Genome.” Genome Research: 756–767.

Burge, Chris, and Samuel Karlin. 1997. “Prediction of Complete Gene Structures in Human Genomic DNA.” Journal of Molecular Biology 268 (1) (April): 78–94.

Castellana, Natalie, and Vineet Bafna. 2010. “Proteogenomics to Discover the Full Coding Content of Genomes: A Computational Perspective.” Journal of Proteomics 73 (11): 2124–2135.

Castellana, Natalie E, Samuel H Payne, Zhouxin Shen, Mario Stanke, Vineet Bafna, and Steven P Briggs. 2008. “Discovery and Revision of Arabidopsis Genes by Proteogenomics.” Proceedings of the National Academy of Sciences of the United States of America 105 (52): 21034–21038.

Castellana, Natalie E, Victoria Pham, David Arnott, Jennie R Lill, and Vineet Bafna. 2010. “Template Proteogenomics: Sequencing Whole Proteins Using an Imperfect Database.” Molecular Cellular Proteomics MCP 9 (6): 1260–1270.

Choi, Hyungwon, and Alexey I Nesvizhskii. 2008. “False Discovery Rates and Related Statistical Concepts in Mass Spectrometry-Based Proteomics.” Journal of Proteome Research 7 (1): 47–50.

Choudhary, Jyoti S, Walter P Blackstock, David M Creasy, and John S Cottrell. 2001. “Interrogating the Human Genome Using Uninterpreted Mass Spectrometry Data.” Proteomics 1 (5): 651–667.

Claassen, Manfred. 2012. “Inference and Validation of Protein Identifications.” Molecular & Cellular Proteomics : MCP 11 (11) (November): 1097–104.

Collins, John E, Melanie E Goward, Charlotte G Cole, Luc J Smink, Elizabeth J Huckle, Sarah Knowles, Jacqueline M Bye, David M Beare, and Ian Dunham. 2003. “Reevaluating Human Gene Annotation : A Second-Generation Analysis of Chromosome 22.” Genome Research: 27–36.

Craig, Robertson, and Ronald C Beavis. 2004. “TANDEM: Matching Proteins with Tandem Mass Spectra.” Bioinformatics 20 (9): 1466–1467.

Dagda, Ruben K, Tamanna Sultana, and James Lyons-Weiler. 2010. “Evaluation of the Consensus of Four Peptide Identification Algorithms for Tandem Mass Spectrometry Based Proteomics.” Journal of Proteomics & Bioinformatics 3 (February): 39–47.

Dandekar, Thomas, Martijn Huynen, Jörg Thomas Regula, Barbara Ueberle, Carl Ulrich Zimmermann, Miguel A Andrade, Tobias Doerks, et al. 2000. “Re-Annotating the *Mycoplasma pneumoniae* Genome Sequence: Adding Value, Function and Reading Frames.” Nucleic Acids Research 28 (17): 3278–3288.

de Souza, Gustavo A, Magnus Ø Arntzen, Suereta Fortuin, Anita C Schürch, Hiwa Målen, Christopher R E McEvoy, Dick van Soolingen, Bernd Thiede, Robin M Warren, and Harald G Wiker. 2011. “Proteogenomic Analysis of Polymorphisms and Gene Annotation Divergences in Prokaryotes Using a Clustered Mass Spectrometry-Friendly Database.” Molecular & Cellular Proteomics : MCP 10 (1): M110.002527.

Derrick, Peter J, and Scott D Patterson. 2001. “Mass Spectrometry and Proteomics.” Proteomics 1 (8): 925–926.

Desiere, Frank, Eric W Deutsch, Alexey I Nesvizhskii, Parag Mallick, Nichole L King, Jimmy K Eng, Alan Aderem, et al. 2005. “Integration with the Human Genome of Peptide Sequences Obtained by High-Throughput Mass Spectrometry.” Genome Biology 6 (1): R9.

Deutsch, Eric W. 2010a. “The PeptideAtlas Project.” Methods in Molecular Biology (Clifton, N.J.) 604 (January): 285–96.

Deutsch, Eric W. 2010b. “Mass Spectrometer Output File Format mzML.” Methods in Molecular Biology (Clifton, N.J.) 604: 319–331.

