Bioinformatics methods for annotating genomes using proteomic data

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Gianluca Daniel Naguib.

Abstract

In recent years the number of genome sequencing projects has been exponentially increasing, leaving genome annotation dependent upon primarily automated tools. Recently, proteogenomics studies have attempted to bridge the gap between genomics and proteomics, by actively using proteomic data during the annotation stage. This project attempts to address some limitations in current bioinformatics approaches, such as the identification of N-terminal peptides and those spanning across exons – so called intron-spanning peptides (ISPs). Additionally it presents approaches for determining the quality of gene models. The results provide insights on the N-terminus of proteins (identification strategies, modifications), quality assessment on available gene annotation and performance of gene finders. A new method has also been developed for the identification of ISPs and, although this technique remains challenging, provides a framework in which future developments can be made.

Acknowledgements

The Biotechnology and Biological Science Research Council (BBSRC) that provided funding for this project.

I would like to thank my supervisor Andy Jones mentorship and the support from the members of my group that helped me throughout the challenges faced during this project.

I also want to thank my family, friends and Samantha Connor for supporting me through the hardships of this thesis.

Abbreviations

Hidden Markov model	HMM
Stretch of DNA on the same frame comprised between 2 stop codons	ORF_SS
Stretch of DNA on the same frame comprised between Methionine and a stop codon	ORF_MS
Short amino acid sequence	TAG
Generic feature format	GFF
Official gene model set	OGM
Alternative gene model generated with GlimmerHMM	GLM
Alternative gene model generated with GeneMark	gM
Alternative gene model generated with FgeneSH	FgSH
Panel of gene model (official and alternative)	P_GM
False discovery rate	FDR
Short amino acid sequence tag	TAG
Prefix Residue Mass	PRM
Intron spanning peptide	ISP
Whole genomic index based on 3 amino acid TAG	TAGdb
Coding sequences, stretches of DNA transcribed and translated	CDS

1 Introduction

1.1 The Biological framework

1.1.1 From DNA information to active proteins

Ever since the double helix structure of the Deoxyribonucleic acid (DNA) was elucidated in 1953 [1, 2] there have been enormous improvements in the understanding of the role of this molecule and how it relates to Ribonucleic acid (RNA) and proteins [3-5]. The basic structure of each strand of DNA is made by a long polymer of nucleotides linked together by a sugar molecule (2-deoxyribose) and phosphate molecules, which act as the backbone. This backbone, held together by asymmetric phosphodiester bonds, creates an inner direction of the polymer (5′ and 3′ prime ends, ending respectively with a phosphate and hydroxyl group). This leads to the assembly of a double helix, where opposite polymers are entwined together by interaction of the nucleobases (Guanine, Adenine, Thymine, Cytosine and Uracil [6-8]) as well as by hydrogen bonds. Each nucleobase from a strand binds specifically to a nucleobase of the opposite strand in a process called base pairing: Adenine (A) binds to Thymine (T), or Uracil (U) in RNAs, while Guanine (G) binds to Cytosine (C). In thermodynamic studies it has been observed that the stability of the double helix is directly affected by the G-C content, as their bond is stronger than A-T [9]. Unlike DNA, in the RNA chain, mostly single stranded [10], the alternative ribose molecule replaces the 2-deoxyribose sugar. Large DNA molecules (chromosomes) are made by millions of these repeated nucleotides; very large chromosomes can span from around $3x10^7$ to $9x10^7$ nucleotide repetitions such as chromosome 1 for *Homo sapiens* [11] and chromosome 3B of common wheat (*Triticum aestivum*) [12].

The information in DNA, or RNA, is carried by stretches of sequence called genes. Eukaryotic and prokaryotic cells differ both in size and structure of the genome and genes [13]. In the former, the structurally organized linear DNA and

DNA-bound proteins are assembled as chromosomes, contained within the cell nucleus. The genes are mostly represented by non-contiguous fragmented stretches of DNA sequence. In prokaryotic unicellular organisms the nucleus is absent and the DNA, of circular structure, has higher gene density and less fragmentation [14-17].

There are a large number of factors that influence gene expression such as non-coding RNA (transcriptional and translational), promoters and proteins. RNA-polymerases, which transcribe the gene sequences, also have the task of proofreading transcribed sequences preventing mutational errors [18-20]. The spliceosomes (large macromolecular complexes) remove the non-coding sequences located within the gene (introns) during the transcription from pre-RNA to messenger-RNA [21, 22]. As the intron is excised at sites 3′ and 5′, the expressed sequences (exons) are joined together [21, 23-25] (Figure 1:1).

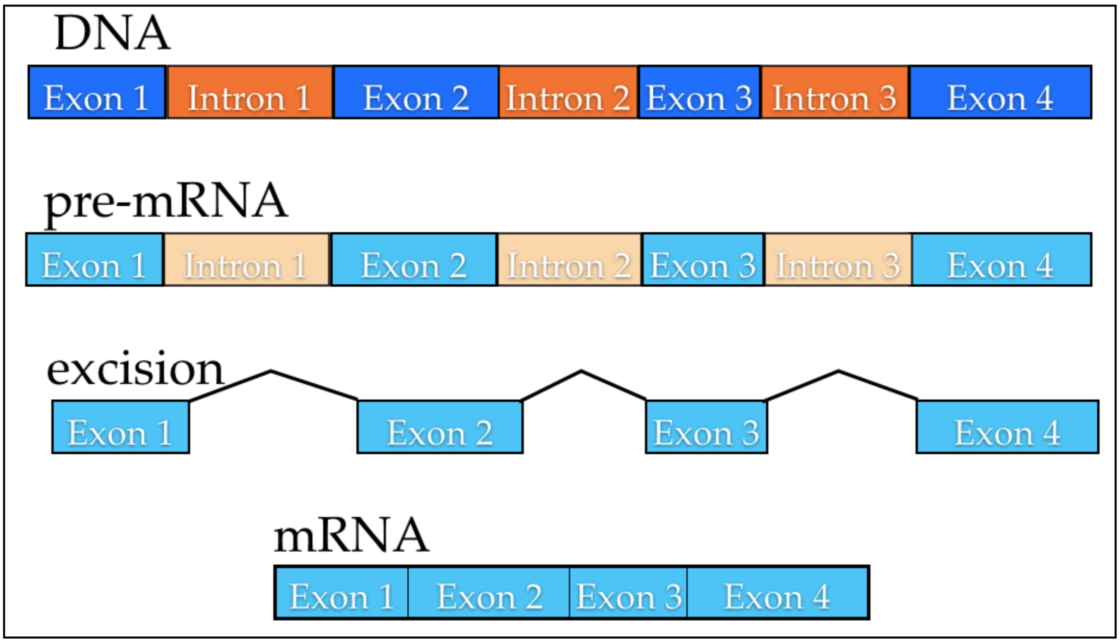


Figure 1:1 During transcription in eukaryotes, the non-coding introns regions are spliced out from the pre-mRNA and the resulting mRNA can then be translated by ribosomal RNA.

After the mRNA is translated by ribosomes (a macromolecule composed of RNA and proteins), the resulting protein can undergo several stable, unstable, reversible and non-reversible modifications. More than 200 different types of modifications have been found and they can be categorized into two groups depending *i*) whether they are formed by chemical linkage to the N or C terminal group or specific side chains, or ii) whether they are formed by processing, such as peptide segments removed by the protein chain (Figure 1:2).

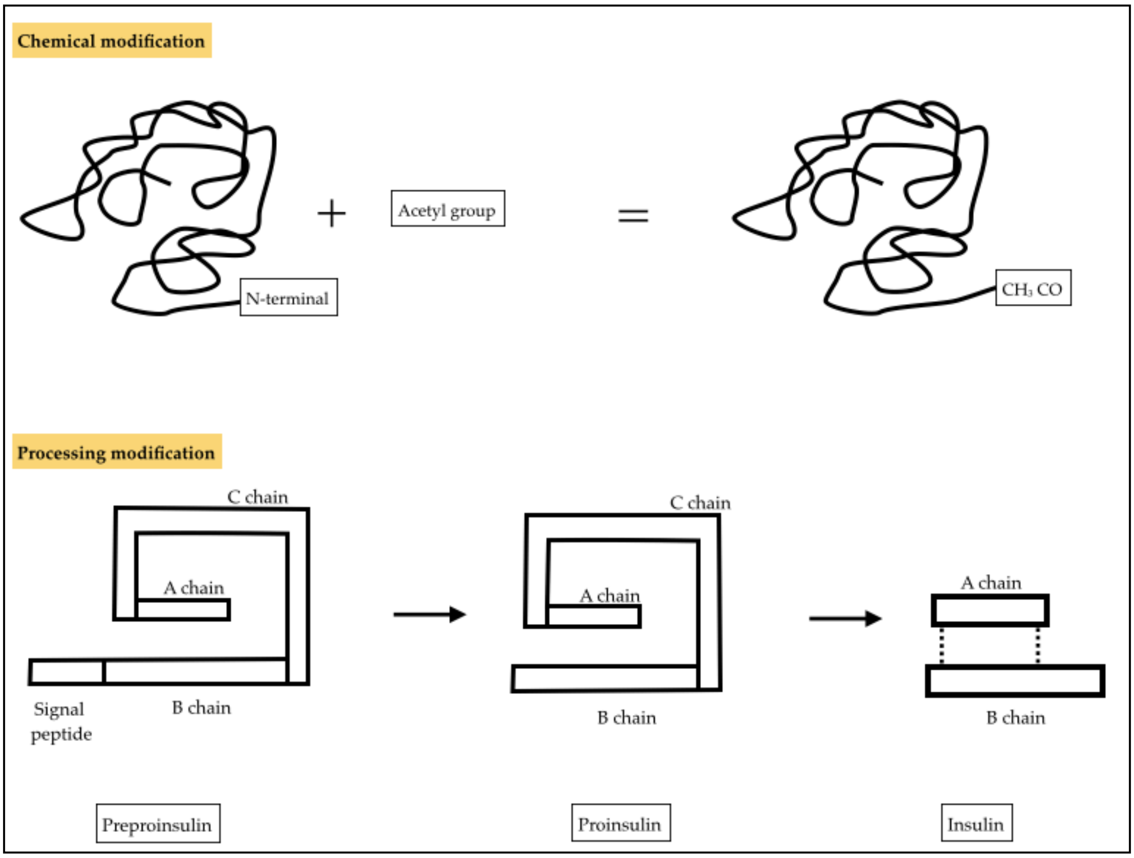


Figure 1:2 Examples of chemical and processing modifications occurring in proteins. The chemical modification displayed involves binding of an acetyl group to the N-terminus (N-terminal acetylation), catalysed by N-terminal acetyltransferase. The second modification shows the conversion of the precursor insulin protein preproinsulin into insulin. The signal peptide is cleaved after insertion in the endoplasmic reticulum; the protein then folds forming the C shape, proinsulin, allowing the A chain to bind to the B chain through disulphide bonds. The C chain is then cleaved leaving only the bound A-B chain, insulin, and thus activating the protein.

1.1.2 Gene models

Presently, for a large number of available raw genome sequences, the annotation of the regions encoding for proteins is performed with automated bioinformatic tools. In this thesis genome annotation is considered as determining the set of gene sequences predicted or confirmed for a given organism; the description of genes and their protein products represent instead the functional annotation. These are largely based on probabilistic interpretations and experimental data [26]. Widely available gene finder algorithms can be separated in two categories depending on which evidence is used for assessments: intrinsic and extrinsic.

In the first group can be found software packages such as Glimmer [27], Tigrscan [28], Genezilla [29], GeneMark [30, 31] and FGENE [32, 33]; these essentially attempt to predict the gene structure based on DNA pattern recognition such as AC-GT (splice and promoter motifs) and isochores (stretches of sequences with high GC content). In different studies isochores have been associated with both gene density and structure [26, 34-36]. Generally the Viterbi algorithm [37] and Markov models [38, 39] are used to rank the coding sequences. Here the automated version of the Markov model considers the unknown states that bring the resulting proteins, thus it is often referred to as a Hidden Markov model (HMM) [40]. Before the gene prediction stage can be carried out, the algorithms need to be tuned for the specific genomic sequence and this is performed with the generation of training sets. The HMM algorithms generally make these training sets by analysing the known data (i.e. confirmed gene sequences) on their genome sequence; then gene finder algorithms are able to evaluate the raw genomic sequence and rank gene structures based on these training sets [41]. Where high confidence data for training is not available, it is possible to use sets of Open Reading Frames (ORFs). These can be extracted from a six-frame translation of the genome, and are represented by long stretches of potentially translated genomic sequence located on the same frame. Throughout different studies these stretches can be comprised either between 2 stop codons or between a Methionine (start codon) and stop codon. As exons do not contain stop codons the ORF dataset can be seen as superset of the annotation, as each exon is contained within an ORF (Figure 1:3). The threshold length of computationally generated Open Reading Frames can be variable although the ORF datasets available on EupathDB online resource [42] are generated with Orf-Finder [43] (minimum length of 50 amino acids). In this particular study both types of ORFs are used: Methionine_Stop (ORF_MS) and Stop_Stop (ORF_SS).

The second group include software packages like Genscan [44, 45] and Genewise [46]; these incorporate sequence alignments [47] from related species (using ESTs, cDNAs, RNA) in their predictions. However the gene models obtained with this method rely on the accuracy of previously curated gene models (for other organism); therefore these generated models are based only on the high ranked homologs, as weak homologs with different conserved regions would result in lower accuracy at exonic level [48].

Other *ab initio* gene finders such as Evigan [49] and Augustus [50-52], also allow the inclusion of data from external sources in the form of proteomic evidence and gene predictions from other software packages. These can provide useful as gene model predictions can be re-evaluated based on peptide sequences identified from the samples.

Although in the recent years the use of transcriptome data has become popular thanks to new techniques such as RNASeq [53] only proteomic data provides definitive evidence that a protein product is made in the cell.

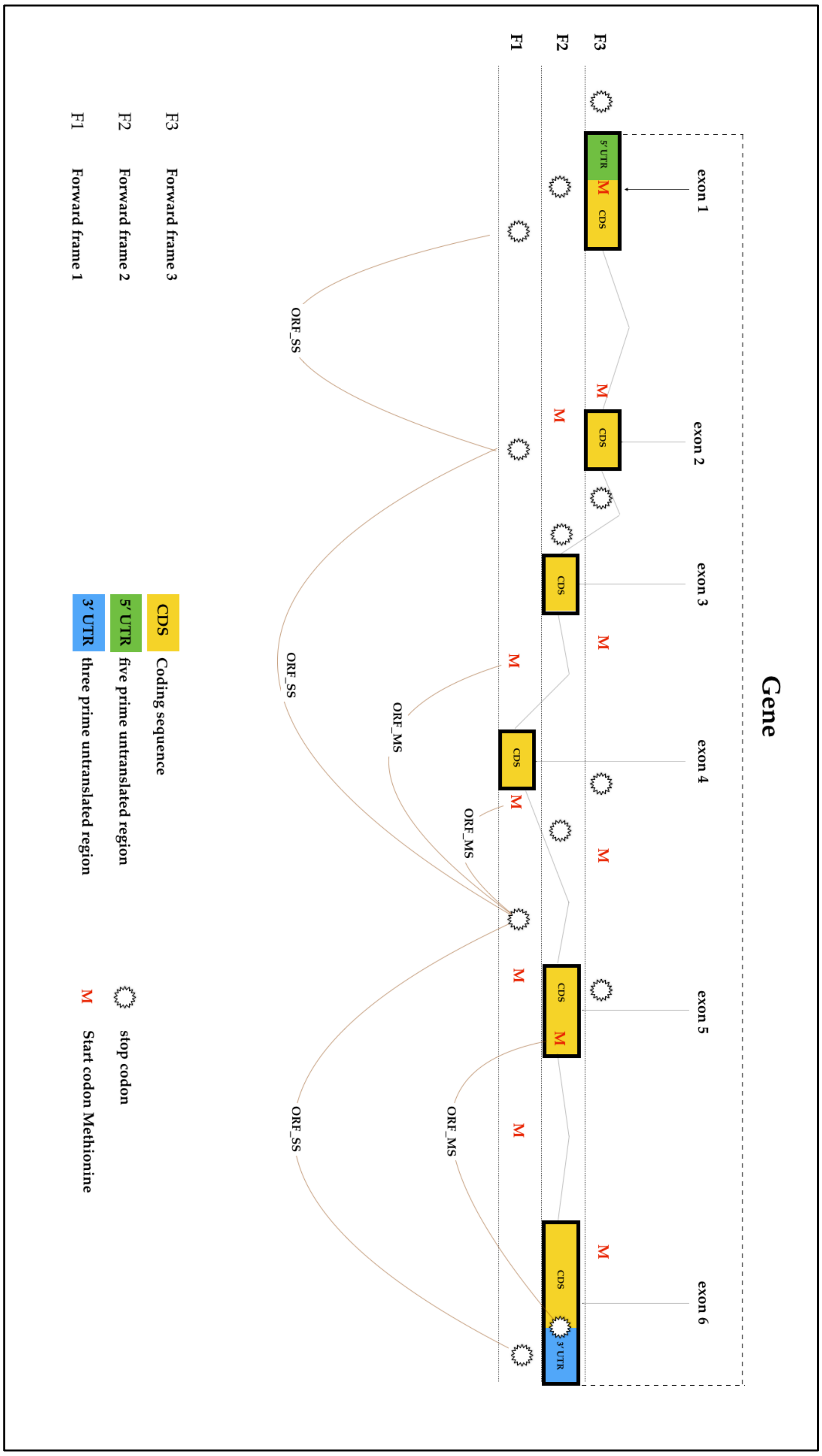


Figure 1:3 The above figure illustrates the complexity of a genomic region containing an example on gene structure and how informative open reading frames can be. To simplify the visualisation, only the forward strand is represented here. The gene is made of 5 exons that are located on different reading frame. The first and the last exons comprise both the coding sequence CDS, in yellow, as well as the five and three prime untranslated regions (5'UTR and 3'UTR) in green and blue respectively. The 5'-UTR precedes the first CDS that begins with the start codon Methionine (M or Met depending on amino acid nomenclature). Similarly the final CDS ends with a stop codon, followed by the 3'-UTR. The internal exons align only

with the CDS. The start and the stop codons are represented by M and star symbols respectively. As illustrated, the start codon can appear within the coding sequences other than the first CDS as it codes for a common amino acid. Differently the stop codon only encodes for the end of translation and as such it cannot be contained within the coding sequences; as such it can appear within the gene region on frames other than the current coding sequence frame (e.g. here the first stop codon in frame one and two are both within the first exon boundaries). The open reading frames can be selected as sequences comprised by a start codon Methionine and a stop codon (ORF_MS) or between two stop codons (ORF_SS) on the same frame. Here the first frame has three ORF_SS and four ORF_MS as each start to stop codon are considered as ORF_MS (second and fourth ORF_MS are subsets of the first and third ORF_MS respectively). Without setting a minimum length threshold for ORF_SS it would be possible to include all exons; instead even if selecting all the possible ORF_MS it is possible to miss exons (i.e. here exon three and the exon five).

1.2 Proteins and proteomics

1.2.1 Protein modifications

The 20 amino acids encoded by codons share a similar molecular structure. The α carbon atom is linked to the carboxyl group and the amine group; this structure is repeated for all amino acids, constituting their backbone. The α carbon also binds to a side chain, which makes up for the different properties of different amino acids. With the exception for Glycine (having only one hydrogen) the carbon structure of side chain (R chain) has Greek nomenclature (β, γ, δ, ε and ω). The type of R chain gives the amino acids its biophysical properties: isoelectric point, hydrophobic and hydrophilic as well as the capacity to bind specific elements [54]. The amino acids monomers are bound together by covalent reactions called peptide bonding during which the amino group from one monomer binds with the carboxyl group of another monomer (Figure 1:4).

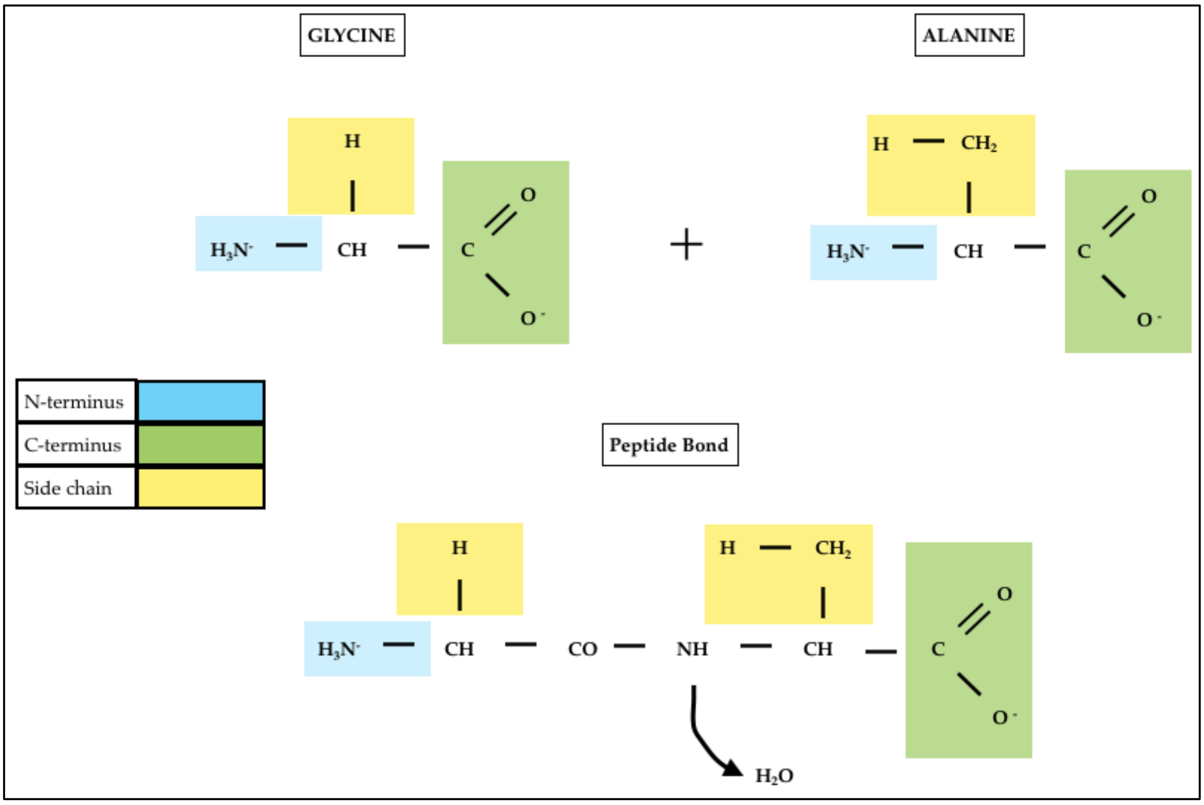


Figure 1:4 The basic structure of amino acid molecules, showing N- C- terminal groups and the side chain. Through peptide bonding (Glycine and Alanine) water is released, keeping the same N- C- structure.

Through the reaction, the hydrogen molecule released from the amino group forms a molecule of water with the hydrogen and oxygen molecules from the carboxyl group. This process, also called condensation, can be reversed through the addition of water (hydrolysis) [55]. This is maintained throughout the length of the monomer chain where one end presents the NH_2- amino group (N-terminal) while the other the COOH- carboxyl group (C-terminal). Peptides are short amino acids chains that are made of 2 or more amino acids (di-peptide, tripeptide and onwards). Single or multiple polypeptide chains make up the protein molecule [56]. As peptides are created the sequence is stabilised by hydrogen bonding, which leads it to the secondary structure with specific geometric shape. In the tertiary structure the different side chains interact in bonding, creating the specific molecular structure of the protein. The quaternary structure is given by multiple polypeptides binding together. The final translated protein can then go through a number of reversible and non-reversible modifications.

The signal peptide can be considered as a processing modification, which sees the amino terminal region of the protein cleaved off upon correct cellular localisation. Signal peptides are generally short sequences (~20-30 amino acid long in eukaryotes), positioned at the N-terminus, which serve as target signals within the cellular environment. Its structure comprises three regions: a positive charged n-terminus, a central hydrophobic region and a lightly polar c-terminal hydrophilic region [57, 58].

The signal peptide is present on secretory proteins and allows these to pass or attach to the ER membrane. The carboxyl region of signal peptides is detected by signal peptidase proteins, which perform the cleavage leading to the mature protein [59]. Different studies have investigated the role of the residue positions within the cleavage site, highlighting the importance of position -3 and -1 of the hydrophilic region as a pattern recognised by signal peptidases [60]. The most frequent residues at these positions have been identified as Alanine–x–Alanine [61].

The identification of signal peptides has been targeted with a number of computational tools (TargetP, SignalP, Signal-3L and Signal-CF [62-68]). These are generally based on a training set of known data and their algorithms include HMM models to identify secretory proteins and the cleavage site. Recent implementations have attempted to overcome the limitations in distinguishing between signal peptides and N-terminal helices of trans-membrane proteins (Phobius [69], Spoctopus[68], MEMSAT-SVM [70], SignalP4.0 [67]). In contrast to other software packages, SignalP4.0 is based on a neural network and works on two networks to assess the final score of cleavage positions (trans-membrane and non trans-membrane networks). However confident the predictions of signal peptides are, it remains a challenge to confirm these through proteomic analysis.

Another frequent modification involved at the N-terminus is the excision of a peptide sequence or precise amino acids. N-terminal methionine excision (NME) seems to be present in high proportion of proteins and during the excision process it appears that the amino acids following methionine may affect whether the cleavage takes place [71]. Other cleavages simply allow the protein to change state until needed (e.g. signal peptide cleavage and self-splicing synthesize insulin protein from its preproinsulin precursor, Figure 1:2).

Through reversible and non-reversible chemical modifications protein activity can adapt and respond to external stimuli. Important for energy transfer and signalling within the cell, phosphorylation (a phosphate group bound to Serine, Threonine, Tyrosine and Histidine by protein kinases) can be frequently observed within the proteome, and for each protein this multi-site modification can be potentially performed and reversed several times [72]. Equally important is the reversible modification involving the addition of the acetyl group to the N-terminus (by N-terminal acetyltransferase proteins), which affects gene regulation and is also frequent [73-78].

Other common modifications include glycosylation [79-81] (oligosaccharide group added to secretory/membrane proteins), methylation [82-84] (chemical link on side chain specific residues Lysine and Arginine [85]) and ubiquitination [86-88] (protein degradation performed by ubiquitin activated enzymes).

The large number of all possible modifications for each protein on the whole proteome scale makes protein prediction and the annotation process increasingly challenging.

1.2.2 Mass spectrometry proteomic workflow

Mass spectrometry (MS) based proteomic studies (Figure 1:5) offer a way to annotate the genes by allowing the annotator to validate the presence of the proteins in a sample [89] in a quantitative or qualitative manner [85].

MS-based proteomics evaluation consists of measuring, with mass analyser instruments, the mass-to-charge ratio versus abundance of ionized peptides;

these can be further fragmented into smaller ions during a two-stage analysis in a Tandem Mass Spectrometer (MS/MS).

The peptides can either be obtained by separation of a digested protein mixture, as in Multidimensional Protein Identification Technology (MudPIT) [90, 91], or for example by in-gel digestion after electrophoretic separation of proteins [92]. This approach for protein identification is also known as bottom-up proteomics. In order to draft the peptide results to list proteins as identified (procedure also referred to as protein inference) it is necessary to assemble these resulting peptides and assess their confidence. To draft this list of peptides considered as reliable one must also consider whether the peptide is contained in multiple protein sequences, or whether only one peptide per protein was identified (problem known as "one-hit-wonder") [93]. In this stud the protein inference step y makes use of these considerations and as such the proteins here identified must contain one unique peptide to be considered confident identifications.

However as statistical methods are used to computationally assess the significance of spectral interpretations, it becomes challenging to validate the correctness of the interpreted sequence within very large datasets [94]. With generally adopted database search approaches the statistical importance of the identifications are given as the probability that each one is incorrect such as p-value estimate. This can be seen as the probability distribution that a null hypothesis is true, such as that a peptide spectrum match (PSM) with the same or better score happens by chance in a sequence database [95, 96].

MS top-down proteomics is a different approach as undigested proteins are analysed with MS; this can be extremely useful to provide precise protein mass, which, in turn, can lead to detailed information regarding possible PTMs. However this approach is still limited and not adept at the annotation process itself because of high resolving power and high mass range needed to perform the analysis. This approach also relies on accurate annotations, which are often not available [97].

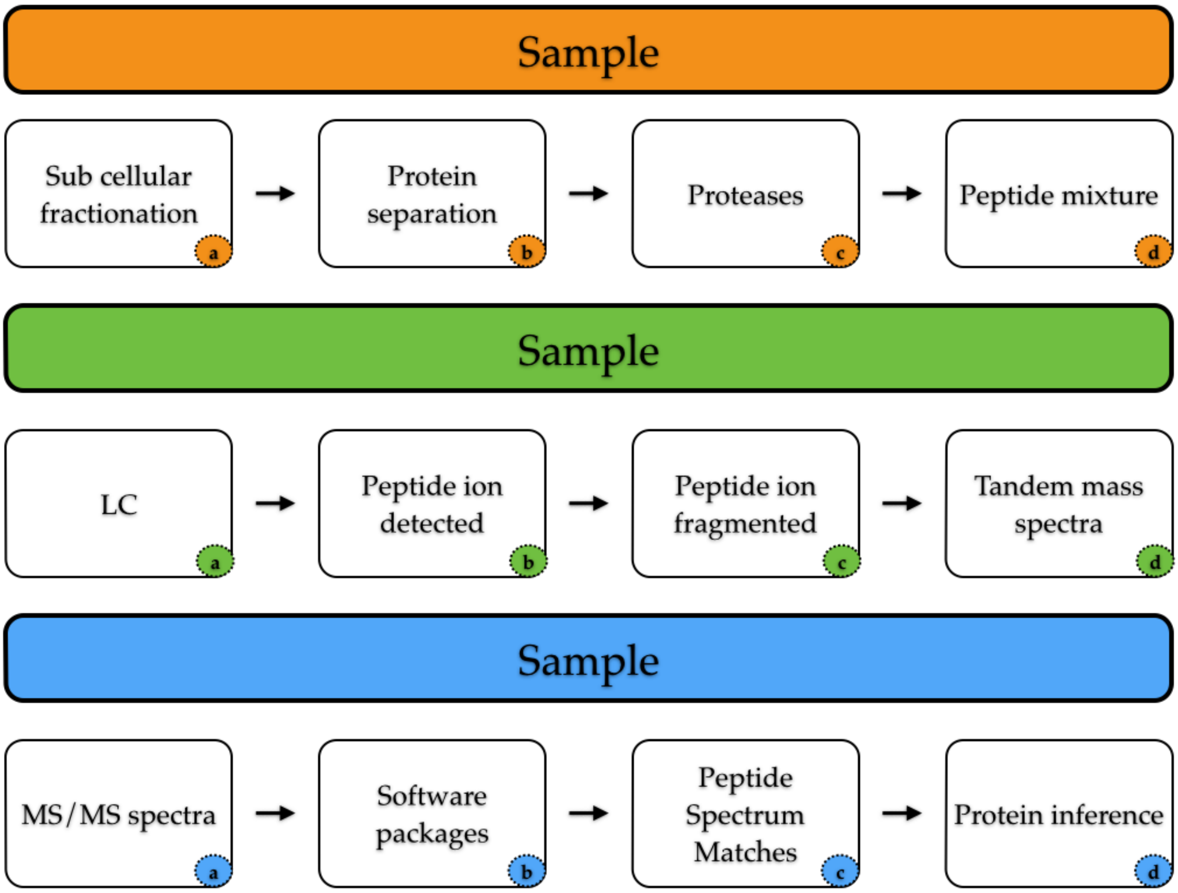


Figure 1:5 Basic proteomic workflow divided in 3 main stages: biological sample preparation (orange) where the sample goes through subcellular fractionation (a), protein separation (b) followed by proteases (c), resulting in peptide mixtures (d). During mass spectrometry stage (green) these peptide mixtures are further separated through the Liquid Chromatography (a); ionised peptides are detected (b) and selected peptide ions are fragmented (c) resulting in tandem mass spectra. Finally, during bioinformatic analyses (blue), the obtained spectral data (a) is processed with (b) specific software packages that identify (c) the peptide spectrum matches (PSMs) and provide protein inference (d).

1.2.3 Protein and peptide separation

Using the properties of different molecular weight (MW) and electric charge of the molecules can attain the separation of polypeptides. One-dimensional electrophoresis (1-DE) separates polypeptides by their mass and although it can efficiently separate a large amount of proteins, generally fails to separate polypeptides completely in complex mixtures [98].

The process of 2-DE addresses this limitation by separating the molecules first by their net charge, isoelectric focusing (IEF), and then by their MW. This separation over 2 different axes achieves a higher resolution [92, 99]. This process has clear advantages like resolving thousands of proteins at the same time on a single 2-DE gel as well as separating/ revealing proteins with PTMs. There are some limitations associated with this method as proteins of very large or small size. Also hydrophobic proteins such as membrane proteins are generally difficult to observe in 2-DE gels. It may be difficult to detect low abundance proteins with conventional staining although protein pre-fractionation preceding 2-DE can simplify protein mixture [100]. Figure 1:6 taken from Xia et al. [101] paper, shows a 2-DE gel with localized spots.

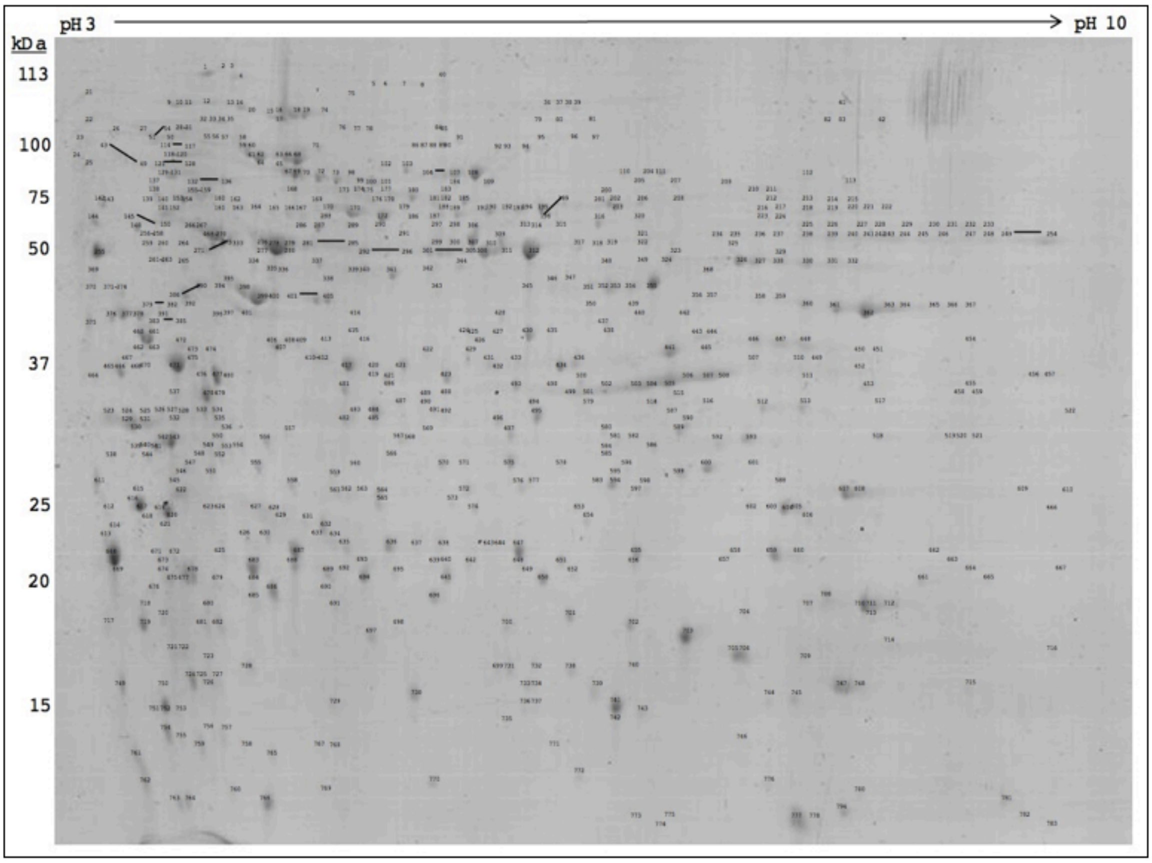


Figure 1:6 Example of a 2DE gel electrophoresis. This has been obtained from *T.gondii* for proteomic annotation in the Xia. *et al* study [101].

Following protein separation proteolysis is performed to digest proteins into peptides, for example using trypsin. Trypsin cleaves the peptide bond at the carboxyl end of arginine and lysine amino acids as long as these are not followed by proline [102].

Prior to mass spectrometry, High Performance Liquid Chromatography (HPLC [103]) is often performed, which makes use of the different hydrophobicity of the peptides to perform the separation.

1.2.4 Tandem Mass Spectrometry

The prepared samples are then ready to be analysed in the Mass Spectrometer and the analysis is performed in 3 phases (Figure 1:5):

- the samples are firstly ionized (using for example MALDI (described below) [104] or Electrospray [103]);
- the ions are then separated (Time-of-flight, Ion trap, Triple quadrupole [105]);
- and finally the ions are detected;

In ion trap mass analysers such as Quadrupole ion trap [106, 107] and Fourier transform ion cyclotron resonance [108] (FT-ICR) molecule detection is accomplished by trapping, within electric or magnetic fields in high vacuum cells, the desired ions of specific mass. The ions are then given a small electronvolts potential that send them towards the detecting end; or as for FT-ICR the ion detection is accomplished by the frequency of the oscillating ions. Ion trap mass analysers have generally higher sensitivity and offer faster scan rates than older quadrupole mass filter analysers [109]. FT-ICR offers high mass accuracy and high resolution, thanks to its strong magnetic field. Similarly to FT-ICR the Orbitrap mass analyser traps the selected peptide ions at specific frequencies; however the oscillating ions revolve around an electrode instead of in a magnetic field [110, 111]. Time of flight mass analyser represent the simplest type of instrument, where accelerated ions travel across vacuum tube to the detector. As the mass affects the molecule velocity, it is calculated based on the time of arrival [106, 112, 113]. The output is a mass spectrum showing the mass-per-charge of the ion series and their intensities.

Peptide ionisation is mainly performed with Matrix Assisted Laser Desorption Ionisation (MALDI [104]) and electrospray ionisation (ESI [103]). The first technique uses a laser and a light absorbing matrix to desorb and ionise the molecules, generating mainly singly/ doubly charged ions. The second technique involves molecules going through a capillary tube contained within an electric field, generating mainly multiply charged ions. By identifying C^{-12} peaks and using mono-isotopic masses it is possible to calculate the charge of the ion and peptide mass.

These peptide ions, confined by an electric field, are separated and detected based on their mass-per-charge *(m/z)* ratio during the MS1 pass. Then the most abundant ion species can be selected to go through further fragmentation into smaller ions (in data dependent acquisition instruments), which are then detected in the MS2 stage, from which peptide sequence information can be derived (Figure 1:5).

The peptide fragmentation can be obtained with different techniques, such as Collisional Induced/Activation Dissociation (CID, CAD [114]) or Electron Transfer Dissociation (ETD) and Electron Capture Dissociation (ECD [115]). Each technique produces predominant ion series (Figure 1:7) from peptide bond fragmentation (i.e. CAD leads to -*b* and -*y* ions while ETD and ECD lead to -*c* and -*z* ions – see below for ion terminology). This makes these different techniques complementary [116] when combined and can produce more information in the output spectra, which can improve the computational interpretation of the spectra. As an example in Figure 1:5 the parent ions have their molecular ions accelerated by an electric field and collide with neutral gas such as helium or argon; then due to the acquired high kinetic energy the collision with the gas breaks the internal bonds making the molecular fragments into smaller fragment ions (CID, CAD).

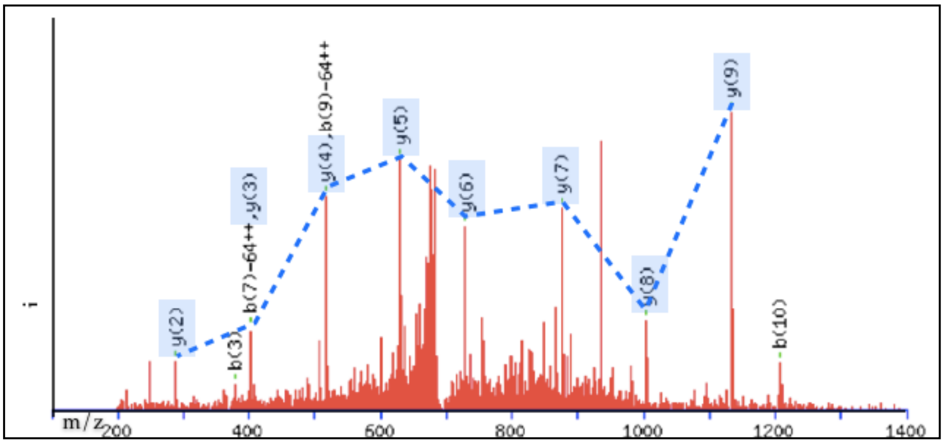
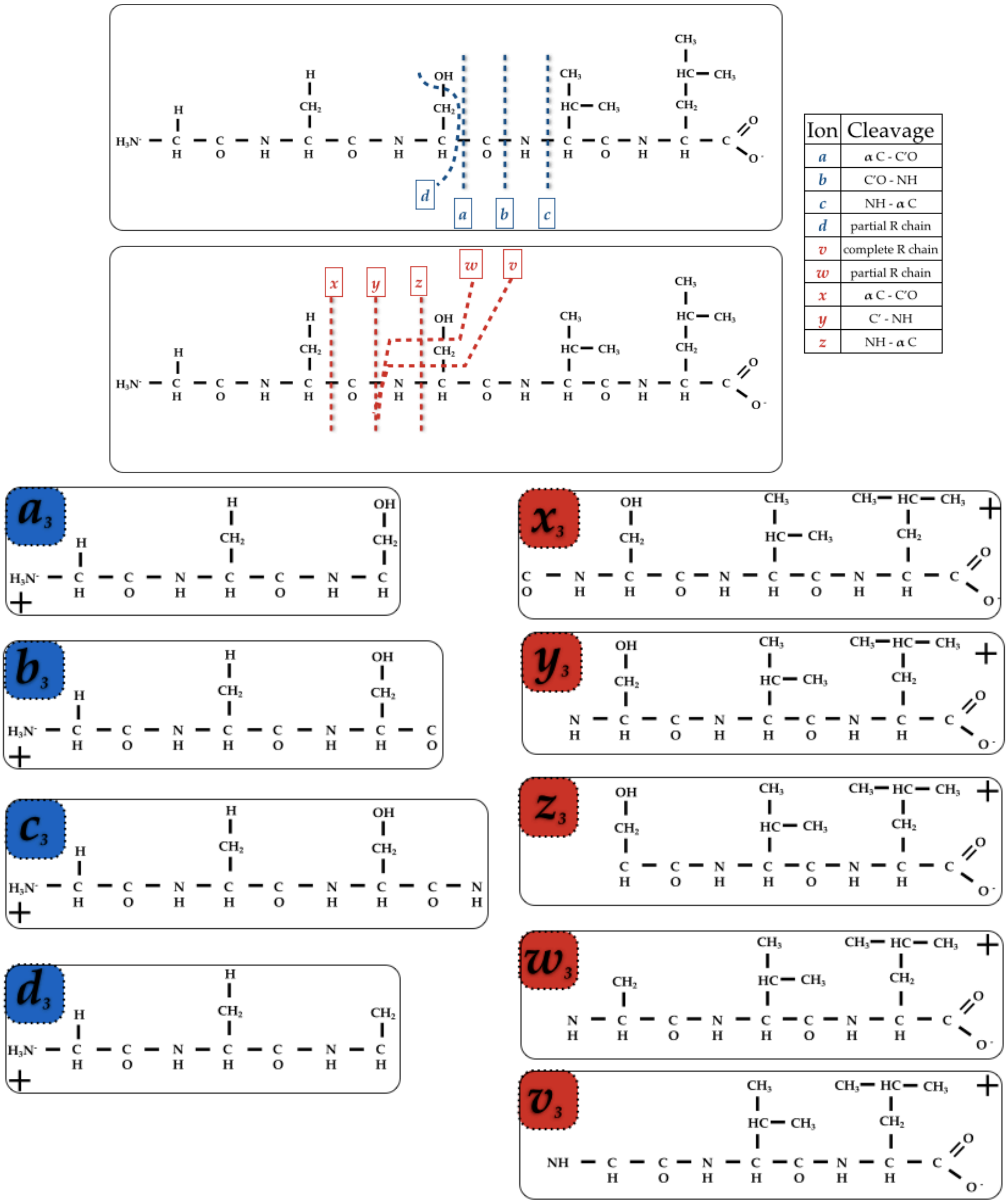


Figure 1:7 Tandem MS spectrum from a peptide, TMEEFVIDLLR, identified by the *y*- ion ladder (MASCOT).

If the fragments do not carry any charge they would simply not be detected. Fragment ions (Figure 1:9) resulting from backbone cleavage either at alpha C-CH, C-N or N-αC are usually indicated with *a*, *b* or *c* if the charge is retained on the N-terminus; instead if the charge is retained on the C-terminus they are indicated with *x*, *y* or *z* (with a subscript indicating how many residues in the fragment). Also present, but not commonly observed as high CID is required, are *d*- *v*- and *w*- ions, resulting from side-chain cleavages.

The MS2 spectrum, shown in Figure 1:7, can be seen as a graph where the Y-axis is the intensity while the X-axis is the mass over charge *m/z* (sometimes reported in Thompson units, *Th*). On the spectra the peaks correspond to the intensity (the tallest is normally labelled the base peak) while the X-axis shows the mass of the molecule divided by charge (if the graph shows a peak at 728.4665 m/z and it is doubly charged, then the MW of the peptide analysed would correspond to 1454.917 Dalton).



Ion	Cleavage
a	α C - C'O
b	C'O - NH
c	NH - α C
d	partial R chain
v	complete R chain
w	partial R chain
x	α C - C'O
y	C' - NH
z	NH - α C

Figure 1:8 Different ions species fragment differently (N- or C- terminus) as illustrated in the cleavage table. The position of the additional proton produces either N-terminal ions (a-, b-, c-, d-) in blue, or C-terminal ions (v-, w-, y-, x-, z-), in red. High energy CID causes side chain fragmentations, partial in d- and w- ions and complete in the v- ion. The example is obtained from residues GASVL.

1.3 Bioinformatics approaches

In bottom-up mass spectrometry proteomic experiments the peptide sequences can be interpreted by different bioinformatics approaches (Figure 1:9). In sequence database searches experimental mass spectra are compared against theoretical fragmented spectra generated by computationally digested protein sequences or six-frame translations [117-123]; the result is a statistically significant identification for a proportion of the collected mass spectra. *de novo* sequencing instead attempts, through probabilistic networks and complex algorithms, to interpret the peptide sequence yielding the spectrum with no previous knowledge of the sequence [124-126]. Hybrid methods combine *de novo* and sequence database search in two steps: initially, computational algorithms provide sets of short amino acid, called TAGs, yielding adjacent peaks in the spectrum, then with this information the algorithms filter the sequence databases and perform database searches on a generated, restricted database [127-130]. In the spectral library search approach, unknown spectra are identified through comparison against compiled database of spectra. In this case the database/ library holds a large collection of observed high-quality spectra with identified peptide sequence [131, 132].

All methods have some limitations: database dependent approaches are limited by the accuracy of the sequence available for searches; *de novo* approaches lack high confidence identification for long peptides [94, 133].

At present, algorithms used in search engine database software packages make use of statistical tools to score comparisons between theoretical and experimental tandem mass spectra. All candidate peptide sequences are weighted based on the estimated probability of random peptide sequences generating MS/MS spectra of the same or higher similarity by chance [134]. This can have repercussions on sensitivity and accuracy of peptide spectrum matches (PSMs) where the scoring system might lead to differences in results for searches of small sequence database versus large ones (i.e. high-quality gene models vs. six frame translation).

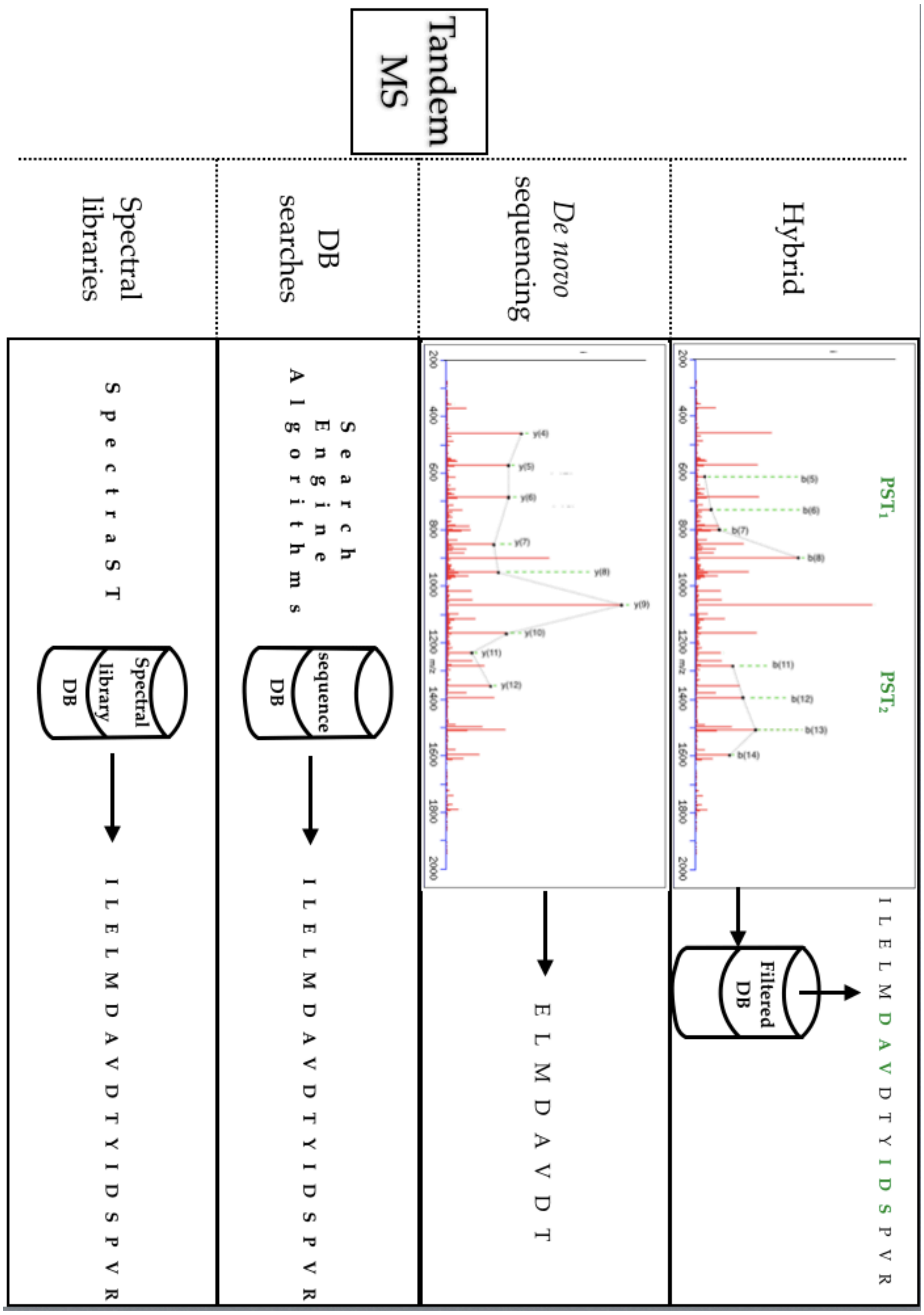


Figure 1:9 Tandem MS identification, in red, through four main bioinformatic approaches, from the top left clockwise: hybrid approaches generates peptide spectrum tags (PSTs, highlighted in green) to filter sequence databases before running the database search; *de novo* sequencing attempts to identify the full length peptide sequence using only the spectrum. With spectral libraries, comprising large sets of high quality spectra of known compounds, allow direct raw spectrum identification. In common database dependent approaches, algorithms computationally digest protein sequences and fragment peptides that are then

used for identification.

1.3.1 Database dependant approaches

The workflow for the first type of approach, typified by the MASCOT search engine [122], is as follows. MASCOT allows both Peptide Mass Fingerprint (PMF [135]) as well as an "MS/MS Ion Search". PMF is used to identify proteins by matching their constituent peptides masses (MS1 only) to the theoretical peptide masses generated from a protein database. The PMF identifications rely on observing a large number of peptides from the same protein at high mass accuracy. It is better used together with 2-DE data where proteins are generally separated into simple mixtures [136].

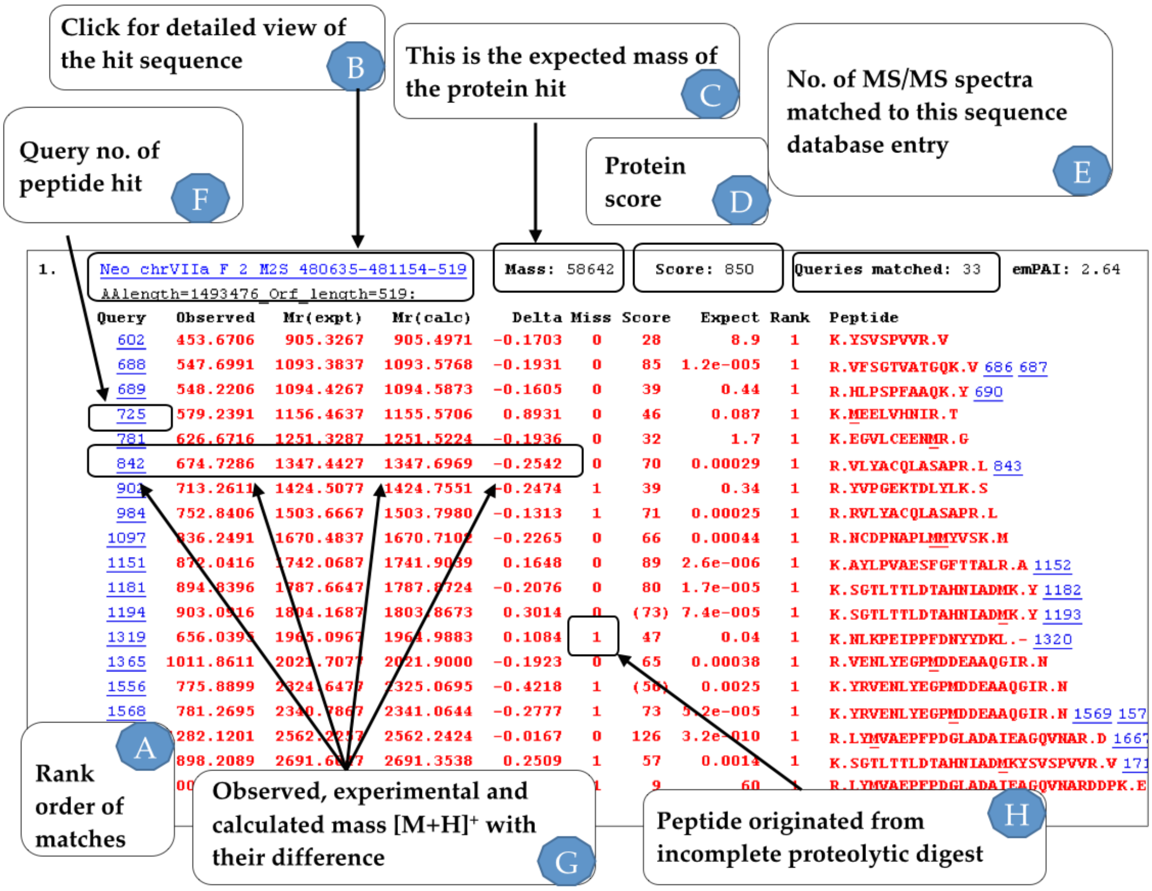
The MS/MS technique can be used to identify a protein from even a single peptide, even though the quality of the result will increase by searching an MS/MS run containing several peptides for a given protein. In the MS/MS search the experimental spectra are used to gather information about the precursor ion masses. In Sequest [119, 123, 137, 138] both experimental and theoretical spectrum are pre-processed where normalization of signal intensities allow the spectra to be comparable. In OMSSA [120] there is no normalization process but instead noise removal.

The sequence database is processed and the search engine generates a set of theoretical spectra for all digested peptide sequences, whose mass falls within the mass tolerance. It then proceeds by fragmenting each sequence in this temporary set by recreating *in silico* fragmentation for each ion (-*a*, -*b*, -*c*, -*x*, -*y*, -*z*). Finally it tries to match each expected value against the experimental value in the original spectrum. Each ion is given a score and all positive matches, within the tolerance window, between the theoretical spectrum and experimental are summed in the final score of the reconstructed the peptide sequence.

The X!Tandem software search engine further analyses a compiled list of high confidence peptides in order to search for modified and non-enzymatic peptides within the protein result [118, 124, 139, 140]. Similarly the Phenyx search engine, commercial software package distributed by GeneBio, Geneva Bioinformatics SA, has a two-step analysis for searching combinatorial modifications and an algorithm to evaluate which of the alternative matches is most probable [121, 141, 142].

MASCOT provides a statistical weighting for each individual PSM, based on the quality of the match between experimental and theoretical spectrum. This probabilistic approach is based on the MOWSE (for MOlecular Weight Search) algorithm [122, 143].

With the protein score in PMF and ion score in MS/MS ion search MASCOT provides an expectation value that corresponds to the frequency of matches having equal or better score that could be obtained by random match. The MASCOT web-interface allows the user to view the search results with scalable level of details. It also provides the tabular form to map the identified peptide sequence to the query it has been matched (from Figure 1:10 to Figure 1:12).



```
1.   Neo chrVIIa F 2 M2S 480635-481154-519        Mass: 58642     Score: 850    Queries matched: 33   emPAI: 2.64
     AAlength=1493476_Orf_length=519:
     Query  Observed   Mr(expt)   Mr(calc)     Delta Miss Score   Expect Rank  Peptide
       602  453.6706   905.3267   905.4971   -0.1703   0    28       8.9    1   K.YSVSPVVR.V
       688  547.6991  1093.3837  1093.5768   -0.1931   0    85  1.2e-005    1   R.VFSGTVATGQK.V 686 687
       689  548.2206  1094.4267  1094.5873   -0.1605   0    39      0.44    1   R.HLPSPFAAQK.Y 690
       725  579.2391  1156.4637  1155.5706    0.8931   0    46     0.087    1   K.MEELVHNIR.T
       781  626.6716  1251.3287  1251.5224   -0.1936   0    32       1.7    1   K.EGVLCEENMR.G
       842  674.7286  1347.4427  1347.6969   -0.2542   0    70   0.00029    1   R.VLYACQLASAPR.L 843
       902  713.2611  1424.5077  1424.7551   -0.2474   1    39      0.34    1   R.YVPGEKTDLYLK.S
       984  752.8406  1503.6667  1503.7980   -0.1313   1    71   0.00025    1   R.RVLYACQLASAPR.L
      1097  836.2491  1670.4837  1670.7102   -0.2265   0    66   0.00044    1   R.NCDPNAPLMMYVSK.M
      1151  872.0416  1742.0687  1741.9039    0.1648   0    89  2.6e-006    1   K.AYLPVAESFGFTTALR.A 1152
      1181  894.8396  1787.6647  1787.8724   -0.2076   0    80  1.7e-005    1   K.SGTLTTLDTAHNIADMK.Y 1182
      1194  903.0916  1804.1687  1803.8673    0.3014   0  (73)  7.4e-005    1   K.SGTLTTLDTAHNIADMK.Y 1193
      1319  656.0395  1965.0967  1964.9883    0.1084   1    47      0.04    1   K.NLKPEIPPFDNYYDKL.- 1320
      1365 1011.8611  2021.7077  2021.9000   -0.1923   0    65   0.00038    1   R.VENLYEGPMDDEAAQGIR.N
      1556  775.8899  2324.6477  2325.0695   -0.4218   1  (56)    0.0025    1   K.YRVENLYEGPMDDEAAQGIR.N
      1568  781.2695  2340.7867  2341.0644   -0.2777   1    73  5.2e-005    1   K.YRVENLYEGPMDDEAAQGIR.N 1569 157
            1282.1201  2562.2257  2562.2424   -0.0167   0   126  3.2e-010    1   R.LYMVAEPFPDGLADAIEAGQVNAR.D 1667
            898.2089  2691.6047  2691.3538    0.2509   1    57    0.0014    1   K.SGTLTTLDTAHNIADMKYSVSPVVR.V 171
            00                                         1     9        60    1   R.LYMVAEPFPDGLADAIEAGQVNARDDPK.E
```

Click for detailed view of the hit sequence
B
This is the expected mass of the protein hit
C
No. of MS/MS spectra matched to this sequence database entry
E
Query no. of peptide hit
F
Protein score
D
Rank order of matches
A
Observed, experimental and calculated mass [M+H]+ with their difference
G
Peptide originated from incomplete proteolytic digest
H

Figure 1:10 MASCOT result summary view:

A: Rank of identified protein sequences by the sum of their peptide score;

B: This is the accession number of the sequence (computationally predicted Open Reading Frame in this case) and the peptides matched against. It leads to another page containing the full sequence and showing how peptides align to it.

C: Expected mass of the protein/sequence in Daltons.

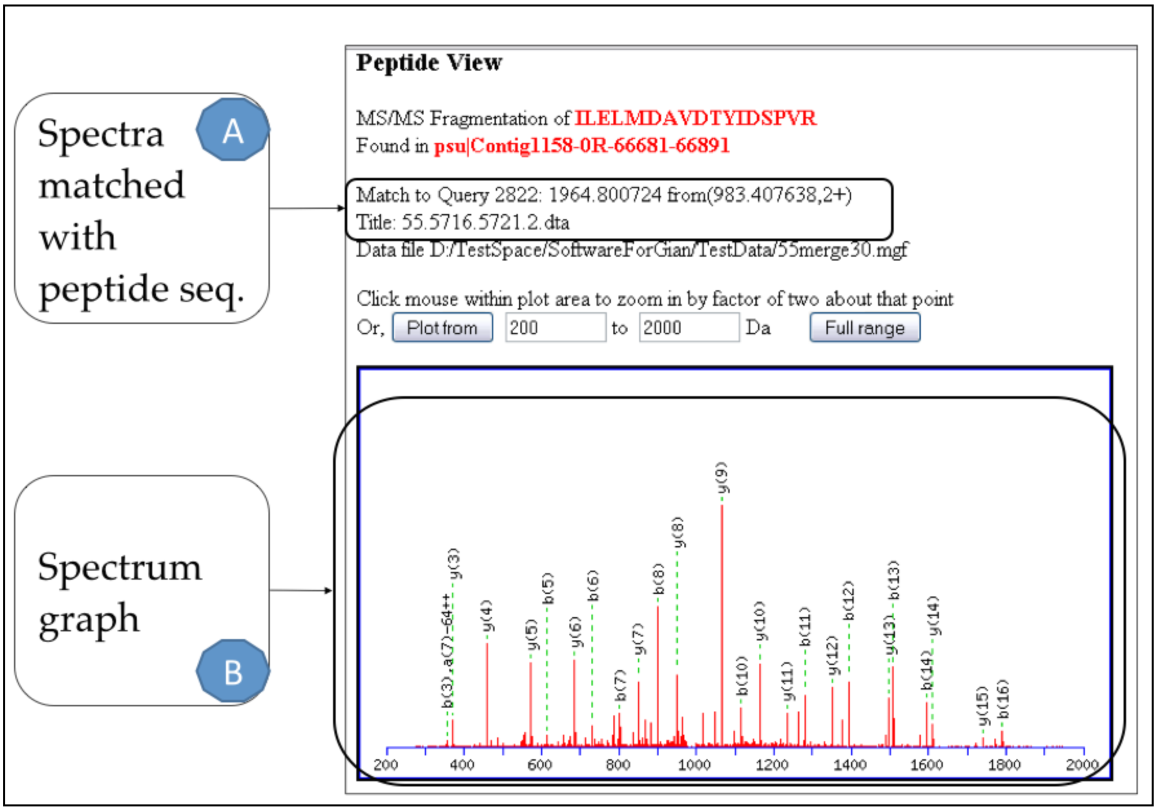
D: The protein score is approximately the sum of each individual peptide score, derived from ion score

E: The number of MS/MS spectra matched

F: The identifier of the query matched against the peptide, it leads to a page with full Peptide view of the spectrum identifier.

G: This is a summary of peptide different mass/per charge, which provides the observed mass queried against its theoretical uncharged mass, the theoretical closest peptide mass matched and their difference.

H: Indicates where the match comes from an incomplete cleaved peptide, as by experience proteolysis usually fails to cleave every peptide, by allowing some partial cleavage. If its value is chosen inappropriately, like above 2, it will simply make MASCOT increase the random matches resulting in indiscriminate hits.



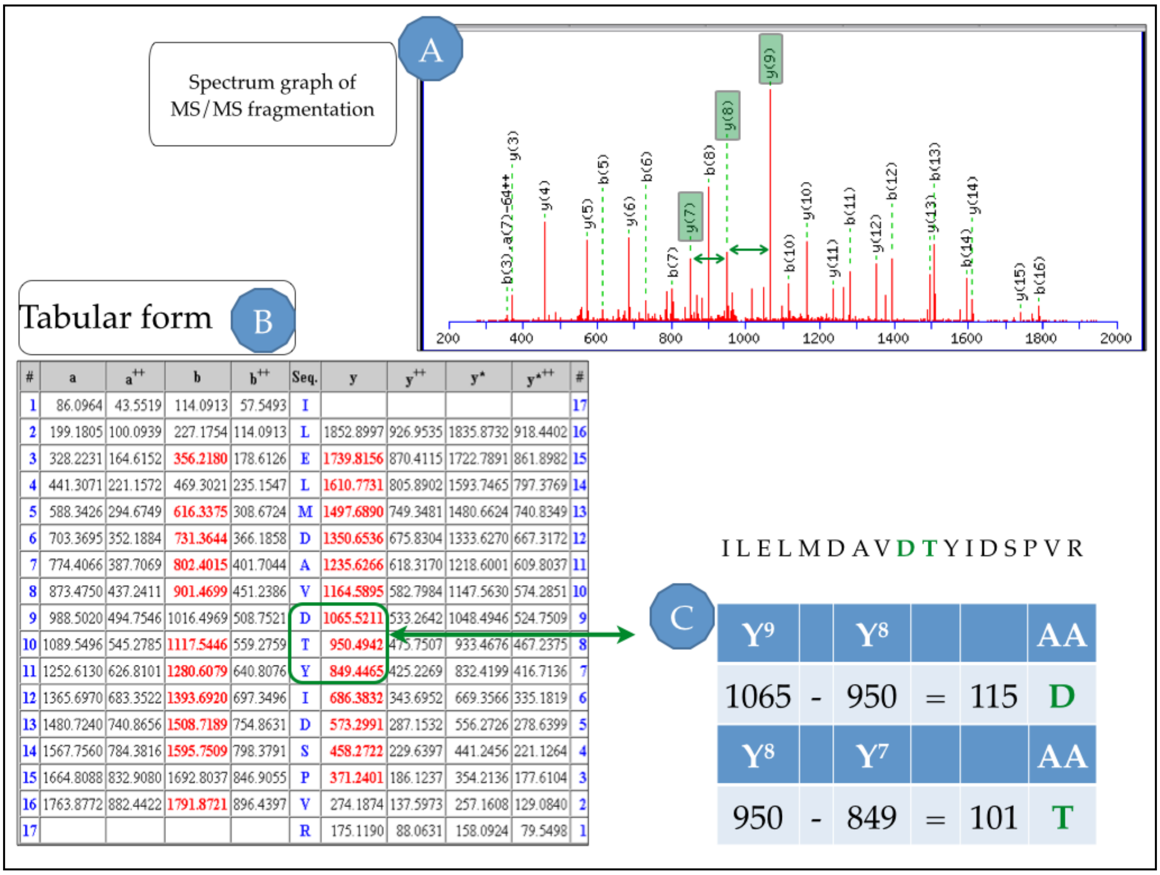
Peptide View

MS/MS Fragmentation of ILELMDAVDTYIDSPVR
Found in psu|Contig1158-0R-66681-66891

Match to Query 2822: 1964.800724 from(983.407638,2+)
Title: 55.5716.5721.2.dta
Data file D:/TestSpace/SoftwareForGian/TestData/55merge30.mgf

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 200 to 2000 Da Full range

Figure 1:11 A: details of the mass spectrum matching with this peptide; B: spectrum graph showing the matched fragment ions, it can be zoomed by a factor of 2 by clicking on it.



