Development of computational methods for analysing proteomic data for genome annotation

University of Cambridge

Darwin College

A thesis submitted for the degree of

Doctor of Philosophy

Markus Brosch

The Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SA,
United Kingdom.

December 2009

Dedicated to my family

Declaration

This thesis describes work carried out between May 2006 and December 2009 under the supervision of Dr Jyoti Choudhary and Dr Tim Hubbard at the Wellcome Trust Sanger Institute, while member of Darwin College, University of Cambridge. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee. This thesis has been typeset in 12pt font using LaTeX2ε according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Markus Brosch
December 2009.

Summary

Current functional genomics relies on known and characterised genes, but despite significant efforts in the field of genome annotation, accurate identification and elucidation of protein coding gene structures remains challenging. Methods are limited to computational predictions and transcript-level experimental evidence, hence translation cannot be verified. Proteomic mass spectrometry is a method that enables sequencing of gene product fragments, enabling the validation and refinement of existing gene annotation as well as the detection of novel protein coding regions. However, the application of proteomics data to genome annotation is hindered by the lack of suitable tools and methods to achieve automatic data processing and genome mapping at high accuracy and throughput. The main objectives of this work are to address these issues and to demonstrate the applicability in a pilot study that validates and refines annotation of *Mus musculus*.

In the first part of this project I evaluate the scoring schemes of “Mascot”, which is a peptide identification software that is routinely used, for low and high mass accuracy data and show these to be not sufficiently accurate. I develop an alternative scoring method that provides more sensitive peptide identification specifically for high accuracy data, while allowing the user to fix the false discovery rate.

Building upon this, I utilise the machine learning algorithm “Percolator” to further extend my Mascot scoring scheme with a large set of orthogonal scoring features that assess the quality of a peptide-spectrum match. I demonstrate very good sensitivity with this approach and highlight the importance of reliable and robust peptide-spectrum match significance measures.

To close the gap between high throughput peptide identification and large scale genome annotation analysis I introduce a proteogenomics pipeline. A comprehensive database is the central element of this pipeline, enabling the efficient mapping of known and predicted peptides to their genomic loci, each of which is associated with supplemental annotation information such as gene and transcript identifiers. Software scripts allow the creation of automated genome annotation analysis reports.

In the last part of my project the pipeline is applied to a large mouse MS dataset. I show the value and the level of coverage that can be achieved for validating genes and gene structures, while also highlighting the limitations of this technique. Moreover, I show where peptide identifications facilitated the correction of existing annotation, such as re-defining the translated regions or splice boundaries. Moreover, I propose a set of novel genes that are identified by the MS analysis pipeline with high confidence, but largely lack transcriptional or conservational evidence.

Acknowledgements

First and foremost I would like to thank my supervisors Jyoti Choudhary and Tim Hubbard for giving me the opportunity to carry out this project and for all their invaluable advice, support and encouragement. Many thanks also to my thesis committee Richard Durbin, Gos Micklem and John Cottrell for their critical and constructive assessment of my work. I thank the Wellcome Trust for my PhD studentship and Matrix Science for funding conference travel.

I extend my gratitude to Lukas Käll for all the fruitful peptide scoring discussions and beyond, as well as for extending the Percolator software package interfaces, without which parts of my project would not have been possible. Special thanks to John Cottrell and David Creasy for extensive support with Mascot related questions and problems. I am also very grateful for their decision to integrate Mascot Percolator, which was developed as part of this project, into their upcoming Mascot 2.3 release, enabling widespread use of this method. Also, I would also like to thank David Fenyo for X!Tandem related scoring discussions as well as Mario Stanke for help with the Augustus gene prediction software. Many thanks also to Eric Deutsch and Zhi Sun for answering Peptide Atlas related questions as well as providing me with a comprehensive mouse dataset used in this work.

Many thanks to Sajani Swamy, who introduced me to the Mascot software and Parthiban Vijayarangakannan for writing a web application for Mascot Percolator. I would like to express my gratitude to Mercedes Pardo, who “tried” to introduce me to the “web-lab world”, as well as Lu Yu and Mark Collins for their great efforts to provide me with the required mass spec data. I am also grateful to the HAVANA team around Jennifer Harrow, who started to investigate my proteogenomic data. Particular thanks to Felix Kokocinski and Jonathan Warren for extensive help with setting up the distributed annotation server and helping with related problems.

I enjoyed working in Jyoti Choudhary and Tim Hubbard’s research group and thank every member and ex-member for the enjoyable time at the Sanger Insitute. Jyoti Choudhary, Tim Hubbard, James Wright, Mark Collins and Daniel James kindly read the draft of this thesis, providing me with valuable feedback. Thank you.

On a personal note, I want to thank my partner Daniela Wieser for supporting me in my endeavour and for putting up with my ridiculous working hours.

Markus Brosch,
Wellcome Trust Sanger Institute,
December 2009.

Contents

List of Figures

Nomenclature

AMT	Adjusted Mascot Threshold
E-value	Expectation Value
FDR	False Discovery Rate
FP	False Positive
MATH	Mass Accuracy-Based THreshold
MHT	Mascot Homology Threshold
MIT	Mascot Identity Threshold
MMD	Maximum Mass Deviation
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
PEP	Posterior Error Probability
PPM	Parts Per Million
PSM	Peptide Spectrum Match
ROC	Receiver Operating Characteristics
SQL	Structured Query Language
TP	True Positive

Chapter 1

Introduction

Current functional genomics relies on known and characterised genes, but despite significant efforts in the field of genome annotation, accurate identification and elucidation of protein coding gene structures remains challenging. Methods are limited to computational predictions and transcript-level experimental evidence, hence translation cannot be verified. Proteomic mass spectrometry is a method that enables sequencing of gene product fragments, enabling the validation and refinement of existing gene annotation as well as the elucidation of novel protein coding regions.

However, the application of proteomics data to genome annotation is hindered by the lack of suitable tools and methods to achieve automatic data processing and genome mapping at high accuracy and throughput. The main objective of this work is to address these issues and to demonstrate its applicability in a pilot study that validates and refines annotation of *Mus musculus*.

This introduction presents the foundations of the work described in this thesis. Section 1.1 is an introduction to the field of protein mass spectrometry and focusses on the importance of reliable peptide identification methods. Section 1.2 describes available genome annotation strategies with a focus on in-house systems such as Ensembl or Vega. A brief history of using proteomics data for genome annotation is presented in section 1.3. Finally, the outline of my work is described in section 1.4.

1.1 Protein mass spectrometry

Mass spectrometry (MS) has become the method of choice for protein identification and quantification (Aebersold and Mann, 2003; Foster *et al.*, 2006; Patterson and Aebersold, 2003; Washburn *et al.*, 2001). The main reasons for this success include the availability of high-throughput technology coupled with high sensitivity, specificity and a good dynamic range (de Godoy *et al.*, 2006). These advantages are achieved by various separation techniques coupled with high performance MS instrumentation.

In a modern bottom-up LC-MS/MS proteomics experiment (Hunt *et al.*, 1992; McCormack *et al.*, 1997), a complex protein mixture is often separated via gel electrophoresis first to simplify the sample (Shevchenko *et al.*, 1996). Subsequently, proteins are digested with a specific enzyme such as trypsin, generating peptides that are amenable for subsequent MS analysis. To further reduce sample complexity, peptides are separated by liquid chromatographic (LC) systems (Wolters *et al.*, 2001), allowing direct analysis without the need for further fractionation: eluents are ionised, separated by their mass over charge ratios and subsequently registered by the detector. In a tandem MS experiment (MS/MS), low energy collision-induced dissociation is used to fragment the precursor ions, usually along the peptide bonds. Product fragments are measured as mass over charge ratios, which commonly reflect the primary structure of the peptide ion (Biemann, 1988; Roepstorff and Fohlman, 1984). This simplified process is illustrated in figure 1.1.

Today this technology allows researchers to identify complex protein mixtures and enables them to build protein expression landscapes of any biological material (Foster *et al.*, 2006). However, protein sequence coverage varies largely (de Godoy *et al.*, 2006; Simpson *et al.*, 2000) while protein inference can be challenging if identified sequences are shared between different proteins (Nesvizhskii and Aebersold, 2004; Nesvizhskii *et al.*, 2003).

The alternative top-down MS approach allows us to identify and sequence intact

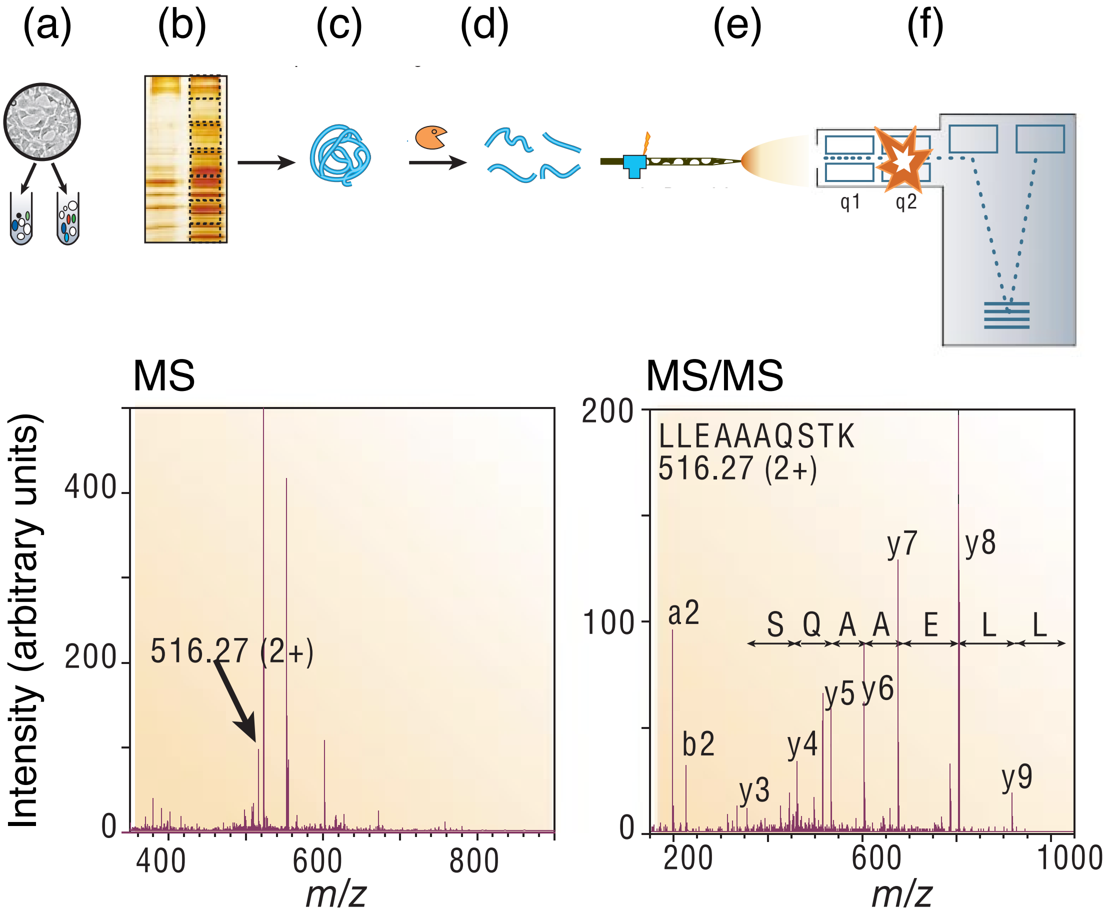


Figure 1.1: Schematic of a generic bottom-up proteomics MS experiment. (a) Sample preparation and fractionation, (b) protein separation via gel-electrophoresis, (c) protein extraction, (d) enzymatic protein digestion, (e) separation of peptides in one or multiple steps of liquid chromatography, followed by ionisation of eluents and (f) tandem mass spectrometry analysis. Here, the mass to charge ratios of the intact peptides are measured, selected peptide ions are fragmented and mass to charge ratios of the product ions are measured. The resulting spectra are recorded accordingly (MS, MS/MS) allowing peptide identification. Adapted from Figure 1 in Aebersold and Mann (2003).

proteins directly and does not limit the analysis to the fraction of detectable enzyme digests (Parks *et al.*, 2007; Roth *et al.*, 2008). However, this method is currently not applicable to complex protein samples in a high throughput fashion. Firstly, there is an insufficiency of efficient whole protein separation techniques and secondly commercially available MS instruments are either limited by efficient fragmentation or by molecular weight restrictions of the analytes (Han *et al.*, 2006).

The most widely used instruments are ion trap mass spectrometers (Douglas *et al.*, 2005), which offer a high data acquisition rate and have generated an enormous amount of data, some of which are available in public repositories (Desiere *et al.*, 2006; Jones *et al.*, 2008; Martens *et al.*, 2005a). Ion trap data is of low resolution and low mass accuracy and therefore the typical rate of confident sequence assignments is low (10-15%) (Elias *et al.*, 2005; Peng *et al.*, 2003).

The recent availability of hybrid-FT mass spectrometers (Hu *et al.*, 2005; Syka *et al.*, 2004) enables high mass resolution (30k-500k) together with very high mass accuracy (in the range of a few parts per million, ppm). On these instruments, throughput and sensitivity is maximised by collecting MS data at a high resolution and accuracy, and MS/MS data is recorded at high speed with low resolution and accuracy (Haas *et al.*, 2006). High resolution spectra enable charge state determination of the precursor ion (Chernushevich *et al.*, 2001; Heeren *et al.*, 2004) and highly restrictive mass tolerance settings lead in database search algorithms to fewer possible peptide candidates because of the limited number of amino acid compositions that fall into a given mass window (see next section). It is expected that the discrimination power of database search engines improves with high accuracy MS data (Clauser *et al.*, 1999; Zubarev, 2006). In chapter 2 of this work I test this hypothesis by evaluating the scoring scheme of two common database search engines with high accuracy data and in chapter 3 I further utilise the discrimination power of these data. For an outline of my work, please refer to section 1.4.

1.1.1 Peptide identification

A large number of computational tools have been developed to support high-throughput peptide and protein identification by automatically assigning sequences to tandem MS spectra (Nesvizhskii *et al.* (2007), table 1). Three types of approaches are used: (a) *de novo* sequencing, (b) database searching and (c) hybrid approaches.

1.1.1.1 *De novo* and hybrid algorithms

De novo algorithms infer the primary sequence directly from the MS/MS spectrum by matching the mass differences between peaks to the masses of corresponding amino acids (Dancik *et al.*, 1999; Taylor and Johnson, 1997). These algorithms do not need *a priori* sequence information and hence can potentially identify protein sequences that are not available in a protein database. However, *de novo* implementations do not yet reach the overall performance of database search algorithms and often only a part of the whole peptide sequence is reliably identified (Mann and Wilm, 1994; Pitzer *et al.*, 2007; Tabb *et al.*, 2003).

High accuracy mass spectrometry circumvents many sequence ambiguities, and *de novo* methods can reach new levels of performance (Frank *et al.*, 2007). Moreover, hybrid algorithms become more important, which build upon the *de novo* algorithms, but compare the generated lists of potential peptides (Bern *et al.*, 2007; Frank and Pevzner, 2005; Kim *et al.*, 2009) or short sequence tags (Tanner *et al.*, 2005) with available protein sequence databases to limit and refine the search results.

With the constant advances in instrument technology and improved algorithms, *de novo* and hybrid methods may have a more important role in the future, however database searching remains the most widely used method for peptide identification.

1.1.1.2 Sequence database search algorithms

Sequence database search algorithms resemble the experimental steps *in silico* (figure 1.2): a protein sequence database is digested into peptides with the same enzyme that is used in the actual experiment, most often trypsin that cuts very specifically after Arginine (R) and Lysine (K) (Olsen *et al.*, 2004; Rodriguez *et al.*, 2007). All peptide sequences (candidates) that match the experimental peptide mass within an allowed maximum mass deviation (MMD) are selected from this *in silico* digested protein sequence database. Each candidate is then further investigated at the MS/MS level by correlating the experimental with the theoretical peptide fragmentation patterns

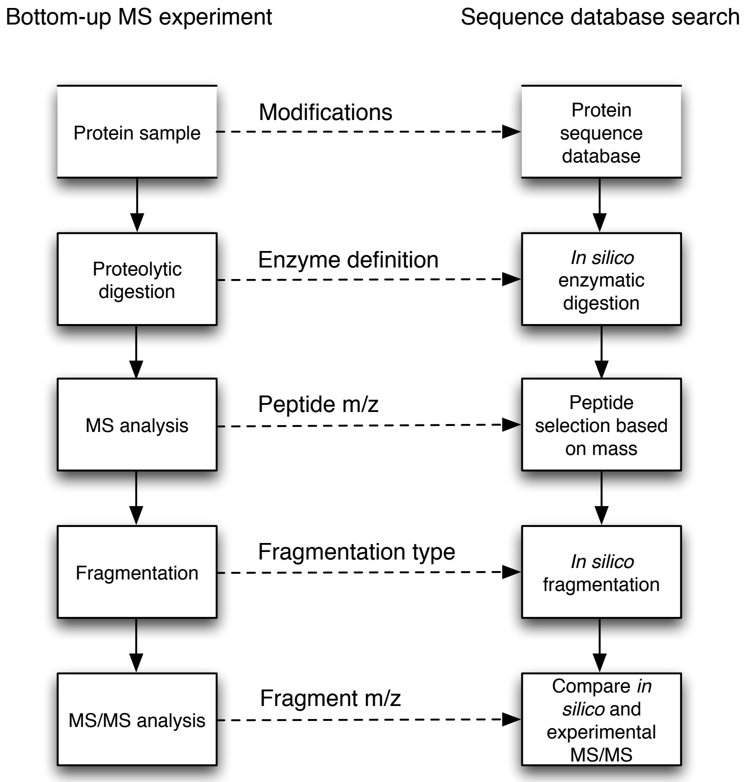


Figure 1.2: Concept of sequence database searching resembles a generic bottom-up MS experiment, as for each stage of the experiment, an *in silico* equivalent component is available.

and scoring the correlation quality (Eng *et al.*, 1994; Kapp *et al.*, 2005; Perkins *et al.*, 1999). It should be noted that the sequence database is usually supplemented with expected experimental contaminant proteins. This avoids spectra that originate from contaminant proteins to incorrectly match to other proteins.