Domon, Bruno, and Ruedi Aebersold. 2006. “Mass Spectrometry and Protein Analysis.” Science (New York, N.Y.) 312 (5771) (April): 212–7.

Dorfer, Viktoria, Peter Pichler, Thomas Stranzl, Johannes Stadlmann, Thomas Taus, Stephan Winkler, and Karl Mechtler. 2014. “MSAmanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra.” Journal of Proteome Research 13 (8): 3679–3684.

Duncan, Robert, Moussa Kourout, Elena Grigorenko, Carolyn Fisher, Robert Duncan, and Moussa Kourout. 2015. “Blood-Borne Pathogens : Promises and Pitfalls Advances in Multiplex Nucleic Acid Diagnostics for Blood-Borne Pathogens : Promises and Pitfalls.” Expert Review of Molecular Diagnostics 7159: 1–13.

Edwards, Nathan J. 2007. “Novel Peptide Identification from Tandem Mass Spectra Using ESTs and Sequence Database Compression.” Molecular Systems Biology 3 (102) (January): 102.

Elias, Joshua E, and Steven P Gygi. 2007. “Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry.” Nature Methods 4 (3) (March): 207–14.

Eng, Jimmy K, Ashley L Mccormack, and John R Yates. 1994. “An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a

Protein Database." Journal of the American Society for Mass Spectrometry 5 (11): 976–989.

Eng, Jimmy K, Brian C Searle, Karl R Clauser, and David L Tabb. 2011. "A Face in the Crowd: Recognizing Peptides through Database Search." Molecular & Cellular Proteomics : MCP 10 (11) (November): R111.009522.

Fenyö, David, and Gerben Menschaert. 2015. "Proteogenomics From A Bioinformatics Angle: A Growing Field." Mass Spectrometry Reviews 9999 (December): 1–16.

Fermin, Damian, Baxter B Allen, Thomas W Blackwell, Rajasree Menon, Marcin Adamski, Yin Xu, Peter Ulintz, Gilbert S Omenn, and David J States. 2006. "Novel Gene and Gene Model Detection Using a Whole Genome Open Reading Frame Analysis in Proteomics." Genome Biology 7 (4): R35.

Flicek, Paul, Ikhlak Ahmed, Ridwan M Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, et al. 2012. "Ensembl 2013." Nucleic Acids Research 41 (Database issue) (November):48-55.

Frank, Ari, and Pavel Pevzner. 2005. "PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling." Analytical Chemistry 77 (4): 964–973.

Frishman, Dmitrij. 2007. "Protein Annotation at Genomic Scale: The Current Status." Chemical Reviews 107 (8) (August): 3448–66.

Gallien, Sébastien, Emmanuel Perrodou, Christine Carapito, Caroline Deshayes, Jean-Marc Reyrat, Alain Van Dorsselaer, Olivier Poch, Christine Schaeffer, and Odile Lecompte. 2009. "Ortho-Proteogenomics: Multiple Proteomes Investigation through Orthology and a New MS-Based Protocol." Genome Research 19 (1): 128–135.

Geer, Lewis Y, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. 2004. "Open Mass Spectrometry Search Algorithm." Journal of Proteome Research 3 (5): 958–64.

Guigó, Roderic, Paul Flicek, Josep F Abril, Alexandre Reymond, Julien Lagarde, France Denoeud, Stylianos Antonarakis, et al. 2006. “EGASP: The Human ENCODE Genome Annotation Assessment Project.” Genome Biology 7 Suppl 1 (January): S2.1–31.

Gupta, Nitin, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. 2011. “Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong.” Journal of the American Society for Mass Spectrometry 22 (7) (July): 1111–20.

Gupta, Nitin, Jamal Benhamida, Vipul Bhargava, Daniel Goodman, Elisabeth Kain, Ian Kerman, Ngan Nguyen, et al. 2008. “Comparative Proteogenomics: Combining Mass Spectrometry and Comparative Genomics to Analyze Multiple Genomes.” Genome Research 18 (7): 1133–1142.

Gupta, Nitin, and Pavel A Pevzner. 2009. “False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule.” Journal of Proteome Research 8 (9) (September): 4173–81.