#	a	a^{++}	b	b^{++}	Seq.	y	y^{++}	y^{*}	y^{*++}	#
1	86.0964	43.5519	114.0913	57.5493	I					17
2	199.1805	100.0939	227.1754	114.0913	L	1852.8997	926.9535	1835.8732	918.4402	16
3	328.2231	164.6152	**356.2180**	178.6126	E	**1739.8156**	870.4115	1722.7891	861.8982	15
4	441.3071	221.1572	469.3021	235.1547	L	**1610.7731**	805.8902	1593.7465	797.3769	14
5	588.3426	294.6749	**616.3375**	308.6724	M	**1497.6890**	749.3481	1480.6624	740.8349	13
6	703.3695	352.1884	**731.3644**	366.1858	D	**1350.6536**	675.8304	1333.6270	667.3172	12
7	774.4066	387.7069	**802.4015**	401.7044	A	**1235.6266**	618.3170	1218.6001	609.8037	11
8	873.4750	437.2411	**901.4699**	451.2386	V	**1164.5895**	582.7984	1147.5630	574.2851	10
9	988.5020	494.7546	1016.4969	508.7521	D	**1065.5211**	533.2642	1048.4946	524.7509	9
10	1089.5496	545.2785	**1117.5446**	559.2759	T	**950.4942**	475.7507	933.4676	467.2375	8
11	1252.6130	626.8101	**1280.6079**	640.8076	Y	**849.4465**	425.2269	832.4199	416.7136	7
12	1365.6970	683.3522	**1393.6920**	697.3496	I	**686.3832**	343.6952	669.3566	335.1819	6
13	1480.7240	740.8656	**1508.7189**	754.8631	D	**573.2991**	287.1532	556.2726	278.6399	5
14	1567.7560	784.3816	**1595.7509**	798.3791	S	**458.2722**	229.6397	441.2456	221.1264	4
15	1664.8088	832.9080	1692.8037	846.9055	P	**371.2401**	186.1237	354.2136	177.6104	3
16	1763.8772	882.4422	**1791.8721**	896.4397	V	274.1874	137.5973	257.1608	129.0840	2
17					R	175.1190	88.0631	158.0924	79.5498	1

ILELMDAVDTYIDSPVR

Y^9		Y^8			AA
1065	-	950	=	115	D
Y^8		Y^7			AA
950	-	849	=	101	T

Figure 1:12 The arrows shows how the spectrum graph (A) and the tabular form (B) can be traced back to the peptide sequence (C) ILELMDAVDTYIDSPVR.

A: From the theoretical fragmentation in the spectrum graph it is possible to view peaks –*y*7, –*y*8 *and* –*y*9.which allow identifying consecutive amino acids D and T.

B: The tabular form for these fragmented ions shows the matches in BOLD RED for each ion type.

C: Highlighted in green is an example of how each amino acid residue can be calculated based on the difference between adjacent ion fragment masses.

1.3.2 Identification reliability: the assessed scores

Some algorithms do not provide information on the process for scoring and ranking peptide identifications; similarly the score assessed by different algorithms and different sequence databases is not comparable [144]. The algorithms generally calculate the p-value and e-value as a method to evaluate the significance of the identified peptides. The p-value is described as being the area under the curve of the tail in the distribution generated by random matches;

in other words, a p-value assigned to a PSM A describes the probability of seeing a PSM B with the same or better score of A if B has been matched by random chance. In proteomic studies it is not uncommon for thousands of spectra to be searched against sequence database and this leads to the need to correct the score for multiple testing. The e-value, similarly to Bonferroni correction, provides multiple testing correction and it is defined as the expected frequency of PSMs having a better or equal score assuming that they have been matched to a given spectrum randomly [145]. It is calculated using the p-value and the search space, number of spectra and database length (the number of peptides, that are contained within the chosen tolerance, from the database); as an example the MASCOT [122] ion score is calculated from the probability P that the observed match is random event and it is reported as $-10\log P$.

Nesvizhskii *et al* [146] observe that the score distribution for each tool is influenced by many factors such as the quality of the mass spectrometer, the data and the database size. Hence by assessing the false discovery rate (FDR) [147] it can be possible to overcome these limitations [148]. The search engine results are re-scored using the FDR-based statistical methods to validate the expected rate of true positives and false positives of PSMs present below a FDR cut-off score. For this method, the original queried databases contain also a decoy set, sequences that are known to be incorrect. There are different methods to generate these sequences: generally the protein sequences are reversed, or the tryptic sites are conserved while the amino acids are shuffled in order. Then this can be interpreted as a binary classification problem in which a hypothesis can either be True (T) or False (F) and their results can either be Positive (P) or Negative (N), leading to 4 possible outcome, as shown on the table in Table 1:1.

	True actual value	False actual value
Predicted positive outcome	True positive (TP)	False Positive (FP)
Predicted negative outcome	False negative (FN)	True negative (TN)

Table 1:1 Binary classification problem: the hypothesis can either be True (T) or False (F) and their results can either be Positive (P) or Negative (N), leading to 4 possible outcome.

From a real true result, if the outcome is also positive, then we can consider it as TP (true positive) while if the outcome is negative it will be a FN (false negative); while in the case of an incorrect (false) actual value, like a decoy sequence for instance, the results can be scored as either positive (FP) or negative (TN).

The equation for the global FDR algorithm used in this thesis is the following:

$$FDR = \frac{FP}{TP + FP}$$

Where the FP (false positives) value is estimated within PSMs by counting the number of "decoy" PSMs above threshold. The TP (true positives) are calculated as $TP = (T - FP)$ where T comprises all PSMs above threshold [148].

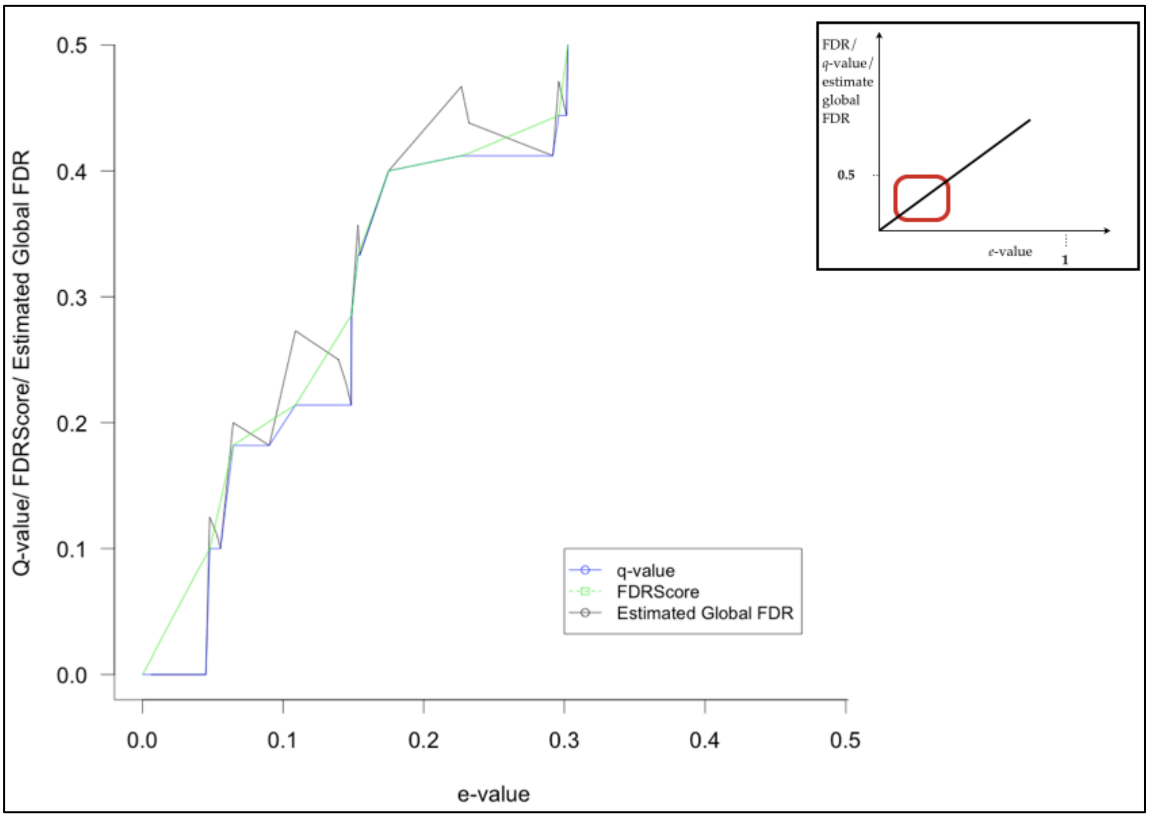


Figure 1:13 Overlay of FDRScore, q-value and estimated global FDR on the right hand side plot and on the left hand side the selected area in red. The data has been obtained from [148] to highlight differences between stepwise q-value score and FDR estimates.

As a practical example, in a dataset, the search engine score such as the e-value is used to sort the PSMs; then, going from the lowest to the highest value, the PSMs are then labelled as FP if they are from the "decoys" otherwise they are calculated as TP ($TP = (T - FP)$); then the FDR is then calculated for each identification. As an example, considering a dataset of 100 PSMs above given FDR threshold, if 5 PSMs came from the decoy database it will lead to a total number of $FP = 5$, while the $TP = 90$; that is within the 95 target PSMs there will be approximately 5 false positives thus $FDR = \frac{5}{95}$. The total number of TP can then be obtained through a fixed FDR, normally 1% or 5%.

As the FDR allows estimating the proportion of FP present at a given threshold in a dataset, the q-value can be defined as the as the minimal FDR value that would render the identification reliable [149]. As described by Käll *et al* the q-value provides a method to score the significance of individual identifications in terms of FDR. That is a q-value of 0.05 for a specific peptide indicates that the peptide can appear within the output at a minimum FDR of 5%. The q-value is considered a global correction method on the dataset, since, as an example, any FP that has a higher rank of the real observed peptide would increase its q-value For instance if a confirmed peptide (NLISENVAFP) is in a dataset of 100 PSMs but its rank is lower than the only FP PSM then this q-value would be 0.01; if instead there are five FP PSMs and they rank above the confirmed peptide (NLISENVAFP) then its q-value would be 0.05. Basically the minimum FDR to set in order to include NLISENVAFP in the dataset has to be respectively 1% and 5% in the two scenarios described [145]. After the FDR has been assigned, the q-value is calculated one-by-one for all identifications, going from the highest to the lowest e-value. It is equal to lowest FDR encountered thus far if this is lower than the current estimated FDR; otherwise the q-value is equal to the estimated FDR, while the lowest FDR is then updated to the current estimated FDR. The FDRScore, showed in Figure 1:13, by Jones *et al* [148] is the FDR estimated between each decoy and it is calculated from the gradient of the connecting line between the step-points where the q-value increases (Figure 1:13). During multiple search engine queries this is used to calculate the combined FDRScore, re-scoring PSMs based on the different sets of search engines that have made the identification.

In addition to FDR is the Posterior Error Probability (PEP), sometimes denoted as local FDR [150]. This assesses the probability that the null hypothesis is null and as such that the observed PSM is incorrect [151] (Figure 1:14). As illustrated by Kall *et al* the non-parametric approach for calculating the PEP for a given PSM can be affected by the bin size by which the target and decoy PSMs have been divided. There have been other developed algorithms for calculating the PEP such as PeptideProphet [152, 153] and Percolator [154].

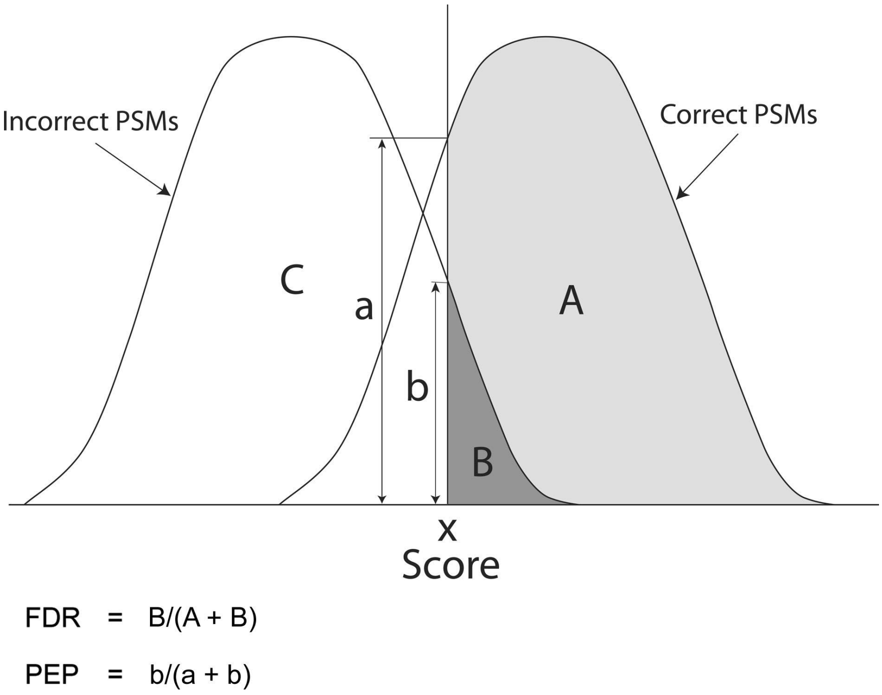


Figure 1:14 The FDR and PEP scoring curves drawn in the study by Käll *et all* [145] visualise the area (A and B) and the heights of the distribution. The FDR is the ratio of the incorrect PSMs with score > *x* (B) to all PSMs (A+B) score > *x*. The PEP is the ratio of the heights of distribution at score = *x* where *b* is the incorrect PSMs at score = *x* while *a* is the number of correct PSM with the same score *x*.

Generally search engine algorithms provide the *e*-value as an estimate that the identified peptide is correct. This is calculated from the *p*-value and the number of sequences *S* in the database (e-value = *p*-value * *S*); as such size of the database can linearly affect *e*-value estimates [155, 156]. Recent studies have stressed the importance for ensuring accurate FDR estimates before validating PSMs identified with target/decoy searches on large genomic databases, such as six-frame translations [157]. Compared to searches against gene annotations, searching 6 frame translations appears to be biased. This is due to the presence of 6 (incorrect) frames (from the decoys) added to the 5 alternative (also incorrect) frames, competing against a single correct frame. Additionally it should be stressed the importance of choosing the correct score. Given its properties the global FDR and q-value are rates related to the whole dataset and as such useful when analysing multiple identification within the dataset. However when the analysis is focused on assessing the reliability of one precise identification, then PEP should be chosen as it provides the probability that given identification is incorrect (Table 1:2).

p-value	It represents the probability (P) of finding a PSM (S) with better or higher score (*x*) by random chance.	$pvalue(S) = P(x \geq S)$
e-value	It is the frequency at which we would expect to see PSMs having equal or better score, if the matched PSM has occurred by random chance. It is calculated by multiplying the p-value with the search space (i.e. the number of searched spectra within given threshold).	$evalue(e) = pvalue \times N$
FDR	Also considered as global FDR, it gives the quantity of false positives that can be expected in a given dataset within a given threshold.	$FDR = \frac{FP}{TP + FP}$
q-value	It is considered as the lowest FDR that can be assigned to the identifications before additional false positives are added to the dataset.	$qvalue = \min\{FDR\}$
PEP	Also known as local FDR, it is the probability that a precise identification is not correct (number of incorrect PSMs with score x).	$\frac{\text{number incorrect PSMs with score x}}{\text{total number of PSMs with score x}}$
FDRScore	FDRScore estimates the FP frequency at given score (*ex*) if this is set as threshold for the search engine. It is based on q-value step-points which are used to calculate the intercept *i* and the gradient *g*.	$FDRScore = ex \times g + i$

Table 1:2 The table shows the denotations of the main scores (p-value, e-value, FDR, PEP, q-value and FDRScore) and their simplified formulas.

1.3.3 *De Novo* sequencing and hybrid approaches

As explained previously, *de novo* sequencing algorithms can provide useful information for unanticipated peptide sequences [158, 159], but they are still limited to very short sequences of amino acids identified with high confidence in the mass spectra. Key to *de novo* reconstruction is the score or probability assigned to peptide sequences such that the highest score reflects the best out of all interpreted peptide sequences [160, 161].

Some scoring algorithms are based on the correlation between observed and theoretical spectrum obtained from candidate peptides. Others evaluate the score from observed peaks by comparing statistical models based on ion fragmentation rules and against fragmentations led by a random process (Figure 1:15) [125]. This is however computationally difficult as the probabilistic distribution appears too complex to be modelled and often leads to errors [162-165].

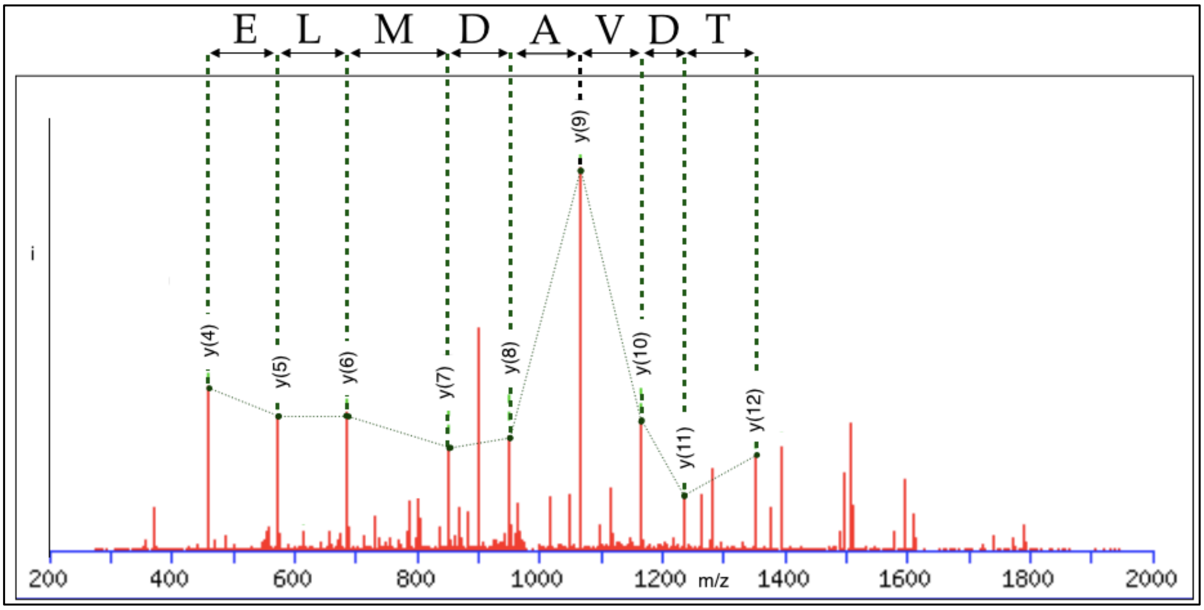


Figure 1:15 From adjacent ions belonging to the same series we can deduct the partial amino acid sequence it originated from. Here the partial sequence ELMDAVDT from the –*y* ion ladder of peptide ILELMDAVDTYIDSPVR (identified by MASCOT).

The *de novo* sequencing software PepNovo [161, 166, 167], like InSPecT (Interpretation of Spectra with PT modifications) software (Figure 1:16), initially analyses the mass spectrum to identify peptide sequence tags (PSTs), short amino acid sequences, and then attempts a full peptide sequence reconstruction. The InSPecT algorithm computes the score from three factors: intensity rank of the peak, isotope pattern and prefix residue mass (PRM) that represent each node from the constructed directed acyclic graph. By default the output is a set of long amino acid sequence, where the high confidence PST is contained. PepNovo algorithms are currently based on a training dataset obtained from mass spectra where the fragment ion has been achieved with collision-induced dissociation (CID).

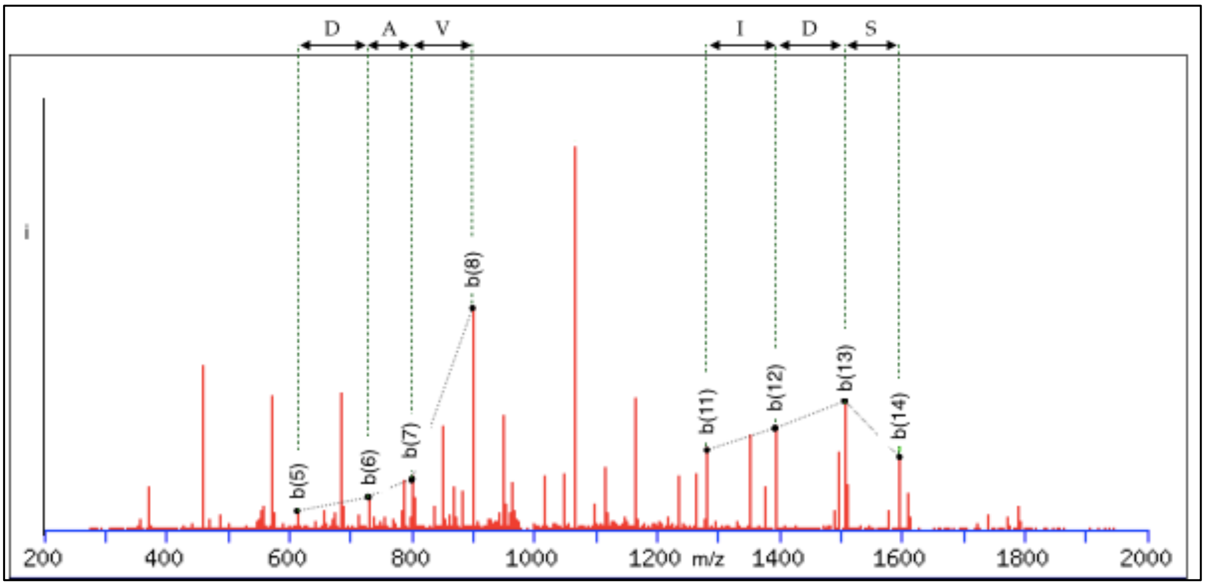


Figure 1:16 Peptide identified with InSPecT as database search engine. In this mode, its algorithm first generates a set of interpretations for high intensity peaks and from these it reconstructs peptide spectrum tags PSTs (typical length of three amino acid). The spectrum above shows DAV and IDS as calculated PSTs. Exploiting the TAG pattern and the spectrum mass the sequence database is filtered. Finally a search database query is performed on this restricted database.

Most of the *de novo* sequencing software packages try to overcome the limitations in accuracy by combining analysis with search databases e.g. used in InSPecT, PEAKS and SPIDER, MS-Dictionary [127, 160, 168-171]. The approach comprises *de novo* algorithms for producing high confidence short amino acid TAGs with their prefix and suffix m/z value (corresponding to -*b* and -*y* ions) [170]; this allows them to place the TAG (Figure 1:16) within the spectrum and with this information the sequence database is filtered in order to maximize peptide discovery of unknown sequences while querying a large starting sequence database [171].

1.3.4 Explainable tandem mass spectra

Shotgun proteomics experiment can generate a very large number of MS/MS spectra (magnitude of $\mathbf{10^5}$) and although computational tools dedicated to spectral identification have been greatly improved, many of the collected spectra fail to be identified due to their low quality or to the inability of the algorithm to identify their matching peptide sequence [172, 173]. To this day different approaches have been implemented to maximize spectral identification by evaluating the quality of the spectra and processing only those identifiable. The collection of spectra can be pre-analysed with machine learning algorithms, such as Support Vector Machines (SVM), to assign to each spectrum specific features that allow the assessment of whether it was generated by some peptide in the digested protein sample (i.e. noticeable peak intensities in the CID spectra) [89, 174]. Another approach developed to improve the tandem mass spectra identification relies upon spectral clustering on the basis that 1DE and 2DE LC-MS/MS runs generate a range of duplicated spectra [175]. This method sees the clustering of duplicate spectra with subsequent redundancy removal in order to increase identification of those unassigned spectra. Pep-Miner [176] algorithms first cluster together spectra with similar characteristics, then for each cluster, they generate a representative spectrum used for subsequent analysis; thus reducing computational resources required to complete the search. However this approach is dependent on the data acquisition and is not yet tailored for rare peptide sequences that have not been fragmented more than once and for different charge state of spectra from the identical peptide. For testing these filtering techniques search engine database software, such as MASCOT and Sequest [119, 122] have been used to identify test datasets and assess high and low quality spectra. It remains challenging to filter out the unexplainable spectra without losing any useful data prior analysis (i.e. for database searches or *de novo* sequencing) where at best removing around 80% of unworthy spectra leads to a 10% loss of explainable spectra [177].

1.4 Proteomics and proteogenomics

1.4.1 Proteomics challenges

Bottom-up proteomics approaches use digested protein sequences for identification. The confidence of protein identification is connected to the number of peptides that were used to identify the protein. The identification of low abundance peptide sequences and the presence of post-translational modifications (PTMs) can prove challenging for proteomic studies. In recent years mass spectrometry proteomic studies have increasingly provided researchers with a high-throughput method for identifying and measuring proteins, targeting their interactions, post translational modification and for annotating genes [123, 138, 178-180].

Unlike prokaryotes, which feature compact genomes with only a small amount of non-coding sequences outside and within gene boundaries, eukaryote genomes present a far more complex structure where genes contain a large number of sequences that are removed between pre-mRNA and mRNA (Figure 1:1). Ever since spliced genes were found, the presence of intragenic regions was considered of particular biological importance, for example as a driver of evolution [181].

Since multiple-exon genes in eukaryotes can be translated into more than one protein product due to alternative splicing, it is crucial to correctly predict intron-exon boundaries [21, 25, 182, 183].

Of crucial importance for proteogenomics, are the peptides that align across introns, which essentially imply that a peptide sequence belongs to consecutive exons. During translation to protein the splice site might be encoded by one single codon, sharing one nucleotide with adjacent coding sequence. At present it is still a huge challenge to find these peptides with bioinformatics approaches as they partly rely on gene finder accuracy. If the gene finders have not predicted the correct splice junction of a gene, it will not be possible to identify the spectra of intron-spanning peptides (ISPs) during analysis on the sequence.

No matter how accurate gene prediction software can be, it is not unusual for them to miss out short single exon genes [184], or fail to detect different protein isoforms due to alternative splice sites [185-187]. If, for example, the mean length of each exon in a multi-exonic gene is about 50 amino acids there is a 25% probability that a tryptic peptide is crossing intronic boundaries [188]. EST approaches such as the Transcript Assemble Program (TAP) [189] have been implemented in order to maximize the identification accuracy of splice variants generating results with good sensitivity. However it remains intrinsically dependent on EST coverage [190] which is often low.

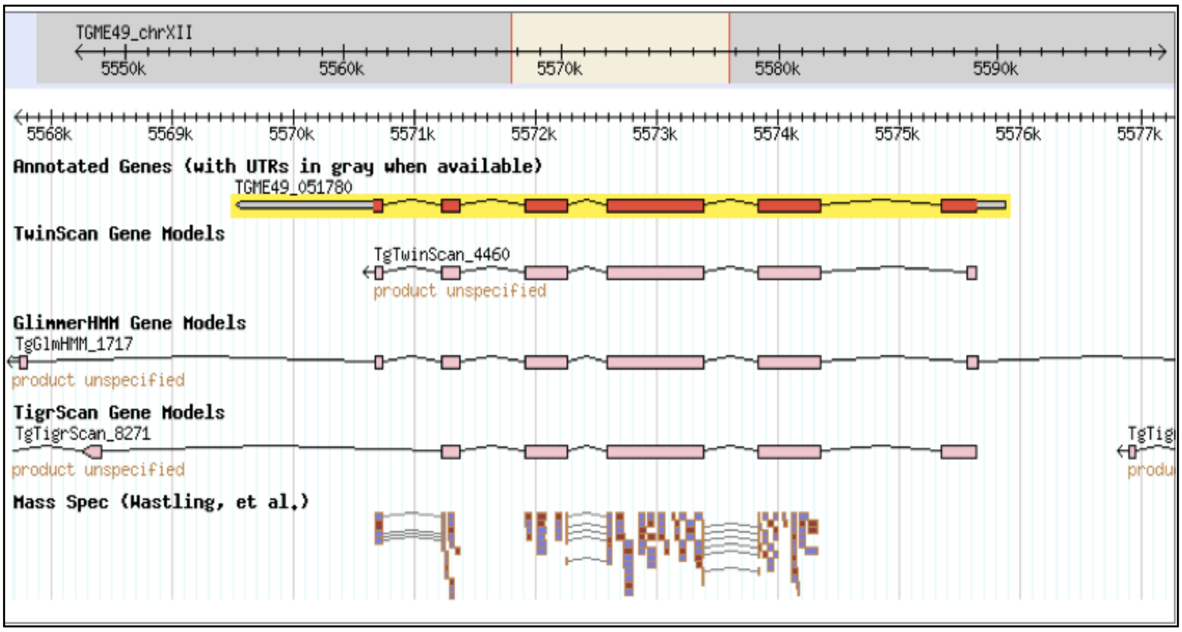


Figure 1:17 View of a genomic region of *Toxoplasma gondii* extracted with Gbrowse [191, 192] from EupathDB [193]. The segments highlighted in red represent the coding regions from gene TGME49_051780 as per the official annotation. Below there are panels of alternative annotations predicted with different software packages (GlimmerHMM, Twinscan and Tigrscan) sharing similarities and divergences on the predicted gene structure. At the bottom the mass spectrometry peptide evidence provided by the Wastling group is shown. As shown, a peptide can confirm the genomic structure by providing evidence of splice sites.

Although a multiple database search approach such as querying Open Reading Frames (ORF_SS) or panels of alternative gene predictions can assist at correcting/ confirming annotation [178, 194, 195] (Figure 1:17), the approach would still be dependent on a protein/peptide sequence being present in the database. Even to this day, it remains a bioinformatic challenge to provide proteomic evidence of intron exon boundaries; the so-called splice sites or splice junctions [196-198], since confirming the presence of these elusive peptides in the experimental data relies on accurately predicted protein sequence [101, 188, 199-201].

Also challenging are the N-terminal peptides, due to possible N-terminal modifications (i.e. signal peptide) and bioinformatic difficulties in correctly predicting the translational start of the protein.

1.4.2 Proteogenomics role

In the recent years mass spectrometry-based proteomics have advanced considerably providing rapidly improving techniques for generating comprehensive collections of data from studied organisms. Dedicated sample preparation (fractionation protocols and multiple proteases) makes it possible to increase depth in sub-cellular proteome analyses and expand sequence coverage [202]. With an ever-increasing number of genome sequencing project throughout the scientific community, thousands of genome sequences are now available for deeper proteomic research studies. Publicly available resources such as Genome Online Database (GOLD [203]) provide an outlook on genome sequencing project across the world.

To this day manual annotation remains the most accurate method to determine a gene model; this however presents a considerable bottleneck in the process given the increasing number of organisms sequenced [204]. In contrast to data from transcriptomics, such as RNASeq [53], the use of proteomic data allows the validation of the models based on protein expression evidence, interactions and post-translational modification (PTMs) at precise life stages of the organisms through sequence database searches [123, 179, 180, 202]. A recent study from

Adamini *et al*, attempts to combine shotgun proteomic evidence with *de novo* transcriptome sequencing [205] to confirm expressed protein-coding genes for *S. mediterranea*. Through a comparison of identified tandem mass spectra (<1% FDR) on an available gene model set and their sequenced transcriptome, they highlight the importance of proteomic evidence for validating previous gene models [206, 207]

Proteogenomic studies are closing the gap between proteomics and genomics, providing gene annotators with tandem mass spectra-derived peptide evidence for protein identification and characterization. In this study Castellana *et al* [208] refine the *A. thaliana* proteome by querying 3D LC tandem mass spectra with InSPecT search engine [171] against 3 distinct sequence databases: official proteome (TAIR), 6 frame translation and a splice graph (for all putative splice events). By providing proteomic evidence for ~12k genes, identifying ~500 novel genes and refining around 1000 gene sequences they bring support to proteogenomics approaches. The lack of data for low abundance proteins in the sample and low quality tandem mass spectra can be seen as limitations to their approach.

In another interesting proteogenomic approach developed by the same group an imperfect genomic template is exploited with tandem mass spectra to successfully return the correct target protein [209]. Here spectra are used to construct the template sequence, where their chain order is given by the genomic sequence (6 frame translation of gene loci). Then anchors are selected by overlapping spectra providing specific amino acid substrings that appear with no mutations on both template and target proteins. The final stage of their method sees the extension of these anchors to create the protein sequence. However this method is still limited by both PTMs and unexpected splice sites present in the sample, as the consensus algorithm would miss these out.

Another useful approach is comparative proteogenomics, whereby the tandem mass spectrometry-based proteomic data is exploited for parallel genome analyses across different species from the same genus [210]. This approach, however powerful, relies on sequence homology and domain conservation of the studied species; it is also based on highly curated genomes to assess identification of the sequence on less curated genomes. Rare unexpected modifications require further evaluation in order to measure their confidence.

Proteomic data can help to confirm not only the presence of gene expression as well as PTMs but also the complex structure of genes [171, 209] and protein-protein interactions [211]. This data has been already exploited for genome annotation of several organisms such as *Toxoplasma gondii* [101], *Plasmodium falciparu*m [212], *Drosophila melanogaster* [213], *Homo sapiens* [214] and *Caenorhabditis elegans*[178].

1.5 Organisms studied

The data sets used to test and refine the approached developed in my project are obtained from protozoan parasites of the phylum *Apicomplexa*. These protists belong to the superphylum of the *Alveolata* of Chromalveolata kingdom and are all obligate intracellular parasites. The apicomplexan parasites differentiate from other *Alveolata* for their unique cellular morphology such as the characteristic apical complex (with secretory organelles) and the apicoplast, used during cellular invasion of host organisms [215, 216]. These eukaryote parasites share many metabolic pathways with the host, which makes the development of therapeutic targets particularly difficult and there are no vaccines for most diseases caused by these parasites. They all have a complex life cycle, often comprising 3 different stages, both asexual and sexual (sporogony, merogony and gametogony); additionally many of these protists can have intermediate hosts (vertebrates and invertebrates) before reaching the definitive host. During initial cell invasion the parasites replicate asexually producing large quantities of sporozoites (one replication). From these, during merogony stage, the parasite replicates generating merozoites; during this stage the replication can occur multiple times. Following this, through the gametogony stage the parasites undergo sexual reproduction, generating gametes that form zygotes. Then the cycle starts over with sporogony stage. Although the whole life cycle can take place within the same tissue/ cell, it often occurs in different host/ tissues.

The Apicomplexan parasites are responsible for major tropical diseases such as Malaria caused by the *plasmodium* parasite, alone responsible for hundreds of thousands of cases resulting in death worldwide [217-221]. *Toxoplasma gondii*, the most common zoonotic parasite, can infect all warm-blooded animals and around the globe more than a billion people are infected [222]. This, among other coccidia, is one of the most extensively studied [223] with several genomes sequenced to this day [224]. *Neospora caninum,* closely related to *toxoplasma,* causes abortion in cattle. It can be transmitted horizontally and vertically, where it can pass consecutively and intermittently to the offspring [225, 226].

Notably, there are genomic structural differences across the eukaryotic parasites of the phylum Apicomplexa such as *Cryptosporidium parvum*, *Plasmodium falciparum*, *Toxoplasma gondii* and *Neospora caninum* [42, 193, 227-229], which are studied in this work. Compared to the *P. falciparum* genome, the *C. parvum* genome appears to be almost two fifths smaller and presents an almost doubled gene density. This divergence increases with the *T. gondii* genome, which, with its 14 chromosomes, is over double in size compared to *P. falciparum* and presents a lower gene density having more introns per gene [230-232].

As currently genome annotation is largely based on predictions of the gene structures, these can dramatically change over time as more accurate predictions are computed and released to the public [184]. Using sequence databases such as UniProtKB, NCBIgi and IPI can lead to out-dated peptide sequences reported in their protein identifications from proteomic studies [233]. Even in a well-annotated genome there is a large percentage of hypothetical proteins (purely based on predictions and without known domains) and putative proteins (having sequence similarity with characterized proteins but not experimentally validated) [234]. As an example, during evaluation on the genome structure of apicomplexan parasites, *P. falciparum* presented only ~0.68% hypothetical and ~32% putative, while for C. parvum, *N. caninum* and *T. gondii* were respectively ~40% and ~5%, ~43% and ~33%, ~64% and ~31% [101, 227, 232, 235].

The data for this project, used for mining or statistical evaluation, comes from: *Cryptosporidium parvum [227]*, *Plasmodium falciparum*, *Toxoplasma gondii*, and

Neospora caninum (Table 1:3).

In the recent years genome sequencing has been performed for these species. In this project, due to the close evolutionary distance, I have been using *Toxoplasma gondii* to provide a statistical insight on the current quality of annotation for the later sequenced *Neospora caninum* genome and how it could be improved (Table 1:3 [193, 227-229, 231, 236-238] - for further sources EupathDB).

Organism	Genome	Genome database	Available proteome data
T. gondii	Completed	ToxoDB - highly annotated gene predictions	Wastling group (2008); Kim and Weiss group (2008) in ToxoDB
P. falciparum	Completed	PlasmoDB - highly annotated gene predictions	Several different groups, e.g. Florens *et al.* in PlasmoDB.
N. caninum	Completed	ToxoDB - highly annotated gene predictions	The welcome Trust Sanger Institute - Wastling group (2012) ToxoDB.org
E. tenella	Whole shotgun contigs 2007	GeneDB - early stage automated predictions	Tomley (IAH) and Wastling groups (2008)
C. parvum	Draft assembly	CryptoDB - high quality automated predictions	Wastling group, see also cryptodb.org

Table 1:3 Genome sequence and proteomic data available for apicomplexan parasites. It takes consideration of the recent annotation improvements of *N. caninum* [239].

1.6 Research objectives

This research focuses on the implementation of novel approaches to maximise peptide identification for gene model validation and improvement. To achieve this target, attention is devoted to optimising database design. Additionally a multi-sequence database approach is devised, similar to a multiple search engine mode, to examine the performance of specific database structures (e.g. gene models, six frame translation or ORF_SS/ ORF_MS). This analysis can be used to improve current database structure resulting in increased peptide identification. This type of approach is then tested on past and presently available genome sequences for *T. gondii* and *N. caninum* to provide a broad picture of genome annotation progress through time.

Furthermore the research focuses on targeted peptide identification to confirm and correct gene structure as well as novel identifications. The peptide sequences in question are those peptides that span across two exons and N-terminal peptides. As only querying gene models can confidently identify the first type, a novel type of method is proposed. This method can be considered as a type of hybrid approach, involving peptide sequence tag identification and database searches; but the result is similar to a *de novo* approach since it works in the absence of gene models for full-length peptide identification.

The challenges of N-terminal identification have been improved due to recent advances in techniques for sample preparation that allow selective enrichment of N-terminal peptides. This research focuses on database design assessment and performance optimisation that can provide lists of valid N-terminal peptides to confirm or correct the translational start of available models. Using these results we attempt to provide further rules to increase the identification confidence.

2 Gene annotation: quality assessment and improvements

2.1 Abstract

As most current gene annotations only result from computational predictions, proteomic data can be used to assess and validate them. This chapter proposes database design strategies to address limitations in both assessments of draft annotations and maximisation of peptide identification from tandem mass spectra. The designed strategy comprises multiple database searches followed by analyses and comparisons of the datasets searched against different databases.

The pipeline compares results from official gene models and open reading frames (ORFs) derived from a six-frame translation, and identifies PSMs that are unique to individual databases for further investigation. This gives an overview on database performances (i.e. gene model accuracy).

Using *Neospora caninum* data it was possible to identify from searches against ORFs 452 peptides that were absent from the official gene model draft (5.1 release*)*. By analysing all the available genomic sequences and draft annotation for *T.gondii* and *N. caninum* it was possible to evaluate the ratio of PSMs made from searches against ORFs versus official gene models. When gene models are high quality we expect the number of PSMs from hits to gene models to be significantly higher than hits to ORFs. This chapter shows the ratio varies from 1:1 to 1:2 (ORF hits:gene model hits)in *T.gondii* and *N. caninum* as gene models improve over time. The change in this ratio acts as a score of gene model quality.

2.2 Introduction

Genome curation is generally performed with automated methods [28-30, 49, 51] and it undergoes multiple refinement stage before an annotation is considered as complete. Although proteomic studies enable protein identification based on draft annotations it remains uncertain how to assess the accuracy of these predicted models. Recent bioinformatics improvements have provided increased confidence in the peptide sequences identified with search engines [118, 119, 122, 148, 168, 171, 188, 234, 240-242]. This facilitates searches for novel genes and corrections on genomic structure through genome wide searches. Since open reading frame sequence databases (ORFs) are generally used for genome-wide database queries the strategies discussed in this chapter comprise analyses to improve ORF prediction for search optimisation [243].

At the time this study was started, the Sanger Institute had recently released the *N. caninum* genomic sequence and highly annotated gene models were available (release 5.0) from the closely related *T.gondii* ME49 [101, 230, 239, 244, 245]. Because of the evolutionary similarities the *T.gondii* annotation was used to provide an initial assessment of *N. caninum* gene models. This enabled the hypothesis to be formulated that *N. caninum* annotation could be improved, since the model was missing out ~2K genes. Through further rounds of annotation at the Sanger Institute [239] we were later provided with more accurate *N. caninum* data to test the theory.

As described in the introductory chapter, Open Reading Frames (ORFs) in the thesis are extracted from the nucleotide sequence translated in all 6 frames (3 forward and 3 reverse complement). Each ORF is then considered as the portion of sequence located on the same reading frame, comprised either between two stop codon (ORF_SS), or between a Methionine and a stop codon (ORF_MS). When generating them computationally, their minimum length is usually chosen between 50 to 100 amino acids, although the real coding sequence length is directly affected by the genomic structure of the organism (such as intron frequency).

One of the objectives here discussed is how ORF_SS prediction could be improved to yield the best set of sequences while improving search performance. This was achieved by evaluating the genomic structure of other apicomplexan organisms: *C. parvum*, *P. falciparum* and *T. gondii* [42, 193, 227-231, 246].

The designed pipeline enables us to compare the PSMs in the datasets and generate Venn diagrams. By tracking PSMs back to specific sequence databases, this can facilitate further design refinements. This approach was used with ORF_SS and gene model databases to adjust the predicted model. Additionally this method proved useful as part of a new protocol for accuracy assessment of gene models using proteomic data.

The multiple database study also included assessment of multiple search engines performance compared to individual search engines. By incorporating alternative gene models in the study it was also possible to highlight their importance in a proteogenomics context; it also showed a rough benchmark of Glimmer and GeneMark software packages.

This chapter has the following aims:

- to improve the design of ORF_SS databases;
- to provide additional evidence for *N. caninum* annotation improvements;
- to measure the improvements of the gene models over time;
- to benchmark gene finding software and the multiple search engine approach;

2.3 Methods

2.3.1 Understanding gene structure of studied organisms

The first step in reducing the size of a six-frame translation database could be achieved by extracting the putative Open Reading Frame (ORF) sequences [243, 247, 248]. The ORF sequences discussed in this chapter are those stretches of sequence, located on the same reading frame, comprised between two stop codons (ORF_SS as referred to in the introduction chapter).

In order to generate an ORF_SS database containing relevant data, while filtering out likely non-coding regions, it is necessary acquire some statistical figures on the genomic structure of the organism analysed. The initial study comprised the annotation release 5.1 for *Neospora caninum* as well as *Toxoplasma gondii* release 5.1 and other species from the same phylum for which the annotation is at a more advanced stage such as *Cryptosporidium parvum* and *Plasmodium falciparum* release 4.1 and 5.5 respectively. The last two organisms were selected to validate the statistical methods used for analysing the others.

2.3.2 Sequence database analysis

The workflow for the database searches and scoring approach was designed to include a False Discovery Rate assessment algorithm [148]. As such, to construct the decoy sequence database the original sequences were reversed with a flag in their accessions (i.e. the word "Rev" prefixed to all headers); this was then concatenated to the original sequence database.

The database search engine software of choice for analysing MS/MS data was X!Tandem [140] and MASCOT [122]. The searches with MASCOT were performed considering: peptide charge of 1+, 2+ and 3+, trypsin digestion, fixed modifications of carbamidomethyl on cysteine and methionine oxidation allowing only 1 missed cleavage, with MS peptide and MS/MS tolerance of ±0.8 Da, after having tested various alternatives, to maximize output at fixed FDR (Appendix B Figure 7:1).

The global FDR re-scoring allows assessment of the output quality and to measure its search performance by using the same MS/MS dataset in multiple parallel queries to different databases or simply with different search parameters. The parent and fragment ion tolerance was then assessed on three

N.caninum 5.1 sequence databases: ORF_SS provided by EupathDB.org, ORF_SS with threshold length of 40 amino acids and the official gene annotation, respectively (see Appendix B Figure 7:1). This method can facilitate the evaluation of sequence database design and optimisation of search parameters.

2.3.3 Alternative gene models

The gene finder software packages used include GlimmerHMM [28, 29, 39] and GeneMark [30, 31, 249]. GlimmerHMM is fully configurable for new organisms, since with enough genomic data it is possible to create new HMM trained models.

Whenever possible only coding sequences from the available gene model were used to generate new GlimmerHMM models. The theory was based on the relation between training data accuracy and HMM model optimisation. For historical evaluation of the official gene models accuracy GlimmerHMM models have been generated using the previously available official release. When the trained models were created with the purpose of contributing to the current annotation, the official model from the same release was used as source data.

For both software packages only default parameters were used, without altering the values for isochores (stretches of genomic sequence with higher content of GC). This was decided as the overall GC abundance throughout the different chromosomes demonstrated little variance across the *T.gondii* genome. In *T.gondii* the difference of GC frequency between coding and non-coding regions is very small and not evident.

Initially the attention was focused on the quality of the predictions that could be generated; ideally the predicted sequence is expected to have a ~80-90% similarity when aligned against the official model. In order to carry out the most appropriate testing it was opted to create both a set of top 5 best predictions (which did contain a fair amount of sequence redundancy) and a set with only the best prediction. We also generated a set of the top 10 best predictions but due to computational time/resources it proved to be inefficient for our purposes, i.e.

slow searches for no gains in PSMs (data not shown). The results showed a general consensus in exon predictions across the top ranking predicted models; instead different ranks are directly connected to the way the exons are assembled into genes (the predicted splicing of exons).

GeneMark does not require previous training for a specific organism for the assessment of hidden states during HMM predictive algorithm. This software package has been designed with Viterbi Algorithm [30, 37] that generates a variable number of hidden states as needed during the computation of gene predictions.

For each species all available sequenced genome releases with their respective official gene model were gathered whenever there was a change between releases. In the case of *T.gondii* the whole shotgun genome sequences were available for the releases: 3.3, 4.x, 5.x and 6.x. The gene models came instead from only three releases: 3.3, 4.x, 5.x (5.x was unchanged in 6.x and 7.x) [250]. The first publicly available gene model, release 3.3, was generated using different gene finding software: Glimmer, Tigrscan (*ab initio*) and Twinscan (*ab initio* and homolog alignments). The second available release for gene model 4.3 was generated by GLEANS, incorporating EST data. To generate predictions for *N.caninum* we used the genomic sequence and models 5.x and 6.x available from EuPathDB. The pre-release 7.x of *T. gondii* was evaluated for additional updates/ changes, but there were none.

As previously explained, GlimmerHMM training sets were generated with known official models, except for *T.gondii* 3.3 as the gene model provided was not compiled as generic feature format (GFF) and it was not possible to acquire the detailed coordinates of the coding sequences. Instead for this particular release the alternative gene model based on Glimmer (provided on ToxoDB) was used [250]. For the same release the genomic sequence was provided in terms of BAC sequences, which were used to generate the ORF sequence database.

2.4 Results

2.4.1 Gene structure of the Apicomplexa

The figures needed to assess the genomic structure were the distribution of lengths of the genes and the length of their initial, internal and final exons. To provide a visual comparison the density plot in Figure 2:1 provides an insight on the different complexity of genomic structure across some apicomplexan species; *C. parvum* for example shows a very simple gene structure where ~95% of its genes are single exons, hence reduces the challenge of finding evidence for gene splicing. *N.caninum* and *T.gondii* share gene structure similarity, as shown on the density plot in Figure 2:1. Although the apex of the density is at single exon genes, for these two organisms, the density is also similar for genes having four to 20 exons.

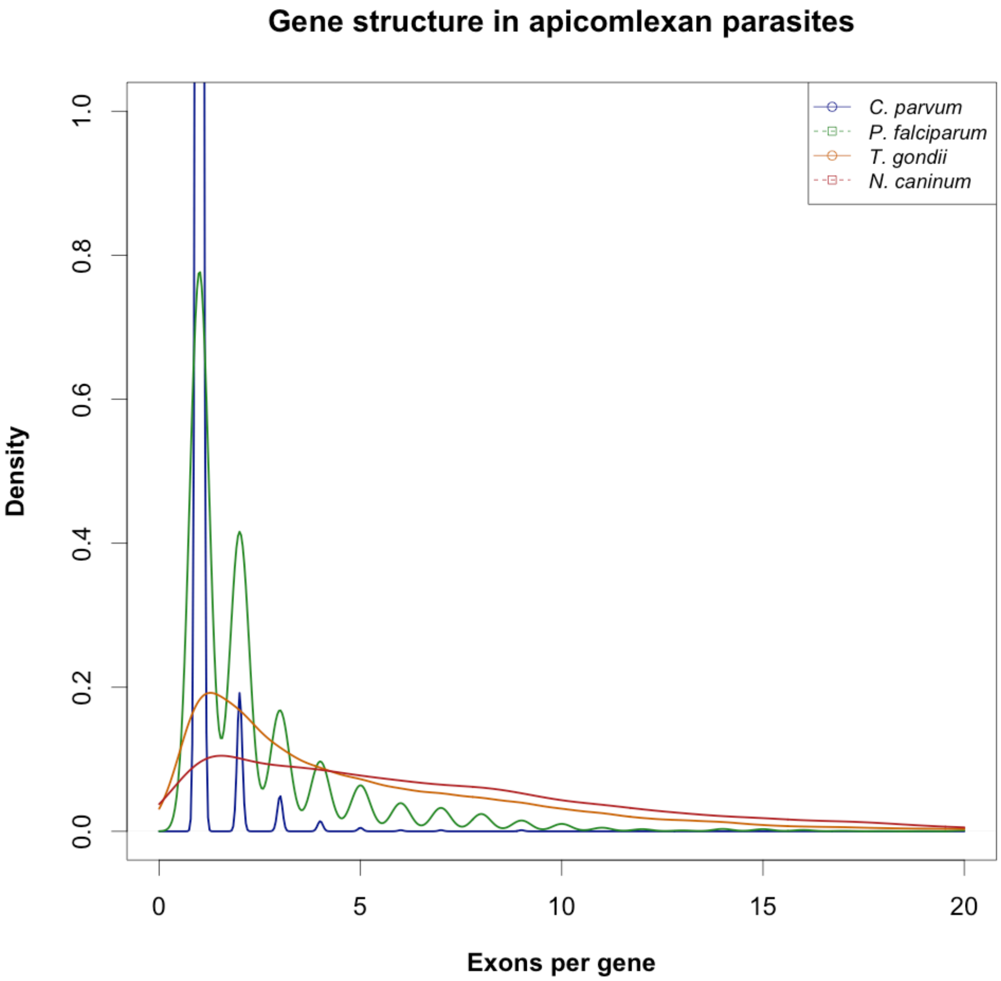


Figure 2:1 The density plot shows the genomic structure of the Apicomplexa *C. parvum* 4.3, *P. falciparum* 5.5, *N. caninum* 5.1 and *T. gondii* 5.1. The plot focuses on the number of exon per gene throughout the whole genome for these parasites. *Cryptosporidium* appears to have mainly single exon genes and no gene with 10 or more exons at all. Plasmodium has a high number of single and double exon genes. Neospora and toxoplasma show a similar density of genes with five to 20 exons, although the latter shows a higher number of genes with less than four exons.

The study aimed to provide information about short coding sequences located in short open reading frames that could potentially be filtered out during ORF_SS database generation. Figure 2:2 presents a boxplot and density plot of the exon lengths across the genome of all four parasites. The median of the boxplot shows the central tendency, which is 186, 162 and 186 bps for *T. gondii, N. caninum* and *P. falciparum* respectively. *C. parvum*, having mainly single exon genes, has a median value of 1191 bps and a lower interquartile of 621 bps. For the other three species the lower interquartile is 104, 96 and 81 bps respectively. The higher interquartile shows similarities between *T. gondii* and *N.* caninum (446 and 339 bps respectively) and similarities between *P. falciparum* and *C. parvum* (983 and 2070 bps). This is related to the number of exon per gene (Figure 2:1). The whiskers are drawn to indicate the minimum and maximum values that for *P. falciparum* and *C. parvum* are considerably more evident – particularly for the longest exon. The exon length and count distribution was considered when determining the minimum threshold for ORF_SS extraction in order to identify novel peptides and possibly novel genes. Given the difficulty of predicting the start of the protein with high accuracy, it appeared important to draw the attention to short exon frequencies, such as exons shorter than 90 bps (~30 amino acids), 120 bps (40 amino acids) or 150 bps (50 amino acids). These are presented in the density plot of Figure 2:2 as enlarged area between one and 2000 bps, with the overlaid ORF_SS thresholds; together with the boxplot it allows evaluating more accurately the minimum thresholds for capturing short exons that would be otherwise missed out (i.e. lower quartile regions). This can be useful as these sequence databases are often generated only computationally, by applying a length filter after extracting sequences from the genome. The length filter applied on the ORF_SS sequence database available on EupathDB.org is 50 amino acids, so more than ~3000 exons could be missed out in *P. falciparum*, *N. caninum* and *T. gondii*.

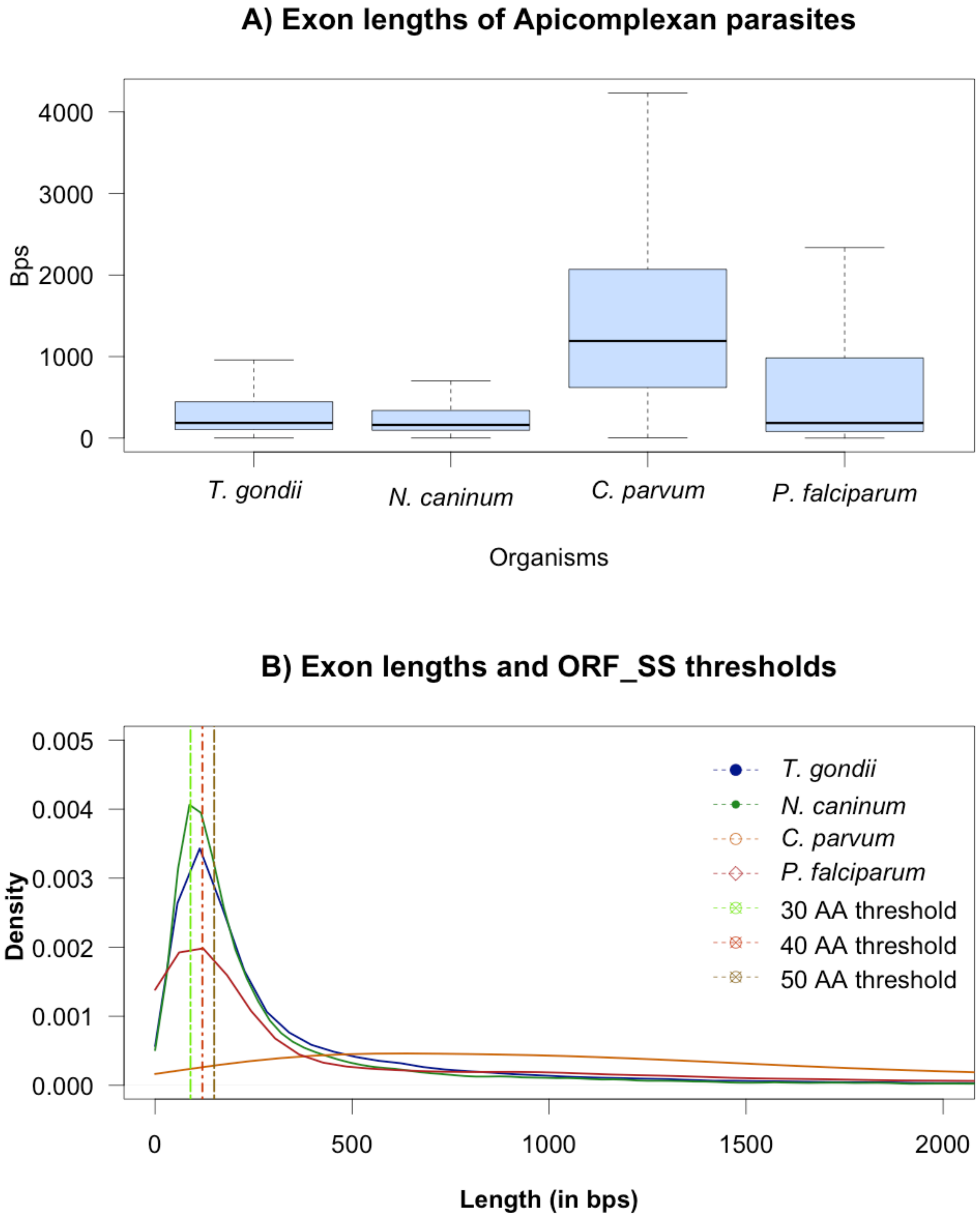


Figure 2:2 A) The boxplot shows the exon distribution and the extreme values (whiskers) across apicomplexan parasites from *C. parvum* 4.3, *P. falciparum* 5.5, *N. caninum* 5.1 and *T. gondii* 5.1. As *C. parvum* 4.3 and *P. falciparum* 5.5 have mainly single or double exon genes their exons appear longer when compared with *N. caninum* 5.1 and *T. gondii* 5.1. As for *T. gondii* 5.1 the exons are generally shorter than 2000 bps the density plot has been focused

within 2000 bps limit. B) The density plot shows the length densities of the exons and from the data it is clear how *T. gondii* and *N. caninum* 5.1 could benefit from precise threshold length for ORF_SS. The ORF_SS thresholds of 30 and 40 amino acids have been displayed here. The different density of lengths can help chose appropriate filter for ORF_SS prediction.

For instance in Figure 2:2 it is possible to note the median length among all exons as well as their interquartile and extreme values of the two species. Although both genomes show a similarity in gene structure (number of exon per gene), the total number of genes between these two species appears to be significantly different. The annotation of *Neospora caninum* 5.1, with 5587 predicted genes, appears to have 2406 genes less than in the *Toxoplasma gondii* (ME49) annotation. These similarities and differences allowed us to make use of known structures from well-curated genomes to aid the gene curation for the newly sequenced organism *Neospora caninum*.

As the Sanger Institute was sequencing the *N. caninum* genome, the curators allowed the pre-release, corresponding to 6.0 [239], to be used in this study. As its generic feature format file (GFF) was accessible during late August 2009, the same statistical analysis was performed to observe divergences and similarities within different annotation releases from the same species. Both annotations were compared between each other as well as against *T. gondii*. The data shows how these organisms have similar genomic structure and provide further insights how *Toxoplasma gondii* data can potentially aid the on-going annotation for *Neospora caninum*. In the new release *N. caninum* presents 1440 more genes (7027 in total) than in the previous release, and the average of exons per gene reduces from 7.3 to 5.9, closer to *T.gondii* 5.1. A similar profile of exon length was observed in *T. gondii* 5.1, *N. caninum* 5.1 and *N. caninum* 6.0 (Figure 2:3).

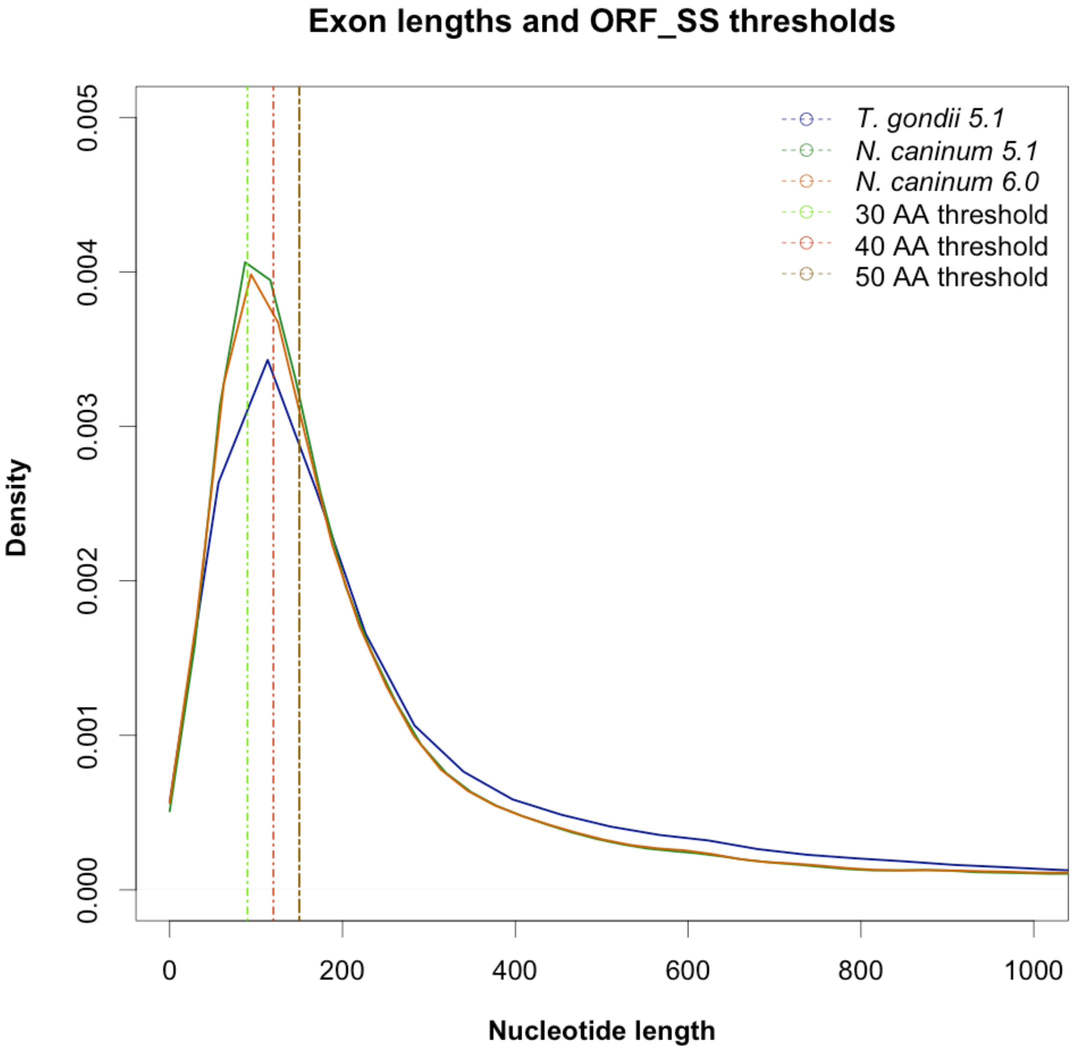


Figure 2:3 The release 6.0 of *N. caninum* shows an increase in similarity towards the *T. gondii* structure with a smaller density of exons 100-120 bps long.

2.4.2 Defining the optimum length of ORF_SS

The optimum ORF_SS length was selected from the statistical figures such as the mean length for genes, exons, introns and how they relate to each other on the sequence. We evaluated the statistics from *N. caninum* in comparison with *T. gondii* (ME49 strain) that, to our knowledge, had a better annotation at the time of the study. The mean length of the first exons was particularly important as MS/MS evidence could potentially confirm the correct start of the gene, or show a different start. This information was used to decide the most appropriate threshold for the minimum length of ORF_SS from a six-frame translation.

The main objective was to keep the threshold as low as possible to include as many potentially valid sequences to be contained in the database, including atypically short ones, while maintaining statistical power in the results of the search. From the data gathered for *N. caninum* release 5.1 about ~27% of all 40,570 exons were sequences below 33 amino acids and about ~31% have length comprised between 30 and 66 amino acids (Figure 2:3).

Our databases were generated using four values as threshold: one with no threshold (to include every sequence comprised between two stop codons) and the other three having respectively 30, 40 and 50 amino acids as the minimum length. The database size was 218Mb for threshold set at zero, and for the others was respectively 140Mb, 121Mb and 105Mb (Table 2:1).

Sequence database	Amino acid count
ORF_SS_All	~78 x 10^6
ORF_SS_ 30	~39 x 10^6
ORF_SS_ 40	~28 x 10^6
ORF_SS_ 50	~22 x 10^6
Official Gene model (OGM) *Neospora caninum* 5.1	~1 x 10^5
Official Gene model (OGM) *Neospora caninum* 6.0	~7 x 10^5

Table 2:1 Comparison of database size between ORF_SS with different thresholds and the official gene models for *N. caninum*. The calculated number of amino acid excludes decoy sequences.

2.4.3 Datasets comparisons

The results from the first tests helped to formulate the best threshold level for a reasonable trade-off between database size and true positive - false positives ratio (Figure 2:5). It appears that all ORF_SS sequence databases returned very similar results (in this case the score is represented by the total number of TP PSMs at a fixed FDR threshold). In order to test the differences in ORF_SS threshold and evaluate how it affects identification of PSMs and peptide sequences it was decided to combine 10 slices from the same 1DE gel for *N. caninum* 5.1 (Figure 2:5). Merging all results into one dataset could potentially alter the statistical scoring of the final dataset. However querying each slice individually could increase the completeness for protein identification as these slices were sourced from different mass spectrometry runs.

Next all the individual results for each database search were re-scored individually by FDR with a maximum threshold of 5%; in the following step all of these were combined into one file per search database where the sequence redundancy had been replaced by a counter of PSMs based on the peptide sequence alone. The final output provided all TP non-redundant peptide sequences found in each database, excluding all false positives, listing the lowest FDR found in case the peptide had been matched multiple times. An internally designed pipeline algorithm was used to perform a post-processing comparison of all sequences on both outputs mutually inclusive and exclusive sets of peptides (see Figure 2:4 for workflow). The final lists highlight whether specific peptide sequences are unique to either the ORF_SS database or to gene model databases, or common to both.

The reason for this was to generate a small set of peptides that can be investigated further to elucidate whether or not they appear on the gene models. Even with increasingly accurate gene models, there can be missed peptide sequences from the gene models such as wrongly predicted splice sites and the N-terminus of the proteins are still difficult to predict with precision.