1.1.2 Scoring of peptide identifications

Most of these database search algorithms provide one or more peptide-spectrum match (PSM) scores that correlate with the quality of the match, but are typically hard to interpret and are not associated with any valid statistical meaning. Researchers face the problem of computing identification error rates or PSM significance measures and need to deal with post-processing software that converts search scores into meaningful statistical measures. Therefore, the following sections are focussed on scoring and assessment of database search results, providing a brief overview of common methods, their advantages and disadvantages.

1.1.2.1 Peptide-spectrum match scores and common thresholds

Sequest (Eng *et al.*, 1994) was the first sequence database search algorithm for tandem MS data and is today, together with Mascot (Perkins *et al.*, 1999) one of the most widely used tools for peptide and protein identification. These are representative of the numerous database search algorithms that report for every PSM, a score that reflects the quality of the cross correlation between the experimental and the computed theoretical peptide spectrum. Although Sequest and Mascot scores are fundamentally different in their calculation, they facilitate good relative PSM ranking: all peptide candidates that were matched against an experimental spectrum are ranked according to the PSM score and only the best matches are reported.

Often only the top hit is considered for further investigation and some search engines like X!Tandem (Craig and Beavis, 2004) exclusively report that very best match. However, not all these identifications are correct. Sorting all top hit PSMs (absolute ranking) according to their score enables the selective investigation of the very best matched PSMs. This approach was initially used to aid manual interpretation and validation. As the field of MS-based proteomics moved towards high-throughput methods, researchers started to define empirical score thresholds.

PSMs scoring above these thresholds were accepted and assumed to be correct, while anything else was classified as incorrect. Depending on how well the underlying PSM score discriminates, the correct and incorrect scores overlap significantly (figure 1.3) and therefore thresholding is always a trade-off between sensitivity (fraction of true positive identifications) and the acceptable error rate (fraction of incorrect identifications). Low score thresholds will accept more PSMs at the cost of a higher error rate and on the other hand a high score threshold reduces the error rate at the cost of sensitivity.

Many groups also apply heuristic rules that combine the score threshold with some other validation properties such as charge state, the difference in score to the second best hit, amongst others. The problem with these methods is that the actual error rate remains unknown and the decision of accepting assignments is only based on judgement of an expert. Moreover, results between laboratories or even between experiments cannot be reliably compared, since different search algorithms, protein databases, search parameters, instrumentation and sample complexity require adaptation of acceptance criteria. A recent HUPO study (States *et al.*, 2006) investigated the reproducibility between laboratories. Amongst the 18 laboratories, each had their own criteria of what was considered a high and low confidence protein identification, which were mostly based on simple heuristic rules and score thresholds (States *et al.* (2006), supplementary table 1). It was found that the number of high confidence assignments between two different laboratories could vary by as much as 50%, despite being based on the same data. As a result, many proteomic journals require the validation and assessment of score thresholds, ideally with significance measures such as presented below.

1.1.2.2 Statistical significance measures

The expected error rates associated with individual or sets of PSMs can be reported as standard statistical significance measures. This allows transformation of specific scoring schemes into generic and unified measures, enabling comparability across any experiment in a consistent and easy to interpret format. In this section I discuss and explain commonly used statistical measures that ideally are reported by every database search algorithm or post-processing software; focusing on the false discovery rate (FDR), its derived q-value and the Posterior Error Probability (PEP), also sometimes referred to as local FDR.

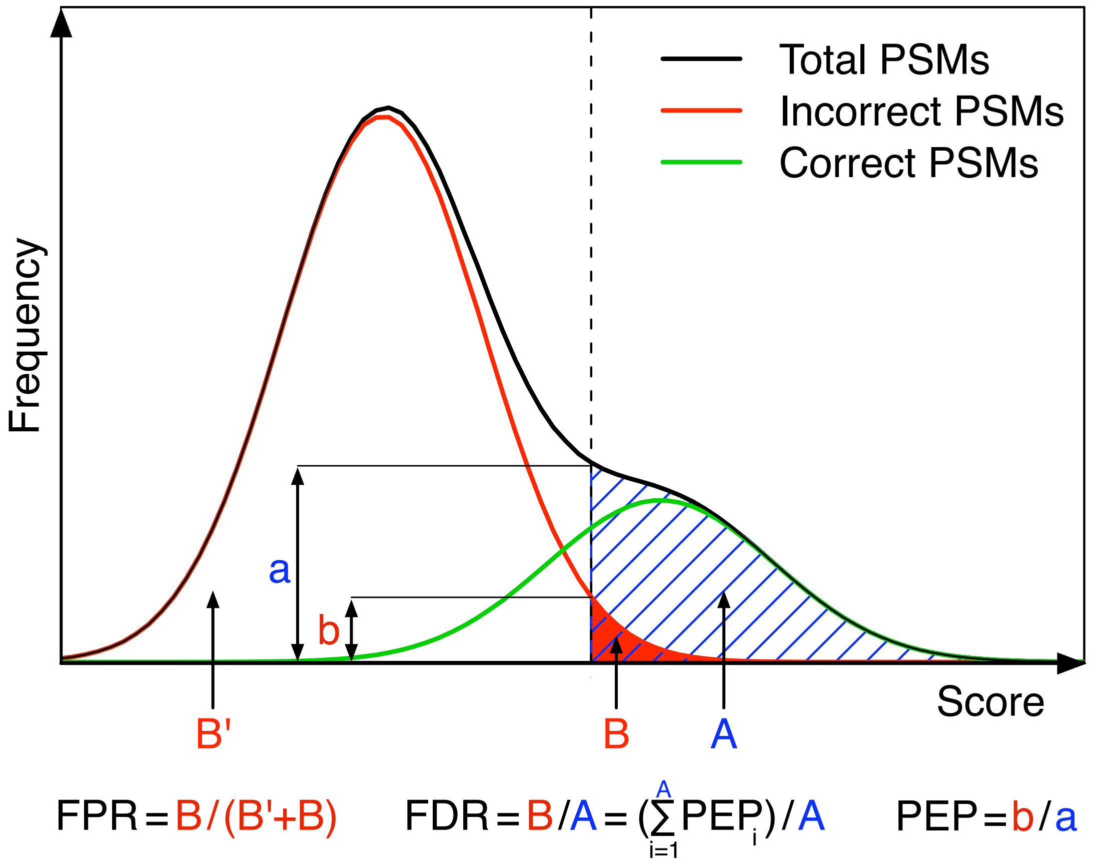


Figure 1.3: A score distribution (black) typically consists of a mixture of two underlying distributions, one representing the correct PSMs (green) and one the incorrect PSMs (red). Above a chosen score threshold (dashed line) the shaded blue area (A) represents all PSMs that were accepted, while the solid red filled area (B) represents the fraction of incorrectly identified PSMs with the chosen acceptance criteria. B together with B' sum up all incorrect PSMs for the whole dataset. The false positive rate (FPR) and the false discovery rate (FDR) can be calculated when the numbers of PSMs in B, B' and A are counted using the presented formulas. The posterior error probability (PEP) can be calculated from the height of the distributions at a given score threshold.

p-values, false discovery rates and q-values

The p-value is a widely used statistical measure for testing the significance of results in the scientific literature. The definition of the p-value in the context of MS database search scores is the probability of observing an incorrect PSM with a given score or higher by chance, hence a low p-value indicates that the probability is small of observing an incorrect PSM. The p-value can be derived from the false positive rate (FPR), which is calculated as the proportion of incorrect PSMs above a certain score threshold over all incorrect PSMs (figure 1.3). The simple calculation of the p-value however is misguiding when this calculation is performed for a large set of PSMs. In this case, we would expect to observe a certain proportion of small p-values simply by chance alone. An example: given 10,000 PSMs at a score threshold that is associated with a p-value of 0.05, we expect $0.05 \times 10,000 = 500$ incorrect PSMs simply by chance. This leads to the well known concept of multiple testing correction, which can be found in its simplest, but conservative, form in the Bonferroni correction (Bonferroni, 1935; Shaffer, 1995). Bonferroni suggested to correct the p-value by the number of tests performed, leading to a p-value of 5×10^{-5} in our example above. However, we have only corrected for the number of spectra, but not for the number of candidate peptides the spectrum was compared against. A correction taking into account both factors leads to extremely conservative score thresholds. However, an alternative well established method for multiple testing correction for large-scale data such as genomics and proteomics is to calculate the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

The FDR is defined as the expected proportion of incorrect predictions amongst a selected set of predictions. Applied to MS, this corresponds to the fraction of incorrect PSMs within a selected set of PSMs above a given score threshold (figure 1.3). As an example, say 1,000 PSMs score above a pre-arranged score threshold, and 100 PSMs were found to be incorrect, the resulting FDR would be 10%. On the other hand, the FDR can be used to direct the trade-off between sensitivity and error rate, depending on the experimental prerequisites. If, for example, a 1% FDR were required, the score threshold can be adapted accordingly.

To uniquely map each score and PSM with its associated FDR, the notion of q-values can be used. This is because two or more different scores may lead to the same FDR, indicating that the FDR is not a function of the underlying score (figure 1.4). Storey and Tibshirani (Storey and Tibshirani, 2003) have therefore proposed a new metric, the q-value, which was introduced into the field of MS proteomics by

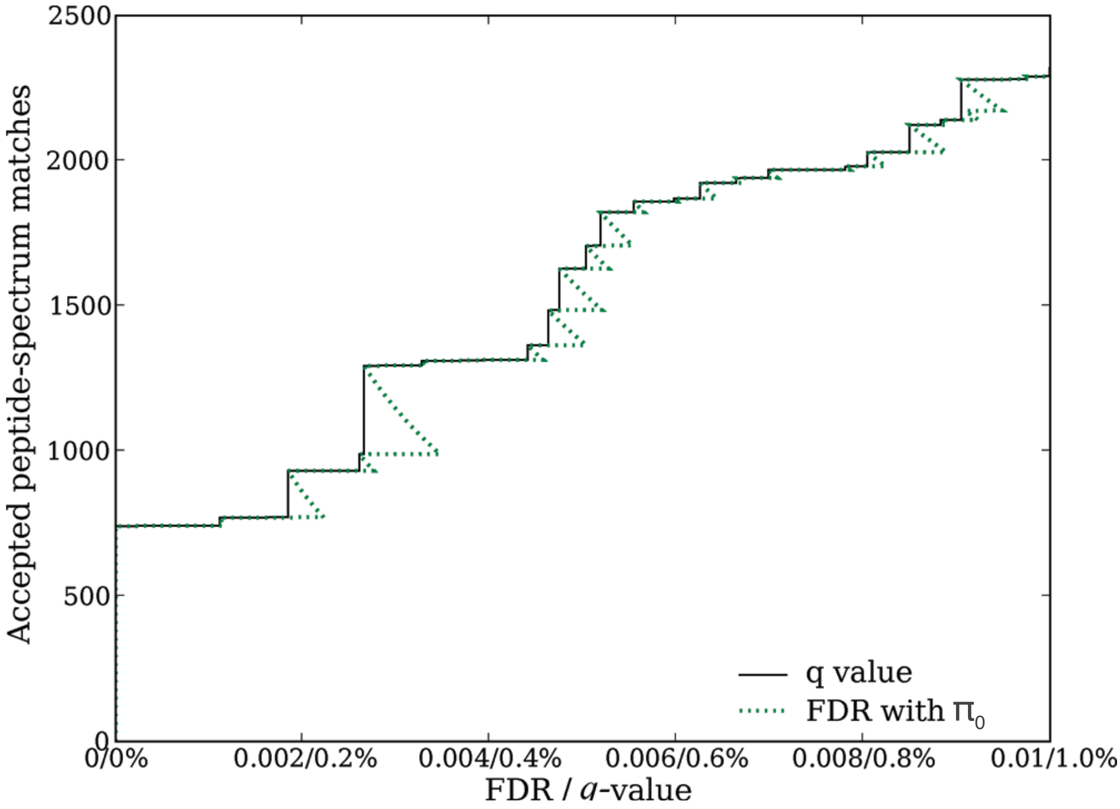


Figure 1.4: FDR compared with q-value: two or more different scores may lead to the same FDR, whereas the q-value is defined as the minimal FDR threshold at which a PSM is accepted, allowing to associate every PSM score with a specific q-value. Adapted from Käll *et al.* (2008a), figure 4b.

Käll *et al.* (2008a,b). In simple terms, the q-value can be understood as the minimal FDR threshold at which a PSM is accepted, thereby transforming the FDR into a monotone function: increasing the score threshold will always lower the FDR and *vice versa*. This property enables the mapping of scores to specific q-values. In Figure 1.5 the q-value is shown for a Mascot search on a high accuracy dataset. At a Mascot Ionscore of 10, 20 and 30 the corresponding q-values were 0.26, 0.04, 0.005 with 19967, 14608, 10879 PSM identifications respectively. It is important to note that for other datasets, instruments and parameter setting, the q-value could be significantly different for the same score and hence the q-value analysis should be performed for any individual search.

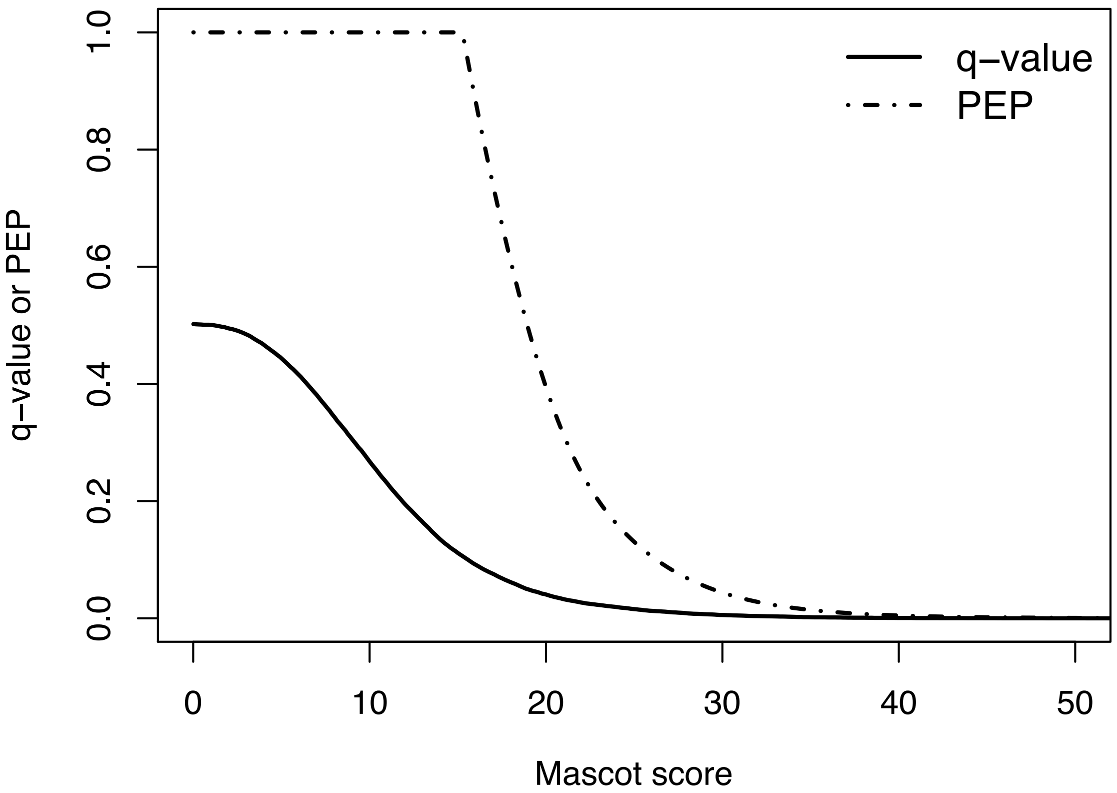


Figure 1.5: Mascot PSM scores were transformed into q-values and posterior error probabilities (PEP) using Qvality (see section 1.1.2.3). A score cut-off of 30 demonstrates the fundamental difference of the two significance measures: the q-value would have reported about 0.5% of all the PSMs as incorrect above that score threshold, whereas the PEP would have reported a 4% chance of a PSM being incorrect at this specific score threshold. Note: The maximum q-value for this dataset is 0.5, since only half of the PSMs are incorrectly assigned even without any score threshold applied due to the use of high quality and high mass accuracy data stemming from an LTQ-FT Ultra instrument. This factor (π_0) is discussed in more detail in figure 1.6.