Harrow, Jennifer, Alinda Nagy, Alexandre Reymond, Tyler Alioto, Laszlo Patthy, Stylianos E Antonarakis, and Roderic Guigó. 2009. “Identifying Protein-Coding Genes in Genomic Sequences.” Genome Biology 10 (1) (January): 201.

Has, Canan, and Jens Allmer. 2016. “PGMiner: Complete Proteogenomics Workflow; from Data Acquisition to Result Visualization.” Information Sciences 384:126-134.

Has, Canan, Şule Yılmaz, and Jens Allmer. 2012. “COMAS: Ant Colony Optimization a De Novo Sequencing Algorithm.” Paper presented at the 11th European Conference on Computational Biology, Basel, Switzerland September 8-12.

Helmy, Mohamed, Naoyuki Sugiyama, Masaru Tomita, and Yasushi Ishihama. 2010. “Onco-Proteogenomics: A Novel Approach to Identify Cancer-Specific Mutations Combining Proteomics and Transcriptome Deep Sequencing.” Genome Biology 11 (Suppl 1): P17.

Helmy, Mohamed, and Masaru Tomita. 2012. “Peptide Identification by Searching Large-Scale Tandem Mass Spectra against Large Databases: Bioinformatics Methods in Proteogenomics.” Genes Genomes and Genomics 6 (1): 76–85.

Helmy, Mohamed, Masaru Tomita, and Yasushi Ishihama. 2011. “OryzaPG-DB: Rice Proteome Database Based on Shotgun Proteogenomics.” BMC Plant Biology 11 (1): 63.

Himmelreich, Ralf, Helga Plagens, Helmut Hilbert, Berta Reiner, and Richard Herrmann. 1997. “Comparative Analysis of the Genomes of the Bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*.” Nucleic Acids Research 25 (4): 701–712.

Hoopmann, Michael R, and Robert L Moritz. 2013. “Current Algorithmic Solutions for Peptide-Based Proteomics Data Generation and Identification.” Current Opinion in Biotechnology 24 (1) (February): 31–8.

Imanishi, Tadashi, Takeshi Itoh, Yutaka Suzuki, Claire O’Donovan, Satoshi Fukuchi, Kanako O Koyanagi, Roberto A Barrero, et al. 2004. “Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones.” PLoS Biology 2 (6) (April): e162.

Ingolia, Nicholas T, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. 2009. “Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling.” Science (New York, N.Y.) 324 (5924) (April): 218–23.

Ingolia, Nicholas T. 2014. “Ribosome Profiling: New Views of Translation, from Single Codons to Genome Scale.” Nature Reviews Genetics 15 (3) (January): 205–213.

Ingolia, Nicholas T, Liana F Lareau, and Jonathan S Weissman. 2011. “Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes.” Cell 147 (4): 789–802.

Jackson, Richard J, Christopher U T Hellen, and Tatyana V Pestova. 2010. “The Mechanism of Eukaryotic Translation Initiation and Principles of Its Regulation.” Nature Reviews Molecular Cell Biology 11 (2): 113–127.

Jaffe, Jacob D, Howard C Berg, and George M Church. 2004. “Proteogenomic Mapping as a Complementary Method to Perform Genome Annotation.” Proteomics 4 (1): 59–77.

Jagtap, Pratik D, James E Johnson, Getiria Onsongo, Fredrik W Sadler, Kevin Murray, Yuanbo Wang, Gloria M Shenykman, Sricharan Bandhakavi, Lloyd M Smith, and Timothy J Griffin. 2014. “Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework.” Journal of Proteome Research 13 (12) (December): 5898–908.

Jagtap, Pratik, Jill Goslinga, Joel A Kooren, Thomas McGowan, Matthew S Wroblewski, Sean L Seymour, and Timothy J Griffin. 2013. “A Two-Step Database Search Method Improves Sensitivity in Peptide Sequence Matches for Metaproteomics and Proteogenomics Studies.” Proteomics 13 (8) (April): 1352–7.

Käll, Lukas, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. 2007. “Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets.” Nature Methods 4 (11) (November): 923–5.

Käll, Lukas, John D Storey, and William Stafford Noble. 2009. “QVALITY: Non-Parametric Estimation of Q-Values and Posterior Error Probabilities.” Bioinformatics (Oxford, England) 25 (7) (April): 964–6.