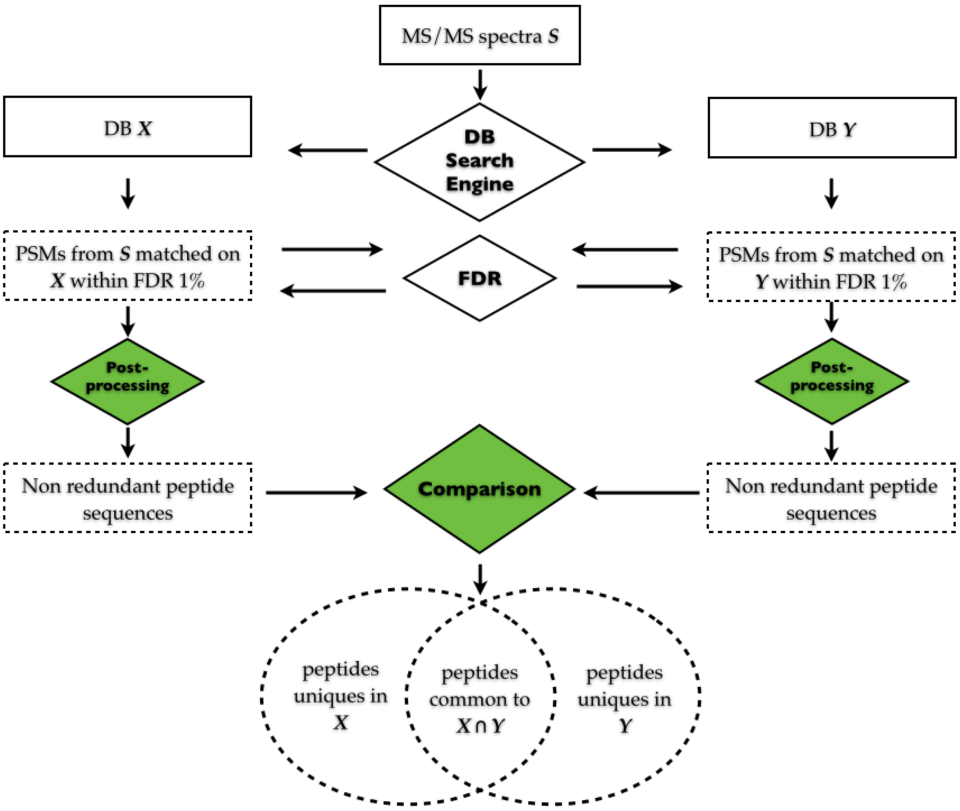

MS/MS spectra *S*
DB *X*
DB Search Engine
DB *Y*
PSMs from *S* matched on *X* within FDR 1%
FDR
PSMs from *S* matched on *Y* within FDR 1%
Post-processing
Post-processing
Non redundant peptide sequences
Comparison
Non redundant peptide sequences
peptides uniques in *X*
peptides common to *X*∩*Y*
peptides uniques in *Y*

Figure 2:4 The workflow for the post-processing algorithm. Tandem MS data are queried with database search engine (DB SE) against different databases. The results are then individually rescored by fixed FDR and the redundancy is removed at peptide level. Finally the results from each datasets are compared organising them into mutually exclusive/inclusive sets of peptides.

A comparison between different dataset can provide the intersection of the datasets, showing which database search yielded specific peptides (Figure 2:5). However as this could be the result of a statistical bias, it was appropriate to evaluate whether the peptides unique to a specific dataset were also unique to given database.

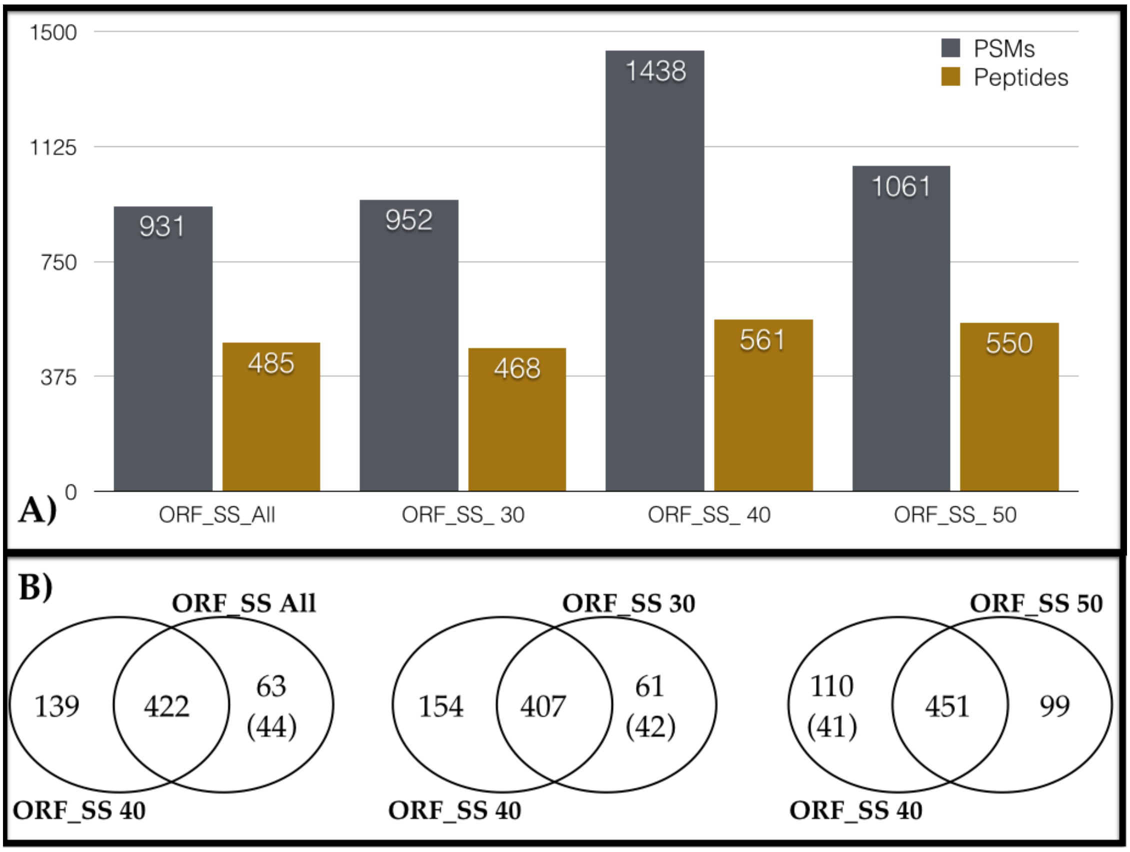
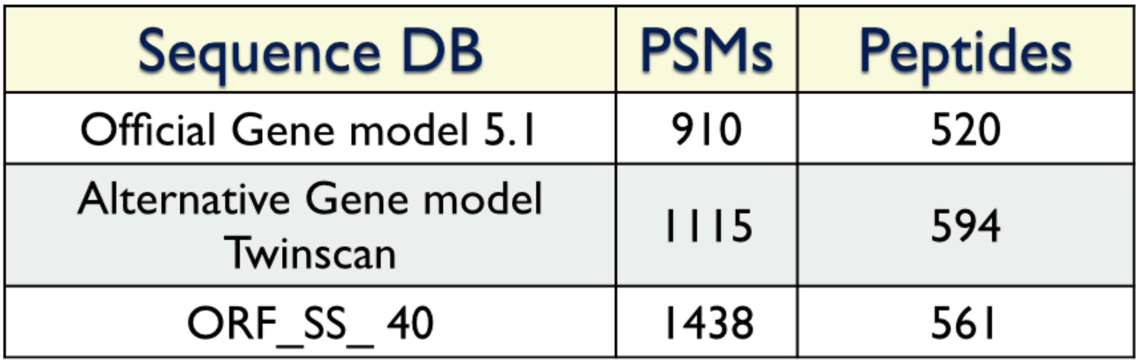


Figure 2:5 A) The bar chart shows the number of identified PSMs at 1% fixed FDR. The redundancy is the removed at peptide level as shown in the last column. The identifications were obtained by querying one spectra file containing 10 different slices from the same 1DE gel. B) The Venn diagrams show the intersection of peptide sequences as identified on the different datasets, as well as the number of peptides unique to the specific dataset. A further study was then conducted to establish if the peptides unique to a dataset were due to being not present in the other database. From ORF_SS ALL to ORF_SS 50, as the minimum length threshold increases, the selection of extracted sequences becomes more stringent. ORF_SS 40 is by construction a subset of ORF_SS; all peptides unique to ORF_SS 40 dataset are present on ORF_SS ALL and ORF_SS 30 databases. Similarly all peptides unique to ORF_SS 50 dataset are also contained an all other databases as they contain also ORF_SS with length below 50 amino acids. As in the diagram shown here, in the comparison between the ORF_SS 40 and ORF_SS ALL, of the 63 peptide sequence uniquely identified on the ORF_SS ALL dataset, only 44 peptides are unique to the ORF_SS ALL database (shorter than 40 amino acids). In the comparison between ORF_SS 30 and ORF_SS 40, of the 61 peptides unique to ORF_SS 30 dataset only 42 are present only on ORF_SS 30 database (minimum length lower than 40 amino acids). In the comparison between ORF_SS 40 and ORF_SS 50, of the 110 unique peptides to ORF_SS 40 only 41 are only present on the ORF_SS 40 database. This shows how statistical evaluation and database design can have direct effect on peptides identified within the score threshold.

The next comparison was the ORF_SS_40 sequence database versus the current available gene model (*N. caninum* 5.1); to provide an insight on the current state of the annotation, as ideally, querying optimal gene models should produce significantly higher number of high confidence PSMs. Table 2:2 shows this comparison and it is clear that the official genome annotation (OGM) has been significantly outperformed by searches against the ORF_SS database, indicating that the current annotation could be improved; it was decided to repeat this comparison on other two organisms *Toxoplasma gondii* and *Cryptosporidium parvum* for which the genome annotation, to our knowledge, had higher quality.

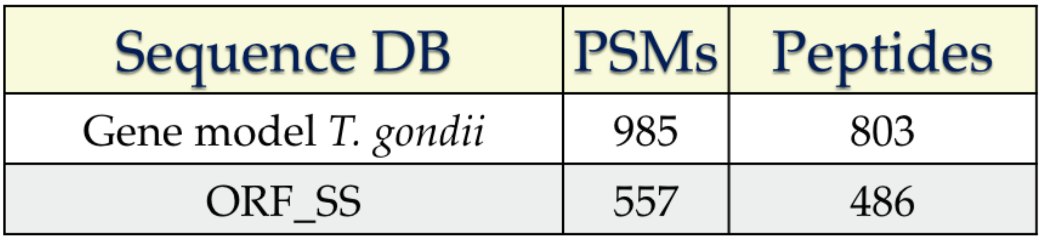


Sequence DB	PSMs	Peptides
Official Gene model 5.1	910	520
Alternative Gene model Twinscan	1115	594
ORF_SS_ 40	1438	561

Table 2:2 The counts of true positive PSMs and the peptide sequences at 1% FDR from 10 slices from the same 1DE gel for *N. caninum*. The results from ORF_SS database score, at PSM level, ~36% better than the official annotation from the same release; at peptide level there is a ~7% gain. The alternative gene model offers similar outcomes (~18% and ~12% gain for PSM and peptide level respectively), indicating that improvements in the gene models are still possible.

The 10 1DE gel slices for *C. parvum* and *T. gondii* were queried against sequence databases and the separate results were merged into one results file. This was post-processed to provide the total number of identified PSMs and the non-redundant total of identified peptides. By comparing the results between different sequence databases, it is possible to see that, for *T. gondii*, ORF_SS 40 yields a lower number of confident PSM and unique peptides at 1% fixed FDR. In *T. gondii* (Figure 2:6), querying the official gene model database yielded larger number of PSMs (increase of ~43.5%) and non-redundant peptides (increase of ~39.4%) compared to queries against the ORF_SS 40 database at 1% fixed FDR. For *C. parvum* this comparison (Figure 2:7) showed a different picture due to the structure of the genome itself as 95% of the genes are single exon genes. The results from querying the official gene model (4.1 release) and the ORF_SS 40 database differ by ~3% for PSMs identification and ~3.6% for non-redundant peptides – with slightly improved performance seen for ORF_SS, potentially indicating some genes have been missed in this annotation.

A)



Sequence DB	PSMs	Peptides
Gene model *T. gondii*	985	803
ORF_SS	557	486

B)

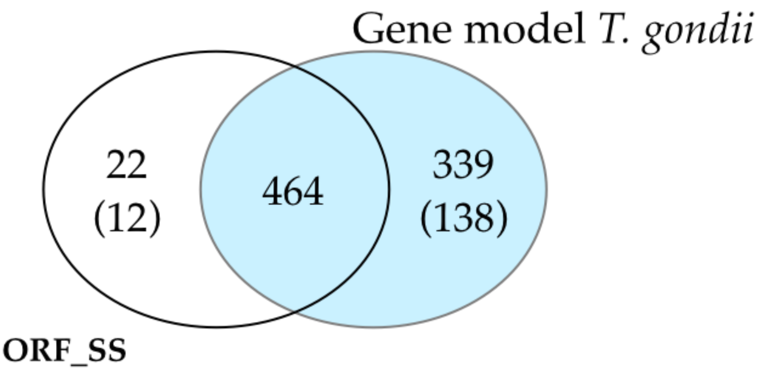
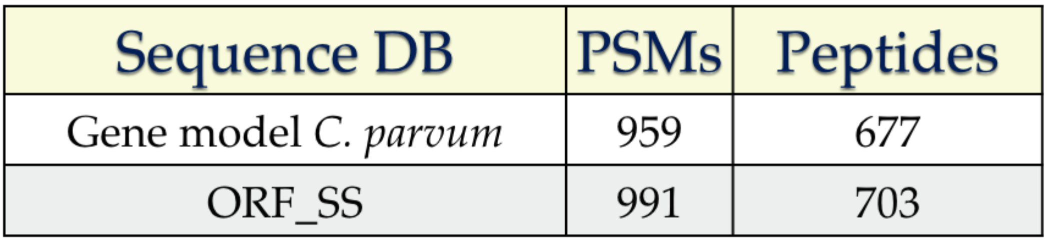


Figure 2:6 A) Merged results from ten 1DE gel slice queried two different sequence databases for *Toxoplasma gondii* (release 5.1): ORF_SS with threshold of 40 amino acid minimum length and official gene models. The number of peptide sequence refers to the non-redundant list of peptides gathered from the PSMs. B) The Venn diagram shows the peptides sequences that are unique identified from this specific dataset. However it was also evaluated whether these unique peptides were also not present on the compared database. Of the 22 peptides unique to ORF_SS dataset, only 12 were missing from the gene model database. Of the 339 peptides unique to the gene model, 138 were unique to the gene model database.

A)



Sequence DB	PSMs	Peptides
Gene model *C. parvum*	959	677
ORF_SS	991	703

B)

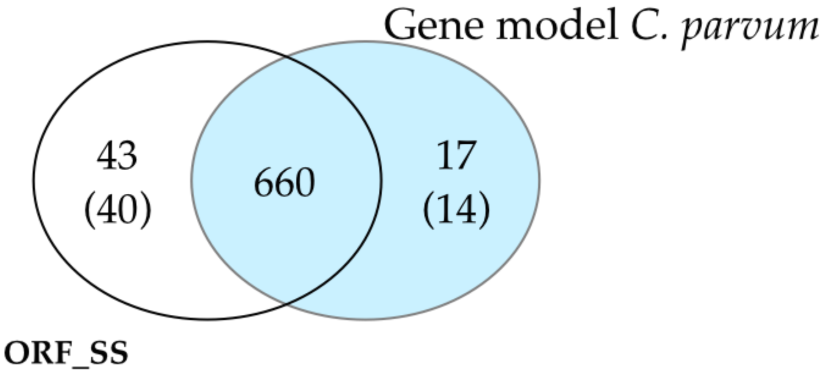
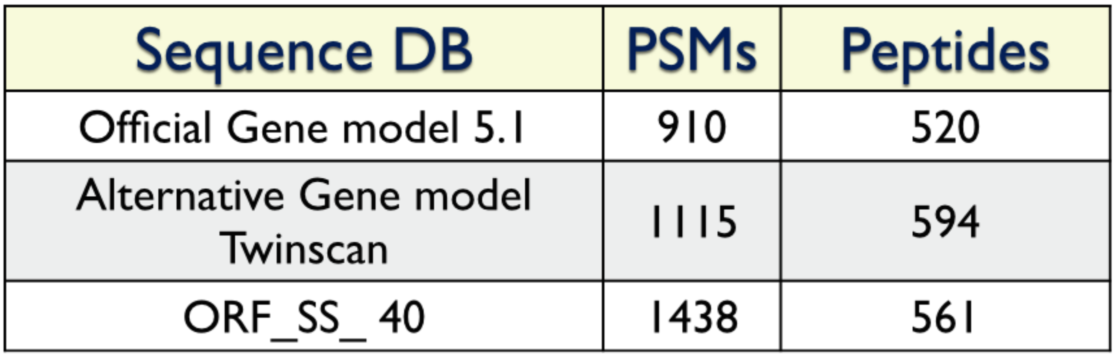


Figure 2:7 A comparison of the combined 10 1DE gel slices from dataset queried against the ORF_SS sequence database and gene models for *Cryptosporidium parvum* (4.1 release). Although genes contain few introns (less than ~5%), the results from ORF_SS database and those obtained with gene model queries differ by ~3% in performance. The Venn diagram shows the intersection of the two dataset and their unique peptides. Of the 43 peptides unique to ORF_SS dataset, 40 were unique to ORF_SS database. Of the 17 peptides unique to gene model dataset, 14 were unique to gene model database.

Together with the official and ORF_SS sequence database, alternative gene models were also tested, downloaded from a public repository (EuPathDB.org for TwinScan). Table 2:2 shows the comparison of PSM and peptide count between the three datasets (official gene model, Twinscan alternative models and ORF_SS); although initial results indicate that ORF_SS may be statistically biased as TwinScan predictions (7588 entries) yielded less PSMs but more unique peptides compared to ORF_SS. This would indicate that further analyses are needed for database design. Alternative gene models were also generated internally by gene finder software GlimmerHMM, trained specifically for this organism (discussed in detail later in the chapter). This achieved the desired results by yielding high confidence PSM results, ranking second after the official gene model 6.0, with only ~10% fewer PSMs on average (in Table 2:3 in the section "Alternative gene model").



Sequence DB	PSMs	Peptides
Official Gene model 5.1	910	520
Alternative Gene model Twinscan	1115	594
ORF_SS_ 40	1438	561

Table 2:2 The table shows the comparison between datasets obtained querying three different database: ORF_SS 40, official gene model and alternative gene model (Twinscan prediction) from *N. caninum* 5.1 release. The dataset was generated by combining 10 1DE gel slices as previously described.

In late August 2009 we were provided with a pre-release of the newly compiled gene annotation for *N. caninum*, then released to the public on EupathDB March 2011. Given the availability it was decided to test whether the new database gave improved results following the same approach. As shown Figure 2:4, it is possible to compare the results from the previous annotation as well as from the ORF_SS databases built from raw genome for both version 5.1 and 6.0. The data is shown (Figure 2:13) and discussed later in the chapter (in paragraph 2.4.5 "Multiple database approach for official model evaluation").

Following this, the same comparison approach was performed on an experiment with a wider scale. All 1-DE gel MS/MS proteomic data available to us, approximately 200,000 spectra, were used to query both the ORF_SS and the annotated proteins database from *N. caninum* release 6.0.

The final dataset for the peptide sequences matched uniquely to the ORF_SS sequence database total 453 unique peptide sequences (from a total of 749 redundant PSMs, listed in Appendix B from Table 7:1 to Table 7:10). While the list with sequences uniquely matched to the official gene model resulted in about

3193 peptide sequences (belonging to 1328 different proteins, data not shown). This appears to be consistent with the improved quality of the newly annotated genome, and the method could be used by curators to evaluate the state of current annotations as well as to use this evidence within on-going annotation processes.

A powerful software package developed by the Welcome Trust, Artemis Genome Viewer [251, 252], can be used to evaluate the peptide sequences, which are unique to the ORF_SS database and how they can be related to current gene models. The gene model sequence with its features is loaded on Artemis, which displays them graphically on the main window. The lower section displays the list of features such as the coordinates of genes, mRNA, exons and coding sequences; above this sits the whole genomic sequence both forward and reverse strand, in nucleotide as well as the three translated frames for each strand. Above it, the same sequences are presented on a much wider range, although the zoom scale and feature highlights can be adjusted to the user preferences; in the images portrayed as examples the stop codons are shown as black pipe symbols and methionines (start codons) with a purple pipe symbol. This feature allows quick recognition of ORF_SS sequences overlaid with the gene model structural sequences such as exons, presented as rectangular shapes situated on the genome and frame respectively, coloured in light blue and those belonging to the same gene are connected by a thin line of the same colour (see example Figure 2:8).

The main objective is to provide the curator with a relatively small dataset that can be feasibly further analysed. These detailed examinations could help explain these sequences correlating them with possible adjustment and correction to the gene model studied. Peptide sequences suggesting a different start of the gene, or the splice sites as well as missed exons would provide strong evidence for reconstructing the gene structure.

In order to assess the reliability of results a few peptide sequences were manually investigated by selecting them by: FDR score, number of PSMs and peptide length. The number of occurrences for a given peptide matched in several spectra would decrease the probability of appearing by chance. Additionally some PSMs with either lower PSM count or with shorter length were inspected depending if they were located nearby official genes.

The first example in Figure 2:8 has two different peptide sequences AFDEAGRTPDGEDGSQTTEQDLR and NLISENVAFPVTDRTGEESR both with FDR 0.0 were matched 11 and 1 time respectively on the same ORF on forward strand on chromosome XI within coordinates 6,061,873 to 6,062,736. Within the same ORF_SS, around 66 amino acids downstream of these PSMs, the first exon of the hypothetical protein-coding gene NCLIV_060250 is located with coordinates from 6,062,455 to 6,063,844 (Figure 2:8). At first it could be noted that upstream of the PSMs there are two potential start codons (purple bar), possibly indicating a different start codon may exist for this gene. However, as a stop codon is present downstream the most 5′ start codon, this cannot be considered as correct start codon.

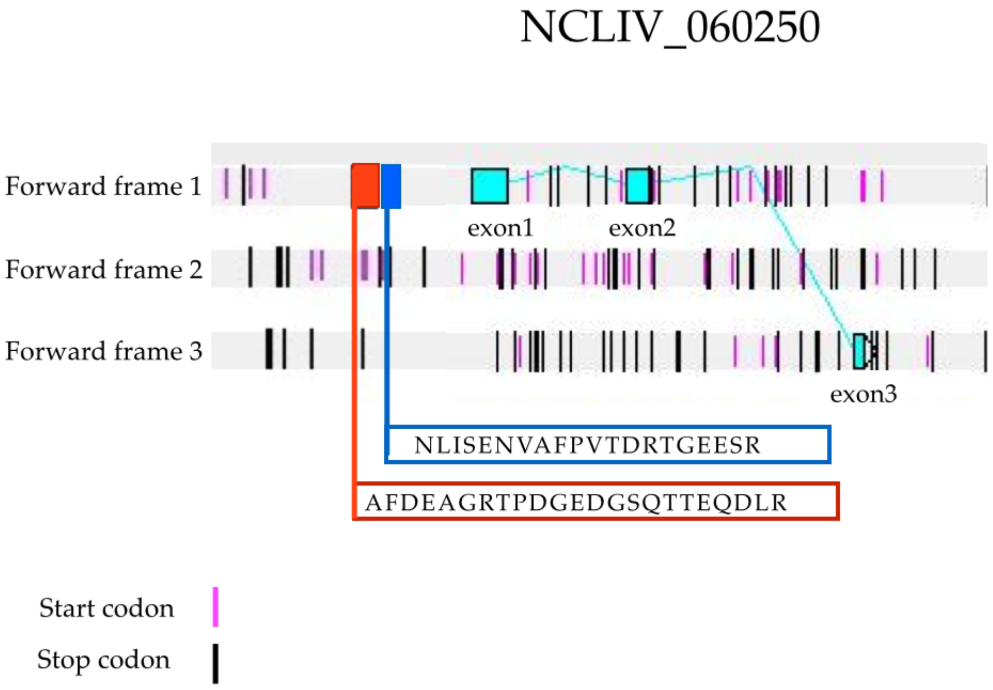


Figure 2:8 The image shows an extract of genomic sequence for the three forward frames for *N. caninum* (from 6,061,873 to 6,063,844 bps) taken with Artemis genome viewer. On each

reading frame potential start and stop codons are highlighted with a purple and black vertical bar respectively. The exons of gene NCLIV_060250 are highlighted in light blue, while the peptide sequences identified from the ORF_SS are highlighted in red and blue respectively. These adjacent peptides sequences (AFDEAGRTPDGEDGSQTTEQDLR and NLISENVAFPVTDRTGEESR respectively) are located on the first forward reading frame and are on the same reading frame of the first exon of the gene NCLIV_060250, which is located downstream. As there are no stop codons between the two peptides and the first exon, this could be further investigated to provide possible evidence for correcting the gene structure of NCLIV_060250.

In other instances PSMs could provide the evidence to adjust and confirm intron/exon boundaries, which are notably challenging to be correctly predicted by gene finders. From the dataset one peptide sequence was observed in more detail: LLRPMEGVPVPER. This was aligned two times against the ORF_SS database, with lowest FDR 0.034. This PSM alignment overlaps with the hypothetical gene NCLIV_057700, sharing 4 codons with the second exon (Figure 2:9).

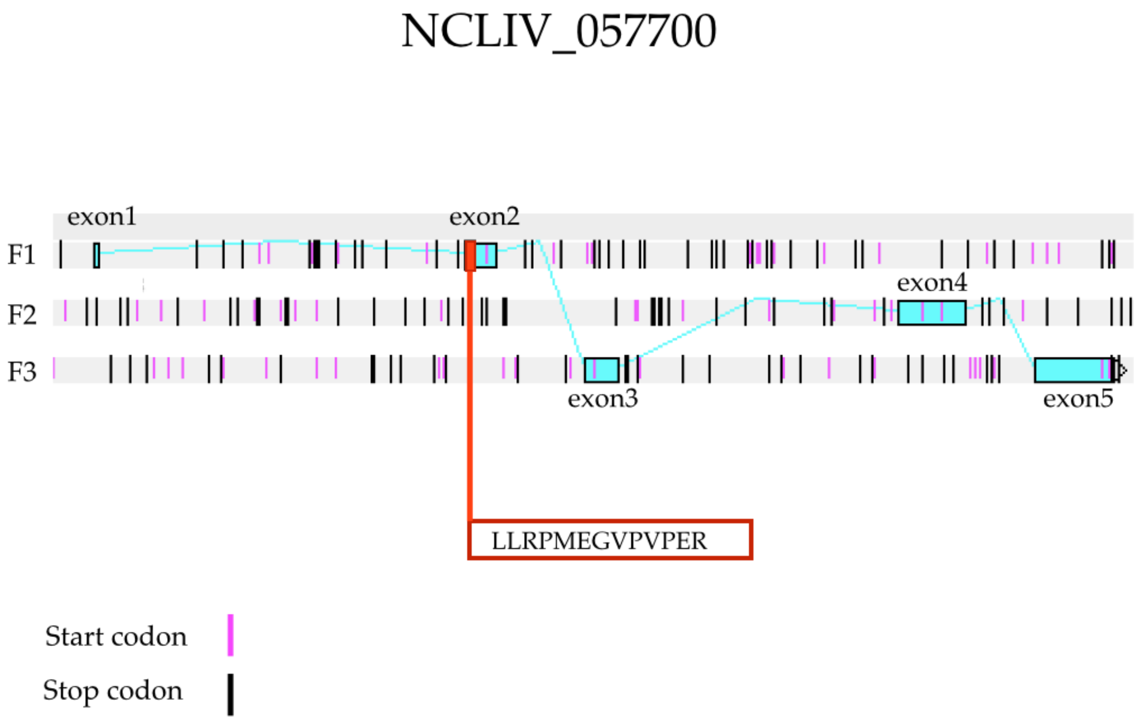


Figure 2:9 The image shows an extract of genomic sequence for the three forward frames for *N. caninum* (from 3,990,610 to 3,994,430 bps) taken with Artemis genome viewer. As in the previous figure, the black and purple vertical bars on each frame indicate the stop and start codon respectively. The gene NCLIV_057700, in light blue, comprises five exons located on the three forward frames. The peptide LLRPMEGVPVPER identified only on ORF_SS aligns with the 5′ of the second exon, suggesting that intron-exon boundaries of this gene structure should be further inspected as potential corrections could be made.

For some cases the PSM evidence from the ORF_SS database could supply the annotator with additional regions to consider while inspecting proteins and their coding sequences for missing exons. An additional peptide sequence (TYCSSPVVNNGDGLVIQLPNAEQK with lowest FDR of 0.0) was matched three times on the ORF_SS database detected in the intronic region 11 codons downstream of the first exon of hypothetical protein NCLIV_068460 on chromosome XII (Figure 2:10).

NCLIV_068460

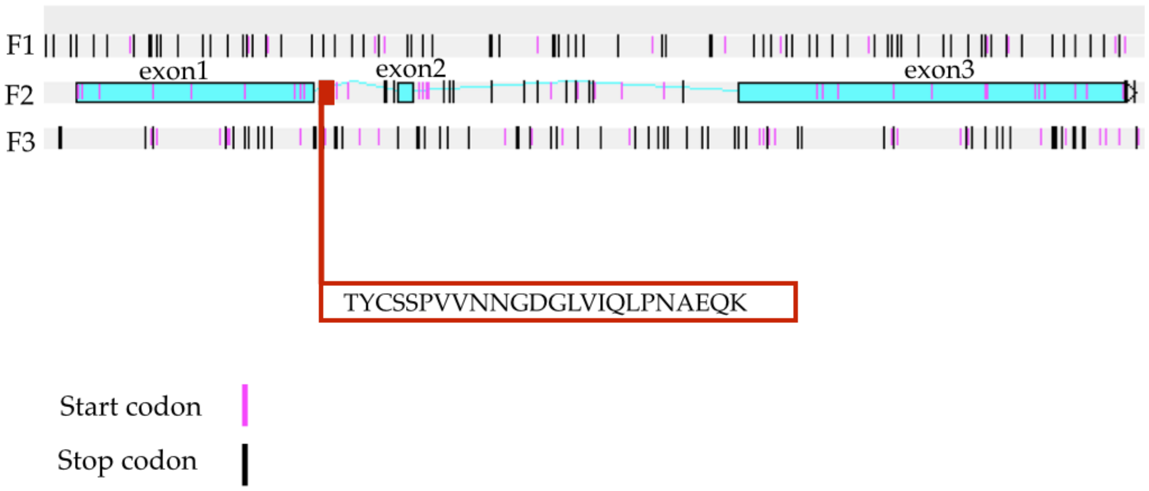


Figure 2:10 The image shows an extract of genomic sequence for the three forward frames for *N. caninum* (from 6,285,470 to 6,290,398 bps) taken with Artemis genome viewer. As in the previous figure, the black and purple vertical bars on each frame indicate the stop and start codon respectively. The structure of the hypothetical gene NCLIV_068460 comprises three exons on the same forward reading frame (second). The peptide sequence identified on the ORF_SS, highlighted in red, is located on the second frame, 11 codons from the 3′ end of the first exon. Both the peptide sequence and the first exon are located within the same ORF_SS; thus this type of identifications, after further inspection, can lead to a correction of the annotation.

2.4.4 Alternative gene models

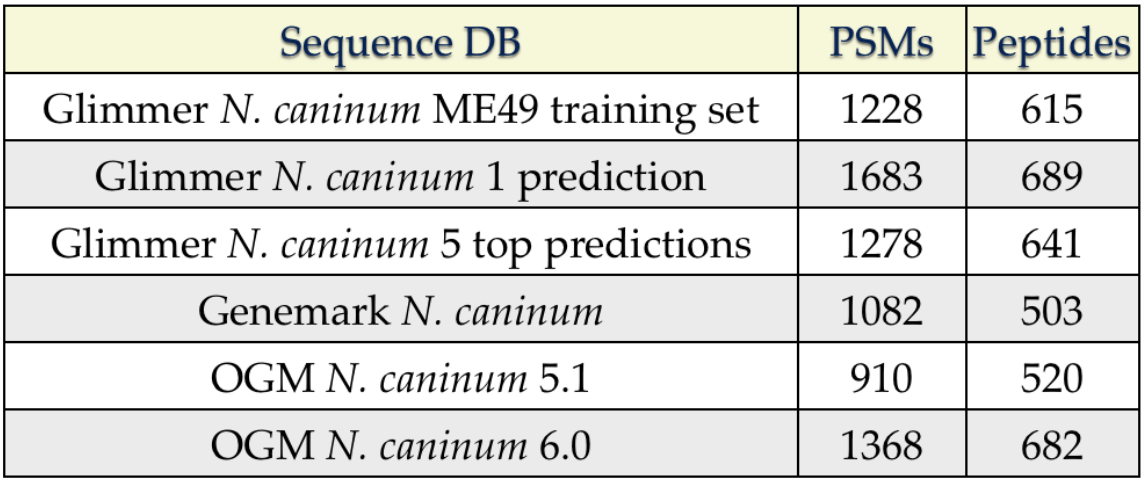
Following the sequence database comparison approach to evaluate both official models and ORF_SS, the attention was directed to the implementation of additional gene models for *N. caninum* that could be exploited. Because of the evolutionary close distance and the similarity of the genomic structure with *Neospora caninum*, a training set for GlimmerHMM was created using genes from *Toxoplasma gondii* (ME49 strain) and this was then used to generate alternative gene models for *N. caninum*.

Sequence database	Amino acid count	Entries
Glimmer *N. caninum* ME49 training set	~1.7 x 10^6	11296
Glimmer *N. caninum* 1 prediction	~1.3 x 10^6	7653
Glimmer *N. caninum* 5 top predictions	~5.6 x 10^6	33084
Genemark *N. caninum*	~3.6 x 10^5	7025
Official Gene model *N. caninum* 5.1	~1 x 10^5	5587
Official Gene model *N. caninum* 6.0	~7 x 10^5	7083

Table 2:3 The table shows how the different gene model predictions differ in size. The *Neospora* prediction "ME49" was based on the training set generated from *T. gondii* official gene models. The other two Glimmer predictions differ in the process used during the assembling of predicted coding sequences. The top first prediction is effectively a subset of the top five which contains four more different predicted splice sequences.

As shown in Table 2:4, the same comparison method was applied to the results obtained by querying these three databases. The two alternative models created using *Neospora* training set are very different in size (i.e. the top five predictions is nearly five times bigger). The third alternative model differs from these two previous models as it was created from a training set based on *T. gondii* (Table 2:3). However different these three alternative models are, they overall lead to rather similar results (supplementary table in appendix B Table 7:11). The comparisons were applied also against the current gene model, the alternative model (Twinscan [57] prediction from ToxoDB.org) and the generated ORF_SS database. From the results in appendix B Table 7:11 it can be seen how the predictions made using a training set for a closely related organism could effectively help in case the gene annotation for the targeted organism had not been well implemented. The training sets that are constructed on either evolutionary related organism or on previous gene model of the same organism; these constructed predictions allowed to identify more PSMs when comparing the ORF_SS sequence database.

Also the Twinscan predictions, downloaded from EuPathDB, were based on the 5.1 version of *Neospora caninum* instead of the later improved 6.0 used elsewhere), which could explain the general low score (Table 7:11), compared to the other databases. In addition, as shown in Table 2:4 and Table 2:5, the difference between the top 5 predictions from GlimmerHMM and the top one only can be considered minimal, hence the overall process could be made faster by generating a smaller set of top predictions (i.e. the top 2 or 3 only).



Sequence DB	PSMs	Peptides
Glimmer *N. caninum* ME49 training set	1228	615
Glimmer *N. caninum* 1 prediction	1683	689
Glimmer *N. caninum* 5 top predictions	1278	641
Genemark *N. caninum*	1082	503
OGM *N. caninum* 5.1	910	520
OGM *N. caninum* 6.0	1368	682

Table 2:4 The comparison of the combined 10 1DE gel slices from *N. caninum* queried against the official and alternative gene model sequence databases. The number of peptide sequence refers to the non-redundant list of peptides gathered from the PSMs at 1% fixed FDR.

Nonetheless the differences between alternative gene models can be visually analysed with the Artemis genome viewer by overlaying them against the genomic sequence. It is possible to see how exon prediction obtained from *Neospora* and *Toxoplasma* training sets respectively was similar (e.g. exon sequences predicted within close genomic regions). However at the same time, the comparison showed some difference in the sequences, such as length differences in predicted exons and splice sites (quantifiable in amino acid count and number of entries in the database). This is also given by different assembly of the predicted exons, which could lead to a future implementation for finding missed sequences (Table 2:4).

In Figure 2:11 it is possible to see an example of sequence similarity between the alternative predictions. The segments highlighted in blue correspond to exons belonging to the *N. caninum* prediction based on *T. gondii* training sets (gene entry 414 in the database); those highlighted in green correspond to the *N. caninum* gene structure predicted based *N. caninum* training set (gene entry 292 in the database). The gene finding algorithm numbers the predicted genes as the exons are predicted and assembled; the differences in numbering can be tracked down to how many genes have been predicted so far. Hence, although geneID 414 and geneID292 show sequence similarities at exon level, *T. gondii* leads to an increase in predicted genes (Figure 2:12). The segments highlighted in yellow correspond to the gene structure of NCLIV_026500 from the official gene model (release 6.0).

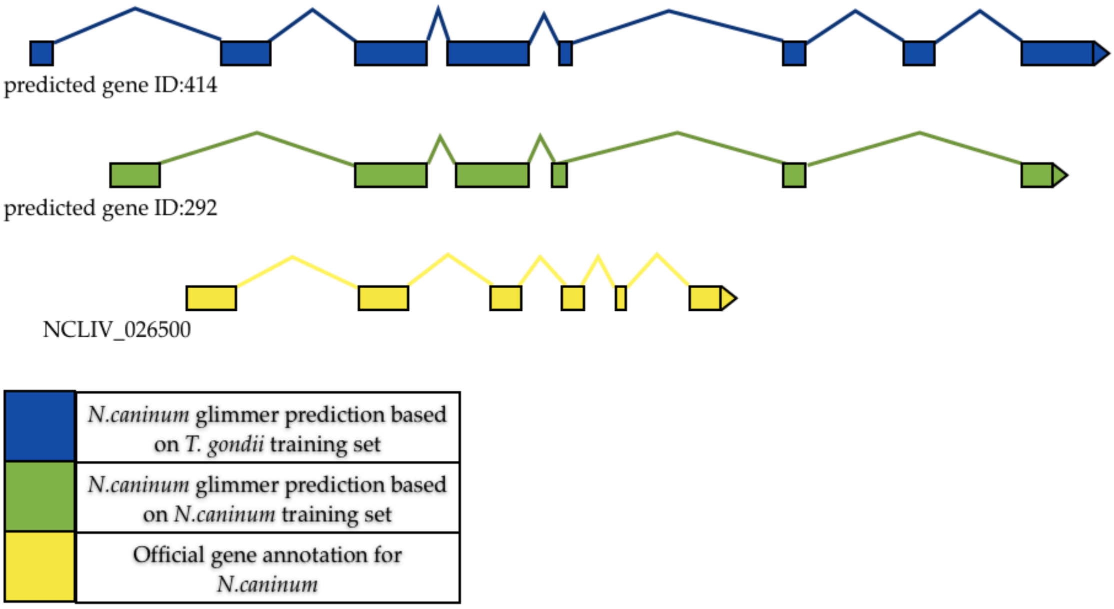

predicted gene ID:414
predicted gene ID:292
NCLIV_026500
N.caninum glimmer prediction based on T. gondii training set
N.caninum glimmer prediction based on N.caninum training set
Official gene annotation for N.caninum

Figure 2:11 Differences between gene models (predicted and official); in blue is the

GlimmerHMM alternative gene model generated using training set based on *T.gondii* ME49, predicted gene ID:414 has 8 exons. In green is the alternative gene model generated with GlimmerHMM using the *N.caninum* training set, the predicted gene ID:292 has 7 exons. In yellow the official annotation is gene NCLIV_026500 with 6 exons. Although the predictions look similar to the official annotation, their intron-exon boundaries and the exon length assessment is still of low quality when compared to the new official annotation release 6.0

In both alternative models, in the same region, a gene has been predicted with similar exon-structure, which was the expected result by generating targeted predictions on *N. caninum* using the statistics from the genomic structure of *T. gondii*.

An additional comparison between different sequence databases was performed on the same spectra file containing 10 concatenated 1DE gel slices from *Neospora caninum*. The query was performed with the multiple search engine mode [148] (OMSSA, X!Tandem and MASCOT). As shown in Figure 2:12, the data generated has clear improvements in TP identification at fixed FDR.

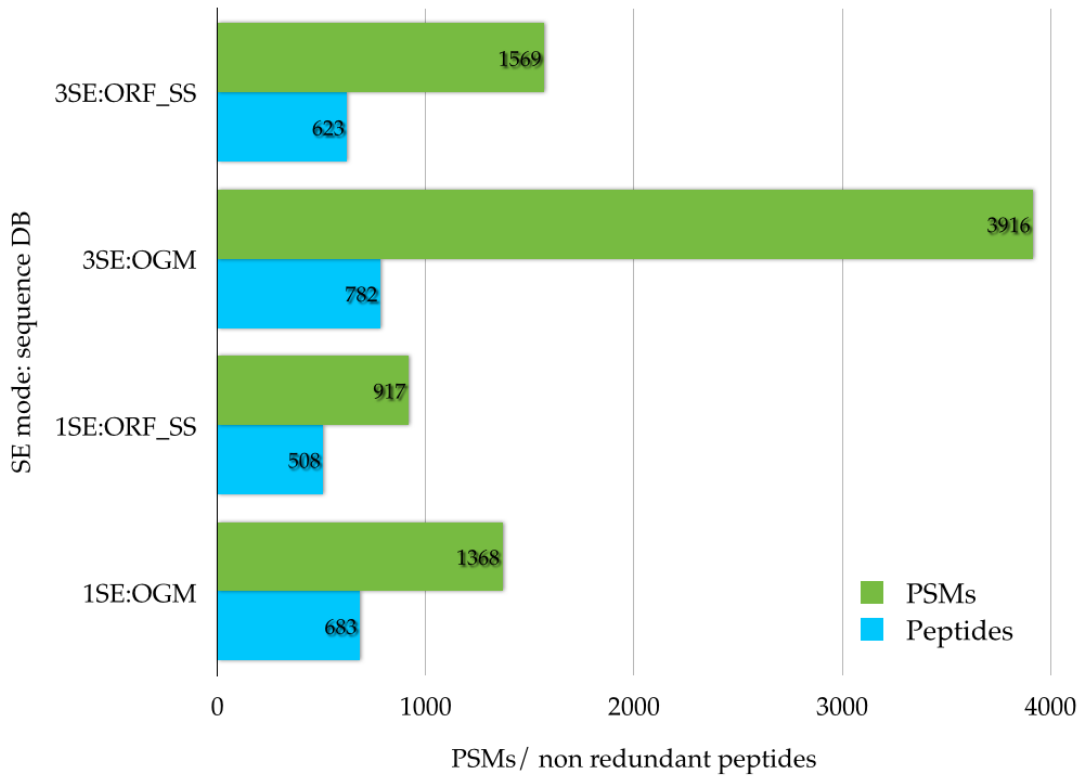


Figure 2:12 The comparison of the combined 10 1DE gel slices from *N. caninum* dataset queried against the ORF_SS sequence database and official gene models (OGM) using a single search engine (MASCOT) and multiple search engines. The number of peptide sequence refers to the non-redundant list of peptides gathered from the PSMs.

2.4.5 Multiple database approach for official model evaluation

The multiple sequence database approach has been used to further assess how it could be exploited to assess which stage the genome annotation for an organism has reached in terms of accuracy.

The study design comprises the generation of ORF_SS sequence databases, as previously discussed, panels of alternative gene models and the official annotation.

The sequence databases concatenated with decoys were searched in two search engine mode (OMSSA and X!Tandem) in order to maximise the true positive

PSMs; at this stage MASCOT was left aside since one of the objectives for this experiment was to create a pipeline that could be automated completely, using open source software.

For each organism, the output from the search engines, for each database, was re-scored by FDR and the individual results from each gel slice were combined into one final output. In the last stage, PSMs redundancy is removed together with false positive identifications. In Figure 2:13 it is possible to view the results for *T.gondii ME49* throughout the genomic sequencing project at 1% fixed FDR, with redundancy removed at PSM level first then at peptide level. The ratio of PSMs in the searches between official annotation (OGM) and ORF_SS changes over time; it varies from 1.4 to 1.9, respectively from releases 3.3 to 6.x, highlighting improvement in gene annotation. The direct comparison between which peptide sequences are unique to either the gene models or ORF_SS database, or commonly found in both datasets is in Figure 2:14.

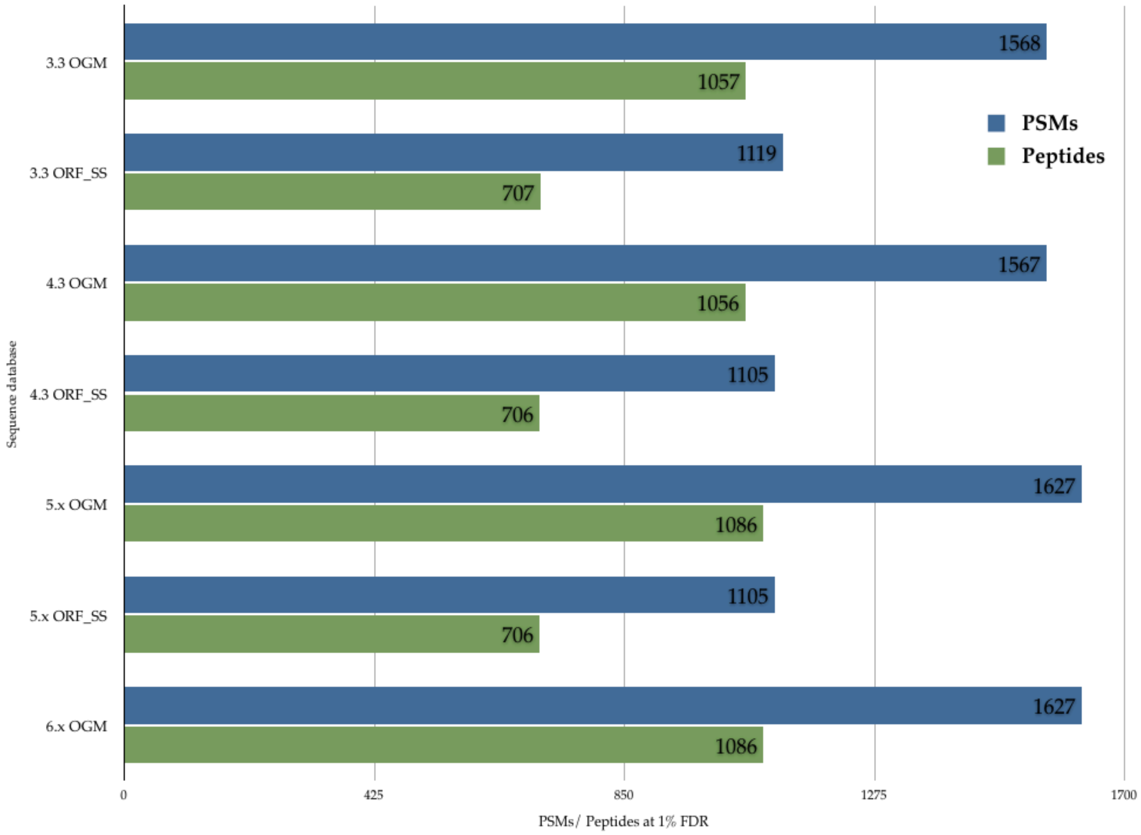


Figure 2:13 The PSMs and peptide counts from the whole dataset of 10 1DE gel slice for each

database query for *T.gondii*. It provides insight on the PSM count as well as non-redundant peptide sequences.

The majority of the peptide sequences that were "uniquely" found a given database searched, were in fact present in both databases (Figure 2:14). This implies that most differences seen in presence/ absence of peptides in the final processed results from these pipelines are due to differential statistical thresholding of results, rather than intrinsic differences in the peptides contained within the databases. As the accuracy of gene models improves, the total number of peptides matched uniquely on the ORF sequences decreases notably. The alternative gene models appear to perform consistently better compared with the ORF_SS database. The Glimmer models (GLM) generated with training sets from previous annotation performed better than the GeneMark gene predictions (gM) (Figure 2:15). Although the predictions were obtained with default parameters on both gene finders, it appears to demonstrate that GlimmerHMM produces higher quality predictions for these species. The final analysis was designed to highlight peptide sequences that appeared uniquely on ORF_SS queries while absent from the gene model from the same release, but were present in the next. On *T. gondii* the comparison was performed on release 3.3 - 4.3, 4.3 – 5.x. However only one peptide sequence resulted uniquely identified on 4.3 and not on the 3.3 gene model, while results from 4.3 against 5.x showed 23 PSMs unique to ORF_SS 4.3 and annotation 5.x, obtained from 8 proteins (Table 2:5).

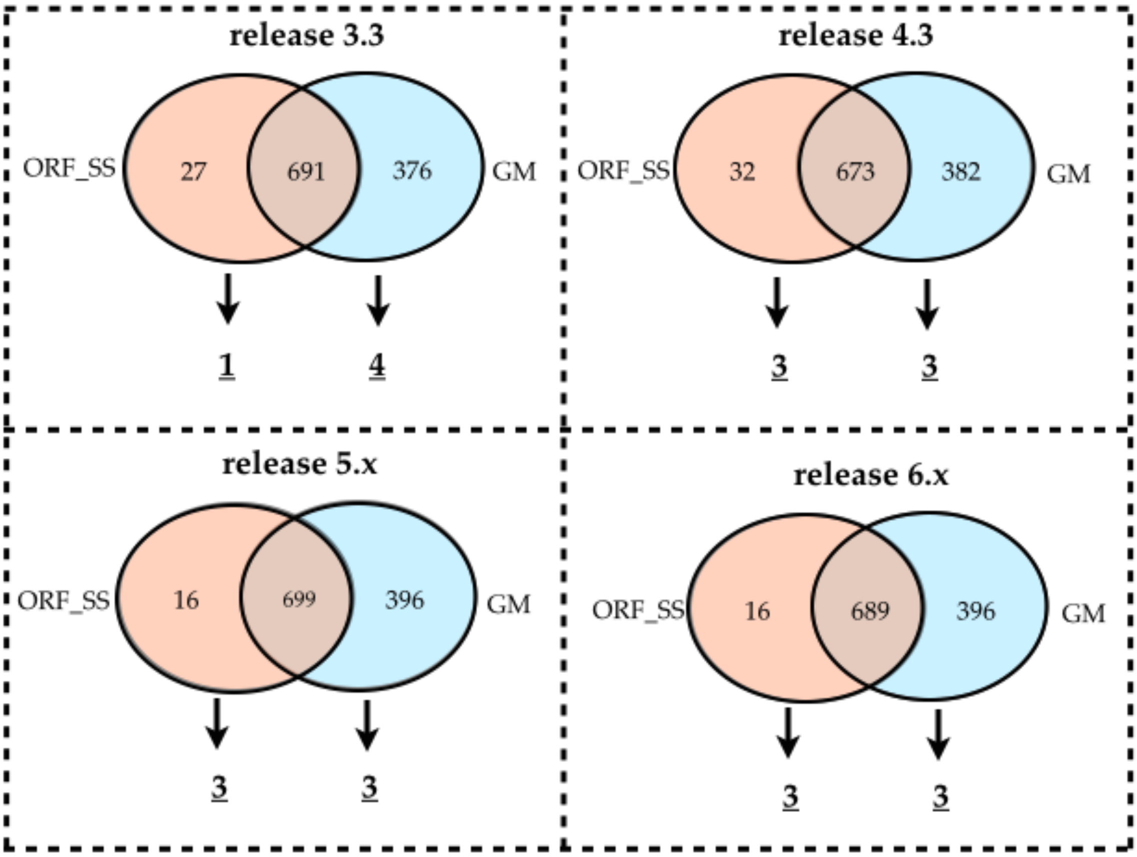


Figure 2:14 Venn diagrams showing the peptide sequences as unique or common to specific datasets searched against sequence databases ORF_SS or official gene model (GM) for *T. gondii*. In clockwise order the datasets are obtained from release 3.3, 4.3, 5.x and 6.x. For each dataset, the arrow indicates how many peptides were truly unique to the sequence database, as statistical biases could lead to peptides being scored outside the threshold window chosen. Compared versus gene models (GM), hits unique to ORF_SS are expected to decrease.

Accession ID	Peptide	PSMs
gb|TGME49_019800	AVVGEEALSPDDLLYLEFTDKFENR	1
gb|TGME49_025320	AAGLASGDSPASR	1
gb|TGME49_026830	QLQEMEAADLR	2
gb|TGME49_029010	SAEGTSESPPVPQLGTPPRPAPR	2
gb|TGME49_029010	VESIIAGSDTTPR	1

<table>
<tr><td>gb|TGME49_029010</td><td>ESSSEDENQPPTTASRPSNGEGESQPPTAAPR</td><td>1</td></tr>
<tr><td>gb|TGME49_067550</td><td>KIEGLSTLSHLR</td><td>1</td></tr>
<tr><td>gb|TGME49_088360</td><td>TPFQDAALSIVKGAACIALSLK</td><td>2</td></tr>
<tr><td>gb|TGME49_088360</td><td>MYQESYLDEFGIPANVK</td><td>2</td></tr>
<tr><td>gb|TGME49_088360</td><td>CSITLGPIEVEPTAAEEALLK</td><td>2</td></tr>
<tr><td>gb|TGME49_088360</td><td>STGQIADIQFESVKFNEEK</td><td>2</td></tr>
<tr><td>gb|TGME49_088360</td><td>LVVLPEWNINANMYPVLK</td><td>2</td></tr>
<tr><td>gb|TGME49_088360</td><td>TPFQDAALSIVK</td><td>1</td></tr>
<tr><td>gb|TGME49_088360</td><td>STGQIADIQFESVK</td><td>1</td></tr>
<tr><td>gb|TGME49_092920</td><td>TPAVVTGFLSSTLR</td><td>1</td></tr>
<tr><td>gb|TGME49_097970</td><td>YGANFLSFVNETGSPYHSVLAVQQR</td><td>1</td></tr>
</table>

Table 2:5 The table PSMs/ peptide sequence results from *T. gondii* from ORF_SS release 4.3 and gene models release 5.x. It shows 23 PSMs (16 unique peptides) that were present in the ORF_SS dataset of release 4.3 but were absent from gene model dataset of the same release. These peptides were also present in the following release 5.x and identified on the official gene model.

At later date, the datasets obtained by querying the two available releases (5.x against 6.x) for official annotation for *N.caninum* were analysed and compared between each other. This provided a first insight into the performance differences between releases and database designs. The ten 1DE gel slices combined were used to query the ORF_SS, the official gene model (OGM) as well as alternative gene models generated with GlimmerHMM (GLM) and GeneMark (gM) Figure 2:15.

By comparing two different releases of gene model databases the annotation improvements are clear, with the average increase of true positive PSMs and unique peptide sequences at 1% fixed FDR up ~50% more than the older release. In the chart (Figure 2:16) it can be noted that the ORF_SS sequence database from the release 5.1 appears within the top ranking databases. This unexpected high scoring in the ORF_SS sequences 5.1 had subsequently been explained with the help of the original curator as *Mycoplasma* contamination in both the genome and proteomic data; the PSMs mapping to *Mycoplasma* were thus excluded in later releases.

The original hypothesis, that ORF_SS database would perform less well when compared to gene models (Figure 2:15), was then tested by comparing it can be seen how in release 6.0, the ORF_SS database performed substantially less well, with an average of ~43% fewer PSMs when compared to the gene model from the same release. This can provide additional evidence that a well-annotated genome would easily outperform any correspondent ORF_SS database, especially for intron rich genomes.

The set of peptides uniquely identified in a given results set to a dataset, was analysed against the presence/ absence of sequence for a given database, to explain the cause (database design or statistical scoring). On release 5.x 126 peptides were uniquely identified on the ORF_SS dataset, of which 122 were actually present only on ORF_SS database (Figure 2:16). The sequences uniquely identified on the ORF_SS on release 5.x were also evaluated against the OGM database for release 6.x in order to assess the confidence of the ORF_SS identifications. Out of these 122 peptides, 103 peptide sequences were present also on the official gene model of the later release 6.x. From these 103 peptides, only 95 were selected for protein identification: the proteins (24) were selected if they contained at least one unique peptide (Table 2:6). The results of this methodology can provide useful for further improvement in designing alternative sequence databases as well as for gene annotation.

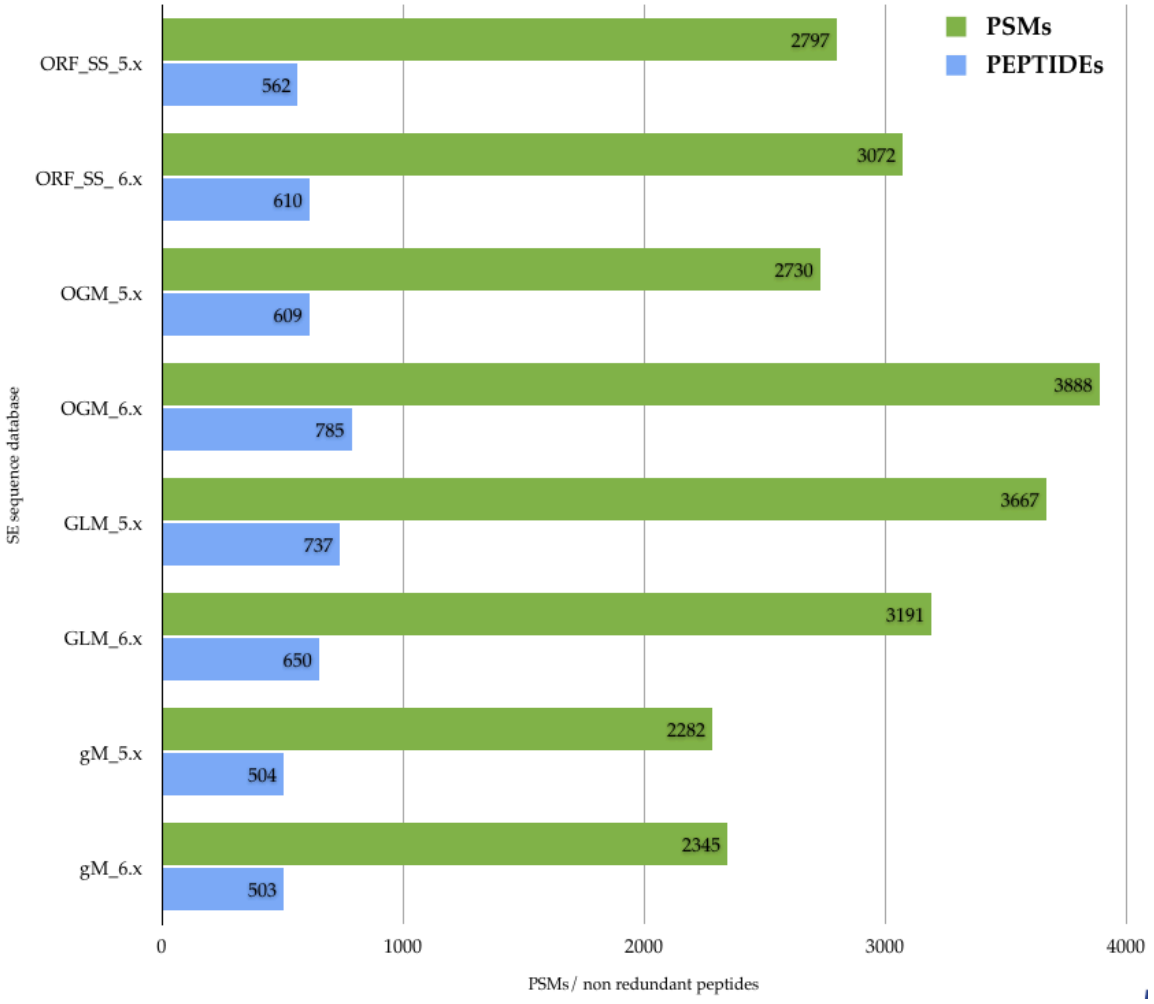


Figure 2:15 The charts shows the PSM and peptide identified at 1% fixed FDR across 2 different gene model/ raw genome sequence releases for *N. caninum*. The spectra file contains the concatenated spectra from 10 slices from the same 1DE gel. The different sequence databases listed are the official annotation (OGM), Glimmer (GLM) and GeneMark (gM) models as well as ORF_SS. It provides insight on the PSM count as well as non-redundant peptide sequences.

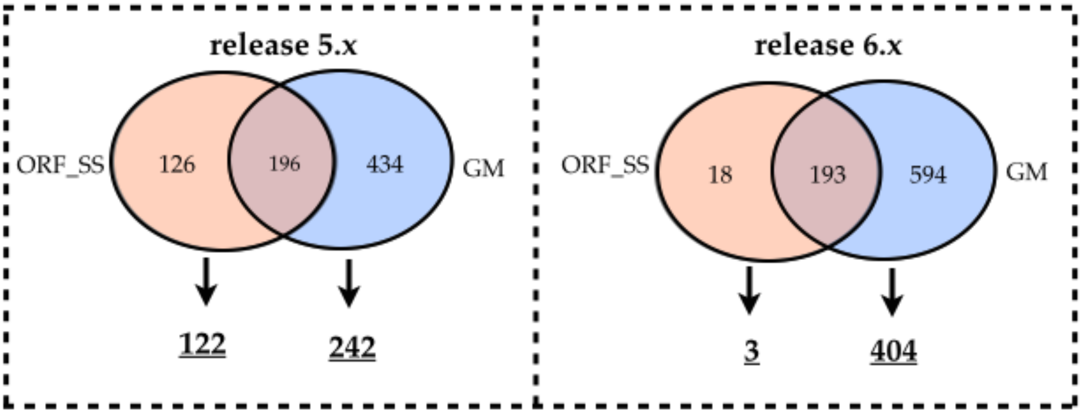


Figure 2:16 Venn diagram of the results from set of *N. caninum* showing the peptide sequences as unique or common to specific datasets searched against sequence databases ORF_SS or official gene model (GM). For each dataset, the arrow indicates how many peptides were truly unique to the sequence database, as statistical biases could lead to peptides being scored outside the threshold window chosen. Compared versus gene models (GM), hits unique to ORF_SS are expected to decrease.

Peptide	ProteinID	PSMs
AGIIPDVLPESACR	NCLIV_002850	1
AVVFLTDPDAPSR	NCLIV_002850	1
AGDIESPQPANDLTECMAR	NCLIV_002940	1
CYLLTAGFSK	NCLIV_002940	1
LYEYPGDLTGSK	NCLIV_002940	3
SGKPDLYSYPGDMTAPR	NCLIV_002940	2
VFVWDYFEK	NCLIV_002940	1
ETAPEIDRNFLADFALTQSR	NCLIV_007770	1
GLVGDMLETGR	NCLIV_007770	1
AIGKDKYACDCPAGYSR	NCLIV_010600	5
CIDDASQPSRYTCECPQDSWR	NCLIV_010600	1
CVQGAEASLAER	NCLIV_010600	1
DAECVEDLNAGGSVR	NCLIV_010600	6
DKYACDCPAGYSR	NCLIV_010600	3
SMTSQSEEKCVQGAEASLAER	NCLIV_010600	1
TGCNAYSEYCNPGR	NCLIV_010600	4
YTCECPQDSWR	NCLIV_010600	1
YTLATDDGTLICAISSEGQPCR	NCLIV_010600	1
LLYLGNTGVAR	NCLIV_011730	1
LTPTLAAFLVK	NCLIV_011730	1
ASYQTLFPK	NCLIV_013260	3
LILDIEKSEEEVVR	NCLIV_013260	4
YACPGEDPNCTETTR	NCLIV_013260	1
CTALDLFMSSPLFAAGRPVSPEPSPGLASEVN	NCLIV_028170	1
FGLAVPLMAGQIR	NCLIV_028170	4
KLPSGFSSEELLNK	NCLIV_028170	2
KNTPEFSPPELVR	NCLIV_028170	2

NTPEFSPPELVR	NCLIV_028170	7
SVMGELEAEDKCVSLVR	NCLIV_028170	2
SVVNFVQVLPVMVCDVQNVR	NCLIV_028170	6
TGQLFLTDFDALVR	NCLIV_028170	5
FGLVQVYTYQPATLK	NCLIV_030820	1
AKPFTDVFPK	NCLIV_033230	14
CPDNSTAVPAALGYPTNR	NCLIV_033230	21
EIPLESLLPGANDSWWSGVDIK	NCLIV_033230	4
EIPLESLLPGANDSWWSGVDIKTGVK	NCLIV_033230	2
FSADWWQGKPDTK	NCLIV_033230	25
FSADWWQGKPDTKDGAK	NCLIV_033230	6
LNHITLKCPDNSTAVPAALGYPTNR	NCLIV_033230	2
LTIPEASFPTTSK	NCLIV_033230	17
LTIPEASFPTTSKSFDVGCVSSDASK	NCLIV_033230	2
SCMVTVTVPPR	NCLIV_033230	17
SEKSPLLVNQVVTCDNEEK	NCLIV_033230	2
SFDVGCVSSDASK	NCLIV_033230	5
SPLLVNQVVTCDNEEK	NCLIV_033230	3
SPLLVNQVVTCDNEEKSSVAVLLSPK	NCLIV_033230	4
SSVAVLLSPK	NCLIV_033230	12
SVSSPEVYCTVQVEAER	NCLIV_033230	8
SVSSPEVYCTVQVEAERASAGIK	NCLIV_033230	11
TGVKLTIPEASFPTTSK	NCLIV_033230	2
DKGETGGENGDSPVLR	NCLIV_033250	14
ESEVIGQVAHCAYSSNVR	NCLIV_033250	21
EWVTGTLQQGIK	NCLIV_033250	7
GEASGVAGATLTIPKDQ	NCLIV_033250	4
IGQVAHCAYSSNVR	NCLIV_033250	1

IPDEHYPATSK	NCLIV_033250	1
ITIPDEHYPATSK	NCLIV_033250	15
ITIPDEHYPATSKAFR	NCLIV_033250	1
KEWVTGTLQQGIK	NCLIV_033250	8
LLSEDDGLIVCNESDGEDECEK	NCLIV_033250	1
LRPITVNPENNGVTLICGPDGK	NCLIV_033250	9
NAAPLSTFLPGAK	NCLIV_033250	7
NAAPLSTFLPGAKK	NCLIV_033250	5
NVCLLNVYVQSR	NCLIV_033250	5
SENEKFTCLPK	NCLIV_033250	3
EFKLPTESYVAPPVWIR	NCLIV_035220	1
SISTAIQVGQAGAALVQNFVHR	NCLIV_041120	1
HDELSQLIK	NCLIV_043270	2
HDELSQLIKEGVVR	NCLIV_043270	1
YCSGFQAAANSYCNK	NCLIV_043270	1
YCSGFQAAANSYCNKR	NCLIV_043270	1
EAFPLNNGSCDTAK	NCLIV_043760	2
GAKPWAELFPGADK	NCLIV_043760	4
HKLEVGETCTIEMLPQNSK	NCLIV_043760	2
LEVGETCTIEMLPQNSK	NCLIV_043760	1
NFAFATTTSSSSLILK	NCLIV_043760	2
VPPHGDGQGFCFILR	NCLIV_043760	4
VTVEPEQLEKEAFPLNNGSCDTAK	NCLIV_043760	1
ARVPFSGYGQEK	NCLIV_045870	1
GLAGLIAAVAVLAAR	NCLIV_045870	3
MRNPPKTFMDEIK	NCLIV_045870	2
TASLFVALPAALFSAVFLSK	NCLIV_045870	1
VPFSGYGQEK	NCLIV_045870	3

ILDLAPSFGEGPAEIVSR	NCLIV_046530	1
FVDPQSSSLIGR	NCLIV_047390	1
SSLPLFIVLPSEHLR	NCLIV_048040	1
MEEADDAPKPVPVR	NCLIV_052880	2
EEVSELNTVLMR	NCLIV_056300	1
EFKNEDEIANAVASLLYTVPAAVAELSAGYR	NCLIV_056300	2
VRPIQLDDIAAVLYGSDPR	NCLIV_059000	1
ITSQHTLFMPDGR	NCLIV_060730	1
LLPMSAIQTPEFR	NCLIV_060730	2
LLPMSAIQTPEFR	NCLIV_060740	2
QADGTVYPLITLPK	NCLIV_062280	2
TLDAPTSGSASFEVAQR	NCLIV_062280	1

Table 2:6 The table presents the peptide sequences that were identified on ORF_SS database but absent from official gene model (*N.caninum* release 5.x). These peptides were however present on the official gene model of the following release 6.0. This list contains the protein identified with one unique peptide at least.

