universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Peptidome-DB: A proteogenomic database to annotate the peptidome“

verfasst von / submitted by

Sebastian Didusch, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2020 / Vienna, 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 910
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Computational Science
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Acknowledgment

First and foremost I would like to thank my supervisors. I want to thank Thomas Rattei for his efforts and support throughout this thesis. The mentoring provided by him through regular meetings and discussions were a tremendous help in facing the challenges and achieving the goals of this work. Special thanks are also due to my co-supervisor Harald Marx who introduces me to the exciting field of proteogenomics and who also invested a lot of his time and efforts and gave advice that helped me to successfully finish my thesis.

Furthermore I would like to acknowledge all members of the CUBE - Florian, Gabi, Patrick, Roman, Susana, Samuel, Javier, Silvia, Lukas, Michael and also Lovro, Jean and Loic for the good discussions and wonderful working environment. I also want to thank the DOMEies for their welcoming environment especially Craig.

Last but not least I want to express my deepest gratitude to my family, to my loving parents Christine and Thomas for their endless support and to my brother Daniel. Special thanks go to my dear friends Julian, Marius, Nico, Adrian, Max, Pascal, Robin, Lisa, Hofer, Franzi, David, Vroni and Karo for their kindness and all the good people from Isaac's for helping me out in difficult times.

Abstract

Proteogenomics leverages next generation sequencing (DNA, RNA) and mass spectrometry (MS)-based proteomics data to refine existing gene models and identify novel gene models, including open reading frames (ORF) encoding bioactive peptides, e.g small ORFs (smORFs). Recent estimates indicate that hundreds of thousands of those ORFs may exist across the tree of life, that would notably expand the known peptidome. However, computational predictions of smORFs are particularly difficult and reference protein database contain only few entries.

To improve efforts in structural genome annotation of eukaryotes an integrated pipeline was build which results output proteogenomic, novel peptides and a complementary search database to Ensembl. Peptidome-DB is an integrated Java pipeline comprising data preprocessing, peptide mapping, ORF clustering and proteogenomic classification. A web prototype allows users to upload results from proteomic database search tools and the respective protein sequence database. The pipeline performs an exact string-matching (Aho-Corasick algorithm) on the peptides from the search engine to their sequence database entries. BLAST and Exonerate align entries against Ensembl databases from various molecular types (PEP, cDNA, ncRNA, DNA). Single-linkage clustering (SLC) groups ORFs from unannotated loci with exons from Ensembl PEP together. An optimal SLC distance threshold from the intragenic and intergenic distance distributions from Ensembl is calculated and a decision tree classifies ORFs from novel peptides into an intragenic or intergenic event.

I developed a pipeline, Peptidome-DB, for the incorporation and curation of proteogenomic data. Peptidome-DB utilizes Ensembl genome assemblies to refine gene models and annotate the peptidome with protein-level evidence.

Kurzfassung

Die Proteogenomik nutzt Next Generation Sequencing (DNA, RNA) und Massenspektrometrie (MS)-basierende Proteomikdaten, um bestehende Genmodelle zu verfeinern und neue Genmodelle zu identifizieren, einschließlich offener Leserahmen (ORF), die für bioaktive Peptide kodieren, z. B. small ORFs (smORFs). Jüngste Schätzungen deuten darauf hin, dass Hunderttausende dieser ORFs im gesamten Lebensbaum existieren könnten, die das bekannte Peptidom erheblich erweitern würden. Computerbasierte Methoden zur Vorhersage von smORFs sind jedoch besonders schwierig und Referenzproteindatenbanken enthalten nur wenige Einträge.

Zur Verbesserung der strukturellen Annotation von eukaryotischen Genomen wurde eine integrierte Pipeline erstellt, die nicht annotierte Peptide und eine komplementäre Suchdatenbank zu Ensembl ausgibt. Peptidome-DB ist eine integrierte Java-Pipeline, die Datenvorverarbeitung, Peptid-Mapping, ORF-Clustering und proteogenomische Klassifizierung umfasst. Ein Web-Prototyp ermöglicht es Benutzern, Ergebnisse von Proteomik-Suchmaschinen und der entsprechenden Proteinsequenzdatenbank hochzuladen. Die Pipeline führt ein exaktes String-Matching (Aho-Corasick-Algorithmus) für die Peptide aus der Suchmaschine zu ihren Sequenzdatenbankeinträgen durch. BLAST und Exonerate gleichen Einträge mit Ensembl-Datenbanken verschiedener molekularer Typen (PEP, cDNA, ncRNA, DNA) ab. Single-Linkage Clustering (SLC) gruppiert ORFs von nicht annotierten Loci mit Exons von Ensembl PEP zusammen. Ein optimaler SLC-Distanzwert aus der intragenen und intergenen Distanzverteilung von Ensembl und ein Entscheidungsbaum klassifizieren jeden ORF eines unannotierten Peptids in ein intragenes oder intergenes Ereignis.

Ich habe eine Pipeline entwickelt, Peptidome-DB, zur Aufnahme und Kuratierung proteogenomischer Daten. Peptidome-DB verwendet Genom-Assemblies von Ensembl, um Genmodelle zu verfeinern und das Peptidom zu erforschen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Proteins are the functional units of an organism participating in practically every process within a cell. They consist of amino acid residues which are linked to each other by peptide bonds. The proteome of an organism is defined as the entirety of all expressed proteins [1] and proteomics is the large-scale study of the proteome of a species. The peptidome is part of the proteome and consists of peptides originating from larger precursor proteins by proteolysis and peptides originating from short open reading frame (ORF) encoding bioactive peptides, sequences smaller than 100 amino acids (aa) which can be directed translated from small open reading frames (smORFs). Endogenous peptides play a critical role in many physiological processes in the intracellular and extracellular space, they are important as antibiotics, signal molecules such as neurotransmitters and hormones, or as regulators of protein expression but may also be linked to disease [2].

The state of the art method for proteomics is a combination of liquid chromatography (LC) and tandem mass spectrometry (MS/MS) referred to as 'shotgun-proteomics' or 'bottom-up proteomics' [3]. Here proteins are enzymatically cleaved into smaller peptides and their MS/MS spectra are measured and most commonly identified by matching observed spectra against theoretical spectra obtained from a protein sequence database. A protein sequence database should be complete and include all correctly annotated protein-coding genomic loci of a species. Despite the great efforts in structural genome annotation by database consortia many peptides are not present in reference databases because of mutations, alternative splice forms or novel protein-coding loci. This is also the case for smORFs which are particularly difficult to predict computationally [4].

Proteogenomics is connecting customized protein sequence databases from genomic or transcriptomic data together with mass spectrometry (MS) - based data to help efforts in structural genome annotation. MS - based proteomics data provides experimental evidence of gene expression hence it can be used to identify novel peptides which are not present in canonical protein sequence databases [3].

The rapid decrease in next-generation sequencing (NGS) cost and the technological advancement of mass spectrometry lead to a magnified interest in proteogenomic re-

search. Proteogenomics is used to refine existing gene models and to annotate novel gene models [5][6] for different domains of life albeit it is especially useful in higher eukaryotes where alternative splicing or the expression of as non-coding annotated transcripts can be validated [7][8][9].

1.1 Proteogenomics

The idea to compare unidentified tandem mass spectra to translated nucleotide sequences came before the high-throughput era in the genomic age had fully begun. A mere fraction of the human genome, 0.6% of its size had been sequenced in 1995 when John R. Yates III, Jimmy K. Eng, and Ashley L. McCormack published the first paper [10] in the yet to be named field of proteogenomics. They concluded that the increasing numbers of available genomic sequences could become very useful for MS-based proteomic data analysis. As a search database nucleotide sequences obtained from different species were translated in all six reading frames and they could demonstrate that their approach was promising to lead to robust peptide identifications.

Since then advancements in NGS technologies lead to a rapid increase in the number of sequenced genomes across all domains of life and automated gene prediction tools helped in annotating the protein-coding genes of these genomes. However, gene predictions are particularly difficult for eukaryotic species, the accuracy of gene predictions decreases from the nucleotide level to exons to complete gene structures [11] and some protein-coding genes do not have any predictions. MS-based evidence of protein expression can help in correcting incorrect or missing gene predictions.

In early 2001 half of the human genome was still in its draft form, when Choudhary et al. [12] worked with this version and an expressed sequence tags database (dbEST) to find novel human genes. In this publication the issue of decreased sensitivity in peptide identifications due to an inflated search space was already discussed, an issue which still remains to this day.

The term *proteogenomics* can first be found in literature in 2004 [13]. While at the beginning the term was defined as using proteomic data to improve structural genome annotation efforts, the meaning nowadays also includes the identification of novel peptides through the usage of customized sequence databases with novel proteins obtained from genomic or transcriptomic sources in general [3].

The applications of proteogenomics are diverse; sequence database consortia are interested in corroborating and refining predicted gene models. Alternative splicing, one of the major sources of cell - and tissue specific variation in proteins can also be identified with a proteogenomic approach. Together with sequence variants like single amino acid variants (SAAVs) this can be very useful for the detection of abnormal protein variants in cancer research [14]. Metaproteomics of microbial communities from environmental

samples is another research area in which proteogenomics is applied, the challenging part here is that usually metagenomic or metatranscriptomic data is needed to build a good search database and to infer the identified peptides to the correct organisms [15]. The same peptides may be expressed by homologous proteins making their identification indistinguishable. On top of that homologous proteins can be present in different species. On account of this a taxonomic classification from metagenomic or metatranscriptomic data can determine the presence of a species in the sample.

1.2 Biological background

Transcripts from mammalian genomes can originate from genic regions or intergenic regions. While most transcripts from RNA-Sequencing studies can be mapped back to annotated genes (the molecular structure of an eukaryotic gene can be seen in Figure 1.1), this is not the case for all of them. Interestingly, most intergenic transcripts from human samples are unspliced and short and can be found in multiple cell lines [16]. The function of those 'dark matter' transcripts is largely unknown, they could be potentially translated into small polypeptides and may harbor yet unannotated smORFs. Transcripts from known gene models may also be important, smORFs can be located upstream of a protein-coding region in the 5' untranslated region (UTR). Furthermore frameshifts in protein-coding sequences may result in the expression of smORFs.

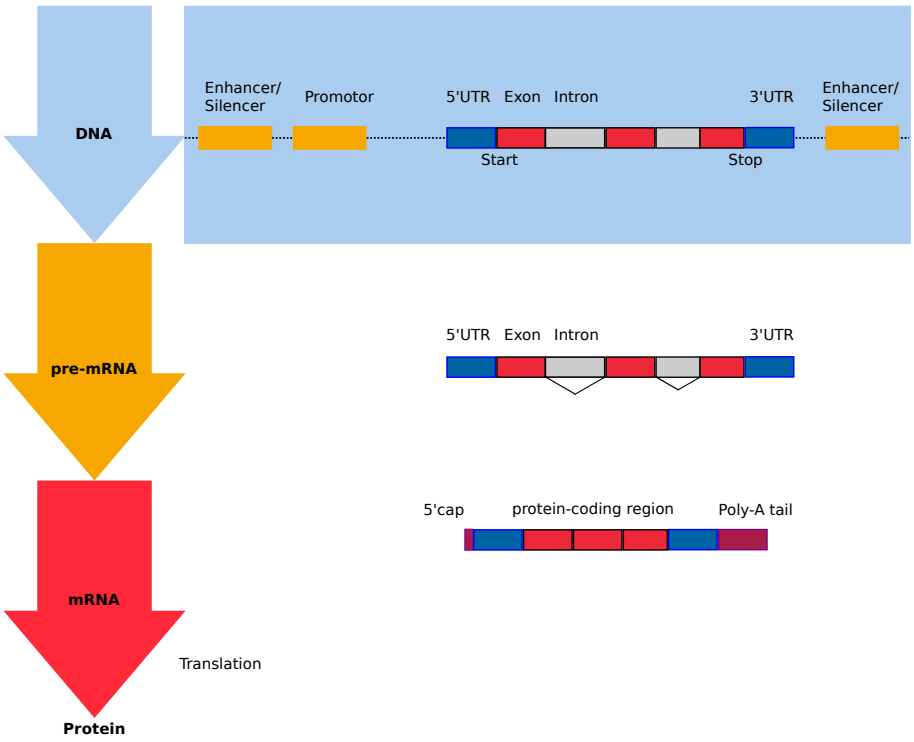


Figure 1.1: Structure of an eukaryotic gene from transcription to translation. A regulatory region with enhancers or silencers can often be found up - or downstream a gene. These elements bind to transcription factors that regulate the transcription from DNA to RNA. An enhancer helps increasing transcription by binding an activator protein, while a silencer on the other hand reduces transcription by binding a repressor protein. Another regulatory sequence is the promotor that helps the RNA polymerase to bind to the DNA with the effort of different transcription factors. After transcription a pre-messenger RNA (pre-mRNA) consists of a 5' and 3' untranslated region. The pre-mRNA contains exons and introns where the latter get removed from the pre-mRNA by a process called splicing. To the finished mRNA a cap is added to its 5'UTR and a poly-A tail is added downstream, this is a sequence that consists of many adenosines. The 5'cap and the poly-A tail are both important for the stability of the mRNA.

1.2.1 Small open reading frames (smORFs)

Classical, well described bioactive peptides such as insulin are cleaved from a precursor protein and determined for the secretory pathway by their N-terminal signal peptides [17]. Small open reading frames, sometimes also referred to as micropeptides, short open reading frames (sORFs), or short open reading frame-encoded polypeptides (SEPs) on the other hand are not cleaved from any precursor protein and lack a signal peptide. They get released directly into the cytoplasm after translation.

Eukaryotic genomes contain millions of smORFs, the public repository of small open reading frames sorfs.org [18] contains over 4 million entries. Most of them were obtained from ribosome profiling, a complementary technique to MS which is used to determine actively translated messenger-RNA (mRNA) sequences. It also includes small

data experimentally corroborated from MS.

The reasons why database consortia are excluding most smORFs from their annotations are manifold. First of all the high number of smORFs is opposed to the relatively stable number of predicted and observed proteins by standard computational and biochemical methods. Because most of these smORFs are lacking any biological relevant function and could be present in genomes by mere chance alone ORF length is a valid parameter in distinguishing coding from non-coding RNAs. The FANTOM consortium for example used an initial length cutoff of 100 aa for protein-coding sequences [19] in their analysis of mouse transcriptomic data.

Another point is that most computational gene predictions are relying on intrinsic (similarity to known proteins from the same organism) and extrinsic (conservation across species) factors. Conservation scores such as the K_a/K_s measure, the ratio of the rate of nonsynonymous (a nucleotide mutation that alters the amino acid) to the rate of synonymous mutations, are more difficult to apply for smaller sequences than for canonical proteins because a ratio that would indicate conservation is more likely to happen by chance [20]. The presence of non-canonical start codons (non AUG) that can be observed by many smORFs [21] further complicates an accurate prediction by computational methods.

The low number of annotated smORFs in public databases leads to issues in proteomic database searches, sequences not present in the database can not be experimentally reassured and sequences which could not be experimentally reassured can not be confidently annotated.

Classification of smORFs

Couso et al. [22] established a classification of smORFs into 5 classes:

1. intergenic smORFs: They make up the largest amount of smORFs in a genome and are considered non-functional.

2. upstream ORFs (uORFs): Located in the 5'UTR of genes. Their expression can lead to a regulation of the downstream gene.

3. lncRNA ORFs: A smORF located in a long noncoding RNA genomic locus.

4. short CDS: annotated smORFs in canonical mRNAs which are below 100 codons.

5. short isoform ORFs: A short polypetide produced by alternative splicing of a canonical mRNA.

These classes differ in their median length, average taxonomic level of ORF conservation, and features of the encoded amino acids, such as the prevalence of transmembrane

α-helices [22]. Intergenic smORFs for example seem to be generated at random out of a stochastic process and have very likely no biological function. Most long noncoding RNAs contain at least one ORF although it is unclear if it is translated. Upstream ORFs are present in 40 % of mammalian mRNAs [23], albeit they often miss a protein domain, are rarely conserved and their translation efficiency is thought to be low.

1.2.2 Examples of smORFs

GCN4 uORFs

The yeast gene GCN4 contains four uORFs in its 5'UTR. Under normal growth conditions the uORFs are translated and the ribosome dissociates. This inhibits the translation of the downstream main ORF. The GCN4 gene gets expressed under limited growth conditions, when amino acids or carbohydrates are missing. Under these conditions the initial uORFs are not translated due to leaky scanning and the ribosome translates the main ORF [24].

tarsal-less (tal)

A new type of noncanonical gene was described by Galindo et al. [25] in 2007 when they discovered that the previously as non-coding RNA annotated tarsal-less (tal) gene was in fact translated into 11 amino acids long peptides (tal-1A, tal-2A, tal-3A and tal-AA) from several 33 nucleotides long ORFs. The short and unprocessed peptides from a polycistronic mRNA have a key function in morphogenesis and development in Drosophila. There are several homologous genes in other species identified for this directly translated gene.

Humanin

This small mitochrondrial-derived peptide plays a crucial role in retrograde signaling, the communication between the mitochondrion and the cell [26]. Humanin, which is also highly conserved across species, exhibits strong cyto-protective protection. It is able to provide protection from Alzheimer's disease - related toxicities. Intra-cellular humanin can prevent apoptosis by interacting with pro-apoptic proteins. Extra-cellular humanin is able to regulate important cellular processes such as metabolism and inflammation.

1.3 Sequence databases

Proteomics usually involves a database search to identify tandem mass spectra. An optimal sequence database should only include correct gene models and be complete. On the other hand the database cannot simply be crammed with all possible putative

coding sequences because of increasing false discovery rates (FDRs) with an enlarged search space.

Database consortia integrate automated computational annotations with manually curated data. The following three databases include protein sequence databases in the FASTA format to download. Ensembl and RefSeq also include genomic and transcriptomic FASTA files.

Ensembl [27] integrates experimental data from various sources (transcriptomics data from cDNA) with an automated pipeline including gene-predictions and comparative genomics to generate gene models. It also includes a multitude of cross references such as species specific databases like the Saccharomyces Genome Database, FlyBase, WormBase. A biotype is assigned on the gene level (i.e protein-coding, pseudogene, lncRNA among others) which is extended on the transcript level (i.e protein-coding, nonsense mediated decay (NMD)). For the best annotated organisms like *Homo sapiens* the GENCODE consortium provides manually refined annotations (HAVANA: Human and Vertebrate Genome Analysis and Annotation) that are regularly integrated into automated Ensembl annotations, which constitute the majority of annotations (Figure 1.2).

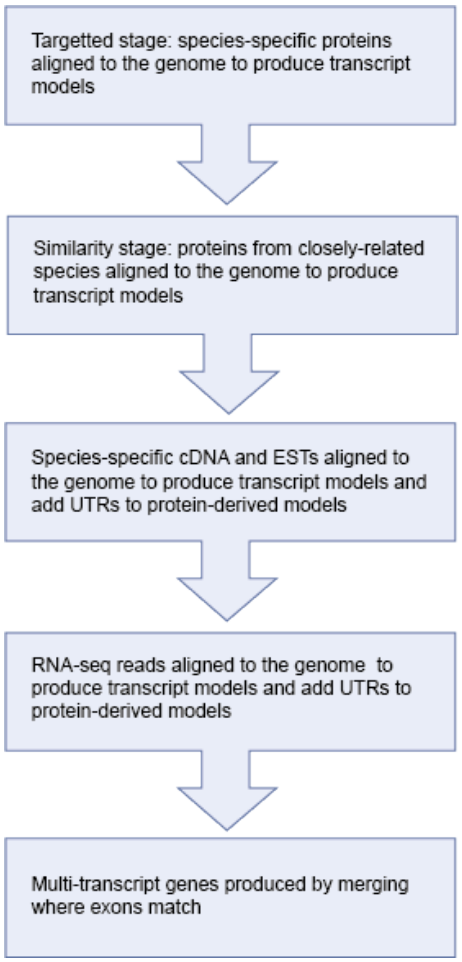


Figure 1.2: Ensembl annotation pipeline. Figure taken from reference [28]. In the **Targeted stage** species-specific proteins are aligned with Exonerate to the genome to build a transcript structure. In the **Similarity stage** homologous proteins from other closely related organisms are used to build a transcript structure where the targeted stage could not produce one. In the next stages species-specific cDNA and RNA-Seq reads are aligned to further extend the transcript models, including UTRs. Some transcripts have the same exons, these are merged to form genes.

RefSeq [29] is an abbreviation for Reference Sequence and is hosted by the NCBI. It holds three levels of annotated sequences, this includes the annotations from a GenBank record of the submitter, curated annotations from species specific databases or the NCBI Eukaryotic Genome Annotation Pipeline. Similar to Ensembl this automated pipeline relies on sequences from closely related organisms or from species-specific source:

- Transcript alignments: taken from various sources, may include known transcripts from other species.
- RNA-Seq read alignments: spliced alignments generated by Splign.
- Protein alignments: taken from various sources, may include known transcripts from other species. alignments generated by ProSplign.

Additionally, these alignments are passed to Gnomon, a gene prediction tool which generates gene models by "chaining" together non-conflicting alignments into putative models [30]. Gnomon is also able to extends the models created by alignments by ab initio gene predictions. Gnomon predictions are included into RefSeq if there is not already a known RefSeq transcript with the same splice sites present for the species.

UniProt [31] comprises functional information on proteins. The UniProt Knowledgebase (UniProtKB) is divided into the manually curated and reviewed Swiss-Prot database and the automatically annotated and not reviewed entries of TrEMBL. Most UniProtKB entries come from the translation of coding sequences (CDS) from submitted sequences to the International Nucleotide Sequence Database Collaboration (INSDC), which were generated by gene predictions or experimentally proven. These CDSs are stored in the TrEMBL section and can be selected for manual curation, which involves following steps:

1. Sequence curation: alignments identify homologous sequences, discrepancies between sequence reports are gathered such as alternative splicing, natural variations, frameshifts, incorrect initiation sites, incorrect exon boundaries and unidentified conflicts to provide a complete and accurate sequence.
2. Sequence analysis: The output of prediction tools for PTMs, subcellular location, transmembrane domains and protein topology, domain identification and protein family classification are manually reviewed.
3. Literature curation
4. Family-based curation: homologous sequences are searched in other organisms
5. Evidence attribution
6. Quality assurance, integration and update

UniProtKB uses a classification system that includes five levels of evidence for a protein entry, taken from reference [32]:

PE 1) Experimental evidence at protein level: identification by X-ray or NMR structure (applied to study the 3D molecular structure of proteins), mass spectrometry, detection by antibodies, good protein-protein interaction or complete Edman sequencing.

PE 2) Experimental evidence at transcript level: existence of expression data (cDNAs, RT-PCR or Northern blots).

PE 3) Protein inferred by homology: clear orthologs exists in related species.

PE 4) Protein predicted: No experimental evidence or evidence by homology present.

PE 5) Protein uncertain: existence is unsure because the protein might be a product of a pseudogene, a dubious CDS prediction or a dubious gene prediction.

1.3.1 Challenges of annotations from ab initio gene prediction tools

Ab initio gene prediction methods can be used as standalone methods in gene finding or together with experimental evidence. On the one hand *ab initio* predictions work with signal sensors such as promotors, ribosomal binding sites (RBSs), start or stop codons, splice donor and acceptor sites and poly-A features. Content sensors on the other hand help in determining exons, introns, UTRs or intergenic regions. Exons for example have different hexamer frequencies of oligonucleotides than intronic or intergenic regions. A hidden markov model (HMM) can be used to model a system with different discrete states and transitions between them [11]. Among the most common challenges are:

- Alternative splicing.
- Long/short genes.
- Long/short introns.
- Non-canonical start codons

The reason why short, intronless genes are difficult to predict computationally is because they have a lower occurrence than longer genes with introns. Even though content sensors such as hexamer frequencies might be present in the short gene it could still get confused as an exon or even as a pseudogene [11].

1.3.2 Challenges of annotation efforts with RNA-Seq

The best evidence for the expression of a protein is self-explanatory to observe it at the protein level, while RNA-Seq has many advantages over proteomics (cost, no physical isolation or separation, copying transcripts by PCR, sequencing, etc.) a true validation of an expressed protein can only be provided by experimental evidence such as proteomics. A complementary DNA (cDNA) can be synthesized from mRNAs with the enzyme reverse transcriptase. Gene predictions based on cDNAs might be erroneous because of premature transcript sequences, partial mRNA/ESTs or sequencing errors from single-pass mRNA sequencing projects [33]. The presence of a transcript does not necessarily indicate that it will be translated into a protein; it could be a non-coding RNA or simply degraded. The absence of a transcript does not lead to the conclusion that no protein was expressed because proteins are slower degraded than transcripts (in mammalians proteins have an average half-life of 46 hours compared to 2 - 7 hours of mRNAs [34]). On the other hand transcriptomics is a complementary technique to proteomics, without the sequencing efforts generating a great many of known and especially novel transcripts the construction of sequence databases for proteogenomics was not so easily possible.