Posterior Error Probability

The q-value is associated with individual PSM scores, although this measure is always a result of all PSMs in a dataset. For illustration, imagine we remove from a large dataset half of the spectra that were incorrectly matched above a given score threshold; after spectral removal the q-value for this same score threshold would be only about 50% of its original value, even though the underlying spectrum and PSM has not changed. Moreover, in an extreme case, a q-value of 1% could be taken to mean that 99 PSMs are perfectly correct and 1 PSM is incorrect. More likely the majority of these PSMs are good, but not perfect matches and a few are weaker matches. Clearly, when the focus of an experiment is based on individual peptide identifications (for example in biomarker discovery, genome annotation, or follow-up research of a key peptide), then it would be useful to compute spectrum specific significance measures that can be represented as the posterior error probability (PEP).

The global FDR or q-value reflects the error rate which is associated with a set of PSMs, whereas the PEP (or sometimes referred to as local FDR) measures the significance of a single spectrum assignment with a specific PSM score (Käll *et al.*, 2008b,c). The PEP is simply the probability of the PSM being incorrect, thus a PEP of 0.01 means that there is 1% chance of that PSM being incorrect. For the previous example where 100 PSMs resulted in a q-value of 1%, the PEPs would have reflected the stronger and weaker matches.

Unlike the FDR and q-value calculations that require minimal distributional assumptions, the PEP can only be calculated with knowledge of the underlying score distributions representing the correct and incorrect PSM identifications (see next section), since the PEP is inferred from the height of the distributions at a given PSM score. Figure 1.3 illustrates again that the PEP is specific to one PSM score, whereas the FDR accounts for the whole set of PSMs that scored at least as good as the PSM at hand. This leads to the fact that the sum of the PEPs above a chosen score threshold divided by the number of selected PSMs results in an alternative way of computing the FDR (Keller *et al.*, 2002).

Figure 1.5 shows the results of the PEP as well as the q-value calculations for a high mass accuracy dataset that was searched with Mascot. For a PSM score threshold of 10, 20 and 30, the associated q-values were 0.26, 0.04 and 0.005 whereas the PEPs were 1.0, 0.39 and 0.04, respectively. This clearly demonstrates the difference between the significance measures: whereas a Mascot score threshold of 30 (this is all PSMs with Mascot scores of 30 and above) led to only 0.5% incorrect

PSMs in this dataset, the individual Mascot score of 30 was associated with a 4% chance of being incorrect.

1.1.2.3 Computing statistical significance measures

Some database search algorithms report statistical measures, but these should be carefully validated and fully understood before being used and interpreted since their significance calculations are often based on pseudo statistical principles (see chapter 2). It is however very easy to obtain well founded significance measures with free post-processing software packages and methods as briefly described below. Finally, the well known effect of ”garbage-in/garbage-out” is also true for MS data analysis, but when tools and methods are applied sensibly, they can be extremely valuable and represent some of the latest developments in shotgun proteomics.

Target/Decoy database searching

A crucial step forward in assessing the reliability of reported PSMs was the introduction of the target/decoy search strategy pioneered by Moore *et al.* (2002): data is not only searched against the standard sequence database (target), but also against a reversed (Moore *et al.*, 2002), randomised (Colinge *et al.*, 2003), or shuffled (Klammer and MacCoss, 2006) database (decoy).

The idea is that PSMs obtained from the decoy database can be used to estimate the number of incorrect target PSMs for any given criteria such as score thresholds or heuristic methods. This enables the calculation of the FDR by simply counting the number of decoy and target PSMs that meet the chosen acceptance criteria (figure 1.3, FDR formula for separate target/decoy searches). A more accurate FDR can be obtained when the fraction of incorrect PSMs (π_0) matching the target database can be estimated and incorporated (figure 1.6). π_0 is equivalent to the ratio of the area under the curve of incorrect target PSMs (figure 1.3, red line) to the area under the curve of all target PSMs (figure 1.3, black line). This ratio can be estimated when

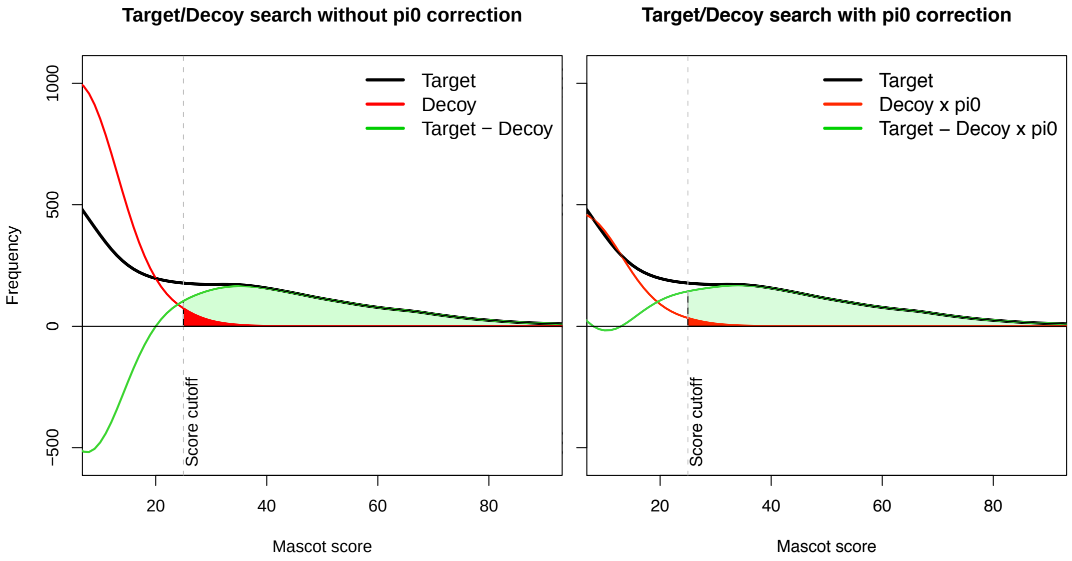


Figure 1.6: Score distributions of a target and decoy search with and without accounting for π_0 (pi0, percentage of target PSMs that are incorrect). Generally, the target score distribution (black) is a mixture of correct (green) and incorrect (red) peptide-spectrum matches, while the decoy matches are meant to be a “proxy” for the incorrect peptide matches obtained in the target run.
When no score thresholds are applied, all matches from the decoy search are counted as incorrect identifications. However, this is not a good proxy for the incorrect target matches, because a certain fraction of target matches are always correct, regardless of the score threshold. This becomes more important for recent data that is obtained from modern hybrid instruments such as the Orbitrap or LTQ-FT (Thermo Fisher Scientific), where even 50% of the peptide assignments can be correct as shown in this illustration. In fact, not accounting for this would mean that the estimated number of true identifications (target minus decoy hits) would become negative (left figure, green). However, incorporating the estimated fraction of peptides that are incorrect (π_0) in the target run, results in a much improved estimate of incorrect (red) and correct (green) peptide identifications (right figure).
This illustration is based on real data from sample 1 of section 2 of this thesis. Spline fits of score distributions were generated with the “smooth.spline” function of the R-project software (`http://www.r-project.org`) using default parameters and setting the degrees of freedom to 15.

decoy and target PSMs are counted for the score intervals [0, n], where 0 is the lowest score and n increases from the lowest to the highest score for each interval. Scores close to zero comprise mostly incorrect target PSMs and therefore the larger the interval the more conservative the π_0 estimate becomes with the variance decreasing (Käll *et al.*, 2008a). Various methods exist to average across these intervals (Hsueh *et al.*, 2003; Jin and Cai, 2006; Meinshausen and Rice, 2006; Storey, 2002; Storey and Tibshirani, 2003), but in the simplest form a straight line is fitted across the different interval ratios to yield a π_0 estimate (Käll *et al.*, 2008a). A formal description of the π_0 estimation procedure used in Percolator and Qvality is discussed in detail in Käll *et al.* (2008c)

It should be noted that there are two accepted concepts of target/decoy database searching and different groups favour one or the other method: either data is searched against a concatenated target/decoy database or data is separately searched against the target and decoy database (Bianco *et al.*, 2009; Elias and Gygi, 2007; Fitzgibbon *et al.*, 2007). A clear consensus as to which method is best is still to be established.

Qvality

Qvality (Käll *et al.*, 2008c) is a software tool that builds upon separate target/decoy database searching together with nonparametric logistic regression, where decoy PSM scores are used as an estimate "proxy" of the underlying null score distribution. It thereby enables transformation of raw and arbitrary PSM scores into meaningful q-values and PEPs. Since no explicit assumptions of the type of the score distributions are made, the method was shown to be robust for many scoring systems and hence is not limited to one specific database search algorithm. Qvality incorporates pi0 estimates into the FDR calculation and is therefore expected to produce more accurate significance metrics than standard target/decoy FDR calculation.

Application of Qvality is straightforward; it only expects two disjoint sets of raw PSM scores as input, one stemming from the target and one from the decoy database.

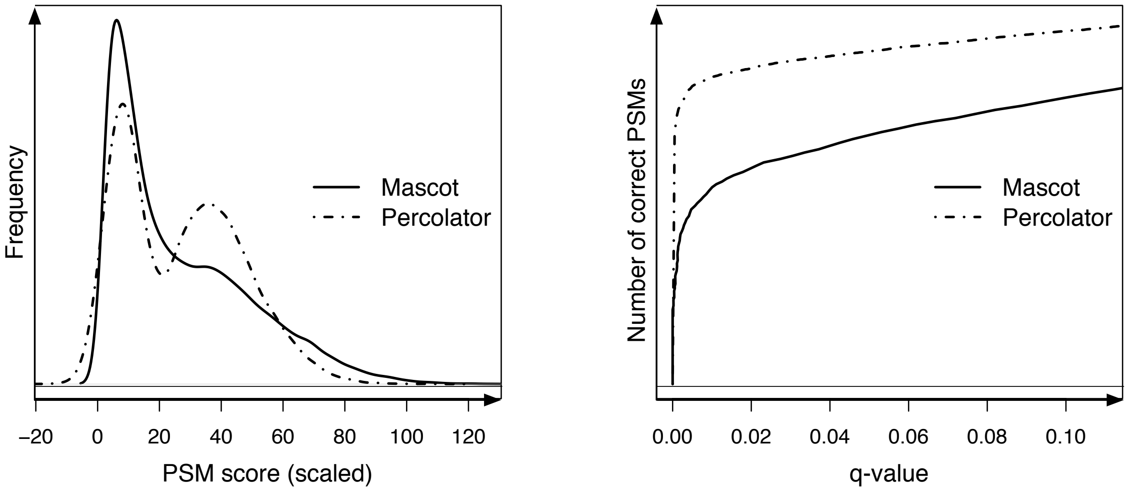


Figure 1.7: Distributions of Mascot and Percolator scores were generated from a high accuracy LTQ-FT Ultra dataset (left). This illustrates the bi-modal nature of PSM matching scores as simulated in figure 1.3 and further demonstrates the discrimination performance improvement between correct and incorrect PSMs for post-processing tools such as Percolator over Mascot. Note: these scores are not on the same scale, but have been normalised and scaled for this illustration.

Data for figure 1.5 was computed with Qvality using the target and decoy Mascot ion scores. Qvality is a small stand-alone command-line application without any external dependencies and is readily applicable `http://noble.gs.washington.edu/proj/qvality/`. Qvality was used for parts of the analysis in chapter 3.

PeptideProphet and Percolator

PeptideProphet and Percolator not only provide meaningful statistics, but also attempt to improve the discrimination performance between correct and incorrect PSMs (figure 1.7) by employing an ensemble of features, several of which are used by experts for manually validating PSMs.

”PeptideProphet” developed by Keller, Nesvizhskii, Kolker, and Aebersold (2002), was the first software that reported spectrum specific probabilities (P), akin to the PEP, as well as FDRs. In order to improve the discrimination performance between correct and incorrect PSMs, PeptideProphet learns from a training dataset a discriminant score which is a function of Sequest specific scores such as XCorr, deltaCn, Sp amongst others. PeptideProphet makes extensive use of the fact that PSM scores, as well as discriminant scores, represent a mixture distribution from the underlying superimposed correct and incorrect score distributions (figure 1.3, 1.6).

The original PeptideProphet algorithm is based on the assumption that the type of these distributions remain the same across experiments and hence were determined from training datasets. However, using an Expectation Maximisation algorithm (Dempster *et al.*, 1977), the parameters of these distributions are adapted for each dataset individually, enabling calculation of the corresponding FDR and P significance measures.

Recent versions of PeptideProphet supplemented this parametric model with a variable component mixture model and a semi-parametric model that incorporate decoy database search results (Choi and Nesvizhskii, 2008; Choi *et al.*, 2008). The rational of this was to provide more robust models for a greater variety of analytical platforms where the type of distribution may vary. PeptideProphet is a widely used and accepted method to compute confidence measures and is available at `http://tools.proteomecenter.org`. However, I have not used this tool in this work, since the Mascot implementation (the algorithm that is installed on our compute farm) does not improve discrimination and only uses the raw Mascot scores (personal communication, Alexey I. Nesvizhskii 2007).

Percolator (Käll *et al.*, 2007) is an alternative post-processing software relying on target/decoy database search results rather than on distributional assumptions to infer the q-value and PEP. This system employs a semi-supervised machine learning method for improving the discrimination performance between correct and incorrect PSMs. In the following the Percolator algorithm is outlined before its use in this work is discussed in more detail.

Target and decoy search results from Sequest (see section 1.1.1.2 and 1.1.2.3) are used as an input dataset for Percolator. In a first step, a vector of 20 features is calculated for every target and decoy PSM from these data, which remain fixed

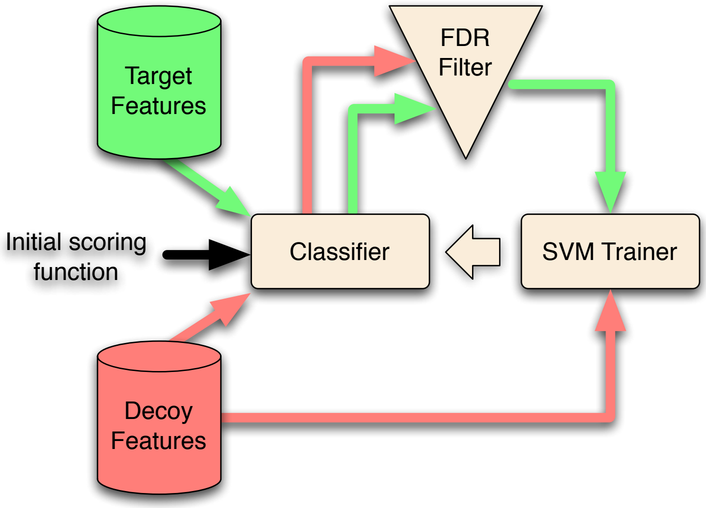


Figure 1.8: Schematic of the iterative learning process as implemented by Percolator

throughout the algorithm execution. Every feature, in isolation or in combination with other features, is reflective of some aspects that relate to the quality of the PSM at hand. The complete list of features is described in Käll *et al.* (2007) (supplementary table 1), which includes PSM scores, score difference between top hit and second best hit, enzyme specificity, peptide properties amongst others.

In the next step, a user defined feature that is known to discriminate well between correct and incorrect PSMs, such as the XCorr Sequest score, is used as an initial scoring function; a FDR filter can utilise this initial scoring function to select all target PSMs at a predefined low FDR. Given that at a 1% FDR setting 99% target PSMs can be assumed to be correct, this PSM subset serves as a positive training dataset, whereas the total set of decoy PSMs, which are known to be incorrect, are used as a negative training set (figure 1.8). Using the pre-calculated features of these training data, a linear support vector machine (SVM) (Ben-Hur *et al.*, 2008) learns to discriminate between the positive and negative training set.

The resulting SVM classifier is then used to re-score the target and decoy PSMs. The FDR filter is applied in another iteration to select all target PSMs at a low FDR, which together with all decoy PSMs are used for SVM training. The algorithm continues this cycle for a few iteration, and in Käll *et al.* (2007) it was shown that after a few iterations the system converges and results in a robust classifier that is then used in a last step to re-score each PSM in the dataset. It should be noted that a three-fold cross validation is performed at each iteration to avoid overtraining, resulting in biased scoring. The combination of features results in significantly better discrimination between correct and incorrect PSMs when compared to raw PSM scores (figure 1.7).