2.5 Conclusion

The study discussed in this chapter focused on alternative database designs for proteogenomics. The results confirmed the importance in examining closely related species and their gene structure in order to build sequence databases. By comparing the identifications from different datasets resulting from ORF_SS and official annotation databases it is possible to give an insight into the quality of gene models using proteomic data. However careful attention needs to be paid at whether uniqueness of a peptide identified is due to statistical scoring or to database design. This approach can provide a shortlist of peptides that do not appear in the annotation but are identified on other database and as such could be further investigated.

The ORF_SS database was designed after analyses on the genomic structure, showing that the length filter could be a variable factor to be tweaked itself during the design. The ratio of hits to ORF_SS versus official annotation could be considered as a possible framework for assessing the quality of the gene models. A future study on intron patterns and protease specificity could potentially yield additional variables for selecting/ discarding ORF_SS during the extraction (i.e. ORF_SS containing a specific pattern would be flagged, long ORF_SS not containing a proteolytic site would be discarded).

Additionally, as alternative gene models are compared against official annotation, we show the importance of generating them with different gene finding software. The algorithms of these, although sharing similarities, interpret the genomic sequence slightly differently. With large panels of alternative gene models it could be possible to increase the confidence of gene structures by evaluating the PSMs common between different models.

However the number of search engine runs, together with some sizable sequence database and the post-processing needed could be considered as one of the constraints of the approach here described. Ideally both the creation of alternative gene models and ORF_SS and the post-processing algorithms need to be fully automated.

3 Intron Spanning Peptides: identification by a blind search strategy

3.1 Abstract

At present, genome annotation is performed mainly with computational predictions. In organisms having mostly multi exon genes, experimental validation is still needed to confirm the gene structure is correctly predicted. MS/MS proteomic validation is crucial to gene annotation with database dependent approaches; however it can prove successful only if the analysed peptide sequences have been correctly predicted, and hence present in the sequence database. Although *de novo* sequencing could help identify these intron spanning peptides from the tandem mass spectra alone, this approach is yet not reliable for full length peptide predictions. Because of this, the task of confirming the intron-exon splice sites and intron spanning peptides is particularly challenging.

This chapter presents a novel approach in identifying these intron spanning peptides (ISPs) by attempting to overcome bioinformatic challenges. This approach, comprising both database dependent and *de novo* sequencing, was designed to provide *de novo* ISPs identification; this new hybrid approach was tested on *T. gondii*.

The hybrid pipeline comprises four main stages: (I) firstly standard sequence database queries are used to filter spectra that are easily explained and highlight genomic regions where PSMs have been identified. (II) Next InSPecT *de* novo algorithm is used to predict short amino acid sequences (TAG) from the spectra that have yet to be explained. These TAGs are then used to anchor the partial spectrum on previously selected genomic regions. (III) The full peptide sequence is then reconstructed by calculating all the possible combinations and splice position from nearby potential cleavage sites. (IV) Finally each spectrum is queried against a custom database of candidate intron spanning peptides (using

OMSSA). From a small dataset of known ISPs the highest identification rate achieved was ~26% sensitivity at 5% fixed FDR. The results as yet do not demonstrate this method is ready for deployment in proteogenomics pipelines. However further analysis of data reveal that accurate TAG generation remains a bottleneck. Thus improvements in this step may help to provide a new tool for proteogenomics.

3.2 Introduction

In a proteomic study the peptide sequences that align onto two exons are of particular importance [183]. These intron-spanning peptides (ISPs) determine how splicing occurs and can confirm the different isoforms generated through alternative splicing. With traditional bioinformatic approaches it is possible to correctly identify the splice site only if it has been predicted in the sequence database [198, 253, 254]. In this chapter, I describe a novel approach for identifying ISPs. The pipeline combines several stages using the output of InSPecT hybrid and *de novo* mode and the OMSSA search engine for the final analysis. InSPecT hybrid mode analyses the data in two stages: firstly, with a *de* novo algorithm, it predicts peptide spectrum tags (TAGs) along with their relevant *prefix* and *suffix mass*. Each TAG can be described as a short sequence of amino acids that correspond to the calculated acyclic path connecting spectrum peaks. Their prefix and suffix mass corresponds to b- and y- ions, thus providing the position of the TAG (and its forming amino acids) within the spectrum. InSPecT hybrid algorithm uses then the TAG data and the parent ion mass to perform the second stage: pre-filtering the sequence database based on predicted TAG data. The second stage concludes with a database search approach on the restricted database to identify PSMs.

As discussed in this chapter, the designed pipeline analyses the mass spectra data in multiple passes (Figure 3:1). Initially standard database searches (e.g. against ORF_SS) provide information such as which spectra have been identified (PSMs) and, inherently, highlight where, on the genome, these identifications have been made. The assumption of this approach is that the returned genomic regions represent the presence of genes encoding the identified proteins. Then *de novo* sequencing (InSPecT tag only mode) is exploited, on the spectra that have yet to be identified, to generate sets of interpreted TAGs. The designed algorithm attempts to align these against the genomic regions that were previously selected. When aligned, these TAGs are then processed to provide all possible full length peptide sequences that originated the spectrum. A final database search (OMSSA) is then used to identify the PSM from the list of full length peptide candidates.

The pipeline described here includes both InSPecT and OMSSA. An earlier design comprised only InSPecT algorithms in an attempt to simplify the process. However, as the final stage of the pipeline consists of a search of a temporary sequence database, OMSSA search engine was ultimately chosen, as test results proved its algorithm and scoring appears to function adequately with unusually small database sizes (as produced here), which was not the case for InSPecT hybrid mode or X!TANDEM (data not shown).

The development of this pipeline offered the opportunity to evaluate other *de novo* algorithms (e.g. PepNovo) available at the time of this study. The test dataset used for ISP pipeline development was also used in this comparison between different approaches. A panel of alternative gene models, generated with different gene finders, was compared against the results of the ISP pipeline and PepNovo. This confirmed that *de novo* algorithms still struggle to identify the full length peptide sequences; in addition the results show how different gene finders can provide divergent gene predictions and how this could affect the predicted exon-exon structure [166].

3.3 Methods

The pipeline comprises four main stages:

1. Database queries generate a list of identified spectra that can highlight genomic regions of interest;

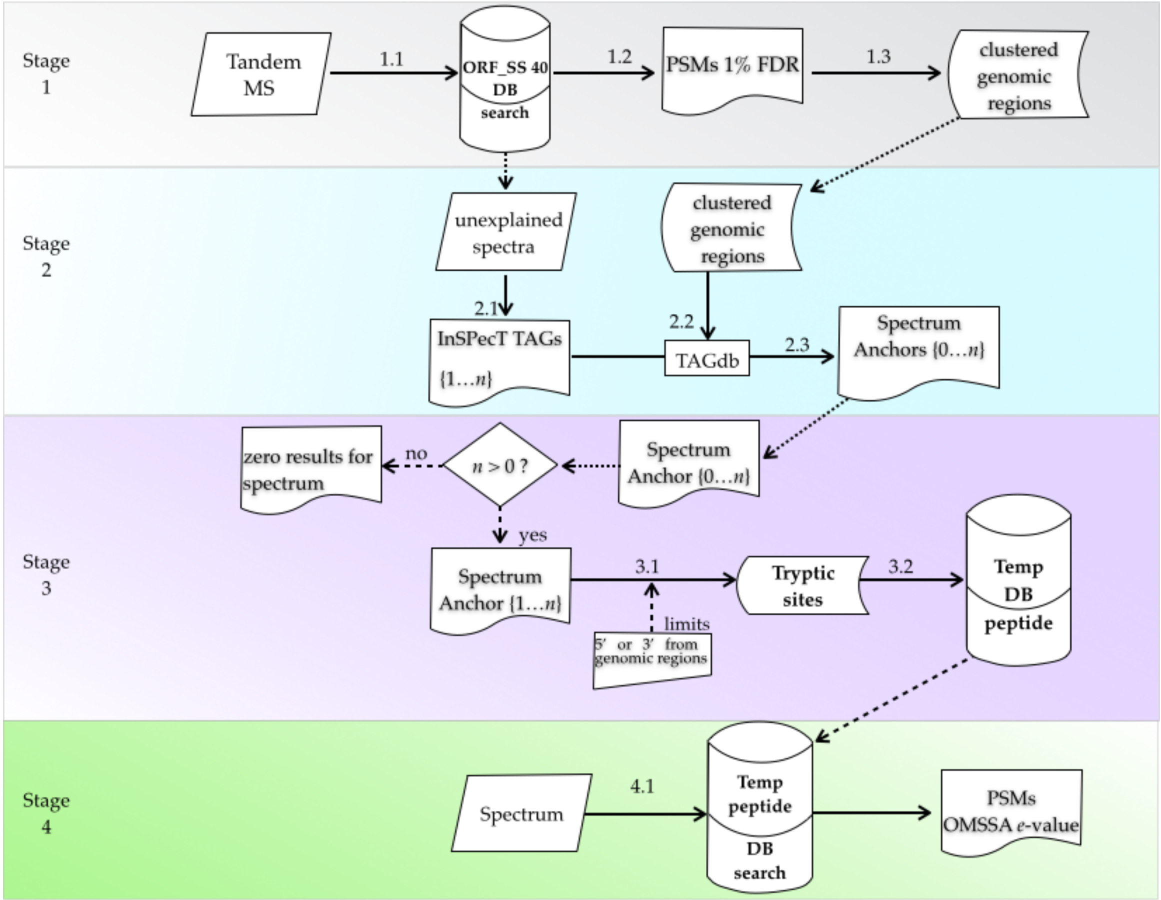
2. Unexplained spectra are processed by InSPecT and generated TAGs are mapped within genomic clusters providing spectrum anchors;
3. Reconstruction of full length peptides with all possible splice sites stored in a temporary DB;
4. The temporary DB is queried with the spectrum;

The workflow in Figure 3:1 shows these four stages, fully discussed in detail in each subsection of the methods. During stage 1.1, tandem mass spectra are analysed with a standard database search engine against ORF_SS sequence database. The identified PSMs are rescored by FDR (stage 1.2) and they are clustered together based on their position on the genome (stage 1.3). These clusters represent the genomic regions ("filtered" database) for the following analyses in the second stage.

In stage 2.1, InSPecT analyses the spectra that have not been identified in stage 1.1, generating sets of interpreted TAGs. In stage 2.2 the algorithm aligns these TAGs against the indexed genome (TAGdb) within the genomic regions selected from stage 1.3. Then the aligned TAGs are used for reconstructing the partial peptide sequence: the *spectrum anchor* (stage 2.3).

During the third stage, the algorithm exploits nearby tryptic sites (stage 3.1) and retrieves the full length peptide sequences. During stage 3.2 an additional process allows for calculating the expected but unknown splice site within the peptide sequence. This data, loaded into a temporary database, is ultimately searched with OMSSA search engine in the fourth and final stage of the pipeline.

The initial approach to the pipeline design included *de novo* sequencing and sequence alignments that followed from step 3.2 (see Figure 3:1). The generated full-length peptides were then aligned against the *de novo* predictions to identify the splice site and shortlist ISP candidate sequences (see Appendix A for methods and results). As this proved to be inefficient it was soon replaced with different algorithms described in this chapter.



	Stage steps
1.1	InSPecT hybrid search against ORF_SS
1.2	PSMs re-scored by FDR at peptide level
1.3	PSMs clustered by their genomic coordinates
2.1	Unexplained spectra queried by InSPecT in TAG only mode (equal to PSTs)
2.2	InSPecT TAGs are mapped onto the TAGdb if contained within identified genomic regions
2.3	The mapped TAGs are used to reconstruct the partial spectrum
3.1	Unidentified mass of the spectrum anchor is extracted from nearby tryptic sites (based on precise limits)
3.2	Full length peptide reconstruction (temporary DB peptide) is performed to include all possible splice sites
4.1	With OMSSA the spectrum is searched against the temporary DB peptide

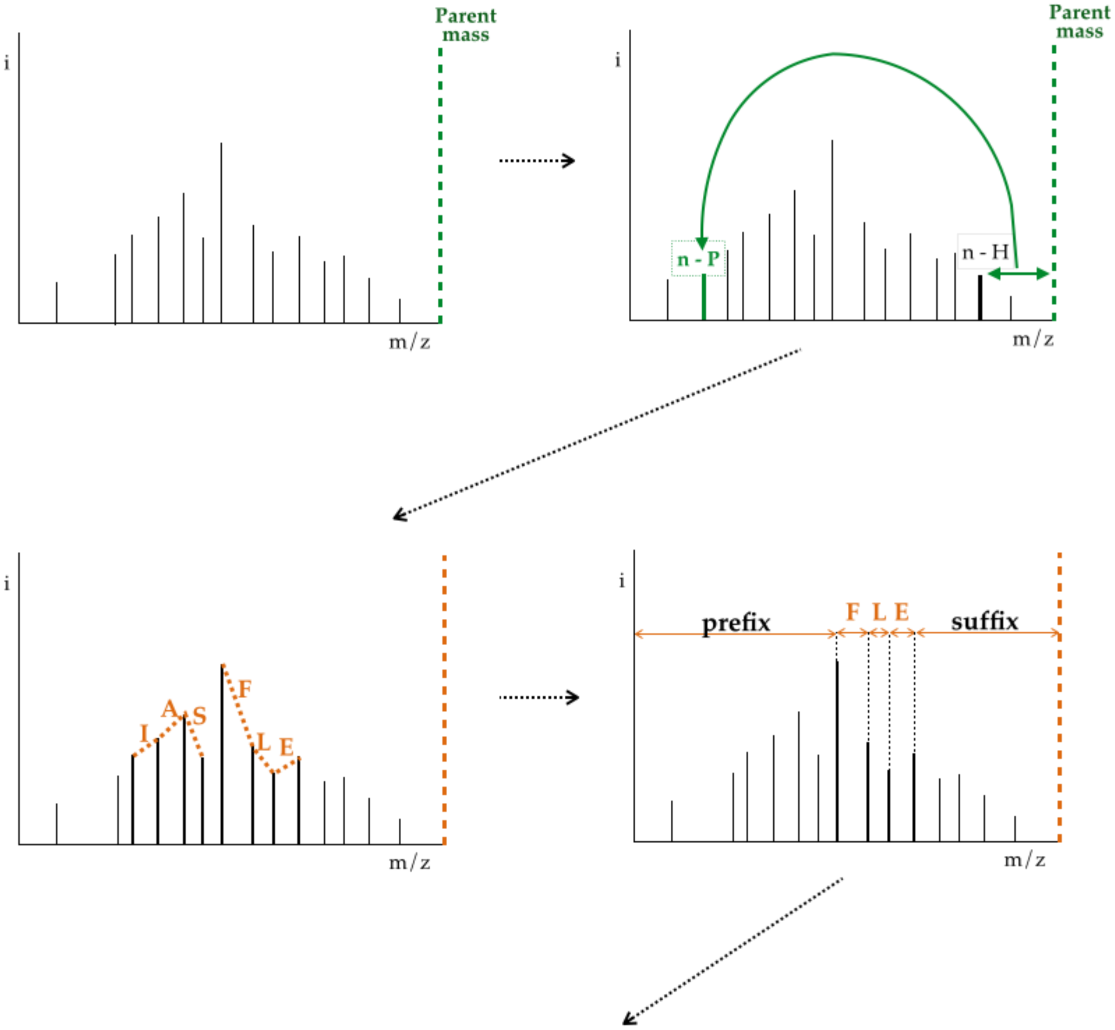
	Terminology
ORF_SS 40	sequence DB used for initial search engine
Unexplained spectra	Spectra that failed to be identified by search engine
TAGdb	whole genomic index based on 3 amino acid TAG
genomic region	selected area comprised within identified PSMs
spectrum anchor	partial peptide sequence with matched either N- or C- terminal

Figure 3:1 The pipeline workflow begins with a standard database search in stage 1.1; here the sequence are ORF_SS longer than 40 amino acids. This step provide a list of PSMs that are rescored by 1% FDR in stage 1.2 and then, in 1.3, these PSMs are clustered together based on their genomic locus. The first stage also filters the unexplained spectra, which in stage 2.1 are processed by InSPecT to yield a set of TAGs. Each TAG is mapped onto TAGdb within the clustered genomic regions in 2.2 and subsequently, the pipeline algorithm generate a set of spectrum anchors in stage 2.3. During the third stage the algorithms uses the nearby tryptic

sites and the unexplained spectrum mass (either N- or C terminal) to reconstruct all possible full-length peptides (stage 3.1). The algorithm uses both the PSMs position within their genomic region and a bps range in order to select the nearby tryptic sites for the unexplained terminus mass. In stage 3.2 all possible splice sites are calculated for each full-length peptide based on their nucleotide sequence from stage 3.1. Before storing this data in a temporary sequence database, the algorithm evaluates all full length reconstructions and discards those that do not meet precise conditions (must be being tryptic peptides and without stop codons inside). The final stage of the pipeline 4.1 comprises a database search with OMSSA of the spectrum used to generate the TAGs; the final results are assessed by their e-value.

The InSPecT *de novo* algorithms are here explained in detail. The algorithms, during a pre-processing step, remove the low intensity peaks by comparing each peak against all peaks within a 25 Da range and retaining only the top six. Then, in the acyclic graph, InSPecT algorithms compute all the possible paths that connect the assessed nodes. As in the paper from Tanner *et al* [171] the algorithm calculates the nodes, from each peak of mass M within parent mass P, as a possible result from b ions $(M - |H|)$ and y- $(P - M)$ ions and it assigns these a prefix residue mass (PRM).

The PRM score $S(N)$ is derived from: i) intensity rank, ii) PRM supporting evidence and iii) isotope pattern. The intensity rank score $S_1(N)$ assesses the probability that the node is the result of either a b or y ion against the probability that it was generated by a random event. The PRM supporting evidence score $S_2(N)$ lists a set of identified ions that would make the node a true positive node. This ion set includes the loss of ammonia and water for a-, b- and y- ions and the doubly charged b- and y- ions. Then the log score $S_2(N)$ is calculated based on probabilistic evaluation that a given peak is randomly matched with b- or y- ions. The third score, isotope pattern $S_3(N)$, is the probability that the calculated relative intensity belongs to a specific isotopic peak. In other words this classification relates directly to the likelihood that given peak belongs to b- or y- ions. The amino acid residues are given by the mass difference of two adjacent nodes, allowing for PTMs. Each edge score $S(E)$ is calculated from the difference between the expected mass and the edge length (Figure 3:2).



GVHPQLIASSFLEASKQSEK

PRM	Prefix Da	TAG	Suffix Da	Expected parent Mass	Spectrum
87.67	989.38	FLE	758.45	2137.02506	44.3993.3993.3.dta
86.06	1136.54	LEA	687.37	2137.073764	44.3993.3993.3.dta
74.33	1249.66	EAS	600.26	2137.031733	44.3993.3993.3.dta
73.72	1093.53	REA	687.37	2137.080814	44.3993.3993.3.dta
52.87	1135.2	LSL	687.37	2135.770149	44.3993.3993.3.dta

Figure 3:2 This workflow shows how InSPecT algorithm generates the set of *de novo* predictions. In a pre-processing stage InSPecT removes the low intensity peaks (by retaining only the top six peaks within 25 Da range) and estimates the spectrum charge (unless the value is in the source file, or manually set for multiple charges to be evaluate). Then InSPecT, evaluating each peak n as a node, computes whether the peak was generated by b or y ion: $N - H$ (as prefix) and $P - N$ (as suffix), where P is the parent mass. An acyclic path is drawn to connect the nodes, based on their respective prefix and suffix; if the mass difference of

adjacent nodes in the path is equal to amino acid residue mass, this is selected as part of the generating short TAG. Here the short extract of generated TAGs reports the respective TAG with the score (PRM) together with the expected prefix, suffix and parent mass.

When generating the set of TAGs InSPecT provides the score for each identified nodes N_k and edges E_k, and retains the top 50 – 100 TAGS, as per Tanner *et al* paper:

$$\sum_{k=1}^{n}\left(S_1(N_k) + S_2(N_k) + S_3(N_k)\right) + \sum_{k=1}^{n} S(E_k) \text{ [171]}$$

When using InSPecT hybrid mode, the algorithms use both the set of generated TAGs and the parent mass to reduce the search space on the database (Figure 3:2). Then the sequence database search on this pre-filtered database also attempts to identify a number of PTMs. The InSPecT algorithms attempt to restrict the database by mapping each TAG together with both their N- and C-terminus (i.e. the full peptide sequence with no splice sites). However this designed pipeline attempts to map the predicted TAG against the genome index, TAGdb, by allowing either of its termini to match (N-terminal as prefix or C-terminal as suffix) Figure 3:1. The theory is that spliced peptides cannot be identified from an ORF database. By first mapping either terminus together with the TAG, it would be possible to then search for the unexplained terminus mass and find the correct splice site. However this type of alignment of TAGs against the genome is challenging because unexpected PTMs can alter prefix/suffix masses; also in particular cases the splice site can be present within the TAG itself, which would make the genomic alignment impossible.

3.3.1 Genome database TAGdb construction

In order to map InSPecT TAGs and their prefix/ suffix onto the whole genome, a pre-processing algorithm was run to create an appropriate index, stored on a server in a MySQL database. The algorithm evaluates the TAGs based on tryptic peptides. Although very short exons could lead to peptides that align across two or more introns this study focuses on peptides that spans across only one intron (i.e. one expected gap of unknown length).

The indexed genomic sequence TAGdb includes all possible combinations of three amino acid TAGs (20^3) from the whole genome of *T. gondii*; the algorithm first converts the nucleotide sequences to a six-frame translation and then extracts the ORF_SS set. There is no length threshold for these extracted ORF_SS although those containing no tryptic sites are discarded. The algorithm, Figure 3:3, traverses each retained ORF_SS sequence and selects the TAGs incrementally from position one to the end of the sequence. For each TAG the algorithm calculated the prefix and suffix mass up to the first tryptic site (i.e. trypsin cleaves after lysine K or arginine R not followed by proline P).

If the ORF_SS sequence contains only one potential tryptic site, then the TAG would have either the prefix or the suffix. This would be the case if the tryptic site was either preceding the TAG or aligned with third amino acid of the TAG itself. Any value that cannot be calculated is flagged on the TAGdb as -1. If the tryptic site was located at position one or two of the TAG, then both prefix and suffix are flagged as -1 on TAGdb. This was chosen in order to avoid discarding data in view of future algorithm developments. This type of flag was assigned also in other cases where the prefix or suffix value could not be calculated: a TAG located between a stop codon and a tryptic site would have -1 as prefix value and the calculated suffix. On the TAGdb each TAG is stored with both its location on the genome (chromosome, strand, coordinates) and its calculated prefix and suffix values. The resulting genome index TAGdb was then loaded into a MySQL database [255-258] for future querying. The table created has 7 columns (primary key, chromosome, strand, TAG, start coordinate, prefix and suffix).

WMALSHDHRGVHPQLIASSFLEASKQSEKDGHEAMSN

Tryptic site position	9	25	29

Genomic coordinate	Chromosome	Strand	TAG position within ORF_SS	prefix	TAG	suffix
3704835	VIII	1	1	-1	WMA	745.36
3704836	VIII	1	2	-1	MAL	632.28
3704837	VIII	1	3	-1	ALS	545.25
3704838	VIII	1	4	-1	LSH	408.19
3704839	VIII	1	5	-1	SHD	293.16
3704840	VIII	1	6	-1	HDH	156.10
3704841	VIII	1	7	-1	DHR	0
3704842	VIII	1	8	-1	HRG	1607.87
3704843	VIII	1	9	-1	RGV	1508.80
3704844	VIII	1	10	0	GVH	1371.74
3704845	VIII	1	11	57.02	VHP	1274.69
3704846	VIII	1	12	156.09	HPQ	1146.63
3704847	VIII	1	13	293.15	PQL	1033.54
3704848	VIII	1	14	390.20	QLI	920.46
3704849	VIII	1	15	518.26	LIA	849.42
3704850	VIII	1	16	631.34	IAS	762.39
3704851	VIII	1	17	744.43	ASS	675.36
3704852	VIII	1	18	815.47	SSF	528.29
3704853	VIII	1	19	902.50	SFL	415.21
3704854	VIII	1	20	**989.53**	**FLE**	**286.16**
3704855	VIII	1	21	1136.60	LEA	215.13
3704856	VIII	1	22	1249.68	EAS	128.09
3704857	VIII	1	23	1378.72	ASK	0
3704858	VIII	1	24	1449.76	SKQ	344.17
3704859	VIII	1	25	1536.79	KQS	257.14
3704860	VIII	1	26	0	QSE	128.09
3704861	VIII	1	27	128.06	SEK	0
3704862	VIII	1	28	215.09	EKD	-1
3704863	VIII	1	29	344.13	KDG	-1
3704864	VIII	1	30	0	DGH	-1
3704865	VIII	1	31	115.03	GHE	-1
3704866	VIII	1	32	172.05	HEA	-1
3704867	VIII	1	33	309.11	EAM	-1
3704868	VIII	1	34	438.15	AMS	-1
3704869	VIII	1	35	509.19	MSN	-1

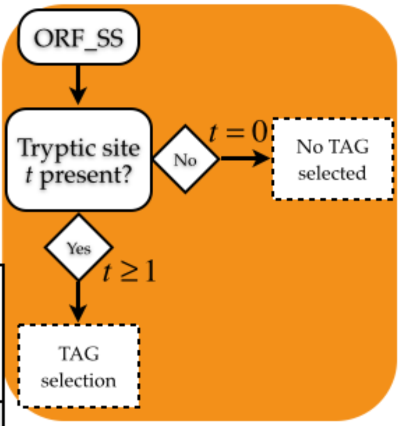


Figure 3:3 The workflow illustrates the algorithm (orange) that generates the TAGdb index, showing how data is retained/ flagged or filtered out. To begin with the algorithm selects ORF_SS from six-frame translation and discards all ORF_SS with no tryptic site. Then the algorithm traverses the ORF_SS extracting all sequence TAGs and calculates their respective prefix/ suffix mass from the nearest tryptic site; if this value is missing, as no tryptic event at the terminus (e.g. TAGs located between the stop codon and the closest tryptic site: positions one to nine and 28 to 35), the value stored on TAGdb is flagged with -1. At this development

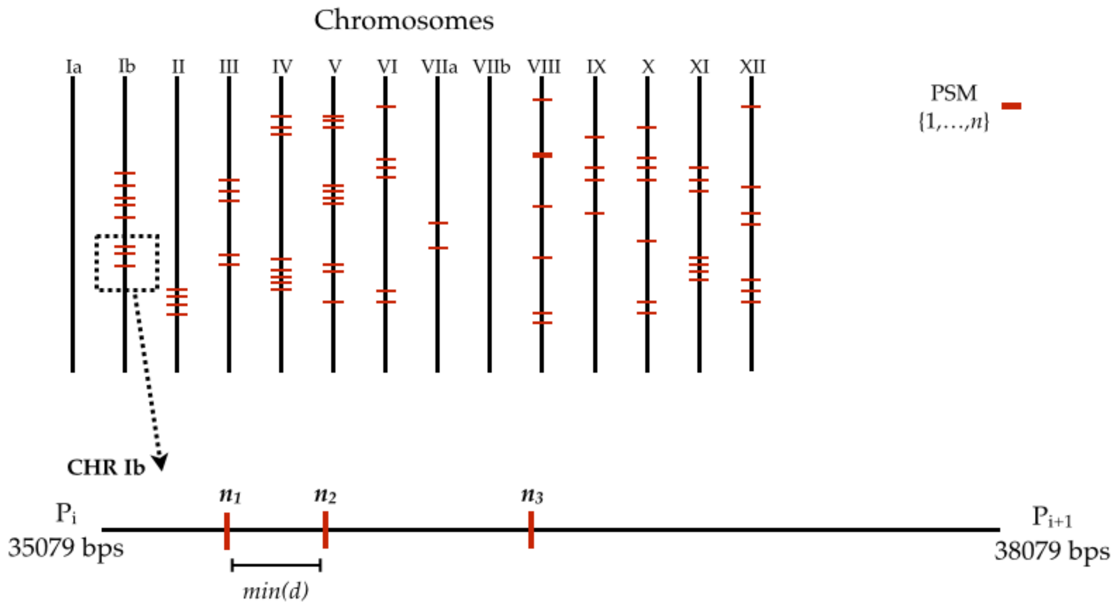
stage the algorithms computes the missed cleavages (t) only if they reside at position 1 or 2 within the sequence TAG and each respective terminus can be located (e.g. here the second tryptic site, position 25).

3.3.2 Pipeline stage 1: from DB search to genomic regions

The aim of this designed pipeline is to identify ISP, without a previous genome annotation existing. This blind approach is designed for those organisms whose genome has been sequenced but their gene annotation is still missing or lacking accuracy. With this in mind the sequence database search (stage 1.1) is meant to be addressed with common search engine algorithms such as MASCOT, X!Tandem, OMSSA and InSPecT hybrid mode. The processes described in this chapter were initially designed to use only one search engine throughout and as such the database searches were performed only with InSPecT hybrid mode. However, stage 1.1 could equally well be performed with other search engine or a combined search engine approach. The ORF_SS database was constructed by extracting open reading frames comprised between two stop codons with minimum length of 40 amino acids as discussed in chapter two. In order to allow calculating the FDR score the ORF_SS database included a decoy database, made of reversed sequences.

The stage 1.2 comprises the FDR rescoring of the PSMs identified with InSPecT searches against the ORF_SS database in the previous stage; the set of PSMs within the 1% FDR threshold is then used for the following stage, after the removal of the decoy hits.

The clustering algorithm used in stage 1.3 organizes the selected PSMs by genomic regions into clusters. To do so, the designed algorithm aligns each PSM onto their correspondent chromosome and then orders the PSMs based on their coordinates. The distance threshold min(d) is based on nucleotide length and it is given at run-time as end-user parameter input. As illustrated on Figure 3:4, for PSMs (n_1, n_2, n_3, n_i) located on the same chromosome (e.g. chromosome Ib) they are first sorted ascending by their coordinate. The algorithm then traverses the chromosome evaluating the distance of each consecutive pair of PSMs and comparing it to the $\min(d)$; if subtraction between the start of n_2 and the end of n_1 is less or equal to the $\min(d)$ then these PSMs would be grouped in the same cluster 1. Then it evaluates if the subtraction between the start of n_3 and the end of n_2 is less or equal to the $\min(d)$, if within the threshold the PSM n_3 is added to cluster 1 otherwise it will be added to new cluster 2. Several tests were performed to assess the optimal value $\min(d)$, such that resulting clusters would contain more than one PSM. After testing different lengths (i.e. 2000, 3000, 4000 and 5000 bps), the 5000 bps value was used for this study.



P	Position on chromosome {i, i+1}
n	PSM {1,2,3}
min(*d*)	minimum distance threshold (user input)

Clustering algorithm:

If (start (n_2) - end(n_1)) ≤ min (d)) → cluster 1 {n_1,n_2}

then

If (start (n_3) - end(n_2)) ≤ min (d)) → cluster 1 {n_1,n_2,n_3}

else (start (n_3) - end(n_2)) ≥ min (d)) → cluster 1 {n_1,n_2}
cluster 2 {n_3}

Figure 3:4 Clustering algorithms groups together the PSMs located within chosen nucleotide range min(*d*) (user manual input). It first orders the PSMs by their chromosome and then sorts them by their position. It then evaluates the position of each pair of consecutive PSMs and if their distance is less or equal to min(*d*) then they are included in the same genomic region (cluster 1), otherwise they are assigned to separate genomic regions (cluster 1 and cluster 2).

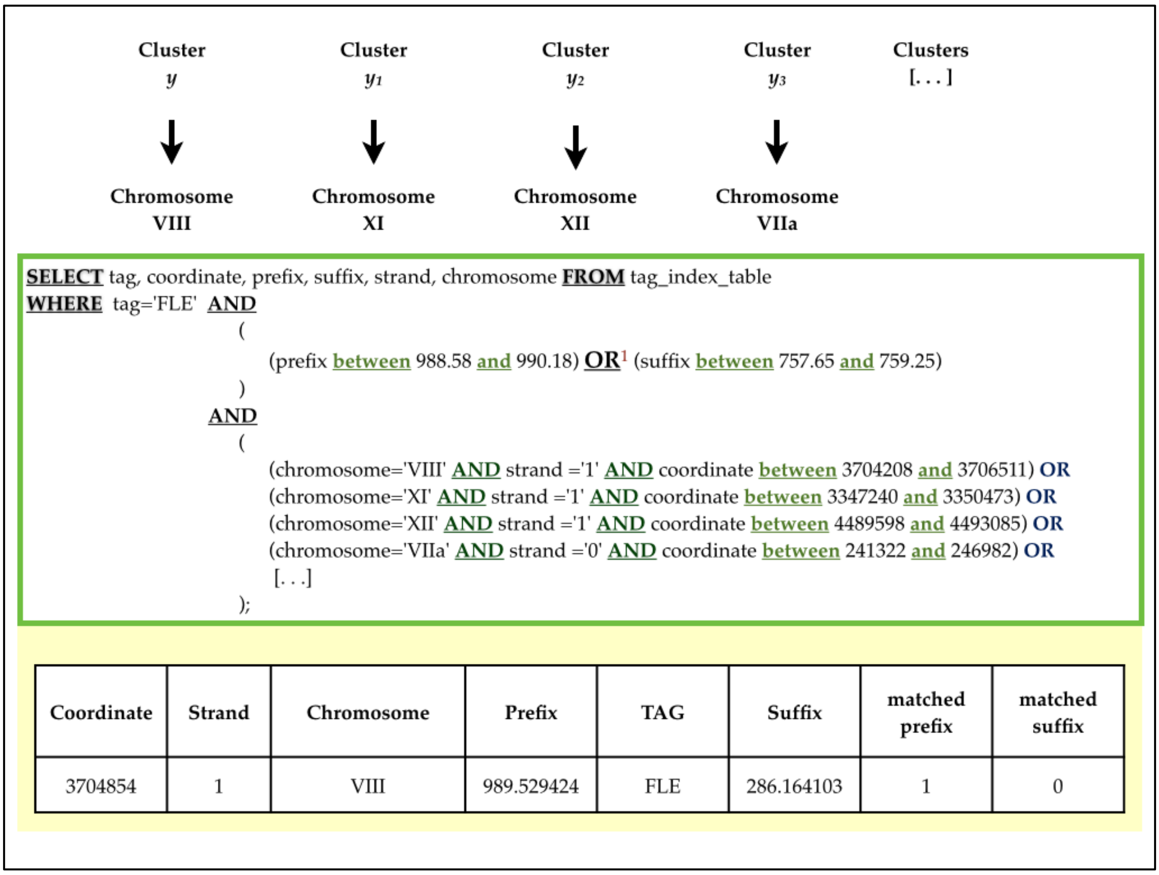
3.3.3 Pipeline stage 2: from InSPecT TAGs to spectrum anchors

During the second stage of the pipeline (stage 2.1) InSPecT is run on the mass spectra that have not been explained during the first stage. In this study InSPecT was set to run with fixed modification: carbamidomethyl (57 Da); variable modification: methionine oxidation (16 Da).

InSPecT performance was also tested on several different sets of predicted TAGs, based on different modifications such as phosphorylation (80 Da), pyroglutamate (-17 Da), dehydration (-18 Da) and acetylation (42 Da). However poorer overall performance was achieved and so these data are not presented.

Additional tests on PRM score thresholding indicated that only TAG with PRM score equal or above 50 were successfully mapped onto the genomic region (data not shown) and therefore PRM score of 50 was chosen as cut-off value for selecting InSPecT TAGs to process in the pipeline.

The resulting set of TAGs (algorithm previously explained) is then queried against the TAGdb (stage 2.2) using a common SQL "select" query statement with the appropriate conditions (Figure 3:5). For each TAG, the query attempts to retrieve its TAG sequence if it matches either the prefix or the suffix with a mass range of ±0.8 Da (as it is LTQ ion trap data); this tolerance range can be selected by the end-user at run-time. The second condition of the query statement is that the TAG should be contained within the genomic regions previously selected (stage 1.3). At this stage only the start and the end of each genomic region is retained, with an additional user input coordinate range. Although the query itself is rather simple, the number of genomic regions can increase the search space and therefore the computational run-time of the algorithm. Whenever a TAG has been matched on the TAGdb, it will be extracted with its matched prefix or suffix; additionally a flag of binary value (0,1) is linked to the prefix and suffix to indicate which value had been matched (positive value indicates a positive match). After the TAG has been mapped on the TAGdb, with either its prefix (N-terminus) or suffix (C-terminus), it is possible to retrieve the amino acid sequence that stretches from the TAG to the matched terminus. With this, the algorithm uses each matched TAG to partially align the correspondent tandem mass spectrum to the genome, hence generating a ***spectrum anchor*** (stage 2.3). This stage can generate a set of number of *spectrum anchors*, dependent upon the InSPecT TAG prediction accuracy and the selection of genomic regions. For this reason, this set can contain zero or more candidate spectrum anchors for each spectrum.



```
SELECT tag, coordinate, prefix, suffix, strand, chromosome FROM tag_index_table
WHERE  tag='FLE'  AND
                   (
                      (prefix between 988.58 and 990.18) OR1 (suffix between 757.65 and 759.25)
                   )
                 AND
                   (
                      (chromosome='VIII' AND strand ='1' AND coordinate between 3704208 and 3706511) OR
                      (chromosome='XI' AND strand ='1' AND coordinate between 3347240 and 3350473) OR
                      (chromosome='XII' AND strand ='1' AND coordinate between 4489598 and 4493085) OR
                      (chromosome='VIIa' AND strand ='0' AND coordinate between 241322 and 246982) OR
                       [. . .]
                   );
```

Coordinate	Strand	Chromosome	Prefix	TAG	Suffix	matched prefix	matched suffix
3704854	1	VIII	989.529424	FLE	286.164103	1	0

Figure 3:5 A query statement (in green) written in SQL language is executed to retrieve data from TAGdb. Highlighted in bold and underlined are the specific query commands and Boolean conditions. The conditions here are that both the TAG pattern and prefix (or suffix mass) and at least one of the genomic regions are matched. The condition for prefix and suffix includes a mass tolerance ± 0.8 Da based on InSPecT generated TAGs (user input). The condition for location includes all genomic regions (as generated from stage 2.3), although here only four are represented. The retrieved data is shown in yellow and each row has tag, coordinate, prefix, suffix, strand and chromosome as selected from TAGdb. An additional binary flag is provided for both prefix and suffix mass matching to retain information on which of these values has been positively matched.

3.3.4 Pipeline stage 3: from spectrum anchors to full length peptides

The third stage of the pipeline comprises the full length reconstruction of peptides from the previously generated spectrum anchors. Unless stage 2.3 generated a null set of spectrum anchors, the algorithm further processes the data as discussed in the following paragraphs.

The genomic regions have been obtained by clustering the PSMs matched against ORF_SS sequence database; this allows the extraction of the start and the end coordinate of each ORF_SS where the PSMs have originally been identified. In this study each identified ORF_SS is considered as a potential exon from a potential gene; this is why the genomic regions retain the coordinate information about matched ORF_SS. The theory is that an intron spanning peptide would have the prefix aligning with exon n while the suffix would align with exon $n + 1$.

The stage 3.1 uses the spectrum anchor to search for the terminus whose mass has yet to be explained. As an example in Figure 3:6, we can consider a genomic region that was obtained by three PSMs from two different ORF_SS entries; in this scenario the first PSM is contained in the first ORF_SS, while the second and third PSM are contained in the second ORF_SS. From this genomic region the algorithm would still retain both the start-end coordinates of the two ORF_SS and the start-end coordinates of the PSMs contained inside them. A TAG was mapped onto the first ORF_SS with a matching prefix located downstream of the C-terminus of the first PSM in the genomic region. Then the suffix value will be searched on a range starting from the TAG coordinate up to the N-terminal coordinate of the second PSM of the same genomic region. This strategy would further reduce the number of false positive from the list of retrieved tryptic sites, while it would still remain coherent with the basis of the ISP search, as no real PSM can be identified in a intron. In the case instead where a TAG was mapped on the first ORF_SS with a matching suffix located before the N-terminus of the first PSM, then the range would be calculated differently. This is due to the assumption that the unexplained N-terminus of the ISP could be located outside the genomic region; to account for this possibility the algorithm takes an additional value for bps length as the manual input (in this study it is 1000 bps). Then the algorithm would effectively search for the N-terminus of the peptide within a coordinate range starting from the start of the ORF_SS minus 1000 bps up to the TAG coordinate. This approach provides the limits for stage 3.1 as the algorithm traverse the genomic sequence, between the TAG and the coordinate limit, extracting all contained tryptic sites. One additional limit based on coordinate range (in bps) is also added to stage 3.1; this takes in account those aligned TAGs that fall between genomic region start/ end and the first/ last element (PSM) of given genomic region.

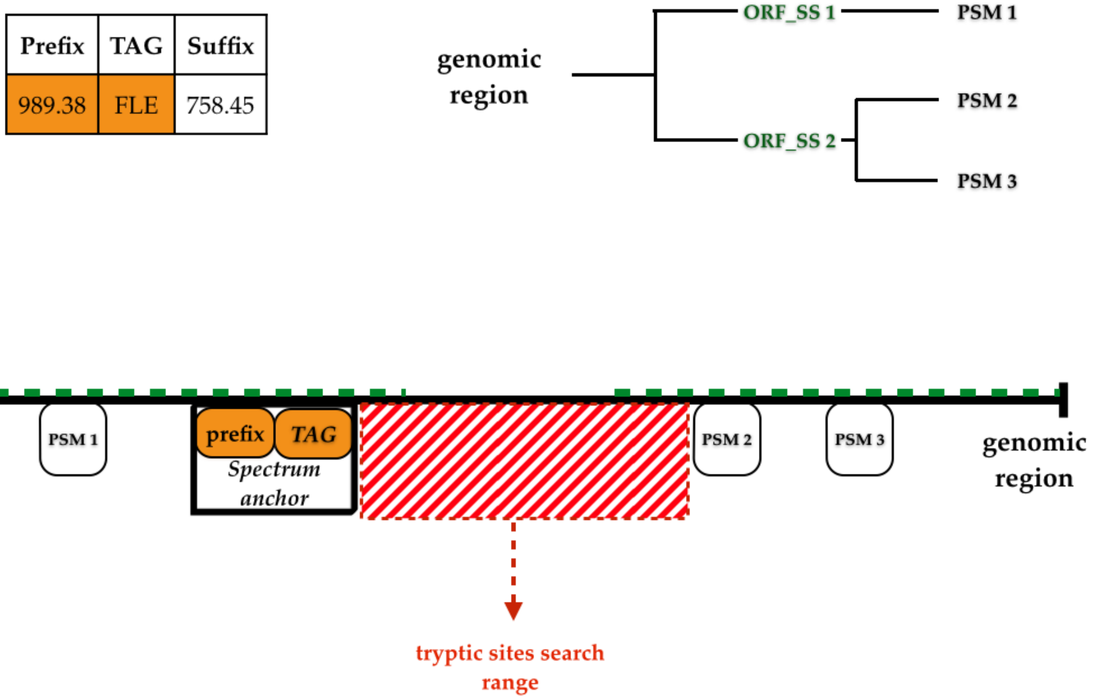
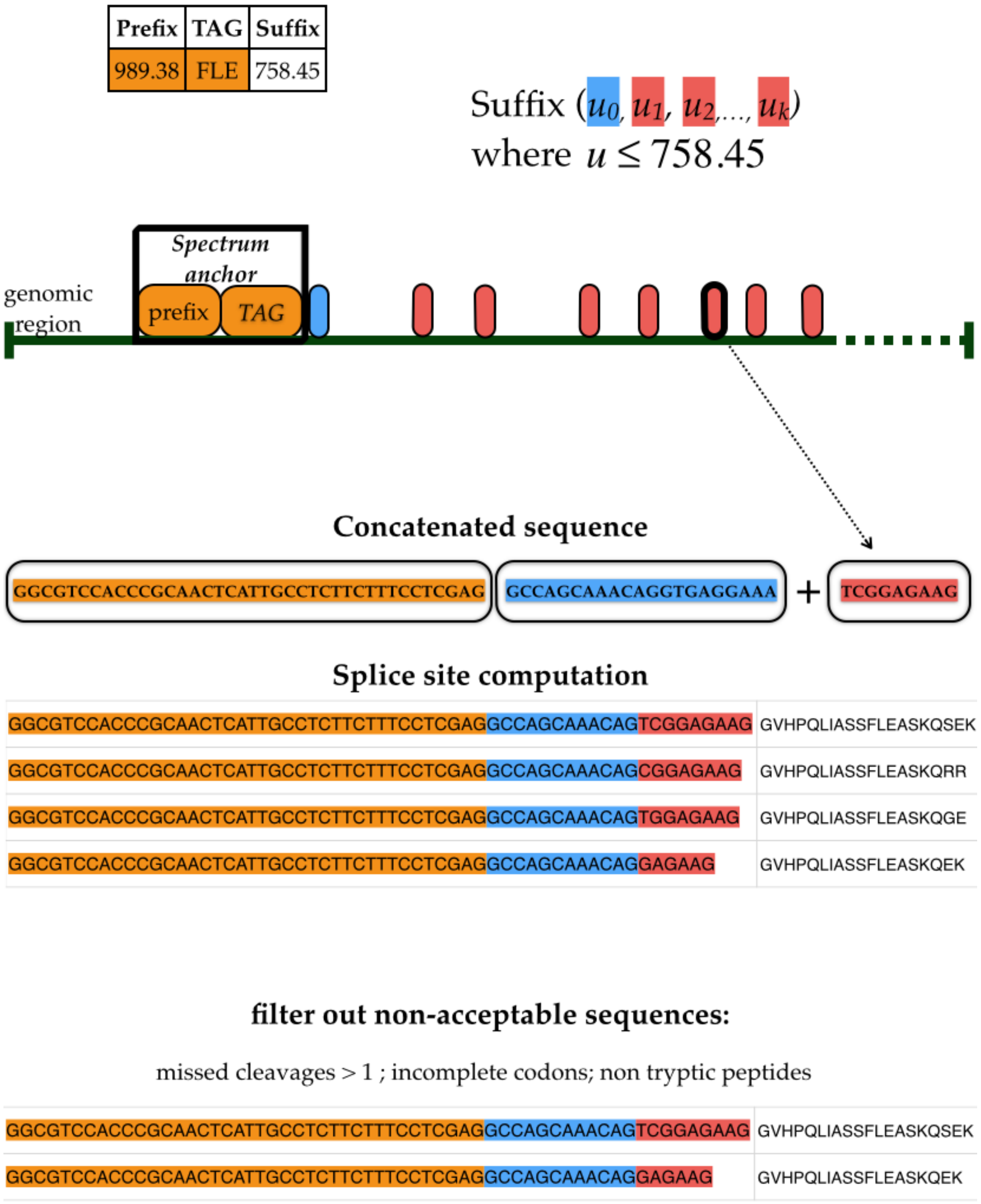


Figure 3:6 To illustrate how the algorithm selects the coordinate range for retrieving tryptic sites, the example above shows a genomic region comprised by three clustered PSMs located on two ORF_SS (in green). Where a TAG has been aligned by its prefix value (in orange), the tryptic site will be searched downstream the spectrum anchor from the TAG position to the 5′ of the nearest PSM (area highlighted in red).

Stage 3.2 comprises the extraction of the partial peptide sequence corresponding to the terminus mass to be explained (Figure 3:7). From each tryptic site the algorithm concatenates one by one the amino acids in order that their total mass is equal or less than the mass value to explain (e.g. suffix). This process allows generating potential partial peptide sequences from all tryptic sites but the algorithm also attempts to identify the unknown location of the splice site. In order to account for this, the algorithm extracts the partial sequence, for the unexplained terminus, that is adjacent to the spectrum anchor. The spectrum anchor is then concatenated with the adjacent partial sequence; the algorithm then generates an initial list of full length peptide sequence by concatenating each tryptic site partial sequence, one by one, with the spectrum anchor and adjacent sequence.

This initial list would in theory contain, among all sequences, the correct N-terminus and the potential C-terminus. As the splice site can occur at the nucleotide level inside a codon (i.e. the first base pair located in exon 1, while the following two base pairs located in exon 2) the algorithm uses the coordinate values of the spectrum anchor, adjacent sequence and tryptic site partial sequence to retrieve the nucleotides which could potentially encode the full length peptide sequence. Then all splice events are calculated from the TAG position to the end of tryptic site partial sequence. Before storing this list of all candidate full length peptides with all possible splice sites, the algorithm discards all peptide sequences if: the nucleotide length is not multiple of three (i.e. it must contain complete codons); if they contain more than one missed cleavage; they contain a stop codon or if they are not tryptic peptides.

The results from stage 3.2 is a temporary database that contains a list of peptide sequences that attempt to explains the unknown terminal mass.



Prefix	TAG	Suffix
989.38	FLE	758.45

Splice site computation

GGCGTCCACCCGCAACTCATTGCCTCTTCTTTCCTCGAGGCCAGCAAACAGTCGGAGAAG	GVHPQLIASSFLEASKQSEK
GGCGTCCACCCGCAACTCATTGCCTCTTCTTTCCTCGAGGCCAGCAAACAGCGGAGAAG	GVHPQLIASSFLEASKQRR
GGCGTCCACCCGCAACTCATTGCCTCTTCTTTCCTCGAGGCCAGCAAACAGTGGAGAAG	GVHPQLIASSFLEASKQGE
GGCGTCCACCCGCAACTCATTGCCTCTTCTTTCCTCGAGGCCAGCAAACAGGAGAAG	GVHPQLIASSFLEASKQEK

filter out non-acceptable sequences:

missed cleavages > 1 ; incomplete codons; non tryptic peptides

GGCGTCCACCCGCAACTCATTGCCTCTTCTTTCCTCGAGGCCAGCAAACAGTCGGAGAAG	GVHPQLIASSFLEASKQSEK
GGCGTCCACCCGCAACTCATTGCCTCTTCTTTCCTCGAGGCCAGCAAACAGGAGAAG	GVHPQLIASSFLEASKQEK

Figure 3:7 As InSPecT TAGs have been mapped within selected genomic regions, their matching value (prefix or suffix) is used to retrieve the partial sequence and generate a spectrum anchor. Here the TAG (FLE) has been aligned within genomic region (green line) and the matched prefix is used to retrieve the spectrum anchor (in orange). Then the unexplained suffix mass is used to retrieve a set of partial sequences from nearby tryptic sites (u_1, u_2, u_k in red). Given the expected presence of the splice site but its unknown position, the partial sequence adjacent to the spectrum anchor is also retrieved (u_o in blue). Both the partial

sequences in red and in blue have to satisfy the suffix mass condition. The mass of the partial sequence in blue, just like the mass of the partial sequence in red, has to be equal or less than the TAG suffix. After these nucleotide sequences have been retrieved and concatenated, the algorithm translates them into amino acid sequence. During this last step a filter discards all sequence that have either: i) incomplete codons; ii) more than one missed cleavage; iii) stop codons or iv) non tryptic peptides.

3.3.5 Pipeline stage 4: search engine against temporary database

The fourth stage of the pipeline comprises a traditional sequence database search with OMSSA. The temporary fasta database is formatted for OMSSA and the current MS/MS spectrum being evaluated is isolated from the original mass spectra file. The sequence database search with OMSSA algorithm generates a shortlist of full length peptides with their respective e-values.

3.3.6 Sample acquisition and test dataset

The data analysed were acquired from tachyzoite samples of *T. gondii* RH strain, harvested 3- 4 days post infection as described in [244]. The protein samples were processed by 1D SDS-PAGE, followed by in-gel digestion (trypsin), and analysed on an LTQ-ion trap mass spectrometer as described in [101].

A test dataset was constructed of *known and correct intron-spanning peptides* as follows. The data was sourced from 10 1D SDS gel slices from *T. gondii* RH strain acquired from [101]. The gene models used for testing were from ME49 strain as to our current knowledge it can be considered a well-annotated organism and all 66 annotated gene for RH are most closely related to ME49. The sample spectra were queried against official annotation 6.3 with the three Search Engine mode (MASCOT, X!Tandem and OMSSA); parent and fragment ion tolerances: 0.8 Da; fixed modification: carbamidomethyl (C); variable modification: methionine oxidation. The combined results were re-scored using FDR estimates with a fixed threshold of 5% FDR. By mapping all peptide sequence back to their constituent genes, it was possible to filter out a set of all the spectra generated from intron spanning peptides. From this dataset 126 PSMs were selected with a total of 113 non-redundant peptides, mapped to 77 proteins, which were matched by all three search engines with low FDR (< 0.001). Methionine oxidation, present in 35 PSMs, is allowed as a PTM as it is included in the search algorithm of InSPecT.

3.4 Results

3.4.1 ISP identification by sequence database searches

The algorithm was evaluated by querying the test dataset (126 spectra known to match true ISPs). The approach was able to identify 26% of the ISPs through the blind search approach (Figure 3:8, details shown in Table 3:1 and appendix Table 7:12). The remaining spectra were either incorrectly assigned during the last stage or were not assigned at all, resulting in null set. A total of 64 spectra were searched with OMSSA, providing correct ISPs for half of these (TP=32). The profile of e-value for TP and FP were substantially different (Figure 3:9); OMSSA algorithm was not only capable of assessing the correct peptides as true positive with confidence, it was also capable of identifying all incorrect identifications with low confidence, thus allowing a threshold of e-value $< Ie^{-5}$ to be selected for further processing. The e-value chosen for the FP corresponded to the lowest e-value available for a given set of results; in other words, where the identifications for ISPs were all incorrect, the best e-value was selected for this analysis. Although the currently designed algorithm achieved only 26% of correct ISP identifications, the results from OMSSA can be considered as partially successful given that the profile of TP and FP are clearly substantially different. Further implementation of the approach would focus on reducing the data loss and improving the other stages of the pipeline.

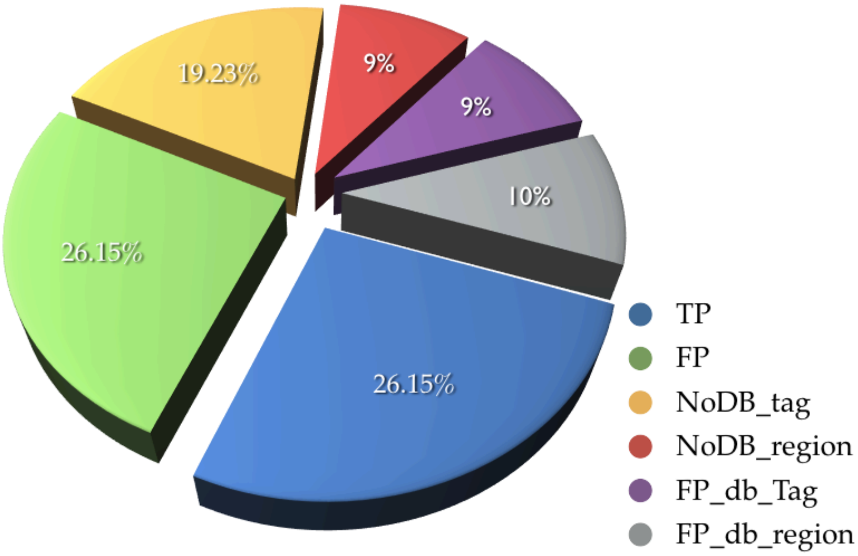


Figure 3:8 From the final test dataset it was possible to identify only 26% ISPs correctly. The remaining spectra were either identified, and considered FP due to OMSSA e-value, or filtered out at different stages in the pipeline. One reason for data loss was incorrect InSPecT TAG prediction, which generated FP ISP candidates (FP_db_Tag) or did not allow the spectrum to be located on the genome (NoDB_tag). Where the spectrum could not be correctly mapped to the genome it was caused by wrong prefix/ suffix mass assessed for the predicted TAG, or suboptimal parameters for genomic region filters. The second reason for data loss was the inaccuracy of genomic region selection, which in turn led to either false positive candidates (FP_db_region) or filtered out correctly predicted TAG (No_db_region).

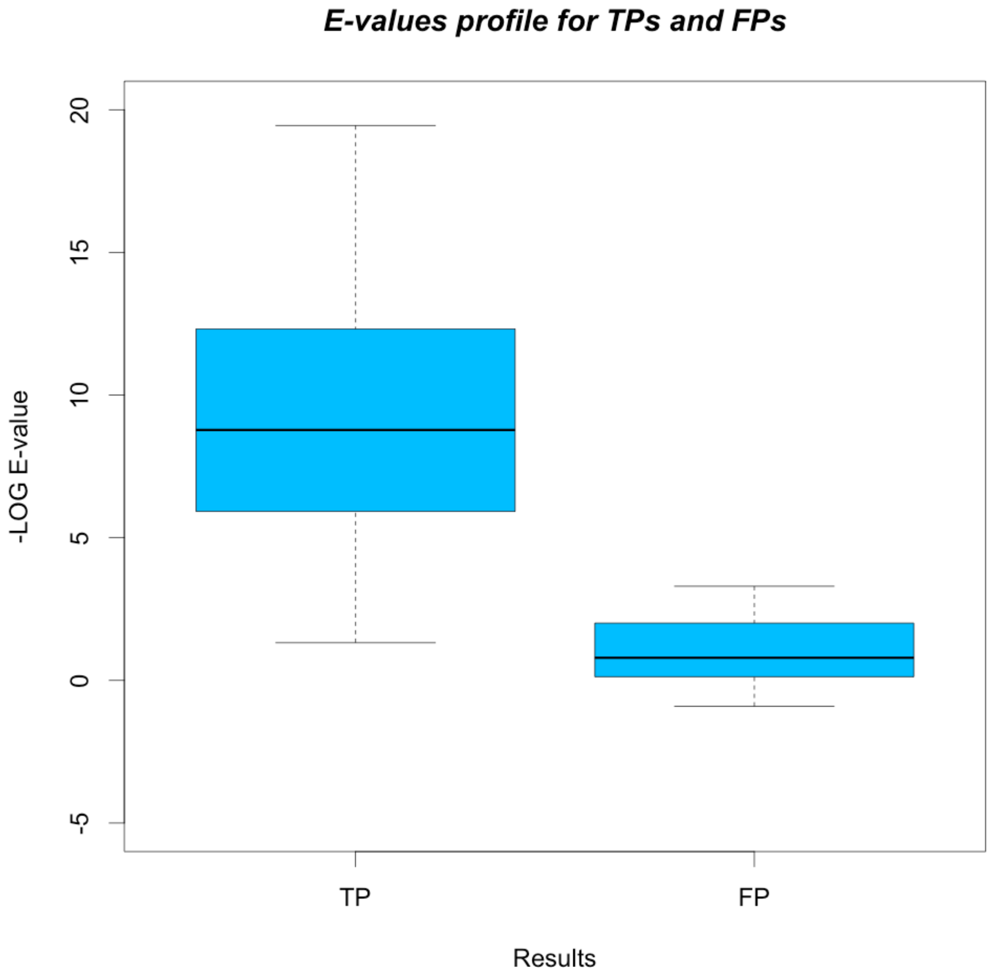


Figure 3:9 Comparison of e-value calculated by OMSSA shows clear difference between accurate and inaccurate ISP identification. The plot shows the –LOG of the e-value for both correct (32) and incorrect (32) matches.

The data loss for a large portion of ISPs was attributed to four different causes. The first reason involves InSPecT sequence TAGs being incorrectly predicted (amino acid composition) or being assigned wrongly to *b-*/*y-* ion (~19%). The second reason involved incorrectly assigned genomic regions (~9%). The third and fourth reason for data loss involved OMSSA not yielding any results (~19%) from the searches against temporary sequence databases. This was the consequence of the correct ISP sequence not being in the temporary database, due to incorrect sequence TAG or genomic region (respectively ~10% and ~9%).

GENE	Peptide	PTM	TAG	R	*e*-value
TGME49_088 360	YLQDVFDVPLVIQLTDDEK	N/A	3	OK	3.55E-20
TGME49_051 780	ISQQAYNQAGSTDSSAGSEGTGSESGDK K	N/A	3	OK	1.93E-18
TGME49_036 540	MIELFPSSKQEMEFAAQGGDPR	M:+16	2	OK	4.40E-18
TGME49_063 180	LQQVEPTADSQELTVQAK	N/A	8	OK	7.75E-16
TGME49_072 910	GVHPQLIASSFLEASKQSEK	N/A	6	OK	4.53E-14
TGME49_061 950	AAPLFADQSTEPGLLQTGIK	N/A	3	OK	6.81E-14
TGME49_090 200	AGGIIGTAFGQGGFDWAMLK	M:+16	4	OK	9.06E-14
TGME49_026 960	TLGEIVTFVADAVK	N/A	7	OK	1.17E-13
TGME49_032 180	SLQTIIEDQTELAVYPHVGEALQR	N/A	3	OK	1.96E-12
TGME49_029 010	SLYGGIANTLETPFADSEAVAK	N/A	4	OK	2.81E-12

Table 3:1 Extract from the list of the 126 ISPs processed through the pipeline (first set) ordered by OMSSA e-value. The first three and the last columns identify gene ID, peptide, PTMs and e-value (OMSSA). The fourth column shows the number of TAGs that align on the peptide (prefix/ suffix mass not considered here). The column "R" specify the reason why the match failed: 'WR-TAG/ REG' indicates false positive ISPs in the temporary DB due to incorrect TAG or wrong genomic region selected; 'WR' indicates that OMSSA failed to match the correct peptide; 'NR-TAG/ REG' indicates that OMSSA yield null result due to incorrect TAG or wrong genomic region selected; 'NDB-TAG/ REG' indicates that the temporary dataset of ISPs candidates was not generated due to incorrect TAG or wrong genomic region selected. The last peptide sequence is flagged with WR as OMSSA failed to correctly identify the ISP although it was contained in the temporary DB. The second last peptide instead is flagged with WR-TAG InSPecT did not provide correct TAGs identification for such spectrum.

3.4.2 Targeting ISPs: comparison of bioinformatic approaches

The same test dataset of 126 spectra was used to compare different bioinformatic approaches dealing with targeted intron spanning peptide identifications. Database searches were performed with three search engine mode with identical parameters as in previous searches: 0.8 Da for both parent and ion tolerance; the results were rescored by FDR (fixed 5% FDR threshold) as previously discussed in chapter two. The sequence databases used were the alternative gene models, generated with GeneMark and GlimmerHMM gene finders described in the previous chapter. To allow FDR rescoring, the 126 spectra were combined with ~1200 random spectra.

The *de novo* sequencing approach (Appendix A) was evaluated by pairwise alignment (Smith-Waterman [47]) between each of the 3 highest ranked peptides predicted by PepNovo against the known ISP. The matrix used for scoring was BLOSUM62 (Block Substitution Matrix) with gap penalty and extension of respectively 1 and 0.5. All *de novo* predictions did not cover the entirety of the spectrum.

The results from different approaches were compared to evaluate how many correct ISPs could be identified. For database dependent approaches it was possible to set a cut-off score at 5% fixed FDR. To enable comparisons with *de novo* predictions, only the peptides with similarity score above 90% were retained; sine in contrast to the other approaches PepNovo does not provide full-length peptide sequences (Table 3:2).

Overall the blind search pipeline can only be considered a limited success, as on this dataset, generating alternative gene models (with GeneMark) appears a more accurate route to ISP identification.

Method	GeneMark	Glimmer	ISP pipeline	PepNovo 1	PepNovo 2	PepNovo
ISPs	94	33	30	36	39	36
Score cut-off	5%FDR	5%FDR	5%FDR	>90% similarity	>90% similarity	>90% similarity

Table 3:2 The tables shows the number of correctly identified ISPs respectively by alternative gene models (GeneMark, Glimmer), blind database search pipeline, and *de novo* predictions by PepNovo. The three best peptides interpreted by PepNovo were aligned against the ISPs; only those with alignment score above 90% were included here.

3.5 Conclusion

The strategy presented here attempts to minimise the weaknesses of database dependent and *de novo* sequencing by combining the complementary information provided from their respective results, and provide an additional analysis stage to perform after common sequence database searches. Similarly to InSPecT hybrid approach, the devised algorithms attempt to exploit the predicted TAGs generated only from the spectra that have not been unexplained through database searches.

A novel database dependent approach for *de novo* identification of ISPs was presented. The development of the pipeline algorithms gave us the opportunity to perform tests on *de novo* sequencing; from these evaluations it was possible to confirm how it is still unable to provide high confidence identifications for those spectra that have not been explained by database. Even generated sequence TAGs can yield large number of false positives that would prevent the identification of those spectra (~50%) that have not yet been identified with database dependent methods.

There are several factors that make the pipeline approach challenging:

- the spectrum anchor length directly affects the confidence of the identification;

- the spectrum anchor can be mapped only if the TAG and either its prefix/ suffix mass has been correctly predicted and it is present within the selected genomic region;

- where present, the PTMs must to be correctly identified on the TAG prefix/ suffix masses;

- the manual selection of ISP candidate from the OMSSA list of reconstructed/ identified full length peptides needs to be automated and its reliability should be further improved;

The pipeline design is still not optimal as there is a considerable data loss throughout the different stage processes. Also, the algorithm can be further implemented to decrease the running time for large-scale analyses. One of the future pipeline improvements would comprise the use of multiple search engine mode (here used to collect the test dataset) during the stage 1.1 for identifying spectra and clustering the PSMs in genomic regions. Additional improvements to the pipeline algorithms could reduce the error rate and data loss occurred when anchoring the spectra to the genome (tag prediction, genomic region filtering). With the availability of high-resolution data, it should be possible to better exploit tag prediction software and optimise the filter for genomic regions of interest.

However our results have proved an initial framework for identification of ISP without a gene model set and an initial method for discriminating between real peptides and false identifications. The e-value assessed by OMSSA search engine can prove useful to efficiently assess the confidence of identified candidate intron spanning peptides. Future improvements in *de novo* sequencing algorithms, even for only the partial peptide reconstructions, could enable a further increase in the confidence for peptides identified during the final stage (OMSSA).