1.3.3 Sequence databases in proteogenomics

In the field of proteogenomics customized protein sequence databases are used. Usually they are combined with one the aforementioned reference sequence databases. The simplest way to generate a customized database is a six-frame translation (6FT) of a genome. This will result in an inflated database that contains a large proportion of sequences without any biological function (a 6FT of the UCSC v.19 genome resulted in a 3.2-gigabase (Gb) database, a 70-fold increase to the corresponding Ensembl reference database of roughly 45 Mb [3][35]). Filtering steps need to be applied to reduce the amount of uninformative data. Furthermore transcriptomic data can be used for database construction, as well as variant sequence databases that might detect single amino acid variations, deletions or insertions. Last but not least specialized databases for pseudogenes or noncoding sequences could also be chosen, depending on the research question.

1.4 Analytical high-throughput methods to detect smORFs

1.4.1 Ribosome profiling

Translation is located on the sites of the ribosomes and it is the process of translating an mRNA template into a protein. Ribosome profiling is a method to measure translation globally and *in vivo*, a translating ribosome is protecting around 30 nucleotides of an mRNA from the activity of an added nuclease [36]. The nuclease destroys unprotected regions and results in fragments, termed 'ribosome footprints'. These fragments are sequenced and mapped back to their mRNAs and subsequently mapped back to the genome. The measured density of them serves as a proxy for the rate of protein synthesis. The fact that the genomic loci is known gives the start and end positions and even the used reading frame of a transcript. Protein-coding transcripts with alternative translation initiation sites or smORFs can be detected with ribosome profiling.

An advantage of this method compared to MS-based methods is that no sequence database is needed. Furthermore the translation initiation site can be detected. One of the disadvantages is that ribosomal occupancy does not tell whether a stable protein product is produced.

1.4.2 Basics of MS and computational MS

The typical workflow of a 'shotgun proteomics' experiment incorporates a combination of liquid chromatography (LC) and tandem mass spectrometry (MS/MS). Before peptides can be eluted through columns of LC to be analyzed by MS/MS proteins must be separated and enzymatically digested (Figure 1.3).

1.4.3 Sample preparation

Proteins are usually present in a complex mixture with other proteins. To conduct a bottom-up proteomic experiment these proteins need to be separated to digest them and analyze their peptides. One general used technique is SDS-PAGE (**S**odium **S**odecyl **S**ulfate **P**oly **A**crylamide **G**el **E**lectrophoresis). Separation occurs on molecule size, proteins migrate inverse proportional to their molecule size.

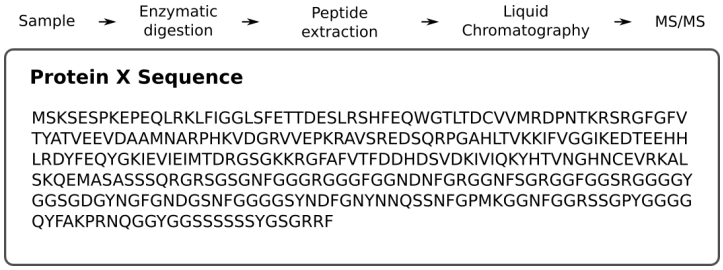
Another technique is isoelectric point focusing (IEF), where proteins are separated according to their isoelectric point, the pH gradient where the net charge of a molecule is zero. SDS-PAGE and IEF can be combined to separate molecules of interest in two dimensions.

1.4.4 Proteolytic enzymes

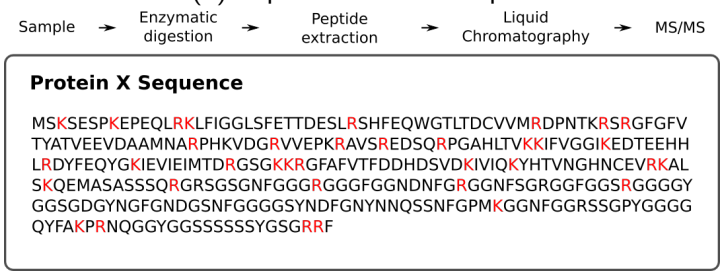
Depending on the experimental design there are several proteolytic enzymes used in proteomics. These so called proteases are found in all organisms. By far the most common used enzyme is trypsin. It is present in the small intestine where it hydrolyzes proteins. Trypsin cleaves exclusively C-terminal to arginine and lysine residues [37]. Its popularity in the community can be explained by its high cleavage specificity with few mis-cleavages and almost no cleavages at random positions. The average length of digested peptides is about 15 amino acids, resulting in optimal masses for the MS to analyze.

1.4.5 Liquid chromatography

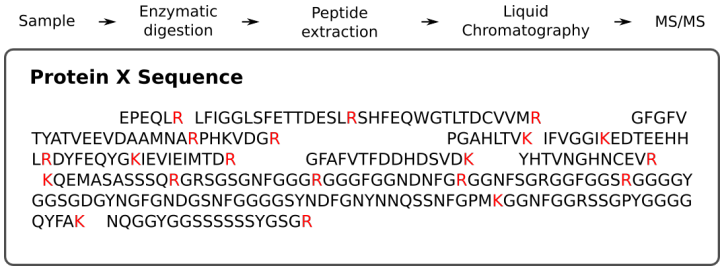
LC is used in proteomics to separate a mixture of peptides into individual components to further analyze them in the MS. This is usually achieved by injecting the mixture in a *mobile phase* into a porous, *stationary phase*. It takes each component a different *retention time* (RT) to separate because individual components differ in their interactions with the two phases. The peptides mixture is then eluted in the mobile phase. The LC system is coupled online with the tandem MS.



(a) A protein in our sample.



(b) All arginine and lysine residues are highlighted in red.



(c) Peptides after cleavage. For visualization and simplicity some peptides have been removed from this Figure.

Figure 1.3: Enzymatic digestion of a protein with trypsin. Adapted from reference [38].

1.4.6 Mass spectrometry

All mass spectrometers consist of three main components. An **ion source** adds additional charge(s) to a compound. This has to be done because mass spectrometers work with electrical or electromagnetic fields. The **mass analyzer** is separating ions according to their mass to charge (m/z) ratio and finally the **detector** analyzes the ions and creates a mass spectrum. The detector is often integrated into the mass analyzer. The mass analyzer and the detector are in a vacuum to avoid measuring neutral compounds. In tandem MS two mass analzers are used. The first one selects ions (a precursor ion) at a certain m/z range (which is often the value of a single peptide of interest) that gets fragmented into fragment ions. The latter get measured at the second mass filter which generates MS/MS spectra. The information of the fragment ion peaks enables the reconstruction of the peptide sequence that generated the spectrum.

Ion source In single MS Matrix-Assisted Laser Desorption Ionization (**MALDI**) is preferred because it usually leads to single-charged ions. In this method peptides get

mixed with other compounds in a matrix. Short pulses of light (energy) are emitted from a laser and absorbed by the matrix. The introduced energy lets small part of the matrix close to the surface evaporate, freeing the embedded single-charged peptides.

In tandem MS multiple charged precursor ions (often 2+) are preferred due to the fact that the peptide backbone breaking in the fragmentation step leads to two fragment ions where each of them can only be detected if it is charged. This can be achieved with electronspray ionization (**ESI**) where high voltage is applied to a liquid containing the analyte. The liquid is sprayed from a heated needle or capillary into an electromagnetic field. This lets small droplets evaporate where charged peptides desorp from the surface.

Fragmentation In collision-induced dissociation (**CID**) or higher-energy collision-induced dissociation (**HCD**) an electrical potential is applied to the precursor ion which then accelerates to collide with neutral molecules (often noble gases), resulting in fragmentation through the conversion of kinetic energy to internal energy.

Fragmentation occurs mainly along the peptide backbone. The most common fragment ion types are b - and y-ions, see Figure 1.4. A single charged precursor ion will fragment into one charged fragment ion and one neutral "sister" fragment therefore multiple charged precursor ions are preferred in tandem MS. If the charge is on the N-terminal fragment the ion type of the ion is a a-, b - or c-ion and if the charge is on the C-terminal fragment the ion type is a x-, y - or z - ion. Other fragment types include internal fragments [39] which do not include the N-term residues nor the C-termn residues, iminium ions [39], water loss, ammonia loss and side chain fragments.

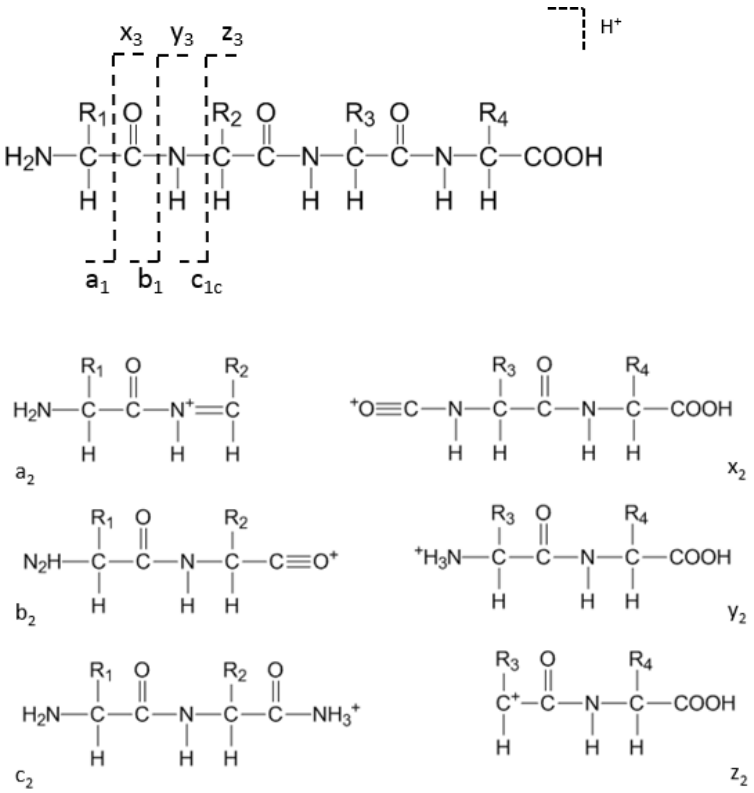


Figure 1.4: Common fragment types regularly observed in mass spectrometry. b-ions are counted from the N-terminus, and y-ions are counted from the C-terminus. Taken from reference [40].

On the computational side it is crucial to predict the intensities of mass spectra for spectral comparisons accurately, see section 1.4.7. Data driven machine learning approaches utilizing a large amount of tandem MS spectra are used to predict fragmentation models [41].

Mass analyzers and detectors The task of a mass analyzer (sometimes also referred as mass filter) is to filter ions on their mass-to-charge ratio where they get passed to a detector that generates a mass spectrum. Three types of mass analyzers are predominantly used in proteomics: trapping type instruments (quadrupole ion trap, linear ion trap, Fourier transform ion cyclotron resonance, and Orbitrap), quadrupole, and time of flight (TOF) instruments [42]. Different types of analyzers are described here:

- Quadrupole: Four metal rods that create an electric field that forces the ions to travel in a oscillatory motion according to their mass-to-charge ratio. If the ion is to light or to heavy it will collide with a metal rod [43].
- Time of flight (TOF) mass analyzer: Ions are separated by their velocity and kinetic energy. The TOF of distinct ions in a flight tube varies due to differences in their mass-to-charge ratio [43].

- Orbitrap: Can be used as mass analyzer and detector, it consists of two outer electrodes and one inner electrode [44]. Ions orbit around the inner electrode and can be separated after their m/z value. The mass-to-charge ratio can be calculated by a Fourier transform, because the angular frequency of an ion is depending on its m/z value.

Some the most common detectors in mass spectrometry are usually electron multipliers, photodiode arrays, microchannel plates, or image current detectors [42].

1.4.7 Peptide identification by tandem MS

All database search tools attempt to solve following problem formulation:

Find the peptides from a set of peptides $P = \{P_1, P_2, ..., P_n\}$ from a sequence database D which generated the spectra from the set of spectra from a spectral data set $S = \{S_1, S_2, ..., S_n\}$, if there is any [45].

The peptide that generated the spectra might come from a protein which is in D, it might come from a protein which has a homolog in D or the spectra might have been generated by a peptide with modifications.

The construction of theoretical peptides from in silico digests is important in spectral comparison. There are a few properties that need to be considered [45]:

- Considered fragment ions (b, y, a, z, H_2O^+)

- Intensities (uniform spectra where all peaks have the same height or intensities calculated from statistical or chemical model)

- Modifications: fixed modifications like carboxymethyl at cysteine changes the mass of a specific residues, variable modifications (like oxidation at methionine) may or may not be present at a single or even multiple residues.

The underlying scoring schemes of the spectral comparison can be divided into two categories, non-probabilistic and probabilistic scoring schemes.

Non-probabilistic scoring

Let I^R denote an intensity vector of an experimental spectra and I^T denote an intensity vector of a theoretical spectra. An intensity vector stores the intensity of the peaks at the index of the corresponding m/z.

Counting matching peaks The simplest comparison is to count the number of matching peaks at a given point of the m/z - axis. To calculate a score the dot product can be considered, where score $S = \sum_{j=1}^{n} I_j^R I_j^T$.

This very simple scoring scheme has the problem that some of the matching may only occur by chance (e.g noise peaks) and is therefore not very accurate [45].

Spectral contrast angle The cosine similarity is a measure of similarity for two non-zero vectors that measures the angle between them can be used for spectral comparison. If the vectors are similar to each other the angle will be close to 0° while an angle of 90° represents maximum dissimilarity. The formula is derived from the Euclidean dot product formula [45]:

$$\cos\theta = \frac{\sum_{j=1}^{n} I_j^R I_j^T}{\sqrt{\sum_j (I_j^R)^2 (I_j^T)^2}}$$

The spectral contrast angle is a good indicator for the correctness of a fragmentation model [46].

Cross-correlation The cross-correlation is often used in signal processing to measure the similarity between two series. It determines the correlation between two spectra. The cross-correlation function can be formulated as sliding dot product $\sum_{j=1}^{n} I_j^R I_{j+\tau}^T$, where τ is a displacement variable. Fourier transformation can be used to determine an appropriate value for τ. This method (among other methods) is used in one of the earliest published database search tools Sequest [47].

Probabilistic scoring

The idea behind probabilistic scoring is that ions of different fragment types (a,b,c,y,z) have different probabilities to be observed [48], which allows one to calculate the probability that spectrum S was generated by peptide P.

Log likelihood ratio Each fragment match is considered an independent Bernouilli event with different probabilities according to its fragment type. Two alternative hypotheses are considered, H_0 (random match) and H_1 (correct match), and the log ratio of them is calculated [45][49]:

$$L = \log \frac{P(E|D, s, H_1)}{P(E|D, s, H_0)},$$

where E is the information obtained from a spectral comparison (mass difference of the spectral comparison, fragment mass matches, charge of the precursor ion), D represent valuable background information like fragmentation statistics and s is the

peptide sequence. The probabilities of H_1 can be learned with statistical methods while the probabilities of H_0 can be estimated as an uniform distribution.

1.4.8 False Discovery Rates

A database search tool outputs a list of peptide spectrum matches (PSMs) between experimental MS/MS spectra and theoretical generated spectra with associated similarity scores for these matches. Each MS/MS spectra needs to be compared to numerous in silico generated spectra. Because also weak matches are provided in the output this raises the question which matches are correct and which are false positives.

The target decoy approach (TDA) [50] is the most established strategy to distinguish incorrect from correct peptide identifications. Here the database gets extended by a decoy database that contains reversed, shuffled or randomly generated sequences from the target database. Hits against the decoy database act as a null hypothesis since they can be considered false identifications. The number of target and decoy PSMs are used to calculate a false discovery rate (FDR) above a score threshold. A global FDR of 1% or lower is generally accepted in bottom-up proteomics.

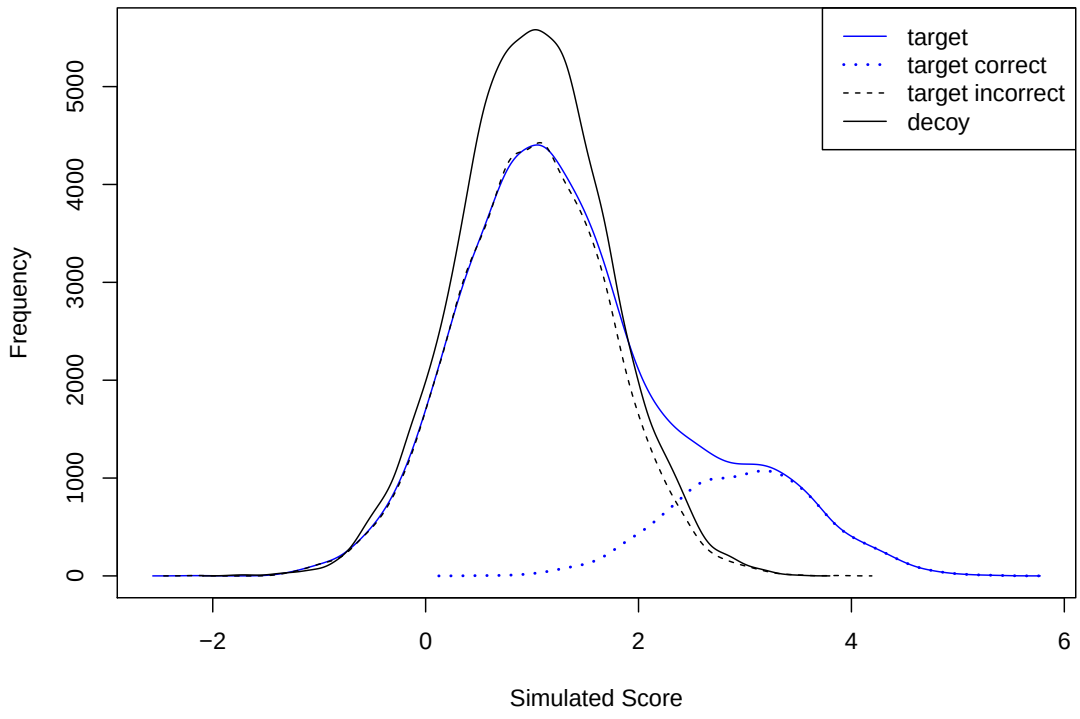


Figure 1.5: Score distribution of target and decoy hits. The target hits consist of correct hits and incorrect hits. Adapted from reference [51].

1.4.9 De Novo Sequencing

Database search tools limit peptide identifications to sequences in the provided protein sequence database which can be incomplete or contain mis-annotated sequences. In contrast large sequence databases (e.g from a six frame translation) consist mostly of non-existing protein sequences which increase the search space and result in difficulties estimating correct FDRs. De Novo Sequencing of tandem mass spectra is another strategy without prior knowledge of the amino acid sequence that is independent from reference databases.

The two most common errors in de novo sequencing include low mass tolerances in instruments and mass ambiguity among multiple amino acids [52]. Isobaric masses (leucine and isoleucine) for example are impossible to distinguish and specific amino acids like lysine and glutamine have a very small mass delta of 0.03638 Da. The combination of multiple residues may also lead to difficulties in accuracy as in the case of one asparagine and two glycines with a mass delta of 0.000002 Da. In some cases the b - and y-fragment ion ladders can become mixed up and and the sequence results in an inversion. The longer the peptide sequence is the higher is the likelihood for an error occurring in de novo sequencing because of larger sequence possibilities.

De Novo Sequencing was not used in this thesis.

1.4.10 Human Proteome Project Mass Spectrometry Data Interpretation Guidelines

The Human Proteome Project (HPP) of the Human Proteome Organization (HUPO) released guidelines to ensure the quality of MS-based proteomic data [53] and its interpretation. These are general guidelines but especially useful for proteogenomic studies where novel findings need to be validated conservatively. Among the guidelines for "extraordinary" detection claims (e.g. novel coding elements) are the following:

- present clearly annotated spectra.
- consider alternative explanations for PSMs.
- present clearly annotated spectra of synthetic peptides.
- consider isobaric masses and known SAAVs.
- provide two or more distinct uniquely-mapping peptide sequences of length ≥ 9 amino acids.

1.5 Tools for proteogenomics

The Proteogenomic Mapping tool [54] provides a graphical user interface (GUI) to map peptides from LC-MS/MS back to a six-frame translation of a genome which is done by an implementation of the Aho-Corasick algorithm. The tool also outputs ORFs from mapped peptides but it does not provide any kind of classification of the peptide.

Another tool called **PoGo** [55] claims to exhibit superior performance over other state of the art mapping tools. PoGo its output includes quantitative information, single nucleotide polymorphisms (SNPs), and post translational modifications (PTMs). PoGo needs a reference protein sequence file and a file with annotations, unfortunately it is only able to map peptides to annotated regions. However, it works very well as a visualization tool and has been integrated into the two well established frameworks for computational MS, namely the OpenMS [56] framework and the PRIDE toolsuit [57]. PRIDE is an acronym for PRoteomics IDEntification Database and is used as a data archive where MS raw files can be downloaded.

The open source software suite **PGTools** [58] goes a step further than the previously described tools. It includes multiple proteomic database search tools and multiple generated peptide database from RNA - Sequencing. In theory it would also allow the usage of custom databases, albeit the user would need some programming skills for that. The authors mention that user contribution is needed to expand the scope of usability of their software suite [58]. The usage of PGTools is limited to human samples only.

The integrated proteogenomics analysis workflow (**IPAW**) [8] extends the previous research by including a curation - and validation - step in their pipeline. The validation comprises several orthogonal sources such as conservation of novel genomic loci and their coding potential. The pipeline takes as input multiple protein sequence databases (reference proteome, variants, six-frame translation of the genome). Unfortunately the workflow is only applicable to human samples.

Last but not least there are many publications that were performed with custom scripts that were not made public or did not result in a finished, comprehensive and complete software. Castellana et al. [5] aided the efforts in structural genome annotation for the model organism plant *Arabidopsis thaliana* while Marx et al. [59] used a similar workflow for finding novel peptides in a quantitative study for the domestic pig *Sus scrofa*. Both publications are quite similar in their workflow to this thesis, even though both of them are not generalized for more than one organism. They include a peptide mapping step using BLAST (described in 2.1.3) and the BLAST - like alignment tool (BLAT) and perform a single-linkage clustering to differentiate between intergenic and intragenic novel peptides.

1.6 Aim of this thesis

Although the workflow of a proteogenomics study was defined by other authors there is a lack of a generalized software solution in the field. The previous work in the field is too often made up of individual scripts that are not available to the public. In many cases the important results of these studies are lost to fellow researchers because of a lacking central repository. Existing open source tools are frequently only applicable for individual organisms (i.e well studied organisms like *Homo sapiens*).

The first aim of this thesis is to automatize the existing workflow of proteogenomic analysis into a comprehensible, integrated pipeline. The second aim of this thesis is to provide the community with an open source pipeline that can deal with all proteogenomic data from annotated organisms in Ensembl resulting in complementary data to Ensembl. The final aim of this thesis is to create a prototype for a website. The front end should enable the community to browse analyzed data and to upload the database search results from their own experiment. The back end consists of a pipeline that was developed in this thesis that outputs un-annotated genomic loci of uploaded experiments. In addition a database was set up to store curated data. To my knowledge there exists no such software in the community and this work would alleviate efforts in structural genome annotation.

Chapter 2

Methods

2.1 Pipeline

A Java pipeline was implemented to facilitate the processing and integration of proteogenomic data. Technically the pipeline could accept different file formats from proteomic database search tools albeit at the moment only the output formats of MS-GF+ [60] and MaxQuant [61] are accepted. Beside a list of PSMs the pipeline needs the sequence database that was used to obtain these results as input, a graphical workflow is depicted in Figure 2.1.

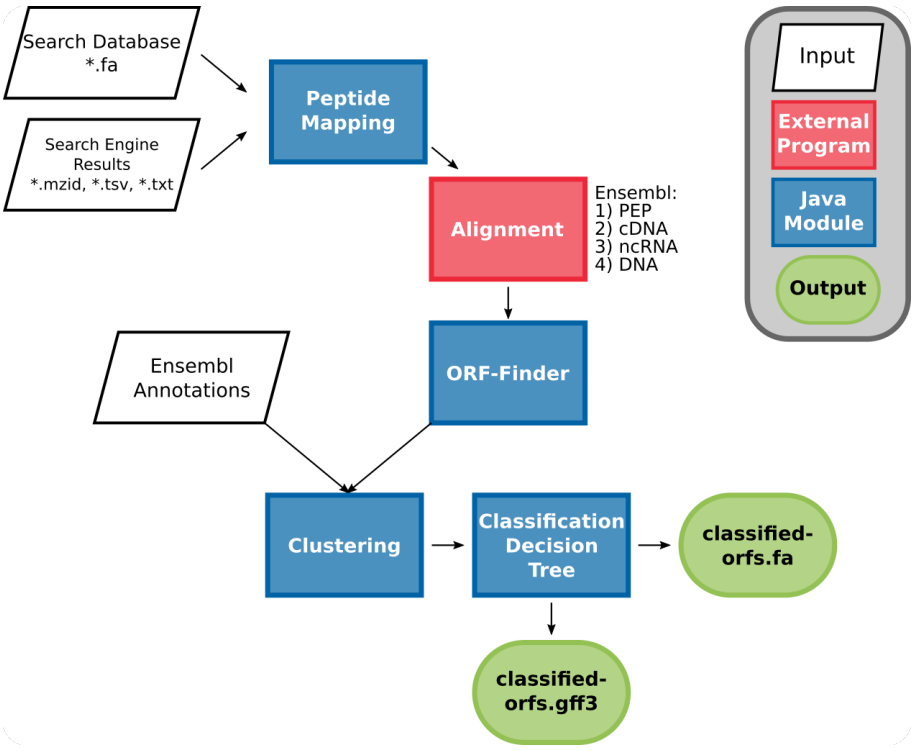


Figure 2.1: The pipeline comprises peptide mapping, alignment, ORF finding, clustering and classification.