For every PSM, the associated q-value as well as the PEP are reported (Käll *et al.*, 2008b,c). The whole process is fully automated and does not require any expert-driven or subjective decisions, thereby eliminating any artificial biases. The learnt classifier is specific and unique to each dataset, thus adapting to variations in data quality, protocols and instrumentation. This was demonstrated in Käll *et al.* (2007) (supplementary figure 2), where feature weights were used as a measure of the importance of individual features. However, it should be noted that feature weights of a SVM are difficult to interpret, since multiple features may be correlated and hence feature weights are divided arbitrary between those. Alternatively, relative importance of a feature could be measured by removing it from the set, but again, correlating feature complicate the interpretation.

Percolator is available under `http://noble.gs.washington.edu/proj/percolator/` and similar to Qvality does not depend on any external dependencies and hence can be readily used. It offers a simple command line interface that requires Sequest results as input and outputs the q-value, PEP, as well as the peptide and associated protein(s) information for each spectrum.

I have developed upon Percolator a Mascot module that uses an extended feature set, including Mascot specific features as well as intensity and ion-series information. This work is discussed in detail in chapter 3. It is available for download under `http://www.sanger.ac.uk/resources/software/mascotpercolator/` and is currently integrated into the official Mascot 2.3 release (see `http://www.matrixscience.com/workshop_2009.html` for more information).

1.2 Genome annotation

1.2.1 Fundamentals of gene transcription and translation

The genomic sequence encodes the blueprint of an organism. The instruction sets are encoded in protein coding and non-coding genes, which are defined stretches of DNA sequence that contain the information required to construct proteins and functional RNA molecules respectively. The realisation of genes is initiated by transcription, whereby genomic DNA is transcribed into RNA.

This premature RNA sequence comprises two different types of segments in eukaryotes, exons and introns, the latter of which is removed during splicing. This process enables the construction of alternative products (alternative splicing) by varying the joining of exons: these can be extended at the 5' donor or 3' acceptor site, one or multiple exons can be skipped or rarely introns can be retained.

Products that are derived from non-coding RNA genes, code for RNA molecules and are not further translated into proteins. These non-coding molecules have been studied extensively in the last decade and are involved in many cellular processes, although the function is unknown for some of these elements (Carninci *et al.*, 2005; Clamp *et al.*, 2007; Claverie, 2005; Washietl *et al.*, 2007). However, the focus of this introduction are the main functional players in a cell: proteins.

Spliced RNA sequence that was derived from protein coding genes is referred to as messenger RNA (mRNA). Mature mRNA comprises the open reading frame (ORF) that codes for the protein and the untranslated sequences (5' UTR upstream and 3' UTR downstream of the ORF). During protein translation, three nucleotides are read at a time (codons) and specific transfer RNAs (tRNA) match these codons with three unpaired complementary bases (anticodon). Each anticodon defines a specific amino acid that is bound to the tRNA, which upon binding of mRNA and tRNA is ligated to the growing polypeptide chain.

The newly synthesised protein must fold to its active three-dimensional structure

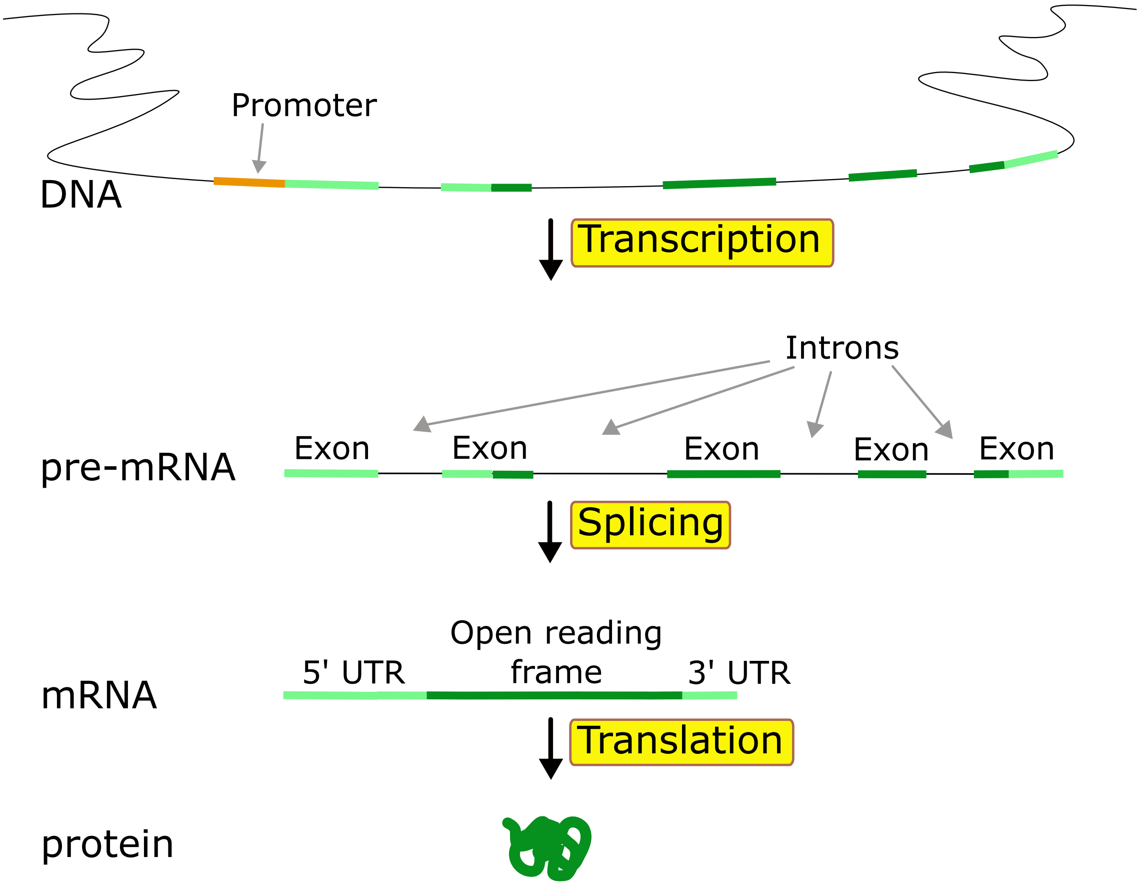


Figure 1.9: Illustration of gene transcription and translation according to the standard model. The figure was adapted from Wikipedia (`http://en.wikipedia.org/wiki/File:Gene2-plain.svg`)

before it can carry out its function. This simplified standard model describing the unfolding of genomic sequence, also known as the “central dogma of molecular biology” (Crick, 1958, 1970), is further illustrated in figure 1.9.

1.2.2 Genome sequencing

Sequencing efforts in the last decade generated a large amount of raw genomic DNA sequence data. To date there are 118 complete eukaryotic genomes sequenced (Liolios *et al.*, 2009) and more sophisticated sequencing technologies will even speed up this data collection process. A project to sequence 10,000 vertebrate species has just been proposed, even though technology is not yet up to it (Pennisi, 2009). Genomes can be large, for example the human genome comprises approximately 3.2×10^9

base pairs, yet only about 1-2% of its DNA codes for proteins (Birney *et al.*, 2007; Claverie, 2005).

1.2.3 Definition of genome annotation

Genome annotation can be defined as augmenting these raw DNA sequences with additional layers of information (Brent, 2005; Stein, 2001). It can be distinguished between structural and functional annotation. The former is the process of identifying important genomic elements such as genes, the precise localisation of genes within the genome and the elucidation of exon/intron structures, while the latter deals with the biological function, regulation and expression analysis of these elements. For clarification, when the term "genome annotation" is used in the remainder of this work, it refers to structural annotation only.

The task of accurately annotating the complete set of protein coding genes and their alternative splice forms is considered one of the hardest and yet most important steps towards understanding a genome, since proteins are central to virtually every biological process in a cell. However, the difficulty of gene identification and gene structure elucidation is determined by the complexity of the underlying genome: for example, identification of ORFs in bacteria, which are not discussed in this work, is relatively easy due to the lack of alternative splicing and a compact genome; simpler eukaryotes, such as yeast with limited splicing and short intronic regions are much easier to annotate than vertebrates, since extensive alternative splicing, long introns and intergenic regions further complicate sensitive and specific annotation.

1.2.4 Genome annotation strategies

With the ever increasing availability of sequenced genomes, automatic genome annotation is an active area of research. Figure 1.10 provides an overview of the different available annotation strategies, which will be briefly discussed.

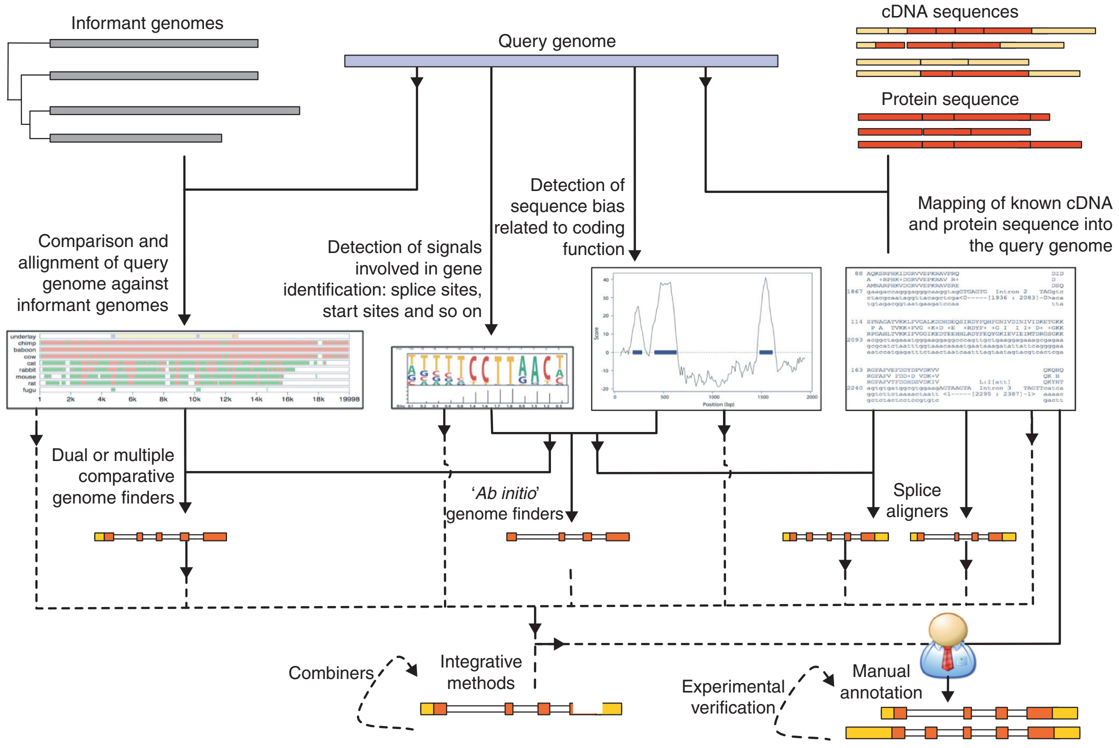

Informant genomes
Query genome
cDNA sequences
Protein sequence
Detection of sequence bias related to coding function
Mapping of known cDNA and protein sequence into the query genome
Comparison and allignment of query genome against informant genomes
Detection of signals involved in gene identification: splice sites, start sites and so on
Dual or multiple comparative genome finders
'Ab initio' genome finders
Splice aligners
Combiners
Integrative methods
Experimental verification
Manual annotation

Figure 1.10: Overview of the different gene-finding strategies. Figure was adapted from Harrow *et al.* 2009, figure 1.

The most reliable gene-finding systems are based on experimental evidence where available complementary DNA (cDNA) (Furuno *et al.*, 2003; Imanishi *et al.*, 2004), expressed sequence tags (EST) (Adams *et al.*, 1991; Parkinson and Blaxter, 2009) and protein sequences are aligned to the genomic sequence by algorithms that can account for splicing, such as GeneWise (Birney and Durbin, 1997; Birney *et al.*, 2004) or Exonerate (Slater and Birney, 2005). However, this approach requires extensive mRNA or protein sequence coverage and since only a fraction of genes are transcribed at any given time for any given cell, complete coverage is hard to achieve. Moreover, the quality of these data is often low, for example the intrinsically short EST sequences contain up to 5% sequencing errors or include contaminant sequences and "full-length" cDNAs can be truncated, which together with SNPs can result in ambiguous or incorrect alignments (Nagaraj *et al.*, 2007).

An additional strategy is the comparative genomics approach. It is known that functional elements undergo mutation at a slower rate and hence regions that are found to be conserved between related genomes such as human and mouse can indicate functional genes (Alexandersson *et al.*, 2003; Korf *et al.*, 2001; Parra *et al.*, 2003). However, many non-coding functional elements are also conserved (Claverie, 2005) and species specific genes can be missed (Knowles and McLysaght, 2009), limiting the approach when used in isolation.

Ab initio gene predictors detect protein coding signals from DNA sequence alone. These signals are either specific sequences that indicate the presence of a nearby gene (e.g. regulatory regions such as promoters), or statistical properties of the protein-coding sequence itself (e.g. GC content). Genscan (Burge and Karlin, 1997), GeneID (Parra *et al.*, 2000) and Augustus (Stanke and Waack, 2003) are popular *ab initio* gene-finders. Inferring annotation from genomic sequence alone is an extremely challenging task, resulting in low sensitivity and specificity and hence is not used directly for annotation but rather for the generation of candidate transcripts. Some of these predictors optionally allow the incorporation of additional extrinsic evidence such as cDNA, EST, protein or sequence conservation data to improve prediction accuracy.

1.2.5 Ensembl and Vega

With the availability of the human genome draft sequence in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001), Ensembl was developed with the aim of providing a robust and high quality automated annotation system yielding reliable information (Hubbard *et al.*, 2002). Ensembl leverages experimental evidence (see previous section), whereby species specific cDNAs and protein data are aligned onto the genome to derive annotation. However, ESTs are not considered in the Ensembl gene build process due to their variable quality and the implied ambiguities. The automatic Ensembl annotation system is described in detail by Curwen *et al.* (2004). Ensembl now expanded to more than 41 vertebrates (Hubbard *et al.*, 2009) as well as to plants, fungi, parasites and bacteria (Kersey *et al.*, 2009).

Moreover, Ensembl offers a stable and rich resource for researchers. It provides a web application that enables researchers to explore the genome of interest with a web browser (figure 1.11), optionally allowing to integrate external annotation data. Lastly, it provides a robust and extensive Perl application programming interface that enables more advanced analysis of the underlying data.

When the first draft of the human genome sequence was published, the number of protein-coding genes was estimated to be around 30,000 to 40,000 (Lander *et al.*, 2001; Venter *et al.*, 2001). Over the years the number of predicted protein coding genes decreased (International Human Genome Sequencing Consortium, 2004) and even today the exact number remains uncertain and is estimated to be between 20,000 and 25,000 (Clamp *et al.*, 2007), with Ensembl release 56 (November 2009) predicting 23,621 protein coding genes. The ENCyclopedia Of DNA Elements (ENCODE) community experiment aims at identifying all functional elements in the human genome with high-throughput methods (The ENCODE Project Consortium, 2004), with the pilot study being completed in 2007, where 1% of the human genome was investigated (Birney *et al.*, 2007).

The GENCODE project produced a high quality "reference" annotation of protein coding genes for these regions through a combination of computational, experimental and manual annotation efforts (Harrow *et al.*, 2006). Based on a reference annotation set produced by GENCODE, the ENCODE Genome Annotation Assessment Project (EGASP) evaluated the accuracy of automatic gene prediction methods, including Ensembl (Guigo *et al.*, 2006). The results confirmed the high quality GENCODE annotation, but also illustrated that automated annotation cannot produce the same level of accuracy: in 30% of the cases, the best predicted transcript per gene did not reproduce the GENCODE reference annotation and accuracy dropped significantly when alternative isoforms were to be considered by Ensembl.