Käll, Lukas, John D Storey, Michael J MacCoss, and William Stafford Noble. 2008. “Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin.” Journal of Proteome Research 7(1) (January):40-4.

Käll, Lukas, and Olga Vitek. 2011. “Computational Mass Spectrometry–Based Proteomics.” Edited by Fran Lewitter. PLoS Computational Biology 7 (12) (December): e1002277.

Kapp, Eugene A, Frédéric Schütz, Lisa M Connolly, John A Chakel, Jose E Meza, Christine A Miller, David Fenyö, et al. 2005. “An Evaluation, Comparison, and Accurate Benchmarking of Several Publicly Available MS/MS Search Algorithms: Sensitivity and Specificity Analysis.” Proteomics 5 (13): 3475–3490.

Keller, Andrew, Alexey I Nesvizhskii, Eugene Kolker, and Rudi Aebersold. 2002. “Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search.” Anal Chem 74 (20) (October): 5383–92.

Keller, Andrew, Samuel Purvine, Alexey I Nesvizhskii, Sergey Stolyar, David R Goodlett, and Eugene Kolker. 2002. “Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis.” Omics 6 (2): 207–12.

Kent, James W. 2002. “BLAT--the BLAST-like Alignment Tool.” Genome Research 12 (4) (April): 656–64.

Kersey, Paul J, Daniel Lawson, Ewan Birney, Paul S Derwent, et al. 2010. “Ensembl Genomes: Extending Ensembl across the Taxonomic Space.” Nucleic Acids Research 38 (Database issue) (January): D563–9.

Kessner, Darren, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. 2008. “ProteoWizard: Open Source Software for Rapid Proteomics Tools Development.” Bioinformatics 24 (21): 2534–2536.

Khatun, Jainab, Yanbao Yu, John A Wrobel, Brian A Risk, Harsha P Gunawardena, Ashley Secrest, Wendy J Spitzer, et al. 2013. “Whole Human Genome Proteogenomic Mapping for ENCODE Cell Line Data: Identifying Protein-Coding Regions.” BMC Genomics 14 (1) (January): 141.

Kim, Sangtae, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert J R Heck, and Pavel A Pevzner. 2010. “The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to

Database Search.” Molecular & Cellular Proteomics : MCP 9 (12) (December): 2840–52.

Knowles, David G, and Aoife McLysaght. 2009. “Recent de Novo Origin of Human Protein-Coding Genes.” Genome Research 19 (10) (October): 1752–1759.

Kochetov, Alex V. 2008. “Alternative Translation Start Sites and Hidden Coding Potential of Eukaryotic mRNAs.” BioEssays News and Reviews in Molecular Cellular and Developmental Biology 30 (7): 683–691.

Kourout, Moussa, Carolyn Fisher, Anjan Purkayastha, Clark Tibbetts, Valerie Winkelman, Phillip Williamson, Hira L Nakhasi, and Robert Duncan. 2016. “Multiplex Detection and Identification of Viral, Bacterial, and Protozoan Pathogens in Human Blood and Plasma Using a High-Density Resequencing Pathogen Microarray Platform.” DONOR INFECTIOUS DISEASE TESTING (June): 1537–1547.

Kozak, Marilyn. 2001. “Constraints on Reinitiation of Translation in Mammals.” Nucleic Acids Research 29 (24) (December): 5226–32.

Kuhring, Mathias, and Bernhard Y Renard. 2012. “iPiG: Integrating Peptide Spectrum Matches into Genome Browser Visualizations.” PloS One 7 (12) (January): e50246.

Kumar, Dhirendra, Amit Kumar Yadav, Puneet Kumar Kadimi, Shivashankar H Nagaraj, Sean M Grimmond, and Debasis Dash. 2013. “Proteogenomic Analysis of *Bradyrhizobium japonicum* USDA110 Using GenoSuite, an Automated Multi-Algorithmic Pipeline.” Molecular & Cellular Proteomics : MCP 12 (11): 3388–97.

Liu, Ming-Jung, Szu-Hsien Wu, Jing-Fen Wu, Wen-Dar Lin, Yi-Chen Wu, Tsung-Ying Tsai, Huang-Lung Tsai, and Shu-Hsing Wu. 2013. “Translational Landscape of Photomorphogenic Arabidopsis.” The Plant Cell 25 (10) (October): 3699–3710.