2.1.1 Input

The output of MS-GF+ is in the Proteomics Standards Initiative's mzIdentML format [62], which can also be converted with the MzidToTsvConverter that comes with the software to a tab separated output format. MS-GF+ outputs one mzIdentML file per spectral data file. MaxQuant's default output format is always a single tab separated text file (called peptides.txt), no matter how many spectral data files were inputted. MaxQuant could also output the tab separated standard from the Proteomics Standards Initiative, the mzTab format [63].

Because all database search tools output a list of PSMs with differences in scoring schemes, statistical parameters (e.g MS-GF+ outputs e values for single PSM which is missing in MaxQuant's output) vary between tools. A general proteomics "lexicon" was established that includes an overlap of values (peptide sequence, score) and search tool independent variables (e values, posterior error probabilities of PSMs, q values, etc.).

The file parsing was implemented as a factory pattern, meaning that single database search tool output parsers implement methods from an interface. A Factory object calls the input reader method from the correct parser without having to specify the specific class. This is very useful if one would want to add a new reader from a different search tool.

2.1.2 Peptide Mapping

A Java implementation of the Aho-Corasick algorithm [64] was modified to map the peptide sequences against the provided sequence database [65].

The Aho-Corasick algorithm takes as input a set of words and creates a trie out of them. Vertices in this tree store their transition states, meaning that a new pattern matching can start immediately after a failed state or that a word has been found in a text. Given a text of length n and m number of characters in all words then the complexity of this algorithm is $O(n + m + z)$, where z is the number of occurrences of words in text. This is a big improvement from a single pattern string matching algorithm (e.g the Boyer-Moore algorithm or the Knuth-Morris-Pratt algorithm) which need to be run for every pattern independently. In proteomics or proteogenomics we may easily end up with 100 thousands to 1 million PSMs and sequence databases with 40 million amino acids, which would result in unfeasible running times for these algorithms. A description of the algorithm after reference [66]:

1. Preprocessing: Before string matching an automaton is build from all peptide sequences, which has three functions:

    - Go To: Next state for current character from the edges of the trie.

- Failure: When the current character doesn't have an edge in the trie, this function stores all edges that follow that state.
- Match: Stores indexes of all peptide sequences that end at the current state.

2. Matching: Traverses an inputted text over the automaton for string matching.

The output of the mapping results in unique peptides (peptide sequences mapping to only one entry of the sequence database) and in shared peptides (peptide sequences mapping to multiple entries, explained because of sequence homology between entries and redundancies in sequence databases). In the pipeline peptide sequences are matches twice:

1. Ensembl PEP: In the first mapping the sequences get mapped against the reference protein sequence database from Ensembl since we are not interested in these annotated and known peptides.

2. Provided sequence database: Peptides that mapped against Ensembl PEP are excluded to ensure that the resulting mappings are on one hand from novel peptides and on the other hand against customized sequence entries.

2.1.3 Alignment

Peptides from the most common cleavage enzyme(s) are usually short. A tryptic digestion of a protein sample for example results in detectable peptides between 6 - 30 amino acid residues. Therefore it was chosen to use the sequence database entries from mapped peptides for alignments because they are longer and might lead to less ambiguous mappings. Alignments against Ensembl PEP, cDNA and ncRNA were performed using BLAST [67] and alignments against Ensembl DNA were performed with exonerate [68], to retrieve splice peptides. Entries were aligned in following chronology:

1. Ensembl PEP all

2. Ensembl PEP abinitio

3. Ensembl cdna

4. Ensembl cdna abinitio

5. Ensembl ncrna

6. Ensembl dna

Aligned entries against Ensembl PEP and PEP abinitio were removed, remaining sequence database entries were retrieved with samtools faidx version 1.9 [69]. Ensembl PEP abinitio and Ensembl cdna abinitio may not be present in all organisms (e.g both are missing for the plant *Medicago truncatula*). To accept an alignment mapped peptide coordinates needed to overlap the aligned query sequence and an identity cutoff of 95% had to be exceeded. After alignment against DNA with exonerate a last alignment with all novel peptides is performed against Ensembl DNA to determine if these peptides are in a genome unique locus. In this final step no mismatches or gaps were allowed and the full peptide sequence had to be aligned.

BLAST

Sequence similarity search tools like BLAST (BLAST+ 2.8.1) search query sequences against a target sequence database and retrieve the best matches against them. Compared to the optimal solution given by the Smith-Waterman algorithm [70] BLAST utilizes heuristics to reduce runtime.

The first step in its algorithm is the filtering of low complexity regions. In the second step so called seeds are created, where all k-mers for each query sequence get retrieved (with the usual choice of k=3 for protein sequences and k=11 for DNA sequence). An efficient search tree gets constructed from this sub-sequences and the target sequence database is searched with this tree. The exact string matches get extended in both directions until its generated scores decrease. Only high-scoring segment pairs (HSPs) above a certain threshold will be reported.

Exonerate

Even though BLAST is capable of gapped alignments it is difficult to apply its algorithm on more complex models such as exon - intron structures [68]. Exonerate works similar to BLAST as it also generates seeds, extends seeds to HSPs which are finally joined together to form the alignments.

Exonerate tries to avoid as much dynamic programming as possible in the alignments. It applies different strategies to build an alignment from HSPs, which can be seen in Figure 2.2. A portal describes a set of states of a set of HSPs, a span describes a looping state for large gapped alignment features such as introns and a SAR (Sub-Alignment region) refers to a rectangular region of an HSP to which dynamic programming is applied. Upper bounds are generated for each SAR to avoid dynamic programming in each of them. Last but not least multiple models are included, ranging from simple ungapped local alignments to global alignments, from cDNA to DNA alignments to Protein to DNA alignments. The protein2genome model incorporates and outputs frameshifts, introns, splice sites and exons.

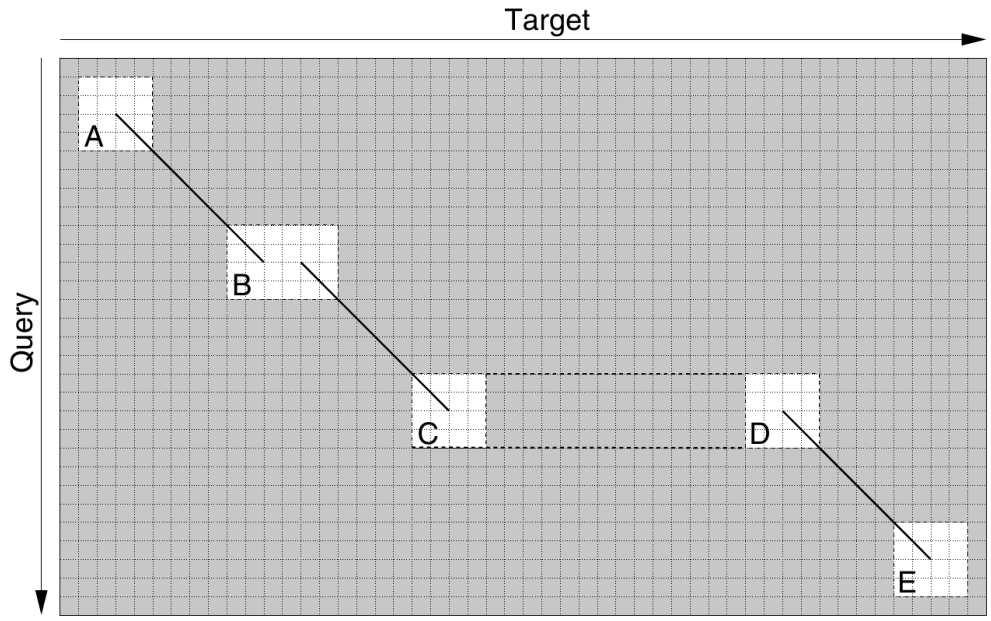


Figure 2.2: Image taken from Reference [68]. (A) is the start of a HSP, (B) is an adjacent HSP joined with (A), (C) and (D) are joined via a span over a larger gap and (E) is the end of an HSP.

2.1.4 Execution

Exonerate version 2.4.0 was used.

Listing 2.1: Exonerate command

```
exonerate --geneseed 250 --seedrepeat 4 --fsmmemory 2000 \
-m protein2genome --score 500 --maxintron "${THRESHOLD}" \
--bestn 1 --showalignment no --showtargetgff yes --verbose 0 \
--showvulgar no -q "${QUERY}" -t "${TARGET}" > "${OUT}"
```

The *geneseed* parameter provides speedup for gapped alignments as it excludes gapped alignment computation where no HSP is scored above this threshold. The *seedrepeat* parameter is the minimum number of HSPs on the same diagonal or reading frame that must be found for an HSP to be extended. The memory for the finite state machine (for the exact string matching in the seeding step) is increased to 2GBs in the parameter *fsmmemory*. All this parameters have been modified to increase the runtime of the program, while the other parameters were all set to output a gff - file. The online manual of exonerate [71] explains the used parameters in more detail.

2.1.5 ORF Finding

The alignments of protein sequence entries or peptide sequences results in genomic coordinates of the identified peptides. The final output should be the ORFs and their

corresponding peptides. Because of this reason and to alleviate a simpler clustering (see section 2.1.7, with less elements to cluster), the coordinates from the ORF have to be determined to put each novel peptide into the interval of the coordinates of an ORF in the correct frame. For this purpose a class was created that translates each codon from all 6 reading frames and saves the coordinates from stop codons, as its easier to define an ORF from stop to stop [72]. If there is a constructed list of ORF coordinates from all six reading frames a binary search could find the correct ORF coordinates for a peptide. The correct frame is calculated by $startRegion \ mod \ 3$. I thought about three ways in what data structure to store these ORFs:

Java ORF class

Listing 2.2: ORF class

```
public class Orf {
    private int startRegion;
    private int stopRegion;
}
```

The most intuitive solution would be to create a class to store a start and an end coordinate. Six lists are created per chromosome that store objects of the ORF class according to their frame. Novel peptides are also stored according to their frames. Each of this peptides is searched against the correct list, an ORF is found when following condition is met:

Listing 2.3: Interval boolean

```
return peptide.getStart() ≥ orf.getStart() &&
    peptide.getEnd() ≤ orf.getEnd();
```

If we consider that a Java object has typically 16 bytes overhead [73] and that the further memory usage consists of the fields of the objects and a padding that the memory usage is a multiple of 8. That means that an ORF object with two integer instance variables (32bit or 4 byte per integer) has approximately 24 bytes.

Lets assume the human genome has 3 billion nucleotides and 1 out of 20 codons are stop codons (3 out of 64). A six frame translation would then result in 300 million ORFs ($\frac{3 \cdot 10^9}{3} \cdot 6 \cdot \frac{1}{20}$, nucleotides divided by codons, multiplied by frames and multiplied by stop codon frequency). The creation of 300 million ORF objects would have a memory usage of 7.2 GB ($300000000 \cdot 24/(1000^3)$).

In practice I would not use an unmasked genome, not store every single ORF, e.g especially if they are below 6 amino acids which is the default minimum peptide length

in most database search tools. However, even if this points would reduce the amount of ORFs to some extend, I would still need to store multiple GBs in memory.

Java ORF bitshift

There is a very interesting way to reduce the memory usage without the creation of any objects. Basically all I need are two integer variables and two integers (32 bit) can be stored by one long (64 bit). With bitshift operations (listing 2.4) the two original integers can be retrieved. Lets declare start the start coordinate of an ORF and end the end coordinate. Then these two integers can be stored in a long variable in the following way:

Listing 2.4: Two Integers in one Long

```
long l = (((long)start) << 32) | (end & 0xffffffffL);
int start = (int)(l >> 32);
int end = (int)l;
```

In the first step start is converted to long and shifted 32 bits to the left such that start is represented by the top 32 bits of long l. The variable end gets converted to a long with a bitwise AND operation with the magic number 0xffffffffL which converts a signed int to a long (end is stored in the bottom 32 bits). This is the Java solution to bitshifting, because there exist no unsigned primitive types. Finally, a bitwise OR combines the top 32 bits with the bottom 32 bits to a single long type. The start coordinate can then be retrieved by shifting all bits 32 positions to the right and casting to integer, while the end coordinate can be retrieved by just casting to integer.

While an ORF object needed 24 bytes of memory, a long variable is stored with 8 bytes. If we repeat the above calculation with the same assumptions I end up with a memory usage of 2.4 GB, three times less the amount of the object oriented solution.

ORF database

A solution that does not require to keep data in memory in general is to store all ORF coordinates in a SQLite database [74]. SQLite is a small SQL database engine that doesn't require any root permissions and is therefore perfectly suited to be included in a portable software.

The workflow to save ORF coordinates is the following, from all reading frames of all chromosomes the stop codon coordinates are detected and written to temporary files, which are read in again and the last two read in coordinates are then combined to an ORF with a start and an end region. A table is created for each frame of each chromosome. Fast data retrieval is granted by an index for a table. The fastest retrieval is granted by a R-Tree index structure that is used for indexing multi-dimensional data, such as coordinates. While the construction of these indexes takes a bit longer (approximately

45 minutes for the human genome), updates on assemblies are not that regularly and that time is well spend considering the retrieval speed of a query. This solution was implemented in Peptidome-DB.

2.1.6 Ensembl annotations

Ensembl hosts public SQL servers that can be used to retrieve annotations [75]. The documentation for the Ensembl core database schema can be found online [76]. This is a straightforward and convenient way to access the huge data sets without having to download large annotation files. To get all exon and gene coordinates from any organisms following query is used:

Listing 2.5: Ensembl SQL query

```
SELECT b.stable_id AS gene_id, c.stable_id AS transcript_id, \
e.stable_id AS exon_id, e.seq_region_strand,f.name, \
e.seq_region_start, e.seq_region_end,e.phase, e.end_phase, \
b.biotype FROM gene b INNER JOIN transcript c ON \
b.gene_id = c.gene_idINNER JOIN exon_transcript d ON \
c.transcript_id=d.transcript_id INNER JOIN exon e ON d.exon_id\
= e.exon_id INNER JOIN seq_region f ON \
e.seq_region_id = f.seq_region_id WHERE \
f.coord_system_id=(SELECT coord_system_id FROM coord_system \
WHERE coord_system.rank=(SELECT MIN(coord_system.rank) \
FROM coord_system)) ORDER BY f.name asc,e.seq_region_strand \
desc, e.seq_region_start;
```

2.1.7 Single Linkage Clustering (SLC)

The previous ORF finding step resulted in a list of ORFs from novel peptides. These ORFs need to be compared to annotated exons for further classification into intragenic or intergenic events. Therefore a hierarchical clustering is applied that clusters ORFs and exons from the same strand of the same chromosomes in distance. This has been defined in the field to refine existing gene models and estimate novel gene models [5][77][59].

In brief, coding elements a and b get clustered together with their next nearest neighbors if they are below a distance threshold T, $distance(a, b) \leq T$. Different methods have been proposed to determine T, while some authors [77][59] chose to look at the intron length distribution (they used the 95th or the 96th percentile), Castellana [5] did not disclose how T was calculated. To avoid misclustering, threshold T should best be larger than the largest intron from the intragenic distance distributions or intron length distributions (ILD) and shorter than the smallest distance between two genes (intergenic distance distribution (IGD). Since these distributions overlap the Kolmogorov–Smirnov

test is used to determine the maximal distance between the cumulative distribution function of the ILD (F_{ILD}) and the IGD (F_{IGD}):

$$T = x(\sup_x |F_{ILD}(x) - F_{IGD}(x)|)$$

The threshold T is the x-intercept of the largest Kolmogorov–Smirnov distance.

Any kind of fixed threshold clustering will lead to misclustering and diverse ILDs [78] make the distance between coding elements an incomplete feature to assure a correct clustering. However, this is not a great concern because the SLC is most important in clustering together intragenic ORFs for the refinement of annotated gene models. In the subsequent classification (section 2.1.8) a binary search against all gene coordinates with novel peptides corrects for misclustering. For organisms were further annotation effort is needed an intergenic cluster could be interpreted as a novel gene model.

A Java implementation of the SLC was adapted from reference [79].

2.1.8 Classification

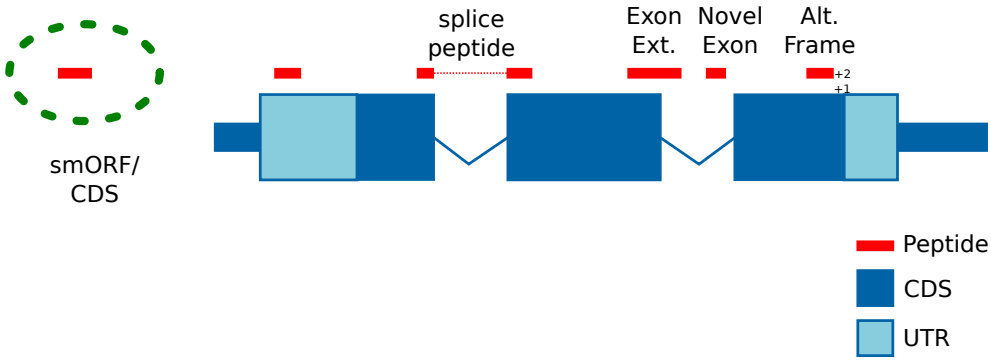


Figure 2.3: Different classes that are determined by the pipeline. The classification is divided into intragenic and intergenic events.

The output of Peptidome-DB includes all identified novel peptides with a classification that either refines known gene models or labels new intergenic coding sequences (Figure 2.3). Intragenic classifications include exon boundary corrections, novel exons, splice peptides, alternative frames, translated untranslated region (UTR) exons and exons. An already known exon is interesting if its biotype is not labeled as protein coding (pseudogenes or non-coding RNAs) or it its start phase is -1 which means it is an expressed UTR exon.

There are two intergenic classification events, smORF if the length of the open reading frame of novel peptide(s) is below 301 nucleotides and CDS otherwise. As previously mentioned, peptides classified as CDS or smORF are searched against gene

coordinates from the same strand to avoid mis-classifications from mis-clustering. The implementation follows the strategy design pattern. A decision tree class evaluates the ORF properties at runtime (figure 2.4) and chooses the next paths for the final classification.

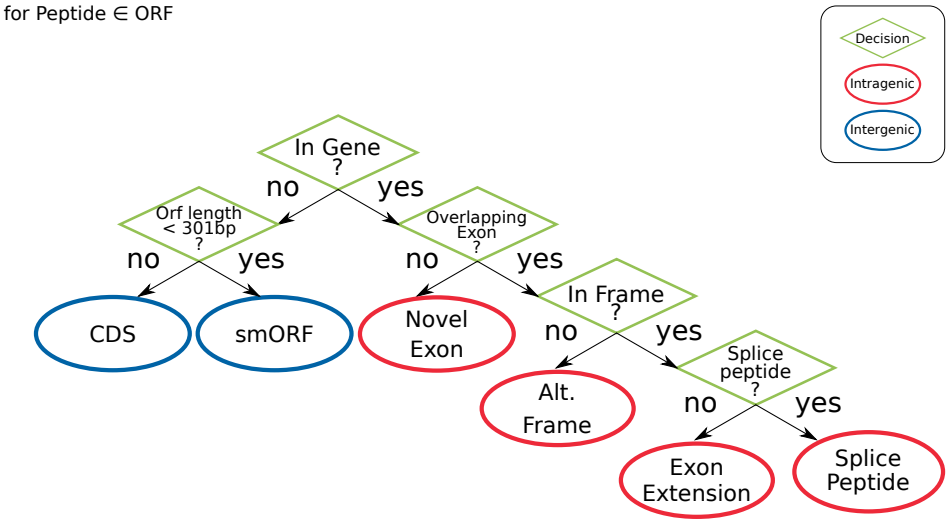


Figure 2.4: Decision tree of the proteogenomic classification.

2.1.9 Physicochemical properties

The most important physicochemical properties of a peptide are calculated by the pipeline, as they might be of relevance in sample separation (isoelectric point, GRAVY value) prior to tandem MS analysis.

GRAVY (Grand Average of Hydropathy) value

The GRAVY value of a peptide $P = p_1, ..., p_n$ comprising n residues is computed by the sum of all hydrophobicity values of amino acid residues divided by the length of the peptide, $\frac{1}{n} \cdot \sum_{i=1}^{n} \text{hydrophobicity}(p_i)$. Hydrophobicity values were taken from reference [80].

Isoelectric point (pI)

The isoelectric point depends on 7 charged amino acid residues in the peptide, the positively charged residues arginine, lysine and histidine, the negatively charged residues aspartic acid, cysteine, glutamic acid and tyrosine. Furthermore the NH_2 - and the COOH - terminal side need to be considered. From the Henderson-Hasselbach equation [81] formulas can be derived to calculate the net charge of a peptide at a given pH. For

negatively charged residues the formula is:

$$\sum_{i=1}^{n} \frac{-1}{1 + 10^{pK_n - pH}}$$

, where pK_n is the acid dissociation constant for a negative charged residue and for positively charged residues:

$$\sum_{i=1}^{n} \frac{1}{1 + 10^{pH - pK_p}}$$

, where pK_p is the acid dissociation constant for a positive charged residue.

To calculate the isoelectric point a bisection algorithm is used to find the solution in an efficient way. Thereby the space of potential solutions is always halved in each iteration until a solution in a predefined threshold is found, in the case of the isoelectric point this threshold is accurate on two decimal places.

The method *float calcNetCharge(float pH)* in the code below returns the computed net charge at a given pH value of a peptide. The amino acid residues of the peptide are stored in a member variable of a physicochemical property calculating class and can be accessed by this method.

Listing 2.6: Bisection algorithm

```
public float isoElectricPoint (float d) {
/*
 * Bisection method to compute the iso-electric point.
 */
        float eps = 0.01f; // convergence criteria
        float upperBound = 14.0f; // max pH
        float lowerBound = 0.0f; // min pH
        float tmp = 0f;
        int it = 0;

        while (it < 1000) { // while(true)
                it++;
                if (Math.abs(upperBound - d) < eps &&
                  Math.abs(d - lowerBound) < eps) {
                        return d;
                }
                float charge = calcNetCharge( d);
                if (charge < 0.0) {
                        tmp = d;
                        d = d - (d-lowerBound)/2;
                        upperBound = tmp;
                }
                else {
                        temp = d;
                        d = d + (upperBound - d)/2;
                        lowerBound = tmp;
                }
        }
        return d;
}
```

Table 2.1: Physicochemical properties of amino acids, values taken from [45] and [82].

Amino acid		Mol. weight	Hydrophobicity index	pI	Occurence in protein (%)
Alanine	A	89.094	1.8	6.01	8.76
Arginine	R	174.203	−4.5	10.76	5.78
Asparagine	N	132.119	−3.5	5.41	3.93
Aspartic acid	D	133.104	−3.5	2.77	5.49
Cysteine	C	121.154	2.5	5.07	1.38
Glutamine	Q	146.146	−3.5	5.65	3.9
Glutamic acid	E	147.131	−3.5	3.22	6.32
Glycine	G	75.067	−0.4	5.97	7.03
Histidine	H	155.156	−3.2	7.59	2.26
Isoleucine	I	131.175	4.5	6.02	5.49
Leucine	L	131.175	3.8	5.98	9.68
Lysine	K	146.189	−3.9	9.74	5.19
Methionine	M	149.208	1.9	5.74	2.32
Phenylalanine	F	165.192	2.8	5.48	3.97
Proline	P	115.132	−1.6	6.48	5.02
Serine	S	105.093	−0.8	5.68	7.14
Threonine	T	119.119	−0.7	5.87	5.53
Tryptophan	W	204.228	−0.9	5.89	1.25
Tyrosine	Y	181.191	−1.3	5.66	2.91
Valine	V	117.148	4.2	5.97	6.73

The **hydrophobicity index** is taken from Kyte and Doolittle [80], it determines the *hydrophobic effect*, which is the 'observed tendency of nonpolar substances to aggregate in an aqueous solution and exclude water molecules' [83]. Negative values indicate hydrophilic behavior, values around 0 indicate a neutral behavior and positive values indicate a hydrophobic behavior.

The **isoelectric point** is the pH value where a molecule is not carrying a net electric charge.

The **occurence in protein (%)** is taken from reference [84] and is calculated as the average occurrence in percent from available proteomes of eukaryota, archaea, bacteria and viruses.

2.1.10 Peptidome-DB output

Classified novel peptides are outputted in a tab separated file with their phyiscochemical properties and database search tool properties (e.g experimental m/z value or MS/MS scan number). Their corresponding ORFs are retrieved from Ensembl chromosome files according to their genomic loci with Emboss transeq [85] version 6.6.0.0. Coordinates from ORFs and peptide are also outputted in a gff3 file.