This illustrates that manual analysis still plays a significant role for high quality

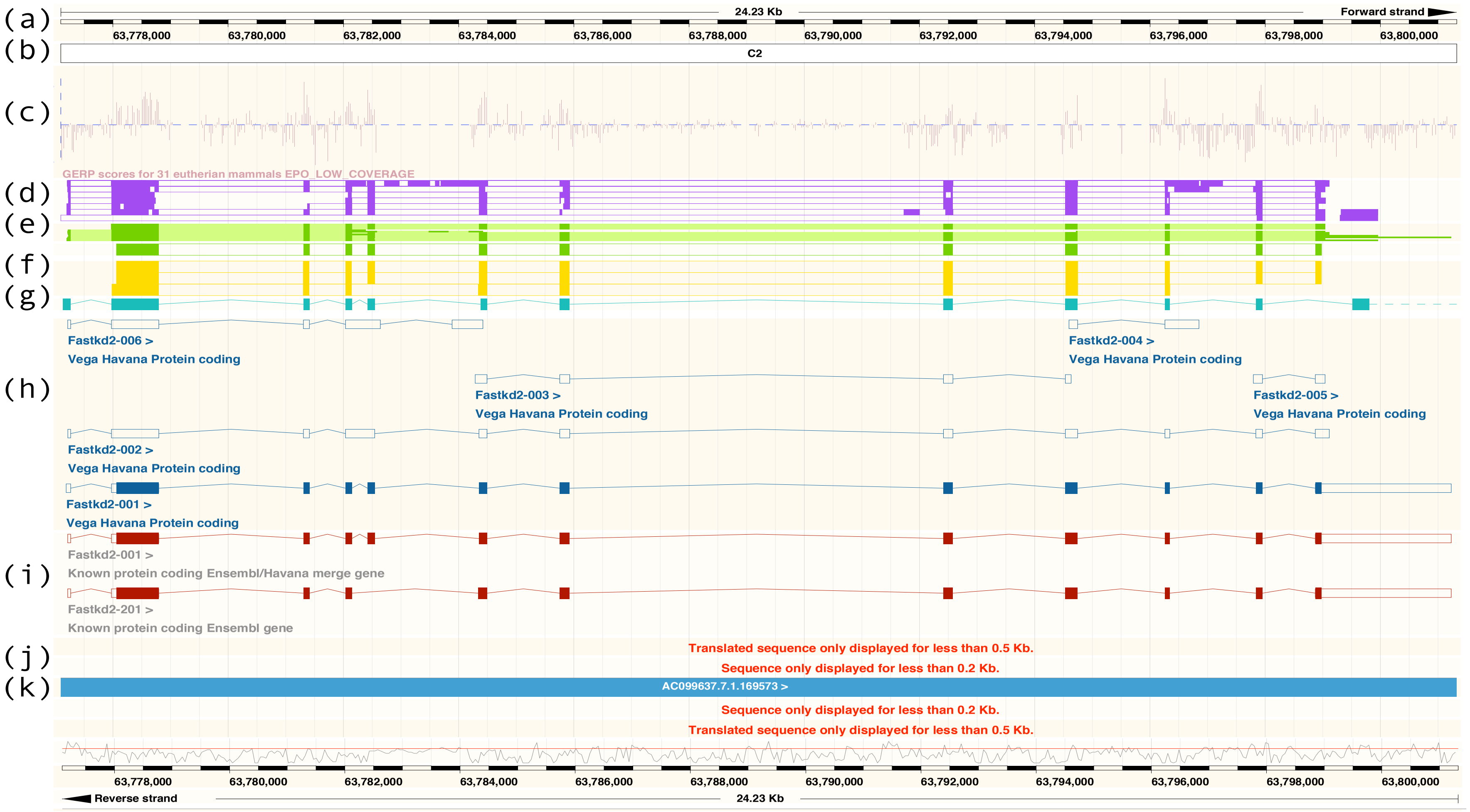

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
24.23 Kb
Forward strand
63,778,000
63,780,000
63,782,000
63,784,000
63,786,000
63,788,000
63,790,000
63,792,000
63,794,000
63,796,000
63,798,000
63,800,000
C2
GERP scores for 31 eutherian mammals EPO_LOW_COVERAGE
Fastkd2-006 >
Vega Havana Protein coding
Fastkd2-004 >
Vega Havana Protein coding
Fastkd2-003 >
Vega Havana Protein coding
Fastkd2-005 >
Vega Havana Protein coding
Fastkd2-002 >
Vega Havana Protein coding
Fastkd2-001 >
Vega Havana Protein coding
Fastkd2-001 >
Known protein coding Ensembl/Havana merge gene
Fastkd2-201 >
Known protein coding Ensembl gene
Translated sequence only displayed for less than 0.5 Kb.
Sequence only displayed for less than 0.2 Kb.
AC099637.7.1.169573 >
Sequence only displayed for less than 0.2 Kb.
Translated sequence only displayed for less than 0.5 Kb.
63,778,000
63,780,000
63,782,000
63,784,000
63,786,000
63,788,000
63,790,000
63,792,000
63,794,000
63,796,000
63,798,000
63,800,000
Reverse strand
24.23 Kb

Figure 1.11: Screenshot of the Ensembl browser representing the *Fastkd2* locus on chromosome 1. (a) Chromosome coordinates. (b) Chromosome bands. (c) Sequence conservation across 31 eutherian mammals. (d) Mouse EST sequence alignments. (e) Full length cDNA sequence alignments. (f) Alignment of UniProtKB proteins. (g) Genscan gene prediction. (h) Manually annotated Vega coding (solid) and noncoding (outline) transcripts. (i) Ensembl transcript predictions (coding represented by solid and noncoding by outline rectangles). (j) Translated and genomic sequence (not shown since region too large). (k) Contigs.

annotation. The HAVANA group at the Wellcome Trust Sanger Institute manually annotates sequences on a clone by clone basis, using a combination of extrinsic evidence, most notably cDNAs/ESTs and protein sequence alignments combined with *ab initio* gene predictions (Genscan, Augustus) and comparative analysis. Thereby the team manually annotates genes by supporting evidence only. The Vertebrate Genome Annotation (Vega) database is a publicly accessible repository for these manually annotated genome sequences (Ashurst *et al.*, 2005; Wilming *et al.*, 2008). Moreover, full length HAVANA transcripts are also merged into Ensembl (Hubbard *et al.*, 2009).

Future work will continue to improve genome annotation quality. For example, experimental validation will continue as part of the GENCODE scale-up project (`http://www.sanger.ac.uk/encode/`), which builds on the success of the GENCODE pilot project (Harrow *et al.*, 2006), but is limited to the human genome. The CCDS (Consensus Coding Sequence, Pruitt *et al.* 2009) project defines a stable set of protein coding gene structures for human and mouse by identifying agreeing annotation between Ensembl/Vega, RefSeq (Pruitt *et al.*, 2006) and UCSC (Kuhn *et al.*, 2009). Lastly, as technology evolves, new and revolutionary methods will be identified that can further aid the genome annotation efforts, such as the recent introduction of next-generation sequencing methods (Fullwood *et al.*, 2009; Wang *et al.*, 2009).

1.3 Proteogenomics

The automatic Ensembl pipeline and the HAVANA manual curation pipeline incorporate protein data from the UniProtKB database (Bairoch and Apweiler, 1997; Wu *et al.*, 2006), where more than 99% of the protein sequences are derived from genomic translations and cDNA sequences, but only 13% are supported by protein level evidence such as mass spectrometry identification (UniProt release notes 15.11, `http://www.uniprot.org/news/2009/11/24/release`). Proteins that are detected by mass spectrometry provide direct experimental evidence for gene translation, which cDNA data cannot offer. Therefore high-throughput tandem mass spectrometry can aid genome annotation efforts on a genome scale, by validating and refining annotated coding sequences and detection of novel ORF. Efforts to combine genome annotation with protein mass spectrometry led to the establishment of a new field, proteogenomics, a term coined by Jaffe *et al.* (2004).

Yates *et al.* (1995) demonstrated the concept of searching MS/MS data directly against a six-frame translation of the genome, but it was Kuster *et al.* (2001) and Choudhary *et al.* (2001a,b) that applied this approach to eukaryotic genomes with the purpose of validating and refining gene annotation as well as the identification of novel genes. In these studies a six-frame translation was used as a search database, however in higher eukaryotes this is problematic: only 1-2% of the human genome encodes proteins (Birney *et al.*, 2007; Claverie, 2005), therefore most of the six-frame translation is essentially random sequence. The inflated search space increases the likelihood of false positive identifications and therefore sensitivity decreases at a constant FDR. In addition, six-frame translation does not account for alternative splicing, which can affect the majority of genes (Wang *et al.*, 2008), and 20-28% of tryptic peptides, depending on the number of allowed missed cleavages, span a splice site.

The Peptide Atlas project (Desiere *et al.*, 2005, 2006), the first large-scale proteogenomics pipeline and MS/MS peak lists and raw data repository, employs the standard International Protein Index (IPI) database (Kersey *et al.*, 2004) as an alternative approach to six-frame translation. IPI provides a minimally redundant yet maximally complete sets of protein sequences from Ensembl, Vega, RefSeq and UniProtKB. Later versions of Peptide Atlas complement the IPI database with protein isoforms from Ensembl. Peptide Atlas comprises an analysis pipeline to processes MS data with Sequest and PeptideProphet and provides access to these peptide identifications, which are persisted in a comprehensive relational database. As an additional feature, Peptide Atlas maps peptide identifications to the genome using the sequence alignment tool BLAST (Altschul *et al.*, 1990). These mappings are made available with a distributed annotation server (DAS) (Dowell *et al.*, 2001), allowing peptide identification results to be integrated into various genome browsers, such as Ensembl. The currently available DAS source (`http://www.peptideatlas.org/setup_genome_browser.php`) does not provide meta-information of the uniqueness of the peptide within the genome, limiting the direct use for annotation, since the peptide could match multiple different genomic loci. The system is not available for download, providing little flexibility for required changes or extensions, such as support of Mascot and Mascot Percolator or different search databases.

The Genome Annotating Proteomic Pipeline (GAPP), developed by Shadforth *et al.* (2006), is an alternative proteogenomic pipeline that unlike PeptideAtlas relies on Ensembl translations for peptide identification, guaranteeing a perfect genomic match of every identified peptide. Another significant difference compared to Peptide Atlas is the peptide scoring scheme: GAPP accepts Mascot, Sequest and X!Tandem peptide identification results, which are subsequently post-processed with the advanced average peptide score (Shadforth *et al.*, 2005), where peptides are given extra credibility when they share a protein that was obtained from within the same experiment (Chepanoske *et al.*, 2005). However, this approach does not provide a significance measure for an individual peptide match, which is required when peptide identifications are used for genome annotation. Moreover, the inherent peptide-protein apportioning further increases scoring complexity (Nesvizhskii and Aebersold, 2004; Nesvizhskii *et al.*, 2003), in particular in respect to target/decoy FDR estimation. The target/decoy approach is extensively tested for peptide level FDR estimation, but when protein level information is incorporated, it requires the decoy database to resemble the target database in terms of peptide-protein composition in order to provide a valid null model. Otherwise the number of protein identifications in the decoy database may deviate from the actual number of incorrect protein identifications.

Although Peptide Atlas and GAPP are the only available high-throughput proteogenomic systems, the following studies are representative of alternative analytical strategies that are employed in this field of research. Tanner *et al.* (2007) developed an exon splice graph database that is build by combining all pairs of predicted exons with subsequent cDNA and EST filtering data to limit the search space. This method is implemented as an extension of the Inspect peptide identification algorithm (Tanner *et al.*, 2005), a peptide sequence tag based approach (Mann and Wilm, 1994). The associated proteogenomics study of Tanner *et al.* (2005) remains the most comprehensive proteogenomics study to date. They searched a corpus of 18.5 million tandem MS spectra (human), enabling the validation of 39,000 exons, 11,000 splice sites (introns) and confirmed 40 alternative splice events. Tress *et al.* (2008) focussed specifically on the analysis of alternative splicing and identified multiple alternative gene products for over a hundred *Drosophila* genes. Castellana *et al.* (2008) has combined the splice graph approach with a six-frame translation and the currently annotated proteome of *Arabidopsis thaliana* and found the majority of peptides to map to existing annotation, although 13% novel peptides were identified.

Further improvements can be expected in the field of proteogenomics when experimental and computational methods integrate. For example, Sevinsky *et al.* (2008) leveraged peptide isoelectric focusing and accurate peptide mass to greatly reduce the peptide search space, enabling highly sensitive peptide identification even on a large six-frame translation of human. Brunner *et al.* (2007) has combined sample diversity, multidimensional fractionation and analysis-driven feedback loops to guide data collection, resulting in unprecedented gene coverage in *Drosophila melanogaster*.

Proteogenomics studies can be focussed on particular problems, as demonstrated by Schandorff *et al.* (2007) and Bunger *et al.* (2007) who validated non-synonymous SNPs, Wright *et al.* (2009) who used proteogenomics on newly sequenced genomes as well as by (Gupta *et al.*, 2008) who introduced comparative proteogenomic studies.

Although proteogenomics is still a relatively novel field of research, the growing interest from both sides, the proteomics and genomics community is apparent. This is facilitated by the readily available proteomics data that provides inherently strong experimental evidence of translated gene products, something that cannot be achieved with transcriptional data.

1.4 Thesis outline

The objectives of my work are to build on and improve the methods introduced in section 1.3 to enable reliable high-throughput proteogenomic data analysis.

In the first results chapter, I evaluate the peptide identification software "Mascot" that is routinely used at the Wellcome Trust Sanger Institute and elsewhere. Since peptide-spectrum matching is a difficult problem, wrong peptide identifications are expected. To address this Mascot provides a scoring scheme with probability thresholds. I have evaluated these for low and high mass accuracy data and showed that they are not sufficiently accurate. I developed an alternative scoring scheme that provides more sensitive peptide identification specifically for high accuracy data, while allowing the user to fix the false discovery rate.

I utilise the machine learning algorithm "Percolator" in the following chapter to further extend my Mascot scoring scheme with a large set of orthogonal scoring features that contribute to the discrimination performance between correct and incorrect peptide-spectrum matches. I demonstrate that this method provides very good sensitivity, while producing reliable and robust significance measures that were validated with protein standard datasets. Sound scoring statistics avoid propagation of wrong peptide identifications into genome annotation pipelines.

My genome annotation pipeline, introduced in chapter 4, closes the gap between high throughput peptide identification and large scale genome annotation analysis. At the core of this pipeline is a comprehensive database, enabling the efficient mapping of known and predicted peptides to their genomic loci, each of which is associated with supplemental annotation information such as gene and transcript identifiers. Software scripts allow the creation of automated genome annotation analysis reports.

In the last results chapter, the pipeline is tested with a large mouse MS dataset. I show the value and the level of coverage that can be achieved for validating genes and gene structures, while also highlighting the limitations of this technique. Moreover, I show where peptide identifications facilitated the correction of existing annotation, such as re-defining the translated regions or splice boundaries. Lastly, I propose a set of novel genes that are identified by the MS analysis pipeline with high confidence, but currently lack transcription or conservational evidence.

Chapter 2

Assessment of Mascot and X!Tandem and development of the Adjusted Mascot Threshold

2.1 Introduction

In the general introduction I have discussed the concept of sequence database searching, that is commonly used to assign sequence information to MS/MS spectra (section 1.1.1). This chapter focusses on the scoring schemes of database search algorithms, which are required to provide sound peptide assignment significance measures in order to minimising incorrect and maximising correct identification. Many different techniques have been applied in the past, from manual heuristic rules to machine learning algorithms that discriminate between correct and incorrect identifications (Anderson *et al.*, 2003; Jones *et al.*, 2009; Resing *et al.*, 2004; Ulintz *et al.*, 2006). The most popular database search engines to date, including Mascot (Perkins *et al.*, 1999) and X!Tandem (Craig and Beavis, 2004), provide theoretically or empirically derived statistical thresholds to help assess the significance of peptide identifications.

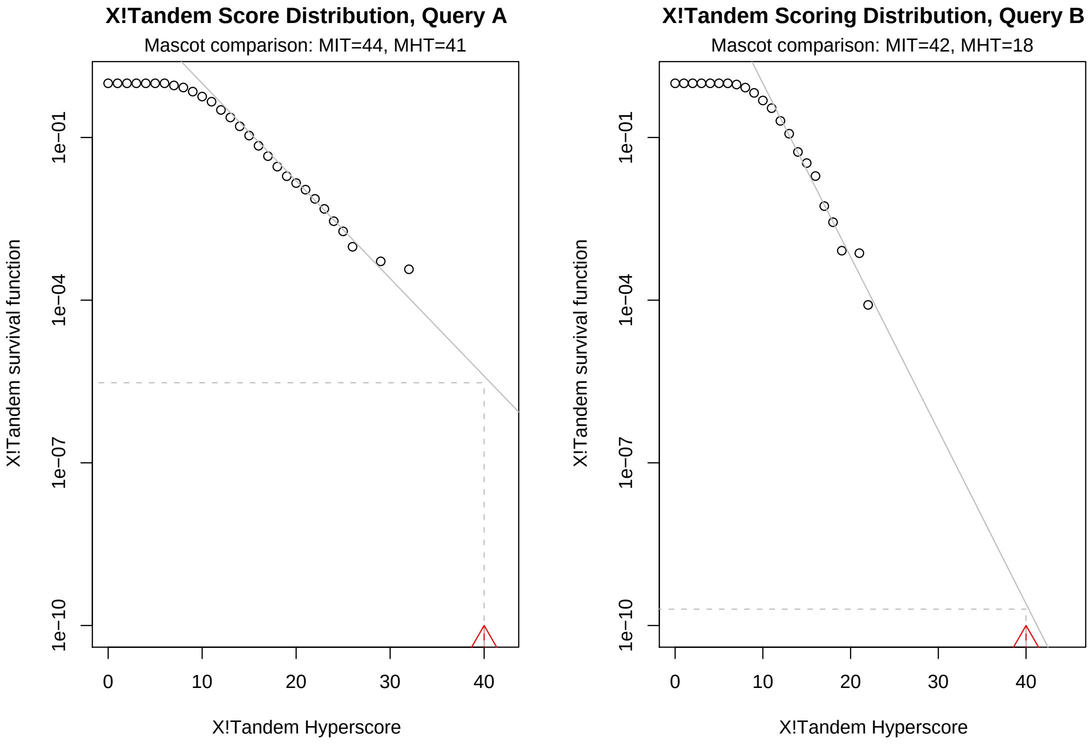


Figure 2.1: Exemplary survival functions from X!Tandem for two spectrum queries A and B. Although the number of peptide candidates for both queries is similar, there are apparent differences in the actual peptide score distributions. The survival functions were extrapolated for a score of 40 that corresponds to a probability of approximately 3×10^{-6} and 2×10^{-10} for query A and B respectively. Given the number of peptides scored were 1×10^{5}, the expectation value of the former would be 0.3 while the expectation value of the latter would be 2×10^{-5} (for a detailed explanation on how the survival function and expectation values are calculated, refer to Fenyo and Beavis, 2003). Therefore, at a significance level of 0.05 the same score would have been considered highly significant for query B, but not for query A.
In contrast, the MIT is inferred from the number of peptide candidates only, resulting in very similar thresholds of 44 and 42 for both queries. A hypothetical Mascot score of 40 would not have been considered significant for either query.
On the other hand, the empirically derived MHT was 41 for query A and 18 for query B, thus classifying the peptide hit for query B as significant which agrees with the X!Tandem extrapolation example. It should be noted that the absolute scores and threshold values of X!Tandem and Mascot are not directly comparable.