A comparison of different approaches provided similar results between the ISP pipeline and GlimmerHMM gene models; GeneMark models appeared to be of higher quality as almost three times more ISPs were identified. Differently designed algorithms for gene finding can cause this. For instance GlimmerHMM analyses the genomic sequence and makes evaluations based on previously trained datasets (obtained from gene models of previous release). Instead GeneMark analyses the genomic sequence and using various algorithms, including Viterbi algorithm, it estimates the number of gene predictions (hidden paths for Hidden Markov Model) without the need for prior training datasets. Although GeneMark appears to have outperformed GlimmerHMM it would be advisable to use panels of alternative gene models to increase confidence of the results.

PepNovo algorithm is largely based on InSPecT algorithm and the peptide prediction accuracy is directly related to the predicted peptide length: the shorter the peptide, the higher the accuracy. However it can present datasets of widely different predictions; because of this, additional evaluations would be required to confidently identify peptides, even after sequence alignments.

4 Optimised bioinformatic processing of N-terminal proteomics data: characterisation of the N-terminome of *Toxoplasma gondii*

4.1 Abstract

To this day the identification of start codons for genome annotation purposes remains a bioinformatic challenge. The difficulties result from an inability to identify with high confidence the correct start codon (ATG) out of a number of possible alternatives [259]. A proteomic analysis of N-terminal peptides can confirm a protein's presence in the samples, as well as provide additional information on signal peptides and other PTMs such as N-terminal methionine excision (NME).

Novel approaches for targeted N-terminal identification have been developed, using protocols to selectively enrich for N-terminal peptides or deplete internal peptides. Now bioinformatic methods can be put in place to analyse this small percentage of the peptidome: which, for example, accounts for less than ~3% of the peptidome of *T.gondii*.

By combining recent improvements in data sampling/ acquisition with new bioinformatics approaches, this chapter presents a study on sequence database design and search query optimisations to maximise N-terminal identifications. Further analyses of the results also provide a deeper understanding of signal peptides and N-terminal methionine excision in the Apicomplexa.

4.2 Introduction

Even with additional transcriptional data and ever more accurate gene predictions, it still remains challenging to identify the N-terminus of protein sequences with high confidence. Their prediction is generally a result of computational prediction by gene finders and, although their accuracy has notably improved in the recent years, the annotations still need to be verified experimentally [28-30, 39, 50, 51, 189, 194, 202, 242, 260]. Determining the correct start of the isoform remains a bioinformatic challenge due to 5′ UTR region and co- or post-translational modifications some proteins undergo. N-terminal peptide identification can provide valuable information about the protein analysed such as confirming the translational start and potentially give insight on its cellular location through the identification of a signal peptide. These short sequences on the amino terminus of secreted proteins play a major role in leading preproteins to their correct cellular compartment. The signal peptide, after having performed its role, is then cleaved off by signal peptidases and undergoes systemic degradation [57].

Different groups have taken on the challenge of N-terminome identification and have provided insights on the future outlook and drawbacks of respective techniques. It has also been argued that N-terminal peptides could be sufficient for large-scale protein identification, since it would simplify the analysis process by minimising the number of peptides needed for protein identification [261, 262].

Of the few methods that have been implemented for identification of proteolytic cleavage sites and removal of internal peptide from sample pool are: COmbined FRActional DIagonal Chromatography (COFRADIC) [263-265], which allows for selective analysis following HPLC and Mass Spectrometry; selective enrichment of N-termini by removal of internal peptides [261, 262]; selective enrichment of analysed peptides [266].

Gevaert *et al* group has successfully developed the COFRADIC [265, 267, 268] approach and combined with strong cation exchange chromatography (SXC)

[269] to increase the portion of the terminome for proteome analysis.

The positive enrichment approach developed by McDonald *et al* [262] is the one used to generate the datasets used in this chapter; this method comprise the immobilisation of acetylated N-termini and following depletion of internal peptides from the sample.

Another study uses negative N-terminal selection via isotope labelling; the approach developed is termed terminal amine isotope labelling of substrate (TAILS) and allows the tagging and removal of internal peptides from the sample [270, 271]. In the study of Mommen *et al*, which is based on phospho-tagging of internal peptides, these are separated instead via titanium dioxide (TiO_2) chromatography [272].

The identification of N-terminal peptides through selective enrichment can however prove to be challenging at the sampling steps. As the majority of peptides belong to internal peptide sequences it is important to reduce the false positive identifications that these could potentially generate during the analysis [273].

By including the signal peptide data within a study of the N-terminome, it can provide further information of biological relevance. As discussed previously, the hydrophobic region of the signal peptide displays high frequency for specific amino acids at the cleavage site (e.g. alanine, glycine, serine, cysteine). A number of software packages are now available, which allow for prediction of signal peptide based on extrinsic known data [274-279].

Confident protein identification from a single peptide sequence could potentially simplify proteomic analysis while increasing space/time performance if the peptide sequence targeted is the N-terminal peptide. For *Toxoplasma gondii* [101, 193, 229, 245, 250] the set of N-terminal peptides corresponds to less than ~3% of the peptidome. Until a few years ago this target was not considered as achievable due to the difficulty in N-terminal peptide extraction during sample preparation. Bioinformatic strategies would have resulted in biased results containing a high abundance of internal peptide sequences. With the availability of an N-terminal dataset (from *T.gondii*), generated with protocols described below [261], we were given the opportunity to develop optimisation methods for both sequence database design and database query performance.

The sequence databases used include alternative gene models generated with GeneMark, GlimmerHMM and FgeneSH [29, 30, 32, 280]. Also the set of all non-redundant protein from all three strains (ME49, GT1 and VEG) for *T. gondii* were processed with signalP to include both preproteins and mature protein sequences into the database. Additionally a specific database design, similar to the one used by Dormeyer *et al* [281], allowed the pre-selection and pre-parsing of N-terminal sequences from gene models to take into consideration unexpected signal protease sites. The "frayed" database [281] contains all potential mature sequences in the first 100 amino acids of the proteins.

The results from these analysis are used to elucidate the role of signal peptide proteases cleavage sites; additionally the cross comparison between signal peptide characterisation and semi specific enzymatic proteolysis can help targeted database design.

4.3 Methods

The N-terminal dataset from *T. gondii* was processed by Dr. Sanya Sanderson (Wastling group).

4.3.1 Sample preparation

Datasets of enriched N-terminal peptides were generated as follows (Figure 4:1). The enrichment was performed by protein N-terminal acetylation. After protein digestion the peptides free from acetylated N- termini are then bound to biotin and, using biotynilation-binding affinity, effectively removed by streptavidin [261]. The tandem mass spectrometry analysis was performed with HPLC (nanoACQUITY-nLC system from Waters MS technologies, Manchester, UK) coupled with LTQ Ion Trap and LTQ-Orbitrap Velos (ThermoFisher Scientific, Bremen, Germany) mass spectrometer fitted with nanospray ion source [110, 111].

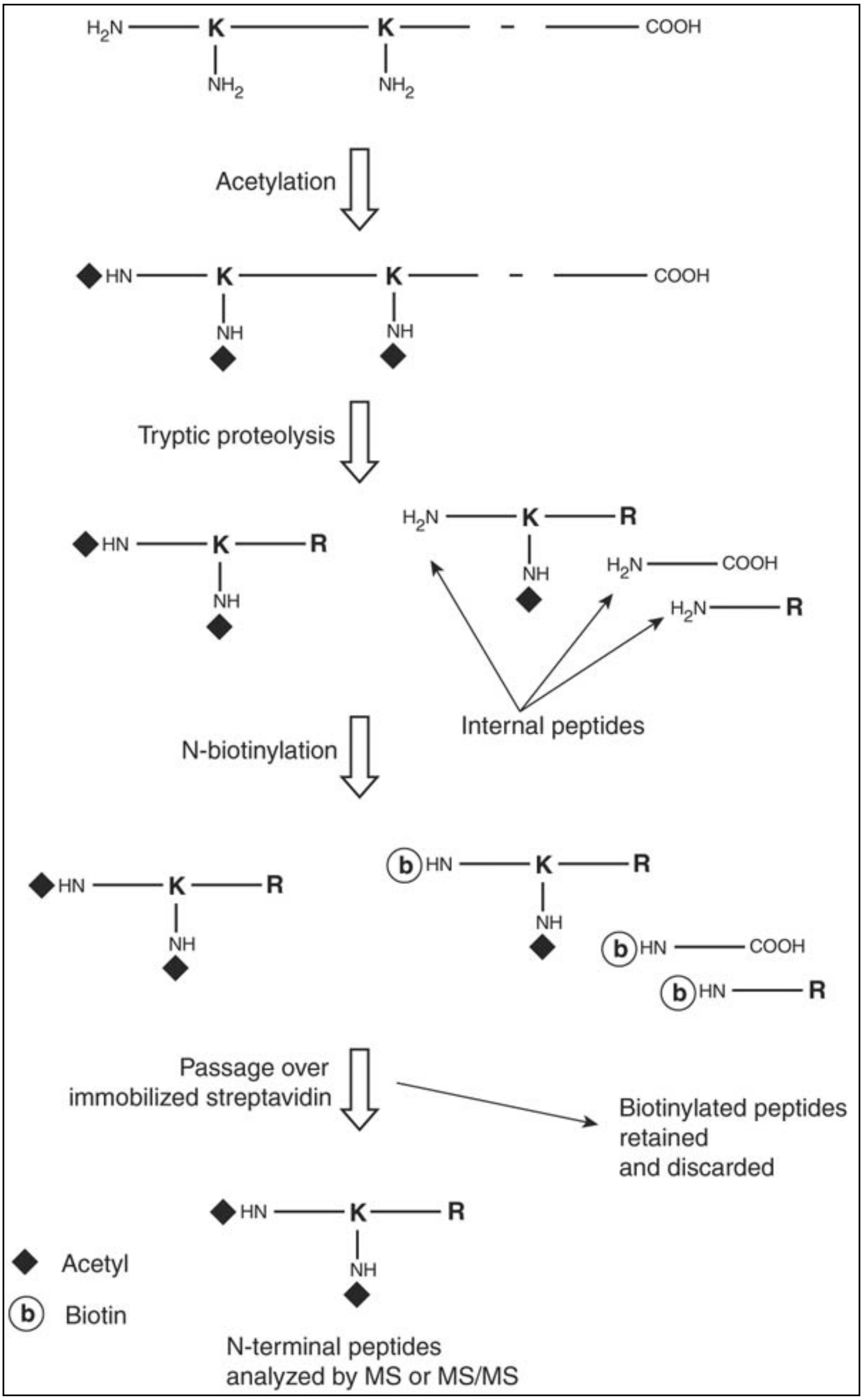


Figure 4:1 Workflow for selective enrichment of N-terminal peptides as described in the study from McDonald *et al* [261].

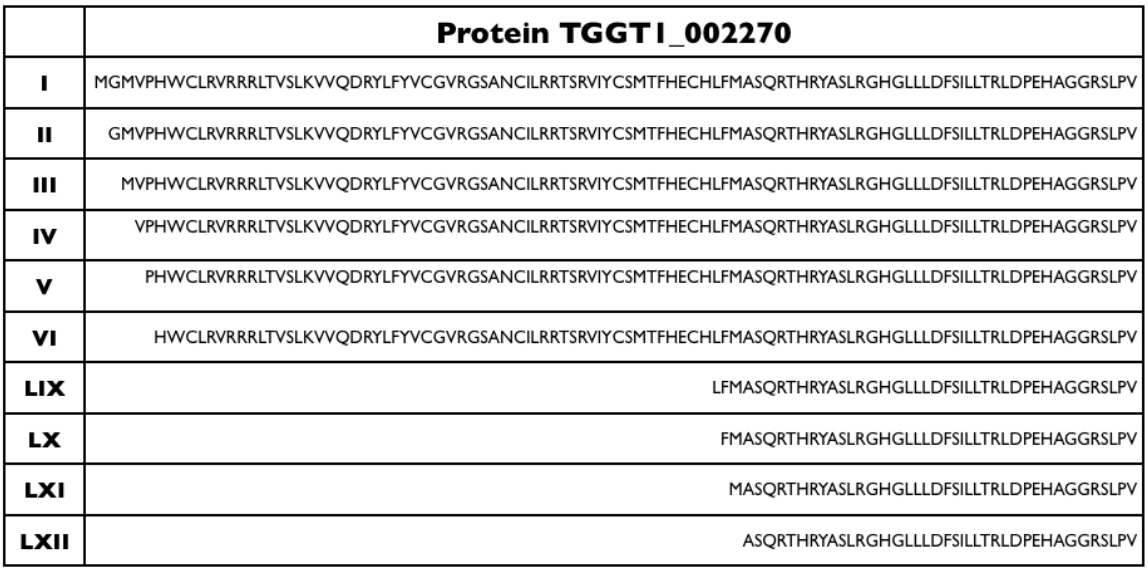
4.3.2 Sequence Database design

The sequence databases, designed to maximise N-terminal peptide discovery, were based on *T.gondii* release 6.2 and comprised: an ORF_MS database, an ORF_MS database concatenated with the official gene models (OGM), the official gene models concatenated with a panel of alternative predictions (P_GM) and a database of frayed sequences of N-terminal regions, simulating all possible signal peptide cleavages.

The ORF_MS were generated only from strain ME49, via six-frame translation of regions flanked by putative start and stop codons, longer than 40 amino acids, as described in chapter one. The official gene models for *T.gondii* included all non-redundant genes from strain ME49, GT1 and VEG as described on EuPathDB.

The panel of alternative gene models were based on the 14 chromosomes sequenced for ME49. The predictions generated with GeneMark (gM) were computed using the default parameters. The predictions generated with GlimmerHMM (GLM) were based on a training set computed on the previously released gene model 5.x; and the default parameters were used when running the algorithm as before. Furthermore the web-based gene finder FgeneSH (FgSH), freely available at softberry.com, was used to make a further set of predictions. Like Glimmer this software algorithm is based on HMM for generating predictions and the publicly available "human model" was used as a training set for *T. gondii* predictions. As with the previous gene finders, the default parameters were used during the prediction.

The official gene model was processed by SignalP 4.0 to provide a set of predicted signal peptides (~2000). This was combined with a set of signal peptides available from EuPathDB (~700) based on signalP 3.0; the result was a non-redundant set of all signal peptides predicted for *T.gondii.* A script was written to create a sequence database that contained both the preproteins and the mature protein sequences including different signal peptide predictions where possible. Different predictions were available for the same protein in some cases, due to the algorithmic differences between release 3.0 and 4.0 of signalP: the latest release being based on neural network while the previous on HMM.

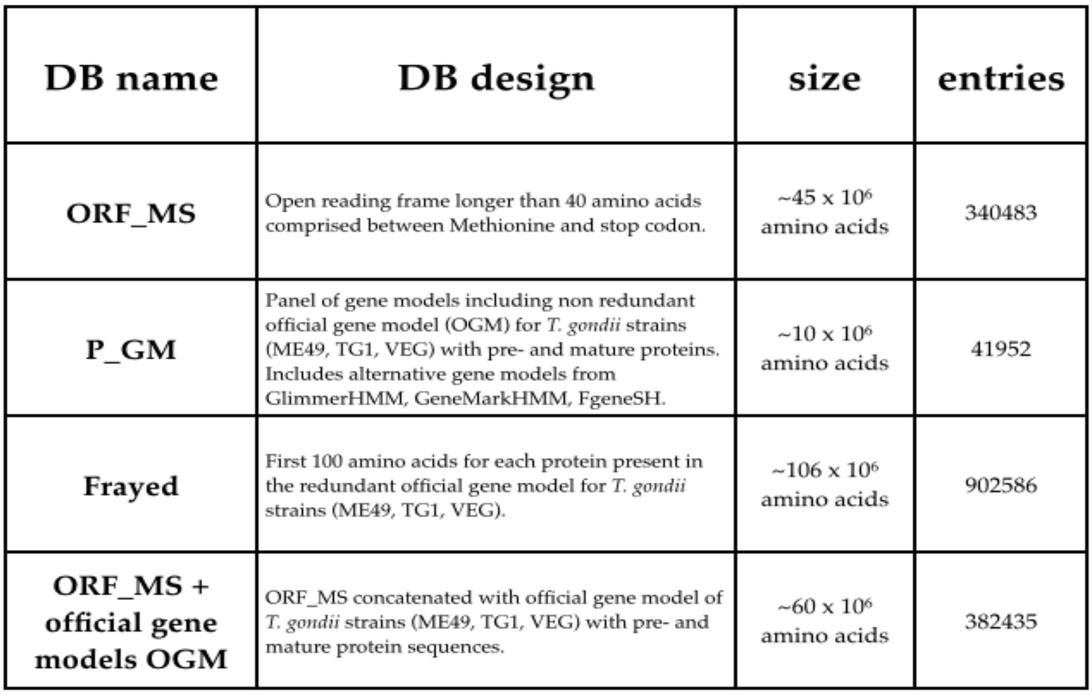


	Protein TGGT1_002270
I	MGMVPHWCLRVRRRLTVSLKVVQDRYLFYVCGVRGSANCILRRTSRVIYCSMTFHECHLFMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
II	GMVPHWCLRVRRRLTVSLKVVQDRYLFYVCGVRGSANCILRRTSRVIYCSMTFHECHLFMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
III	MVPHWCLRVRRRLTVSLKVVQDRYLFYVCGVRGSANCILRRTSRVIYCSMTFHECHLFMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
IV	VPHWCLRVRRRLTVSLKVVQDRYLFYVCGVRGSANCILRRTSRVIYCSMTFHECHLFMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
V	PHWCLRVRRRLTVSLKVVQDRYLFYVCGVRGSANCILRRTSRVIYCSMTFHECHLFMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
VI	HWCLRVRRRLTVSLKVVQDRYLFYVCGVRGSANCILRRTSRVIYCSMTFHECHLFMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
LIX	LFMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
LX	FMASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
LXI	MASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV
LXII	ASQRTHRYASLRGHGLLLDFSILLTRLDPEHAGGRSLPV

Table 4:1 Extract of the frayed sequence database. The example shows protein TGGT1_002270 first 100 amino acids. During each iteration (here starting position I to VI and LIX to LXII in the protein sequence) the first amino acid is cleaved in an attempt to provide the possible unpredicted cleaved signal peptide.

The frayed database was designed to fit specifically the semi specific Arg-C protease used. For each protein sequence from the official gene model the first 100 amino acid from the amino-terminal region were extracted. Then from each of these, a set of potential novel N-termini were created by incremental removal of each N-terminal residue (Table 4:1). This would enable a search for N-terminal peptides to include unexpected signal peptide cleavage sites or new N-termini due to different start codons.

These four sequence databases designed vary in size and number of entries (Table 4:2) and they include also decoy sequences (made by reversed sequences).



DB name	DB design	size	entries
ORF_MS	Open reading frame longer than 40 amino acids comprised between Methionine and stop codon.	~45 x 10^6 amino acids	340483
P_GM	Panel of gene models including non redundant official gene model (OGM) for *T. gondii* strains (ME49, TG1, VEG) with pre- and mature proteins. Includes alternative gene models from GlimmerHMM, GeneMarkHMM, FgeneSH.	~10 x 10^6 amino acids	41952
Frayed	First 100 amino acids for each protein present in the redundant official gene model for *T. gondii* strains (ME49, TG1, VEG).	~106 x 10^6 amino acids	902586
ORF_MS + official gene models OGM	ORF_MS concatenated with official gene model of *T. gondii* strains (ME49, TG1, VEG) with pre- and mature protein sequences.	~60 x 10^6 amino acids	382435

Table 4:2 The four database used in this study and their respective composition. The calculated size and number of fasta entries does not include the decoy sequences.

4.3.3 Database searches

The N-terminal dataset included Tandem MS performed with the LTQ and Orbitrap mass spectrometers. For the database searches the same parameters were considered for fixed modification: N-terminal and lysine acetylation, carbamidomethyl C; variable modification: methionine oxidation. The data obtained with Orbitrap was analysed allowing parent and fragment ion tolerance set at ±0.8 Da and 10 ppm respectively. The parent and fragment ion tolerance for LTQ data were instead set to ± 0.8Da and ±1.5 Da respectively. "Semi specific Arg-C" was set as the enzyme for searches against sequence databases: frayed and ORF combined with official gene models. The ORF database and the panel of gene models (official and alternatives) including signal peptide predictions were searched using "full Arg-C" protease. All search results were then post-processed by the multiple search engine pipeline and rescored by fixed FDR [148].

An initial test on the Orbitrap data (experiment 13) was performed to evaluate the best approach for search performance and output maximisation; the search parameters described above, with semi specific protease were used with the addition of variable modification of serine acetylation. The sequence database tested was the one comprising alternative gene models, the official annotation including preproteins and mature proteins. The test results, assessed by 1% fixed FDR, indicated an increase in TP for the searches with no additional variable modification. For both MASCOT and X!Tandem there was an increase in peptide identifications without serine acetylation, respectively 321 and 169 PSMs; when including this modification the peptide identification dropped to 260 and 126 PSMs respectively.

4.3.4 Signal peptide: study across species

To study the distribution of signal peptides and their cleavage motifs, signalP4.0 was used to generate a set of predicted signal peptide for five other eukaryotic organisms: *A. thaliana* [196, 282], *C. elegans* [283-285], *D. melanogaster* [195, 286, 287], *H. sapiens* [288], *M. musculus* [289]. Compared to the set of predicted signal peptides for *T.gondii*, the number of predicted signal peptides for the other organisms was considerably larger as it included all available protein isoforms available from public repositories: ~101K for *H. sapiens*, ~59K for *M. musculus,* 39K for *A. thaliana,* ~26K for *D. melanogaster* and *C. elegans.* The signal peptide composition was studied to understand the length distribution statistics; as well as providing additional information on the specificity and frequency of the amino acids located at the cleavage site position (-1, -2, -3). This evidence was then used during the N-terminal study on *T.gondii* in an attempt to provide additional evidence for confirming, correcting and proposing candidate signal peptides through N-terminal peptide identifications from the MS data.

4.3.5 Study on signal peptide cleavage

Analyses performed on signal peptides across different eukaryotic organisms showed the high degree of similarity in both lengths of signal peptides and their cleavage sites. As already shown in other studies the signal peptides generally have a length between 20 and 50 amino acids, visible as the median in Figure 4:2. This information can play an important role when analysing specific PSMs in search of candidate N-terminal peptides. As an example, if a PSM is expected to have been generated from an N-terminal peptide (as it has a fixed modification of N-terminal acetylation), it would have a higher probability of having arisen from signal peptide cleavage if it is located within the expected position of the protein sequence. This method could thus either correct signal peptide predictions or identify new signal peptides (not predicted by algorithms).

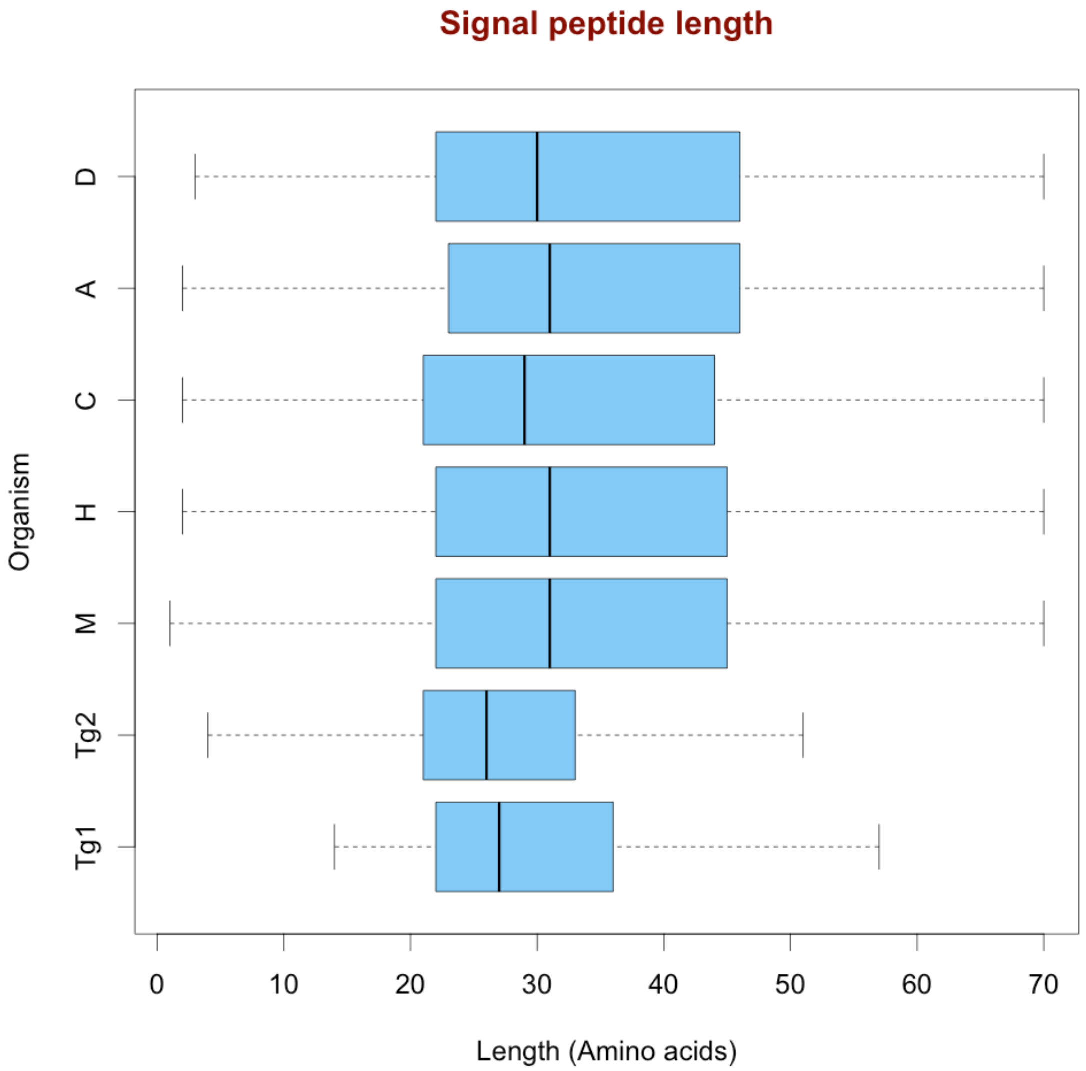


Figure 4:2 The boxplot shows the graph of the signal peptide lengths as predicted by SignalP The organisms are *D. melanogaster* (D), *A. thaliana* (A), *C. elegans* (C), *H. sapiens* (H), *M. musculus* (M). For *T. gondii,* the predictions are respectively from SignalP 3.0 (Tg2) and

SignalP 4.0 (Tg1).

Additionally the cleavage site were analysed in depth revealing amino acid frequencies at positions -1, -2, and -3, across different species (Figure 4:3). The last three amino acids from signal peptide predictions were counted and the three most frequent ones were selected.

In order of frequency - alanine, serine, glycine, and lysine appear consistently at position -1 in all species (Figure 4:3). At position -2 there appears to be a weaker motif, although serine is consistently reported. Position -3 appears to have similar importance as position -1, given the higher consensus of amino acids across species. The four most abundant are valine, alanine, threonine and serine.

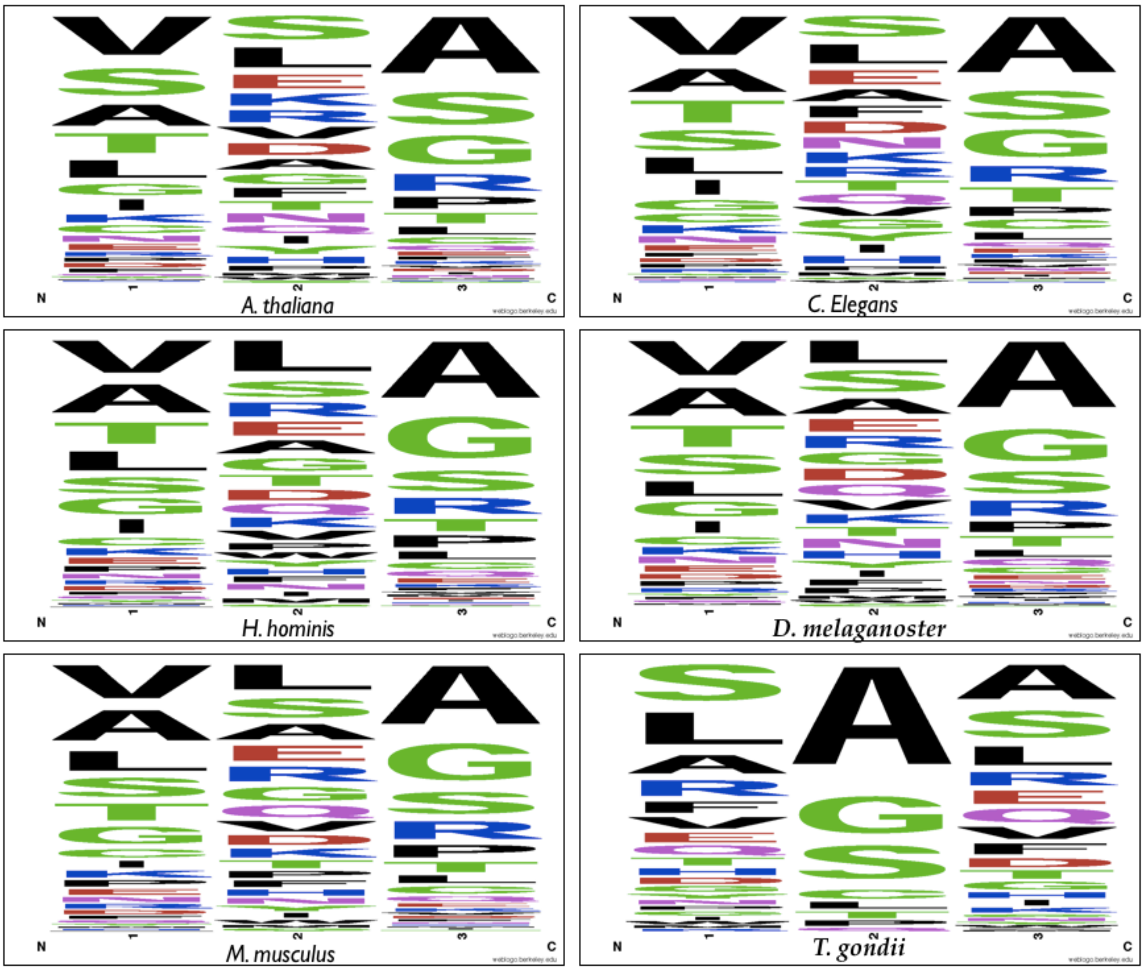


Figure 4:3 For the eukaryote species analysed, the graph generated with weblogo [290] **reflects the pattern frequency of cleavage sites. From left to right they represent the last three**

amino acids of signal peptides 1, 2, 3 (corresponding respectively to -3, -2, -1). The size of each letter represents the frequency at specific position.

4.3.6 Dataset post-processing algorithms

The data on signal peptides has been used to shortlist specific peptide sequences to confirm the likelihood of an identified peptides being an N-terminal peptide. The final results were parsed and organised depending on the peptide sequence composition (Figure 4:4). The remaining peptides were further analysed in order to discriminate between novel N-terminal peptides and internal peptides. An algorithm was devised to use the collected data on signal peptides for *T. gondii* to process these peptides (internal/ unexpected signal peptides). The statistical data also comprised the 20 most abundant three amino acid combinations from cleavage sites.

The identified peptides were filtered by residue arrangement first to extract all possible candidate N-terminal sequences (with methionine or cleaved by NME). The peptides left were then judged as candidate N-terminal peptides after signal peptidase cleavage (Figure 4:4). If a signal peptide (SP) had been predicted, it is a matter of confirming the N-terminus after cleavage by sequence alignment. If the alignment was not successful then the peptide was checked if the coordinate range and the preceding residue pattern fall within the studied cleavage sites. If there was no evidence for being a signal peptide it was considered as an internal peptide. The final assessment was made for identified peptide in proteins with no predicted signal peptide; by analysing the coordinate range and preceding residue pattern it is possible to shortlist candidate N-terminal peptide with novel SP and discard the likely internal peptides. All output files were retained for a final manual evaluation.

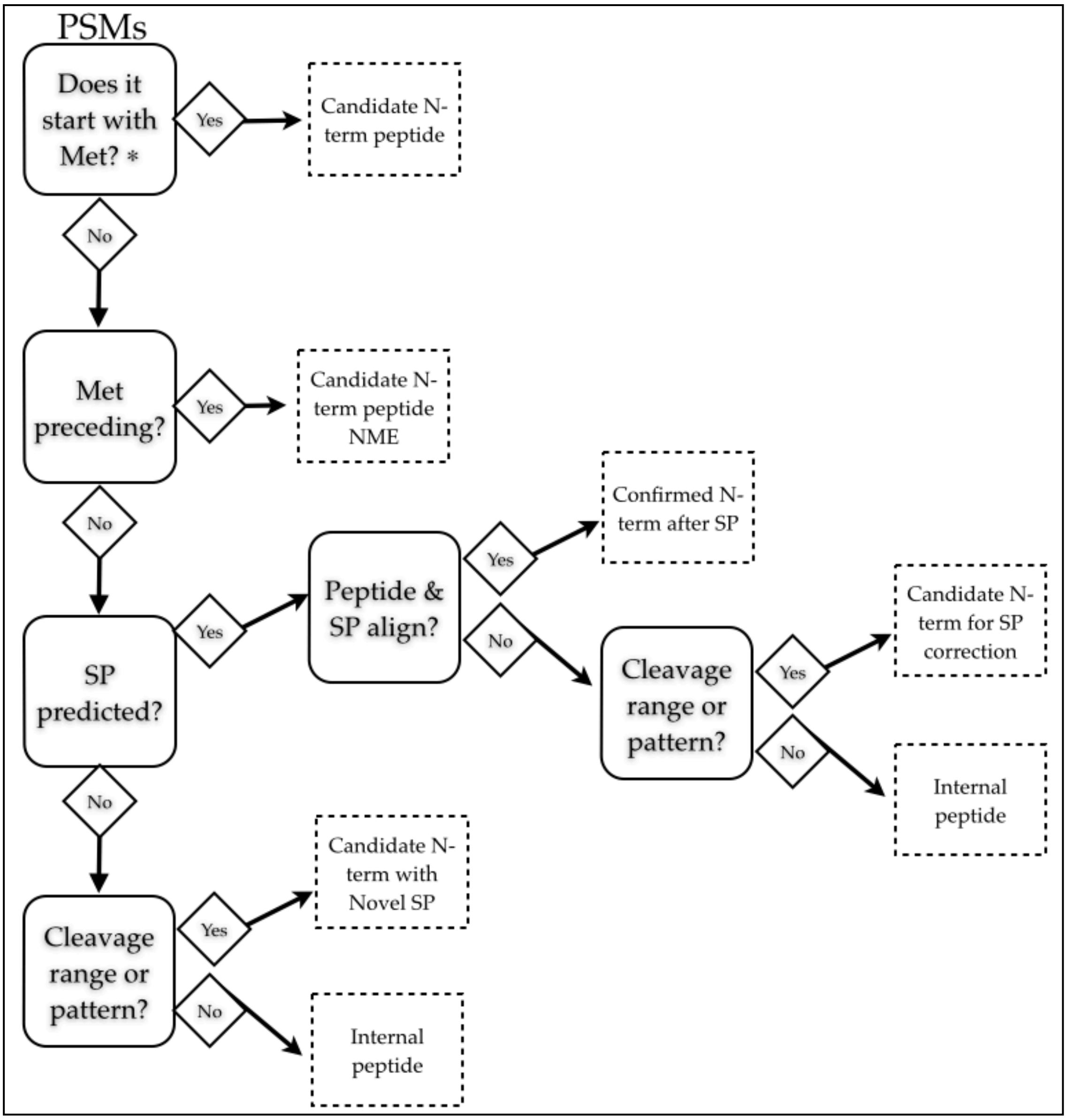

PSMs
Does it start with Met? *
Yes
Candidate N-term peptide
No
Met preceding?
Yes
Candidate N-term peptide NME
No
SP predicted?
Yes
Peptide & SP align?
Yes
Confirmed N-term after SP
No
Cleavage range or pattern?
Yes
Candidate N-term for SP correction
No
Internal peptide
No
Cleavage range or pattern?
Yes
Candidate N-term with Novel SP
No
Internal peptide

Figure 4:4 Workflow for the assessment of identified peptide sequences. These were analysed to evaluate their position within the protein to predict/ correct signal peptide cleavage site. This enabled also the evaluation of the frequency of NME. Although the search engine can already take into account methionine excision, the designed algorithm also evaluates the peptide position within the protein sequence to highlight for potential novel N-terminal peptide (*the first condition is that the Methionine is not preceded by a tryptic site).

4.4 Results

4.4.1 How database design influences search engine performance

The N-terminal dataset comprised 14 analyses on the LTQ and 3 analyses with the Orbitrap. The combined results (MASCOT and X!Tandem), rescored by FDR, were filtered at 1% fixed FDR. Through the results obtained from the traditional sequence database (official gene models only) it was possible to identify ~1999 PSMs from ~386 proteins; this can be viewed as a baseline during assessment of different approaches. The designed database, comprising the frayed amino-termini sequences of official gene models, led to the identification of 2135 PSMs from ~388 genes. Of the designed sequence databases studied the ORF_MS produced the lowest results (1717 PSMs) while the largest number of identification was generated with ORF_MS concatenated with the panel of alternative gene models (2601 PSMs) (Figure 4:5).

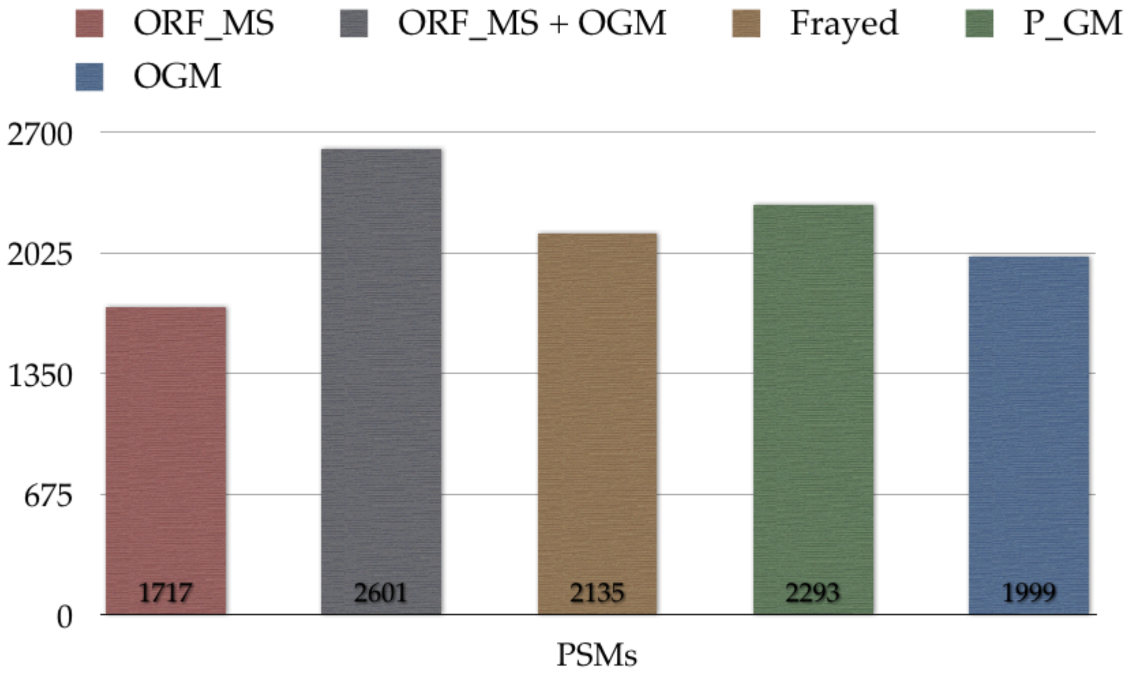


Figure 4:5 **The chart shows the complete set of PSMs from N-terminal dataset queries against different sequence databases. These PSMs have not been filtered to extract true N-terminal candidate peptides from internal peptide sequences. Given the current quality of *T. gondii* annotation, it is not unexpected to see the ORF_MS results at lowest rank. The top-**

ranked results belong to ORF_MS-gene models database.

From these results, with additional post-processing, it was possible to remove redundancy at peptide level and highlight specific peptides common to ORF_MS database, frayed database and the ORF_MS with concatenated official gene models (Figure 4:6).

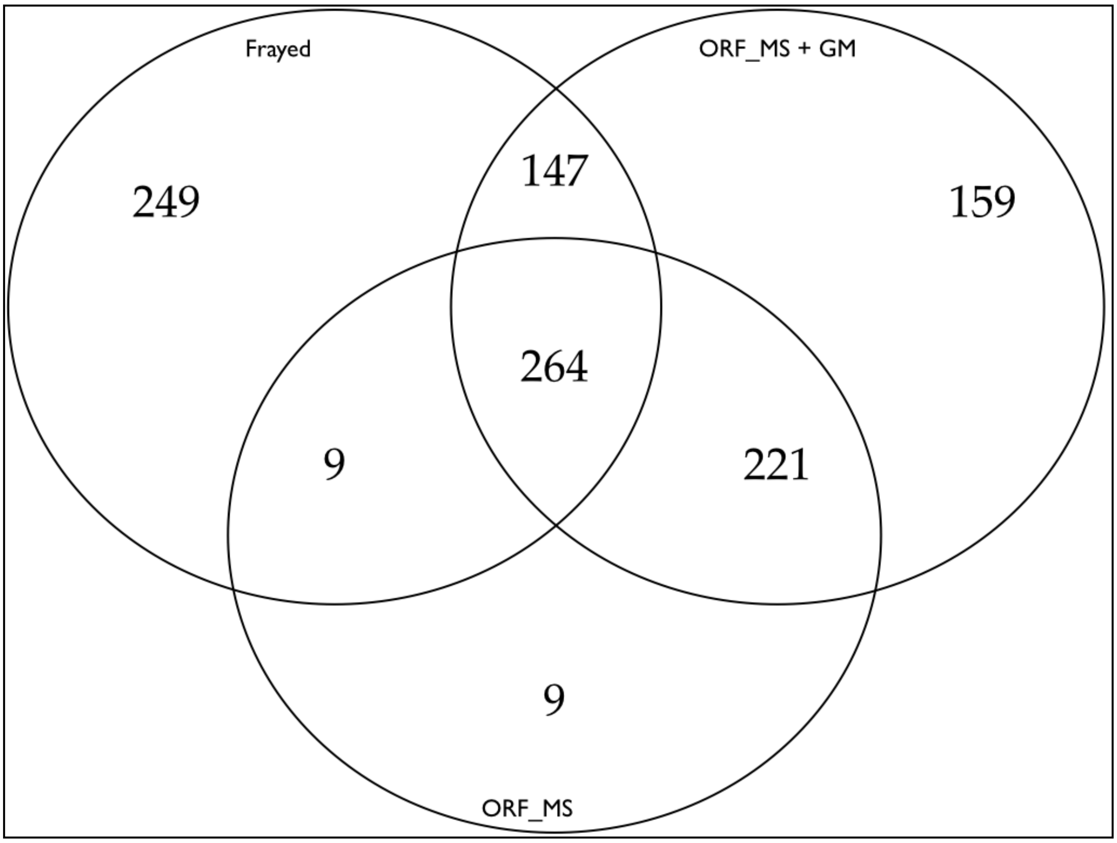


Figure 4:6 The Venn diagram shows the overlap across datasets obtained from different database designs: frayed, ORF_MS with gene models (ORF_MS+GM), ORF_MS. A script was used to remove the redundancy at peptide level. The majority of peptides appear to be common to all three datasets; although the dataset obtained from querying the frayed database yielded the second highest number of peptides. Due to the design of this database, all peptides from this dataset are also present on the other two databases; hence their sole presence on the frayed dataset is the result of statistics (i.e. search engine and scoring algorithms). 159 peptides are uniquely identified on ORF_MS+GM dataset, but out of these 159 only 25 peptides are not present on the official gene model (subset of ORF_MS+GM sequence database); however, after further examination of the tryptic and methionine sites, these 25 peptides were assessed as belonging to internal sequences. Of the 159 peptides 134 were also aligned on the official gene model but their starting position was found to be higher

than the length threshold used for the frayed database: 100 amino acids. Only three out of the eight peptides uniquely identified on the ORF_MS dataset were also present on the official gene model database. In particular these three peptides aligned with internal exons at position outside frayed database threshold; similarly the five peptides uniquely present on the ORF_MS dataset and database aligned with an intragenic region.

From Figure 4:6 it is possible to identify the small number of peptide sequences identified on ORF_MS, which do not align with the panel of gene models. The frayed database appears to rank third, next to the panel of gene models. The frayed database intrinsically provides a way for N-terminal peptide candidate filtering by discarding internal peptide sequences (>100 amino acids from the N-terminus). In Figure 4:6 there is a list of non-redundant peptides gathered from the datasets; the number of peptides here discussed is based on a non-redundant list of peptide sequences. All the peptide sequences unique to searches of the frayed database are also present on the other sequence databases. Of the peptide sequences unique to ORF_MS dataset (9), three of these are also present on official annotation database but aligned with an internal exon, hence did not appear on the frayed database. The other five peptide sequences uniquely present on the ORF_MS dataset were also unique to this database. Of the peptide sequences uniquely identified on the ORF_MS+GM (159), only 25 peptides do not appear on the official gene annotation; however, by examining the tryptic sites and methionine position preceding the peptides, these were not assessed as N-terminal sequences. The peptide sequences uniquely present on the frayed dataset (249) were also present on the other databases, but did not appear on other dataset as result of statistical evaluations of the search engines and rescoring algorithms. The peptide sequences common to ORF_MS and ORF_MS+GM identified 221 peptides. Out of these 31 peptides were not present on the frayed database but half of these were present on the official gene model. The peptide sequences common to ORF_MS+GM and frayed datasets (147 peptide sequences) are all identified on the official annotation within ORF_MS+GM; this indicates that these peptides can be located within the first 100 amino acids of the protein sequences from the official annotation. Although out of these 147 peptides, 34 have been identified also on ORF_ MS present on ORF_MS+GM they were missing from the ORF_ MS dataset due to statistical and search engine scoring and thresholding. The same can be said about the peptide sequences common to frayed and ORF_ MS datasets as, although missing from the third dataset, they were common to all three databases. The peptides identified on the gene models database can potentially lead to prediction of novel gene structure and identification of signal peptides; however the results would need additional analysis to eliminate internal peptides.

Listed in Table 4:3 non-redundant PSMs, obtained for each analysis and each sequence database, allow direct comparisons of the performance of different techniques. In this table the peptide redundancy has yet to be removed as this visualisation of the identified spectra is meant to provide a method to examine data acquisition protocols across all experiments. Focusing on Orbitrap data results, notably of higher quality, the importance of database design becomes evident as for the ratio of hits of the novel versus official annotation database is greatly increased.

These results were further analysed to provide positional information of the non-redundant peptides within the proteins of interest. The object of these analyses was to confirm peptide sequences as being N-terminal peptides, filtering out residual internal peptides (considered as FP), and to produce an overview of amino terminal cleavages (Signal peptides, NME).

Total PSMs at 1% FDR					
Databases	Frayed	ORF_MS	ORF_MS + official gene models (OGM)	Panel of gene models: alternative, officials and pre-proteins (P_GM)	Official Gene models (OGM)
Experiments					
5_1	71	53	69	81	65
5_2	122	63	120	107	88
5_3	103	62	102	116	87
9_5micro	32	26	43	24	15
9_10	87	50	91	66	58
10_5	190	175	271	274	260
10_10	184	189	311	310	277
10_extr_10	109	163	248	267	250
10_extract_5	93	124	195	193	187
11_5micro	41	32	45	40	36
12_5micro	61	26	40	43	37
13_post_r1	5	5	7	4	2
13_post_r2	9	10	12	9	4
ORBTR_SA_SS_11	315	243	345	287	226
ORBTR_SA_SS_12	391	231	377	292	252
ORBTR_SA_SA_13	325	272	328	183	172

Table 4:3 A comparison of the results for each experiment queried with multiple search engines against different sequence databases. This overview can help to evaluate the effectiveness of specific experiments (left-most column) and the performance of each database (top row). The count of PSMs listed is the estimated TP at 1% fixed FDR although peptide redundancy has yet to be removed.

By just evaluating the dataset from frayed database I attempted to explain the identified peptide as N-terminal. With this approach the frayed dataset highlighted 669 non-redundant peptides as N-terminal candidates; next their residue pattern and position, within the gene, was examined to discard possible internal peptides. With this process I shortlisted a dataset of 26 peptide sequences as novel candidate N-terminal peptides with SP cleavage (supplementary data in appendix B Table 7:13). For these no signal peptide had been predicted. Around half of these belonged to putative proteins while the other half to hypothetical ones. A list of 20 peptides (from 17 proteins) presented candidate N-terminal peptides with adjusted SP cleavage site (supplementary data in Table 7:14). For these proteins the signal peptides were predicted, but their cleavage site did not completely align with the identified peptides. Around 18 peptide sequences were confirmed to be N-terminal peptides, confirming the correctly predicted signal peptides (supplementary data in Table 7:15). The remaining peptide sequences were assessed to be internal peptides. The same approach was performed on the other datasets: ORF_MS with panel of gene models (P_GM). However the vast majority of the peptide sequences were assessed to be internal peptides (preceded by proteolytic site Arginine) (Table 4:4).

	Novel N-terminal candidate	**Corrected N-terminal candidate**	**Confirmed N-terminal candidate**
Frayed db	26	20	18
ORF and gene models db	28	0	0
panel gene models db	13	21	0

Table 4:4 The table lists the peptide sequence located around/ after the signal peptide cleavage site identified in the databases but not in official gene models. These were assessed as being N-terminal candidate peptides where signal peptide had been confirmed, not predicted or possibly to be corrected.

By reviewing together the results from the three databases (frayed, ORF_MS+GM, panel of gene models) it was possible to analyse the peptides that can confirm/ correct predicted signal peptides and suggest novel predictions. Of particular interest are the peptide sequences that appear to overlap within the predicted signal peptides. For example the identified peptide SVAHAQTAASEAEAATKVPDFR overlaps with one signal peptide MLSSALRSVRPAASAASRRFASVAHAQTAASEAEAA (signaP4.0) and does not align with MLSSALRSVRPAASAA (signalP3.0), protein TGME49_015280, potentially indicating that neither prediction is correct.

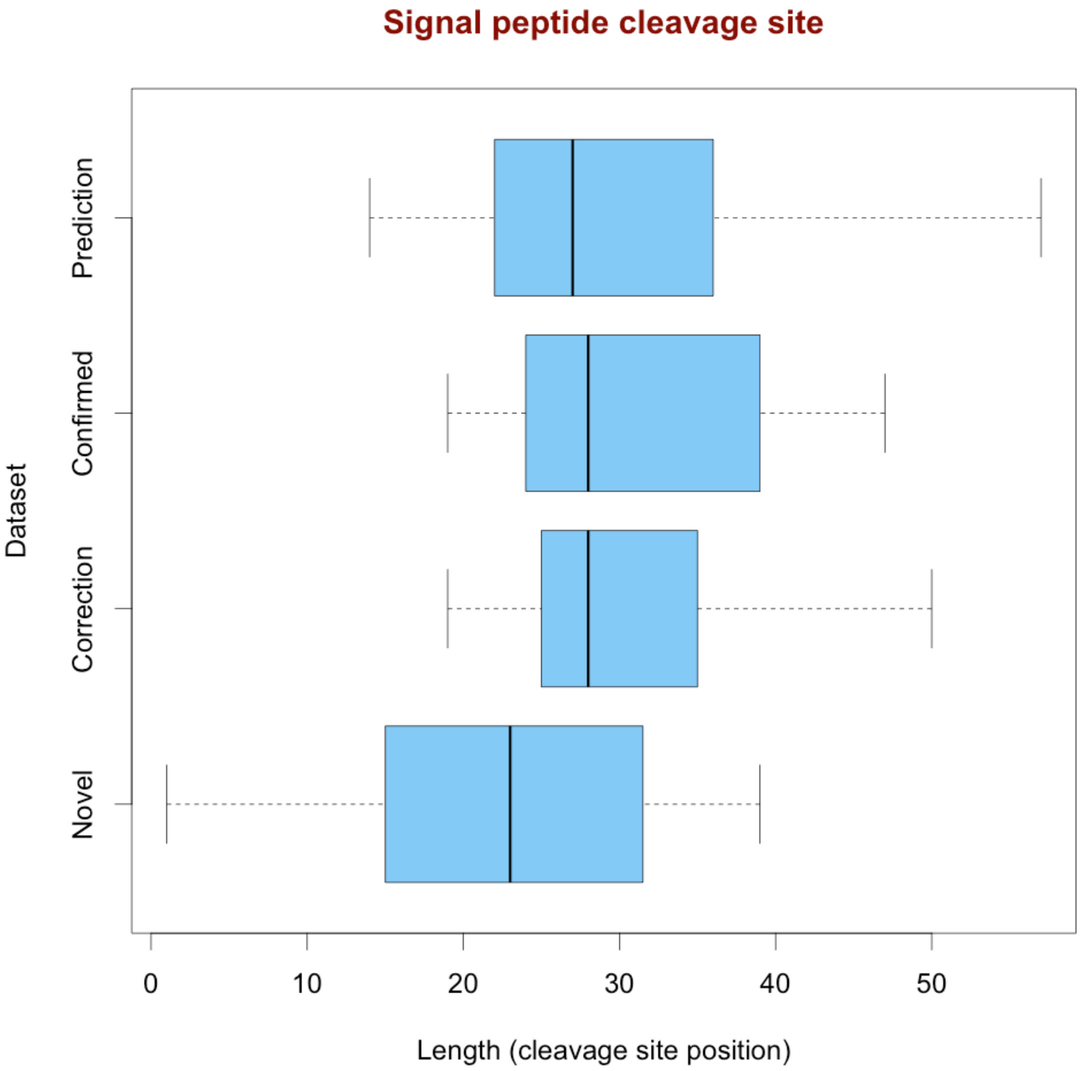


Figure 4:7 Signal peptide cleavage site from N-terminal peptides evaluated as confirmed, corrected and not predicted.

We compared the peptides that confirmed signal peptides against both candidates for correction of the signal peptide cleavage and novel signal peptides. The idea was to examine how the length of signal peptides and the patterns of cleavage sites were allocated (Table 4:4, Figure 4:7 and Figure 4:8). The pattern of amino acids for confirmed signal peptides appears to be reasonable stable, with a high frequency at position -1 and -3 of alanine, glycine and valine. For both the candidate correction and novel signal peptides the pattern at position -1 and -3 display similar frequencies to each other; additionally there are present amino acid patterns with very low frequency (e.g. threonine, glutamic acid, lysine, glutamine). This could be caused by a bias in the small dataset, by there being real unexpected cleavage sites or by internal peptide contaminating the dataset. Additional experiments could help to discriminate with them high confidence.

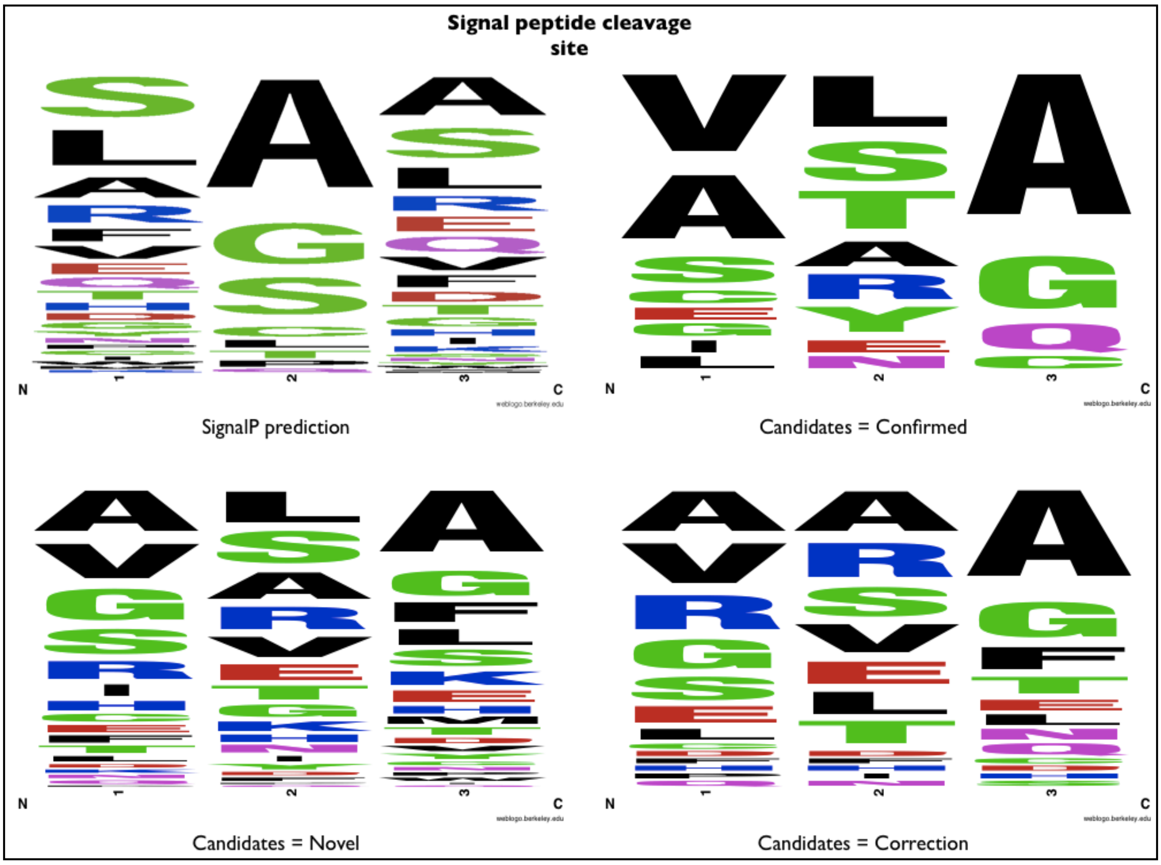
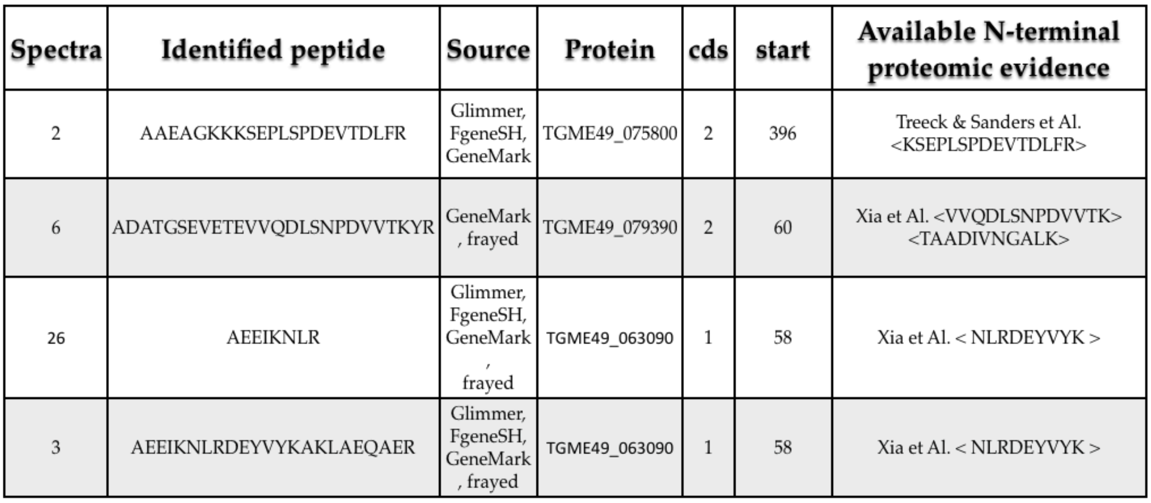


Figure 4:8 The amino acid frequency extracted from the cleavage site of signal peptides for the three set considered (confirmed SP 18 PSMs, candidate for correction SP 40 PSMs, candidate for novel SP 112 PSMs). The high abundance of alanine, glycine and valine is coherent with the study performed on other eukaryotic species.

From this pool of data it was possible to extract and analyse the peptide sequences uniquely identified from alternative gene models (Table 4:5). In total there were identified 4 peptide sequences; these were assessed by official model as internal peptides while on the alternative gene models were considered as N-terminal peptides. 3 of these peptides had also been identified on the frayed database.



Spectra	Identified peptide	Source	Protein	cds	start	Available N-terminal proteomic evidence
2	AAEAGKKKSEPLSPDEVTDLFR	Glimmer, FgeneSH, GeneMark	TGME49_075800	2	396	Treeck & Sanders et Al. <KSEPLSPDEVTDLFR>
6	ADATGSEVETEVVQDLSNPDVVTKYR	GeneMark , frayed	TGME49_079390	2	60	Xia et Al. <VVQDLSNPDVVTK> <TAADIVNGALK>
26	AEEIKNLR	Glimmer, FgeneSH, GeneMark , frayed	TGME49_063090	1	58	Xia et Al. < NLRDEYVYK >
3	AEEIKNLRDEYVYKAKLAEQAER	Glimmer, FgeneSH, GeneMark , frayed	TGME49_063090	1	58	Xia et Al. < NLRDEYVYK >

Table 4:5 The table displays the list of peptide sequence identified with novel database design and the number of spectra identified. From the left, the first column points out the number of spectra used for peptide identification (second column). The source column refers to the sequence database or gene model it was aligned on. The CDS column refers to the coding sequence (exon) of the protein from the official gene model. The "start" column locates the start of the identified peptide on the protein sequence. The last column shows the currently available proteomic evidence, viewable on EuPathDB (Gbrowse). All proteins are still described as putative. The first peptide was matched also on the official annotation but the selective algorithm did not shortlist as it was considered an internal peptide.

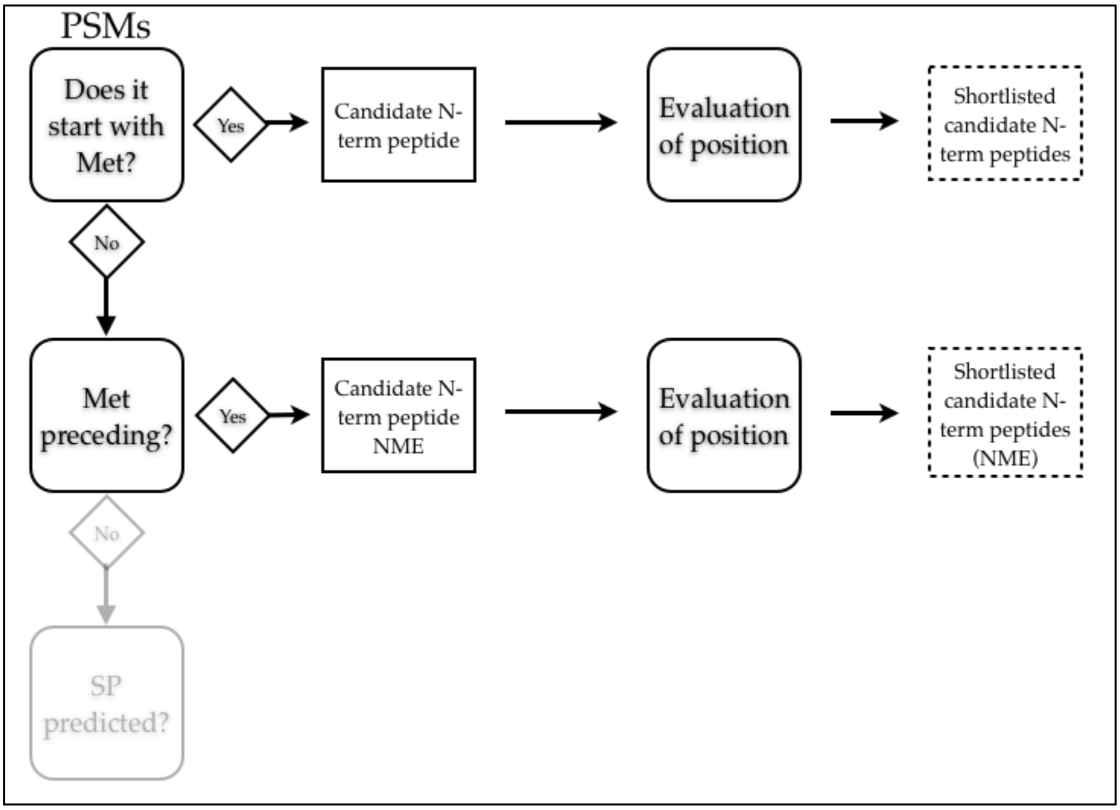


Figure 4:9 Workflow of steps in analysing N-terminal candidate peptides that present Methionine at the N-terminus.

4.4.2 Analysis of the N-terminal Methionine

By combining the results from four datasets (official annotation only, frayed, ORF_MS with official and alternative gene models, panel of gene models) it is possible to see in Figure 4:10 the overall peptide position within the protein sequences from the identified peptide sequences. I used previously described post-processing algorithms to generate a list of selected non-redundant N-terminal peptides. After filtering out all dubious internal peptides an additional step grouped the results by peptides that retained the N-terminal Methionine (85 PSMs) and those presenting NME event (204 PSMs). The data is obtained from the results across the three unified datasets where also each peptide's relative position has been recalculated based on protein sequences. Although post-processing analyses allowed discarding internal peptides, the density plot shows the general trend of position in the N-terminal region.

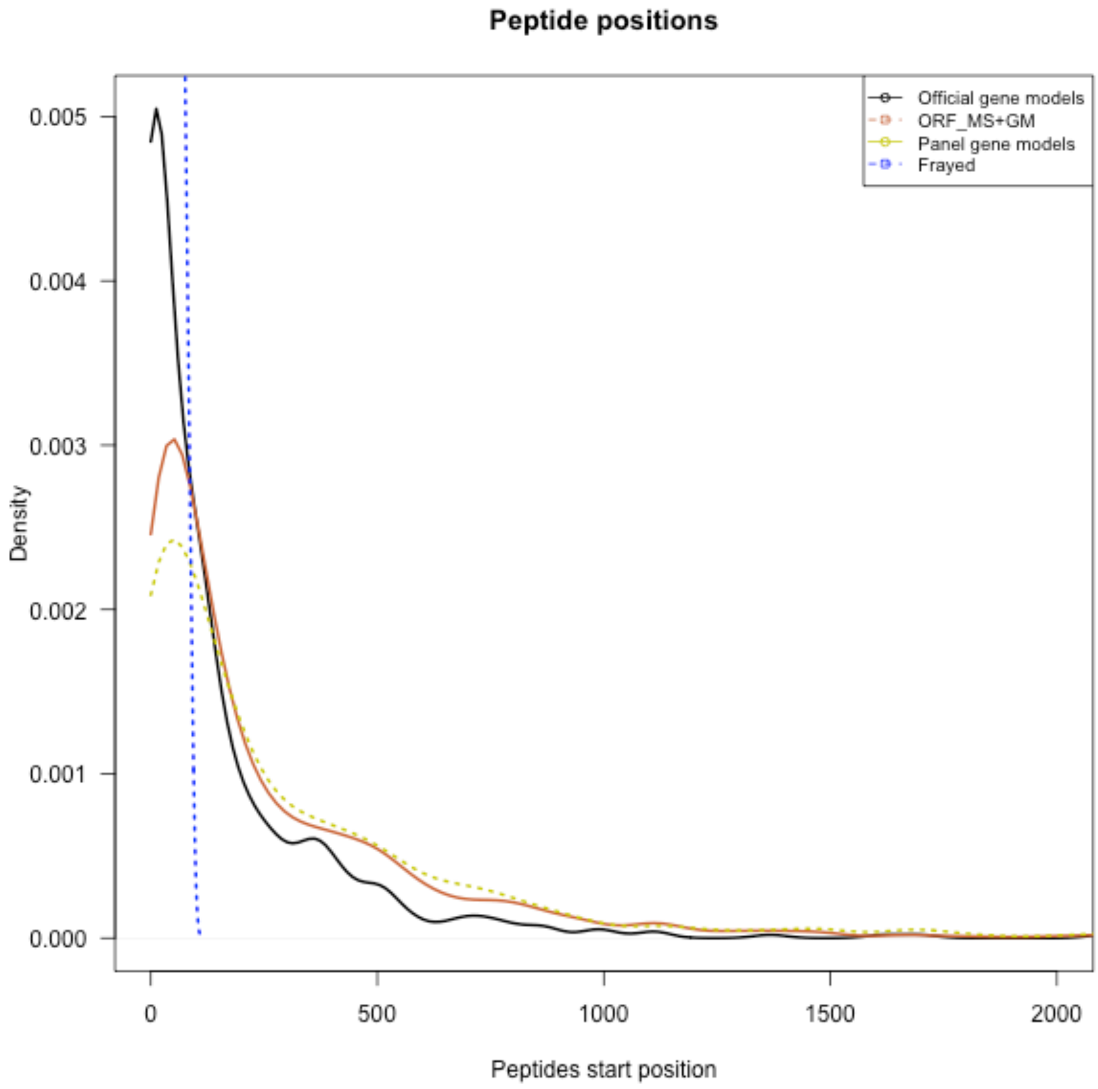


Figure 4:10 The plot shows four dataset obtained respectively from the sequence database: official gene annotation, frayed, ORF_MS with official and alternative gene models, panel of gene models displaying the density of the start position of identified peptide sequences across all four datasets. The overall results from frayed dataset can be seen as database pre-filtering technique as the position are strictly within 100 amino acids from the start of the proteins. Additionally from the frayed dataset, it can be seen start position density is higher towards the 5′ of the gene. From the other dataset it is possible to view the density of internal peptides that are still being identified through search engines.