2.2 MS-GF+

The database search tool MS-GF+ provides an universal and sensitive solution to analyze tandem mass spectra [60]. It can process various spectral types from different fragmentation methods, instruments, protocols or cleavage enzymes and provides a robust probabilistic model that works for many data sets. On one hand MS-GF+ generating function approach [86] offers a way to assign statistical significance to peptide identifications without the use of a decoy database, on the other hand MS-GF+ works well with statistical modeling tools for post-processing, such as Percolator [87].
MS-GF+ takes as input a spectral data set and a protein sequence database, its workflow includes the generation of spectral vectors, the searching of spectra against the sequence database, the scoring of PSMs and the computation of e values and the estimation of FDRs.

2.2.1 Scoring

Database search tools provide a scoring function for a PSM, that in the best case should assign the best score to that peptide P that generated spectrum S. These scores can then be used to establish statistical significance of individual PSMs. The construction of spectral vectors and peptide vectors allows MS-GF+ an inexpensive and yet accurate way to compute scores, namely the dot product of these two vectors. The correctness of this simple calculation is given by the probabilistic information in the spectral vector.

$$MSGFScore(P,S) = \mathbf{P}\cdot\mathbf{S} = \sum_{i=1}^{k} p_i \cdot s_i \text{ if } Mass(P) = PrecursorMass(S) \text{ and } -\infty \text{ otherwise.}$$

2.2.2 Database search

The database search compares a spectral data set $Spectra$ and an extended sequence database $ProteinDB^{+}$, the set of all variants allowing modifications, and finds for each spectra $S \in Spectra$ a variant $PV_{S,ProteinDB}$ such that

$$PV_{S,ProteinDB} = \underset{PV \in ProteinDB^{+}}{\operatorname{argmax}} MSGFScore(PV,S).$$

To perform this search efficiently all generated spectral vectors are kept in memory and a suffix array is constructed for peptides. Instead of searching peptides according to their appearance in $ProteinDB$ they are searched according to their ordering in the suffix array. And since most protein databases contain similar proteins, resulting in redundant peptide sequences, the Longest Common Prefix (LCP) data structure is used to score each non-redundant peptide only once.

Execution

MS-GF+ version v2019.04.18 was used. The issued command is displayed here:

Listing 2.7: MS-GF+ command

```
msgfplus -d "${DBNAME}" -s "${MGFDIR}" -t "${PCMTOL}" \
-inst "${INST}" -maxMissedCleavages 2 -m "${FRAGMETHOD}" \
 -mod mods.mod -tda 0 -addFeatures 1
```

The variable $DBNAME is set to the path of the sequence database, $MGFDIR to the directory of the spectra data files, $PCMTOL is the precursor mass tolerance, $INST is the specified instrument and $FRAGMETHOD is the parameter for the used fragmentation method. mods.mod is a text-file that contains all variable and fixed modifications. Each search was performed with the option -tda set to 1, telling MS-GF+ to perform Target decoy approach and tda set to 0 where only a search against a target database is performed. The flag -addFeatures was set to 1, this includes feature like Energy in the MS-GF+ output which are needed as features by the post-processing tool Percolator.

2.3 Handling MS data: ProteoWizard

MS data is usually stored in vendor specific raw file formats. Most database search tools cannot access these proprietary formats, for this reason the raw files need to be converted to an open format that can be used by a database search tool. The ProteoWizard toolkit [88] is an open-source framework that comprises various libraries for MS-based proteomic data analysis. The tool msConvert (version 3.0) integrates multiple vendor dependent libraries to access raw data. It was used to convert all raw files used for database searches with MS-GF+ for the application of Peptidome-DB in this thesis. Following command was used for conversion:

Listing 2.8: msConvert command

```
MSConvert.exe DatasetName.raw --filter "peakPicking true 1-" --mgf --32
```

At the time when this work was done the vendor converter (only ThermoFisher raw files were used as input) in the tool msConvert was limited to the Windows operating system. The filter option `peakPicking true 1-` centroids peaks at the MS1 level. Centroiding converts the shape of mass peaks to a weighted average, ideally to produce symmetrical and smoothed peaks at their respective local maxima. The `--mgf` flag is set for the open mgf (Mascot Generic Format) [89] output format and the `--32` flag tells the tool to use 32 bit floating point numbers.

2.4 Search database construction

A six frame translation of the human genome (GRCh38 assembly from Ensembl, v2018.06) was split into 79 fasta files with an average file size of 139 MB and an average database size of 68.891.533 amino acids (1.67 times the database size of Ensembl PEP, with 41.2 million amino acids (aa) and 70MB in file size). These sizes were chosen that all searches with them would have low runtime with a high search space (Figure 2.5). The search space in an indivual search is neglectable because these searches were run without a target decoy approach and MS-GF+ provides various features in its output that are independent from the search space, e.g RawScore.

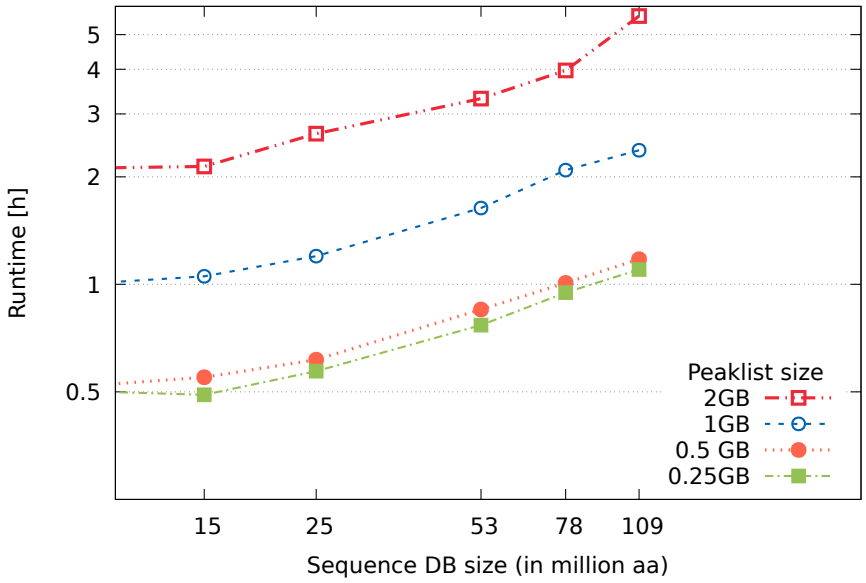


Figure 2.5: Benchmark of MS-GF+ with different search space and peaklist file sizes which were chosen according to the file sizes from the peaklist files of the projects.

2.5 First Search

The basic idea behind the first search is to use the search results, with the six-frame translation of the human genome as sequence database, as input for the pipeline which processes these results and outputs a multi fasta file [90]. The entries of the multi fasta file are the translated ORFs of their corresponding novel peptides. Since most PSMs from a search against a six-frame translation will be of low quality (low scores and high spectral e values) a filtering criteria to reduce the size inflation of the final concatenated database had to be chosen. This criteria should be independent of database size and was found in the MS-GF+ spectral e value (SpecEValue). A proteomic search against Swiss-Prot resulted in Figure 2.6 and led to the filtering criteria of minimum spectral e value of 1e-08 for a PSM.

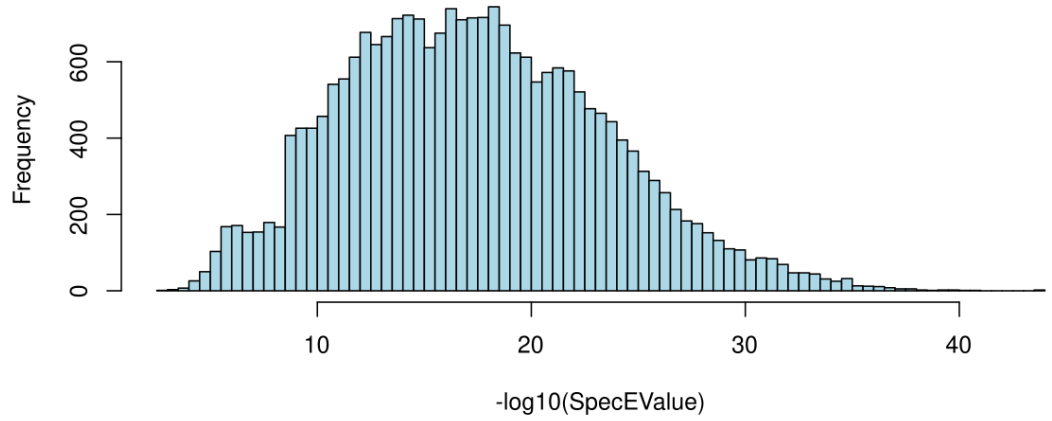


Figure 2.6: Histogram of spectral e-values of PSMs (q-value $< 1\%$). The median is 9.2566e-18. A specEValue cutoff of 1e-08 contains over 97% of PSMs from a search against Swiss-Prot.

2.6 Percolator

Post-processing search engine results helps in better discrimination between correct and incorrect identifications in proteomics. The standard approach to set confidence in PSMs is to control the False Positive Rate (FDR), the rate of False Positives among the set of all positive identifications. Machine – learning based methods such as Percolator use various search engine specific features to re-evaluate q values of individual PSMs [91].

The necessary input are the results from two searches, one against a target database and one against a decoy database. A semi – supervised SVM learning approach takes selected features from the database search tool to distinguish between the positive and negative set. Percolator outputs a list of PSMs with their sequences, q values, PEP (Posterior Error Probability) and p values.

Execution

Percolator version 3.2.1 was used. It works with meta files which contain all mzid - files from a MS-GF+ search, denoted as $TARGET_META and $DECOY_META below. The variable $PIN is the file name of the converted pin (percolator-in) format.

Listing 2.9: Percolator command

```
msgf2pin -o "${PIN}" -P "XXX_" "${TARGET_META}" "${DECOY_META}"
percolator -j "${PIN}"-N 500000 -r "${TARGET_OUT}" -B "${DECOY_OUT}"
```

Decoy fasta files were generated by DecoyPYrat [92], a non-redundant decoy sequence generator for large scale proteomics. The check for redundancy among target and decoy peptide sequences is especially useful for larger search spaces, DecoyPYrat was executed using following command (the -l flag is the minimum peptide length):

Listing 2.10: DecoyPYrat command

```
python3 decoyPYrat.py -l 6 "${FASTA}" -o "${OUT}"
```

2.7 PhyloCSF

PhyloCSF helps in distinguishing coding from non-coding sequences. It finds conserved regions by examining evolutionary characteristics such as a higher number of synonymous substitutions in codon substitution frequencies (CSF) from a multiple sequence alignment (MSA) [93].

PhyloCSF track hubs are available online at reference [94] where a PhyoCSF score was calculated for every codon from all 6 frames of the human genome for multiple MSAs. To estimate the coding potential of a genomic locus a script [95] was modified that calculates PhyloCSF scores from genomic loci from a gff file. This solution is only applicable for very few species (human, mouse, chicken, nematode and fruit fly).

2.8 Prosit

Novel peptide identifications can be validated by comparing them to synthetic peptides. Prosit [46] utilizes the publicly available data from ProteomeTools [96] which contains more than 21 million MS/MS spectra from more than 576.000 synthetic peptides with different sequences, modifications and charge states which were analyzed using different fragmentation techniques. This data comprises 98.5% of human protein coding genes and 63% of SwissProt annotated isoforms.

Prosits neural network architecture is split into two models: an encoder which learns a feature representation for the given input (peptide sequence, charge and normalized collision energy used during fragmentation). And a decoder which learns the fragment ion intensities from the feature representation of the encoder. Apart from fragment ion intensities Prosit is also able to predict retention times of eluting peptides. The online version [97] of Prosit was used.

2.9 Basic metrics

For the evaluation of the clustering and classification these metrics were used:

Precision The precision (sometimes also referred to as positive predictive value) is the fraction of the total number of relevant documents among retrieved documents.

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall The recall (or sensitivity) is the fraction of of the total number of relevant documents that were successfully retrieved.

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Purity Purity of a clustering is calculated by assigning each cluster to the class of its most frequent member class. The number of correctly identified members in each cluster are summed and this sum is divided by the total number of documents.

$$Purity = \frac{1}{N} \sum_{i=1}^{k} max_j |c_i \cap t_j|$$

2.10 Website

A web prototype was developed to present the novel findings of this thesis but also to integrate the pipeline into a web service. Therefore a web framework had to be chosen that was easy to use and fast to develop but also reliable and robust. The Django framework (version 2.2.1) written in the Python programming language was selected for this task [98] as it has several advantages over other frameworks; first of all it is very well documented which makes the set up uncomplicated and very fast, the developers claim that a webpage can be brought from "concept to completion" in mere hours. In addition to that community driven open source applications get extended constantly, which allows developers to build their projects on a well tested codebase. Django's Model-View-Template design pattern is loosely based on the Model-View-Controller design pattern, where data is stored in a database in the model and the user can access data with a controller via a view. This design pattern separates the data storage and access on the server side from the input on the client side.

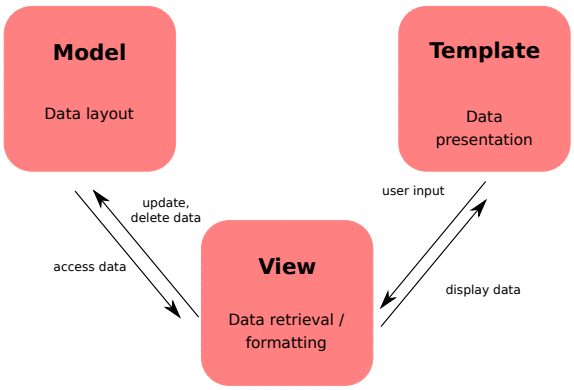


Figure 2.7: The model in django can be seen as an abstracted database layout. Views allow developers to disclose information. Templates are on one hand the presentation and on the other hand transmit requests to views.

An entity relationship model in supplementary Figure S5 shows the interrelated entities that act as a database. In this schema a *Peptide* only contains stable information such as sequence and physicochemical properties. Dynamic information such as which organism the peptide was identified in, genomic coordinates or database search tool specific scores are stored in a *Peptide Properties* entity. The *ORF* entity stores the genomic coordinates of the open reading frame of an identified peptide, the ORF classification, potential biotype (if the ORF is located inside an annotated gene with an Ensembl biotype). The *Experiment* entity contains all data that was used for the database search and experimental layout (e.g which instrument was used). *ORF*s and *Experiment*s are linked together.

Chapter 3

Results

3.1 Simulation of the implemented Clustering

The single-linkage clustering (SLC) is important for the subsequent proteogenomic classification of novel peptides. Therefore a correct clustering ensures a correct classification. SLC as it was first described in a proteogenomic context by Castellana et al. [5] for the plant *Arabidopsis thaliana* was used to link unannotated peptides from the same strand together in distance to distinguish between peptides overlapping protein-coding genes and peptides from novel gene models. Misclustered clusters were split and separated with the help of *ab initio* gene predictions. The distance threshold for the clustering could therefore be chosen from an educated guess because of these corrections.

Marx et al. [59] extended the idea for their proteogenomic study of *Sus scrofa*. The choice for the distance threshold was more elaborate, as they did not use gene predictions to correct for misclustering. The 95th percentile of the intron length distribution was chosen.

These studies could show that a SLC can be used to correctly classify novel peptides into intragenic and intergenic events for a single organism. However, previous studies lacked an evaluation of their clustering and the subsequent classifications and were not generalized for more than one species. On the basis of the results from these publications it can be hypothesized that the SLC implemented in this thesis performs well in clustering and thus distinguishing peptides from intragenic and intergenic regions for further classifications and that the calculation of the distance threshold is an applicable and generic solution for multiple organisms.

To estimate the correctness of the implemented SLC the clustering was tested with annotated exons from Ensembl PEP for two organisms (*Homo sapiens* and *Medicago truncatula*). These organisms were chosen to display different stages of annotation efforts; while the human organism is a very well annotated eukaryotic organism, efforts for the nitrifying model plant are still ongoing, the JCVI assembled version Mt4.0 contains

50,444 protein-coding gene loci, where $\sim$ 31,500 are high confidence genes (based on experimental evidence e.g EST/RNA-seq/protein support) and $\sim$ 19,000 are low confidence genes (based on synteny to related organisms). The calculated distance threshold for *H. sapiens* was 10,977 bp and for *M. truncatula* it was 1,355 bp as can be seen in Figure 3.1.

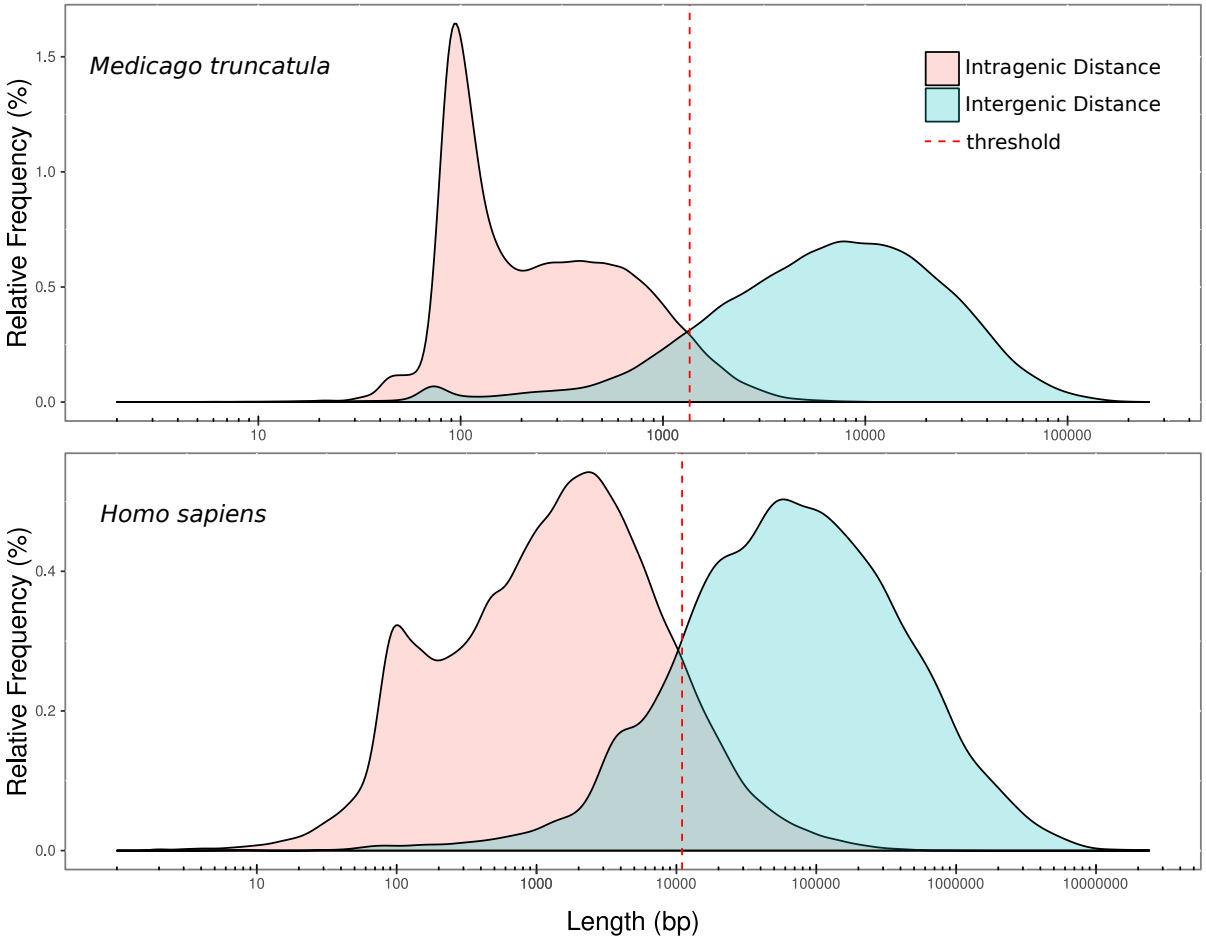


Figure 3.1: The intragenic and intergenic distances of all protein-coding genes were computed and plotted in their densities for *Homo sapiens* and *Medicago truncatula*. The intragenic distribution (intron length distribution [78]) is of particular interest for gene predictions because this distribution needs to be modeled to compute the probabilities of an occurring exon in a gene. The overlap of the two distributions determines the correctness of the clustering; the bigger the overlap the worse the clustering is performed.

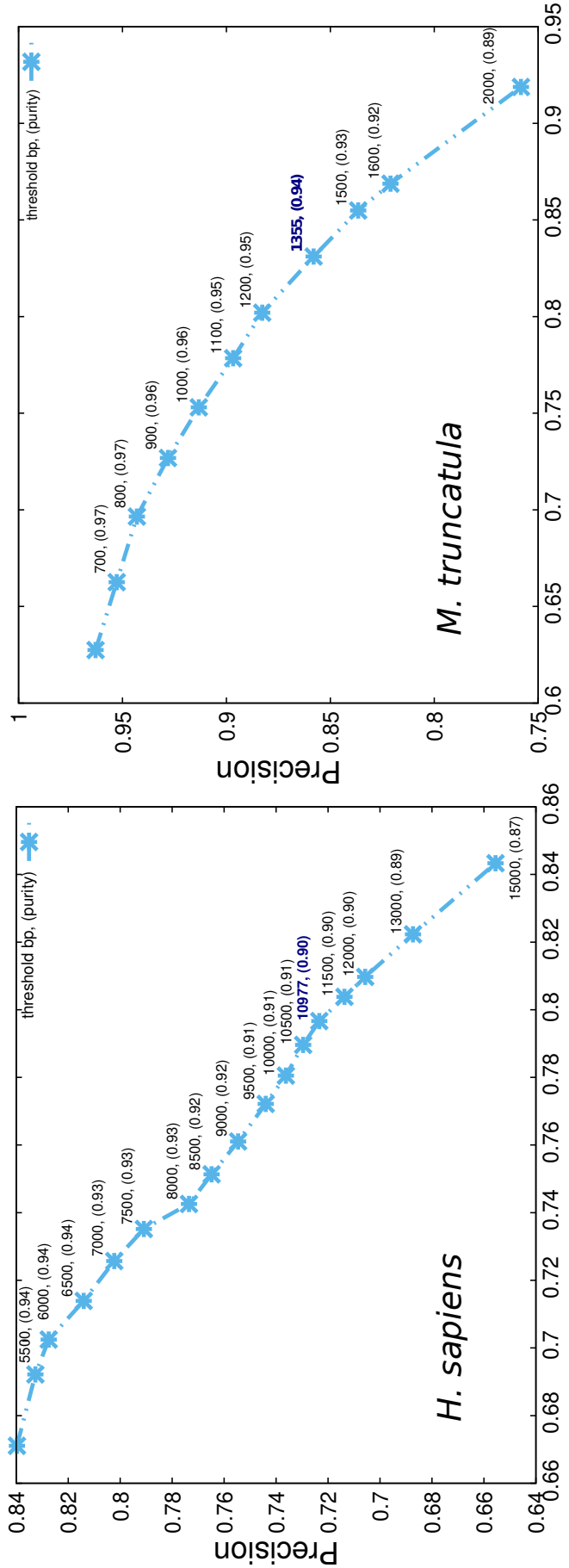


Figure 3.2: The precision and recall of the SLC was calculated for various reasonable thresholds. The lower the threshold for the clustering is chosen the better the purity becomes. In theory a perfect purity could be achieved if each element would be in its own cluster, thus this metric does not accurately determine the quality of a clustering . Lower thresholds splits exons from the same gene into different clusters, while exons from different genes are seldom mixed up. This results in a lower recall (meaning more false negatives) and a higher precision (less false positives).

The proposed method of choosing the distance threshold where the highest Kolmogorov-Smirnov distance between the intragenic distance distribution and intergenic distance distribution occurs leads to an optimal solution. The inclusion of the latter distance distribution is extending the idea from Marx et al. [59], were only the intragenic distance distribution was considered. It is also generally applicable and does not rely on predicted gene models as was described in Castellana et al. [5]. Albeit any fixed length clustering is prone to errors (misclustering) the main objective of the SLC is to separate peptides overlapping known gene models from peptides originating from intergenic loci.

To my knowledge the evaluation of the clustering approach shown in Figure 3.2 is the first of its kind. The results of this precision-recall plot are satisfying for the aforementioned objective of the clustering. Noteworthy are the lower values for precision and recall for the SLC of human exons compared to the SLC of *M. truncatula* exons. One possible explanation for this could be that human genes are better annotated. While the clustering follows the chosen intragenic distance distribution of predicted gene models this distribution seems to become different the more experimental evidence is present for the corroboration of gene models. That is why the precision-recall curve could tend to be smoother for *M. truncatula* than for *H. sapiens*.

3.2 Simulation of the implemented Classification

Different types or classes of proteogenomic peptides are described in references [3][5][59]. These classifications can be separated into intragenic and intergenic events, depending on the clustering - if there is no annotated exon in the cluster the peptide very likely originated from an intergenic locus. In the case of long introns (or if the intron length is bigger than the distance threshold) peptides mapping in between two exons might be mistakenly classified as intergenic. To account for this the genomic coordinates from the peptide are searched in a binary search against the coordinates of all genes from the same strand to avoid a missed overlapping gene. Thus the greatest weakness of the clustering, misclustering due to a fixed distance threshold followed by a false classification is eliminated. On grounds of this implementation I hypothesized that the classification will be accurate in distinguishing intragenic from intergenic events and accurate in calling the correct single event.