Mascot reports a probability-based Mascot Identity Threshold (MIT) for each individual spectrum query. A Mascot score above MIT is considered to be a significant peptide assignment. The MIT is defined as $-10 \times log_{10}(20 \times p \times n)$, where p is the probability of a random peptide match and n corresponds to the actual number of peptide candidates. For example, if a 1 in 20 chance of obtaining a false positive is acceptable ($p = 0.05$) and there are 10000, 1000, 100 and 10 peptide candidates for a given mass window in the sequence database, the MIT would be 40, 30, 20 and 10 respectively. For a peptide match with a score that equals the MIT ($p = 0.05$), the expectation value (E-value) of this hit is also 0.05, but if the score exceeds the MIT by e.g. 10, the E-value drops to 0.005. The E-value in Mascot is defined as $p \times 10^{(MIT-score)/10}$ and corresponds to the number of times one would expect this score by chance alone (`http://www.matrixscience.com/pdf/2005WKSHP4.pdf`). Therefore the MIT only reflects changes in search space, defined by the number of peptide candidates, and would be affected by various factors such as the maximum mass deviation (MMD) settings, the number of allowed missed cleavages, enzyme specificity and variable modifications.

Mascot also reports an empirical Mascot Homology Threshold (MHT). A Mascot score exceeding this threshold can be considered a significant outlier from the distribution of all candidate peptide-spectrum match scores, but an exact definition of the MHT was not published. Similarly, X!Tandem employs score distributions, but extrapolates empirical E-values to assess the significance of a peptide match (Craig and Beavis, 2004; Fenyo and Beavis, 2003). It is important to note that the E-values derived by Mascot and X!Tandem are based on completely different assumptions and may therefore lead to significantly different scoring results even for the same peptide spectrum match; as described above, the Mascot E-value is based on a theoretical statistical model, whereas the X!Tandem E-value is an empirical outlier determination. In figure 2.1 I illustrate the similarities and differences between the X!Tandem, MHT and MIT scoring scheme.

With high accuracy MMD settings the search space can decrease significantly leading to insufficient data points of the score distributions to reliably extrapolate E-values. To compensate for this, X!Tandem uses cyclic permutations of all peptide candidates that are scored and used to pad the score distribution (optional). In general, empirical scoring schemes that utilise the peptide candidate score distributions for thresholding or E-value extrapolation are more robust to changing MS/MS data quality such as signal to noise, mass accuracy or fragmentation quality.

It is anticipated that reducing the search space should improve the performance of algorithms for peptide identification (Zubarev and Mann, 2007). For example, with high mass accuracy data in the range of a few ppm, the search space can be reduced by orders of magnitude in comparison to low accuracy data acquired typically on ion trap instruments (Elias and Gygi, 2007).

Established database search algorithms, and in particular their scoring schemes, were not specifically developed for high mass accuracy data. Rudnick *et al.* (2005) evaluated the effects of MMD settings on Mascot performance and proposed an empirical Mass Accuracy based THreshold (MATH) that provided improved sensitivity at a user-defined false discovery rate (FDR). They applied a range of global cut-off thresholds and determined the associated FDRs. A linear regression over the logarithms of these FDRs and the cut-off values enabled an empirical threshold extrapolation at a predefined FDR. However, the Mascot evaluation was exclusively limited to the MIT. Savitski *et al.* (2005) have developed a database size independent scoring scheme for high accuracy data. This work is based on complementary fragmentation techniques, and cannot be applied solely on standard collision induced dissociation data (Biemann, 1988; Roepstorff and Fohlman, 1984). Gygi and co-workers proposed to exploit high accuracy MS data by searching at relaxed mass tolerance settings followed by mass accuracy filtering (Beausoleil *et al.*, 2006; Everley *et al.*, 2006). Combined with a moderate threshold on peptide-spectrum correlation scores, they found this strategy to serve as a good discriminator between correct and incorrect

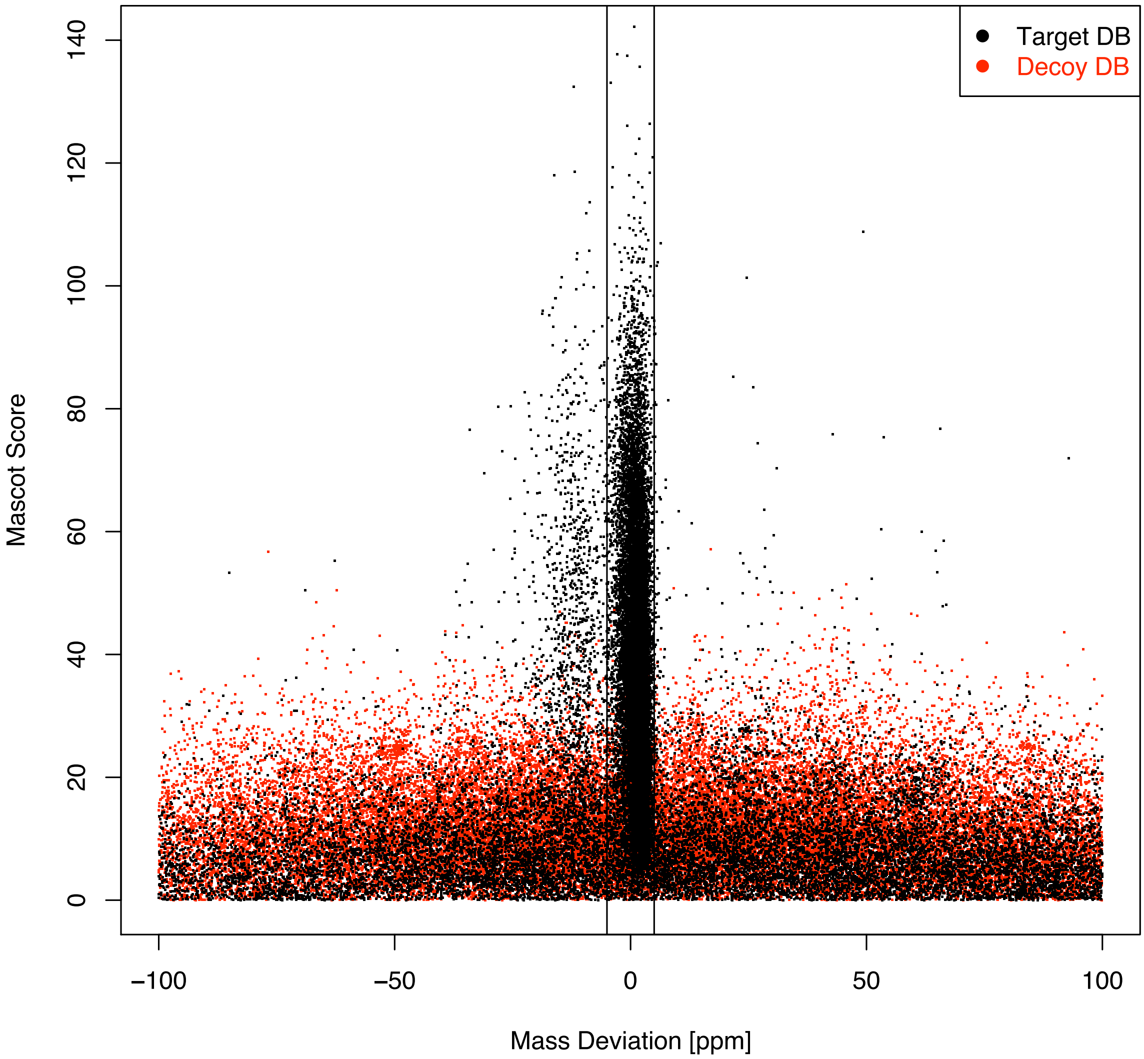


Figure 2.2: Distribution of all peptide matches obtained from a 1 Da MMD target and decoy database search of sample 1, showing the Mascot score and the mass deviations in ppm for a small window of ±100 ppm. Most mass deviations of high scoring peptide-spectra matches fell within the experimental mass errors that have been reported previously, 99% fell within ±20 ppm and 90% fell within ±5 ppm. The mass outliers between -5 and -20 ppm seem to be an experimental artefact for this particular sample.

peptide assignments. The rationale behind this is that the chance of finding a strong peptide match in a relaxed mass window with many peptide candidates is greater than for a very stringent mass window with only a few peptide candidates. A correct and strong match is likely to remain the same, regardless of the size of the search space. On the other hand, it is more likely for a weak match arising from a poor spectra or from an incorrect peptide correlation to find a better alternative in a larger search space. A subsequent mass accuracy filtering step, which limits the matches to the experimental mass deviations, serves as useful discriminator between correct and incorrect matches. This is further illustrated in Figure 2.2 using the data of this study. Overall, these studies indicate that a more detailed evaluation and optimisation of established search algorithms for high accuracy mass spectrometry is still required.

In this chapter I have investigated the performance of Mascot and X!Tandem for varying MMD settings common for low and high accuracy MS. I show that the MIT is highly dependent on the search space and affects false discovery and identification rates. I also show that the empirical scoring scheme in X!Tandem is more robust across different mass tolerance settings. The Mascot equivalent empirical MHT outperforms X!Tandem for ion trap data, but is not comprehensively applicable for very stringent MMD settings. I demonstrate that searching high accuracy data at relaxed MMD windows followed by peptide mass accuracy filtering serves as a good discriminator between correct and incorrect assignments. I propose an alternative empirical Adjusted Mascot Threshold (AMT[1]), applicable to low accuracy data and, in combination with peptide mass accuracy filtering, also to high accuracy data. In addition, the AMT enables the user to freely select the best trade-off between sensitivity and specificity by defining the actual FDR.

Parts of this chapter were published in Molecular Cellular Proteomics (Brosch *et al.*, 2008) by the author of this thesis (Markus Brosch) and my supervisors (Tim Hubbard, Jyoti Choudhary) as well as by Sajani Swamy, who introduced me to the field of computational proteomics. Markus Brosch performed the work and wrote the manuscript. Lu Yu (acknowledgements) run the mass spectrometry experiments (specifically indicated in the relevant sections).

[1] Same abbreviation used for the accurate mass and time tag approach (Pasa-Tolic *et al.*, 2004)

2.2 Experimental Procedures

2.2.1 Sample preparation

Sample 1: A nuclear protein extract of murine embryonic stem cells (2 mg/mL) was reduced with 1 mM dithiothreitol (Sigma) at 70 C for 10 min, followed by alkylation with 20 mM iodoacetamide (Sigma) at room temperature for 30 min. 10 μg of total protein was separated on a NuPAGE Novex 4-12% Bis-Tris polyacrylamide gel (Invitrogen). The gel was stained with colloidal Coomassie Blue (Sigma). The entire gel lane was excised into 48 bands, de-stained with 50% acetonitrile and subsequently digested with sequencing grade trypsin (Roche) overnight. Peptides were extracted with 5% formic acid / 50% acetonitrile twice and vacuum dried in a SpeedVac (Thermo Fisher Scientific). Peptides were redissolved in 0.5% formic acid and subjected to LC-MS/MS. This work was carried out as part of my two month web-lab rotation and was guided by Mercedes Pardo (Team 17 at the Wellcome Trust Sanger Institute).

Sample 2: A standard protein set of 48 human proteins (Sigma, Universal Proteomics Standard Set UPS1) was reduced with Tris(2-carboxyethyl)phosphine hydrochloride (TCEP), alkylated with iodoacetamide as above, followed by digestion in solution with sequencing grade trypsin (Roche Applied Science) overnight. To minimise the chance of detection of low abundance contaminants in the protein standard sample, a very low concentration of 10 fmol (per protein) was directly subjected to the LC-MS/MS. This work was carried out by Lu Yu (Team 17, Wellcome Trust Sanger Institute).

2.2.2 LC-MS/MS analysis

Peptides were analysed with on-line nanoLC-MS/MS on a LTQ FT (Thermo Fisher Scientific), a hybrid linear ion trap and a 7 Tesla Fourier transform ion cyclotron resonance mass spectrometer, coupled with an Ultimate 3000 Nano/Capillary LC

System (Dionex).

Samples were first loaded and desalted on a trap (0.3 mm id x 5 mm) at 20 μL/min with 0.1% formic acid for 5 min then separated on an analytical column (75 μm id x 15 cm) (both PepMap C18, LC Packings) over a 30 min linear gradient of 4-40% CH_3CN/0.1% formic acid. The flow rate through the column was 300 nL/min. The LTQ FT mass spectrometer was operated in standard data dependent mode controlled by Xcalibur 1.4 software. The survey scans (m/z 400-2000) were acquired on the FT-ICR at a resolution of 100,000 at m/z 400 and one microscan was acquired per spectrum. The top three (top five for sample 2) most abundant multiply charged ions with a minimal intensity at 1000 counts were subject to MS/MS in the linear ion trap at an isolation width of 3 Th.

Precursor activation was performed with an activation time of 30 msec and the activation Q was set at 0.25. The normalised collision energy was set at 35%. The dynamic exclusion width was set at ±5 ppm with 2 repeats and a duration of 30 sec. To achieve high mass accuracy, the automatic gain control (AGC) target value was regulated at 4E5 for FT and 1E4 for the ion trap, with a maximum injection time of 1000 ms for FT, and 100 msec for ion trap respectively. The instrument was externally calibrated using the standard calibration mixture of caffeine, MRFA and Ultramark 1600.

All LC and MS related work was carried out by Lu Yu (Team 17, Wellcome Trust Sanger Institute) and was used to introduce me to the basics of practical mass spectrometry during my wet-lab rotation project.

2.2.3 Raw data processing

LTQ FT MS raw data files were processed to peak lists with BioWorks 3.2 (Thermo Fisher Scientific). Parameters were as follows: precursor masses were set to 800-4500 Da, grouping was enabled allowing 50 intermediate scans, and a precursor mass tolerance setting of 10 ppm in BioWorks was applied. The number of minimum scans per group was set to 1. For sample 2 grouping was disabled.

RAW data, peak lists (with and without mass error correction) and Mascot results for both samples are available through ftp under the address: `ftp://ftp.sanger.ac.uk/pub/mb8/mcp2008/`

2.2.4 Database search parameters

Sample 1: Mascot 2.1 (Matrix Science, London, UK) and X!Tandem 2007.07.01 (The Global Proteome Machine Organization) were used for analysing the data. Parameters used in Mascot and X!Tandem searches were: enzyme = trypsin; variable modifications = carbamidomethylation of cysteine, oxidation of methionine; maximum missed cleavages = 1; peptide mass tolerance settings/windows were as indicated in the individual experiments (between 2 Da and 5 ppm); product mass tolerance = 0.5 Da. Probability p of random matches for MIT calculations in Mascot was set to the default value of 0.05.

Specific X!Tandem parameters were: spectrum dynamic range was set to 1000, refinement was disabled, maximum valid E-value for reported peptides was set to 100 (E-values were limited in the data analysis steps) and cyclic permutations to compensate for small search spaces was enabled, with remaining parameters at default.

The protein sequence database used by Mascot and X!Tandem was built from an non-identical superset of Ensembl peptides, UniProtKB and RefSeq sequences for *Mus musculus*, including common external contaminants from cRAP (a maintained list of contaminants, laboratory proteins and protein standards provided through the Global Proteome Machine Organization, `http://www.thegpm.org/crap/index.html`) and contains 94,524 sequences and 42,765,694 residues. For false positive discovery assessment, a separate decoy database was generated from the target database using the Perl decoy.pl script provided by MatrixScience. This script randomises each entry, but retains the average amino acid composition and length of the entries. 0.1%

of sequences were common in both target and decoy database, including K/Q and L/I isoforms that are indistinguishable above 0.04 Da MMD.