The first list yielded 85 peptides with Methionine at the N-terminus; by additional positional filtering it was possible to discard 6 false positives (internal peptides) as located outside SP region range on known proteins (supplementary data in Figure 7:16 - the appendix B). Although discarded at this stage some of these peptides might well be a true positive if the gene model is considerably wrong. The remaining 79 peptides identified 66 known proteins (official annotation), 5 alternative protein sequences and 5 ORF_MS.

The second list of non-redundant peptides provided an opportunity for identifying N-terminal peptides and for evaluating the role of methionine amino-peptidase, which results in N-terminal methionine excision. From the dataset of 204 peptides (Table 7:17 - appendix B), it appears that methionine immediately precedes the peptide ~98% (196 peptides) while for the remaining 2% (8 peptides) it is at position -2. Of all identified peptides whose N-terminus was located within the first three amino acids of the proteins (289 peptides), the NME modification is present with very high frequency: ~70%.

We identified 192 proteins from the 204 short-listed peptides (supplementary data viewable in Table 7:17 in the appendix B). Of these peptides, one was found to align with the predicted SP suggesting that either the prediction was a false positive or the protein had not been processed by signal peptidase. Five peptide sequences were uniquely identified from the ORF_MS database. For the remaining eight peptides, although the identified protein did not have signal peptide, their position was within range of confirmed signal peptides.

In this work we were able to identify with confidence 308 non-redundant N-terminal peptides (from 1747 PSMs) from the official gene models. Substantial evidence for confirming and predicting the signal peptide cleavage sites was presented for 152 identified peptides. The NME frequency throughout the identified N-terminome of *T. gondii* was evaluated to be ~70%. We provide 4 candidates for novel protein N-termini as peptides were identified on alternative gene models at the start site. Additional peptide sequences might need further experimental proof to assess whether they are internal peptides.

4.5 Conclusion

In this chapter we show how advances in targeted proteomics at the sample preparation stage can be coupled with improvements in bioinformatics approaches for targeted peptide identification. The methodology described represents an attempt at maximising proteomic identification of amino terminal peptide identification. Although only one protease was used for these analyses, as proved successfully in other studies [291], the use of multiple proteases would generate datasets of complementary PSMs. These could potentially produce overlapping peptides resulting in higher confidence. Additional data generated with available software packages, such as signal peptide predictions with alternative gene models, could be further analysed with known structures. In this study the signal peptide sequences were generated with default parameters and were not analysed against known data, such as the Pfam database [292, 293]. As signal peptide predictions can differ, this information, together with available proteomic evidence from previous experiments (e.g. data from EupathDB), could be used to highlight specific proteins for database design and dataset post-processing.

The frayed database designed in this study was based on the official gene models and it can be considered as pre-filter for N-terminal peptides searches to also assess signal peptide predictions. For this reason, this approach should be applied to alternative gene models in order to restrict post-processing analysis; cross comparisons would allow to analyse those identified dubious peptides matched only on alternative models. Additionally, pre-processing sequence databases in this way could be useful for correcting N-terminal sequences and signal peptides for previously predicted proteins from newly sequenced genomes. Although this database is an attempt to pre-filter the data from possible internal peptides, these could potentially be used to increase confidence in low abundance protein identification for novel N-terminal peptides, achievable with sequence database comparison.

Generating a panel of gene models can help confirm and correct the amino terminal structure of annotated genes, even for organisms on which detailed manual curation has been already performed. Signal peptide predictions validated with proteomic N-terminal peptide sequences would allow to determine the mature protein sequence from gene model. However it has been shown how, for a highly annotated genome, database design can directly affect the performance of the search as designing targeted database can reduce search time and maximise results with higher confidence. The frayed database returned a higher number of PSMs uniquely identified on the dataset although the sequence database could be considered as subset of ORF_MS together with the panel of gene models database.

With this study it was possible to provide: proteomic validation for amino terminal peptides, informative data on possible adjustments and confirmation of signal peptides based on the official gene model. Also it provided proteomic validation of the statistical importance of NME event and the high frequency across *T. gondii* genome. Additionally it was possible to evaluate the different performances of different database designs; for instance, statistical scoring can be affected by the size of the database searched and as such, combining panels of gene models with ORF_MS can yield results with lower significance. By designing frayed sequences the database search can be effectively restricted at a positional level, and this would ideally be suitable also for alternative gene models. Similarly the signal peptide predictions can prove to be effective when performed also on alternative gene models with the predicted dataset compared against the frayed design. The latter can effectively generate all possible signal peptide splice site even for protein sequence lacking a predicted signal peptide. Because of this, the design of an official gene model containing both the pre-protein and the mature sequence would need to be accurately evaluated based on the ratio of predicted versus expected proteins with signal peptide (e.g. analyses with Pfam).

With the additional search engine to the multiple search mode for future work it would be possible to reduce the time of post-processing analyses as the dataset would yield higher confidence identifications. Given the large quantity of data analysed it is mostly important to decrease manual validation but it is nevertheless required in order to correctly estimate FDR.

As demonstrated in other studies, analyses across species can help defining generalisations that can be abstracted in bioinformatic algorithms. Targeting signal peptides from a proteomic point of view remains still challenging, as additional validation is needed to corroborate the findings.

5 Final discussion

5.1 Overview on the thesis:

At present the number of genome sequencing projects for various organisms is steadily increasing. As manual genome annotation can be considered as a bottleneck we need improvements in automated methods to generate gene models with validated sequences. The identification of splice sites and N-terminal peptides is crucial within the proteogenomic context.

In this thesis I worked on proteogenomic approaches to provide methods to identify specific peptides from MS/MS data. I have attempted to exploit and combine together different bioinformatic techniques to target these specific peptides.

In this study I focused on the construction of alternative gene models and ORF_SS database to look at the current gene model state. With this I also assessed whether these generated sequence database can be efficiently used to correct gene sequences or predict novel genes. Additionally, I conducted an evaluation on alternative gene models to provide an insight on the differences between gene finder and whether the training set used for the HMM algorithm affects the quality of gene model predictions. I tried to overcome the bioinformatic challenges given by ISPs identification, using a new hybrid "blind" approach. In the final project of this thesis, I focused on maximizing N-terminal peptide identification and I evaluated how signal peptide analyses can provide further confidence on these identifications.

5.1.1 Achievements

As demonstrated throughout this work, database design can play an important role in peptide identification for validating predicted exons structures and assessing the quality of gene models. The presence of sequences on the database does not directly imply that the peptides identified by search engines will be assessed as significant. This can translate into identified peptides having been assigned low confidence score, hence been missed out from the final dataset even though correctly present in the database, which becomes more evident as the database size increases. As manual validation is still required to a certain degree, a comparative analysis, evaluating mutually exclusive PSMs in the dataset and their source database, would generate a short list of unique peptide sequence to be investigated. Additionally, databases designed tailored for specific peptide sequences, such as N-terminal peptides, can reduce both the database search space and the manual post-processing required.

With the work I performed in the second chapter I attempted to provide an additional filter to improve the quality of ORF_SS database. Although a six-frame translation can be biased towards false positives, due to the 5:1 ratio of FPs to TPs present in the sequence database, it is useful for organisms for which there is no accurate gene model. By using a comparison approach I have been able to identify ~400 peptide sequences that were uniquely matched on the ORF_SS database and not present on results from gene models for *N. caninum* release 5.1. By repeating these analyses, on a smaller set of spectra, on the following release 6.1, I have been able to shortlist 95 peptide sequences (from 24 proteins) that were later added to the gene models. These peptides, while being absent from the previous model (5.0), were successfully identified on the ORF_SS database (release 5.0). I used this approach, tested on *T. gondii*, to provide a novel method for assessing the quality of the gene models with proteomic data.

With the work described in the third chapter I presented a novel approach, for ISPs identification, for proteogenomics pipelines. Although the study proved that blind identification of ISPs can be made from proteomic data, overall the pipeline could not be considered 100% successful at this stage. However, even considering the future framework for the designed pipeline, the algorithm was able to successfully identify ISPs with 26% sensitivity at 5% fixed FDR. This result, achievable by including OMSSA in the final step in the pipeline, presented a framework that can be improved upon in the future. The search engine scoring algorithm was able to differentiate between false and true positive identifications. During the development of the designed pipeline I compared it against commonly used bioinformatics approaches such as *de novo* sequencing and database dependent methods, with alternative gene models. The results of this comparison showed the importance of not relying on one approach only: *de novo* sequencing can provide low confidence identifications with low coherence between the predictions; database approaches only rely on correctly predicted gene models, but these can be highly different due to algorithmic differences of the gene finder used.

In chapter four I demonstrated the importance of database design for targeting N-terminal peptide sequences. Combining different sequence databases I was able to identify with high confidence a set of ~300 N-terminal peptides (from ~1700 PSMs). Using the scripts I wrote for analysing signal peptides cleavage sites (position and amino acid frequency) I was able to shortlist ~150 N-terminal peptide sequences; these candidate peptides can provide evidence for correcting the predicted signal peptides and suggest novel predictions. Additionally with a further analysis of the identified N-terminal peptides I was able to assess the percentage of NME, within the found N-terminome for *T. gondii*, to be ~70%. Given the importance of NME as an irreversible co-translational modification present in various metabolic pathways, this finding could be further exploited for pharmaceutical drug delivery and targeting [294]. Studies on eukaryotic plants have identified that NME inhibition can affect key metabolic complexes within the cell life span [295, 296].

5.1.2 Shortcomings

As the algorithms I designed post-process datasets from various bioinformatic software packages, the time needed to fine-tune them and make them compatible with all different formats could be considered a downside. One of the challenges faced was the lack of system interoperability of the different software packages, as most of them are compiled to run exclusively on one OS (i.e. Windows, Unix). Although most of the software packages can be run on the same machine, this OS incompatibility can lead to slow performance as the analyses as cannot be run in parallel. Additionally not all publicly available bioinformatics tools are provided with conspicuous support; several have been published and made available but their support is not up to date (e.g. GeneZilla [29]). After testing various Unix systems, Biolinux [297], freely available to bioinformatic community, was found to be the best option for gene finding software packages as it is pre-loaded with compilers, libraries and several bioinformatics applications ready to use.

Careful attention needs to be paid when comparing the datasets obtained by querying different sequence databases. The uniqueness of peptide identified may be due to statistical algorithms and not to efficient database design.

The algorithms written for ISPs searches are still too slow and insufficiently accurate to be considered useful for large-scale analyses. Depending on the size of the genome and the computational resources available it may take a few hours to generate a fully indexed table for all short sequence tags, although it needs to be done only once. The time needed for processing each spectrum vary largely depending on the number of generated anchors (from ~2min to ~10min). Also during initial tests the purely *De novo* approach proved unsuccessful suggesting that high confidence *De novo* sequencing cannot be achieved alone. The results from ISP identifications are currently outperformed by the best gene finders, when accurate genomic sequence and annotations are available. However for some genomes, a purely *De novo* approach may be preferable using sequence alignments between different *De novo* algorithms.

5.1.3 Framework for future developments

The protocol for assessing gene models by direct comparison of PSMs against ORF_SS/ ORF_MS could be validly applied to other species with larger genomes. Additional filtering techniques, based on sample protocols, can be applied, when mining ORF_SS; as a practical example, short ORF_SS that do not meet minimum threshold are discarded during our analyses. However an additional filtering step could retain these short ORF_SS should they contain the same enzymatic cleavage sites used during sample preparation. This could help identifying short coding sequences that would otherwise not be in the final sequence database.

Producing alternative gene models with multiple gene finders could be automated including evaluation of genomic sequence and gene structures. For example by using proteomic identifications we can isolate a set of highly annotated genes to both train HMM models and evaluate the resulting predictions.

By continuously improving the annotation of the amino terminus of the proteins we can obtain supplementary data to increase confidence in signal peptide prediction. This information could be added during the design of the frayed database by analysing the entire genomic sequence looking for the motif of signal peptide cleavage sites. This would enable searches for the N-terminus of the protein within the predicted and unpredicted sequence.

Given the current limitations for ISPs blind searches, the core of the approach could be improved by limiting the search space. By using the official and panel of alternative gene models it would be possible to extract all alternative splices. This could be used to further test the results from the ISP pipeline to discriminate candidate ISPs identified.

These methods and approaches studies in this thesis could be, at later date, further improved and combined into one pipeline. Targeted sampling and whole cell fractionation analysed with tandem MS can be analysed on alternative gene models, ORF_SS, official gene models and specifically designed sequence databases, such as frayed database, also applicable to alternative gene models. The high confidence identifications could then be used to analyse the genome for improved alternative gene models and refined filtered ORF_SS to be queried again. As the process could require increased computational resources it would be ideal to place a given pipeline on a cloud computing resource. A simplified interface would additionally increase the usability and availability of the pipeline by researchers.

5.2 Proteomics: work in progress

5.2.1 Latest advances

In recent years shotgun proteomics has been increasingly recognised to generate large quantities of data. High-resolution mass spectrometers have contributed to the increase of the identification confidence. Novel protocols (enrichment/ depletion) have been devised to selectively extract peptide sequences from specific protein regions, to identify gene boundaries and PTMs. These improvements in data acquisition and the increased abundance of genomic sequences available have produced a requirement for notable advances in bioinformatic methods for predicting exon sequences, exon-exon structure (gene finders), increasing PSM confidence (search databases and posterior true positive estimates) while reducing false positive rates (e.g. through improved database designs). More studies used shotgun proteomics to: validate models of protein-coding genes; suggest evidence for completely novel genes and assess protein-coding genes from transcripts [205, 298]. Innovations in transcriptomics have enabled researchers to generate higher quality transcriptomes, used both at gene curation level and database searches with shotgun data.

5.2.2 Future goals and challenges

Although fairly recent, proteogenomics has already seen a number of successful strategies for evaluation and implementation by several groups [129, 157, 202, 208-210, 299-305]. The principal goal of proteogenomics is to effectively use shotgun proteomic datasets for constructing and validating the models of protein-coding genes. However using proteomic data to provide identification on a large scale, such as entire peptidomes, is still a considerable challenge. Low abundance peptides, unexpected PTMs and splice sites represent a few of the limitations in peptide identification by bioinformatics strategies. Recent bioinformatic resources have been implemented to enable even small research groups to perform analyses with high computational power through cloud computing [306].

Combining together several approaches can increase the confidence of identified peptides and expand the final dataset; however different challenges are presented due to the resulting database size. Accurate design of the algorithms is needed to increase the software performance and reduce the error rate. New protocols to correctly assign with confidence the identified peptides to the correct genomic structure are currently being developed by several groups. To facilitate the reproducibility of the experiments and posterior analyses, the bioinformatic community is headed towards the adoption of standard data formats. However there are still a number of software packages that are not entirely supported or up to date, making it hard to combine multiple algorithms.

Future bioinformatic approaches can prove invaluable to generate *in silico* peptide predictions that biologists could attempt to explain through targeted proteomic sampling. Instead for bioinformatics to provide methods to understand biological data, it should start to lead by providing theories that proteomic analyses could prove or discard. Additional challenges include external cellular factors and specific life stages that influence gene expression and the low dynamic range of proteins analysed in the experiments often prevent the detection of low abundance proteins. Similar to bioinformatics, the best approach might rely in combining the results from different sampling protocols and techniques (i.e. enzymes, selective enrichment of specific peptides, separation techniques) to provide complementary data and help overcome difficulties in detecting a small portion of the peptidome.

Appendix: A

6 Original ISP pipeline design

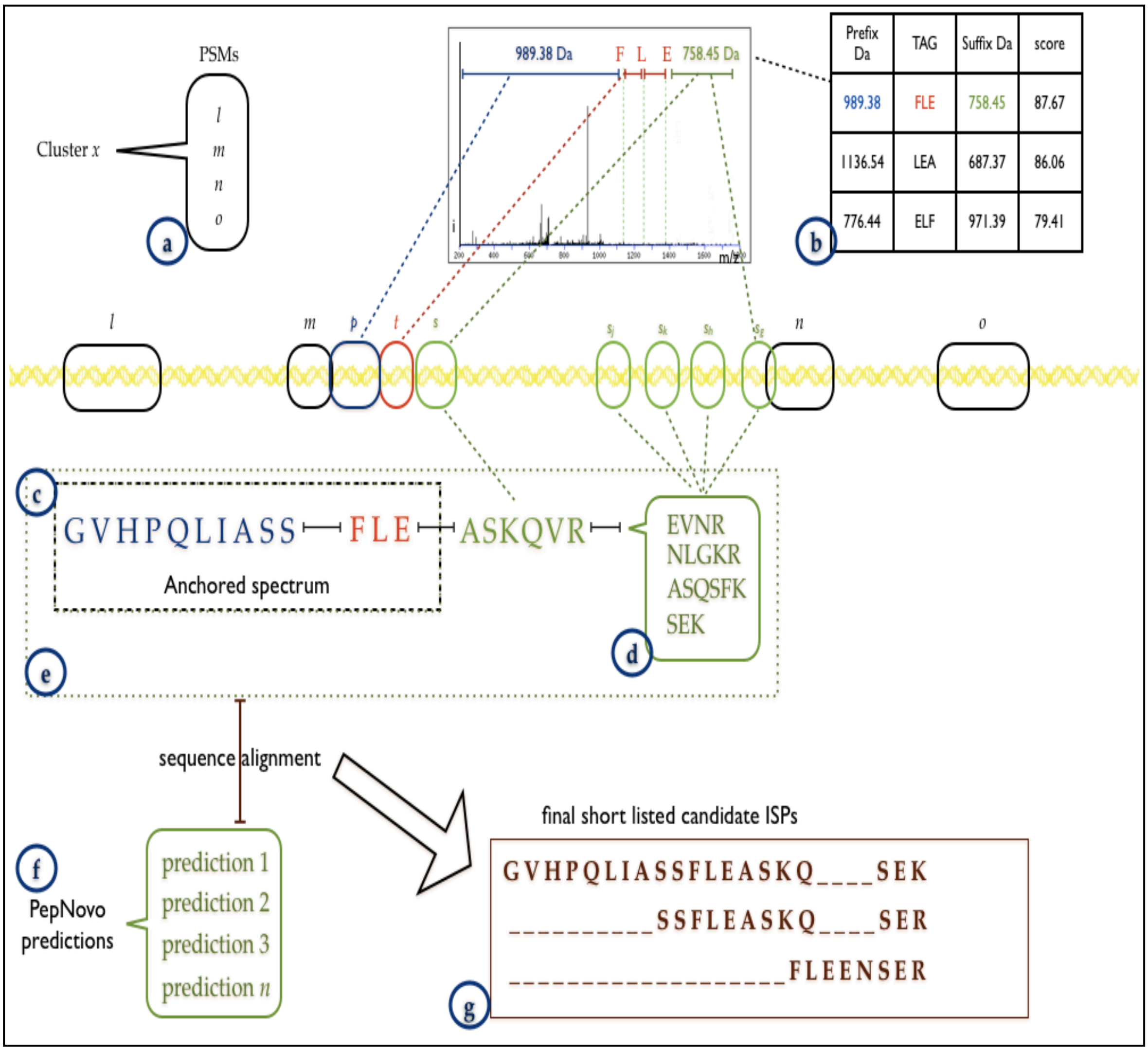
In this appendix I present the first design developed for intron spanning peptides using only *de novo* approach. This was initially devised as an approach to be used by curators of newly sequenced genome with no gene model available. The first stage of the pipeline is based on database searches against the whole genome to select the genomic regions where identifications have been made and select the mass spectra that have yet to be identified. Then pipeline attempts to identify the tandem mass spectra not identified through sequence database searches by anchoring their predicted TAG with InSPecT. The full peptide reconstruction is generated in the same way it is explained in chapter 3. However to shortlist the intron spanning peptide candidate is made by aligning the full length peptide predictions against each of the *de novo* MS/MS interpretations generated by PepNovo. The score of the alignment would ideally allow identifying the candidate ISP while the gap in the alignment would indicate the splice site.

6.1 Methods

6.1.1 Original pipeline design: *De Novo* sequencing and sequence alignments for full-length peptide identification

The original design of the pipeline included sequence alignments with *De Novo* interpretations (PepNovo) as a method to highlight final full-length peptide candidates as intron spanning peptides. The theory was that a peptide sequence that spans across 2 exons would result in a gapped sequence alignment where ideally only one gap, of variable length, would be present. The theory was that the anchored spectrum peptide would be concatenated with an adjacent sequence of residues (with mass within that of the unexplained terminus from

InSPecT tag predictions). This was used to generate a list of candidate peptides, whose mass would have been larger than the spectrum mass, by concatenating the missing terminus obtained from nearby region. The final stage of the pipeline would have seen each of the PepNovo interpretations (up to 20) being aligned against each long peptide sequence from the generated list. By keeping track of the score of the alignment it would have been possible to shortlist only few candidate intron spanning peptide with splice position highlighted by a gap in the alignment (Figure 6:1).



Prefix Da	TAG	Suffix Da	score
989.38	FLE	758.45	87.67
1136.54	LEA	687.37	86.06
776.44	ELF	971.39	79.41

Figure 6:1 As in the previous design (figure 3:4) the sequence TAG with its data (b) is used to anchor the spectrum (c) within genomic regions and positional details from the given cluster (a). The unexplained terminus is extracted from sequence adjacent to the anchor and concatenated to extracted termini from nearby proteolytic sites (d). This generates a list of

peptide reconstructions whose mass is larger than the analysed spectrum mass. Each reconstructed peptide from the list is aligned against each PepNovo prediction (f) and the cumulative score used to shortlist the final candidate intron spanning peptide with the gap information (g).

During the initial testing using the sequence alignment algorithm, each candidate sequence from tryptic sites was simply concatenated to the anchored sequence. It was assumed that given accurate *De novo* predicted peptide sequences, the sequence alignment algorithm would have been able to find the peptide sequence yielding the analysed spectrum, as well as highlighting the splice site.

Finally by looping through the list of sequences from tryptic sites, each candidate sequence was concatenated with the elongated anchored sequence. The approach for this last stage is to recreate all possible splice sites between the DNA sequences prior concatenating them for translation to amino acids.

6.1.2 Original pipeline design: performance evaluation of PepNovo

The *De novo* sequencing software PepNovo relies on a training set generated using ion fragmentation rules in order to assess the probability of each candidate peptide yielding a given spectrum. The only training set currently available in the PepNovo algorithm is CID [160, 161, 166], which is used as a default. Although it is possible to generate new training sets based on different instrumentation, this task would require hundreds of thousands of high accuracy interpreted spectra.

PepNovo analyses used the following parameters: peptide tolerance of 2.5 Da and fragment ion tolerance of 0.5 Da, both set in the only available model (CID-IT-TRYP), fixed modification: carbamidomethyl (C) and variable modification: methionine oxidation (M). In this case the presence or absence of PTM during the evaluation would not affect the pipeline strategy as the numeric values of residue modifications in PepNovo peptide sequences (i.e. M+16QQASTEQQAGEQK indicating methionine oxidized) are removed. Also these modifications tend to appear once or twice within a PepNovo set of interpretation (~20 per spectrum) and the alignment score, from sequence alignment between interpreted peptides sequence and ISP candidates, was the sum of all individual alignments.

PepNovo performance was tested by producing predictions of different length, from a minimum 3 residues to the maximum length that could be interpreted. Not surprisingly the shorter predicted sequence had higher sensitivity, specificity and their similarity to InSPecT sequence TAGs become evident thus defeating the purpose of the *De Novo* sequencing approach being included in the pipeline for sequence identifications. It appears that the longer the predicted sequence, the lower the accuracy of amino acid residues predicted [94, 125, 126, 163, 167, 169].

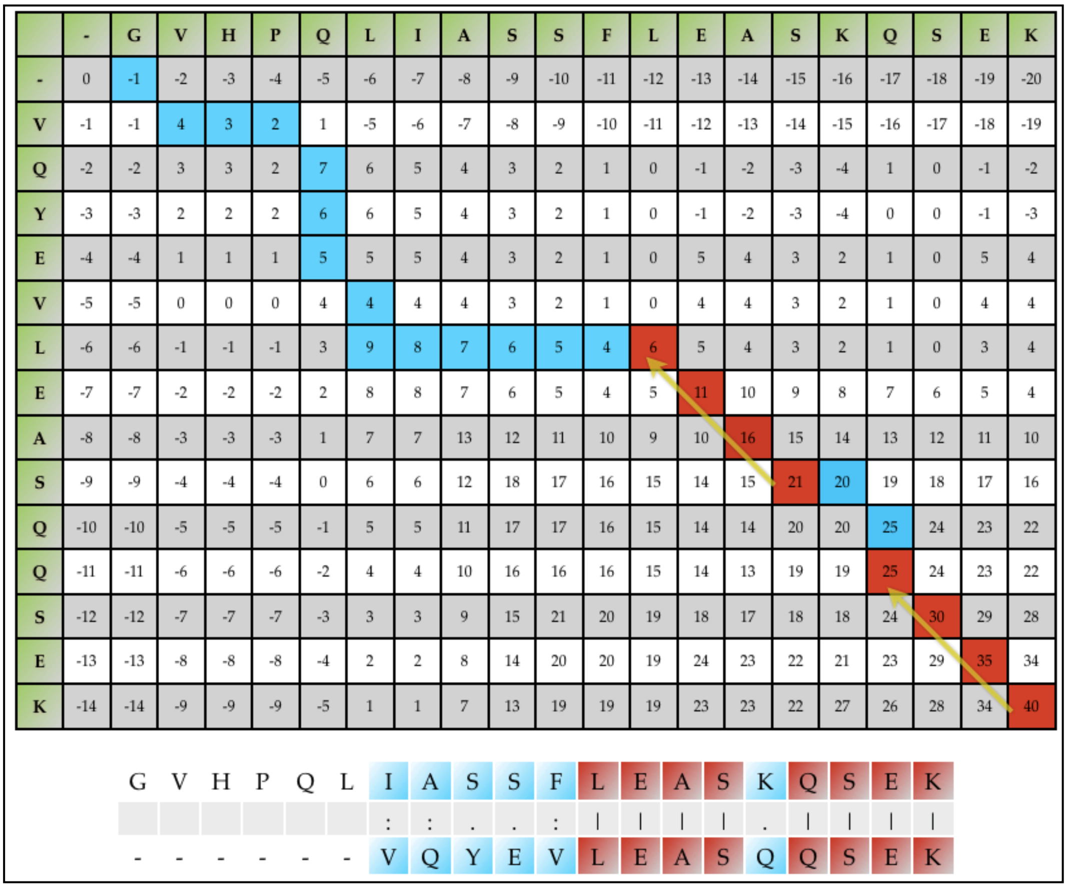
6.1.3 Original pipeline design: Sequence alignment

Sequence alignment algorithms evaluate the homology of two distinct sequences by generating all possible alignments and assessing the score of each of them to find the optimal alignment. Three main algorithms can be used depending on the type of sequence and alignment:

Global alignment (based on Needleman-Wunsch algorithm [307-309], which aligns two sequences completely minimizing the edit distance defined by gap penalty and substitution matrix; this method is generally applicable to sequences similar in length and pattern.

Local alignment (based on Smith-Waterman algorithm [47]) which attempts to find sub-sequences that have minimal distance among all sub-sequences; a method generally applicable to sequences that have similar sub-sequences.

Ends-free sequence alignment is a particular case of the global alignment [310]. This method assign no gap penalty to gap extending to the beginning or end of the sequence; commonly applied where one of the sequences is contained in the other.



	-	G	V	H	P	Q	L	I	A	S	S	F	L	E	A	S	K	Q	S	E	K
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19	-20
V	-1	-1	4	3	2	1	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19
Q	-2	-2	3	3	2	7	6	5	4	3	2	1	0	-1	-2	-3	-4	1	0	-1	-2
Y	-3	-3	2	2	2	6	6	5	4	3	2	1	0	-1	-2	-3	-4	0	0	-1	-3
E	-4	-4	1	1	1	5	5	5	4	3	2	1	0	5	4	3	2	1	0	5	4
V	-5	-5	0	0	0	4	4	4	4	3	2	1	0	4	4	3	2	1	0	4	4
L	-6	-6	-1	-1	-1	3	9	8	7	6	5	4	6	5	4	3	2	1	0	3	4
E	-7	-7	-2	-2	-2	2	8	8	7	6	5	4	5	11	10	9	8	7	6	5	4
A	-8	-8	-3	-3	-3	1	7	7	13	12	11	10	9	10	16	15	14	13	12	11	10
S	-9	-9	-4	-4	-4	0	6	6	12	18	17	16	15	14	15	21	20	19	18	17	16
Q	-10	-10	-5	-5	-5	-1	5	5	11	17	17	16	15	14	14	20	20	25	24	23	22
Q	-11	-11	-6	-6	-6	-2	4	4	10	16	16	16	15	14	13	19	19	25	24	23	22
S	-12	-12	-7	-7	-7	-3	3	3	9	15	21	20	19	18	17	18	18	24	30	29	28
E	-13	-13	-8	-8	-8	-4	2	2	8	14	20	20	19	24	23	22	21	23	29	35	34
K	-14	-14	-9	-9	-9	-5	1	1	7	13	19	19	19	23	23	22	27	26	28	34	40

```
G V H P Q L I A S S F L E A S K Q S E K
            : : . . : | | | | . | | | |
- - - - - - V Q Y E V L E A S Q Q S E K
```

Table 6:1 Table shows the constructed matrix alignment between intron spanning peptide GVHPQLIASSFLEASKQSEK and best *De Novo* prediction from PepNovo VQYEVLEASQQSEK. The scoring algorithm is based on Global alignment (Needleman-Wunsch[309]) with match value set a 5, while mismatch and gap cost are set at -1. The matrix is constructed by evaluating each cell, and setting its value depending on diagonal left, left and top cell values. The match value is implemented as (diagonal cell + match value), (top cell – gap cost), (left cell – gap cost); the higher value between these three is chosen for each cell evaluated. To assign the residues to the sequence aligned (highlighted in red), the final sequence is constructed by backtracking from the last bottom right cell in a way to find the path with highest score. Column and semi-column symbols (highlighted in blue) and hyphen describe whether conserved or semi-conserved substitution or gap has been observed.

The algorithm assesses the score by aligning one letter from one sequence to the other and adding points or penalties for matches, gaps, and gap length (Table 6:1). The algorithm of choice implemented in the initial pipeline was based on the local alignment as ISPs were expected to have internal gap caused by splice site. For this the scoring system was adjusted not to penalize one gap in the sequence with no concern to its length, while deeply penalize the presence of more than one gap, even if very short.

The algorithm for processing and scoring sequence alignments was built to align each PepNovo prediction (~ 19 for each spectrum) against all candidate ISPs generated from the same spectrum. The final score for a candidate ISP was given by the sum of all scores across the PepNovo prediction. In order to provide additional information on the sequence alignment consensus a designed algorithm would score each residue position across all alignments. This would provide visual insight showing how the residues of the candidate sequence aligned with *De novo* predicted sequences. This part of the strategy was later replaced by algorithm described in the previous paragraphs, since it relied too much on accuracy of *De novo* sequencing.

6.2 Results

6.2.1 Original pipeline design:

From the test dataset a small subset of only 24 spectra yielding intron spanning peptides from 20 different proteins was created. This small set was put through the pipeline with genomic regions already restricted to the proteins of interest. The sequence TAGs generated by InSPecT were to be mapped within much smaller area than the final pipeline design. This was specific to test sequence alignment with *De novo* predictions, as the final shortlist of ISP comprised only 6 peptide sequences, per spectrum, ranking highest after all sequence alignments.

From the initial 24 spectra processed through the pipeline, for only 11 spectra the ISP was included within the shortlisted of candidate peptides, although in not one of them the splice site was correctly highlighted in the sequence alignment result (Figure 6:2, Table 6:2). Although the ISP peptide was correctly constructed and present in the list of candidates for 23 spectra, only 10 full-length peptide sequences were correctly identified after sequence alignments. However for a number of cases simple sequence alignments between ISP candidates and *De novo* predictions was not optimal for correct separation between real and false matches, without a manual evaluation. The quality of the sequence tags (from InSPecT) influenced the number of full-length ISP candidates while the *De novo* (PepNovo) predictions for a given spectrum did not always proved to be coherent with each other. In one case InSPecT wrongly assigned the spectrum charge resulting to only partial peptide identification. Additionally an incorrect sequence tag generated a false result, not distinguishable from real peptide sequences by automated algorithms.

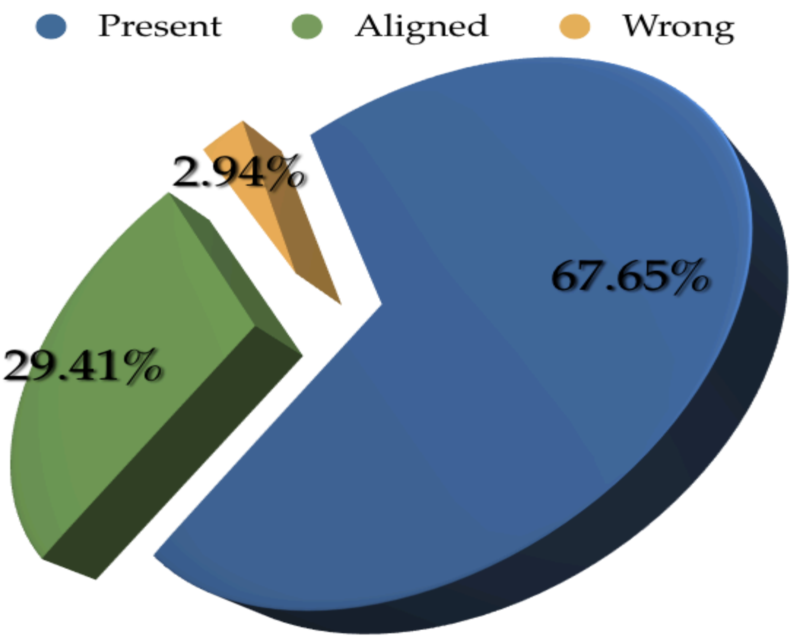
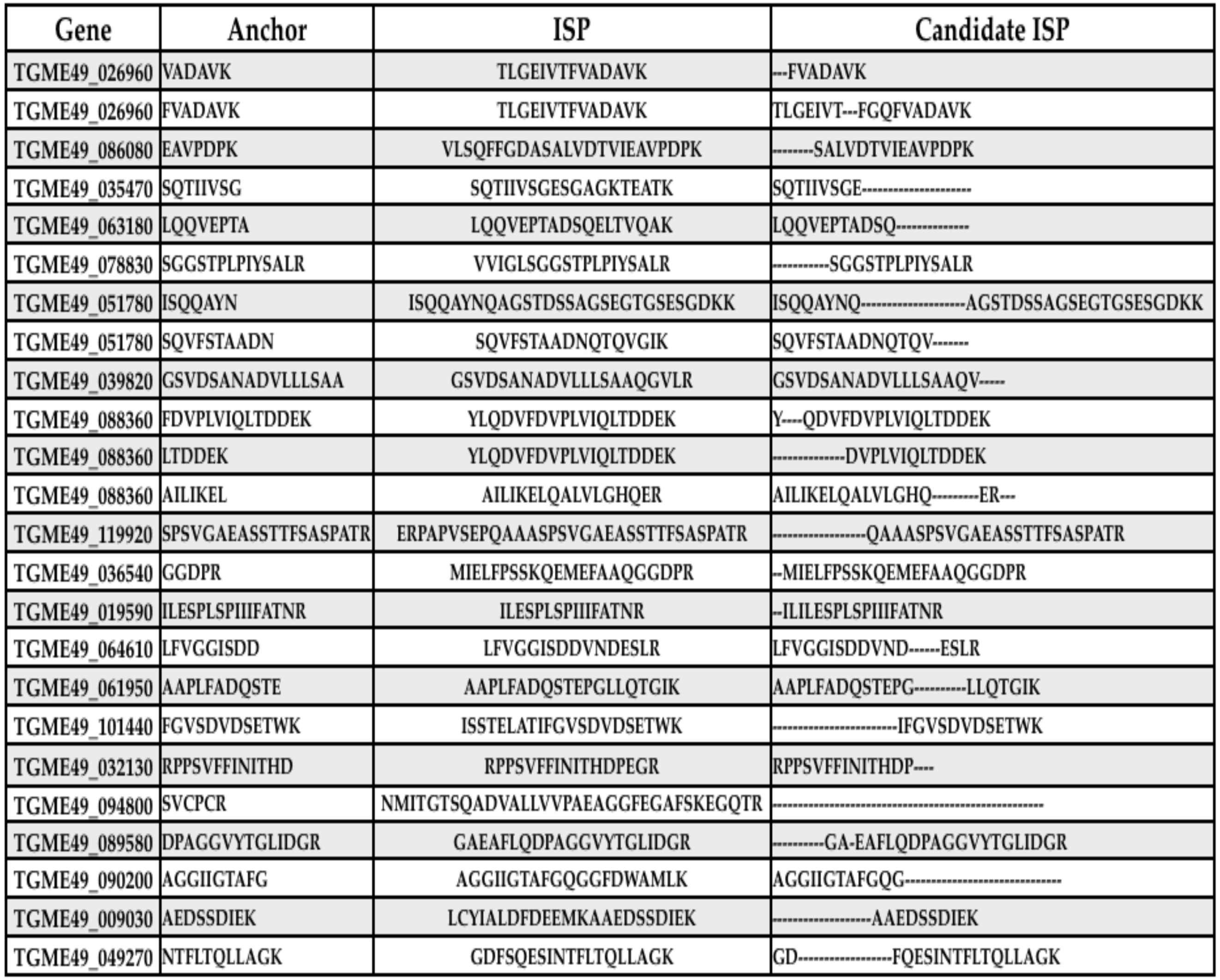


Figure 6:2 From a small set of 24 ISPs, only a sub-set of the reconstructed full-length peptides were identified with sequence alignments. For a large majority the sequence, present in the temporary database, did not make it through the final selection. For a small minority the spectrum was improperly located on the genome due to an incorrect sequence TAG. Virtually in every positive result, seeing the correct candidate being shortlisted, the sequence alignment stage did not enable the correct identification of the splice site or to appropriately discriminate between false and true positive.

A shortcoming of this approach was caused by PepNovo lack of consensus between its interpretations; in general the set of predicted peptides from a spectrum comprised large variation in amino acid composition of the peptides. Even in cases where PepNovo predicted the partial sequence correctly, if the majority of all reconstructions is not correct then sequence alignments would generate a bias towards FP. Similarly, due to low quantity of high confidence predictions of reasonable length, sequence alignment did not prove successful in highlighting the splice positions within the candidate ISPs. The gap and mismatched from the alignments were inclined on either terminal of the peptide.

Although the initial test was not able to correctly identify full-length peptides, it verified the ability to anchor the majority of spectra. For this reason a further pipeline test involved algorithms to automate the selection of the genomic regions of interest. The dataset for this test comprised only 12 spectra and the range for genomic region discovery was set to 5000 bps. Most of the spectra could not be correctly anchored and from the list of ISP candidates, generated with correct anchors, the sequence alignment stage did not prove successful. Successive algorithms replaced the stage involving *De novo* predictions and sequence alignment for splice site generation and OMSSA searches.



Gene	Anchor	ISP	Candidate ISP
TGME49_026960	VADAVK	TLGEIVTFVADAVK	---FVADAVK
TGME49_026960	FVADAVK	TLGEIVTFVADAVK	TLGEIVT---FGQFVADAVK
TGME49_086080	EAVPDPK	VLSQFFGDASALVDTVIEAVPDPK	---------SALVDTVIEAVPDPK
TGME49_035470	SQTIIVSG	SQTIIVSGESGAGKTEATK	SQTIIVSGE------------------------
TGME49_063180	LQQVEPTA	LQQVEPTADSQELTVQAK	LQQVEPTADSQ----------------
TGME49_078830	SGGSTPLPIYSALR	VVIGLSGGSTPLPIYSALR	------------SGGSTPLPIYSALR
TGME49_051780	ISQQAYN	ISQQAYNQAGSTDSSAGSEGTGSESGDKK	ISQQAYNQ-----------------------AGSTDSSAGSEGTGSESGDKK
TGME49_051780	SQVFSTAADN	SQVFSTAADNQTQVGIK	SQVFSTAADNQTQV--------
TGME49_039820	GSVDSANADVLLLSAA	GSVDSANADVLLLSAAQGVLR	GSVDSANADVLLLSAAQV------
TGME49_088360	FDVPLVIQLTDDEK	YLQDVFDVPLVIQLTDDEK	Y-----QDVFDVPLVIQLTDDEK
TGME49_088360	LTDDEK	YLQDVFDVPLVIQLTDDEK	----------------DVPLVIQLTDDEK
TGME49_088360	AILIKEL	AILIKELQALVLGHQER	AILIKELQALVLGHQ----------ER---
TGME49_119920	SPSVGAEASSTTFSASPATR	ERPAPVSEPQAAASPSVGAEASSTTFSASPATR	--------------------QAAASPSVGAEASSTTFSASPATR
TGME49_036540	GGDPR	MIELFPSSKQEMEFAAQGGDPR	--MIELFPSSKQEMEFAAQGGDPR
TGME49_019590	ILESPLSPIIIFATNR	ILESPLSPIIIFATNR	--ILILESPLSPIIIFATNR
TGME49_064610	LFVGGISDD	LFVGGISDDVNDESLR	LFVGGISDDVND------ESLR
TGME49_061950	AAPLFADQSTE	AAPLFADQSTEPGLLQTGIK	AAPLFADQSTEPG----------LLQTGIK
TGME49_101440	FGVSDVDSETWK	ISSTELATIFGVSDVDSETWK	---------------------------IFGVSDVDSETWK
TGME49_032130	RPPSVFFINITHD	RPPSVFFINITHDPEGR	RPPSVFFINITHDP----
TGME49_094800	SVCPCR	NMITGTSQADVALLVVPAEAGGFEGAFSKEGQTR	--
TGME49_089580	DPAGGVYTGLIDGR	GAEAFLQDPAGGVYTGLIDGR	----------GA-EAFLQDPAGGVYTGLIDGR
TGME49_090200	AGGIIGTAFG	AGGIIGTAFGQGGFDWAMLK	AGGIIGTAFGQG----------------------------------
TGME49_009030	AEDSSDIEK	LCYIALDFDEEMKAAEDSSDIEK	---------------------AAEDSSDIEK
TGME49_049270	NTFLTQLLAGK	GDFSQESINTFLTQLLAGK	GD--------------------FQESINTFLTQLLAGK

Table 6:2 The table lists identified ISPs with pipeline genomic regions selected based on gene coordinates and sequence alignments with PepNovo predictions. One peptide could not be mapped on the genome due to incorrect TAG; only 11 peptides were finally shortlisted after alignments due to a lack of coherence in PepNovo predictions.

7 Appendix: B

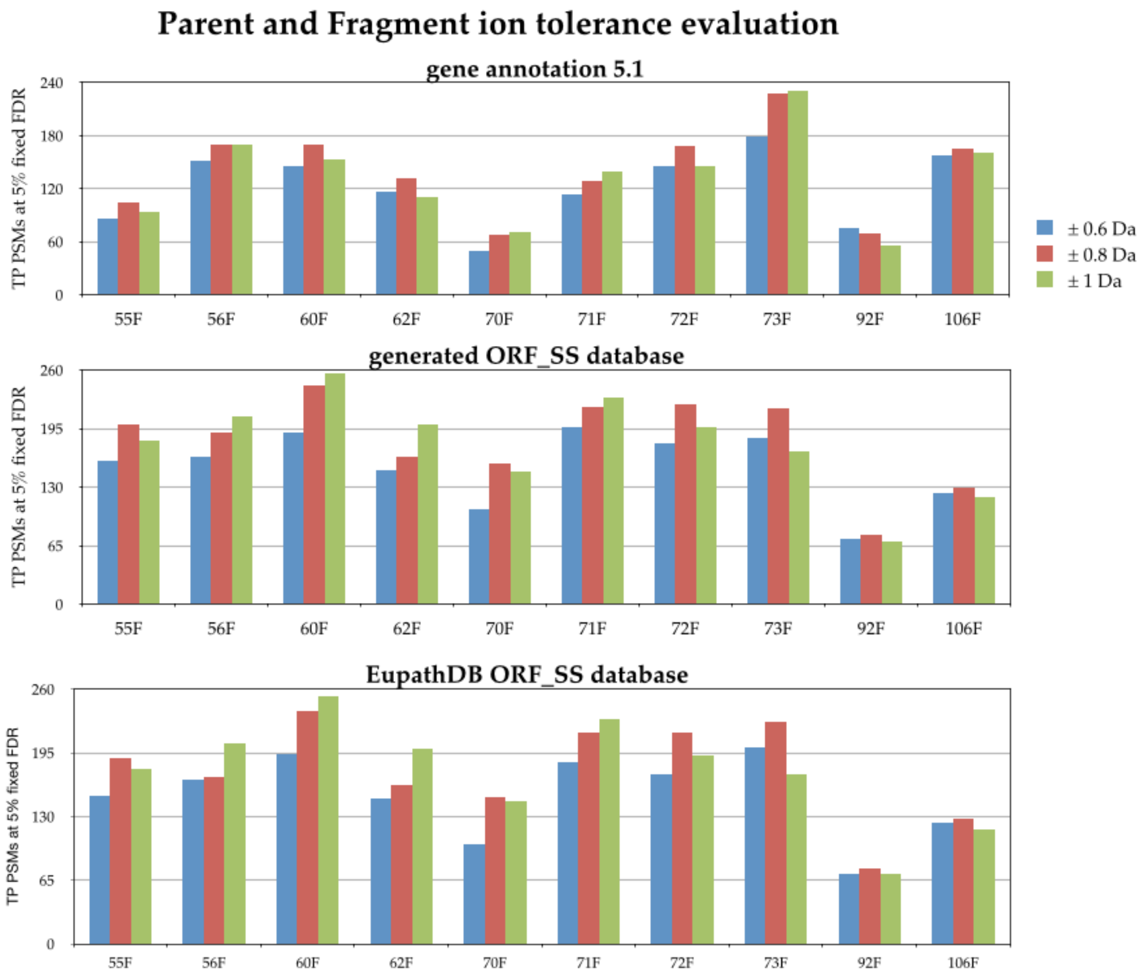


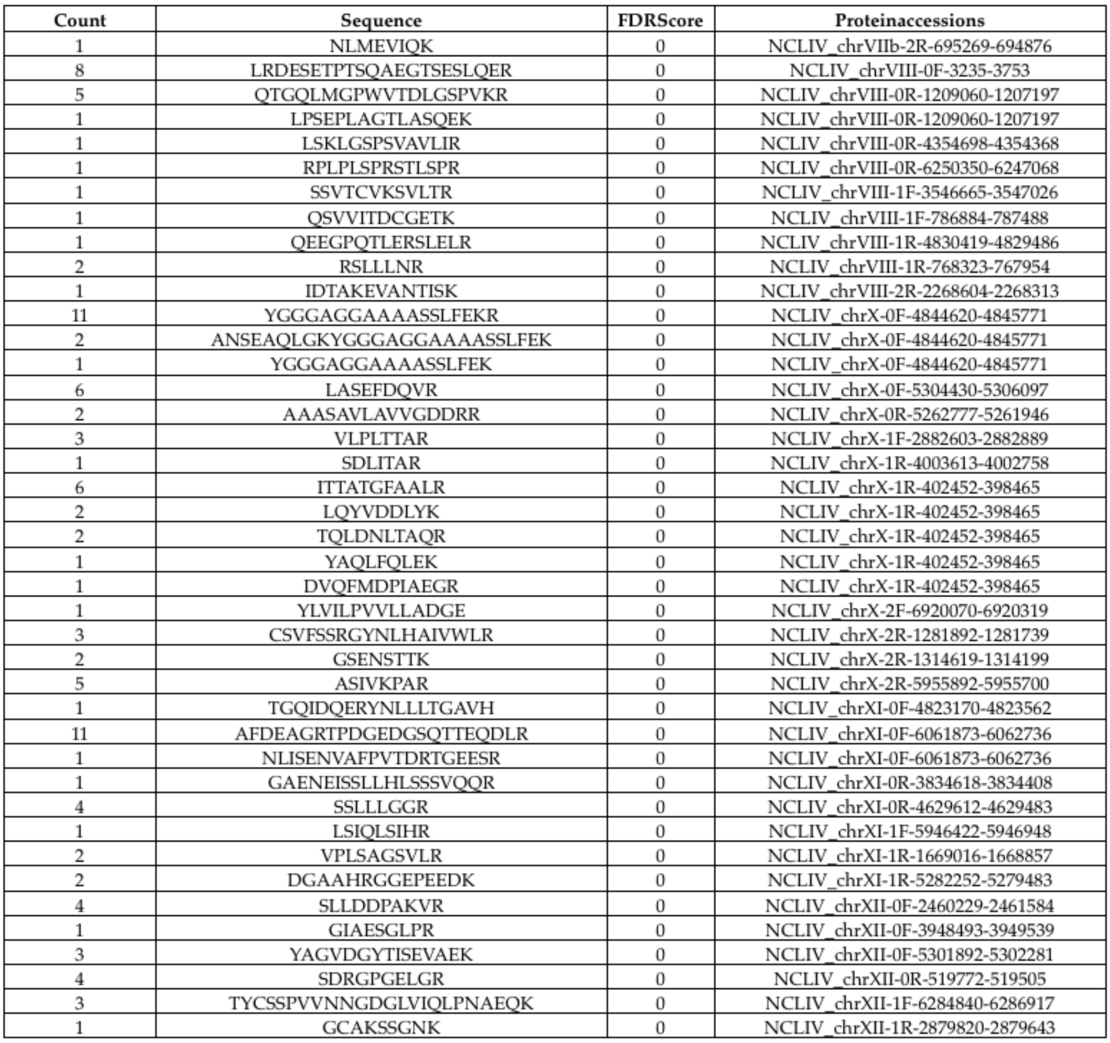
Figure 7:1 10 1-DE gel slice from *N. caninum* were used to query official ORF sequence database available on EupathDB (longer than 50 amino acids). The output was rescored at 5% fixed FDR. Parent and fragment ion tolerance both set at ±0.6, ±0.8 and ±1 Da.

Count	Sequence	FDRScore	Proteinaccessions
3	RLPTAGSFER	0	NCLIV_chrIa-0R-1462260-1461285
1	QTISLGYVNLR	0	NCLIV_chrIa-1F-1004390-1005096
1	SIAATGPTREEAPR	0	NCLIV_chrIa-1R-1318115-1317869
1	AQEGADLLSVAENLEQR	0	NCLIV_chrIa-1R-1857302-1856555
15	GIPAIQKLCR	0	NCLIV_chrIa-1R-627989-627749
3	VHNSVPFFAKR	0	NCLIV_chrIa-2R-1305757-1305391
4	GLQAGLTR	0	NCLIV_chrIb-0R-56639-55937
3	ARQEAMNVNIELSR	0	NCLIV_chrIb-0R-56639-55937
2	YQMETAQELMNR	0	NCLIV_chrIb-0R-56639-55937
1	ALGEYLEQVR	0	NCLIV_chrIb-0R-56639-55937
2	GLAAPLPATLADVYATVFAAMPAK	0	NCLIV_chrIb-1F-1578734-1579203
2	NLQEGLLPVSLR	0	NCLIV_chrIb-1F-1579592-1579932
3	DTSVLQFNQFFTNILK	0	NCLIV_chrIb-1F-1580564-1581000
1	QSPAEYQTVSGTEVIAPLFEGEGK	0	NCLIV_chrIb-1F-1581293-1581624
2	FPSLLLRR	0	NCLIV_chrIb-2F-980304-980607
1	SLFSASPSPSLPASGNLFVVK	0	NCLIV_chrIb-2R-1592913-1591353
2	RAGSSLLPEAGLNCACTR	0	NCLIV_chrII-0F-887134-887601
1	HLSERVER	0	NCLIV_chrII-1R-153517-153058
1	QGAGAASQLPGLPRTSIGK	0	NCLIV_chrII-2F-1718175-1718292
1	VAVGVCAMALFVLR	0	NCLIV_chrII-2R-1118943-1118670
2	TAISDIEVDVEEIDKPK	0	NCLIV_chrIII-0F-682288-682689
1	VCVLCVPVCLEPPQK	0	NCLIV_chrIII-1F-250070-250677
1	AHIHVDVSGANGLGILMIDDTK	0	NCLIV_chrIII-1F-744794-744972
3	LIAELQDSDPASTQGATSDHLTPLK	0	NCLIV_chrIII-1R-192206-191615
2	IYFINLGGSSR	0	NCLIV_chrIII-1R-192206-191615
1	IGVAQHTAADMETLIQQISTIDEENETR	0	NCLIV_chrIII-1R-192206-191615
1	LASSGLWQVRPWPR	0	NCLIV_chrIV-0F-2129371-2129799
2	STSQDDESASHDEPSAATLEAVEK	0	NCLIV_chrIV-0F-2229826-2230416
5	RLIENLTK	0	NCLIV_chrIV-0R-1074117-1073919
1	ISFAFLR	0	NCLIV_chrIV-0R-1593930-1593273
1	DCFVLVLQELR	0	NCLIV_chrIV-0R-2287218-2286861
4	GYLGTLSR	0	NCLIV_chrIX-0F-2869078-2869680
1	NTASSLIASLR	0	NCLIV_chrIX-0F-3114973-3115335
2	EMDTTALR	0	NCLIV_chrIX-0F-3465289-3465567
1	ALCASLQAVSAPDDLAFFR	0	NCLIV_chrIX-0F-578719-579141
1	SSVAVVLRKPRPSQER	0	NCLIV_chrIX-0R-11751-11406
3	SSFPKFEIR	0	NCLIV_chrIX-0R-2168865-2168328
1	FPLQLIK	0	NCLIV_chrIX-0R-4342635-4342272
1	SRFSVEK	0	NCLIV_chrIX-0R-5463819-5463513
29	SSGTGTSRMNR	0	NCLIV_chrIX-1F-2987768-2988165

Table 7:1 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

Count	Sequence	FDRScore	Proteinaccessions
1	NTLDLILR	0	NCLIV_chrIX-1R-4190240-4189958
4	LLSLLVSLSR	0	NCLIV_chrIX-1R-5263157-5262485
1	ELDGSALLSEFVNKEEPLLLAPK	0	NCLIV_chrIX-2F-4487589-4488174
1	NVQFPYNVAQR	0	NCLIV_chrV-0R-1655615-1655441
1	ACPLQPTMARK	0	NCLIV_chrV-0R-757268-757139
1	CGWSVQLTRNAK	0	NCLIV_chrV-1R-662635-662455
1	GQEGGGDQPR	0	NCLIV_chrV-2R-739398-739224
1	QTCLLNK	0	NCLIV_chrVI-0F-1961257-1961571
1	GVAVAAEGPCLGTR	0	NCLIV_chrVI-0R-2577966-2577402
2	AALQVLKTGGR	0	NCLIV_chrVI-0R-3013533-3013029
1	ERVDVVGR	0	NCLIV_chrVI-1R-1384691-1384520
1	NFLGTPIYLQMPSR	0	NCLIV_chrVI-1R-871745-871388
1	AAGISEGPR	0	NCLIV_chrVI-2R-2762197-2761831
1	CIQPYIGIGER	0	NCLIV_chrVI-2R-353980-353617
1	QTFTSGPVPLPDVEEDNNAGAAPAASQK	0	NCLIV_chrVIIa-0R-1715421-1714911
1	QTFTSGPVPLPDVEEDNNAGAAPAASQKK	0	NCLIV_chrVIIa-0R-1715421-1714911
2	NALSFYDTR	0	NCLIV_chrVIIa-0R-2610144-2609898
2	TGFFSVK	0	NCLIV_chrVIIa-0R-265902-265737
1	RSPHFGDR	0	NCLIV_chrVIIa-0R-3467286-3467148
1	EDSFGSFR	0	NCLIV_chrVIIa-1F-513731-513885
1	APAADTPR	0	NCLIV_chrVIIa-1R-2115437-2115254
1	FRPLNLLPRHLLR	0	NCLIV_chrVIIa-2F-2396760-2396967
2	RVNAVMVR	0	NCLIV_chrVIIa-2R-24850-24688
1	AWTLSAVSTSHAR	0	NCLIV_chrVIIa-2R-2545093-2544712
1	SPSAPNETAPQDPSR	0	NCLIV_chrVIIa-2R-3889594-3888910
1	VFREGFWASTK	0	NCLIV_chrVIIa-2R-3895006-3894505
1	LPVIVVQR	0	NCLIV_chrVIIb-0F-3450475-3451740
11	GGSTHAALVSGDSPATVHCRNER	0	NCLIV_chrVIIb-0F-921832-922668
2	GVGGYGMATTGVLGTILALYSGAK	0	NCLIV_chrVIIb-0R-2958288-2957526
1	IGDDLEMSMR	0	NCLIV_chrVIIb-0R-3290883-3290535
3	GPVYFQEAQPVYLPSVVPR	0	NCLIV_chrVIIb-1F-2063726-2064363
3	SPLVSGPVAGVYPVVYESHDEDLLDQR	0	NCLIV_chrVIIb-1F-2063726-2064363
3	SVAYAGGPAFPLPLR	0	NCLIV_chrVIIb-1F-2063726-2064363
10	LVLIGDSGVGK	0	NCLIV_chrVIIb-1F-3389573-3389715
1	RILTNR	0	NCLIV_chrVIIb-1R-237164-236552
2	YAGQDVTSGSSHDTQSVGGR	0	NCLIV_chrVIIb-1R-2962157-2960684
1	QMVSGGGPQVSALR	0	NCLIV_chrVIIb-1R-745694-745454
2	NPPINMDVVVMPR	0	NCLIV_chrVIIb-2F-2061702-2062476
2	VVTDAQGVPLYIR	0	NCLIV_chrVIIb-2F-4190988-4191468
2	MSTPSYASASK	0	NCLIV_chrVIIb-2R-4122823-4122535

Table 7:2 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10



Count	Sequence	FDRScore	Proteinaccessions
1	NLMEVIQK	0	NCLIV_chrVIIb-2R-695269-694876
8	LRDESETPTSQAEGTSESLQER	0	NCLIV_chrVIII-0F-3235-3753
5	QTGQLMGPWVTDLGSPVKR	0	NCLIV_chrVIII-0R-1209060-1207197
1	LPSEPLAGTLASQEK	0	NCLIV_chrVIII-0R-1209060-1207197
1	LSKLGSPSVAVLIR	0	NCLIV_chrVIII-0R-4354698-4354368
1	RPLPLSPRSTLSPR	0	NCLIV_chrVIII-0R-6250350-6247068
1	SSVTCVKSVLTR	0	NCLIV_chrVIII-1F-3546665-3547026
1	QSVVITDCGETK	0	NCLIV_chrVIII-1F-786884-787488
1	QEEGPQTLERSLELR	0	NCLIV_chrVIII-1R-4830419-4829486
2	RSLLLNR	0	NCLIV_chrVIII-1R-768323-767954
1	IDTAKEVANTISK	0	NCLIV_chrVIII-2R-2268604-2268313
11	YGGGAGGAAAASSLFEKR	0	NCLIV_chrX-0F-4844620-4845771
2	ANSEAQLGKYGGGAGGAAAASSLFEK	0	NCLIV_chrX-0F-4844620-4845771
1	YGGGAGGAAAASSLFEK	0	NCLIV_chrX-0F-4844620-4845771
6	LASEFDQVR	0	NCLIV_chrX-0F-5304430-5306097
2	AAASAVLAVVGDDRR	0	NCLIV_chrX-0R-5262777-5261946
3	VLPLTTAR	0	NCLIV_chrX-1F-2882603-2882889
1	SDLITAR	0	NCLIV_chrX-1R-4003613-4002758
6	ITTATGFAALR	0	NCLIV_chrX-1R-402452-398465
2	LQYVDDLYK	0	NCLIV_chrX-1R-402452-398465
2	TQLDNLTAQR	0	NCLIV_chrX-1R-402452-398465
1	YAQLFQLEK	0	NCLIV_chrX-1R-402452-398465
1	DVQFMDPIAEGR	0	NCLIV_chrX-1R-402452-398465
1	YLVILPVVLLADGE	0	NCLIV_chrX-2F-6920070-6920319
3	CSVFSSRGYNLHAIVWLR	0	NCLIV_chrX-2R-1281892-1281739
2	GSENSTTK	0	NCLIV_chrX-2R-1314619-1314199
5	ASIVKPAR	0	NCLIV_chrX-2R-5955892-5955700
1	TGQIDQERYNLLLTGAVH	0	NCLIV_chrXI-0F-4823170-4823562
11	AFDEAGRTPDGEDGSQTTEQDLR	0	NCLIV_chrXI-0F-6061873-6062736
1	NLISENVAFPVTDRTGEESR	0	NCLIV_chrXI-0F-6061873-6062736
1	GAENEISSLLHLSSSVQQR	0	NCLIV_chrXI-0R-3834618-3834408
4	SSLLLGGR	0	NCLIV_chrXI-0R-4629612-4629483
1	LSIQLSIHR	0	NCLIV_chrXI-1F-5946422-5946948
2	VPLSAGSVLR	0	NCLIV_chrXI-1R-1669016-1668857
2	DGAAHRGGEPEEDK	0	NCLIV_chrXI-1R-5282252-5279483
4	SLLDDPAKVR	0	NCLIV_chrXII-0F-2460229-2461584
1	GIAESGLPR	0	NCLIV_chrXII-0F-3948493-3949539
3	YAGVDGYTISEVAEK	0	NCLIV_chrXII-0F-5301892-5302281
4	SDRGPGELGR	0	NCLIV_chrXII-0R-519772-519505
3	TYCSSPVVNNGDGLVIQLPNAEQK	0	NCLIV_chrXII-1F-6284840-6286917
1	GCAKSSGNK	0	NCLIV_chrXII-1R-2879820-2879643

Table 7:3 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

Count	Sequence	FDRScore	Proteinaccessions
1	ASVTPLLR	0	NCLIV_chrXII-2F-1355697-1356084
1	LSLLPRVLR	0	NCLIV_chrXII-2R-1227782-1227554
4	QRAAAALR	0	unknown-0F-1568263-1571940
1	DENILLWGSGQIR	0	unknown-1R-3128010-3127821
1	QDEAHAHR	0.004830918	NCLIV_chrX-0R-5402337-5402178
1	QKAQAVLR	0.005	NCLIV_chrXI-0F-4103317-4103841
1	ASLSSLAVSGKNPR	0.005714286	NCLIV_chrX-0R-5646729-5646399
1	LLCAMHLLSQMAK	0.00617284	NCLIV_chrVIIb-1R-988715-988472
2	AENWLTGDIVTFEHVR	0.006802721	NCLIV_chrX-2F-153351-153633
1	CVPLARLSVFPLK	0.007633588	NCLIV_chrXI-0R-2011161-2010312
1	LLALLFSK	0.007692308	NCLIV_chrIX-1R-3203846-3203687
1	LAGDACHRGGR	0.007751938	NCLIV_chrVIIb-2F-4308852-4309167
1	LLLFSLRLR	0.008196721	NCLIV_chrII-2F-1151979-1152399
1	RASVVSVPAR	0.008403361	NCLIV_chrIX-0R-4582146-4581930
1	ASSGATADLAASR	0.008403361	NCLIV_chrV-2R-745152-744507
1	DDAVPPCLQAECSAER	0.00877193	NCLIV_chrX-1R-707114-706205
1	ALPANLGNAAAEELATK	0.008849558	NCLIV_chrXII-2F-1021383-1021527
2	SSAILHTNAPR	0.008928571	NCLIV_chrVIII-2F-5922858-5923008
10	EFLCSGRIEK	0.008928571	NCLIV_chrX-2R-5889466-5889058
1	VCVSVSVDGR	0.008928571	NCLIV_chrXII-2F-2987286-2987412
1	NSIVSPSFVPVPWASR	0.009090909	NCLIV_chrVIIa-1R-1207613-1207184
1	VNTLAK	0.009174312	NCLIV_chrIX-1R-1798907-1798784
1	GRHSAGSSGPLR	0.009174312	NCLIV_chrXII-2R-660887-660404
2	GASGQGGGGAR	0.009259259	NCLIV_chrVIIb-1R-1352795-1352513
1	DTKQGSLAPEESL	0.009259259	NCLIV_chrVIIb-2F-2359170-2359611
2	SPRFSLCK	0.009259259	NCLIV_chrXII-2R-377264-377096
4	GSHNSAGSGESGAFTLS	0.009302326	NCLIV_chrVIII-0R-4075086-4074879
2	LLPLNSSLGVAR	0.009345794	NCLIV_chrVIIb-2R-2282395-2282017
7	SRFVISLVPK	0.009433962	NCLIV_chrVIIb-2F-4738965-4739370
2	TTNPPFGGFEGR	0.00952381	NCLIV_chrIII-1F-1254839-1255014
1	ADTFEGPLR	0.009615385	NCLIV_chrX-1F-3752189-3752448
3	ARLLAVSFQR	0.009708738	NCLIV_chrV-0R-2315927-2315615
4	CQEKFLR	0.009708738	NCLIV_chrV-1F-1913507-1913631
3	TDSTATLDVR	0.009803922	NCLIV_chrIb-2F-781893-782193
1	SSLSEAPVAGLQGK	0.00990099	NCLIV_chrXII-2R-1700849-1700357
1	GTGGLAGEINAAPR	0.01	NCLIV_chrVIII-1F-6171734-6171966
1	EDPQAGSGK	0.010152284	NCLIV_chrX-0R-2348127-2347926
1	SAGGEEKR	0.010204082	NCLIV_chrIV-2R-934654-934303
1	ISSSAQFLSVQ	0.010416667	NCLIV_chrIX-2R-2554684-2553709
1	ADALPISK	0.010526316	NCLIV_chrVIII-1R-3645131-3644252