To simulate the correctness of the classification with ground truth data a fraction of exons (1%, 3%, 5%, 10% and 15%) were assigned a class (UTR, Alternative Frame, Novel exon, Exon, Exon extension or CDS) from their associated gene. Then they were removed from the set of exons from Ensembl PEP and treated as novel ORFs with an simulated identified peptide spanning the whole exons from its genomic start to end coordinate. The removed exons were then clustered with the remaining (annotated) exons and the percentage of correctly classified ORFs could be determined. This was

repeated 10 times with randomly chosen exons, the results are displayed in Figure 3.3 and Figure 3.4.

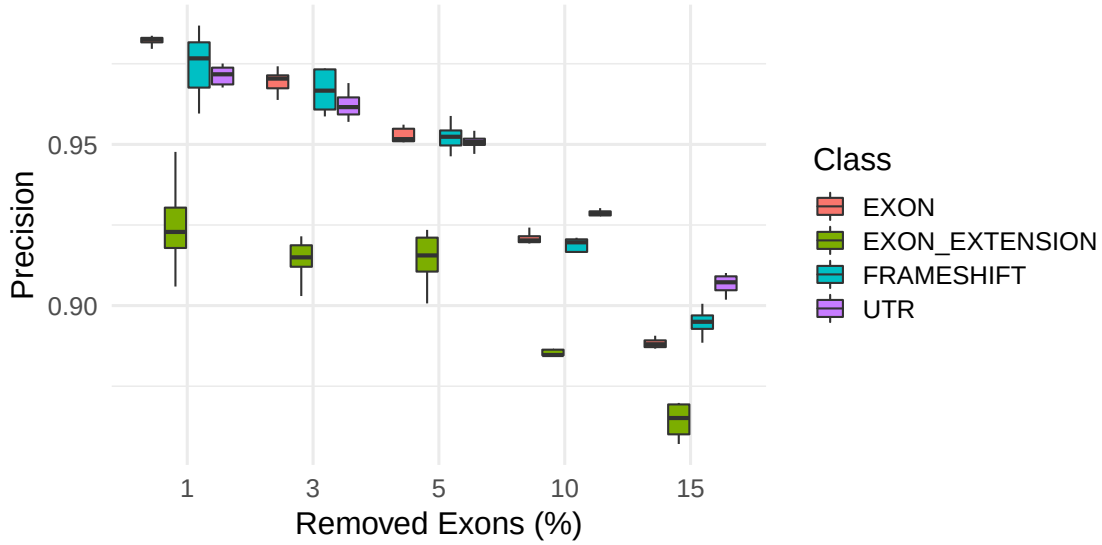


Figure 3.3: Simulation of overlapping classification events. High precision is achieved for all events.

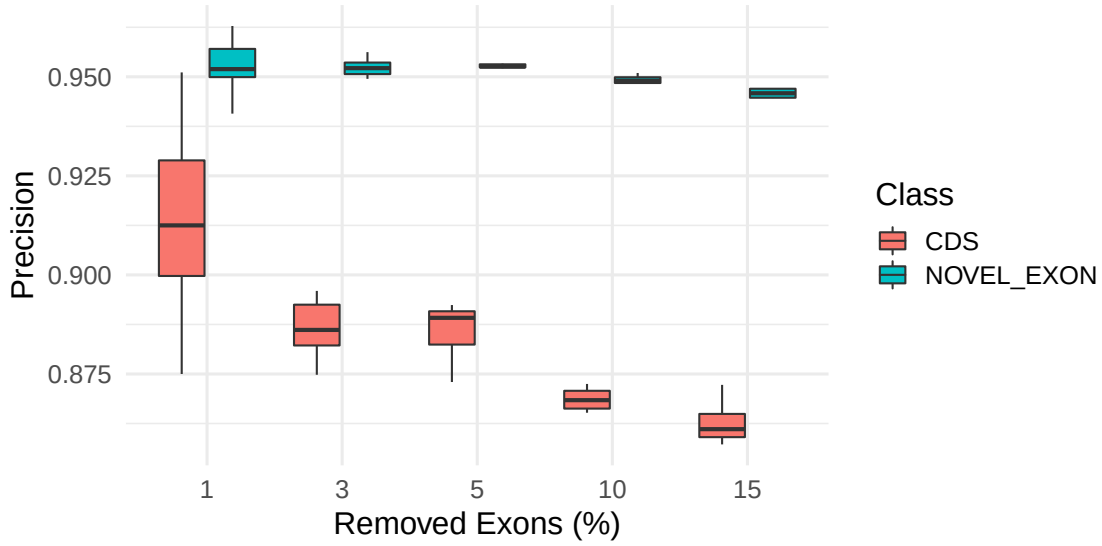


Figure 3.4: Simulation of non-overlapping classification events. The event novel exons has a very stable precision even with greater number of exons removed.

After showing that the clusters from the SLC are sufficiently accurate the subsequent classification and its evaluation provides information on the correctness of this approach.

The four intragenic classification events in Figure 3.3 have in common that the coordinates of an ORF need to overlap the coordinates of an already annotated exon. The correctness of their simulated classification is therefore solely based on the presence

of the annotated exon. It follows that an increased percentage of removed exons leads to more mis-classifications.

The two classification events NOVEL_EXON and CDS in Figure 3.4 have in common that they are not depending on another exon overlapping with the coordinates of the ORF. The class NOVEL_EXON is rarely mis-classified even if a high percentage of exons gets removed. The reason behind this is that the coordinates of every classified CDS are checked if they overlap an already existing gene model to prevent mis-clustering. If there is an overlapping gene the classification of the CDS is wrong and it is changed to a NOVEL_EXON (because it cannot be any other intragenic class as its coordinates would have overlapped with an annotated exon).

Summing up the evaluation of the classification works very well with ground truth data, a high precision was achieved for all classification events, especially for intragenic events that could mostly be used for refining existing gene models.

3.3 Application: Data sets

To test the developed pipeline with real data, data sets from lesser researched tissues and fluids were chosen or data sets with a great sample depth. This was done to facilitate the identification of novel peptides. The following four studies were used in Wilhelm et al. [99], one of the first drafts of the human proteome. Their spectral data sets were uploaded to the PRIDE repository. The original publications did not use a proteogenomic approach and are therefore well suited for the application to Peptidome-DB.

PXD002273 [100]
This study concentrated their efforts on the choroid, a vascular layer of the eye lying between retina and sclera. This study was of particular interest as it provides the first in-depth proteomic inside in its tissue. The samples comprise 24 spectral data files.

PXD007745 [101]
This study conducted MS - based experiments on human regulatory T cells from peripheral blood mononuclear cells (PBMCs). It contained 12 spectral data files.

PXD008934 [102]
Chen et al. had samples of the left ventricle tissue for proteomic analysis. They main focus was on detyrosinated microtubules that can hinder the motion of contracting cardiomyocytes. They had samples from 34 healthy and unhealthy hearts.

PXD009142 [103] The tissue in this study involved corneal endothelial cells which are important in maintaining clarity of the cornea. The samples comprise 24 spectral data files.

3.4 Constructed databases

After showing the correctness and accuracy of the clustering and more importantly the classification real data sets were reanalyzed with Peptidome-DB. What constitutes a proteogenomic study is the usage of a customized protein sequence database for peptide identification. Any database search will miss identifications if the peptide sequence is not present in the database. But with the addition of an increased amount of sequences into the search space sensitivity is lost because the best scoring PSM cannot be guaranteed anymore to be the correct match from a multitude of other similar scoring peptide sequence matches [3]. Consequently proteogenomic studies of human samples utilizing a six-frame translation as sequence database identified 30% less peptides than searches with reference databases, while the number of identified novel peptides remained relatively small [35]. Thus one needs to account for this trade-off between completeness and size inflation of the database. In the exploratory setting I have chosen for the proteomic database searches in this thesis I expect that an approach with a first search filtering out numerous incorrect PSMs can be used to provide the right balance for the aforementioned trade-off.

The first search was used for the construction of the individual databases for each spectral data set of the four different PRIDE projects PXD002273, PXD007745, PXD008934 and PXD009142. The spectral data sets was searched with MS-GF+ against a six-frame translation of the human genome and all ORFs from PSMs over a certain quality threshold (described in section 2.5) were extracted and translated to form the final database. The filtering is crucial to lower the search space and thus increase the sensitivity of peptide identifications. Incidentally the first search with the project specific spectral data created specific databases further reducing database size inflation.

The constructed ORF – databases of the four PRIDE projects differ in sizes (Figure 3.5). The project PXD002273 had the least amount of spectra in its peaklist files and thus the first search of this spectral data set resulted in the smallest database size. The biggest database resulted from the project PXD008934 with a size three times that of Ensembl PEP. These sizes are consistent with the proteoform space [104] and used sequence databases in literature [59][105][106].

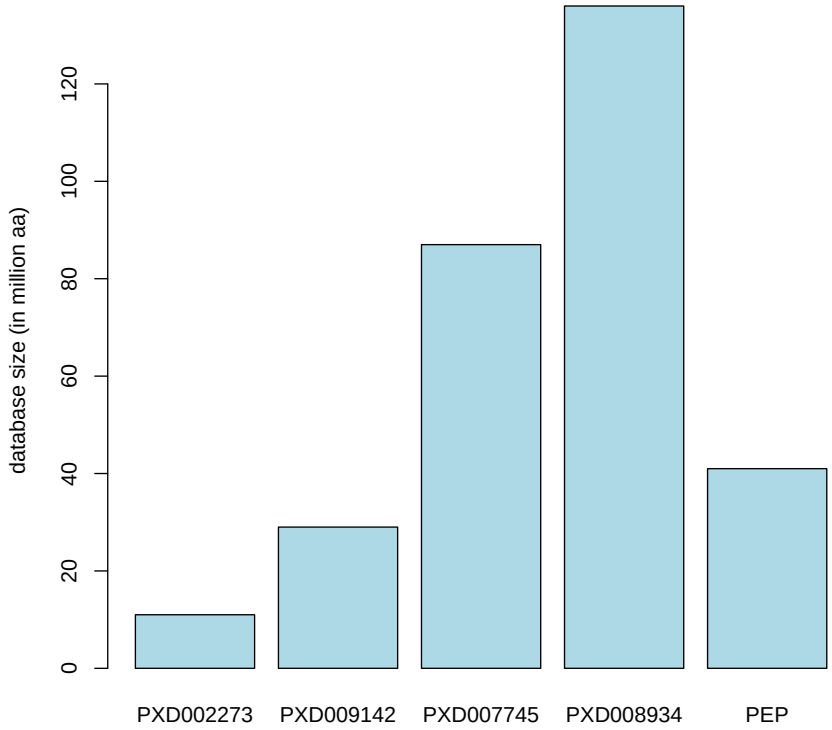


Figure 3.5: Search space sizes compared to Ensembl PEP. PXD002273 contains ∼11 million aa, PXD009142 contains ∼29 million aa, PXD007745 contains ∼86 million aa and PXD008934 contains ∼136 million aa.

3.5 MS-GF+ results

In the next step the database search engine MS-GF+ was used for peptide identification in the four different project samples. Because the four chosen projects were published and uploaded their proteomic search results to the PRIDE repository I can compare their results with mine. Of course this comparison is imprecise as proteomic search engines differ in the algorithms they use for peptide identification and thus their results can vary. However, the numbers in table 3.2 give a rough estimate on a lower boundary for the number of identified peptides. This assumption can be made because on the one hand MS-GF+ outperforms other database search tools in the number of identified peptides [60] and on the other hand the inclusion of a proteogenomic ORF database should increase the number of identified peptides. The same database search parameters were applied to each project as it was described in the respective original publication (table 3.1).

Table 3.1: MS-GF+ search parameters were taken from the original publications.

Project	Instrument	Fragmentation Method	Precursor Mass Tolerance
PXD002273	Q-Exactive	HCD	10ppm
PXD007745	Orbitrap	HCD	10ppm
PXD008934	Q-Exactive	HCD	20ppm
PXD009142	Orbitrap	HCD	20ppm

Table 3.2: Number of identified peptides according to the results in the publication at a FDR of 1%.
* only average identifications was provided in the publication.
** taken from published database search output file, because this information was not provided in the publications.

	PXD002273	PXD007745*	PXD008934**	PXD009142**
#of peptides	26,503	35,744	39,421	144,480

The constructed ORF databases from each project were concatenated with Ensembl PEP (Homo_sapiens.GRCh38.pep.all.fa, release-96) to ensure that spurious PSMs match rather to their corresponding peptides in Ensembl PEP. These concatenated sequence databases were used as input together with the spectral data set that generated them in the first place for MS-GF+.

The numbers from the table 3.3 are from all PSMs which have a q value below 1% from a MS-GF+ target decoy search (hence the FDR is at 1% as it is proposed by the Human Proteome Project Mass Spectrometry Data Interpretation Guidelines [53]). The same database search parameters (precursor mass tolerance and included modifications) from the original publications were used. Furthermore all projects used trypsin as a proteolytic enzyme with up to two missed cleavages allowed. The ratios of unique and shared peptides mapping to Ensembl PEP (20% unique compared to 80% shared) and mapping to the ORF DB (99% unique compared to 1% shared) are constant over all four search results.

The number of total identified peptides is approximately the number of mapped database entries in the ORF DBs from all projects, which means that almost all entries which are declared present in the sample are one hit wonders because of the high number of unique peptides. This is not the case for Ensembl PEP which contains a lot of isoforms, hence the high number of shared peptides, whereas the constructed proteogenomic databases are non-redundant.

Table 3.3: MS-GF+ database search results from the four projects. Reported are the number of identified peptides (unique,shared and total) and the number of mapped search database entries compared to Ensembl PEP and the constructed ORF database (ORF DB).

Project	Unique peptides	Shared peptides	Total peptides	Database entries
PXD002273				
Ensembl PEP	4,745 (19.09%)	20,106 (80.91%)	24,851	30,640
ORF DB	31,806 (98.74%)	406 (1.26%)	32,212	32,611
			57,063	33,251
PXD007745				
Ensembl PEP	6,896 (18.76%)	29,864 (81.24%)	36,760	32,611
ORF DB	2,146 (99.49%)	11 (0.51%)	2,157	2,090
			38,917	34,701
PXD008934				
Ensembl PEP	7,952 (19.17%)	33,527 (80.83%)	41,479	27,375
ORF DB	81,893 (99.32%)	561 (0.68%)	82,454	82,800
			123,933	110,175
PXD009142				
Ensembl PEP	28,042 (19.84%)	113,264 (80.16%)	141,306	68,071
ORF DB	46,516 (99.83%)	81 (0.17%)	46,597	46,373
			187,903	114,444

The total number of identified peptides in table 3.3 are all higher than the published numbers from table 3.2 supporting the aforementioned assumption that a proteogenomic apporach can increase the number of identifications. On the other side the database search results for the projects PXD002273, PXD008934 and PXD009142 seem very unlikely - almost twice as many novel peptides identified compared to the known peptides from the reference database as it is the case for the project PXD008934 shows a high number of false positive identifications. This indicates the importance of post-processing proteomic data to further filter out spurious PSMs.

On another note the very conservative human proteome project data interpretation guidelines asks researchers to provide two or more distinct uniquely-mapping peptide sequences of at least 9 amino acids in length to infer the expression of a protein. However, the detection of more than one peptide in a small ORF encoding bioactive peptide is unlikely compared to longer proteins, hence most identified entries from the ORF databases are "one-hit wonders". Furthermore it has been shown by Gupta and Pevzner [107] that this commonly used "two-peptide" rule compromises the sensitivity-specificity tradeoff in protein identifications. According to their results two spurious, random hits in the decoy database have a higher probability than a single high scoring peptide hit in the target database. Thus this guideline is unfeasible to detect most smORFs and would lead them to get discarded, Gupta and Pevzner argue to abandon the two-peptide rule

for the single-peptide rule in association with control of the FDR. Because of this the "two-peptide" rule was not applied to the database search results.

3.6 Percolator results

The calculation of FDRs relies on scores from incorrect hits against the decoy database to calculate a global error rate for hits against the target database. Post processing tools such as Percolator are able to discriminate better between decoy and target hit distribution than plain search engine scores by including further properties of PSMs into features for a support vector machine (SVM). This makes them a crucial part of proteomic data analysis since they reduce the number of false positives and increase the number of credible assignments.

In Percolators approach target PSMs get split into two groups: correct target hits and incorrect target hits which score similar to decoy hits. In a proteogenomic approach it can be expected that peptides from canonical reference sequence databases that have been well corroborated will score better than most spurious peptides from intergenic regions [105]. Following this separation of target hits into two groups I hypothesize that the number of identified peptides should roughly stay the same for peptides mapping to Ensembl PEP while the number of identified peptides from the ORF database should decrease.

3.6.1 Search results

Percolator outputs a list of sequences from PSMs with recalculated q values and posterior error probabilities. It also writes the obtained weights from the used features of the SVM to stdout, selected features from the third cross-validation split are presented in table 3.4. RawScore is a measure for how well the theoretical and experimental spectra match, lnEvalue is the logarithm of the computed e value and Energy is the difference between RawScore and DeNovoScore (the optimal score a PSM could obtain). The positive weights of all projects in their RawScore (with the exception of PXD008934 where the weight is close to zero) show that a higher RawScore is indicative of obtaining a better score while negative weights in Energy describe that PSMs which RawScore is closer to its DeNovoScore are given better scores. The weight of the feature absdM is also negative in all projects, indicating that a larger difference between observed and calculated mass gives a worse hit. The variable pi_0 gives an estimation of incorrect target PSMs in percent, for example 40.36% of target PSMs are incorrect in the project PXD002273.

Table 3.4: Weights of features after the third cross-validation split per data set.

Project	RawScore	Energy	lnEvalue	absdM	pi_0
PXD002273	0.8710	−1.3603	3.7030	−0.0142	0.4036
PXD007745	0.2578	−1.6904	0.4526	−0.8685	0.6222
PXD008934	−0.0162	−0.5071	3.2223	−0.3261	0.4765
PXD009142	0.7738	−0.8290	3.9258	−0.4338	0.6586

In table 3.5 the numbers of all non-redundant peptides at a FDR of 1% after Percolator was used are listed. The ratios of unique to shared peptides stayed the same after using Percolator compared to the original search results by MS-GF+. On the contrary the number of identified peptides in the ORF database and in total changed for all projects (Figure 3.6).

Table 3.5: Percolator results. Reported are the number of identified peptides (unique,shared and total) and the number of mapped search database entries compared to Ensembl PEP and the constructed ORF database (ORF DB). The single-peptide rule was also applied here.

Project	Unique peptides	Shared peptides	Total peptides	Database entries
PXD002273				
Ensembl PEP	4,759 (19.06%)	20,214 (80.94%)	24,973	29,892
ORF DB	19,972 (97.38%)	537 (2.62%)	20,509	21,117
			45,482	51,009
PXD007745				
Ensembl PEP	8,028 (18.56%)	35,229 (81.44%)	43,257	35,099
ORF DB	5,468 (99.65%)	19 (0.35%)	5,487	5,502
			48,744	40,601
PXD008934				
Ensembl PEP	7,360 (18.96%)	31,451 (81.04%)	38,811	24,535
ORF DB	5,689 (98.75%)	72 (1.25%)	5,761	5,787
			44,572	30,322
PXD009142				
Ensembl PEP	27,782 (19.71%)	113,151 (80.29%)	140,933	66,851
ORF DB	8,257 (99.54%)	38 (0.46%)	8,295	8,235
			148,328	75,086

The re-evaluation of PSMs of Percolator can be distinguished into two groups [105]. One group comprises all peptides from the canonical protein sequence database (Ensembl PEP) and the other group consists of peptides from the ORF database(s). While the number of identified peptides does not change too much between MS-GF+ and Percolator for peptides from Ensembl PEP (ranging from a 6% decrease in PXD008934 to a 17% increase in identified peptides in PXD007745 after Percolator) there is a major

difference in the ORF databases, the projects PXD002273, PXD008934 and PXD009142 lose 36.33%, 93.01% and 82.2% of identified peptides after post-processing of the data. Only for the project PXD007745 there is an increase in identifications to both databases (154.4% to the ORF DB).

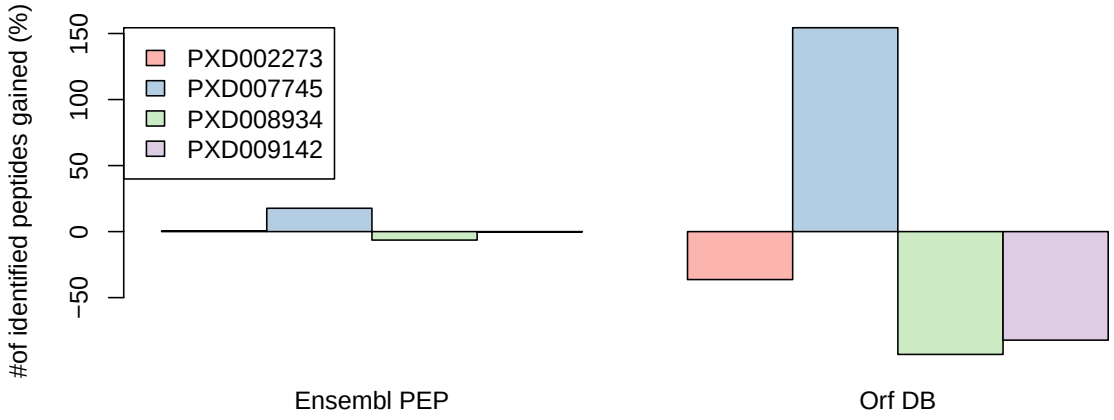


Figure 3.6: Changes compared to MS-GF+. The left side shows the changes in number of identified peptides of Ensembl PEP. On the right side the changes of identified peptides from the ORF database are shown.

With the exception of the project PXD007745 the number of identified novel peptides decreased to a great extend as it was hypothesized in the beginning of this section. PXD007745 has not only the greatest gain of identifications compared to its ORF DB but also compared to Ensembl PEP. In contrast to the other projects PXD007745 had the highest negative weight for absdM (absolute difference between experimental and calculated mass) and the smallest weight for lnEvalue in table 3.4. Compared to the weights of the other projects Percolator seems to find the most discriminating power from these features. This project also had by far the lowest number of identified peptides by MS-GF+ in its ORF DB. Whereas the number of its novel peptides raise from 2157 to 5487 after post-processing, all other projects lose an order of magnitude more peptides.

This is perhaps the greatest surprise in these results, that tens of thousands of PSMs passing the FDR-threshold of MS-GF+ are no correct target PSMs according to Percolator. And while the number of identifications remain relatively stable for Ensembl PEP these PSMs were mostly found in the ORF databases. This indicates that post processing is of high importance in proteogenomics because it helps primary in the correct identification of novel peptides.

3.7 PhyloCSF

Even though MS-based proteomics provides direct evidence of protein-level expression additional orthogonal evidence should be provided to further corroborate novel peptides identified by proteogenomics. To find and to assess the coding potential of these peptides PhyloCSF was used.

The genomic coordinates of novel peptides were stored in a gff3 file and provided to the script described in section 2.7, resulting in 682 novel peptides from conserved regions. These peptides had at least two mismatches to Ensembl PEP making sure the novel identification was not mistaken for a highly abundant, homologous peptide following Nesvizhskii's advice [3] of removing homologous peptides. This resulted in 682 novel peptides located on all chromosomes except the Y-chromosome which are displayed in a manhattan plot in Figure 3.7.

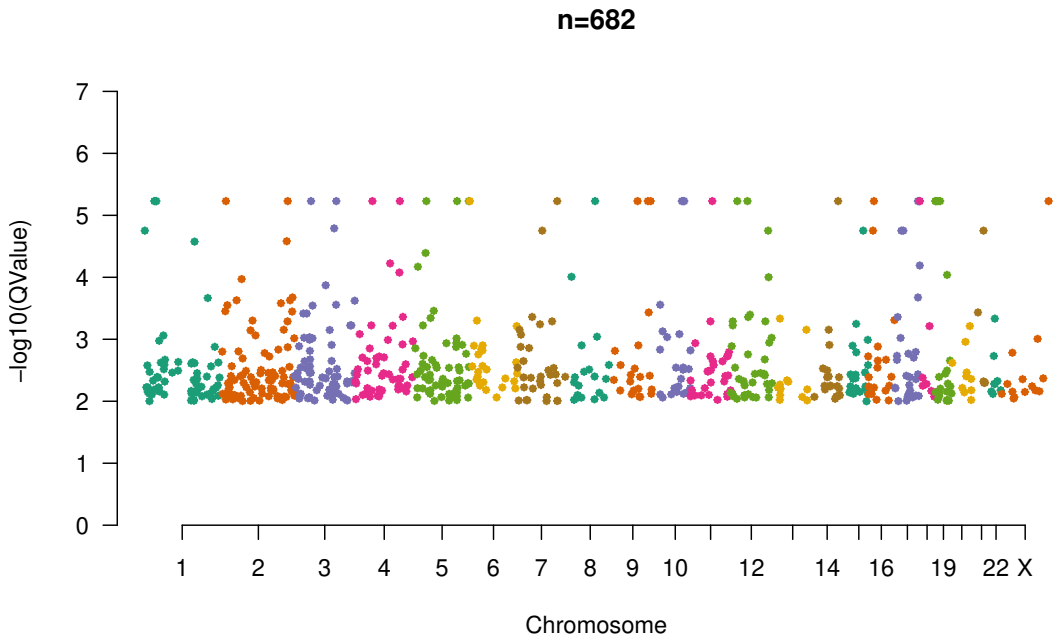


Figure 3.7: Manhattan plot of PhyloCSF conserved peptide sequences. Each dot of the manhattan plot is the QValue of the PSM of the peptide sequence. QValues on the y-axis were recomputed by Percolator.