Lögdberg, Lennart, Marion E Reid, and Teresa Zelinski. 2011. “Human Blood Group Genes 2010: Chromosomal Locations and Cloning Strategies Revisited.” Transfusion Medicine Reviews 25 (1): 36–46.

Malys, Naglis, and John E G McCarthy. 2011. “Translation Initiation: Variations in the Mechanism Can Be Anticipated.” Cellular and Molecular Life Sciences CMLS 68 (6): 991–1003.

Marx, Harald, Simone Lemeer, Jan Erik Schliep, Lucrece Matheron, Shabaz Mohammed, Jürgen Cox, Matthias Mann, Albert J R Heck, and Bernhard Kuster. 2013. “A Large Synthetic Peptide and Phosphopeptide Reference Library for Mass Spectrometry-Based Proteomics.” Nature Biotechnology 31 (6) (June): 557–64.

McHugh, Leo, and Jonathan W Arthur. 2008. “Computational Methods for Protein Identification from Mass Spectrometry Data.” PLoS Computational Biology 4 (2) (February): e12.

Merchant, Sabeeha S, Simon E Prochnik, Olivier Vallon, Elizabeth H Harris, Steven J Karpowicz, George B Witman, Astrid Terry, et al. 2007. “The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions.” Science (New York, N.Y.) 318 (5848) (October): 245–50.

Merrihew, Gennifer E, Colleen Davis, Brent Ewing, Gary Williams, Lukas Käll, Barbara E Frewen, William Stafford Noble, Phil Green, James H Thomas, and Michael J MacCoss. 2008. “Use of Shotgun Proteomics for the Identification, Confirmation, and Correction of *C. Elegans* Gene Annotations.” Genome Research 18 (10): 1660–1669.

Millioni, Renato, Serena Tolin, Lucia Puricelli, Stefano Sbrignadello, Gian Paolo Fadini, Paolo Tessari, and Giorgio Arrigoni. 2011. “High Abundance Proteins Depletion vs Low Abundance Proteins Enrichment: Comparison of Methods to Reduce the Plasma Proteome Complexity.” PLoS One 6 (5) (May): e19603.

Mo, Fan, Xu Hong, Feng Gao, Lin Du, Jun Wang, Gilbert S Omenn, and Biaoyang Lin. 2008. “A Compatible Exon-Exon Junction Database for the Identification of Exon Skipping Events Using Tandem Mass Spectrum Data.” BMC Bioinformatics 9 (1): 537.

Nagaraj, Shivashankar H, Robin B Gasser, and Shoba Ranganathan. 2006. "A Hitchhiker's Guide to Expressed Sequence Tag (EST) Analysis." Briefings in Bioinformatics 8 (1) (May): 6–21.

Nagaraj, Shivashankar H, Nicola Waddell, Anil K Madugundu, Scott Wood, Alun Jones, Ramya A Mandyam, Katia Nones, John V Pearson, and Sean M Grimmond. 2015. "PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization." Journal of Proteome Research 14 (5): 2255–2266.

Nahnsen, Sven, Andreas Bertsch, Jörg Rahnenführer, Alfred Nordheim, and Oliver Kohlbacher. 2011. "Probabilistic Consensus Scoring Improves Tandem Mass Spectrometry Peptide Identification." Journal of Proteome Research 10 (8) (August): 3332–43.

Napierala, Ma. 2012. "What Is the Bonferroni Correction ?" AAOS Now April: 1–3.

Nesvizhskii, Alexey I. 2014. "Proteogenomics: Concepts, Applications and Computational Strategies." Nature Methods 11 (11) (November): 1114–25.

Ning, Kang, and Alexey I Nesvizhskii. 2010. "The Utility of Mass Spectrometry-Based Proteomic Data for Validation of Novel Alternative Splice Forms Reconstructed from RNA-Seq Data: A Preliminary Assessment." BMC Bioinformatics 11 (Suppl 11): S14.