Table 7:4 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

Count	Sequence	FDRScore	Proteinaccessions
8	IAPFALKSVTSMVADR	0.010638298	NCLIV_chrIa-2R-1585993-1585864
3	LIEPLEKVR	0.010638298	NCLIV_chrX-2F-2276595-2279694
1	AAAHLATAVPGLQDAMGKDER	0.010638298	NCLIV_chrXII-2F-5440368-5441715
1	VYASPDLAASR	0.011111111	NCLIV_chrXII-0R-2101810-2101402
1	EAIEKNALR	0.011299435	NCLIV_chrVIIa-2F-1182546-1182684
2	GRTSSPMDER	0.011494253	NCLIV_chrVIII-1R-4317713-4317503
3	VALLASSALPR	0.011904762	NCLIV_chrXI-2R-3744835-3744715
2	HSFFVAASGIDK	0.012195122	NCLIV_chrVIII-2R-4380001-4379821
1	FSLLLNR	0.0125	NCLIV_chrIb-0R-1847159-1847024
1	GTRQDWDLTTAMR	0.012820513	NCLIV_chrVIIb-2R-4095349-4094455
1	LADEVRYVLGLR	0.012987013	NCLIV_chrXI-2F-1959876-1959999
2	RLGGLAGEER	0.013157895	NCLIV_chrII-0F-934159-934371
1	SSEDIYR	0.013333333	NCLIV_chrVIIb-1R-13778-13562
1	GSLLQLGEPTPR	0.013333333	NCLIV_chrVIII-1F-3244895-3245064
1	FSSWGGGGGRSVHSLCLR	0.013636364	NCLIV_chrVIIa-2F-364077-364611
1	EGSGGSFSR	0.014705882	NCLIV_chrIb-0F-1487911-1488366
1	HPCMALLPFAK	0.014705882	NCLIV_chrIX-2F-932445-933087
1	GTGTAMPAASSPAHAVK	0.014705882	NCLIV_chrVIII-0F-2867986-2868384
1	AGSGSETGPR	0.014705882	NCLIV_chrVIII-0R-4042488-4042329
2	TISAAKVR	0.015873016	unknown-2R-1190918-1190636
2	RTLMIVPR	0.016216216	NCLIV_chrVIII-0F-2133730-2133882
2	IPNISR	0.016260163	NCLIV_chrVIII-0F-5960164-5960385
1	FPPCTADPAR	0.016304348	NCLIV_chrXI-0F-912004-912855
1	RPCGQIIDTR	0.016393443	NCLIV_chrIV-0F-962488-962652
2	FDGRAHGVLSR	0.016528926	NCLIV_chrXI-1R-6007130-6006401
1	TPGALLLRGTLR	0.017094017	NCLIV_chrIII-2F-832281-833169
3	CATGSREGCGR	0.017094017	NCLIV_chrXII-1F-4199714-4199928
1	TPTYAAVCLHLR	0.017241379	NCLIV_chrIa-0F-783769-783951
1	GVIAPPER	0.017699115	NCLIV_chrVIIa-2R-3248203-3248029
1	VVSESPYSCPWGMLSFRR	0.01793722	NCLIV_chrIb-2R-294297-294045
1	LDERLIR	0.018018018	unknown-0R-480982-480622
1	SGVSEATLAVACLDSTLSLLSR	0.018181818	NCLIV_chrIII-0R-1210902-1210782
1	WRDTQTNISVVAFGNVSK	0.018181818	NCLIV_chrVIIa-0R-2165277-2165076
1	QEAMNVNIELSR	0.018518519	NCLIV_chrIb-0R-56639-55937
1	QVDVPSLQR	0.018518519	NCLIV_chrVI-2R-335707-335278
1	YDNIDQ	0.01863354	NCLIV_chrXI-2R-2918488-2918308
1	FLAANAGR	0.018867925	NCLIV_chrIII-2F-736197-736320
1	ASTTAIQNLSVSPR	0.018867925	NCLIV_chrVIIa-1R-2508704-2508119
1	FRLLSPILSLQR	0.019047619	NCLIV_chrVIII-1F-2617409-2617821
1	MTGIGSNK	0.01910828	NCLIV_chrVIII-2F-935172-935412

Table 7:5 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

Count	Sequence	FDRScore	Proteinaccessions
1	LKMPVEGQVAEAGSLPVDK	0.019230769	NCLIV_chrVIII-1R-3302765-3302126
1	AQRAACLALTLLR	0.019230769	NCLIV_chrVIII-2R-2915620-2915239
1	VQMARPR	0.019230769	NCLIV_chrXI-2F-2527749-2527989
1	RPTYSTLSSLASFLSIAWSR	0.019230769	NCLIV_chrXI-2F-3974271-3974568
1	RQSGCGILALGGR	0.019607843	NCLIV_chrXII-2F-3680436-3681300
1	IVELIR	0.01986755	NCLIV_chrVIIb-1F-3171491-3171630
1	DRTGWWNLWR	0.02	NCLIV_chrIX-0R-2168865-2168328
1	EAQQNVQNESK	0.020512821	NCLIV_chrIX-2R-2357389-2357140
2	LYKFAIR	0.020833333	NCLIV_chrX-2F-5992617-5993286
1	VAILSRTAR	0.020833333	NCLIV_chrXII-2F-765411-765801
1	LSSPSRGLEPK	0.021276596	NCLIV_chrX-2F-2056071-2056725
1	NVESALSGVHLNQNTR	0.021276596	NCLIV_chrXII-1R-2812863-2812260
1	SSVRAGDTGVPR	0.021428571	NCLIV_chrVIII-2F-6212388-6212901
1	LLEFSSAWR	0.021505376	NCLIV_chrV-0R-298673-298352
2	YFPQGGERGLR	0.021582734	NCLIV_chrIb-2F-1455948-1456506
1	EVVLAAKDAFEAQR	0.02173913	NCLIV_chrIb-0R-56639-55937
12	ESLLSSFSSSFVDSR	0.02173913	NCLIV_chrIV-0F-284140-284871
1	LSSLEVMAESAR	0.02173913	NCLIV_chrIX-1R-1884140-1883288
1	NAATEFLLLLLIR	0.02173913	NCLIV_chrVIIb-1R-824942-824411
1	LSGAASFR	0.02173913	NCLIV_chrX-0R-6423576-6423285
1	GGGGGIDSRGSSPFQR	0.021978022	NCLIV_chrIX-2R-3167050-3166858
1	FLTNVYGVAAATLVQLHAR	0.022058824	NCLIV_chrVIII-1F-2729414-2730096
1	SDGRAQCGLGQALR	0.022222222	NCLIV_chrV-1R-758371-758170
1	LLQYVAAQSAARPPSPPPYAPPVR	0.02247191	NCLIV_chrIa-0R-1863018-1862271
1	EAAGRQLAVQER	0.022727273	NCLIV_chrXII-1F-4871960-4872210
4	GVCSECSGLEK	0.023529412	NCLIV_chrX-0R-289026-288210
1	MPLDSDIFRALVALVR	0.023809524	NCLIV_chrVIIa-1F-2999708-2999853
1	NLLLRPAHAR	0.024	NCLIV_chrIb-0F-895798-896775
1	ETHGAGTR	0.024096386	NCLIV_chrIX-1F-3308765-3309024
1	MQTQLSLLDLLSR	0.024242424	NCLIV_chrIa-1F-94412-94539
1	SSSVELVKPAE	0.024242424	NCLIV_chrVIIa-0R-37848-37623
1	MSAVTHGTASTAVVPMSR	0.024390244	NCLIV_chrVI-1F-3279413-3279657
3	RSSVVQLPK	0.024390244	NCLIV_chrX-0R-2736339-2736192
1	GVCRLPALPR	0.024630542	NCLIV_chrXI-2F-4352115-4352547
1	VCYLDGSARK	0.024752475	NCLIV_chrX-1R-1068626-1067801
1	GGAYGPFHAQNRTR	0.02484472	NCLIV_chrXI-1F-946289-946449
1	LSRLAFLVK	0.025	NCLIV_chrVIIb-1F-1330178-1330404
1	LLRCDTGNTPPTMK	0.025	NCLIV_chrXI-2F-5421168-5421495
1	GGLKCCAVSNIVTQR	0.025157233	NCLIV_chrVIII-0R-1084020-1083900
1	VVFRASGQTSLNK	0.025316456	NCLIV_chrVIII-2F-1898691-1898823

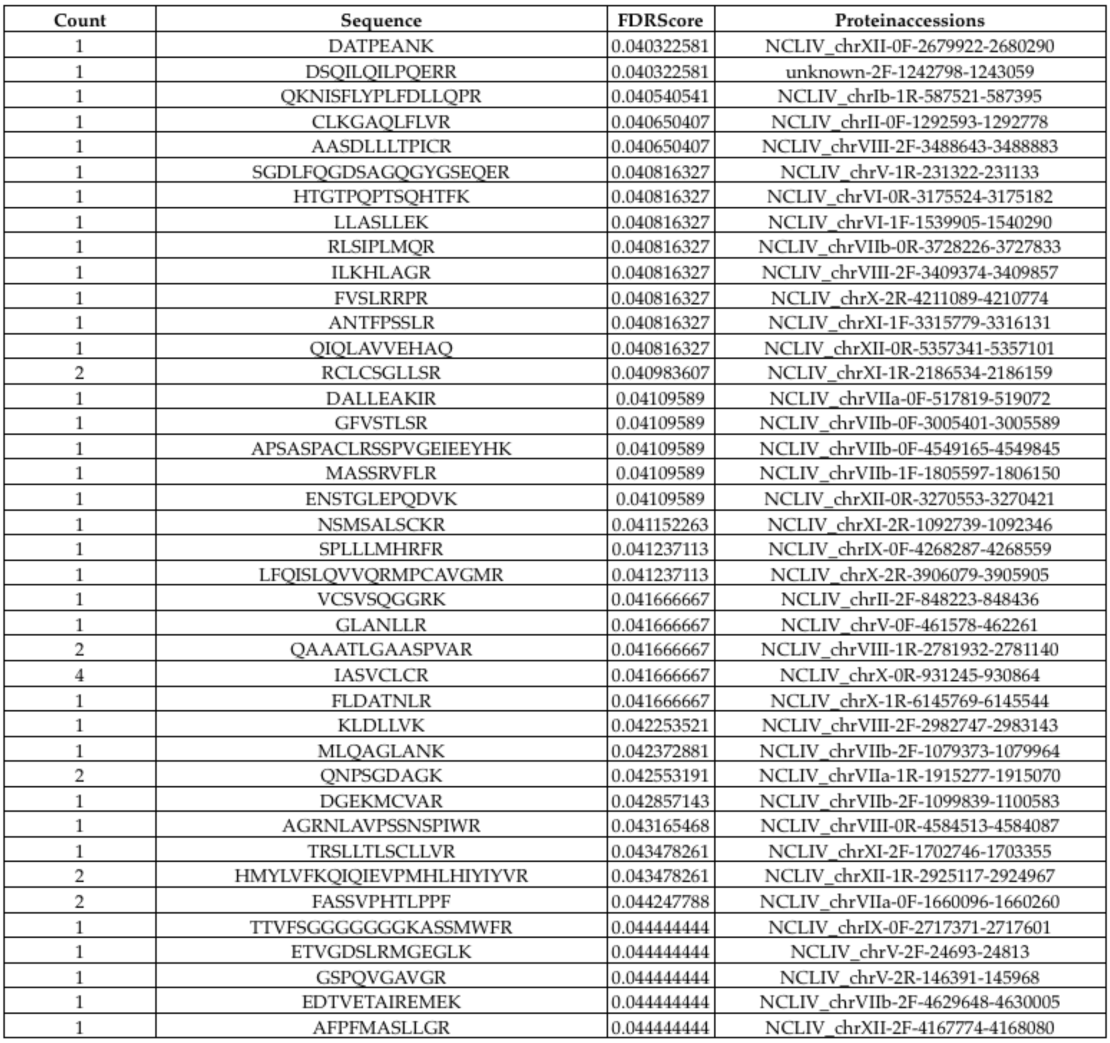
Table 7:6 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

Count	Sequence	FDRScore	Proteinaccessions
1	CPISSVATVSLVR	0.025531915	NCLIV_chrV-1F-1047689-1047927
2	SSILVELR	0.025641026	NCLIV_chrIX-0R-709959-709701
1	DAGLQRK	0.025641026	NCLIV_chrX-1R-1102637-1101809
1	QFVQLNILRAR	0.025751073	NCLIV_chrVIIb-2R-2621227-2621038
2	TFDSDELR	0.026315789	NCLIV_chrIb-0R-56639-55937
1	SMNRFSFDLPGSR	0.026315789	NCLIV_chrVI-2F-2304237-2304369
1	IKQSPAEYQTVSGTEVIAPLFEGEGK	0.026548673	NCLIV_chrIb-1F-1581293-1581624
1	SCPRLCAAVLAAQGDR	0.026666667	NCLIV_chrIb-2F-317424-317601
1	VRKPDGTLGPYQWK	0.026666667	NCLIV_chrVI-0R-2577966-2577402
1	SRADLLSR	0.026666667	NCLIV_chrX-0R-332586-332337
1	GFPSSRTEQK	0.027272727	NCLIV_chrVIIb-2R-3080215-3079783
1	YNVSAGLVAVAPSLK	0.02739726	NCLIV_chrX-2F-5253153-5253846
1	QGLLSRGER	0.02739726	NCLIV_chrX-2R-6377332-6376501
2	EPTELLK	0.027777778	NCLIV_chrVI-1F-903875-904410
1	SGMFRSVSTLLGICFLLER	0.027777778	NCLIV_chrXI-0R-4174431-4173591
1	KPSQALLVAR	0.027777778	NCLIV_chrXI-2F-4213785-4213965
1	EVKAAILK	0.027777778	NCLIV_chrXII-0F-1070221-1070502
2	LYLLVAK	0.028169014	NCLIV_chrIV-0F-1086913-1087236
1	LSSTMYEILNK	0.028169014	NCLIV_chrIX-0F-2167549-2167968
1	GVSTTAIR	0.028409091	NCLIV_chrVIIa-0F-3233470-3233616
1	GSAAGTARLGQPMK	0.028571429	NCLIV_chrIII-1F-908453-908607
1	GAPGLGLGGGSQPAFR	0.028571429	NCLIV_chrV-0F-1152802-1153107
1	IIAHASGTCIMMGEGCRGK	0.028571429	NCLIV_chrVIIb-2R-4663552-4662811
2	NAGTTKSTPSASSHSSA	0.028571429	NCLIV_chrXII-0R-3629755-3629449
1	RVTQGGGGGGSSALPVSR	0.028735632	NCLIV_chrIb-2R-1606251-1605975
2	MQLSSDYLPC	0.028776978	NCLIV_chrVIII-0R-2039787-2039658
1	MLKLDAGDEK	0.028846154	NCLIV_chrV-0R-948806-948062
1	KYALNLLFHLR	0.028901734	NCLIV_chrIb-2R-960120-959961
2	VSISIR	0.029069767	NCLIV_chrIX-2F-5379951-5380455
1	SPGTSSLSDRNSSR	0.029069767	NCLIV_chrX-2R-791824-790918
1	KENEGRPR	0.029126214	NCLIV_chrIII-2R-200476-200200
2	HDDATGSAHGGAWKR	0.029288703	NCLIV_chrIa-1R-1991495-1990763
1	SCSSENGEKNVT	0.029411765	NCLIV_chrV-1R-662092-661933
1	LVLRGFSR	0.02962963	NCLIV_chrIX-1R-601886-601424
1	SSFGGVRACLR	0.03030303	NCLIV_chrV-2R-1422450-1422174
1	AFPDKIIIADGK	0.03030303	unknown-0F-56260-56598
1	RMLALLR	0.030434783	NCLIV_chrXII-2F-4313610-4314030
2	LSLLRGSLR	0.030567686	NCLIV_chrIX-0F-1500457-1500849
1	ILMRAEGGCTER	0.030612245	NCLIV_chrVIIa-2F-2747916-2748597
1	AFLVLSR	0.030864198	NCLIV_chrVIIa-2F-1275879-1276059

Table 7:7 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

Count	Sequence	FDRScore	Proteinaccessions
1	RIMVPELLVPPLLR	0.030927835	NCLIV_chrIX-1F-148760-149016
1	CVVIIGSLLLK	0.030927835	NCLIV_chrVIII-1F-1433423-1433622
1	RVANALCR	0.031007752	NCLIV_chrVIII-0R-3337023-3336558
1	VNSTQPADVAK	0.03125	NCLIV_chrVIIa-0R-2610144-2609898
1	TEGALIQTVR	0.031578947	NCLIV_chrIX-0F-2481610-2481807
1	EAEGNGQR	0.031746032	NCLIV_chrXI-1F-1456487-1456635
1	DDFVTQILQECILAESTQTR	0.032258065	NCLIV_chrII-2F-278880-279189
2	RFSAVGFPK	0.032258065	NCLIV_chrVIIb-1F-4761920-4762116
4	GDNGILMLK	0.032407407	NCLIV_chrV-2R-2116971-2116800
1	IPSESPEKK	0.032467532	NCLIV_chrX-2F-214731-215301
2	LLASLLEKK	0.032520325	NCLIV_chrVI-1F-1539905-1540290
2	ETEGFAEGR	0.03271028	NCLIV_chrXI-1F-3506846-3506982
1	GTRPGGAAAR	0.032786885	NCLIV_chrVI-0R-1946283-1945770
1	ERPAIGHGESQRTK	0.033057851	NCLIV_chrVIII-1F-1003379-1003788
1	EDEEAGK	0.033175355	NCLIV_chrXI-2R-3171505-3171211
1	TPLGVSVSR	0.033333333	NCLIV_chrIb-1R-1390444-1390171
1	SSFIPLNLR	0.033333333	NCLIV_chrVIIa-0F-1675585-1675866
1	EVVMKANVAVER	0.033333333	NCLIV_chrX-1F-603290-603678
3	AATRHGEAAQLR	0.033333333	NCLIV_chrXII-2R-2821271-2820761
1	MDNCGGSESVVLRRPAR	0.033472803	NCLIV_chrXII-2R-4628465-4628342
1	IPSIASVFR	0.033519553	NCLIV_chrVI-0R-3312915-3312747
1	FPLPLSLR	0.033613445	NCLIV_chrV-2R-679899-679587
1	GPVGGGEERSPGVCTVK	0.033707865	NCLIV_chrXI-1R-4005566-4005245
1	ASVDRGVPLCVFSSR	0.034042553	NCLIV_chrVIIb-0F-4049548-4050807
1	SSLFSSSVSSRSSSSMR	0.034090909	NCLIV_chrVIIa-0R-3881652-3881064
1	SRLSVTSVAENR	0.034090909	NCLIV_chrXI-2R-3749503-3749281
1	SRQIKPPTSR	0.034090909	unknown-0F-1760245-1760769
1	WGAVAGDSK	0.034188034	NCLIV_chrVIII-1F-4594031-4594200
3	APVSSLKR	0.034482759	NCLIV_chrIX-1R-895802-892685
2	LLRPMEGVPVPER	0.034482759	NCLIV_chrXI-0F-3991981-3992232
2	VIYVLK	0.034722222	NCLIV_chrVIII-1F-808676-808854
3	TRSMVSMGVDGE	0.034883721	unknown-0R-2127955-2127832
1	SSTLAFFSALSTR	0.034965035	NCLIV_chrIII-0F-1033597-1033935
1	GLRLLSK	0.035211268	NCLIV_chrVIIb-1R-1750232-1750001
2	TPVRTCEFFK	0.035294118	NCLIV_chrV-2R-1841229-1840944
1	HLLSPPLPLCR	0.035294118	NCLIV_chrVIIa-0R-995991-995265
1	ILTGCILLLTLSPK	0.03539823	NCLIV_chrXII-2F-5335383-5335686
1	TQHRVTGK	0.035460993	NCLIV_chrVIII-1R-2551754-2551544
2	NPILPLPFFAR	0.035714286	NCLIV_chrII-0R-635912-635264
1	GRTLVATALLR	0.035714286	NCLIV_chrIX-0R-312417-311952

Table 7:8 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.



Count	Sequence	FDRScore	Proteinaccessions
1	DATPEANK	0.040322581	NCLIV_chrXII-0F-2679922-2680290
1	DSQILQILPQERR	0.040322581	unknown-2F-1242798-1243059
1	QKNISFLYPLFDLLQPR	0.040540541	NCLIV_chrIb-1R-587521-587395
1	CLKGAQLFLVR	0.040650407	NCLIV_chrII-0F-1292593-1292778
1	AASDLLLTPICR	0.040650407	NCLIV_chrVIII-2F-3488643-3488883
1	SGDLFQGDSAGQGYGSEQER	0.040816327	NCLIV_chrV-1R-231322-231133
1	HTGTPQPTSQHTFK	0.040816327	NCLIV_chrVI-0R-3175524-3175182
1	LLASLLEK	0.040816327	NCLIV_chrVI-1F-1539905-1540290
1	RLSIPLMQR	0.040816327	NCLIV_chrVIIb-0R-3728226-3727833
1	ILKHLAGR	0.040816327	NCLIV_chrVIII-2F-3409374-3409857
1	FVSLRRPR	0.040816327	NCLIV_chrX-2R-4211089-4210774
1	ANTFPSSLR	0.040816327	NCLIV_chrXI-1F-3315779-3316131
1	QIQLAVVEHAQ	0.040816327	NCLIV_chrXII-0R-5357341-5357101
2	RCLCSGLLSR	0.040983607	NCLIV_chrXI-1R-2186534-2186159
1	DALLEAKIR	0.04109589	NCLIV_chrVIIa-0F-517819-519072
1	GFVSTLSR	0.04109589	NCLIV_chrVIIb-0F-3005401-3005589
1	APSASPACLRSSPVGEIEEYHK	0.04109589	NCLIV_chrVIIb-0F-4549165-4549845
1	MASSRVFLR	0.04109589	NCLIV_chrVIIb-1F-1805597-1806150
1	ENSTGLEPQDVK	0.04109589	NCLIV_chrXII-0R-3270553-3270421
1	NSMSALSCKR	0.041152263	NCLIV_chrXI-2R-1092739-1092346
1	SPLLLMHRFR	0.041237113	NCLIV_chrIX-0F-4268287-4268559
1	LFQISLQVVQRMPCAVGMR	0.041237113	NCLIV_chrX-2R-3906079-3905905
1	VCSVSQGGRK	0.041666667	NCLIV_chrII-2F-848223-848436
1	GLANLLR	0.041666667	NCLIV_chrV-0F-461578-462261
2	QAAATLGAASPVAR	0.041666667	NCLIV_chrVIII-1R-2781932-2781140
4	IASVCLCR	0.041666667	NCLIV_chrX-0R-931245-930864
1	FLDATNLR	0.041666667	NCLIV_chrX-1R-6145769-6145544
1	KLDLLVK	0.042253521	NCLIV_chrVIII-2F-2982747-2983143
1	MLQAGLANK	0.042372881	NCLIV_chrVIIb-2F-1079373-1079964
2	QNPSGDAGK	0.042553191	NCLIV_chrVIIa-1R-1915277-1915070
1	DGEKMCVAR	0.042857143	NCLIV_chrVIIb-2F-1099839-1100583
1	AGRNLAVPSSNSPIWR	0.043165468	NCLIV_chrVIII-0R-4584513-4584087
1	TRSLLTLSCLLVR	0.043478261	NCLIV_chrXI-2F-1702746-1703355
2	HMYLVFKQIQIEVPMHLHIYIYVR	0.043478261	NCLIV_chrXII-1R-2925117-2924967
2	FASSVPHTLPPF	0.044247788	NCLIV_chrVIIa-0F-1660096-1660260
1	TTVFSGGGGGGGKASSMWFR	0.044444444	NCLIV_chrIX-0F-2717371-2717601
1	ETVGDSLRMGEGLK	0.044444444	NCLIV_chrV-2F-24693-24813
1	GSPQVGAVGR	0.044444444	NCLIV_chrV-2R-146391-145968
1	EDTVETAIREMEK	0.044444444	NCLIV_chrVIIb-2F-4629648-4630005
1	AFPFMASLLGR	0.044444444	NCLIV_chrXII-2F-4167774-4168080

Table 7:9 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

Count	Sequence	FDRScore	Proteinaccessions
1	FAVAHLSPSGR	0.044534413	NCLIV_chrVIIb-0F-4423969-4424409
1	VDDSAFVSRGADR	0.044642857	NCLIV_chrXI-1R-1487129-1486634
1	HLLQLSTR	0.044776119	NCLIV_chrIX-0F-2289472-2289996
1	FLVVSSSVLAR	0.044776119	NCLIV_chrV-0R-1044248-1043924
1	LQRASNASGDMR	0.045045045	NCLIV_chrIX-1R-152651-152006
1	TPFSALSDGDR	0.045112782	NCLIV_chrVI-2F-1950984-1951116
1	LEKEPLDMTDIIR	0.045454545	NCLIV_chrIV-0F-2229826-2230416
1	GGCIATAR	0.045454545	NCLIV_chrV-2R-372189-371586
1	PITLSKCLGDR	0.045454545	NCLIV_chrVIII-1F-216296-216657
1	LLLRGPR	0.045454545	NCLIV_chrXI-0R-609648-609186
2	TGSLDAPDR	0.045454545	NCLIV_chrXII-2F-4381722-4382112
1	LSPPHGSER	0.045454545	NCLIV_chrXII-2R-6375659-6375470
1	QSYLFGPGK	0.04587156	NCLIV_chrIV-1R-770126-769838
1	EVVDLLLPLSNDSTDFVR	0.04587156	NCLIV_chrVIIb-2R-783916-782851
1	THVAASVDGRVPVEDSR	0.045918367	NCLIV_chrIa-1F-710201-710379
1	DDNCGGSSSPK	0.046153846	NCLIV_chrVIIb-1F-3946325-3946779
2	GMVLLER	0.046296296	NCLIV_chrVIII-0F-2970643-2971488
1	MPVLLLDHPENPLGAGPR	0.046296296	NCLIV_chrX-2R-229900-229540
1	LSPGATPSKLR	0.046357616	NCLIV_chrXII-0F-5092816-5093283
1	VAHTRESSPASLLR	0.046511628	NCLIV_chrVI-0F-1437889-1438224
2	TLILVMNTR	0.046728972	NCLIV_chrVI-2R-2291074-2290888
2	FRALSSLSLSR	0.046728972	NCLIV_chrVIII-0F-2526289-2526501
1	LLACGQKGGFPQNATK	0.046875	NCLIV_chrV-2R-12537-12285
1	SSSASTGSWR	0.046875	NCLIV_chrXI-2R-1625584-1625335
1	SFEIAASSR	0.046875	NCLIV_chrXII-0F-2700292-2700810
1	RSIFLEHSR	0.046979866	NCLIV_chrIX-0R-478218-478083
1	EIGPGPVK	0.047120419	NCLIV_chrVIIa-2R-3606811-3606499
1	QPFGADLLGGNITVAR	0.047169811	NCLIV_chrXI-0F-2027176-2027490
1	TPPEGR	0.047297297	NCLIV_chrVIIa-0R-3732063-3731898
1	VQELLFSPPALK	0.04743083	NCLIV_chrVIIa-0F-1166479-1167447
1	RMQAHSHIR	0.047619048	NCLIV_chrII-2F-1493331-1493592
1	MKVVEAVSALK	0.047619048	NCLIV_chrIX-1F-2188547-2188902
2	ETMMGLAKASGR	0.047619048	NCLIV_chrV-2R-2216565-2215722
1	LLEVQK	0.047619048	NCLIV_chrVIIa-1R-1749467-1749275
1	QHVVDRDDFAQSPAPAAR	0.048	NCLIV_chrIX-2F-4043625-4043949
1	SETGTGK	0.048	NCLIV_chrVIIa-1F-3723482-3724524
1	SPATQENGLARK	0.048076923	NCLIV_chrIII-2F-81969-82353
1	QLIARGALHR	0.048076923	NCLIV_chrVI-0R-869256-868896
1	SLGSSGGISPAGARER	0.048192771	NCLIV_chrVIIb-1F-1557407-1557669
1	QDTELDATR	0.048309179	NCLIV_chrIII-2F-764649-764775
1	MSASGSAPR	0.048484848	NCLIV_chrV-0R-543209-542690
1	TVERCAIAHR	0.048543689	NCLIV_chrIX-0F-1296433-1296627
1	IVFRDTVEATK	0.048543689	NCLIV_chrVIIb-2F-3097683-3098181
1	SPPVASSSSASGSTPR	0.048648649	NCLIV_chrIX-0R-1205736-1202442
1	HIRTQALR	0.048780488	NCLIV_chrII-1F-1628087-1628436
1	SSSPTGHR	0.048913043	NCLIV_chrV-2R-2471310-2470989
1	DPDSLR	0.048951049	NCLIV_chrIX-2R-2318521-2317852
1	DESAYLSNVNR	0.049019608	NCLIV_chrX-0F-76444-76578
1	VPLSQGREVHFLR	0.049079755	NCLIV_chrVI-1F-165380-165537
1	IETGIGGHR	0.049107143	NCLIV_chrVIII-2F-5124828-5125233
1	FLGLLLRLLCR	0.04964539	NCLIV_chrXI-0R-3581820-3581553

Table 7:10 The result table for dataset from *N. caninum* release 5.1 lists the peptides uniquely identified on the ORF_SS database. *The complete list includes Table 7:1 to Table 7:10.

	Slice 55	Slice 56	Slice 60	Slice 62	Slice 70	Slice 71	Slice 72	Slice 73	Slice 92	Slice 106
gene model 6.0	82	178	198	150	122	177	242	263	81	180
ORFs 6.0	81	132	156	111	99	162	184	207	57	164
Twinscan Prediction	81	131	141	109	90	153	162	158	17	123
GLMHMM 5 Predictions	77	168	179	147	99	171	219	233	50	160
GLMHMM 1 Prediction	78	166	175	144	98	170	226	231	50	152
GLMHMM Prediction MEG49 training	113	154	115	114	86	152	177	194	36	123

Table 7:11 The table shows comparative results from querying 10 1D gel from a *N. caninum* dataset; these represent the TPs at % fixed FDR. The different gene models databases comprise the official gene model, the ORF_SS and alternative models. Twinscan prediction was based on release 5.1 generated by Sanger Institute. The predictions generated with Glimmer were based on *N. caninum* and *T. gondii* training sets. For *N. caninum* we evaluated the differences between different sets of predictions (including differently spliced genes).

Table 7:12 Here below the list of the remaining 116 ISPs (from the set of 126) processed through the pipeline (first set) ordered by OMSSA e-value. The first three and the last columns identify gene ID, peptide, PTMs and e-value (OMSSA). The fourth column shows the number of TAGs that align on the peptide (prefix/ suffix mass not considered here). The column "R" specify the reason why the match failed: 'WR-TAG/ REG' indicates false positive ISPs in the temporary DB due to incorrect TAG or wrong genomic region selected; 'WR' indicates that OMSSA failed to match the correct peptide; 'NR-TAG/ REG' indicates that OMSSA yield null result due to incorrect TAG or wrong genomic region selected; 'NDB-TAG/ REG' indicates that the temporary dataset of ISPs candidates was not generated due to incorrect TAG or wrong genomic region selected. The last peptide sequence is flagged with WR as OMSSA failed to correctly identify the ISP although it was contained in the temporary DB. The second last peptide instead is flagged with WR-TAG InSPecT did not provide correct TAGs identification for such spectrum.

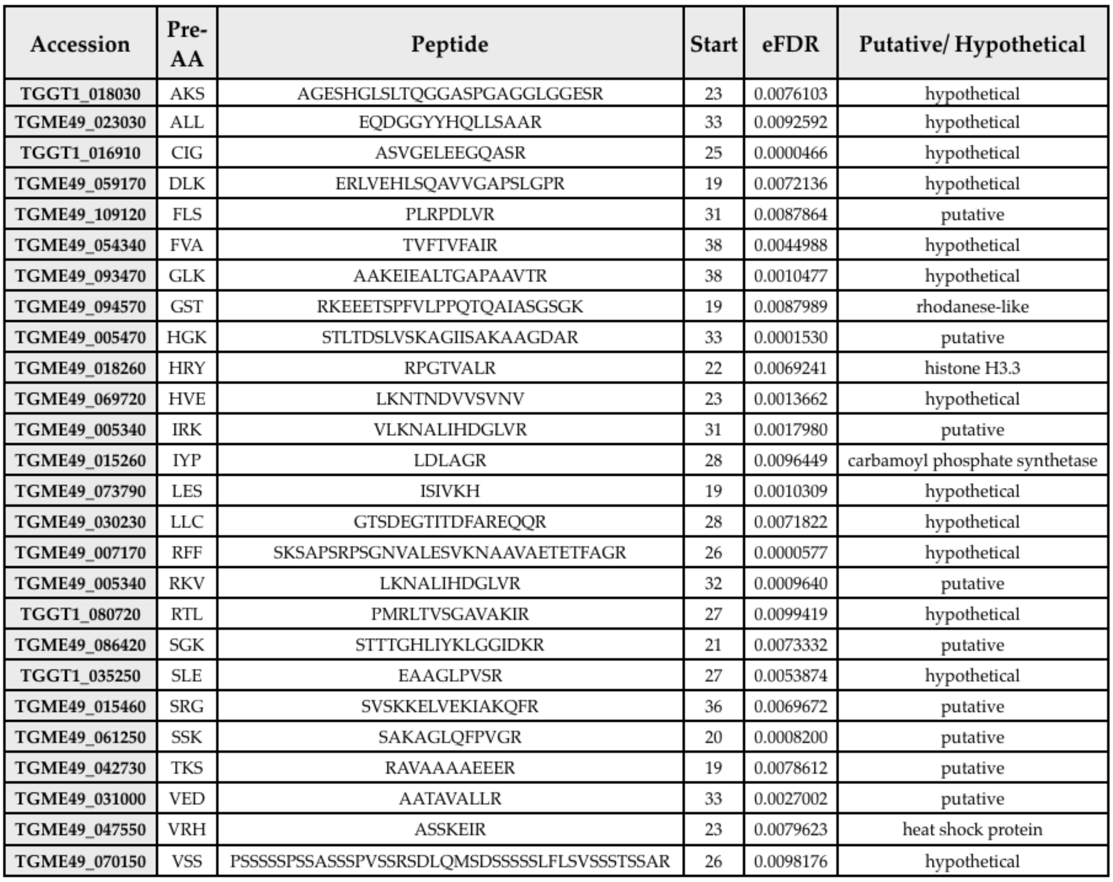
GENE	Peptide	PTM	TAG	R	*e*-value
TGME49_088 360	AILIKELQALVLGHQER	N/A	4	OK	1.13E-11
TGME49_090 670	FDMGGAAAVLGAAR	M:+16	2	OK	1.16E-10
TGME49_031 640	QWVAITAYQPIDTVTK	N/A	3	OK	1.24E-10
TGME49_029 010	SLYGGIANTLETPFADSEAVAK	N/A	8	OK	1.74E-10
TGME49_004 400	TQTSTEEVGRVVSVGDGIAR	N/A	1	OK	6.72E-10
TGME49_005 510	NVELSPLLPDEIAAEVK	N/A	8	OK	1.34E-09
TGME49_064 610	LFVGGISDDVNDESLR	N/A	3	OK	2.08E-09
TGME49_050 770	ELAQQIQKVVLALGDYLQVR	N/A	9	OK	1.48E-08
TGME49_039 820	EALLGGLPQSAVNLQCVR	N/A	3	OK	6.10E-08
TGME49_064 610	HTVDGTQVEVR	N/A	7	OK	3.65E-07
TGME49_072 910	LLAPIAVDAVMK	M:+16	1	OK	3.76E-07
TGME49_091 950	TAELIQGPPGTPGGAAAAGA DLSAQSR	N/A	7	OK	6.06E-07
TGME49_088 360	ELQALVLGHQER	N/A	8	OK	8.57E-07
TGME49_065 450	AIVNDTVGTLVSCAYQR	N/A	2	OK	9.69E-07
TGME49_087 500	ALGATAVVR	N/A	3	OK	1.48E-06
TGME49_034 500	DVQDTFFLQAPK	N/A	1	OK	3.02E-06
TGME49_066 990	TMEEFVIDLLR	M:+16	5	OK	1.12E-05
TGME49_048 810	EGHMVVGDESAVITLK	M:+16	2	OK	1.94E-05

TGME49_088500	NFLFEVPKLNTVLFANDVK	N/A	5	OK	2.69E-05
TGME49_090670	MGSQVFVR	M:+16	3	WR	0.0005
TGME49_026960	DAGTFSGSFSPLCFEFSDSLR	N/A	2	OK	0.0006
TGME49_075690	RVVGQDHAVQVVAEAIQR	N/A	2	WR-TAG	0.001
TGME49_111720	NAVVTVPAYFNDAQR	N/A	7	WR	0.0015
TGME49_029360	TNETNGGAAQNALEALR	N/A	6	WR-REG	0.0015
TGME49_072910	LLAPIAVDAVMK	M:+16	1	OK	0.0022
TGME49_063180	DVLRPEIIEITR	N/A	0	WR-TAG	0.0038
TGME49_019320	KSIDAFNFVSQLPEVR	N/A	0	WR-TAG	0.0048
TGME49_019590	TAPYTISEVVQVLK	N/A	1	WR-REG	0.006
TGME49_119920	LADIGEGIAQVELLK	N/A	0	WR-TAG	0.0089
TGME49_039820	TIGIIGLGQVGTHVAR	N/A	0	WR-TAG	0.0108
TGME49_035470	LPSEEYQLGK	N/A	6	WR-REG	0.0113
TGME49_080380	IDELFPSGLQYITPENPQVHLR	N/A	1	WR	0.0163
TGME49_028210	TLMELLNQLDGFDELGAVK	M:+16	2	WR-REG	0.041
TGME49_034500	GNEEASLDVAAIK	N/A	10	WR-TAG	0.0466
TGME49_009950	FTKLDADEHLSK	N/A	8	OK	0.0479
TGME49_049180	KTDDAATAEPSNAMSSLTSTR	M:+16	2	WR-TAG	0.0712
TGME49_032130	RPPSVFFINITHDPEGR	N/A	2	WR-REG	0.0886
TGME49_026960	TLGEIVTFVADAVK	N/A	4	WR-REG	0.1587
TGME49_051780	KSQVFSTAADNQTQVGIK	N/A	1	WR-REG	0.1644
TGME49_004400	TAVAVDAIINQK	N/A	4	WR-TAG	0.1729
TGME49_057990	ILDSALVEAAQLADRYITSR	N/A	3	WR-TAG	0.3173
TGME49_019540	QTLGMAEGNFPK	M:+16	3	WR-TAG	0.351
TGME49_028210	VALDMTTLTVMR	M:+16	0	WR-TAG	0.3582
TGME49_119920	ERPAPVSEPQAAASPSVGAEASSTTFSASPATR	N/A	10	WR-TAG	0.4148
TGME49_029990	TANAAAVQSIANILR	N/A	0	WR-TAG	0.5549
TGME49_073090	GQVVVIGATNR	N/A	4	WR-TAG	0.6177

TGME49_010730	TSHHLNTMNANVGAR	M:+16	5	WR-TAG	0.905
TGME49_109750	GAGGPILIGSAR	N/A	1	WR-REG	1.1351
TGME49_072910	ASNNLMLDETER	M:+16	3	WR-REG	1.2718
TGME49_051780	SQVFSTAADNQTQVGIK	N/A	6	WR-REG	1.9657
TGME49_101440	MFDSDNSGKISSTELATIFGVSDVDSETWK	M:+16	9	WR-REG	3.5227
TGME49_094800	NMITGTSQADVALLVVPAEAGGFEGAFSKEGQTR	M:+16	2	WR-TAG	5.5918
TGME49_039820	GSVDSANADVLLLSAAQGVLR	N/A	5	WR	5.742
TGME49_007620	DGAIVTDPLLRLPANPDVFVAGDIAAYPYVK	N/A	4	WR	8.1389
TGME49_002370	DVYDADKELLVHAAMTALGSK	M:+16	2	NR-TAG	
TGME49_023680	QMMPMMQNVLDNPELLR	M:+16	3	NDB-TAG	
TGME49_023680	TFLNPQMMQASLQMQQAMQNMQR	M:+16	0	NDB-TAG	
TGME49_023930	MAFVEFYDLR	M:+16	6	NDB-REG	
TGME49_026960	TLGEIVTFVADAVK	N/A	3	NDB-TAG	
TGME49_026960	TLGEIVTFVADAVK	N/A	5	NDB-REG	
TGME49_027950	ALVPTKDEGTVGEGFVIPTQVNYVGLGGR	N/A	0	NDB-TAG	
TGME49_028490	SLAEAIANQGNVGSLLK	N/A	0	NDB-TAG	
TGME49_029010	SLYGGIANTLETPFADSEAVAK	N/A	4	NDB-TAG	
TGME49_029010	SVDTGSGSDASTEQQAGGQKVVTPIPASK	N/A	0	NR-TAG	
TGME49_031640	REAEMVHFPAVEGAPPLPTIPK	M:+16	3	NDB-REG	
TGME49_031640	TVPIGEETERQWVAITAYQPIDTVTK	N/A	3	NDB-TAG	
TGME49_031640	EAEMVHFPAVEGAPPLPTIPKVEQVFKPK	M:+16	0	NR-TAG	
TGME49_032280	TIITEAGGFFGTPVA	N/A	1	NR-REG	
TGME49_035970	SFDVNKPGEEATNLQGGVAGGSISQGVLK	N/A	0	NDB-TAG	
TGME49_039820	SQQLVSLVHLAEILGR	N/A	1	NDB-TAG	
TGME49_039820	NVALETLLASSDFITLHVPLLDKTR	N/A	0	NR-TAG	
TGME49_043800	LLDDMQTLEPAVFSSVPR	M:+16	4	NDB-TAG	
TGME49_049270	LAGKIDAGTDAKPSEK	N/A	4	NDB-REG	
TGME49_049270	GDFSQESINTFLTQLLAGK	N/A	3	NR-TAG	

TGME49_049270	GDFSQESINTFLTQLLAGK	N/A	3	NR-REG	
TGME49_049390	VSQDDRDVVLPDGTAVASGFEFR	N/A	0	NDB-TAG	
TGME49_049530	FVALQILENTIQTR	N/A	2	NR-TAG	
TGME49_051780	SQVFSTAADNQTQVGIK	N/A	3	NDB-TAG	
TGME49_053730	LAAADVAQSKDLVEEVTDVQGSVQR	N/A	2	NDB-TAG	
TGME49_059010	SLAEVASIVSETDIELMR	M:+16	2	NR-REG	
TGME49_061210	QLAQTVQTMEAK	M:+16	0	NR-TAG	
TGME49_061950	LVLEVAQHLGENTVR	N/A	1	NR-TAG	
TGME49_061950	VVDTGAPIQVPVGVETLGR	N/A	0	NR-TAG	
TGME49_063130	KDSEFSPGLEGVVAGESAISSVGPASLAGLTYR	N/A	6	NDB-REG	
TGME49_063180	STSISADEYALGKTMVFLKPQAAK	M:+16	1	NDB-REG	
TGME49_067550	TAGPPASGGAPQESR	N/A	7	NR-TAG	
TGME49_072910	DAVNDLSLDYLAK	N/A	3	NDB-TAG	
TGME49_077500	TMLEIVNQLDGFEAR	M:+16	4	NDB-TAG	
TGME49_078830	VVIGLSGGSTPLPIYSALR	N/A	10	NDB-TAG	
TGME49_079390	TAADIVNGALKK	N/A	0	NDB-TAG	
TGME49_086920	DLLPSLFVDSGLPAAELEER	N/A	4	NR-REG	
TGME49_088360	GADLDIDIPFQYLTFILNDDDQLKEIGEK	N/A	0	NDB-TAG	
TGME49_088360	MYQESYLDEFGIPANVKEVR	M:+16	12	NDB-REG	
TGME49_088360	MYQESYLDEFGIPANVKEVR	M:+16	0	NDB-TAG	
TGME49_088360	AILIKELQALVLGHQER	N/A	7	NR-TAG	
TGME49_089580	ILYPLTDFGPLSSALDALLTK	N/A	7	NR-REG	
TGME49_090200	AGGIIGTAFGQGGFDWAMLK	M:+16	3	NDB-TAG	
TGME49_090200	DFLNATVVPGNMGQPVR	M:+16	3	NDB-REG	
TGME49_090670	VVTSFLETLLVELQPDLR	N/A	0	NDB-TAG	
TGME49_090670	TGGAQIELMKFDMGGAAAVLGAAR	M:+16	0	NR-TAG	
TGME49_094200	ESNYDFPGNSLILEVQPHPSVR	N/A	2	NDB-REG	
TGME49_097470	LFYDEAMQDAFPEEGTMR	M:+16	0	NDB-TAG	

TGME49_097500	LPIGDLATQYFADRDIFCAGR	N/A	2	NDB-REG	
TGME49_101440	ISSTELATIFGVSDVDSETWK	N/A	6	NR-REG	
TGME49_101440	ISSTELATIFGVSDVDSETWK	N/A	6	NR-REG	
TGME49_104710	LSVLSAITSTQQR	N/A	2	NDB-REG	
TGME49_108860	LAGAPGPSIVIPATSLLSK	N/A	3	NR-TAG	
TGME49_109750	GLVAAAKEVGFDKPVVLR	N/A	5	NDB-REG	
TGME49_109750	MPIDINQGISEPR	M:+16	0	NR-TAG	
TGME49_111310	LNDPITVVGDIHGQFYDLLK	N/A	0	NDB-TAG	
TGME49_111470	VTASGPQLTSVADLDTQFKEIPDLVLR	N/A	1	NR-REG	
TGME49_113230	AALQAGQEVGDDEVTINIK	N/A	0	NR-TAG	
TGME49_113410	LQDPEPGVVALALQTLSVQLK	N/A	1	NDB-REG	
TGME49_113410	LQDPEPGVVALALQTLSVQLK	N/A	4	NR-TAG	
TGME49_114740	GAGFRTPAGVTVSLNPNEMEQEGVFTADVIR	M:+16	4	NDB-TAG	
TGME49_118230	LGIQDVGAQLTGK	N/A	1	NR-REG	



Accession	Pre-AA	Peptide	Start	eFDR	Putative/ Hypothetical
TGGT1_018030	AKS	AGESHGLSLTQGGASPGAGGLGGESR	23	0.0076103	hypothetical
TGME49_023030	ALL	EQDGGYYHQLLSAAR	33	0.0092592	hypothetical
TGGT1_016910	CIG	ASVGELEEGQASR	25	0.0000466	hypothetical
TGME49_059170	DLK	ERLVEHLSQAVVGAPSLGPR	19	0.0072136	hypothetical
TGME49_109120	FLS	PLRPDLVR	31	0.0087864	putative
TGME49_054340	FVA	TVFTVFAIR	38	0.0044988	hypothetical
TGME49_093470	GLK	AAKEIEALTGAPAAVTR	38	0.0010477	hypothetical
TGME49_094570	GST	RKEEETSPFVLPPQTQAIASGSGK	19	0.0087989	rhodanese-like
TGME49_005470	HGK	STLTDSLVSKAGIISAKAAGDAR	33	0.0001530	putative
TGME49_018260	HRY	RPGTVALR	22	0.0069241	histone H3.3
TGME49_069720	HVE	LKNTNDVVSVNV	23	0.0013662	hypothetical
TGME49_005340	IRK	VLKNALIHDGLVR	31	0.0017980	putative
TGME49_015260	IYP	LDLAGR	28	0.0096449	carbamoyl phosphate synthetase
TGME49_073790	LES	ISIVKH	19	0.0010309	hypothetical
TGME49_030230	LLC	GTSDEGTITDFAREQQR	28	0.0071822	hypothetical
TGME49_007170	RFF	SKSAPSRPSGNVALESVKNAAVAETETFAGR	26	0.0000577	hypothetical
TGME49_005340	RKV	LKNALIHDGLVR	32	0.0009640	putative
TGGT1_080720	RTL	PMRLTVSGAVAKIR	27	0.0099419	hypothetical
TGME49_086420	SGK	STTTGHLIYKLGGIDKR	21	0.0073332	putative
TGGT1_035250	SLE	EAAGLPVSR	27	0.0053874	hypothetical
TGME49_015460	SRG	SVSKKELVEKIAKQFR	36	0.0069672	putative
TGME49_061250	SSK	SAKAGLQFPVGR	20	0.0008200	putative
TGME49_042730	TKS	RAVAAAAEEER	19	0.0078612	putative
TGME49_031000	VED	AATAVALLR	33	0.0027002	putative
TGME49_047550	VRH	ASSKEIR	23	0.0079623	heat shock protein
TGME49_070150	VSS	PSSSSSPSSASSSPVSSRSDLQMSDSSSSSLFLSVSSSTSSAR	26	0.0098176	hypothetical

Table 7:13 The table lists N-terminal candidate peptides, for a *T. gondii* dataset, with novel signal peptide cleavage obtained with the frayed database (chapter four). It provides a view on: the peptide sequence and the preceding residues, estimated FDR and current annotation status.

Accession	Pre AA	Peptide	PS	SP site	eFDR	Signal Peptides
TGME49_110010	SSR	AAANGSEGGVAQSEQER	39	23	0.00021	MKTPFGCVFFFCLIALWGFSAAL
TGME49_120630	ASG	AQTEPAVDDSAVQQGSTAETSVAPTHLR	35	21	0.00138	MQVLACVLGIVCLYLRTAPAS
TGME49_044280	GSA	ASSVAAEQR	38	31	0.00037	MEICQLPRVRHLRMMVAVALGALLFLSICGE
TGME49_097810	ARL	ATAAAAGPAR	9	23 ǀ 21	0.00762	MLRSCARLATAAAAGPARAAAR/ MLRSCARLATAAAAGPARAA
TGME49_108080	RTN	DLASGTPHVAR	36	26	0.00318	MATKLARLATWLVLVGCLLWRAGAV
TGME49_036210	SRG	FFSAAPAAATAGVSPLAR	36	58 ǀ 13	0.00002	MMFRFLPRVASGASSLSVSQRRLRASFSSSLQSRGFFSAAPAAATAGVSPLARSVDAA/ MMFRFLPRVASGA
TGME49_031910	RNF	GAGDLKIVAAR	39	33	0.00095	MAGLASLSSVGALRGMRLVPAAHLLPLHSAFGQ
TGME49_022670	LFR	GFLNVAWIIAAFLR	22	44 ǀ 40	0.00853	MYTQGGSPMYGADGLWLNLFRGFLNVAWIIAAFLRCLYACVQGA/ MYTQGGSPMYGADGLWLNLFRGFLNVAWIIAAFLRCLYAC
TGME49_054620	QAR	GSIKGLSLKKHLGR	37	29 ǀ 26	0.00002	MARCSGTQPPLFVVLTLIGGDFAKATNAS/ MARCSGTQPPLFVVLTLIGGDFAKAT
TGME49_009980	PVQ	GTLGPDVADTAGEPVVLQVAR	31	27	0.00588	MRTSVALFFAAAGGISLVCAPQVSLAA
TGME49_077270	SLR	GVDADTEKR	33	26	0.00437	MWLPVHVPLLLVFGVSLSLPHGSLGT
TGME49_119920	RAG	IAQAFVPRPLAPLTR	21	26	0.00267	MLATRRVFSASPRLVCARAGIAQAF
TGME49_021210	GVR	KAYMDIDIDGEHAGR	24	18	0.00000	MKLVLLFLALAVSGAVAE
TGME49_086000	FSL	SAPLLVVRI	38	49 ǀ 39	0.00251	METASTFRSRQCGLVSKGVAWRAVSLTLRPLILIFSLSAPLLVVRIPAV/ METASTFRSRQCGLVSKGVAWRAVSLTLRPLILIFSLSA
TGME49_003310	ATA	SDDELMSR	30	27	0.00005	MARHAIFFALCVLGLVAAALPQFATAA
TGME49_108080	DLA	SGTPHVAR	39	26	0.00173	MATKLARLATWLVLVGCLLWRAGAV
TGME49_027620	AEF	SGVVNQGPVDVPFSGKPLDER	27	25	0.00029	MFAVKHCLLVVAVGALVNVSVRAA
TGME49_044280	SAA	SSVAAEQR	39	31	0.00056	MEICQLPRVRHLRMMVAVALGALLFLSICGE
TGME49_015280	RFA	SVAHAQTAASEAEAATKVPDFR	22	36 ǀ 16	0.00153	MLSSALRSVRPAASAASRRFASVAHAQTAASEAEAA/ MLSSALRSVRPAASAA
TGME49_108080	NSR	TNDLASGTPHVAR	34	26	0.00023	MATKLARLATWLVLVGCLLWRAGAV

Table 7:14 The table lists PSMs (for *T. gondii* dataset), from the frayed database, for correcting the signal peptide cleavage site. For each peptide we can evaluate the preceding residues (Pre-AA) and compare the peptide starting position (PS) with the signal peptide cleavage site (SP site).

Accession	Pre AA	Peptide	PS	SP site	eFDR	SignalPeptides
TGME49_027620	VRA	AEFSGVVNQGPVDVPFSGKPLDER	24	24	0.0001480000	MFAVKHCLLVVAVGALVNVSVRAA
TGME49_070250	AYA	AEGGDNQSSAVSDR	25	25	0.0000032100	MVRVSAIVGAAASVFVCLSAGAYAA
TGME49_100100	TDA	AEPDSDATPGLRPQPSPR	47	47	0.0019616880	MTKRAGLPLGRAFLVLILLSAADSLFFSSFPRSALQLFSSVLFTDAA/ MTKRAGLPLGRAFLVLILLSAADSL
TGME49_037880	VRG	APDQAQAAVSDKESESR	46	46 | 38	0.0000193000	MLHHSVCFCSARGFIPAMRKFTIVAVVVFALLGCVSWQESAFVRGA/ MLHHSVCFCSARGFIPAMRKFTIVAVVVFALLGCVSWQ
TGME49_009980	SLA	APVQGTLGPDVADTAGEPVVLQVAR	27	27	0.0032763500	MRTSVALFFAAAGGISLVCAPQVSLAA
TGME49_003310	ATA	ATASDDELMSR	27	27	0.0000296000	MARHAIFFALCVLGLVAAALPQFATAA
TGME49_073320	CSA	DTPSLDSDPSASPLR	19	19	0.0018960420	MKVLFFLGLLAGGLLCSAD
TGME49_089800	AAA	FQAVIELAPGQKR	47	47	0.0018700460	MSVSQLTASRRSSRGSRGLCSCPRSFLLFALGLGLASLALTPPAAAF
TGME49_089800	AAA	FQAVIELAPGQKRCVGEQLS	47	47	0.0025965140	MSVSQLTASRRSSRGSRGLCSCPRSFLLFALGLGLASLALTPPAAAF
TGME49_064080	SYG	FVSPGLIR	30	30	0.0081905650	MEMHPRNAGRKTLLALALFMATSIASSYGF
TGME49_110790	IEA	GETPEETEAVVAATEQGAAEVDEATDEHEEDDDDDHR	28	28	0.0000000000	MTRGFAFCLLFLALFGLFSVAPHSIEAG
TGME49_047520	VTA	SDQKQGSQNPAGGKGGSGPHGGR	36	36	0.0000611000	MKTETRQRSRGGGKRLSLCVVFALVSISSVSFVTAS
TGME49_001390	ATA	SHGTTFQDAGAR	23	23	0.0002160000	MKTLHLLLFAVVAIALCGLATAS
TGME49_068790	VSC	SIWRPQGTPEVGSLGHDAADTAAAEAAR	47	47 | 40	0.0003170000	MMIRISGHVCGGGRPGIPRWKSVLRKYIASFCFFVACKTWPGFVSCS/ MMIRISGHVCGGGRPGIPRWKSVLRKYIASFCFFVACKTW
TGME49_079100	GLG	SQMSDSVGR	28	28 | 24	0.0003430000	MWRIWRCRLSFLFATGCLLGALTAGLGS/ MWRIWRCRLSFLFATGCLLGALTA
TGME49_044560	VAA	TETDAAEPLTAEEAPR	44	44 | 25	0.0000025500	MSPAGRRTPKKLAFAALLLGVSVACTSSFFSASVSPSALWVAAT/ MSPAGRRTPKKLAFAALLLGVSVAC
TGME49_004530	VNG	VSEGVVVPVR	23	23	0.0006060000	MQLKKLSVVSITLLGLFKFVNGV
TGME49_070240	LSQ	RVPELPEVESFDEVGTGAR	32	31 | 26	0.0003336640	MDCGQCRRQLHAAGVLGLFVTLATATVGLSQ/ MDCGQCRRQLHAAGVLGLFVTLATAT

Table 7:15 The table lists PSMs (for *T. gondii* dataset) from the frayed database, for confirming the signal peptide cleavage site. For each peptide we can evaluate the preceding residues (Pre-AA) and compare the peptide starting position (PS) with the signal peptide cleavage site (SP site).

Table 7:16 The table here below shows the list of the N-terminal peptides (for *T. gondii* dataset) containing methionine at the N-terminus. It shows the sequence database entry and peptide identified in the first two columns. The third column "Start" represents the position of peptide within protein, while the fourth and fifth columns represent the eFDR score and the signal peptide (SP) sequence if predicted.

Accession	Peptide	Start	eFDR	SP
glimmerhmm|TGME49_chrIb_gene.251_rev_comp_3216	METNHSGPEAAR	1	0.000137	N\A
glimmerhmm|TGME49_chrIb_gene.32_forward_348	MQPEEFAAAAR	1	0.0041596	N\A
glimmerhmm|TGME49_chrIX_gene.169_forward_1506	MEGGAEHVER	1	0.000337	N\A
glimmerhmm|TGME49_chrIX_gene.726_rev_comp_1023	MDPTVSLAAGAEPVAGEKR	32	0.00028	N\A
glimmerhmm|TGME49_chrVI_gene.497_forward_1353	MDASKKSEKDGAPAAPGVSDIELISNR	1	0.0052851	N\A
glimmerhmm|TGME49_chrVIIa_gene.235_forward_1653	MEKLPTIILPGGKAVDETPLSGR	1	0.0044926	N\A
glimmerhmm|TGME49_chrVIIa_gene.500_rev_comp_918	MFLTYVVRPGEAPEGR	1	0.0032259	N\A
TGGT1_028510	MMMNNDPSPAR	14	0.0033403	N\A
TGME49_002370	MNIATDEFGNPFIILR	1	0.0024026	N\A
TGME49_002500	MYFTYVVR	1	0.0077205	N\A
TGME49_002500	MYFTYVVRPGEAPEGR	1	0.0045446	N\A
TGME49_005440	MIRPQGPVLVLKQNTKR	1	0.000163	N\A
TGME49_005470	MVNFSVEQMR	1	0.0020068	N\A
TGME49_009290	MEQPKLAKVEKVLGR	1	0.0042309	N\A
TGME49_010840	MQALNVQVKEAFR	1	0.000745	N\A
TGME49_011670	MFDDEFGEAFDPR	1	0.0019586	N\A
TGME49_014440	MNTELLSLTDEPVILVR	1	0.0000557	N\A
TGME49_016410	METLDEEKAEALLR	1	0.000464	N\A
TGME49_017570	MEIDLLHPDPKVEASKHKLKR	1	0.000715	N\A
TGME49_018780	MEEADLSSLAATER	1	0.0015695	N\A
TGME49_021470	MFNPNATMDWIR	64	0.000485	N\A
TGME49_021630	MPPLEFEESFEV	2	0.0091932	N\A
TGME49_025310	MDAQTASFFKQLR	14	0.000675	N\A

TGME49_026550	MIGSEEFWKTEADAPLLNR	1	0.000577	N\A
TGME49_029250	MVSSELLWQCVR	1	0.0025516	N\A
TGME49_031850	MKSSAEIR	1	0.000576	N\A
TGME49_032030	MQTDILEEHEQLMR	69	0.0068078	N\A
TGME49_033680	MDTPTLDEAMTDSR	1	0.0021405	N\A
TGME49_034450	MNVLADCLKTLVNAEKR	69	0.0000054	N\A
TGME49_036570	MKAKMSHEALTETAR	1	0.0000133	N\A
TGME49_036950	MNVLAYGTAEQR	1	0.000261	N\A
TGME49_037140	MDQAATSAAASQR	1	0.0013254	N\A
TGME49_040700	MEDHSQPR	1	0.0032165	N\A
TGME49_042660	MIMEHDQEKLLDEASAVVKEQAR	1	0.0089405	N\A
TGME49_042730	MLAAAADANALSAAATKSR	1	0.0052638	N\A
TGME49_045460	MPVHQKKR	18	0.0018426	N\A
TGME49_047460	MLEAKLQHASVLR	1	0.0041742	N\A
TGME49_048390	MKFSSQVSSSR	1	0.000562	N\A
TGME49_049250	MKKSGTKQPVR	1	0.000214	N\A
TGME49_053820	MEGAKR	1	0.0080879	N\A
TGME49_054140	MNVGGGGMGR	36	0.0033611	N\A
TGME49_054900	METVIGIR	1	0.000313	N\A
TGME49_055340	MEEAAFSMVR	1	0.000905	N\A
TGME49_055900	MNVFEQYNQR	1	0.0031044	N\A
TGME49_057090	MEEELTPEILAAR	1	0.0022829	N\A
TGME49_057530	MESTEATMVER	1	0.0032109	N\A
TGME49_058720	MIEVILNDR	1	0.00063	N\A
TGME49_059240	MQNDEGR	1	0.000831	N\A
TGME49_061030	MNPLSAALAAKPR	1	0.0046171	N\A
TGME49_062670	MKADPTLQQKISQY	1	0.0000098	N\A
TGME49_062670	MKADPTLQQKISQYQVVGR	1	0.0016691	N\A
TGME49_062690	MVKLLKSGR	1	0.0000231	N\A
TGME49_062980	MNSADAASRPEAEGASGR	1	0.000838	N\A
TGME49_065180	MDAVMVVHQLQR	1	0.000329	N\A
TGME49_067420	MTSGEEEDFYLR	1	0.0000436	N\A
TGME49_068850	MVAIKDITAR	32	0.000969	N\A
TGME49_070830	MDEKIVALNPNR	1	0.0024105	N\A
TGME49_071440	MEKLVVLR	1	0.00019	N\A
TGME49_073950	MLWVDKHAPR	1	0.000266	N\A
TGME49_078540	MEQPGHPGSSVAPASGR	1	0.000483	N\A
TGME49_089210	MDVEVTEEAQSR	1	0.0000663	N\A
TGME49_089830	MRPLFLMGH	1	0.000574	N\A
TGME49_089830	MRPLFLMGHAR	1	0.0051979	N\A

TGME49_093740	MDVAEEPQIQATADR	10	0.0015799	N\A
TGME49_097970	MLGTSSAVAAALLEGGR	13	0.0000179	N\A
TGME49_098970	MDPTLLLQEPLDIVR	1	0.0000164	N\A
TGME49_099030	MEPGELIVHR	1	0.000451	N\A
TGME49_100140	MKLLTPKDDVR	1	0.0021016	N\A
TGME49_107810	MWSIFAPEALTSTSEPATKPAGGATR	1	0.0014666	N\A
TGME49_111240	MYFGSFPFGDDMR	1	0.000338	N\A
TGME49_111310	MEPLADPLHDR	1	0.002797	N\A
TGME49_115150	MKEAVAVPLASVEEKER	1	0.000145	N\A
TGME49_118230	MLANKLGIQDVGAQLTGKSVLIR	1	0.000261	N\A
TGME49_120600	MEDQIQR	83	0.000253	MGVSSSKVFGWGWFSLHSRARSK
TGME49_chrIX-0R_2502455-2502610-156	MFALQDAEQLQLQR	1	0.0000048	N\A
TGME49_chrVIII-2F_1006917-1007049-135	MLLVGMTLVLLR	1	0.0035475	N\A
TGME49_chrXI-2F_578328-578976-651	MSLQTPEASAAASGTQSRNFFCS	1	0.0063973	N\A
TGME49_chrXII-1R_6711199-6711621-423	MDTQNDVESAGR	100	0.000134	N\A
TGME49_chrXII-2F_6249996-6250131-138	MKNEFLGIR	1	0.0085351	N\A

Table 7:17 The table, here below, from *T. gondii* dataset shows candidate N-terminal peptides with NME pattern/ frequency (preAA), their start position within protein, eFDR score and signal peptide (SP if predicted).