Table 3.6 depicts the different biotypes of these 682 novel peptides. The two most frequent classifications were ALT_FRAMEs and smORFs. Most refinements of annotated exons (intragenic classifications) came from protein coding genes. 64 UTR exons were conserved and 8 exons or novel exons in a long intergenic noncoding gene having a genome unique peptide mapped to them and could also be assessed to be in conserved regions.

Table 3.6: Classifications of conserved novel peptides and their corresponding biotypes. CDSs and smORFs cannot have a biotype as they are defined as intergenic regions by the classification.

Classification	#	biotype(s)
EXON	4	2 x pseudogene, 1 x antisense, 1 x lincRNA
UTR	64	64 x protein_coding
ALT_FRAME	158	145 x protein_coding, 4 x antisense, 9 x pseudogene
EXON_EXT	66	63 x protein_coding, 3 x pseudogene
NOVEL_EXON	47	17 x protein_coding, 20 x antisense
		7 x lincRNA, 3 x processed_transcript
CDS	81	
smORF	262	

52 of 64 ORFs from UTR peptides were below 300 nucleotides in length, defining them as upstream ORFs (uORFs) or smORFs from the 3'UTR. uORFs constitute a very abundant class of smORFs that are translation initiation sites upstream of coding sequences. The expression of uORFs have an effect on the regulation of protein expression of the downstream coding sequence. 144 of the 343 intergenic classifications (CDS and smORF) were overlapping a gene on the opposing strand.

The peptide EALNEFLTR from a smORF found in the MS-based peptidomic study from Slavoff et al. [21] was the only overlap from their study with this thesis. And this identification could not be unambiguously identified because the highly abundant peptide AELNEFLTR from the RPS3 gene (ribosomal protein S3) had the same scores for its MS/MS spectrum in the database search. Although they also removed peptides with 1 mismatch to Ensembl PEP such a case is missed because an inversion of the first two amino acids results in 2 mismatches in the alignment. The low number of overlapping peptides could be explained by the different analyzed sample origins as Slavoff et al. used a human leukemia cell line (K562) and the cell line HEK293.

38 novel peptides had an overlap with the proteogenomic study (Integrated proteogenomic analysis workflow, IPAW) from Zhu et al. [8]. More evidence was used in their study to assess the coding potential of novel identifications than PhyloCSF (they could also access RNA-Seq data from their analyzed cell lines), 4 of these peptides had a positive PhyloCSF prediction. Overlapping peptides had following biotypes: 27 were pseudogenes, 10 protein coding and 1 TEC (To be Experimentally Confirmed). In general, most of their novel peptides mapped to pseudogenes, in contrast to the results in this thesis were only 2% of novel peptides with PhyloCSF evidence mapped to pseudogenes. This is because IPAW added pseudogene entries to their search database which was not done in my searches. Another explanation might be that pseudogenes have high RNA-Seq expression levels in cancer such as the A431 cell line used in the IPAW study.

While some Ribo-Seq experiments state that a large number of lincRNAs might be

translated [108] this has not been the case in the reanalysis of the four PRIDE projects. Low numbers of novel peptides from lincRNAs have also been found by Slavoff et al. and the IPAW study.

3.8 Examples of proteogenomic peptides

In this section differently classified novel peptides are presented as examples. Even though all proteomic data shown in this thesis was post processed to reinforce peptide spectrum matches and the peptides given here as examples are genome and database unique the question remains if they are true positives and have really been expressed. Therefore clearly annotated synthetic spectra should be provided for such strong claims like the novel peptides identified in proteogenomics.

Prosit, a recently developed fragment ion intensity prediction tool [46] was used to create reference spectra for spectral comparison. It provides near-synthetic peptide quality tandem mass spectra and its predicted spectra can therefore serve as a substitute for synthetic peptides in this thesis. This allows me to show if the experimental spectra depicted here have been generated by the peptide sequences assigned to them.

As a measure of similarity of an experimental and a predicted MS/MS spectra the normalized spectral contrast angle (SA) was used. The SA of two fragment ion intensity vectors I^R and I^T is calculated in the following way:

$$DP(I^R, I^T) = \frac{\sum_{j=1}^{n} I_j^R I_j^T}{\sqrt{\sum_j (I_j^R)^2 (I_j^T)^2}},$$

$$SA = 1 - 2\frac{cos^{-1} DP(I^R, I^T)}{\pi}$$

It serves as a similarity measure between two MS/MS spectra in the range $[0, 1]$ where 0 means no similarity and 1 means both spectra are the same.

N-terminal extension in UTR exon

The N-terminal extension of of the C1orf122 gene is one of very few examples were two or more novel peptides corroborate the expression of a protein in this thesis (Figure 3.8).

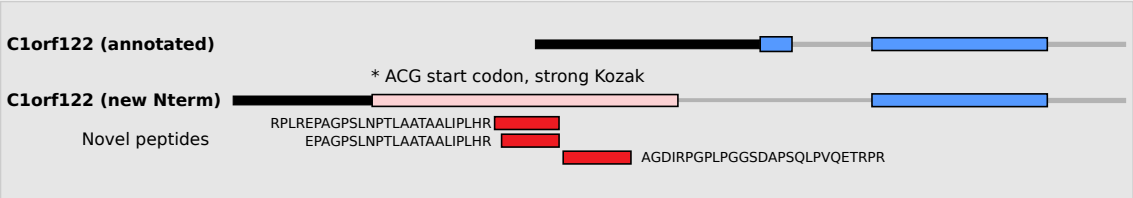


Figure 3.8: N-terminal extension of the C1orf122 gene (uncharacterized protein). Black boxes are UTR regions, blue boxes exons, grey lines introns and red boxes are peptides. Three novel peptides have been identified at the locus on the forward strand on Chromosome 1:37806979-37809454, overlapping an UTR exon. The peptides AGDIRPGPLPGGSDAPSQLPVQETRPR and EPAGPSLNPTLAATAALIPLHR have been identified by Zhu et al., the peptide RPLREPAGPSLNPTLAATAALIPLHR has first been observed in this thesis. The usage of a non-AUG near-cognate translation start codon, in this case a ACG codon, has been determined by Zhu et al. Adapted from [8].

Compared to predicted spectra from Prosit all three novel peptides achieved high similarities, EPAGPSLNPTLAATAALIPLHR its normalized spectral contrast angle (SA) is 0.86, RPLREPAGPSLNPTLAATAALIPLHR its SA is 0.82 and AGDIRPGPLPGGSDAPSQLPVQETRPR its SA is 0.69.

Exon extension

A novel peptide from the HNRNPA1 gene was identified that extended one of its exons by 4 amino acids upstream (Figure 3.9). HNRNPA1's protein products are RNA-binding proteins that associate with pre-mRNAs in the nucleus and influence pre-mRNA processing, as well as other aspects of mRNA metabolism and transport [109].

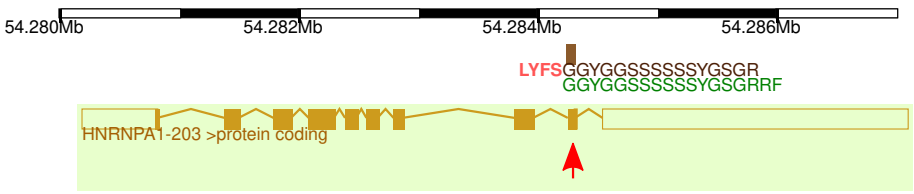


Figure 3.9: Exon extension of the HNRNPA1 gene (heterogeneous nuclear ribonucleoprotein A1). Locus on the forward strand on chromosome 12:54284245-54284304. Amino acids in green show the full exon sequence, the peptide LYFSGGYGGSSSSSSYGSGR was identified, amino acids in red show the exon extension.

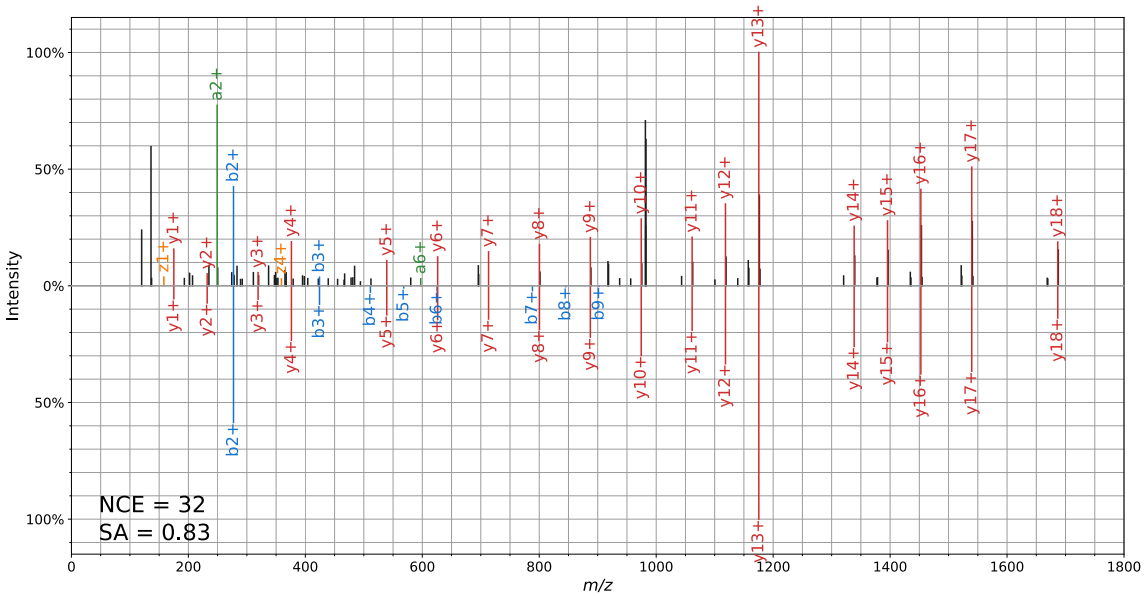


Figure 3.10: Mirror plot of the experimental spectrum of the peptide LYFSGGYGGSSSSSSYGSGR at the top and the predicted spectrum from Prosit at the bottom. B-ions are colored in blue, y-ions are colored in red, a-ions are colored in green and z-ions are colored in orange. This color scheme will be used throughout all mirror plots. Almost all predicted main y-ions are present and have similar intensities compared to the spectrum at the bottom resulting in a very high normalized spectral contrast angle of 0.83. The normalized collision energy (NCE) is a parameter of the MS instrument used to optimize fragment ion intensities, it is needed by Prosit for its predictions.

Novel exon

An identified novel peptide mapped between the second and third exon of the TUBA4B (putative tubulin-like protein alpha-4B) gene. Tubulins are proteins that build the microtubuli, a structure in the cytoskeleton of eukaryotic cells involved in many import processes inside the cell such as intracellular transport. There are multiple stop codons located up - and downstream its locus, indicating that its not part of the previous or subsequent exon.

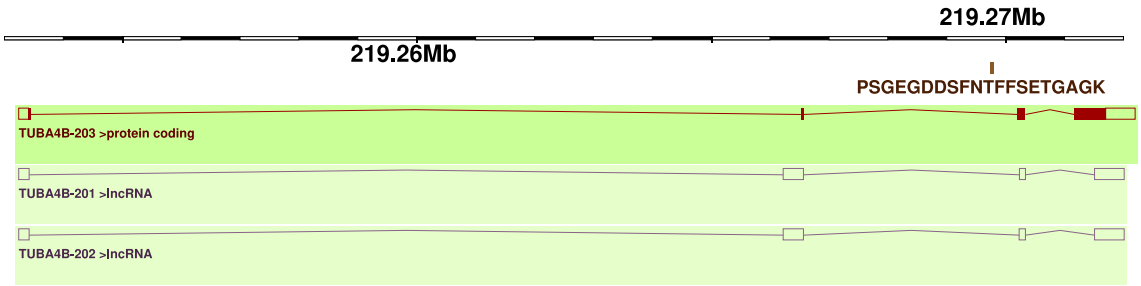


Figure 3.11: The peptide PSGEGDDSFNTFFSETGAGK was identified in the intragenic locus on the forward strand on chromosome 2: 219269510-219270103.

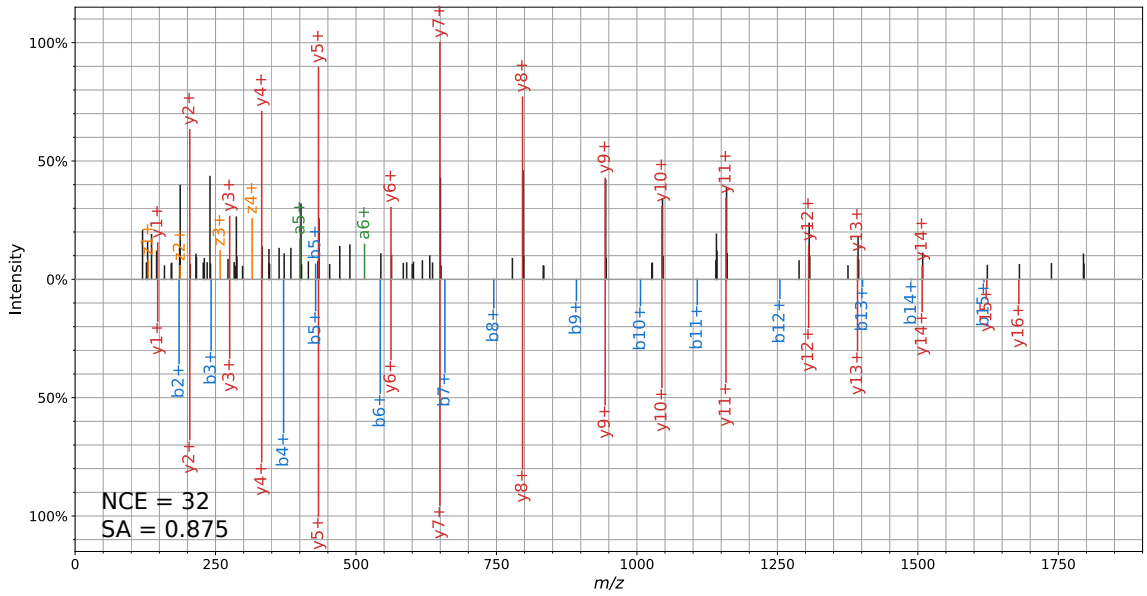


Figure 3.12: Mirror plot of the experimental spectrum of the peptide PSGEGDDSFNTFFSETGAGK at the top and the predicted spectrum from Prosit at the bottom. The ions align very well between the experimental and predicted spectrum resulting in a very high normalized spectral contrast angle of 0.875.

Small Open Reading Frame with Ribo-Seq evidence

The peptide AQGLESLLEHGLR was found in a 82 amino acid long ORF in an alternative reading frame of the cold inducible RNA binding protein (CIRBP) in the project PXD009142. Proteins from the CIRBP gene play a protective role on various genotoxic, cellular stresses such as UV-radiation [110]. They act as a translational repressor by binding to the 3'UTR region of stress-responsive transcripts. The locus of its ORF is on the forward strand on chromosome 19:1274512-1274784. This smORF maps to an alternative reading frame of a 3'UTR exon from multiple CIRBP transcripts and its length and expression has been determined by Ribosome-Profiling [111]. Prosit could further validate this peptides identification, the spectral comparison resulted in a very high normalized spectral contrast angle of 0.852. There was no peptide in the same sample that had a similar m/z value that could have originated this MS/MS spectrum.

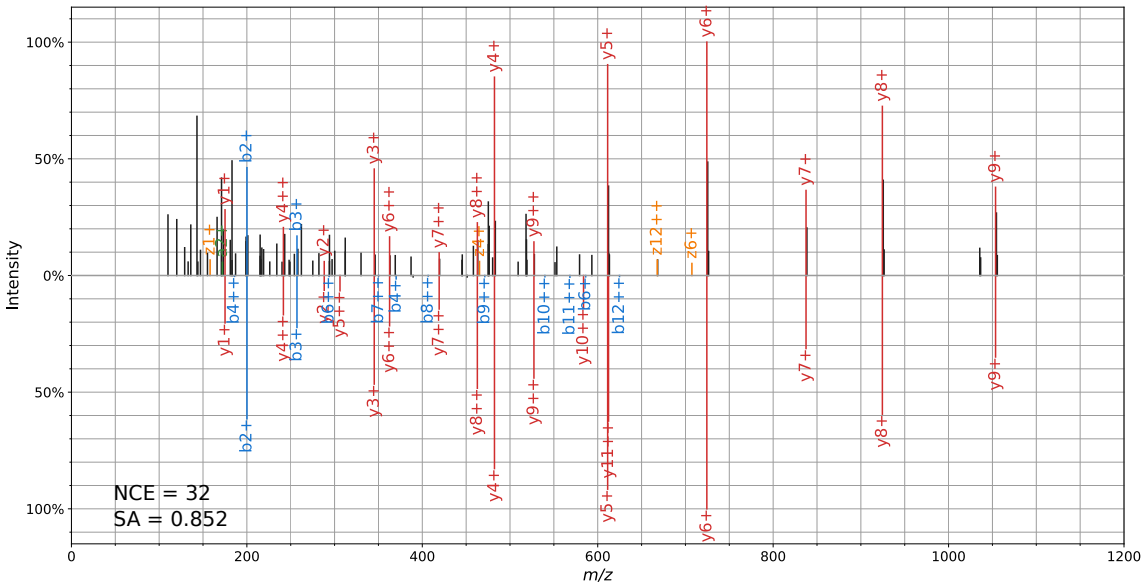


Figure 3.13: Mirror plot of the peptide AQGLESLLEHGLR. The most intense fragment ions are explained and correlate very well with the predicted fragment ion intensities.

Listing 3.1: The ORF of the identified peptide from start to stop codon.

```
atg gcg ggc gcc ttt cgg tgt gtg ttg ggt gca ggg cag cgc ctc ccg gga gcg ccg ggt
 M   A   G   A   F   R   C   V   L   G   A   G   Q   R   L   P   G   A   P   G
ccc ccg cct gga gcc cgc gcc tgt tct ccc tcc ctt cct cct cct tcc agg agg cgc ttc
 P   P   P   G   A   R   A   C   S   P   S   L   P   P   P   S   R   R   R   F
gcc agt gag gtg cgg gct cag ggc ctc gag tct ctc ctg gag cac ggg ctg cgg tgc gcc
 A   S   E   V   R   A   Q   G   L   E   S   L   L   E   H   G   L   R   C   A
ggc agc tta cgg ggc ggc cag tcc ttg ccc aca acg atg tgg agc cct gtg aaa gtc gga
 G   S   L   R   G   G   Q   S   L   P   T   T   M   W   S   P   V   K   V   G
ttc gaa
 F   E
```

A difficult to find false positive identification

The peptide IVVVVSSPR was identified in an intergenic ORF at the locus on the forward strand on chromosome 15:89683734-89683760. The neighboring genes are the TICRR gene which is 52kb upstream and the WDR93 gene located 7kb downstream. An exon from the PEX11A gene is located on the opposite forward strand.

The ORF it is located in is 217 nucleotides long (from stop to stop codon) and does not contain a triplet coding for methionine. However, the ORF also lacks a strong Kozak consensus sequence surrounding an alternative start codon. This raises the question if it belongs to a missing transcript in Ensembl of the upstream gene TICRR or if the PSM was wrongly assigned to a false peptide sequence in the database search by MS-GF+.

Prosit was again used to predict the MS/MS spectra from all peptide sequences with similar experimental m/z values ($\Delta m/z \leq$ precursor mass difference used in the search, in this case 20ppm) found in the sample. The peptide IVVVVSSPR could only

be confidently identified if none of these other predicted MS/MS spectra had a similar value or even a higher value for the normalized spectral contrast angle compared to the experimental spectrum from the original PSM. The peptide ITLVSAAPGK met these criteria and its predicted fragment ions were also compared against the experimental MS/MS spectrum (comparisons in Figure 3.14). Both spectral comparisons resulted in high similarities but the sequence ITLVSAAPGK (SA of 0.84) gave a better explanation for the peptide identification than the assigned sequence IVVVVSSPR (SA of 0.736).

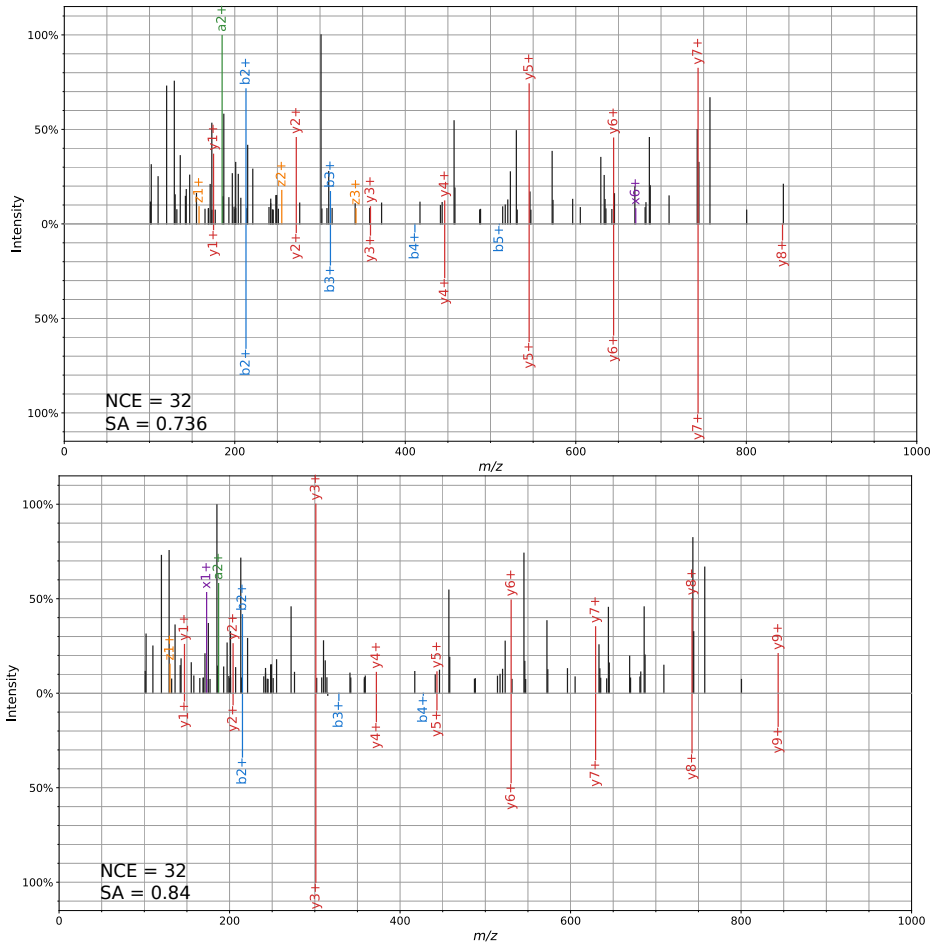


Figure 3.14: The two mirror plots show the experimental spectrum at top and different predicted spectra at the bottom. In the above mirror plot the peptide IVVVVSSPR is getting compared with its predicted spectrum and in the mirror plot below the peptide ITLVSAAPGK is getting compared. Both PSMs have high normalized spectral contrast angles and explain their predicted spectra well but the latter spectral comparison results in a better SA value.

Table 3.7: Scores from MS-GF+ values of the presented peptides and their normalized spectral contrast angle (SA) compared to the predicted spectra from Prosit. If RawScore is equal to DeNovoScore that means that the optimal peptide to that spectrum was found.

Sequence	Score	DeNovoScore	QValue	SpecEValue	SA
UTR					
RPLREPAGPSLNPTLAATAALIPLHR	137	226	$5.8e{-6}$	$1.50e{-16}$	0.82
EPAGPSLNPTLAATAALIPLHR	180	211	$5.8e{-6}$	$1.31e{-20}$	0.86
AGDIRPGPLPGGSDAPSQLPVQETRPR	199	242	$5.8e{-6}$	$2.34e{-24}$	0.69
ALT FRAME					
LLLLPCGPAEEAAHPGVAAQLLPAVAEDGFR	226	245	$5.8e{-6}$	$2.92e{-32}$	
EXON EXT.					
LYFSGGYGGSSSSSSYGSGR	254	254	$5.8e{-6}$	$5.80e{-25}$	0.83
NOVEL EXON					
PSGEGDDSFNTFFSETGAGK	178	217	$5.8e{-6}$	$3.95e{-20}$	0.875
smORF					
AQGLESLLEHGLR	100	102	$5.8e{-6}$	$2.05e{-15}$	0.852
False positive					
IVVVVSSPR	87	91	0.001	$4.66e{-12}$	0.736

The first three peptides in table 3.7 corroborate the expression of a N-terminal extension of the uncharacterized protein C1orf122. The high scores from MS-GF+ and the very high similarities to their predicted spectra confirm this. The peptides from the exon extension and novel exon examples also obtained very good scores from MS-GF+ and very high values for normalized spectral contrast angles and could thus be verified.