Parkinson, John, and Mark Blaxter. 2009. "Expressed Sequence Tags: An Overview." In *Expressed Sequence Tags (ESTs)*, edited by John Parkinson, 1–12. Methods in Molecular Biology Humana Press

Parra, G. 2003. "Comparative Gene Prediction in Human and Mouse." Genome Research 13 (1) (January): 108–117.

Pedrioli, Patrick G A, Jimmy K Eng, Robert Hubley, Mathijs Vogelzang, Eric W Deutsch, Brian Raught, Brian Pratt, et al. 2004. "A Common Open Representation of Mass Spectrometry Data and Its Application to Proteomics Research." Nature Biotechnology 22 (11) (November): 1459–66.

Perkins, David N, Darryl J Pappin, David M Creasy, and John S Cottrell. 1999. “Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data.” Electrophoresis 20 (18): 3551–67.

Qian, Wei-Jun, David T Kaleta, Brianne O Petritis, Hongliang Jiang, Tao Liu, Xu Zhang, Heather M Mottaz, et al. 2008. “Enhanced Detection of Low Abundance Human Plasma Proteins Using a Tandem IgY12-SuperMix Immunoaffinity Separation Strategy.” Molecular & Cellular Proteomics 7 (10) (April): 1963–1973.

Quandt, Andreas, Lucia Espona, Akos Balasko, Hendrik Weisser, Mi Youn Brusniak, Peter Kunszt, Ruedi Aebersold, and Lars Malmström. 2014. “Using Synthetic Peptides to Benchmark Peptide Identification Software and Search Parameters for MS/MS Data Analysis.” EuPA Open Proteomics 5: 21–31.

Reidegeld, Kai A, Martin Eisenacher, Michael Kohl, Daniel Chamrad, Gerhard Körting, Martin Blüggel, Helmut E Meyer, and Christian Stephan. 2008. “An Easy-to-Use Decoy Database Builder Software Tool, Implementing Different Decoy Strategies for False Discovery Rate Calculation in Automated MS/MS Protein Identifications.” Proteomics 8 (6) (March): 1129–37.

Renuse, Santosh, Raghothama Chaerkady, and Akhilesh Pandey. 2011. “Proteogenomics.” Proteomics 11 (4) (February): 620–30.

Risk, Brian A, Wendy J Spitzer, and Morgan C Giddings. 2013. “Peppy: Proteogenomic Search Software.” Journal of Proteome Research 12 (6) (June): 3019–25.

Rutherford Kim, Julian Parkhill, James Crook, Terry Horsnell, Peter Rice. 2000. “Artemis: Sequence Visualization and Annotation.” Bioinformatics 16: 944–945.

Sanders, William S, Nan Wang, Susan M Bridges, Brandon M Malone, Yoginder S Dandass, Fiona M McCarthy, Bindu Nanduri, Mark L Lawrence, and Shane C Burgess. 2011. “The Proteogenomic Mapping Tool.” BMC Bioinformatics 12 (1): 115.

Seifert, Jana, Florian-Alexander Herbst, Per Halkjaer Nielsen, Francisco J Planes, Nico Jehmlich, Manuel Ferrer, and Martin von Bergen. 2013. “Bioinformatic Progress and Applications in Metaproteogenomics for Bridging the Gap between Genomic Sequences and Metabolic Functions in Microbial Communities.” PROTEOMICS 13 (18-19) (August): 2786–804.

Serang, Oliver, Michael J MacCoss, and William Stafford Noble. 2010. “Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data.” Journal of Proteome Research 9 (10): 5346–5357.

Shadforth, Ian, Daniel Crowther, and Conrad Bessant. 2005. “Protein and Peptide Identification Algorithms Using MS for Use in High-Throughput, Automated Pipelines.” Proteomics 5 (16) (November): 4082–95.

Shmatkov, Anton M, Arik A Melikyan, Felix L Chernousko, and Mark Borodovsky. 1999. “Finding Prokaryotic Genes by the ‘Frame-by-Frame’ Algorithm: Targeting Gene Starts and Overlapping Genes.” Bioinformatics 15 (11): 874–886.

Stanke, Mario, Ana Tzvetkova, and Burkhard Morgenstern. 2006. “AUGUSTUS at EGASP: Using EST, Protein and Genomic Alignments for Improved Gene Prediction in the Human Genome.” Genome Biology 7 Suppl 1: S11.1–8.