Peak lists of sample 2 (8,190 spectra) were searched with Mascot and X!Tandem against human IPI (June 2007, 68,322 sequences, 28,806,780 residues) including common external contaminants from cRAP. To minimise unexpected contaminants from the protein standard set (Klimek *et al.*, 2007), a very low concentration of 10 fmol was used. Parameters used: enzyme = trypsin; variable modifications = carbamidomethylation of cysteine, oxidation of methionine and deamidation of aspargine and glutamine; maximum missed cleavages = 2; peptide mass tolerance = 1 Da; product mass tolerance = 0.5 Da. A random and a reversed version of the sequence database was generated and searched under the same conditions.

2.2.5 Data analysis

Mascot results ($p < 1.0$) were exported to pepXML using the Mascot export tool and X!Tandem results (E-value < 100) were stored as X!Tandem XML. An in-house Java tool was used for the data analysis. Results from Mascot and X!Tandem were imported and filters on score thresholds and mass tolerances were applied. Only doubly and triply charged ions and the first hit rank per spectrum were considered for analysis.

For FDR estimation I chose to search the target and decoy database separately to avoid affecting the MIT scoring by changing database size. The decoy database used was a randomised version of the target database, which was found to be the best approximation based on evaluations of sample 2 (see figure 2.3). All estimated FDRs in this work were calculated using the same target/decoy approach, enabling consistent comparison of results.

Estimated FDRs were calculated by counting all peptide assignments obtained from the decoy database (proxy for false positives, FP), divided by the number of peptide assignments that were obtained from the target database (TP+FP), given

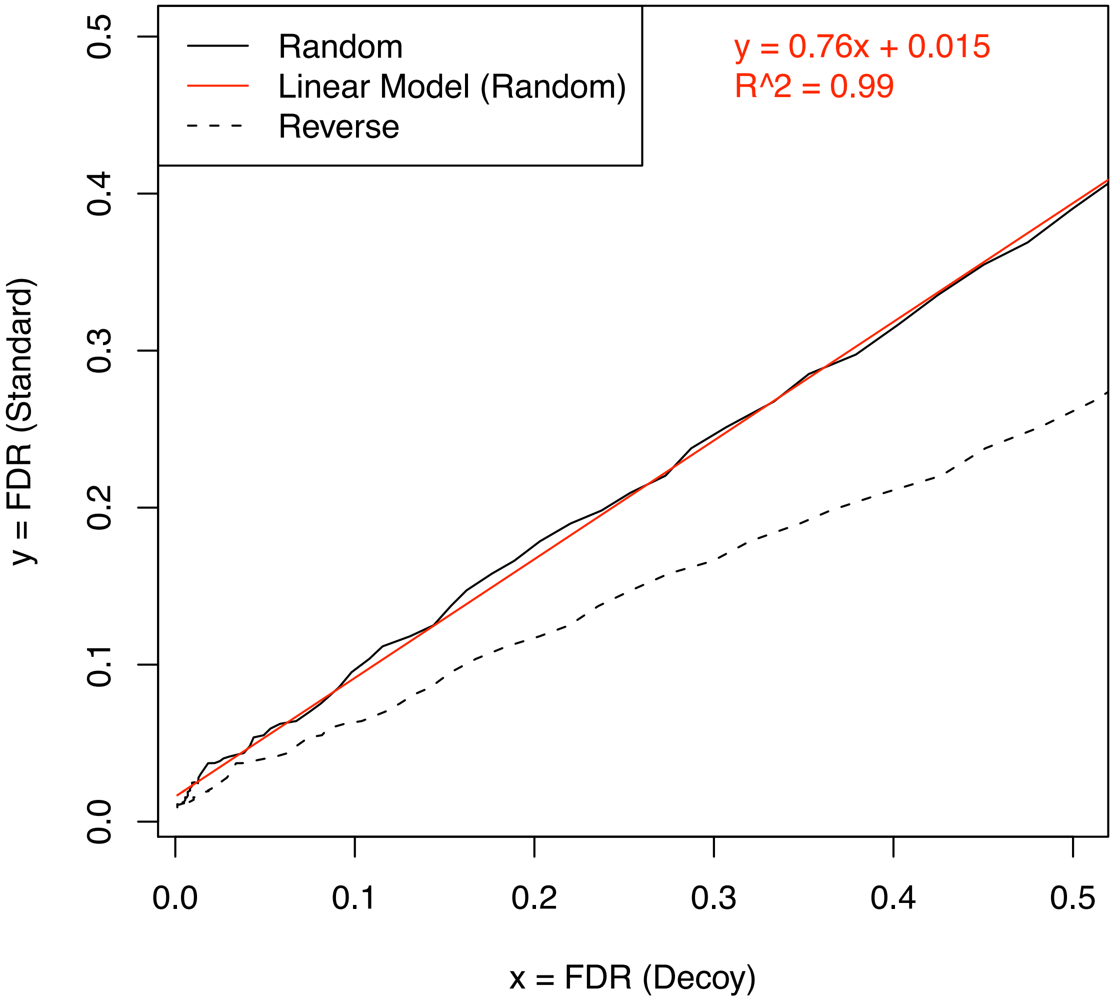


Figure 2.3: An experimental FDR, based on the known proteins of the set, can be determined as follows: any peptide hit that did not match against any of the 48 standard proteins or any of the external contaminants was considered a false positive hit. The FDR rates were determined for a range of Mascot score cutoffs (10-50). Similarly, the estimated FDRs based on target/decoy searching were determined for both the randomised and reversed database. This enabled a comparison of actual FDRs with estimated FDRs, which is interesting since there is no consensus in the proteomics community concerning the different decoy strategies (discussed in section 1.1.2.3). Nevertheless, both decoy strategies (randomized/reversed) tested in this work show a linear relationship between the FDR determined by the protein standard and the target/decoy estimation, validating the target/decoy approach. However, the FDRs derived by the random database were closer to what was reported by the protein standard, which let me to chose the random database as a decoy database for this study. The linear regression of the random database ($R^2 = 0.99$) indicates a small offset of 1.5% which can be explained by unexpected contaminations in the protein standard.

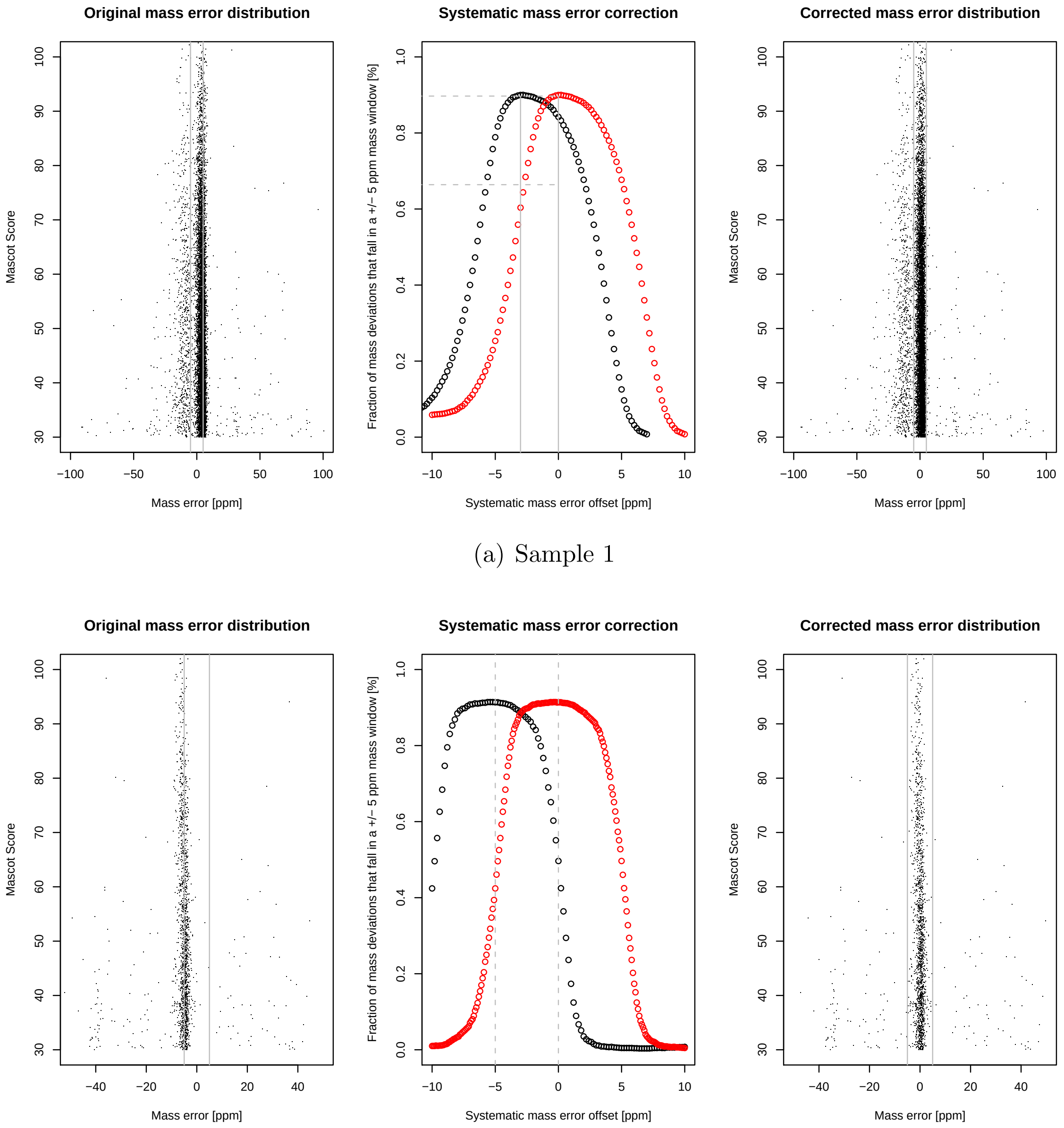
the same parameter and threshold settings. The estimated number of true positive hits (TP) was calculated by counting the number of all peptide hits against the target database minus all hits against the decoy database search. FDR assessment was limited to the peptide level only, since I was interested in the quality of matching individual spectra to peptide sequences. Furthermore it avoids comparison of protein inference strategies (Nesvizhskii and Aebersold, 2004), which is a separate issue.

2.2.6 Correction of systematic mass error

Data from sample 1 was searched in a first pass with Mascot at 100 ppm MMD in order to determine the mass accuracy for the experiment. Only peptide hits with a Mascot score greater than 30 were used for the mass accuracy assessment (10634 queries) to exclude mass deviations of incorrect matches. 99% of hits had mass deviations within a (3±20) ppm mass window (systematic mass error ± peptide mass error), while 90% of mass deviations fell within (3±5) ppm. In order to allow the best possible mass tolerance settings of (0±5) ppm in Mascot and X!Tandem, the precursor masses were corrected by 3 ppm (figure 2.4a). A similar mass error correction method was described by Zubarev and Mann (2007). The mass outliers between -5 and -20 ppm seem to be an experimental artefact for this particular sample. For this study I deliberately accepted a loss of identifications for 5 ppm MMD settings in order to study the effects of stringent mass settings on Mascot and X!Tandem. Mass error correction was applied in the same way to sample 2, where the peptide masses were corrected by 5 ppm (figure 2.4b).

2.3 Results and discussion

If not stated otherwise, all subsequent results are based on sample 1, which is a large complex dataset and representative of typical proteomics experiments.



(a) Sample 1

(b) Sample 2

Figure 2.4: Mass error determination and correction of systematic mass errors. Left: the original mass deviations of all highly significant peptide matches. Centre: Systematic mass error correction that maximises the peptide assignments within a 5 ppm mass window. Right: After correction of the systematic mass error.

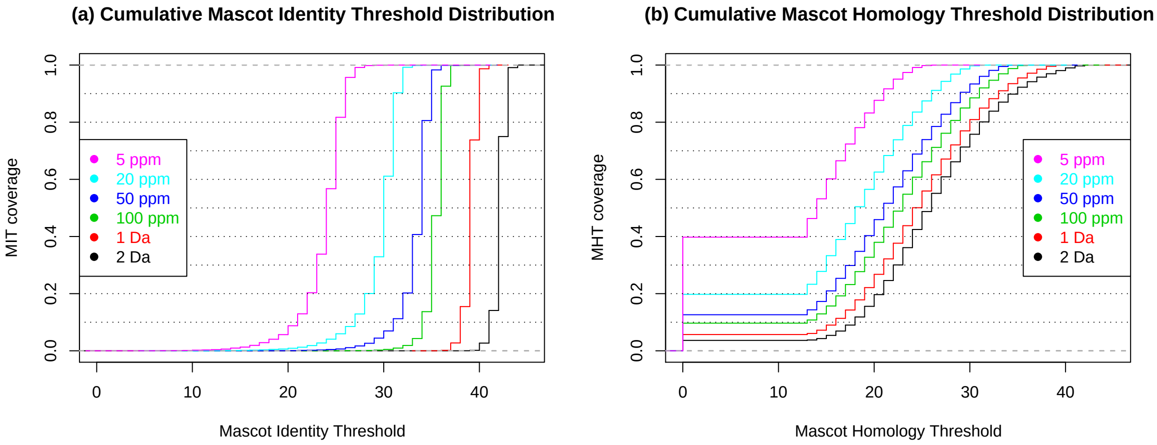


Figure 2.5: (a) Cumulative MIT distributions for different peptide mass tolerance settings. Only MITs from queries with a peptide assignment across all searches were used to enable comparison. With more stringent MMD settings, the MIT tends to decrease, accommodating for the smaller search space. *Vice versa* it increases for more relaxed MMD windows. (b) Cumulative MHT distributions over the range of MMD settings. The MHT is not reported for every query. All MHTs exceeding the MIT are omitted by Mascot and reported as 0 in the HTML and XML result files (personal communication, John Cottrell, Matrix Science). The minimum MHT reported by Mascot is 13 and the maximum MHT is limited by the corresponding MIT.

2.3.1 Performance of the Mascot Identity Threshold

Mass error corrected spectra were submitted to Mascot and searched at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm MMD settings, while all other parameters were fixed.

Spectra that were assigned across all searches (23,080 out of 38,058 queries) were used to draw the MIT distribution for each MMD setting (Figure 2.5a). From this analysis the median MIT values for relaxed MMD settings were 42 at 2 Da MMD and 39 at 1 Da MMD with an inter-quartile range of 1. Under more stringent settings (5 ppm) the MIT median decreased to 24 while the inter-quartile range increased to 2. These results suggest that the MIT adapts with changing search space and performs more like a global cut-off based on the narrow variation in thresholds.

To evaluate the effects of MIT adaptions on the peptide identifications performance at different MMD settings, the rates of incorrect and correct peptide-spectrum

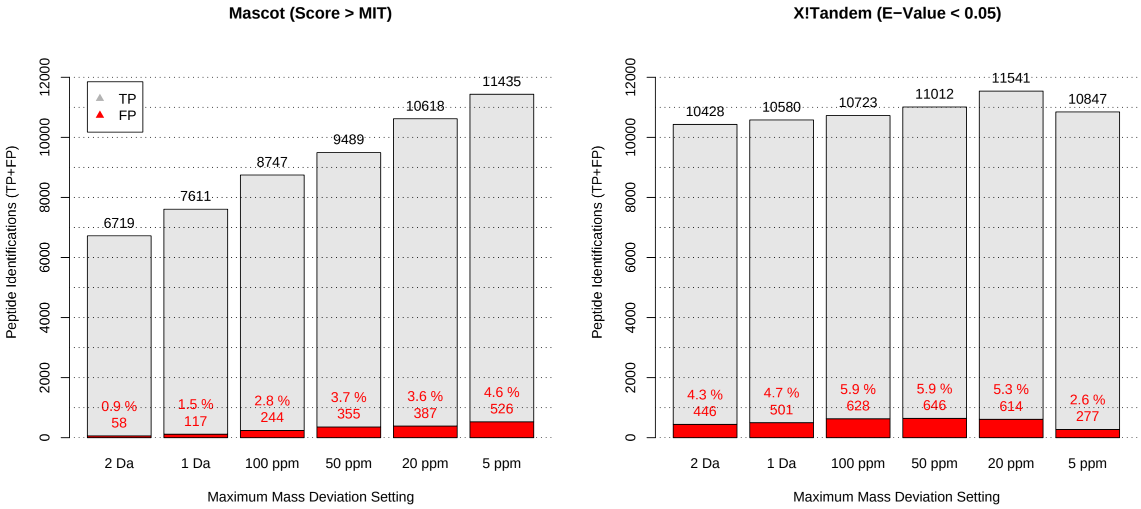


Figure 2.6: Comparative evaluation of Mascot and X!Tandem performance. Mascot and X!Tandem searches were performed against a target and decoy database at different MMD settings. The total number of identifications is reported, the estimated number of true identifications is indicated in grey, while the estimated number of incorrect assignments is highlighted in red.

matches were determined by target/decoy FDR estimations, under identical search and threshold parameters for all spectra (Figure 2.6, Mascot). Using the MIT as a score cut-off, 10,909 and 6,661 estimated TP peptide identifications were made at 2 Da and 5 ppm MMD settings respectively. Relative to the 5 ppm search, this suggests 4,248 (39%) false negative peptide assignments for the 2 Da search. For the same MMD settings, the FDR increased from 0.9% to 4.6% respectively, failing to maintain the specified (5%) rate of random (incorrect) assignments.