The peptide LLLLPCGPAEEAAHPGVAAQLLPAVAEDGFR mapped to an alternative reading frame of multiple transcripts (Ensembl transcript name SLC26A11-214, SLC26A11-210, SLC26A11-205, SLC26A11-211 and SLC26A11-203) of the SLC26A11 (Sodium-independent sulfate anion transporter) gene. The locus of its ORF is on the forward strand on chromosome 17:80221518-80221769. It had the highest spectral e value from MS-GF+ of all novel peptides in this thesis but unfortunately Prosit could not predict its fragment ion intensities because it is limited to a maximum peptide length of 30, whilr this sequence consists of 31 amino acids residues. Hence it could not be validated completely, a clearly annotated MS/MS spectrum can be seen in supplementary Figure S6.

The likely false positive identified peptide IVVVVSSPR gives a good example why novel identifications need to be validated very carefully as it has no homologous sequence in Ensembl PEP, good scores from MS-GF+, passed the re-ranking of Percolator and has a very high normalized spectral contrast angle (0.736) compared to its predicted

MS/MS spectrum from Prosit. However, a different peptide sequence from Ensembl PEP explains its spectrum even better, leading to the conclusion that peptides with similar experimental m/z values and the same charge state should also be compared to the experimental spectrum in question. This task is done by the database search tool and this might be a small caveat of MS-GF+ because it only outputs the best matching peptide spectrum match without any information how well the second ranked PSM scored as it is the case for other database search tools. This information would facilitate the validation process.

The peptide AQGLESLLEHGLR was therefore compared under the same strict criteria. It lacked a homologous sequence in Ensembl PEP and no peptide found in the same samples with similar experimental m/z value could explain its experimental MS/MS spectrum as well. Its identification is especially interesting because it does further corroborate the translation of its ORF, as has already been demonstrated by a ribosome-profiling experiment, serving as additional orthogonal evidence. However, ribosome-profiling does not provide evidence for a stable translated protein product.

3.9 Web prototype

The pipeline was extended by a web prototype to create a platform to display the results from Peptidome-DB and to allow for potential user interactions such as browsing a database or uploading proteogenomic results. Uploads would subsequently be processed and used as input by the pipeline.

A form for an upload procedure was created, alongside file uploads of the sequence database and database search results it included all necessary experimental parameters and an optional text field to the PRIDE archive (displayed in Figure 3.15). This cross reference is useful if the data had already been published by the user him - or herself or a different user re-analyzed the raw files uploaded to the PRIDE database.

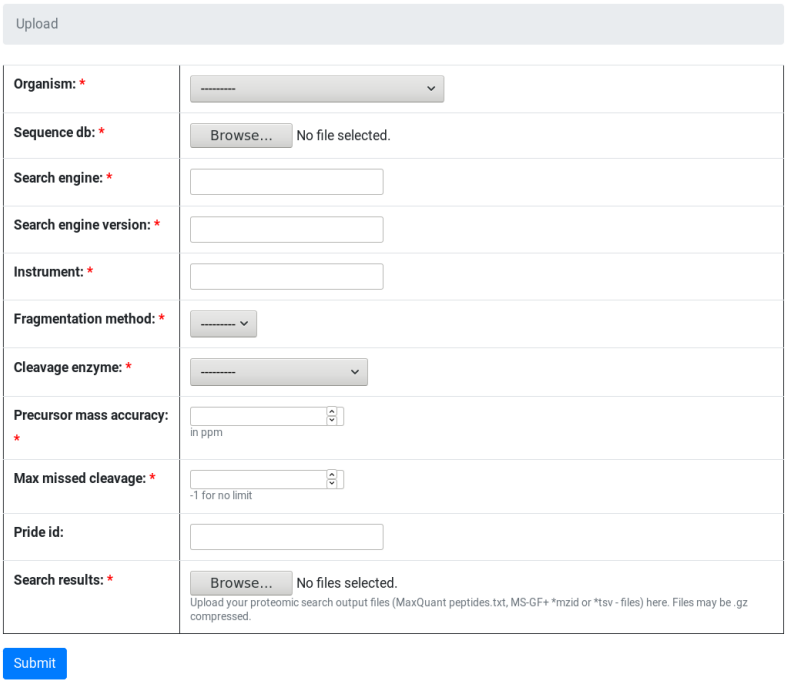


Figure 3.15: The implemented upload form. All parameters used by the search engine in the database search must be included. Peptide identification results can change under different versions used, hence this information must also be entered. The mandatory `Organism` field is displayed as a select list with all available species in Ensembl.

To every successfully submitted upload an unique key is assigned which is used as name for a newly created directory where all uploaded files are saved and this is also the directory where the pipeline is writing its output files. It is displayed in the browser to the user and serves as an identification key and process id. After completion of Peptidome-DB with the uploaded data as input the pipeline results get displayed to the user. Theoretically, an implemented upload function would compare the novel results to entries in the database, store new entries and increment a counter for multiple identified peptides or ORFs and link them to the experiments (database schema in Figure S5). However, inputted user data can not be controlled (e.g if the database search results came from size inflated databases) which would lead to an accumulation of spurious entries from peptide identifications with low confidence in the database.

The front-end was kept very lightweight. A responsive layout gives a satisfying visualization on different devices (e.g tablets and smartphones). The frontpage consists of three columns (see Figure 3.16): the first column leads to the browse page, the second column allows to upload data and the third column could be used to link to an information page.

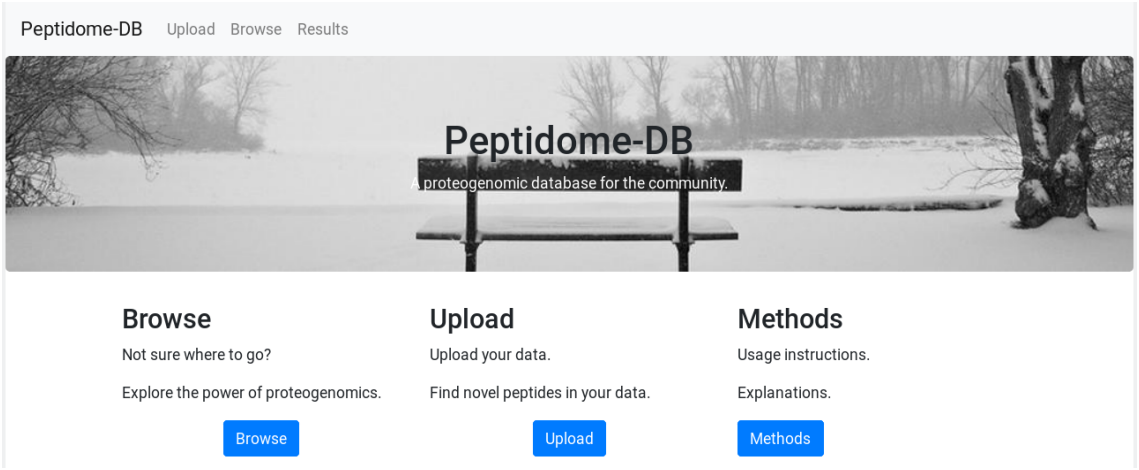
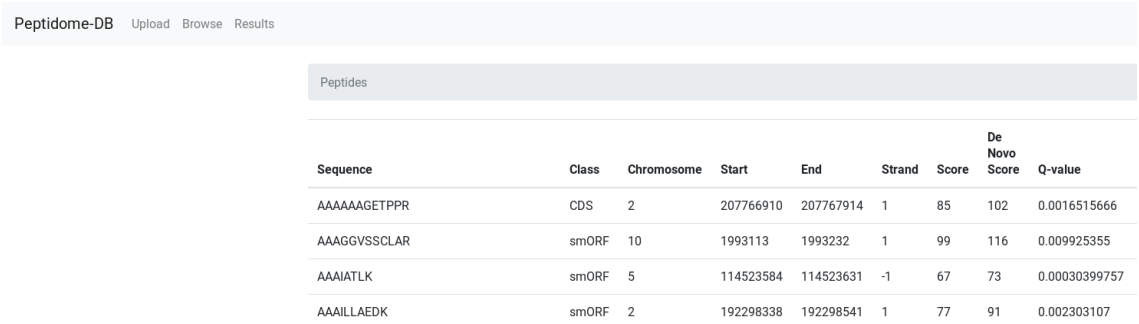


Figure 3.16: The responsive frontpage. The three column view gives a quick overview which information can be retrieved from the website.



Sequence	Class	Chromosome	Start	End	Strand	Score	De Novo Score	Q-value
AAAAAAGETPPR	CDS	2	207766910	207767914	1	85	102	0.0016515666
AAAGGVSSCLAR	smORF	10	1993113	1993232	1	99	116	0.009925355
AAAIATLK	smORF	5	114523584	114523631	-1	67	73	0.00030399757
AAAILLAEDK	smORF	2	192298338	192298541	1	77	91	0.002303107

Figure 3.17: The `Browse` page displays novel identified peptides. Genomic loci and search engine scores can be viewed.

The implementation of the web prototype has basic functionality such as data upload and the browsing of stored data. The front-end development was kept minimalist albeit it contains the most important pages. Considering the fact that the general implementation is just a prototype this is a reasonable compromise.

Chapter 4

Conclusion

The aim of this thesis was to implement a software tool that utilizes proteomic, genomic and transcriptomic data for the identification of novel peptides missing in reference protein databases. Here I could present Peptidome-DB, an integrated Java pipeline comprising proteomic data integration from different database search tools and output formats, genome mapping against multiple databases from Ensembl from different molecule types (proteins, coding and non-coding transcripts, and the genome), clustering of unannotated ORFs with exons and classification of novel peptides and their ORFs compared to Ensembl annotations, resulting in complementary data to Ensembl. The access to proteomic database search result files was implemented using the generic factory pattern which allows the integration of additional output format from different database search tools in the future. The alignment step is fully automated and the clustering and subsequent proteogenomic classification were carefully evaluated with ground truth data. An optimal method to determine the distance threshold for the single-linkage clustering was described that is applicable to all eukaryotic organisms.

A six-frame translation of the human genome could generate specific protein sequence databases by an iterative database search. The re-analyzed data sets from the PRIDE archive highlight the importance of post-processing in proteogenomics, Percolator reduced a high number of spurious peptide identifications compared to MS-GF+. Peptidome-DB was successfully tested on these database search results and could show its usefulness in the identification of novel peptides, expanding the known peptidome and refining known gene models. Novel peptides could be found in annotated transcripts (alternative frames, novel exons and exon extensions) and in intergenic regions. The utilization of Prosit, a tool for the prediction of fragment ion intensities of tandem mass spectra, could corroborate given examples found by Peptidome-DB.

The web prototype consists of the the most informative pages and their functions including a browse page and an upload page. The data found in this work is stored in the database implemented for the web prototype, creating a new database extending the known peptidome. The novel proteogenomic findings identified in this thesis have only

been structurally annotated, their biological importance remains unknown.

An experimental limitation is the difficulty in analyzing low abundant compounds in MS-based bottom-up proteomics where only the most intense precursors are selected for further MS/MS analysis which might miss expressed smORFs. Another limitation are hydrophobic sequences (e.g transmembrane proteins), too hydrophobic peptides might stick to gel or adsorb to tubes and fail to elute from columns and are not detected. Furthermore, the search for smORFs by proteogenomics would be alleviated by more publicly available samples with better sample preparation enriching for low molecular weight proteins for example by fractionation of protein samples before LC-MS/MS analysis [3].

The possible integration of fragment ion intensity prediction tools like Prosit into database search tools could facilitate peptide identifications in the future. Together with the developed similarity measures of spectral comparisons by proteomic search engines this would help in eliminating spurious peptide spectrum matches which can only be beneficial to proteogenomics.

Bibliography

[1] Matthias Mann, Nils A Kulak, Nagarjuna Nagaraj, and Jürgen Cox. The coming age of complete, accurate, and ubiquitous proteomes. *Molecular cell*, 49(4):583–590, 2013.

[2] Arthur D Tinoco and Alan Saghatelian. Investigating endogenous peptides and peptidases using peptidomics. *Biochemistry*, 50(35):7447–7461, 2011.

[3] Alexey I Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature methods*, 11(11):1114, 2014.

[4] Xiaohan Yang, Timothy J Tschaplinski, Gregory B Hurst, Sara Jawdy, Paul E Abraham, Patricia K Lankford, Rachel M Adams, Manesh B Shah, Robert L Hettich, Erika Lindquist, et al. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome research*, 21(4):634–641, 2011.

[5] Natalie E Castellana, Samuel H Payne, Zhouxin Shen, Mario Stanke, Vineet Bafna, and Steven P Briggs. Discovery and revision of arabidopsis genes by proteogenomics. *Proceedings of the national academy of sciences*, 105(52):21034–21038, 2008.

[6] Ulrich Omasits, Adithi R Varadarajan, Michael Schmid, Sandra Goetze, Damianos Melidis, Marc Bourqui, Olga Nikolayeva, Maxime Québatte, Andrea Patrignani, Christoph Dehio, et al. An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome research*, 27(12):2083–2095, 2017.

[7] Malgorzata A Komor, Thang V Pham, Annemieke C Hiemstra, Sander R Piersma, Anne S Bolijn, Tim Schelfhorst, Pien M Delis-van Diemen, Marianne Tijssen, Robert P Sebra, Meredith Ashby, et al. Identification of differentially expressed splice variants by the proteogenomic pipeline splicify. *Molecular & Cellular Proteomics*, 16(10):1850–1863, 2017.

[8] Yafeng Zhu, Lukas M Orre, Henrik J Johansson, Mikael Huss, Jorrit Boekel, Mattias Vesterlund, Alejandro Fernandez-Woodbridge, Rui MM Branca, and Janne

Lehtiö. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nature communications*, 9(1):903, 2018.

[9] James C Wright, Jonathan Mudge, Hendrik Weisser, Mitra P Barzine, Jose M Gonzalez, Alvis Brazma, Jyoti S Choudhary, and Jennifer Harrow. Improving gencode reference gene annotation using a high-stringency proteogenomics workflow. *Nature communications*, 7:11778, 2016.

[10] John R Yates, Jimmy K Eng, and Ashley L McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Analytical chemistry*, 67(18):3202–3210, 1995.

[11] Michael Q Zhang. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3(9):698, 2002.

[12] Jyoti S Choudhary, Walter P Blackstock, David M Creasy, and John S Cottrell. Matching peptide mass spectra to est and genomic dna databases. *Trends in biotechnology*, 19:17–22, 2001.

[13] Jacob D Jaffe, Howard C Berg, and George M Church. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, 4(1):59–77, 2004.

[14] Henry Rodriguez and Stephen R Pennington. Revolutionizing precision oncology through collaborative proteogenomics and data sharing. *Cell*, 173(3):535–539, 2018.

[15] Pierre-Alain Maron, Lionel Ranjard, Christophe Mougel, and Philippe Lemanceau. Metaproteomics: a new approach for studying functional microbial ecology. *Microbial ecology*, 53(3):486–493, 2007.

[16] Harm van Bakel, Corey Nislow, Benjamin J Blencowe, and Timothy R Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS biology*, 8(5):e1000371, 2010.

[17] Jeroen Crappé, Wim Van Criekinge, and Gerben Menschaert. Little things make big things happen: a summary of micropeptide encoding genes. *EuPA Open Proteomics*, 3:128–137, 2014.

[18] Volodimir Olexiouk, Wim Van Criekinge, and Gerben Menschaert. An update on sorfs. org: a repository of small orfs identified by ribosome profiling. *Nucleic acids research*, 46(D1):D497–D502, 2017.

[19] Marcel E Dinger, Ken C Pang, Tim R Mercer, and John S Mattick. Differentiating protein-coding and noncoding rna: challenges and ambiguities. *PLoS computational biology*, 4(11):e1000176, 2008.

[20] Emmanuel Ladoukakis, Vini Pereira, Emile G Magny, Adam Eyre-Walker, and Juan Pablo Couso. Hundreds of putatively functional small open reading frames in drosophila. *Genome biology*, 12(11):R118, 2011.

[21] Sarah A Slavoff, Andrew J Mitchell, Adam G Schwaid, Moran N Cabili, Jiao Ma, Joshua Z Levin, Amir D Karger, Bogdan A Budnik, John L Rinn, and Alan Saghatelian. Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature chemical biology*, 9(1):59, 2013.

[22] Juan-Pablo Couso and Pedro Patraquim. Classification and function of small open reading frames. *Nature reviews Molecular cell biology*, 18(9):575, 2017.

[23] Joanna Somers, Tuija Pöyry, and Anne E Willis. A perspective on mammalian upstream open reading frame function. *The international journal of biochemistry & cell biology*, 45(8):1690–1700, 2013.

[24] Yeast genome. Gcn4 uorf. `https://www.yeastgenome.org/locus/GCN4/regulation`, 2019. [Online; accessed 10-September-2019].

[25] Máximo Ibo Galindo, Jose Ignacio Pueyo, Sylvaine Fouix, Sarah Anne Bishop, and Juan Pablo Couso. Peptides encoded by short orfs control development and define a new eukaryotic gene family. *PLoS biology*, 5(5):e106, 2007.

[26] Changhan Lee, Kelvin Yen, and Pinchas Cohen. Humanin: a harbinger of mitochondrial-derived peptides? *Trends in Endocrinology & Metabolism*, 24(5):222–228, 2013.

[27] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761, 2017.

[28] Ensembl. Automatic annotation of coding genes. `https://www.ensembl.org/info/genome/genebuild/automatic_coding.html`, 2019. [Online; accessed 4-October-2019].

[29] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current

status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2015.

[30] NCBI. The ncbi eukaryotic genome annotation pipeline. `https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/`, 2019. [Online; accessed 4-October-2019].

[31] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2018.

[32] UniProt Consortium. Protein existence. `https://www.uniprot.org/help/protein_existence`, 2019. [Online; accessed 13-December-2019].

[33] Santosh Renuse, Raghothama Chaerkady, and Akhilesh Pandey. Proteogenomics. *Proteomics*, 11(4):620–630, 2011.

[34] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227, 2012.

[35] Jainab Khatun, Yanbao Yu, John A Wrobel, Brian A Risk, Harsha P Gunawardena, Ashley Secrest, Wendy J Spitzer, Ling Xie, Li Wang, Xian Chen, et al. Whole human genome proteogenomic mapping for encode cell line data: identifying protein-coding regions. *BMC genomics*, 14(1):141, 2013.

[36] Gloria A Brar and Jonathan S Weissman. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature reviews Molecular cell biology*, 16(11):651, 2015.

[37] Jesper V Olsen, Shao-En Ong, and Matthias Mann. Trypsin cleaves exclusively c-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics*, 3(6):608–614, 2004.

[38] Harald Marx. Practical course in modelling and systems biology. University Lecture, 2018.

[39] Kermit K. Murray. Mass spectrometry terms. `https://kermitmurray.com/msterms/index.php?title=Category:Final`, 2019. [Online; accessed 29-July-2019].

[40] Wikipedia contributors. De novo peptide sequencing — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=De_novo_peptide_sequencing&oldid=907761604`, 2019. [Online; accessed 29-July-2019].

[41] Ari M Frank. Predicting intensity ranks of peptide fragment ions. *Journal of proteome research*, 8(5):2226–2240, 2009.

[42] Alisa G Woods and Costel C Darie. *Advancements of Mass Spectrometry in Biomedical Research*, volume 806. Springer, 2014.

[43] Ommaima Khan. Mass analyzers (mass spectrometry). `https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Instrumental_Analysis/Mass_Spectrometry/Mass_Spectrometers_(Instrumentation)/Mass_Analyzers_(Mass_Spectrometry)`, 2019. [Online; accessed 30-July-2019].

[44] Thermo Fischer. Orbitrap. `https://www.thermofisher.com/at/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-systems/orbitrap-lc-ms.html#what-is-orbitrap`, 2019. [Online; accessed 30-July-2019].

[45] L. Martens I. Eidhammer, K. Flikka and S.-O. Mikalsen. *Computational mass spectrometry based proteomics*. Wiley, 2007.

[46] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, 16(6):509, 2019.

[47] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. 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, Nov 1994.

[48] Jacques Colinge and Keiryn L Bennett. Introduction to computational proteomics. *PLoS computational biology*, 3(7):e114, 2007.

[49] Jacques Colinge, Alexandre Masselot, Marc Giron, Thierry Dessingy, and Jérôme Magnin. Olav: Towards high-throughput tandem mass spectrometry data identification. *PROTEOMICS: International Edition*, 3(8):1454–1463, 2003.

[50] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207, 2007.

[51] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923, 2007.

[52] Thilo Muth and Bernhard Y Renard. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in bioinformatics*, 19(5):954–970, 2017.

[53] Eric W Deutsch, Christopher M Overall, Jennifer E Van Eyk, Mark S Baker, Young-Ki Paik, Susan T Weintraub, Lydie Lane, Lennart Martens, Yves Vandenbrouck, Ulrike Kusebauch, et al. Human proteome project mass spectrometry data interpretation guidelines 2.1. *Journal of proteome research*, 15(11):3961–3970, 2016.

[54] William S Sanders, Nan Wang, Susan M Bridges, Brandon M Malone, Yoginder S Dandass, Fiona M McCarthy, Bindu Nanduri, Mark L Lawrence, and Shane C Burgess. The proteogenomic mapping tool. *BMC bioinformatics*, 12(1):115, 2011.

[55] Christoph N Schlaffner, Georg J Pirklbauer, Andreas Bender, and Jyoti S Choudhary. Fast, quantitative and variant enabled mapping of peptides to genomes. *Cell systems*, 5(2):152–156, 2017.

[56] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, et al. Openms: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods*, 13(9):741, 2016.

[57] Yasset Perez-Riverol, Qing-Wei Xu, Rui Wang, Julian Uszkoreit, Johannes Griss, Aniel Sanchez, Florian Reisinger, Attila Csordas, Tobias Ternent, Noemi del Toro, et al. Pride inspector toolsuite: moving toward a universal visualization tool for proteomics data standard formats and quality assessment of proteomexchange datasets. *Molecular & Cellular Proteomics*, 15(1):305–317, 2016.

[58] Shivashankar H Nagaraj, Nicola Waddell, Anil K Madugundu, Scott Wood, Alun Jones, Ramya A Mandyam, Katia Nones, John V Pearson, and Sean M Grimmond. Pgtools: a software suite for proteogenomic data analysis and visualization. *Journal of proteome research*, 14(5):2255–2266, 2015.

[59] Harald Marx, Hannes Hahne, Susanne E Ulbrich, Angelika Schnieke, Oswald Rottmann, Dmitrij Frishman, and Bernhard Kuster. Annotation of the domestic pig genome by quantitative proteogenomics. *Journal of proteome research*, 16(8):2887–2898, 2017.

[60] Sangtae Kim and Pavel A Pevzner. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications*, 5:5277, 2014.

[61] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367, 2008.

[62] Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, Jennifer Siepen, Simon J Hubbard, Julian N Selley, Brian C Searle, James Shofstahl, Sean L Seymour, et al. The mzidentml data standard for mass spectrometry-based proteomics results. *Molecular & Cellular Proteomics*, 11(7):M111–014381, 2012.

[63] Johannes Griss, Andrew R Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G Thallinger, Reza M Salek, Christoph Steinbeck, Nadin Neuhauser, et al. The mztab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics*, 13(10):2765–2775, 2014.

[64] Alfred V Aho and Margaret J Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

[65] Lahodiuk, Yurii. aho-corasick-optimized. `https://devhub.io/repos/lagodiuk-aho-corasick-optimized`, 2019. [Online; accessed 4-October-2019].

[66] Ayush Govil. Aho-corasick algorithm for pattern searching. `https://www.geeksforgeeks.org/aho-corasick-algorithm-pattern-searching/`, 2019. [Online; accessed 25-November-2019].

[67] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[68] Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6(1):31, 2005.

[69] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[70] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[71] Guy St C Slater. Exonerate manual. `https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate-manual`, 2019. Online; accessed 22-July-2019.

[72] Patricia Sieber, Matthias Platzer, and Stefan Schuster. The definition of open reading frame revisited. *Trends in Genetics*, 34(3):167–170, 2018.

[73] Robert Sedgewick and Kevin Wayne. Analysis of algorithms. `https://algs4.cs.princeton.edu/14analysis`, 2019. Online; accessed 2-July-2019.

[74] The SQLite Consortium. Sqlite. `https://www.sqlite.org/index.html`, 2019. Online; accessed 2-July-2019.

[75] Ensembl. The ensembl public mysql servers. `https://www.ensembl.org/info/data/mysql.html`, 2019. [Online; accessed 4-October-2019].

[76] Ensembl. Ensembl core - schema documentation. `https://www.ensembl.org/info/docs/api/core/core_schema.html`, 2019. [Online; accessed 4-October-2019].