Sturm, Marc, Andreas Bertsch, Clemens Gröpl, Andreas Hildebrandt, Rene Hussong, Eva Lange, Nico Pfeifer, et al. 2008. “OpenMS – An Open-Source Software Framework for Mass Spectrometry.” BMC Bioinformatics 9 (1): 163.

Sultana, Tamanna, Rick Jordan, and James Lyons-Weiler. 2009. “Optimization of the Use of Consensus Methods for the Detection and Putative Identification of Peptides via Mass Spectrometry Using Protein Standard Mixtures.” Journal of Proteomics & Bioinformatics 2 (6) (June): 262–273.

Tabb, David L, Ze-Qiang Ma, Daniel B Martin, Amy-Joan L Ham, and Matthew C Chambers. 2008. “DirecTag: Accurate Sequence Tags from Peptide MS/MS through Statistical Scoring.” Journal of Proteome Research 7 (9) (September): 3838–46.

Tabb, David L, Christopher G Fernando, and Matthew C Chambers. 2007. “MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis.” Journal of Proteome Research 6 (2): 654–661.

Tanca, Alessandro, Massimo Deligios, Maria Filippa Addis, and Sergio Uzzau. 2013. “High Throughput Genomic and Proteomic Technologies in the Fight against Infectious Diseases.” The Journal of Infection in Developing Countries 7 (3) (March):182-90.

Tanner, Stephen, Ari Frank, Ling-Chi Wang, Marc Mumby, and Pavel A Pevzner. 2005. “InsPecT : Fast and Accurate Identification of Post-Translationally Modified Peptides from Tandem Mass Spectra.” Science 4978: 1–22.

Taylor, J Alex, and Richard S Johnson. 2001. “Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry.” Analytical Chemistry 73 (11) (June 1): 2594–604.

Thorvaldsdóttir, Helga, James T Robinson, and Jill P Mesirov. 2013. “Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration.” Briefings in Bioinformatics 14 (2): 178–192.

Tu, Chengjian, Paul A Rudnick, Misti Y Martinez, Kristin L Cheek, Stephen E Stein, Robbert J C Slebos, and Daniel C Liebler. 2010. “Depletion of Abundant Plasma Proteins and Limitations of Plasma Proteomics.” Journal of Proteome Research 9 (10) (October): 4982–4991.

Uszkoreit, Julian, Nicole Plohnke, Sascha Rexroth, Katrin Marcus, and Martin Eisenacher. 2014. “The Bacterial Proteogenomic Pipeline.” BMC Genomics 15 Suppl 9 (Suppl 9): S19.

Van Damme, Petra, Daria Gawron, Wim Van Criekinge, and Gerben Menschaert. 2014. “N-Terminal Proteomics and Ribosome Profiling Provide a Comprehensive View of the Alternative Translation Initiation Landscape in Mice and Men.” Molecular & Cellular Proteomics 13 (5) (May): 1245–1261.

Vanderperre, Benoît, Jean-François Lucier, Cyntia Bissonnette, Julie Motard, Guillaume Tremblay, Solène Vanderperre, Maxence Wisztorski, Michel Salzet, François-Michel Boisvert, and Xavier Roucou. 2013. "Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome." PLoS One 8 (8) (August): e70698.

Venter, Eli, Richard D Smith, and Samuel H Payne. 2011. "Proteogenomic Analysis of Bacteria and Archaea: A 46 Organism Case Study." PloS One 6 (11) (January): e27587.

Wang, Nan, Shane Burgess, Mark Lawrence, and Susan Bridges. 2009. "Proteogenomic Mapping for Structural Annotation of Prokaryote Genomes." 2009 International Joint Conference on Bioinformatics Systems Biology and Intelligent Computing: 103–106.

Wang, Xiaojing, Robbert J C Slebos, Dong Wang, Patrick J Halvey, David L Tabb, Daniel C Liebler, and Bing Zhang. 2012. "Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data." Journal of Proteome Research 11 (2): 1009–17.

Wenger, Craig D, and Joshua J Coon. 2013. "A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra." Journal of Proteome Research 12 (3) (March): 1377–86.