Accession	Pre AA	Peptide	Start	eFDR	SP
dna.fa_293.lst-chr_X-gene_4174	M	AVKVLVPVAHDSEEIEAVSIIDTLR	2	0.001133054	N\A
dna.fa_31.lst-chr_II-gene_4819	M	STAATESDLDPTKGEGLFVTLTSGFSKAR	2	0.006044464	N\A
dna.fa_352.lst-chr_XI-gene_5873	M	SSLSVAAASPLAAAKSQWDAR	2	0.000092	N\A
dna.fa_99.lst-chr_V-gene_6989	M	VAAGVSHGNR	2	0.0000156	N\A
glimmerhmm|TGME49_chrIa_gene.117_rev_comp_651	M	PELSTADAGVAVKENER	2	1.64E-08	N\A
glimmerhmm|TGME49_chrIa_gene.295_rev_comp_495	M	AAKQQER	2	0.004290744	N\A
glimmerhmm|TGME49_chrIb_gene.302_forward_681	M	ADAQVHSPQ	2	0.007640813	N\A
glimmerhmm|TGME49_chrII_gene.226_forward_357	M	SKLMKGGLEGEEQR	2	0.00000695	N\A
glimmerhmm|TGME49_chrIII_gene.37_rev_comp_2334	M	AAEAGKKKSEPLSPDEVTDLFR	2	0.004417482	N\A
glimmerhmm|TGME49_chrIX_gene.342_forward_654	M	ADILQENFQDLVHSPGGGR	2	0.003105129	N\A
glimmerhmm|TGME49_chrIX_gene.444_rev_comp_1527	M	AAAAPAVGGGIR	2	0.000636	N\A
glimmerhmm|TGME49_chrIX_gene.742_rev_comp_612	M	TFKKVVVIDCQGHLLGR	2	0.000000833	N\A
glimmerhmm|TGME49_chrIX_gene.782_forward_285	M	APKKTIVKKTKAKKDPNAPKRPLSAFIFFSKDKR	2	0.000119	N\A
glimmerhmm|TGME49_chrIX_gene.958_forward_4518	M	TDSNTNPALKFQR	2	0.002947293	N\A
glimmerhmm|TGME49_chrVI_gene.35_forward_306	M	SGIAVGLKR	2	0.002728356	N\A
glimmerhmm|TGME49_chrVIIa_gene.178_forward_1494	M	ASTKSGALPLFWSAAELAANPR	2	0.001135946	N\A
glimmerhmm|TGME49_chrVIIb_ge	M	VSKNNVLPNVHLHKWWQR	2	0.005501076	N\A

ne.162_rev_comp_639					
glimmerhmm|TGME49_chrVIIb_gene.261_forward_810	M	SGAASATTPPLAAQVQALLQAPELR	2	0.009711108	N\A
glimmerhmm|TGME49_chrVIIb_gene.321_rev_comp_249	M	ATSAKKSPSDFLQKVIGQR	2	0.001282555	N\A
glimmerhmm|TGME49_chrVIII_gene.304_forward_990	M	APALVQR	2	0.003985192	N\A
glimmerhmm|TGME49_chrXII_gene.833_rev_comp_2007	M	AVDSSNSATGPMR	2	0.00086	N\A
TGME49_002870	M	PSAEGNAATQGASYASMKVQELKDLLSQR	2	0.0000305	N\A
TGME49_002980	M	GTLKAPDRLR	2	0.008029858	N\A
TGME49_003390	M	AAPHER	2	0.001810441	N\A
TGME49_003450	M	TSKPESPQR	2	0.009114309	N\A
TGME49_003810	M	VTKAGSPSEDGPSR	2	0.002567807	N\A
TGME49_005320	M	AATVLATETQPR	2	0.0000632	N\A
TGME49_005470	M	VNFSVEQMR	2	0.000165	N\A
TGME49_007770	M	TSVTAVASGSPPAADDSAKKLEELAAR	2	0.002342724	N\A
TGME49_009030	MA	DEEVQALVVDNGSGNVKAGVAGDDAPR	3	0.008345375	N\A
TGME49_009140	M	SVVNVTNIR	2	0.001936967	N\A
TGME49_009290	TFM	ADETDLAGR	31	0.007815744	N\A
TGME49_009910	M	SGKGPAQKSQAAKKTAGKSLGPR	2	0.000148	N\A
TGME49_012290	M	SNPAYLYETPLETR	2	0.003168262	N\A
TGME49_013350	M	ADAGDAAANQPKR	2	0.000192	N\A
TGME49_013350	M	ADAGDAAANQPKRR	2	0.005555759	N\A
TGME49_013410	M	ATVTPVNPKPFLTSLTGR	2	0.0000319	N\A
TGME49_014260	M	AQKGHTDAEAPDVR	2	0.0000359	N\A
TGME49_014350	M	APKKKEQAEEKILLGR	2	0.00118682	N\A
TGME49_015470	M	SKLSTDGLKKAIGEILEGSR	2	0.00000192	N\A
TGME49_015950	M	AQSATTQLDSSAHR	2	0.0000158	N\A
TGME49_016000	M	SDAGTPPAVQGELSQPQER	2	0.006309498	N\A
TGME49_016050	M	AHKTAGDPGR	2	0.001049028	N\A
TGME49_016260	MA	PSAATSAPQTPAGSTEAR	3	0.000313	N\A
TGME49_016450	M	AGTGSGYDLSVSTFSPDG	2	0.002011987	N\A

		R			
TGME49_016790	M	APKQKKETEAVDDAR	2	0.000249	N\A
TGME49_016810	M	PLASTESAAPPESDR	2	0.000497	N\A
TGME49_016860	M	TTLEQNPADELVDYEEDEQNDAKEKGVEDVVVGR	2	0.000917	N\A
TGME49_017460	M	ALSPGAVLR	2	0.000506	MALSPGAVLRITRLASLPPLSTVTDVVLAY
TGME49_018210	MS	SVKVAVR	3	0.002393191	N\A
TGME49_018410	M	AGPKGKSDKR	2	0.000051	N\A
TGME49_018820	M	ADVAAAPAPNATSQAANSTEGDAAAGSAR	2	0.004772766	N\A
TGME49_019690	M	AKPNDLAGLEKALNKNDKIDLAR	2	0.00000814	N\A
TGME49_019690	M	AKPNDLAGLEKALNKNDKIDLARTDTFVER	2	0.000581	N\A
TGME49_019800	M	AAQAAKADAALAH	2	0.000168	N\A
TGME49_019800	M	AAQAAKADAALAHAAAASR	2	0.0000288	N\A
TGME49_019800	MA	AQAAKADAALAHAAAASRD	3	0.004518653	N\A
TGME49_019850	M	ATNSDVSVLSAEEQR	2	0.00000535	N\A
TGME49_020140	M	AGPPAVSEQHR	2	0.000243	N\A
TGME49_020400	M	ASGMGVDENCVAR	2	0.0000184	N\A
TGME49_021950	M	SASLLEHLR	2	0.008806594	N\A
TGME49_022970	M	STVNPADAVGEAKPGPEVTVEFVQAIAR	2	0.004839344	N\A
TGME49_024900	M	AAPSGKR	2	0.002228297	N\A
TGME49_025050	M	AEQSKVKDLTLSAFGR	2	0.000000285	N\A
TGME49_025080	M	SIQVSNNQDFQHILR	2	0.0000973	N\A
TGME49_026680	M	ATLAASSAAGPPSSR	2	0.001461294	N\A
TGME49_026970	M	ATADVQTER	2	0.0000403	N\A
TGME49_026980	M	SSWEDEADEILEAEER	2	0.001386147	N\A
TGME49_029250	M	VSSELLWQCVR	2	0.000851	N\A
TGME49_029360	M	PQSKKR	2	0.001490024	N\A
TGME49_029490	M	GGVSKAKGATR	2	0.008511908	N\A
TGME49_029930	M	SIAGVFQSYTQGKGDMDSR	2	0.000393	N\A
TGME49_029990	M	ALAIFGDR	2	0.000854	N\A
TGME49_031140	M	APKEKKTKEQIAAAAAAGSR	2	0.0000297	N\A
TGME49_032030	M	TDETEPQEQMPLPEPPES	2	0.001496409	N\A

		ITQR			
TGME49_032230	M	AKKAKKSGSEGINSR	2	0.000000516	N\A
TGME49_032410	M	SQPVFASPLNVEKR	2	0.0000083	N\A
TGME49_032550	M	SAAVDAQAVPLGGQR	2	0.000921	N\A
TGME49_032710	M	AIGKNKR	2	0.006409937	N\A
TGME49_032940	TRM	SCCGGTVAEHEVVLDNTGDMDEMLPDQLIPSVPR	9	0.000513	N\A
TGME49_033200	M	GDSPSPDPPETFR	2	0.001126748	N\A
TGME49_033410	M	VKVKVIHR	2	0.002065637	N\A
TGME49_035470	M	ASKTTSEELKTATALKKR	2	0.001950849	N\A
TGME49_035930	M	AQEEAEDVKMDR	2	0.000322	N\A
TGME49_035970	M	ANATTDHLRPQDLETLDISKLTPLSPDVISR	2	0.000000034	N\A
TGME49_036570	AKM	SHEALTETAR	6	0.003433355	N\A
TGME49_038950	M	ALLTEIAALWVR	2	0.0000182	MALLTEIAALWVRDLLRRTLGG
TGME49_040500	M	ANIDSTAANTNR	2	0.000543	N\A
TGME49_042330	M	ATEAKLFGR	2	0.000877	N\A
TGME49_042340	M	TNLFNTRPKKFGPGSR	2	0.005137833	N\A
TGME49_042380	M	APTIVDAPLIQLLADGYGQYR	2	0.007307383	N\A
TGME49_044650	M	ALVNIPR	2	0.001249253	N\A
TGME49_045610	M	AETALYYQELSR	2	0.001376772	N\A
TGME49_047510	HTM	AASGHPIPELGEFIIANKEKLR	33	6.88E-08	N\A
TGME49_048340	M	AAAAAQAVPEFKLILVGDGGVGKTTLVKR	2	0.0000597	N\A
TGME49_048370	M	SQEQLTEAMR	2	0.002419961	N\A
TGME49_050830	M	APGVTQAEFQR	2	0.000213	N\A
TGME49_051550	M	ASQEEFER	2	0.003222488	N\A
TGME49_051620	M	GIKGLGKFVGDFAPR	2	0.0000146	N\A
TGME49_051690	M	TIDVNLLR	2	0.003146296	N\A
TGME49_051810	M	SDAEDVTFETADAGASH	2	1.34E-08	N\A
TGME49_053700	M	VAADAHPR	2	0.002529116	N\A
TGME49_053730	M	SAAPAAGGAPGDLQALAAQLPLASLLEETLAANPAAIR	2	0.004412633	MSAAPAAGGAPGDLQALAAQLPLASLLEETLAA
TGME49_054120	M	PSIRDEVSFEKR	2	0.004330414	N\A

TGME49_054390	M	SAATSPSGLPEQPATVHADDFR	2	0.002144606	N\A
TGME49_054440	M	APKFDPSEVKYIYLR	2	0.001481504	N\A
TGME49_054520	M	ASSKPGGASKAGVDAVQEISMMAR	2	0.005948032	N\A
TGME49_055890	M	VSLSGTLNGEEALER	2	0.000995	N\A
TGME49_056050	M	VLADNNVFLEELGR	2	0.001670713	N\A
TGME49_057310	M	VDYSKWER	2	0.005911399	N\A
TGME49_057480	M	SSAFTDTTASSIAKTR	2	0.000078	N\A
TGME49_057740	DAM	AATNTIESGTTR	16	0.000404	N\A
TGME49_057750	M	PHAGFTDDILLLDGGLGTHLR	2	0.00000086	N\A
TGME49_058070	M	AEALEAERPEGAFR	2	0.008811688	N\A
TGME49_058170	M	TDNAQTTSAEAGAANPSGEHPSAGAKR	2	0.000208	N\A
TGME49_059630	M	AEQSEVTR	2	0.006316285	N\A
TGME49_059660	M	SSGNLSVSR	2	0.0000946	N\A
TGME49_061240	APM	SGGIKKPHR	33	0.001517565	N\A
TGME49_062480	M	SEVEETLNR	2	0.007857869	N\A
TGME49_062620	M	PADEQQQQLPR	2	0.0000324	N\A
TGME49_062690	M	VKLLKSGR	2	0.000677	N\A
TGME49_062720	M	PKNKGKGGKNR	2	0.00000724	N\A
TGME49_062730	RYM	SFEEAQKASEAAKR	27	0.000149	MKVTTKGLAFALALLFCTRCATAR
TGME49_063080	M	PAPAASGAAAVLSKDIAR	2	0.00000612	N\A
TGME49_063530	PKM	AANAASKFIPLLDR	24	0.0007	N\A
TGME49_063700	M	APKKSAKAATGEEGEAQGSGLGPATR	2	0.000000005	N\A
TGME49_063720	M	AKDAAAGEEKKR	2	0.0000174	N\A
TGME49_063850	M	SYNPSYGGQFQGLNAAR	2	0.000000831	N\A
TGME49_064450	M	APAVLMVAEKPSIAETIAR	2	0.000554	N\A
TGME49_066460	M	SDDKKDDAGEKEHMQLKVR	2	0.002134891	N\A
TGME49_067400	M	APVSTVKR	2	0.000993	N\A
TGME49_067420	M	TSGEEEDFYLR	2	0.00000693	N\A
TGME49_068850	NKM	VAIKDITAR	33	0.000129	N\A
TGME49_073460	M	SGEGQVADAGSLPVEKR	2	0.001042676	N\A
TGME49_073900	M	ANSGINWPGLYR	2	0.000151	N\A

TGME49_075750	M	SGGVMSNKKLQKIMTQPINLIFR	2	0.001599131	N\A
TGME49_075810	M	TVPGMKFSLIPKANR	2	0.0000749	N\A
TGME49_077510	M	ASMSSAEASGALPAAGEHELLQEQQR	2	0.000419	N\A
TGME49_078530	M	SFQDWTPVSWNKTGQR	2	0.000901	N\A
TGME49_078950	M	VLPLTLLR	2	0.000196	N\A
TGME49_079400	M	AETLASEAH	2	0.000093	N\A
TGME49_079400	M	AETLASEAHVAEWADR	2	0.00000545	N\A
TGME49_079430	M	TQRPLDYLDVGEHSQVILR	2	0.000706	N\A
TGME49_079450	M	AVSVEGSQDER	2	0.001969912	N\A
TGME49_080550	M	AGLSVSGVALLDSEGER	2	0.000103	MAGLSVSGVALLDSEGE
TGME49_080750	M	SHTILLVQFSDR	2	0.0000639	N\A
TGME49_085510	M	SGDSVAPHQR	2	0.000548	N\A
TGME49_086750	M	ASNKALGKKMSLLEEELR	2	0.000177	N\A
TGME49_089600	M	ADSSGPGDAR	2	0.000063	N\A
TGME49_089690	M	VCKLGINGFGR	2	0.0000141	N\A
TGME49_090200	M	GAPSPVAAGVAAR	2	0.000236	MGAPSPVAAGVAARTKSPLLTVCVCGGGNSAHAVAA / MGAPSVAAGVAARTKSPLLTV
TGME49_090290	M	ATPQESSGAAAHIDTDLYSR	2	0.003491643	N\A
TGME49_090850	M	ATNLEIDSADVIR	2	0.000425	N\A
TGME49_090890	M	AKKVALVTGGNKGIGFGVTR	2	0.0000347	N\A
TGME49_091330	M	TQSMLDMSLDDIVAAHR	2	0.000147	N\A
TGME49_093580	M	ALAAASSASASSDQKR	2	0.0000157	N\A
TGME49_094800	M	GKEKTHINLVVIGH	2	0.000805	N\A
TGME49_095040	M	ATAGQTDEGDR	2	0.000184	N\A
TGME49_095730	M	ATDSQAPASR	2	0.000273	N\A
TGME49_097060	GNM	AKAKYTLVLIR	15	0.000147	N\A
TGME49_097500	M	SHLLNAPIILLKDGVDTSQGR	2	0.0000353	N\A

TGME49_100040	NKM	AGSTISNHQVSSNR	5	0.000404	N\A
TGME49_104710	M	ASDGDVDTNIEQWKIKR	2	0.000095	N\A
TGME49_104760	M	ATTSLEEEEYMEGERDEGWGEAGEQR	2	0	N\A
TGME49_105030	M	ALDTPATSLAAR	2	0.00736	N\A
TGME49_105050	M	SSVEQKAR	2	0.00071	N\A
TGME49_105160	M	VAKKSAKSAKPKASGKSGKGKKKR	2	0.00002	N\A
TGME49_105290	M	AMDDAEAQR	2	0.00049	N\A
TGME49_105510	M	AVAKLDGKTLPALR	2	0.00676	MAVAKLDGKTLPALRLALCGEGNAVTTWWAAMLLQSAAFEVVAAW
TGME49_105820	M	AATEQKRPQETSISR	2	0.00116	N\A
TGME49_108050	M	AISSALIQQR	2	0.00025	N\A
TGME49_109120	M	ATARPLVSVYKPEDGTASGTSLMPSVFLSPLRPDLVR	2	0	N\A
TGME49_109820	M	VKKGEENPMR	2	0.00313	N\A
TGME49_110030	GTM	AKTGAEQLELHGDTWR	38	0.00298	N\A
TGME49_110070	M	TAYGKVDYWDER	2	0.00006	N\A
TGME49_110640	M	SELKGKNIFLTPDGR	2	0.00257	N\A
TGME49_110860	M	ALVAANAAGAALSVAPADAPSALAQNAR	2	0.00003	MALVAANAAGAALSVAPADAPSALAQ
TGME49_111690	M	TMEGQQDLTVIPPLSHQDADRR	2	0.00001	N\A
TGME49_112200	M	AEGEGVSGEAATSQDTR	2	0.00001	N\A
TGME49_112530	M	GDLDFDEVEKLLDSR	2	0.00052	N\A
TGME49_113100	M	VLAELGEQISGALR	2	0.00062	N\A
TGME49_113260	M	SGTGGAGSGPLGSAGGAR	2	0.00772	N\A
TGME49_113390	M	APTAAALAKKR	2	0.00034	MAPTAAALAKKRLRTRKPKRQLYKSPAGAAKRMAKLRSSITPGTVLILLSGG

					HRGK / MAPTAAALAKKRLRTR
TGME49_113390	MA	PTAAALAKKR	3	0.00203	MAPTAAALAKKRLRTRKPKRQLYKSPAGAAKRMAKLRSSITPGTVLILLSGGHRGK / MAPTAAALAKKRLRTR
TGME49_113560	M	ANEDGETAASKMTYLSPIASPLLDGKSLR	2	0.00005	N\A
TGME49_114070	M	AAAAEKVAYGPEDEAR	2	0	N\A
TGME49_115110	M	AAIAAGAASQAPR	2	0	MAAIAAGAASQAPRSPASLASLSLQQLVGV
TGME49_115110	MA	AIAAGAASQAPR	3	0.00076	MAAIAAGAASQAPRSPASLASLSLQQLVGV
TGME49_115270	M	APKKKGISVNLR	2	0.00931	N\A
TGME49_115610	M	VAKKAKTGPASR	2	0.00004	N\A
TGME49_115780	M	SNVVRPIKLQEQHLR	2	0.00032	N\A
TGME49_118410	M	VSIVNAKADVLR	2	0.00115	N\A
TGME49_118750	M	ATEQIYKQFTSR	2	0.00053	N\A
TGME49_119730	M	AFGSSSSSER	2	0.0006	N\A
TGME49_120020	M	AAAQETAMVVPNGSGLELQNR	2	0.00001	N\A
TGME49_120050	M	AFVKALKNKA	2	0.00049	N\A
TGME49_120050	M	AFVKALKNKAY	2	0.00072	N\A
TGME49_120050	M	AFVKALKNKAYFKR	2	0.00375	N\A
TGME49_120570	M	SGFVFNPNASVFVPGGVSSAPPPPPPASEDPAR	2	0.00049	N\A
TGME49_chrIb-0F_1400812-1401393-582	MA	PPAVTQSPGQR	3	0.00077	N\A
TGME49_chrIII-1F_412433-412749-318	M	SATEAAQALKAKGN	2	0.00324	N\A

TGME49_chrVIIb-0R_4431405-4431836-432	M	ASKQPQTLSAGAVESGR	2	0.00178	N\A
TGME49_chrVIII-0F_2098510-2098746-237	CMK	TLEEKLAAAKELQATKAAR	23	0	N\A
TGME49_chrVIII-1F_2235251-2235471-222	KLM	ELHGEAEDVGR	45	0.00113	N\A

Bibliography

1. Crick FH: **The Complementary Structure of DNA**. *Proc Natl Acad Sci U S A* 1954, **40**(8):756-758.
2. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**(4356):737-738.
3. Crick FH: **On protein synthesis**. *Symp Soc Exp Biol* 1958, **12**:138-163.
4. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ: **General nature of the genetic code for proteins**. *Nature* 1961, **192**:1227-1232.
5. Crick FH, Brenner S, Klug A, Pieczenik G: **A speculation on the origin of protein synthesis**. *Origins of life* 1976, **7**(4):389-397.
6. Hanson AA, Rogan EG, Cavalieri EL: **Synthesis of adducts formed by iodine oxidation of aromatic hydrocarbons in the presence of deoxyribonucleosides and nucleobases**. *Chem Res Toxicol* 1998, **11**(10):1201-1208.
7. Rutledge LR, Durst HF, Wetmore SD: **Computational comparison of the stacking interactions between the aromatic amino acids and the natural or (cationic) methylated nucleobases**. *Phys Chem Chem Phys* 2008, **10**(19):2801-2812.
8. Cysewski P, Szefler B: **Environment influences on the aromatic character of nucleobases and amino acids**. *J Mol Model* 2010, **16**(11):1709-1720.
9. Chalikian TV, Volker J, Plum GE, Breslauer KJ: **A more unified picture for the thermodynamics of nucleic acid duplex melting: a characterization by calorimetric and volumetric techniques**. *Proc Natl Acad Sci U S A* 1999, **96**(14):7853-7858.
10. Patton JT, Spencer E: **Genome replication and packaging of segmented double-stranded RNA viruses**. *Virology* 2000, **277**(2):217-225.
11. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G *et al*: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**(3):241-247.
12. Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W *et al*: **A physical map of the 1-gigabase bread wheat chromosome 3B**. *Science* 2008, **322**(5898):101-104.
13. Burge CB, Karlin S: **Finding the genes in genomic DNA**. *Curr Opin Struct Biol* 1998, **8**(3):346-354.
14. Lake JA: **Origin of the eukaryotic nucleus: eukaryotes and eocytes are genotypically related**. *Canadian journal of microbiology* 1989, **35**(1):109-118.
15. Ribeiro S, Golding GB: **The mosaic nature of the eukaryotic nucleus**. *Mol Biol Evol* 1998, **15**(7):779-788.
16. Dolan MF, Melnitsky H, Margulis L, Kolnicki R: **Motility proteins and the origin of the nucleus**. *The Anatomical record* 2002, **268**(3):290-301.
17. Huber MD, Gerace L: **The size-wise nucleus: nuclear volume control in eukaryotes**. *The Journal of cell biology* 2007, **179**(4):583-584.
18. Archambault J, Friesen JD: **Genetics of eukaryotic RNA polymerases I, II, and III**. *Microbiological reviews* 1993, **57**(3):703-724.
19. Roeder RG: **Nuclear RNA polymerases: role of general initiation factors and cofactors in eukaryotic transcription**. *Methods Enzymol* 1996, **273**:165-171.

20. Cramer P, Armache KJ, Baumli S, Benkert S, Brueckner F, Buchen C, Damsma GE, Dengl S, Geiger SR, Jasiak AJ *et al*: **Structure of eukaryotic RNA polymerases**. *Annual review of biophysics* 2008, **37**:337-352.
21. Black DL: **Mechanisms of alternative pre-messenger RNA splicing**. *Annu Rev Biochem* 2003, **72**:291-336.
22. Jurica MS, Moore MJ: **Pre-mRNA splicing: Awash in a sea of proteins**. *Molecular Cell* 2003, **12**(1):5-14.
23. Crick F: **Split genes and RNA splicing**. *Science* 1979, **204**(4390):264-271.
24. Murray V, Holliday R: **Mechanism for RNA splicing of gene transcripts**. *FEBS Lett* 1979, **106**(1):5-7.
25. Hastings ML, Krainer AR: **Pre-mRNA splicing in the new millennium**. *Curr Opin Cell Biol* 2001, **13**(3):302-309.
26. Oliver JL, Bernaola-Galvan P, Carpena P, Roman-Roldan R: **Isochore chromosome maps of eukaryotic genomes**. *Gene* 2001, **276**(1-2):47-56.
27. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER**. *Nucleic Acids Res* 1999, **27**(23):4636-4641.
28. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders**. *Bioinformatics* 2004, **20**(16):2878-2879.
29. Allen JE, Majoros WH, Pertea M, Salzberg SL: **JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions**. *Genome Biol* 2006, **7 Suppl 1**:S9 1-13.
30. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding**. *Nucleic Acids Res* 1998, **26**(4):1107-1115.
31. Besemer J, Borodovsky M: **Heuristic approach to deriving models for gene finding**. *Nucleic Acids Res* 1999, **27**(19):3911-3920.
32. Salamov AA, Solovyev VV: **Ab initio gene finding in Drosophila genomic DNA**. *Genome Res* 2000, **10**(4):516-522.
33. Solovyev V, Salamov A: **The Gene-Finder computer tools for analysis of human and model organisms genome sequences**. *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:294-302.
34. Schmidt T, Frishman D: **Assignment of isochores for all completely sequenced vertebrate genomes using a consensus**. *Genome Biology* 2008, **9**(6).
35. Zhang W, Wu W, Lin W, Zhou P, Dai L, Zhang Y, Huang J, Zhang D: **Deciphering heterogeneity in pig genome assembly Sscrofa9 by isochore and isochore-like region analyses**. *PLoS One* 2010, **5**(10):e13303.
36. Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P: **IsoFinder: computational prediction of isochores in genome sequences**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W287-292.
37. Rabiner LR: **A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition**. *P Ieee* 1989, **77**(2):257-286.
38. Newberg LA: **Error statistics of hidden Markov model and hidden Boltzmann model results**. *BMC Bioinformatics* 2009, **10**:212.
39. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H: **Interpolated Markov models for eukaryotic gene finding**. *Genomics* 1999, **59**(1):24-31.
40. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling**. *J Mol Biol* 1994, **235**(5):1501-1531.
41. Zhang MQ: **Computational prediction of eukaryotic protein-coding**

genes. *Nature Reviews Genetics* 2002, **3**(9):698-709.

42. Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M *et al*: **EuPathDB: a portal to eukaryotic pathogen databases**. *Nucleic Acids Res* 2010, **38**(Database issue):D415-419.
43. Rombel IT, Sykes KF, Rayner S, Johnston SA: **ORF-FINDER: a vector for high-throughput gene identification**. *Gene* 2002, **282**(1-2):33-41.
44. Stifanic M, Batel R: **Genscan for Arabidopsis is a valuable tool for predicting sponge coding sequences**. *Biologia* 2007, **62**(2):124-127.
45. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**(1):78-94.
46. Birney E, Durbin R: **Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison**. *Ismb-97 - Fifth International Conference on Intelligent Systems for Molecular Biology, Proceedings* 1997:56-64.
47. Sadi MS, Sami AZM, Ahmed IU, Ruhunnabi ABM, Das N: **Bioinformatics: Implementation of a proposed upgraded Smith-Waterman Algorithm for local alignment**. *Cibcb: 2009 Ieee Symposium on Computational Intelligence in Bioinformatics and Computational Biology* 2009:87-91.
48. Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW: **An assessment of gene prediction accuracy in large DNA sequences**. *Genome Res* 2000, **10**(10):1631-1642.
49. Liu Q, Mackey AJ, Roos DS, Pereira FC: **Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction**. *Bioinformatics* 2008, **24**(5):597-605.
50. Stanke M, Morgenstern B: **AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints**. *Nucleic Acids Research* 2005, **33**:W465-W467.
51. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts**. *Nucleic Acids Research* 2006, **34**:W435-W439.
52. Keller O, Kollmar M, Stanke M, Waack S: **A novel hybrid gene prediction method employing protein multiple sequence alignments**. *Bioinformatics* 2011, **27**(6):757-763.
53. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.
54. Pauling L, Corey RB, Branson HR: **The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain**. *Proc Natl Acad Sci U S A* 1951, **37**(4):205-211.
55. Hansen JL, Schmeing TM, Moore PB, Steitz TA: **Structural insights into peptide bond formation**. *P Natl Acad Sci USA* 2002, **99**(18):11670-11675.
56. Payne JW: **Peptides and micro-organisms**. *Advances in microbial physiology* 1976, **13**:55-113.
57. von Heijne G: **The signal peptide**. *J Membr Biol* 1990, **115**(3):195-201.
58. von Heijne G: **The structure of signal peptides from bacterial lipoproteins**. *Protein Eng* 1989, **2**(7):531-534.
59. Paetzel M, Karla A, Strynadka NC, Dalbey RE: **Signal peptidases**. *Chem Rev* 2002, **102**(12):4549-4580.
60. von Heijne G: **A new method for predicting signal sequence cleavage sites**. *Nucleic Acids Res* 1986, **14**(11):4683-4690.
61. Fikes JD, Barkocy-Gallagher GA, Klapper DG, Bassford PJ, Jr.: **Maturation**

of Escherichia coli maltose-binding protein by signal peptidase I in vivo. Sequence requirements for efficient processing and demonstration of an alternate cleavage site. *J Biol Chem* 1990, **265**(6):3417-3423.

62. Nielsen H, Engelbrecht J, von Heijne G, Brunak S: **Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site**. *Proteins* 1996, **24**(2):165-177.
63. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0**. *J Mol Biol* 2004, **340**(4):783-795.
64. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools**. *Nat Protoc* 2007, **2**(4):953-971.
65. Shen HB, Chou KC: **Signal-3L: A 3-layer approach for predicting signal peptides**. *Biochem Biophys Res Commun* 2007, **363**(2):297-303.
66. Chou KC, Shen HB: **Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides**. *Biochem Biophys Res Commun* 2007, **357**(3):633-640.
67. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat Methods* 2011, **8**(10):785-786.
68. Viklund H, Bernsel A, Skwark M, Elofsson A: **SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology**. *Bioinformatics* 2008, **24**(24):2928-2929.
69. Kall L, Krogh A, Sonnhammer EL: **Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W429-432.
70. Jones DT: **Improving the accuracy of transmembrane protein topology prediction using evolutionary information**. *Bioinformatics* 2007, **23**(5):538-544.
71. Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, Meinnel T: **The proteomics of N-terminal methionine cleavage**. *Molecular & Cellular Proteomics* 2006, **5**(12):2336-2349.
72. Cohen P: **The regulation of protein function by multisite phosphorylation--a 25 year update**. *Trends Biochem Sci* 2000, **25**(12):596-601.
73. Glozak MA, Sengupta N, Zhang X, Seto E: **Acetylation and deacetylation of non-histone proteins**. *Gene* 2005, **363**:15-23.
74. Deng L, de la Fuente C, Fu P, Wang L, Donnelly R, Wade JD, Lambert P, Li H, Lee CG, Kashanchi F: **Acetylation of HIV-1 Tat by CBP/P300 increases transcription of integrated HIV-1 genome and enhances binding to core histones**. *Virology* 2000, **277**(2):278-295.
75. Tiwari R, Koffel R, Schneiter R: **An acetylation/deacetylation cycle controls the export of sterols and steroids from S. cerevisiae**. *EMBO J* 2007, **26**(24):5109-5119.
76. Shahbazian MD, Grunstein M: **Functions of site-specific histone acetylation and deacetylation**. *Annu Rev Biochem* 2007, **76**:75-100.
77. Henriksen P, Wagner SA, Weinert BT, Sharma S, Bacinskaja G, Rehman M, Juffer AH, Walther TC, Lisby M, Choudhary C: **Proteome-wide analysis of lysine acetylation suggests its broad regulatory scope in Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2012.
78. Tian L, Fong MP, Wang JJ, Wei NE, Jiang H, Doerge RW, Chen ZJ:

Reversible histone acetylation and deacetylation mediate genome-wide, promoter-dependent and locus-specific changes in gene expression during plant development. *Genetics* 2005, **169**(1):337-345.

79. Kamemura K, Hart GW: **Dynamic interplay between O-glycosylation and O-phosphorylation of nucleocytoplasmic proteins: a new paradigm for metabolic control of signal transduction and transcription**. *Progress in nucleic acid research and molecular biology* 2003, **73**:107-136.
80. Wells L, Vosseller K, Hart GW: **Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc**. *Science* 2001, **291**(5512):2376-2378.
81. Spiro RG: **Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds**. *Glycobiology* 2002, **12**(4):43R-56R.
82. Torres-Padilla ME, Parfitt DE, Kouzarides T, Zernicka-Goetz M: **Histone arginine methylation regulates pluripotency in the early mouse embryo**. *Nature* 2007, **445**(7124):214-218.
83. Uhlmann T, Geoghegan VL, Thomas B, Ridlova G, Trudgian DC, Acuto O: **A method for large-scale identification of protein arginine methylation**. *Mol Cell Proteomics* 2012.
84. Lee DY, Teyssier C, Strahl BD, Stallcup MR: **Role of protein methylation in regulation of transcription**. *Endocr Rev* 2005, **26**(2):147-170.
85. Mann RAM: **Mass spectrometry-based proteomics**. *NATURE* 2003, **422**:198-207.
86. Hochstrasser M: **Ubiquitin, proteasomes, and the regulation of intracellular protein degradation**. *Curr Opin Cell Biol* 1995, **7**(2):215-223.
87. Hochstrasser M: **Ubiquitin-dependent protein degradation**. *Annu Rev Genet* 1996, **30**:405-439.
88. Rabut G: **Introduction to the pervasive role of ubiquitin-dependent protein degradation in cell regulation**. *Seminars in cell & developmental biology* 2012, **23**(5):481.
89. Zhou K, Panisko EA, Magnuson JK, Baker SE, Grigoriev IV: **Proteomics for validation of automated gene model predictions**. *Methods Mol Biol* 2009, **492**:447-452.
90. Delahunty CM, Yates JR, 3rd: **MudPIT: multidimensional protein identification technology**. *Biotechniques* 2007, **43**(5):563, 565, 567 passim.
91. Liu HB, Lin DY, Yates JR: **Multidimensional separations for protein/peptide analysis in the post-genomic era**. *Biotechniques* 2002, **32**(4):898-+.
92. Klose J: **From 2-D electrophoresis to proteomics**. *Electrophoresis* 2009, **30 Suppl 1**:S142-149.
93. Huang T, Wang J, Yu W, He Z: **Protein inference: a review**. *Brief Bioinform* 2012, **13**(5):586-614.
94. Reinders J, Lewandrowski U, Moebius J, Wagner Y, Sickmann A: **Challenges in mass spectrometry-based proteomics**. *Proteomics* 2004, **4**(12):3686-3703.
95. Hochberg Y: **A Sharper Bonferroni Procedure for Multiple Tests of Significance**. *Biometrika* 1988, **75**(4):800-802.
96. Rice WR: **Analyzing Tables of Statistical Tests**. *Evolution* 1989, **43**(1):223-225.
97. Armirotti A, Damonte G: **Achievements and perspectives of top-down proteomics**. *Proteomics* 2010, **10**(20):3566-3576.

98. Gallagher SR: **One-dimensional SDS gel electrophoresis of proteins**. *Curr Protoc Cell Biol* 2007, **Chapter 6**:Unit 6 1.

99. Gorg A, Weiss W, Dunn MJ: **Current two-dimensional electrophoresis technology for proteomics**. *Proteomics* 2004, **4**(12):3665-3685.

100. Beranova-Giorgianni S: **Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations**. *TrAC Trends in Analytical Chemistry* 2003, **22**(5):273-281.

101. Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP *et al*: **The proteome of Toxoplasma gondii: integration with the genome provides novel insights into gene expression and annotation**. *Genome Biol* 2008, **9**(7):R116.

102. Rodriguez J, Gupta N, Smith RD, Pevzner PA: **Does trypsin cut before proline?** *J Proteome Res* 2008, **7**(1):300-305.

103. Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB: **Electrospray Interface for Liquid Chromatographs and Mass Spectrometers**. *Analytical Chemistry* 1985, **57**(3):675-679.

104. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray ionization for mass spectrometry of large biomolecules**. *Science* 1989, **246**(4926):64-71.

105. Morris HR, Paxton T, Dell A, Langhorne J, Berg M, Bordoli RS, Hoyes J, Bateman RH: **High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer**. *Rapid Commun Mass Spectrom* 1996, **10**(8):889-896.

106. Chernushevich IV, Loboda AV, Thomson BA: **An introduction to quadrupole-time-of-flight mass spectrometry**. *J Mass Spectrom* 2001, **36**(8):849-865.

107. Douglas DJ, Frank AJ, Mao D: **Linear ion traps in mass spectrometry**. *Mass Spectrom Rev* 2005, **24**(1):1-29.

108. Marshall AG, Hendrickson CL, Jackson GS: **Fourier transform ion cyclotron resonance mass spectrometry: a primer**. *Mass Spectrom Rev* 1998, **17**(1):1-35.

109. Yost RA, Enke CG: **Selected Ion Fragmentation with a Tandem Quadrupole Mass-Spectrometer**. *Journal of the American Chemical Society* 1978, **100**(7):2274-2275.

110. Frese CK, Altelaar AFM, Hennrich ML, Nolting D, Zeller M, Griep-Raming J, Heck AJR, Mohammed S: **Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos**. *Journal of Proteome Research* 2011, **10**(5):2377-2388.

111. Hu QZ, Noll RJ, Li HY, Makarov A, Hardman M, Cooks RG: **The Orbitrap: a new mass spectrometer**. *Journal of Mass Spectrometry* 2005, **40**(4):430-443.

112. Cotter RJ, Iltchenko S, Wang D, Gundry R: **Tandem Time-of-Flight (TOF/TOF) Mass Spectrometry and Proteomics**. *Journal of the Mass Spectrometry Society of Japan* 2005, **53**(1):7-17.

113. Vestal ML, Campbell JM: **Tandem time-of-flight mass spectrometry**. *Methods Enzymol* 2005, **402**:79-108.

114. Wells JM, McLuckey SA: **Collision-induced dissociation (CID) of peptides and proteins**. *Methods Enzymol* 2005, **402**:148-185.

115. Hayes RN, Gross ML: **Collision-induced dissociation**. *Methods Enzymol* 1990, **193**:237-263.

116. Mikesh LM, Ueberheide B, Chi A, Coon JJ, Syka JE, Shabanowitz J, Hunt DF: **The utility of ETD mass spectrometry in proteomic analysis**. *Biochim Biophys Acta* 2006, **1764**(12):1811-1822.

117. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS: **Matching peptide mass spectra to EST and genomic DNA databases**. *Trends in Biotechnology* 2001, **19**(10):S17-S22.

118. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics* 2004, **20**(9):1466-1467.

119. Eng JK, Fischer B, Grossmann J, Maccoss MJ: **A fast SEQUEST cross correlation algorithm**. *J Proteome Res* 2008, **7**(10):4598-4602.

120. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *J Proteome Res* 2004, **3**(5):958-964.

121. Palagi PM, Lisacek F, Appel RD: **Database interrogation algorithms for identification of proteins in proteomic separations**. *Methods Mol Biol* 2009, **519**:515-531.

122. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**(18):3551-3567.

123. Yates JR, Eng JK, Mccormack AL: **Mining Genomes - Correlating Tandem Mass-Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases**. *Analytical Chemistry* 1995, **67**(18):3202-3210.

124. Chen T, Kao MY, Tepel M, Rush J, Church GM: **A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry**. *Journal of Computational Biology* 2001, **8**(3):325-337.

125. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA: **De novo peptide sequencing via tandem mass spectrometry**. *J Comput Biol* 1999, **6**(3-4):327-342.

126. Lu B, Chen T: **A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry**. *J Comput Biol* 2003, **10**(1):1-12.

127. Han YH, Ma B, Zhang KZ: **SPIDER: Software for protein identification from sequence tags with De Novo sequencing error**. *2004 Ieee Computational Systems Bioinformatics Conference, Proceedings* 2004:206-215

128. Johnson RS, Taylor JA: **Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry**. *Methods Mol Biol* 2000, **146**:41-61.

129. Kim S, Gupta N, Bandeira N, Pevzner PA: **Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra**. *Mol Cell Proteomics* 2009, **8**(1):53-69.

130. Taylor JA, Johnson RS: **Sequence database searches via de novo peptide sequencing by tandem mass spectrometry**. *Rapid Commun Mass Spectrom* 1997, **11**(9):1067-1075.

131. Craig R, Cortens JC, Fenyo D, Beavis RC: **Using annotated peptide mass spectrum libraries for protein identification**. *Journal of Proteome Research* 2006, **5**(8):1843-1849.

132. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ: **Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries**. *Anal Chem* 2006, **78**(16):5678-5684.

133. Wisniewski JR: **Mass spectrometry-based proteomics: principles, perspectives, and challenges**. *Arch Pathol Lab Med* 2008, **132**(10):1566-

1569.

134. Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS: **Rapid and accurate peptide identification from tandem mass spectra**. *J Proteome Res* 2008, **7**(7):3022-3027.

135. Pappin DJ, Hojrup P, Bleasby AJ: **Rapid identification of proteins by peptide-mass fingerprinting**. *Curr Biol* 1993, **3**(6):327-332.

136. Thiede B, Hohenwarter W, Krah A, Mattow J, Schmid M, Schmidt F, Jungblut PR: **Peptide mass fingerprinting**. *Methods* 2005, **35**(3):237-247.

137. Craig R, Cortens JP, Beavis RC: **The use of proteotypic peptide libraries for protein identification**. *Rapid Commun Mass Spectrom* 2005, **19**(13):1844-1850.

138. Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database**. *Anal Chem* 1995, **67**(8):1426-1436.

139. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data**. *J Proteome Res* 2004, **3**(6):1234-1242.

140. Muth T, Vaudel M, Barsnes H, Martens L, Sickmann A: **XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results**. *Proteomics* 2010, **10**(7):1522-1524.

141. Ahrne E, Muller M, Lisacek F: **Unrestricted identification of modified proteins using MS/MS**. *Proteomics* 2010, **10**(4):671-686.

142. Heller M, Ye M, Michel PE, Morier P, Stalder D, Junger MA, Aebersold R, Reymond F, Rossier JS: **Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides**. *J Proteome Res* 2005, **4**(6):2273-2282.

143. Koskinen VR, Emery PA, Creasy DM, Cottrell JS: **Hierarchical clustering of shotgun proteomics data**. *Mol Cell Proteomics* 2011, **10**(6):M110 003822.

144. Balgley BM, Laudeman T, Yang L, Song T, Lee CS: **Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy**. *Molecular & Cellular Proteomics* 2007, **6**(9):1599-1608.

145. Kall L, Storey JD, MacCoss MJ, Noble WS: **Posterior error probabilities and false discovery rates: two sides of the same coin**. *J Proteome Res* 2008, **7**(1):40-44.

146. Nesvizhskii AI, Vitek O, Aebersold R: **Analysis and validation of proteomic data generated by tandem mass spectrometry**. *Nat Methods* 2007, **4**(10):787-797.

147. Storey JD: **A direct approach to false discovery rates**. *J Roy Stat Soc B* 2002, **64**:479-498.

148. Jones AR, Siepen JA, Hubbard SJ, Paton NW: **Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines**. *Proteomics* 2009, **9**(5):1220-1229.

149. Kall L, Storey JD, Noble WS: **QVALITY: non-parametric estimation of q-values and posterior error probabilities**. *Bioinformatics* 2009, **25**(7):964-966.

150. Efron B, Tibshirani R: **Empirical bayes methods and false discovery rates for microarrays**. *Genet Epidemiol* 2002, **23**(1):70-86.

151. Kall L, Storey JD, Noble WS: **Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry**. *Bioinformatics* 2008, **24**(16):i42-48.

152. Choi H, Nesvizhskii AI: **Semisupervised model-based validation of**

peptide identifications in mass spectrometry-based proteomics. *J Proteome Res* 2008, **7**(1):254-265.

153. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search**. *Anal Chem* 2002, **74**(20):5383-5392.

154. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ: **Semi-supervised learning for peptide identification from shotgun proteomics datasets**. *Nat Methods* 2007, **4**(11):923-925.

155. Reich JG, Drabsch H, Daumler A: **On the statistical assessment of similarities in DNA sequences**. *Nucleic Acids Res* 1984, **12**(13):5529-5543.

156. Pearson WR: **Comparison of methods for searching protein sequence databases**. *Protein Sci* 1995, **4**(6):1145-1160.

157. Blakeley P, Overton IM, Hubbard SJ: **Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies**. *J Proteome Res* 2012.

158. Guan SH, Burlingame AL: **Data Processing Algorithms for Analysis of High Resolution MSMS Spectra of Peptides with Complex Patterns of Posttranslational Modifications**. *Molecular & Cellular Proteomics* 2010, **9**(5):804-810.

159. Mann M, Jensen ON: **Proteomic analysis of post-translational modifications**. *Nat Biotechnol* 2003, **21**(3):255-261.

160. Frank AM: **A ranking-based scoring function for peptide-spectrum matches**. *J Proteome Res* 2009, **8**(5):2241-2252.

161. Frank AM: **Predicting Intensity Ranks of Peptide Fragment Ions**. *Journal of Proteome Research* 2009, **8**(5):2226-2240.

162. Bern M, Goldberg D: **EigenMS: De novo analysis of peptide tandem mass spectra by spectral graph partitioning**. *Research in Computational Molecular Biology, Proceedings* 2005, **3500**:357-372.

163. Bern M, Goldberg D: **De novo analysis of peptide tandem mass spectra by spectral graph partitioning**. *Journal of Computational Biology* 2006, **13**(2):364-378.

164. Fischer B, Roth V, Roos F, Grossmann J, Baginsky S, Widmayer P, Gruissem W, Buhmann JM: **NovoHMM: A hidden Markov model for de novo peptide sequencing**. *Analytical Chemistry* 2005, **77**(22):7265-7273.

165. Mo L, Dutta D, Wan Y, Chen T: **MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry**. *Anal Chem* 2007, **79**(13):4870-4878.

166. Frank A, Pevzner P: **PepNovo: de novo peptide sequencing via probabilistic network modeling**. *Anal Chem* 2005, **77**(4):964-973.

167. Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA: **De novo peptide sequencing and identification with precision mass spectrometry**. *Journal of Proteome Research* 2007, **6**(1):114-123.

168. Jeong K, Kim S, Bandeira N, Pevzner PA: **Gapped Spectral Dictionaries and Their Applications for Database Searches of Tandem Mass Spectra**. *Molecular & Cellular Proteomics* 2011, **10**(6).

169. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G: **PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry**. *Rapid Commun Mass Spectrom* 2003, **17**(20):2337-2342.

170. Tabb DL, Ma ZQ, Martin DB, Ham AJ, Chambers MC: **DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring**. *J Proteome Res* 2008, **7**(9):3838-3846.

171. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V: **InsPecT: identification of posttranslationally modified peptides from tandem mass spectra**. *Anal Chem* 2005, **77**(14):4626-4639.

172. Cox J, Hubner NC, Mann M: **How much peptide sequence information is contained in ion trap tandem mass spectra?** *J Am Soc Mass Spectrom* 2008, **19**(12):1813-1820.

173. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R: **Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data - Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides**. *Molecular & Cellular Proteomics* 2006, **5**(4):652-670.

174. Wu FX, Gagne P, Droit A, Poirier GG: **Quality assessment of peptide tandem mass spectra**. *First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2006), Proceedings, Vol 1* 2006:243-250.

175. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR: **Similarity among tandem mass spectra from proteomic experiments: Eetection, significance, and utility**. *Analytical Chemistry* 2003, **75**(10):2470-2477.

176. Beer I, Barnea E, Ziv T, Admon A: **Improving large-scale proteomics by clustering of mass spectrometry data**. *Proteomics* 2004, **4**(4):950-960.

177. Flikka K, Martens L, Vandekerckhoe J, Gevaert K, Eidhammer I: **Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering**. *Proteomics* 2006, **6**(7):2086-2094.

178. Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ: **Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations**. *Genome Research* 2008, **18**(10):1660-1669.

179. Pandey A, Lewitter F: **Nucleotide sequence databases: a gold mine for biologists**. *Trends in Biochemical Sciences* 1999, **24**(7):276-280.

180. Pandey A, Mann M: **Proteomics to study genes and genomes**. *Nature* 2000, **405**(6788):837-846.

181. Gilbert W: **Why Genes in Pieces**. *Nature* 1978, **271**(5645):501-501.

182. Lebrilla CB, Mahal LK: **Post-translation modifications**. *Current Opinion in Chemical Biology* 2009, **13**(4):373-374.

183. Schmidt U, Im KB, Benzing C, Janjetovic S, Rippe K, Lichter P, Wachsmuth M: **Assembly and mobility of exon-exon junction complexes in living cells**. *Rna* 2009, **15**(5):862-876.

184. Mathe C, Sagot MF, Schiex T, Rouze P: **Current methods of gene prediction, their strengths and weaknesses**. *Nucleic Acids Research* 2002, **30**(19):4103-4117.

185. Brenner SE: **Errors in genome annotation**. *Trends in Genetics* 1999, **15**(4):132-133.

186. Devos D, Valencia A: **Intrinsic errors in genome annotation**. *Trends in Genetics* 2001, **17**(8):429-431.

187. Warren AS, Archuleta J, Feng WC, Setubal JC: **Missing genes in the annotation of prokaryotic genomes**. *BMC Bioinformatics* 2010, **11**.

188. Tanner S, Shen ZX, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry**. *Genome Research* 2007, **17**(2):231-239.

189. Kan Z, Rouchka EC, Gish WR, States DJ: **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs**. *Genome*

Res 2001, **11**(5):889-900.

190. Kan Z, States D, Gish W: **Selecting for functional alternative splices in ESTs**. *Genome Res* 2002, **12**(12):1837-1845.

191. Donlin MJ: **Using the Generic Genome Browser (GBrowse)**. *Curr Protoc Bioinformatics* 2009, **Chapter 9**:Unit 9 9.

192. Podicheti R, Dong Q: **Using WebGBrowse to visualize genome annotation on GBrowse**. *Cold Spring Harb Protoc* 2010, **2010**(3):pdb prot5392.

193. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS: **ToxoDB: accessing the Toxoplasma gondii genome**. *Nucleic Acids Res* 2003, **31**(1):234-236.

194. Kalume DE, Peri S, Reddy R, Zhong J, Okulate M, Kumar N, Pandey A: **Genome annotation of Anopheles gambiae using mass spectrometry-derived data**. *BMC Genomics* 2005, **6**.

195. Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmstrom J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE *et al*: **Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes**. *Plos Biol* 2009, **7**(3):616-627.

196. Pertea M, Mount SM, Salzberg SL: **A computational survey of candidate exonic splicing enhancer motifs in the model plant Arabidopsis thaliana**. *BMC Bioinformatics* 2007, **8**:159.

197. Power KA, McRedmond JP, de Stefani A, Gallagher WM, Gaora PO: **High-Throughput Proteomics Detection of Novel Splice Isoforms in Human Platelets**. *PLoS One* 2009, **4**(3).

198. Sonnenburg S, Schweikert G, Philips P, Behr J, Ratsch G: **Accurate splice site prediction using support vector machines**. *BMC Bioinformatics* 2007, **8 Suppl 10**:S7.

199. Dogan RI, Getoor L, Wilbur WJ, Mount SM: **Features generated for computational splice-site prediction correspond to functional elements**. *BMC Bioinformatics* 2007, **8**:410.

200. Ivashchenko AT, Tauasarova MI, Atambayeva SA: **Exon-intron structure of genes in complete fungal genomes**. *Mol Biol+* 2009, **43**(1):24-31.

201. Novichkov PS, Gelfand MS, Mironov AA: **Prediction of the exon-intron structure by comparison of genomic sequences**. *Mol Biol+* 2000, **34**(2):200-206.

202. Renuse S, Chaerkady R, Pandey A: **Proteogenomics**. *Proteomics* 2011, **11**(4):620-630.

203. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata**. *Nucleic Acids Res* 2010, **38**(Database issue):D346-354.

204. Frishman D: **Protein annotation at genomic scale: The current status**. *Chemical Reviews* 2007, **107**(8):3448-3466.

205. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, Dodt M, Mackowiak SD, Gogol-Doering A, Oenal P *et al*: **De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics**. *Genome Res* 2011, **21**(7):1193-1200.

206. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes**. *Genome Res* 2008, **18**(1):188-196.

207. Robb SM, Ross E, Sanchez Alvarado A: **SmedGD: the Schmidtea**

mediterranea genome database. *Nucleic Acids Res* 2008, **36**(Database issue):D599-606.

208. Castellana NE, Payne SH, Shen ZX, Stanke M, Bafna V, Briggs SP: **Discovery and revision of Arabidopsis genes by proteogenomics**. *P Natl Acad Sci USA* 2008, **105**(52):21034-21038.

209. Castellana NE, Pham V, Arnott D, Lill JR, Bafna V: **Template Proteogenomics: Sequencing Whole Proteins Using an Imperfect Database**. *Molecular & Cellular Proteomics* 2010, **9**(6):1260-1270.

210. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J *et al*: **Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes**. *Genome Res* 2008, **18**(7):1133-1142.

211. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N *et al*: **Towards a proteome-scale map of the human protein-protein interaction network**. *Nature* 2005, **437**(7062):1173-1178.

212. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG *et al*: **Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry**. *Nature* 2002, **419**(6906):537-542.

213. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, de Lichtenberg U, Rinner O *et al*: **A high-quality catalog of the Drosophila melanogaster proteome**. *Nat Biotechnol* 2007, **25**(5):576-583.

214. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S *et al*: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry**. *Genome Biol* 2005, **6**(1):R9.

215. Aikawa M: **Parasitological review. Plasmodium: the fine structure of malarial parasites**. *Exp Parasitol* 1971, **30**(2):284-320.

216. Morrissette NS, Sibley LD: **Cytoskeleton of apicomplexan parasites**. *Microbiology and molecular biology reviews : MMBR* 2002, **66**(1):21-38; table of contents.

217. Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, Tarning J, Lwin KM, Ariey F, Hanpithakpong W, Lee SJ *et al*: **Artemisinin resistance in Plasmodium falciparum malaria**. *The New England journal of medicine* 2009, **361**(5):455-467.

218. Guiguemde WA, Shelat AA, Bouck D, Duffy S, Crowther GJ, Davis PH, Smithson DC, Connelly M, Clark J, Zhu F *et al*: **Chemical genetics of Plasmodium falciparum**. *Nature* 2010, **465**(7296):311-315.

219. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV, Merino EF, Amedeo P *et al*: **Comparative genomics of the neglected human malaria parasite Plasmodium vivax**. *Nature* 2008, **455**(7214):757-763.

220. Cox-Singh J, Davis TM, Lee KS, Shamsul SS, Matusop A, Ratnam S, Rahman HA, Conway DJ, Singh B: **Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening**. *Clin Infect Dis* 2008, **46**(2):165-171.

221. Kochar DK, Das A, Kochar SK, Saxena V, Sirohi P, Garg S, Kochar A, Khatri MP, Gupta V: **Severe Plasmodium vivax malaria: a report on**

serial cases from Bikaner in northwestern India. *Am J Trop Med Hyg* 2009, **80**(2):194-198.

222. Prandota J: **The importance of toxoplasma gondii infection in diseases presenting with headaches. Headaches and aseptic meningitis may be manifestations of the Jarisch-Herxheimer reaction**. *The International journal of neuroscience* 2009, **119**(12):2144-2182.
223. Tenter AM, Heckeroth AR, Weiss LM: **Toxoplasma gondii: from animals to humans**. *Int J Parasitol* 2000, **30**(12-13):1217-1258.
224. Shanmugasundram A, Gonzalez-Galarza FF, Wastling JM, Vasieva O, Jones AR: **Library of Apicomplexan Metabolic Pathways: a manually curated database for metabolic pathways of apicomplexan parasites**. *Nucleic Acids Res* 2013, **41**(Database issue):D706-713.
225. Dubey JP, Schares G, Ortega-Mora LM: **Epidemiology and control of neosporosis and Neospora caninum**. *Clin Microbiol Rev* 2007, **20**(2):323-+.
226. Trees AJ, Davison HC, Innes EA, Wastling JM: **Towards evaluating the economic impact of bovine neosporosis**. *International Journal for Parasitology* 1999, **29**(8):1195-1200.
227. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng MQ, Liu C, Widmer G, Tzipori S *et al*: **Complete genome sequence of the apicomplexan, Cryptosporidium parvum**. *Science* 2004, **304**(5669):441-445.
228. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P *et al*: **PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data**. *Nucleic Acids Res* 2003, **31**(1):212-215.
229. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ *et al*: **ToxoDB: an integrated Toxoplasma gondii database resource**. *Nucleic Acids Res* 2008, **36**(Database issue):D553-556.
230. Khan A, Taylor S, Su C, Mackey AJ, Boyle J, Cole R, Glover D, Tang K, Paulsen IT, Berriman M *et al*: **Composite genome map and recombination parameters derived from three archetypal lineages of Toxoplasma gondii**. *Nucleic Acids Research* 2005, **33**(9):2980-2992.
231. Sanderson SJ, Xia D, Prieto H, Yates J, Heiges M, Kissinger JC, Bromley E, Lal K, Sinden RE, Tomley F *et al*: **Determining the protein repertoire of Cryptosporidium parvum sporozoites**. *Proteomics* 2008, **8**(7):1398-1414.
232. Keeling PJ: **Reduction and compaction in the genome of the apicomplexan parasite Ctyptosporidium parvum**. *Dev Cell* 2004, **6**(5):614-616.
233. Griss J, Cote RG, Gerner C, Hermjakob H, Vizcaino JA: **Published and Perished? The Influence of the Searched Protein Database on the Long-Term Storage of Proteomics Data**. *Molecular & Cellular Proteomics* 2011, **10**(9).
234. Lubec G, Afjehi-Sadat L, Yang JW, John JPP: **Searching for hypothetical proteins: Theory and practice based upon original data and literature**. *Prog Neurobiol* 2005, **77**(1-2):90-127.
235. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ *et al*: **The genome of Cryptosporidium hominis**. *Nature* 2004, **431**(7012):1107-1112.
236. Lal K, Bromley E, Oakes R, Prieto JH, Sanderson SJ, Kurian D, Hunt L, Yates JR, 3rd, Wastling JM, Sinden RE *et al*: **Proteomic comparison of four**

Eimeria tenella life-cycle stages: unsporulated oocyst, sporulated oocyst, sporozoite and second-generation merozoite. *Proteomics* 2009, **9**(19):4566-4576.

237. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL *et al*: **A proteomic view of the Plasmodium falciparum life cycle**. *Nature* 2002, **419**(6906):520-526.
238. Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, Su XZ: **cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome**. *BMC Genomics* 2007, **8**:255.
239. Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Konen-Waisman S, Latham SM, Mourier T, Norton R, Quail MA *et al*: **Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy**. *PLoS Pathog* 2012, **8**(3):e1002567.
240. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing**. *J Roy Stat Soc B Met* 1995, **57**(1):289-300.
241. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P *et al*: **EMBL Nucleotide Sequence Database in 2006**. *Nucleic Acids Res* 2007, **35**(Database issue):D16-20.
242. Bindu Nanduri NW, Mark L. Lawrence, Susan M. Bridges and Shane C. Burgess: **Gene model detection using mass spectrometry**. 2009.
243. Cheng H, Chan WS, Li Z, Wang D, Liu S, Zhou Y: **Small open reading frames: current prediction techniques and future prospect**. *Current protein & peptide science* 2011, **12**(6):503-507.
244. Cohen AM, Rumpel K, Coombs GH, Wastling JM: **Characterisation of global protein expression by two-dimensional electrophoresis and mass spectrometry: proteomics of Toxoplasma gondii**. *Int J Parasitol* 2002, **32**(1):39-51.
245. Dybas JM, Madrid-Aliste CJ, Che FY, Nieves E, Rykunov D, Angeletti RH, Weiss LM, Kim K, Fiser A: **Computational analysis and experimental validation of gene predictions in Toxoplasma gondii**. *PLoS One* 2008, **3**(12):e3899.
246. Carucci DJ: **Plasmodium post-genomics: an update**. *Trends Parasitol* 2004, **20**(12):558-561.
247. Basrai MA, Hieter P, Boeke JD: **Small open reading frames: beautiful needles in the haystack**. *Genome Res* 1997, **7**(8):768-771.
248. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ: **Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics**. *Genome Biol* 2006, **7**(4):R35.
249. Borodovsky M, McIninch J: **Recognition of genes in DNA sequence with ambiguities**. *Biosystems* 1993, **30**(1-3):161-171.
250. Che FY, Madrid-Aliste C, Burd B, Zhang HS, Nieves E, Kim K, Fiser A, Angeletti RH, Weiss LM: **Comprehensive Proteomic Analysis of Membrane Proteins in Toxoplasma gondii**. *Molecular & Cellular Proteomics* 2011, **10**(1).
251. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation**. *Bioinformatics* 2000, **16**(10):944-945.

252. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA: **Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data**. *Bioinformatics* 2012, **28**(4):464-469.

253. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV: **Analysis of evolution of exon-intron structure of eukaryotic genes**. *Briefings in Bioinformatics* 2005, **6**(2):118-134.

254. Cole WG, Chiodo AA, Lamande SR, Janeczko R, Ramirez F, Dahl HH, Chan D, Bateman JF: **A base substitution at a splice site in the COL3A1 gene causes exon skipping and generates abnormal type III procollagen in a patient with Ehlers-Danlos syndrome type IV**. *J Biol Chem* 1990, **265**(28):17070-17077.

255. Gordon RS: **High performance MySQL: Optimization, backups, replication & load balancing.** *Libr J* 2004, **129**(20):152-152.

256. Gordon RS: **MySQL: The complete reference.** *Libr J* 2004, **129**(20):152-152.

257. Sutton J: **PHP and MySQL web development, 2nd edition.** *Libr J* 2003, **128**(20):155-155.

258. Gordon RS: **Web database applications with PHP and MySQL.** *Libr J* 2004, **129**(20):152-152.

259. Link AJ, Robison K, Church GM: **Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12**. *Electrophoresis* 1997, **18**(8):1259-1313.

260. van Baren MJ, Koebbe BC, Brent MR: **Using N-SCAN or TWINSCAN to predict gene structures in genomic DNA sequences**. *Curr Protoc Bioinformatics* 2007, **Chapter 4**:Unit 4 8.

261. McDonald L, Robertson DH, Hurst JL, Beynon RJ: **Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides**. *Nat Methods* 2005, **2**(12):955-957.

262. McDonald L, Beynon RJ: **Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization**. *Nat Protoc* 2006, **1**(4):1790-1798.

263. Staes A, Van Damme P, Helsens K, Demol H, Vandekerckhove J, Gevaert K: **Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC)**. *Proteomics* 2008, **8**(7):1362-1370.

264. Sandra K, Verleysen K, Labeur C, Vanneste L, D'Hondt F, Thomas G, Kas K, Gevaert K, Vandekerckhove J, Sandra P: **Combination of COFRADIC and high temperature-extended column length conventional liquid chromatography: a very efficient way to tackle complex protein samples, such as serum**. *J Sep Sci* 2007, **30**(5):658-668.

265. Van Damme P, Van Damme J, Demol H, Staes A, Vandekerckhove J, Gevaert K: **A review of COFRADIC techniques targeting protein N-terminal acetylation**. *BMC proceedings* 2009, **3 Suppl 6**:S6.

266. Timmer JC, Enoksson M, Wildfang E, Zhu W, Igarashi Y, Denault JB, Ma Y, Dummitt B, Chang YH, Mast AE *et al*: **Profiling constitutive proteolytic events in vivo**. *Biochem J* 2007, **407**(1):41-48.

267. Gevaert K, Van Damme P, Martens L, Vandekerckhove J: **Diagonal reverse-phase chromatography applications in peptide-centric proteomics: ahead of catalogue-omics?** *Anal Biochem* 2005, **345**(1):18-29.

268. Gevaert K, Impens F, Van Damme P, Ghesquiere B, Hanoulle X, Vandekerckhove J: **Applications of diagonal chromatography for proteome-wide characterization of protein modifications and activity-**

based analyses. *FEBS J* 2007, **274**(24):6277-6289.

269. Gevaert K, Van Damme P, Ghesquiere B, Vandekerckhove J: **Protein processing and other modifications analyzed by diagonal peptide chromatography**. *Biochim Biophys Acta* 2006, **1764**(12):1801-1810.

270. Kleifeld O, Doucet A, Prudova A, auf dem Keller U, Gioia M, Kizhakkedathu JN, Overall CM: **Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates**. *Nat Protoc* 2011, **6**(10):1578-1611.

271. Kleifeld O, Doucet A, auf dem Keller U, Prudova A, Schilling O, Kainthan RK, Starr AE, Foster LJ, Kizhakkedathu JN, Overall CM: **Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products**. *Nat Biotechnol* 2010, **28**(3):281-288.

272. Mommen GP, van de Waterbeemd B, Meiring HD, Kersten G, Heck AJ, de Jong AP: **Unbiased selective isolation of protein N-terminal peptides from complex proteome samples using phospho tagging (PTAG) and TiO(2)-based depletion**. *Mol Cell Proteomics* 2012, **11**(9):832-842.

273. Agard NJ, Wells JA: **Methods for the proteomic identification of protease substrates**. *Curr Opin Chem Biol* 2009, **13**(5-6):503-509.

274. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot**. *Methods Mol Biol* 2007, **406**:89-112.

275. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites**. *Int J Neural Syst* 1997, **8**(5-6):581-599.

276. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites**. *Protein Eng* 1997, **10**(1):1-6.

277. Nielsen H, Krogh A: **Prediction of signal peptides and signal anchors by a hidden Markov model**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:122-130.

278. Zhang Z, Henzel WJ: **Signal peptide prediction based on analysis of experimentally verified cleavage sites**. *Protein Sci* 2004, **13**(10):2819-2824.

279. Nugent T, Jones DT: **Detecting pore-lining regions in transmembrane protein sequences**. *BMC Bioinformatics* 2012, **13**:169.

280. Rogic S, Mackworth AK, Ouellette FB: **Evaluation of gene-finding programs on mammalian sequences**. *Genome Res* 2001, **11**(5):817-832.

281. Dormeyer W, Mohammed S, Breukelen B, Krijgsveld J, Heck AJ: **Targeted analysis of protein termini**. *J Proteome Res* 2007, **6**(12):4634-4645.

282. Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S: **Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics**. *Science* 2008, **320**(5878):938-941.

283. Rogers A, Antoshechkin I, Bieri T, Blasiar D, Bastiani C, Canaran P, Chan J, Chen WJ, Davis P, Fernandes J *et al*: **WormBase 2007**. *Nucleic Acids Res* 2008, **36**(Database issue):D612-617.

284. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, de la Cruz N, Duong A, Fang R *et al*: **WormBase 2012: more genomes, more data, new website**. *Nucleic Acids Res* 2012, **40**(Database issue):D735-741.

285. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R *et al*: **WormBase: a comprehensive resource for nematode research**. *Nucleic Acids Res* 2010, **38**(Database issue):D463-467.

286. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE *et al*: **Annotation of the Drosophila melanogaster euchromatic genome: a systematic review**. *Genome Biol* 2002, **3**(12):RESEARCH0083.

287. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE: **Genome annotation assessment in Drosophila melanogaster**. *Genome Research* 2000, **10**(4):483-501.

288. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2012, **40**(Database issue):D13-25.

289. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

290. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.

291. Helbig AO, Gauci S, Raijmakers R, van Breukelen B, Slijper M, Mohammed S, Heck AJ: **Profiling of N-acetylated protein termini provides in-depth insights into the N-terminal nature of the proteome**. *Mol Cell Proteomics* 2010, **9**(5):928-939.

292. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res* 2002, **30**(1):276-280.

293. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.

294. Giglione C, Boularot A, Meinnel T: **Protein N-terminal methionine excision**. *Cell Mol Life Sci* 2004, **61**(12):1455-1474.

295. Giglione C, Vallon O, Meinnel T: **Control of protein life-span by N-terminal methionine excision**. *EMBO J* 2003, **22**(1):13-23.

296. Giglione C, Meinnel T: **Organellar peptide deformylases: universality of the N-terminal methionine cleavage mechanism**. *Trends in plant science* 2001, **6**(12):566-572.

297. Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N, Thurston M: **Open software for biologists: from famine to feast**. *Nat Biotechnol* 2006, **24**(7):801-803.

298. Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, Verstraten R, Adams DJ, Harrow J, Choudhary JS *et al*: **Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome**. *Genome Res* 2011, **21**(5):756-767.

299. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD: **Proteogenomics: needs and roles to be filled by proteomics in genome annotation**. *Brief Funct Genomic Proteomic* 2008, **7**(1):50-62.

300. Armengaud J: **Proteogenomics and systems biology: quest for the ultimate missing parts**. *Expert Rev Proteomics* 2010, **7**(1):65-77.

301. Castellana N, Bafna V: **Proteogenomics to discover the full coding content of genomes: A computational perspective**. *Journal of Proteomics* 2010, **73**(11):2124-2135.

302. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **Mass spectrum**

sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. *Genes Cells* 2012, **17**(8):633-644.

303. Helmy M, Sugiyama N, Tomita M, Ishihama Y: **The Rice Proteogenomics Database OryzaPG-DB: Development, Expansion, and New Features**. *Frontiers in plant science* 2012, **3**:65.

304. Christie-Oleza JA, Miotello G, Armengaud J: **High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade**. *BMC Genomics* 2012, **13**:73.

305. Helmy M, Tomita M, Ishihama Y: **OryzaPG-DB: rice proteome database based on shotgun proteogenomics**. *Bmc Plant Biol* 2011, **11**:63.

306. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson KE: **Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community**. *BMC Bioinformatics* 2012, **13**:42.

307. Murata M: **3-Way Needleman-Wunsch Algorithm**. *Methods in Enzymology* 1990, **183**:365-375.

308. Rose J, Eisenmenger F: **A Fast Unbiased Comparison of Protein Structures by Means of the Needleman-Wunsch Algorithm**. *J Mol Evol* 1991, **32**(4):340-354.

309. Du ZH, Lin F: **Improvement of the Needleman-Wunsch algorithm**. *Rough Sets and Current Trends in Computing* 2004, **3066**:792-797.

310. Huang XQ: **On Global Sequence Alignment**. *Computer Applications in the Biosciences* 1994, **10**(3):227-235.