[77] Jeremy D Volkening, Derek J Bailey, Christopher M Rose, Paul A Grimsrud, Maegen Howes-Podoll, Muthusubramanian Venkateshwaran, Michael S Westphall, Jean-Michel Ane, Joshua J Coon, and Michael R Sussman. A proteogenomic survey of the medicago truncatula genome. *Molecular & Cellular Proteomics*, 11(10):933–944, 2012.

[78] Osamu Gotoh. Modeling one thousand intron length distributions with fitild. *Bioinformatics*, 34(19):3258–3264, 2018.

[79] Robert Sedgewick and Kevin Wayne. Singlelink.java. `https://algs4.cs.princeton.edu/94euclidean/SingleLink.java.html`, 2019. [Online; accessed 4-October-2019].

[80] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

[81] Karl Albert Hasselbalch. *Die Berechnung der Wasserstoffzahl des Blutes aus der freien und gebundenen Kohlensäure desselben, und die Sauerstoffbindung des Blutes als Funktion der Wasserstoffzahl*. Julius Springer, 1916.

[82] Wikipedia contributors. Amino acid — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Amino_acid&oldid=918355292`, 2019. [Online; accessed 4-October-2019].

[83] IUPAC. Compendium of chemical terminology. `http://goldbook.iupac.org/terms/view/H02907`, 2019. [Online; accessed 4-October-2019].

[84] Lukasz P Kozlowski. Proteome-pi: proteome isoelectric point database. *Nucleic acids research*, 45(D1):D1112–D1116, 2016.

[85] Fábio Madeira, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian Tivey, Simon C Potter, Robert D Finn, Rodrigo Lopez, et al. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 2019.

[86] Sangtae Kim, Nitin Gupta, and Pavel A Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of proteome research*, 7(8):3354–3363, 2008.

[87] Viktor Granholm, Sangtae Kim, José CF Navarro, Erik Sjolund, Richard D Smith, and Lukas Kall. Fast and accurate database searches with ms-gf+ percolator. *Journal of proteome research*, 13(2):890–897, 2013.

[88] Matthew C Chambers, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology*, 30(10):918, 2012.

[89] matrix Science. Data file format. `http://www.matrixscience.com/help/data_file_help.html`, 2019. [Online; accessed 4-October-2019].

[90] Pratik Jagtap, Jill Goslinga, Joel A Kooren, Thomas McGowan, Matthew S Wroblewski, Sean L Seymour, and Timothy J Griffin. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, 13(8):1352–1357, 2013.

[91] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, 7(01):29–34, 2007.

[92] James C Wright and Jyoti S Choudhary. Decoypyrat: fast non-redundant hybrid decoy sequence generation for large scale proteomics. *Journal of proteomics & bioinformatics*, 9(6):176, 2016.

[93] Michael F Lin, Irwin Jungreis, and Manolis Kellis. Phylocsf: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, 2011.

[94] Broad Institute. Phylocsf track hub. `https://data.broadinstitute.org/compbio1/PhyloCSFtracks/trackHub/hub.DOC.html`, 2019. [Online; accessed 4-October-2019].

[95] Mikael Hussius. gff-phylocsf-human. `https://github.com/hussius/gff-phylocsf-human`, 2019. [Online; accessed 4-October-2019].

[96] Daniel P Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, et al. Building proteometools based on a complete synthetic human proteome. *Nature methods*, 14(3):259, 2017.

[97] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, et al. Prosit. `https://www.proteomicsdb.org/prosit`, 2019. [Online; accessed 25-November-2019].

[98] Django Software Foundation. Django - the web framework for perfectionists with deadlines. `https://www.djangoproject.com/`, 2019. [Online; accessed 4-October-2019].

[99] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582, 2014.

[100] Manjunath Dammalli, Krishna R Murthy, Sneha M Pinto, Kalpana Babu Murthy, Raja Sekhar Nirujogi, Anil K Madugundu, Gourav Dey, Bipin Nair, Harsha Gowda, and Thottethodi Subrahmanya Keshava Prasad. Toward postgenomics ophthalmology: A proteomic map of the human choroid–retinal pigment epithelium tissue. *Omics: a journal of integrative biology*, 21(2):114–122, 2017.

[101] Eloy Cuadrado, Maartje van den Biggelaar, Sander de Kivit, Yi-yen Chen, Manon Slot, Ihsane Doubal, Alexander Meijer, Rene AW van Lier, Jannie Borst, and Derk Amsen. Proteomic analyses of human regulatory t cells reveal adaptations in signaling pathways that protect cellular identity. *Immunity*, 48(5):1046–1059, 2018.

[102] Christina Yingxian Chen, Matthew A Caporizzo, Kenneth Bedi, Alexia Vite, Alexey I Bogush, Patrick Robison, Julie G Heffler, Alex K Salomon, Neil A Kelly, Apoorva Babu, et al. Suppression of detyrosinated microtubules improves cardiomyocyte function in human heart failure. *Nature medicine*, 24(8):1225, 2018.

[103] Muhammad Ali, Shahid Y Khan, Shivakumar Vasanth, Mariya R Ahmed, Ruiqiang Chen, Chan Hyun Na, Jason J Thomson, Caihong Qiu, John D Gottsch, and S Amer Riazuddin. Generation and proteome profiling of pbmc-originated, ipsc-derived corneal endothelial cells. *Investigative ophthalmology & visual science*, 59(6):2437–2444, 2018.

[104] Ruedi Aebersold, Jeffrey N Agar, I Jonathan Amster, Mark S Baker, Carolyn R Bertozzi, Emily S Boja, Catherine E Costello, Benjamin F Cravatt, Catherine Fenselau, Benjamin A Garcia, et al. How many human proteoforms are there? *Nature chemical biology*, 14(3):206, 2018.

[105] Honglan Li, Yoon Sung Joh, Hyunwoo Kim, Eunok Paek, Sang-Won Lee, and Kyu-Baek Hwang. Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC genomics*, 17(13):1031, 2016.

[106] Sunghee Woo, Seong Won Cha, Gennifer Merrihew, Yupeng He, Natalie Castellana, Clark Guest, Michael MacCoss, and Vineet Bafna. Proteogenomic database construction driven from large scale rna-seq data. *Journal of proteome research*, 13(1):21–28, 2013.

[107] Nitin Gupta and Pavel A Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of proteome research*, 8(9):4173–4181, 2009.

[108] Zhe Ji, Ruisheng Song, Aviv Regev, and Kevin Struhl. Many lncrnas, 5'utrs, and pseudogenes are translated and some are likely to express functional proteins. *elife*, 4:e08890, 2015.

[109] NCBI Genes. Hnrnpa1 heterogeneous nuclear ribonucleoprotein a1. `https://www.ncbi.nlm.nih.gov/gene/3178`, 2020. [Online; accessed 11-January-2020].

[110] UniProtKB. Cold-inducible rna-binding protein. `https://www.uniprot.org/uniprot/Q14011#function`, 2020. [Online; accessed 11-January-2020].

[111] Thomas F Martinez, Qian Chu, Cynthia Donaldson, Dan Tan, Maxim N Shokhirev, and Alan Saghatelian. Accurate annotation of human protein-coding small open reading frames. *Nature Chemical Biology*, pages 1–11, 2019.

Supplement

Single-Linkage Clustering

The intergenic and intragenic distance distributions for different species across the kingdoms of eukaryotes are plotted in this section.

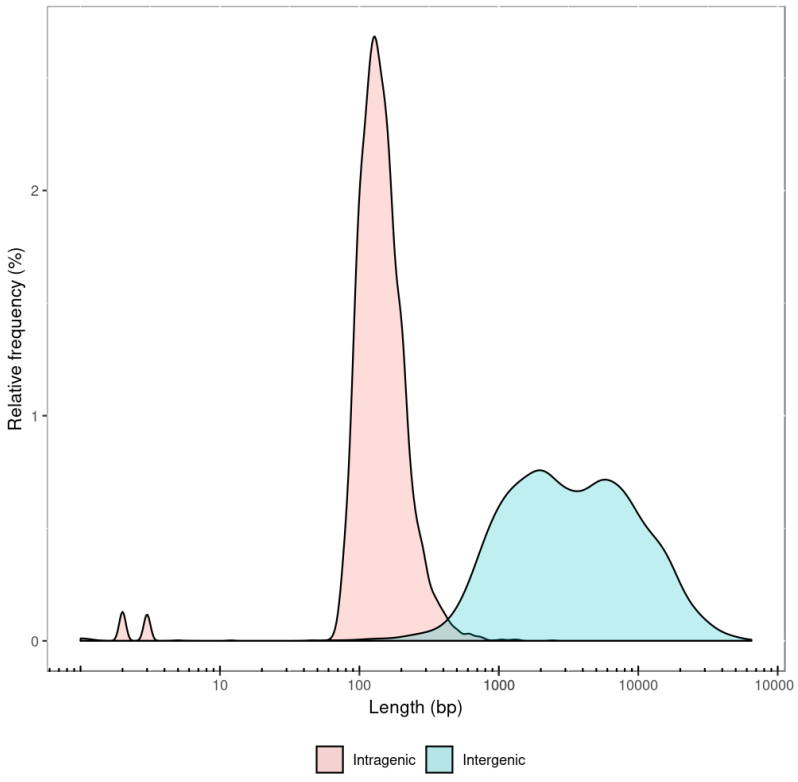


Figure S1: SLC distribution for the protist *Plasmodium falciparum*.

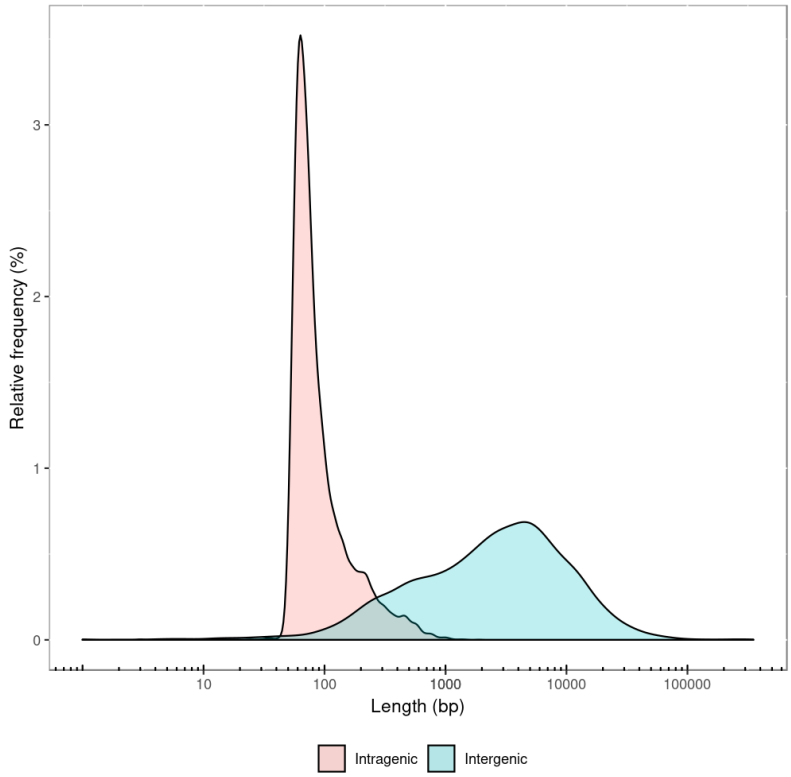


Figure S2: SLC distribution for the fungi *Neurospora crassa*.

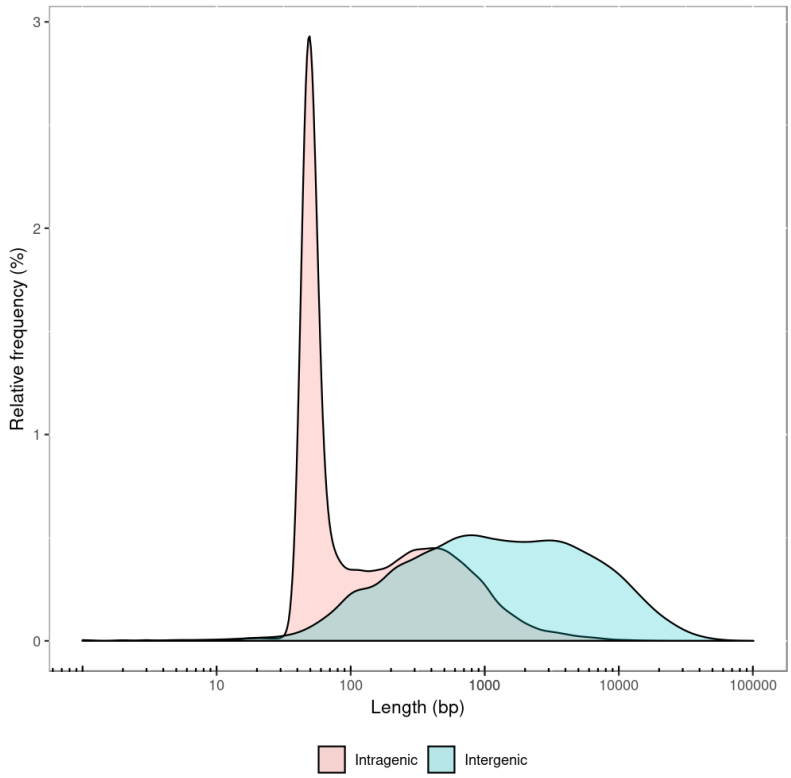


Figure S3: SLC distribution for the nematode *C. elegans*.

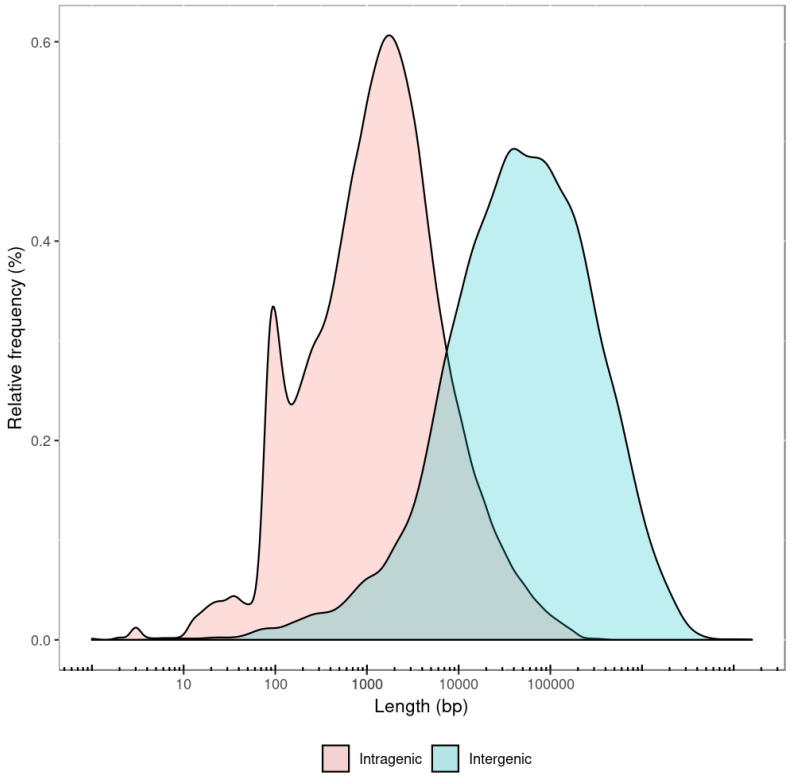


Figure S4: SLC distribution for the mammal *Sus scrofa*.

Web prototype: django model

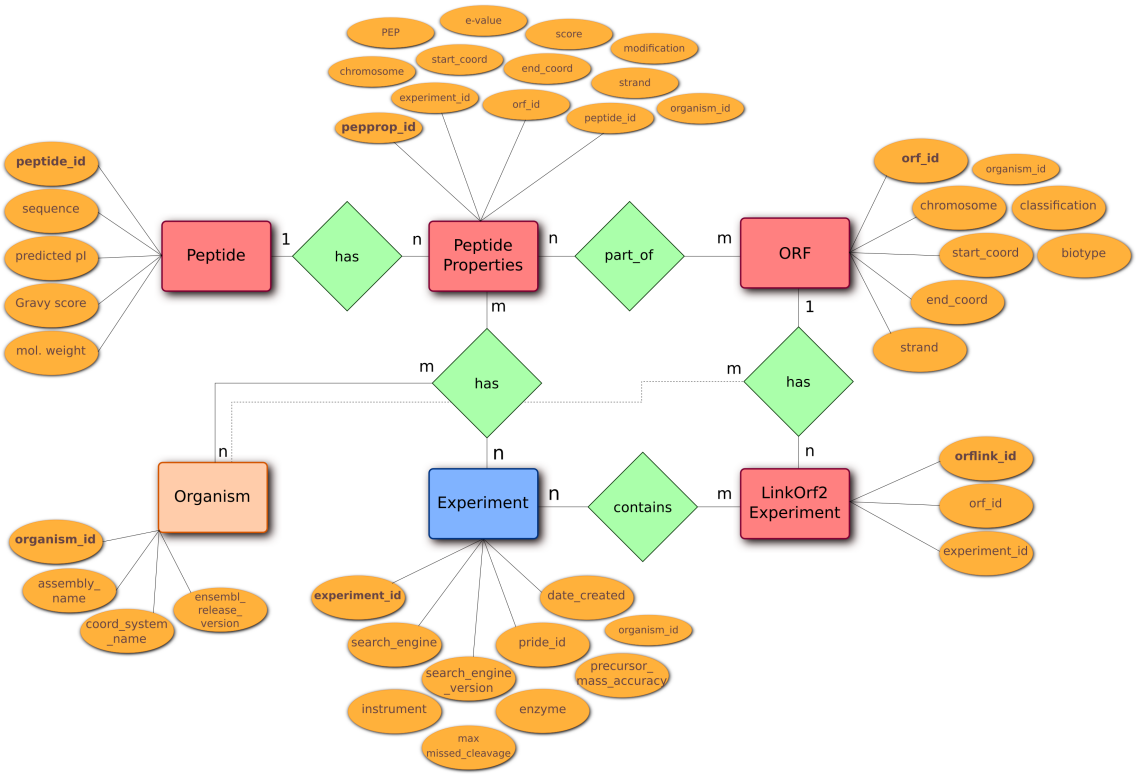


Figure S5: Entity relationship model of the django model. Each entity is represented as a class and the entity attributes are defined as class variables in a django model. An abstraction layer allows the developer to set foreign keys.

Proteogenomic examples: Alternative Frame

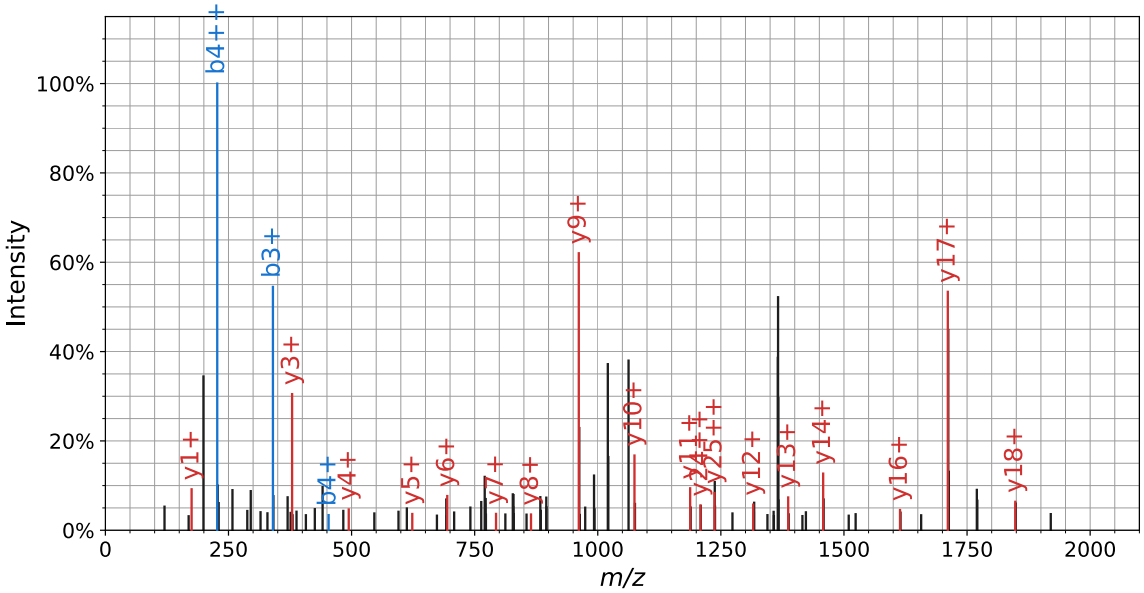


Figure S6: The annotated spectrum of the highest scoring peptide LLLLPCGPAEEAAH-PGVAAQLLPAVAEDGFR from Peptidome-DB. Prosit couldn't provide predicted fragment ion intensities because the peptide length exceeded the length cutoff of 30 amino acids.

Intergenic small Open Reading Frame example

The peptide DGALILELR was identified in the intergenic ORF at the locus on the reverse strand on chromosome 5:56865855-56865929. The neighboring genes are the lincRNA AC008937.1 gene which is 34kb upstream and the AC008937.2 lincRNA gene located 3.6kp downstream. An exon from the MAP3K1 gene is located on the opposite forward strand.

To exclude to possibility that the spectrum could have originated from a sequence that is very different to DGALILELR but has the a m/z value smaller than the set precursor mass tolerance of 20ppm compared to this experimental spectrum. BLAST was only used to remove novel peptides which had homologous sequences to Ensembl PEP and could therefore miss such a case. Two potential candidates were found with a similar experimental m/z value and Prosit was used to predicted their spectra which were used to compare them with the experimental spectrum. The first candidate MELVQVLK had no matching fragment ions under the used fragment ion mass tolerance of 5ppm for comparison and only two matching fragment ions under a relaxed tolerance of 20ppm, resulting in a low normalized spectral contrast angle of 0.36. In addition this peptides retention time (RT) was 55 minutes longer than the predicted RT from Prosit.

The second candidate IGQGDLIKR had only four matching main ions compared to the experimental spectrum (as opposed to 12 matching fragment ions by DGALILELR

which explain the whole spectrum). Even though this resulted in a high SA of 0.7, the RT was 105 minutes longer than the predicted RT by Prosit and could thus also be excluded to explain this spectrum better than DGALILELR, which had a different RT of 10 minutes compared to Prosit and a SA of 0.756, verifying its identification.

Listing 4.1: The ORF of the identified peptide from stop to stop codon.

```
gaa ttt gac atg cgt gaa aca aac ttc tgg atg gtg tta cga gat gga gct ttg atc ctt
 E   F   D   M   R   E   T   N   F   W   M   V   L   R   D   G   A   L   I   L
gag cta cgc cta ctg
 E   L   R   L   L
```

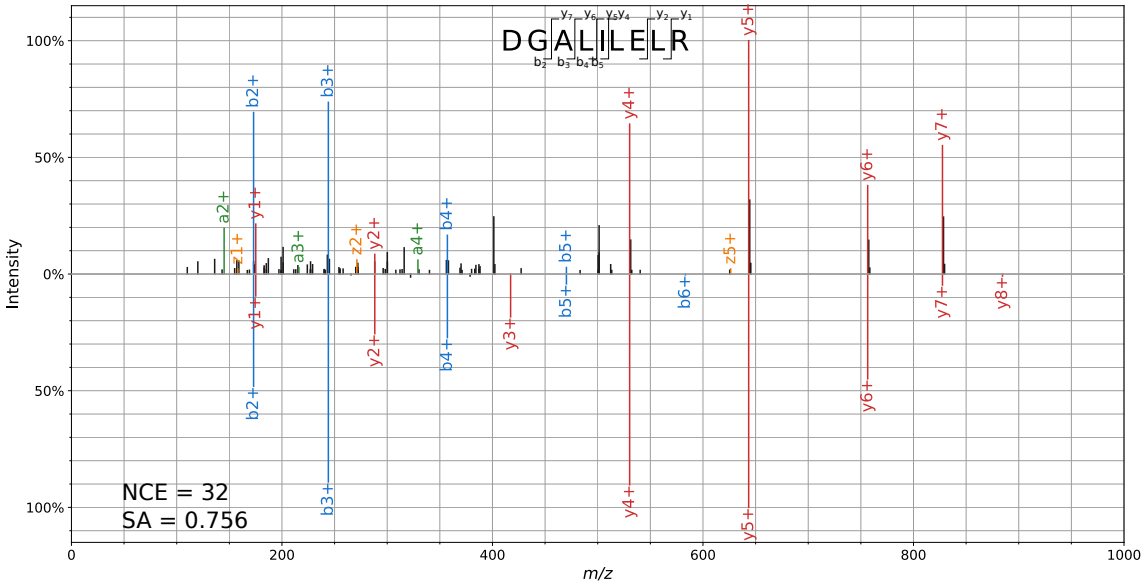


Figure S7: Mirror plot of the peptide DGALILELR. All observed main ions are very similar to the main ions of the predicted spectrum, the normalized spectral contrast angle shows a high similarity of 0.756, validating the experimental spectrum.