Wilkins, Michael J, Nathan C Verberkmoes, Kenneth H Williams, Stephen J Callister, Paula J Mouser, Hila Elifantz, A Lucie N'guessan, et al. 2009. "Proteogenomic Monitoring of Geobacter Physiology during Stimulated Uranium Bioremediation." Applied and Environmental Microbiology 75 (20): 6591–6599.

Wilming, Laurens, and Jennifer Harrow. 2012. "Annotation Guidelines." Accessed July 19 http://www.sanger.ac.uk/science/projects/manual-annotation.

Woo, Sunghee, Seong Won Cha, Gennifer Merrihew, Yupeng He, Natalie Castellana, Clark Guest, Michael MacCoss, and Vineet Bafna. 2013. "Proteogenomic Database

Construction Driven from Large Scale RNA-Seq Data." Journal of Proteome Research 13 (1) (July 17): 21–28.

Wright, James C, Mark O Collins, Lu Yu, Lukas Kall, Markus Brosch, and Jyoti S Choudhary. 2012. "Enhanced Peptide Identification by Electron Transfer Dissociation Using an Improved Mascot Percolator." Molecular & Cellular Proteomics : MCP (April 6): 478–491.

Wu, Sun, and Udi Manber. 1994. "A Fast Algorithm for Multi-Pattern Searching." Tech. Rep. TR94 (17): 1 – 11.

Yates, John R. 2000. "Mass Spectrometry. From Genomics to Proteomics." Trends in Genetics : TIG 16 (1) (January): 5–8.

Yates, John R, Jimmy K Eng, and Ashley L McCormack. 1995. "Mining Genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases." Analytical Chemistry 67 (18) (September): 3202–10.

Yilmaz, Şule, Bjorn Victor, Niels Hulstaert, Elien Vandermarliere, Harald Barsnes, Sven Degroeve, Surya Gupta, et al. 2016. "A Pipeline for Differential Proteomics in Unsequenced Species." Journal of Proteome Research 15 (6): 1963–1970.

Zhang, Jing, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles a Lajoie, and Bin Ma. 2012. "PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification." Molecular & Cellular Proteomics : MCP 11 (4) (April): M111.010587.

Zhou, Chen, Hao Chi, Le-Heng Wang, You Li, Yan-Jie Wu, Yan Fu, Rui-Xiang Sun, and Si-Min He. 2010. "Speeding up Tandem Mass Spectrometry-Based Database Searching by Longest Common Prefix." BMC Bioinformatics 11 (1): 577.

Zickmann, Franziska, and Bernhard Y. Renard. 2015. "MSProGene: Integrative Proteogenomics beyond Six-Frames and Single Nucleotide Polymorphisms." Bioinformatics 31 (12): i106–i115.

VITA

Date of Birth: March 12, 1986, İzmir

Work Experience:

December 2010-Present: Research asistant-İzmir Institute of Technology, Molecular Biyology and Genetics Department

Education and Training:

2012-2016 PhD- İzmir Institute of Technology, Molecular Biology and Genetics Department

2009-2012 MSc.- İzmir Institute of Technology, Molecular Biology and Genetics Department

2006-2007 BSc.- Erasmus Student University of Copenhagen, Genetics Department

2004-2009 BSc.- Istanbul University, Molecular Biology and Genetics Department

Certificates and Traning:

2015 summer- Software Development Internship, KNIME.com AG, Berlin

2013 summer- Visiting Scientist, Algorithmic Bioinformatics, Informatics Institute, Freie University, Berlin, Germany) with DAAD Short Term Research Stay Fellowship

Scientific Papers:

Has C, Lashin S.A, Kochetov A, Allmer J PGMiner reloaded, fully automated proteogenomic annotation tool linking genomes to proteomes, Journal of Integrative Bioinformatics

Has C, and Allmer J, PGMiner: Complete Proteogenomics Workflow; from Data Acquisition to Result Visualization, Information Sciences, doi: 10.1016/j.ins.2016.08.005.

Has C, Kundakci CU, Altay A, and Allmer J, Ranking Tandem Mass Spectra: and the Impact of Database Size and Scoring Function on Peptide Spectrum Matches, IEEE Xplorer, doi: 10.1109/HIBIT.2013.6661686.