The MIT is based on a probabilistic model that attempts to maintain a constant rate of random (false) identifications and hence is dependent on search space. However, I found a correlation between FDRs and MMD settings, indicating that the MIT does not adhere to the predefined FDR. This trend is also mirrored in the number of correct identifications. At relaxed mass tolerances (large search space) used for ion trap data, the MIT tends to become very conservative resulting in excellent specificity but hindering sensitivity. With more stringent mass tolerances (smaller search space) sensitivity increases at the cost of specificity. The results reported here represent a snapshot of many possible combinations of search parameters that directly affect the search space, for example: sequence database size, allowed variable modifications, allowed missed cleavages and enzyme specificity. This highlights the necessity to individually assess the FDR via a target/decoy database search.

2.3.2 Performance of the X!Tandem scoring scheme

Spectra were searched in X!Tandem using MMD settings as described in the previous section. FDRs were calculated on the basis of target and decoy database searches using identical search parameters.

Using an E-value cut-off value of 0.05, which is in-line with that used for the MIT evaluation discussed above, only moderate changes (9%) in sensitivity over all MMD settings were detected, varying between 9,982 TPs at 2 Da and up to 10,927 TP at 20 ppm (Figure 2.6, X!Tandem). A constant FDR for varying MMD settings was not delivered by X!Tandem. The FDRs increased from 4.3% to 5.9% between the 2 Da and 100 ppm MMD, and an inverse trend was observed below 100 ppm, with a minimum of 2.6% FDR at 5 ppm MMD. FDRs show no clear correlation with mass tolerance settings, suggesting no direct dependency. The E-Value distributions of these searches were very similar over the whole range, further supporting the robustness of the X!Tandem scoring (figure 2.7).

Overall, X!Tandem appears to maintain sensitive peptide identification at varying MMD settings. The FDRs were close to the defined E-Values, but were not constant over changing mass tolerance settings. However, there appears to be no direct correlation between the FDRs and search space. These results indicate that the empirical X!Tandem scoring, based on peptide-spectra match score distributions, is more robust over the search space dependent probabilistic scoring model of the MIT.

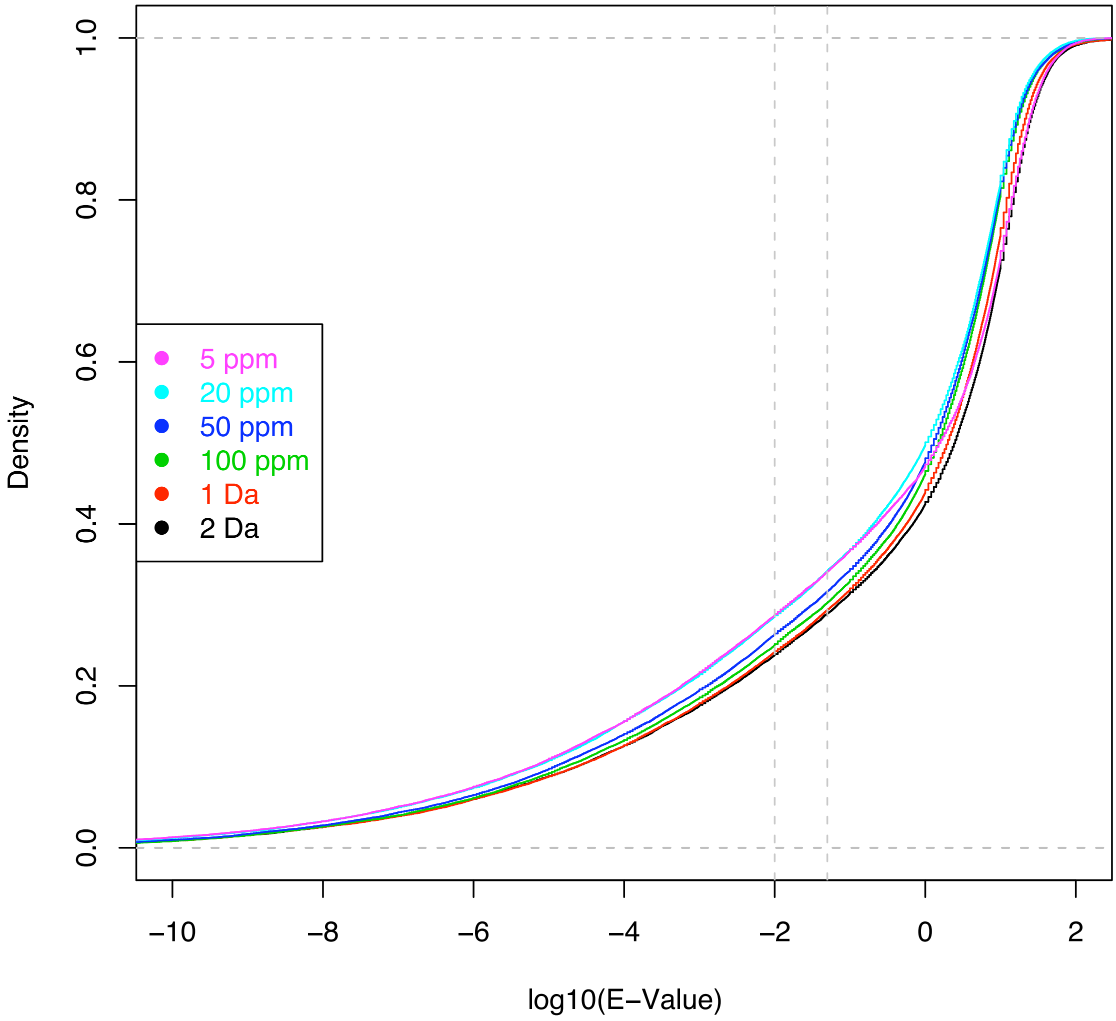


Figure 2.7: Spectra were searched with X!Tandem at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm MMD settings, while all other parameters were fixed. For each search the E-value distribution was drawn, indicating that the X!Tandem scoring is very robust over changes in search space. The E-values 0.01 and 0.05 are highlighted. The plot is in concordance with the ROC curve presented in the paper. Personal communication with Dr David Fenyo (The Rockefeller University) explained the robustness of the E-value distributions: Each E-value depends on the survival function and on the number of sequences scored (Fenyo and Beavis (2003), equation 2). For X!Tandem in its current format, the term "number of sequences scored" refers to the whole sequence database, regardless of the peptide mass tolerance setting and hence all variations seen in the E-values are the result of the slight differences in survival functions only.

2.3.3 Performance of the Mascot Homology Threshold

Similarly to X!Tandem, the empirical MHT also utilises peptide-spectra match score distributions. Using the results from the above Mascot searches, I plotted MHT distributions at different MMD settings (Figure 2.5b). Only spectra that were assigned across all searches (23,080 out of 38,058 queries) were used for comparison.

As stated earlier, the MHT is not always reported. A MHT value was reported for about 95% of the considered queries at relaxed MMD settings of 1 or 2 Da. For stringent MMD settings (5 ppm), MHTs were only reported for less than 60% of queries, limiting its applicability. The MHT median for a 1 Da MMD setting was 24, compared with a MIT median of 39 for the same setting, while the inter-quartile ranges were 9 and 1, respectively. The wide MHT variation observed would be reflective of a query specific thresholding.

Using the MHT as a cut-off score for a 1 Da MMD search, 11,315 TPs were identified at the given FDR of 3.1%. This corresponds to 51% more TP identifications than using the MIT at the given 1.5% FDR and 12% more TP identifications than X!Tandem at the given FDR of 4.7%.

Overall, I observe the MHT to be significantly more sensitive than the MIT and X!Tandem at the given FDRs. However, the FDR is pre-imposed and does not allow the user to select a fixed rate. Furthermore, Mascot omits any MHT which exceeds the MIT to prevent conservative thresholds that arise, for example from score distributions with insufficient data points. This effect is further compounded since the MIT values decrease for a smaller search space. Sufficient search space is required for the MHT to be comprehensively applicable, for example a larger or smaller database would need a more or less restrictive MMD setting to compensate for this effect.

2.3.4 Peptide mass accuracy filtering

An alternative approach for using high mass accuracy for peptide identification is to search under relaxed mass tolerance settings and subsequently apply mass accuracy filters. To evaluate this approach, data was searched with Mascot at a 1 Da MMD setting against the target and decoy databases, where approximately 95% of queries obtained a MHT.

As shown in figure 2.2, peptide-spectrum matches with high scores mostly lie within the experimental mass errors discussed previously, while low scoring matches were distributed evenly across the whole mass window. Mass accuracy filtering of the 1 Da search using 50, 20 and 5 ppm cut-offs, without imposing any other constraints such as MIT or MHT, limits the FDRs to 65%, 35% and 12% respectively. This clearly indicates that mass accuracy based filtering alone can reduce incorrect sequence assignments. However, the effectiveness of this discriminator is confined by experimentally derived mass error deviations. Significantly, 13,273 TP were identified with a 5 ppm mass filter, more than obtained by any method tested here, showing this to be a very sensitive approach for peptide identification with high accuracy data.

The 12% FDR observed at 5 ppm mass accuracy filtering suggests that even higher mass accuracy would be required for lower FDRs. An extrapolation from a regression over 10 data points ranging from 5 to 50 ppm ($r^2 = 0.99$) suggests a 5% FDR for 1.5 ppm, however this prediction would need to be verified experimentally. It should be noted that the use of ultra high mass accuracies cannot further improve FDR once mass accuracies resolve elemental compositions.

If mass accuracies cannot be achieved at this stringent level, an alternative would be to introduce a moderate thresholding on the peptide-spectrum match scores. I therefore tested mass accuracy filtering in combination with the MHT score cut-off. For this, data was searched at 1 Da MMD, then filtered at 5 ppm to exclude all peptide assignments with a larger mass deviation, and subsequently constrained by the MHT. In instances where the MHT was not reported, the MIT was used. This two-step filtering identified 10,338 TP peptide assignments and reduced the FDR to only 0.2%, which is a 60-fold improvement over the mass accuracy filtering alone, although the TPs were reduced significantly (22%). In comparison with the Mascot search using 5 ppm MMD setting with the MIT score cut-off, where a FDR of 4.8% and 10,909 TP was previously reported, the two-step filtering improved the FDR by 23-fold, while the TPs were reduced by only 6%.

These results suggest that mass accuracy filtering on its own might be a valuable and very sensitive approach, however sub-ppm mass errors would be needed for highly specific identification. Alternatively, a combination with a threshold such as the MHT serves as a very strong discriminator between correct and incorrect peptide assignments. In comparison with a direct high accuracy Mascot search, the two-step filtering strategy leads to highly specific identifications without significantly compromising sensitivity. A less restrictive and adjustable thresholding would increase sensitivity for peptide identification from high accuracy data.

2.3.5 The Adjusted Mascot Threshold (AMT)

Applying either the MIT, MHT or the two-step filtering provides pre-imposed FDRs that are not directly adjustable by the user. However, it is often desirable to be able to select and fix the FDR.

To achieve this I have implemented the Adjusted Mascot Threshold (AMT). This is a similar strategy to the MATH threshold introduced by Rudnick *et al.* (2005), which uses a global threshold that defines a cut-off value for all queries. However, I favour the use of individual query specific thresholds based on the MHT, since I have found it to be very sensitive in my above evaluations. The AMT is defined as the sum of the query specific MHT and a global offset value. FDRs are determined for a range of offset values that are used to calculate a linear regression in order to approximate an offset value for a user defined FDR (Figure 2.8).

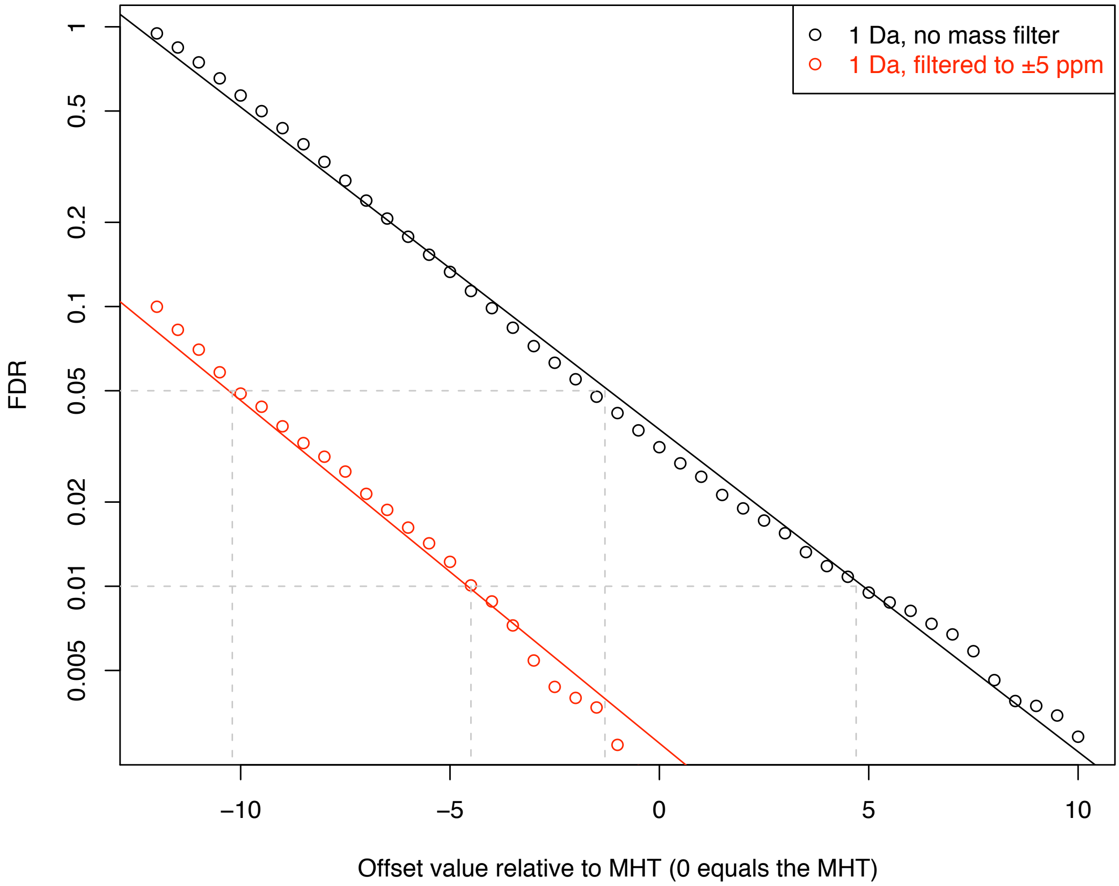


Figure 2.8: Regression for extrapolating the AMT thresholds. Data was searched at a 1 Da MMD setting against the target and decoy database. A range of offset values was applied that were added to the MHT and used as cut-off thresholds. For each new threshold the associated FDR was determined. A linear regression between the logarithm of the FDR and the offset values was calculated (r^2=0.99). The method was also applied to the mass accuracy filtered dataset (5 ppm). A new Adjusted Mascot Threshold can be extrapolated based on a user defined FDR for each dataset. The AMT adapts for the preceding mass accuracy filtering. The offset values for a FDR value of 1% and 5% are indicated as dashed lines.

For the 1 Da search, described in the previous section, the regression was calculated for an offset range of -12 to +10, indicating a strong linear correlation between the logarithm of FDRs and the offset values with a correlation coefficient of $r^2 = 0.99$. For the 5 ppm mass filtered dataset, a second regression was calculated ($r^2 = 0.99$). Offset values of 4.7 and -1.3 were reported for a target FDR of 1% and 5% using the 1 Da search data and for the 5 ppm mass accuracy filtered dataset these values were -4.5 and -10.2 respectively. The slope of both regressions was found to be very similar, but the difference between the offsets was approximately -9, which compensates for the inherent specificity of the mass accuracy filtered dataset by moderating these offset values.

Our proposed AMT is an adjustable and query specific cut-off value. It is calculated based on the MHT and a global offset value, the latter is derived from FDR estimates through target/decoy database searching and thus is no longer dependent on search parameters affecting search space. AMT can be extrapolated for either low or high accuracy (using mass filtered data), and combines the benefits of a highly sensitive MHT with a user defined FDR.

2.3.6 Comparison of the AMT with MIT, MHT, MATH and X!Tandem

I then tested the performance of the AMT. Search results obtained by application of AMT were compared to those from MIT, MHT, X!Tandem and MATH using a receiver operator characteristic (ROC) representing the number of true identifications at various FDRs. ROC curves (Figure 2.9) were calculated using varying thresholds of MATH (global cut-off value), X!Tandem (E-values) and AMT (offset values relative to MHT). Since the MIT and MHT are not variable, they define a single point in the diagram.

For low accuracy MMD settings (Figure 2.9a) applying the MIT identified 7,494

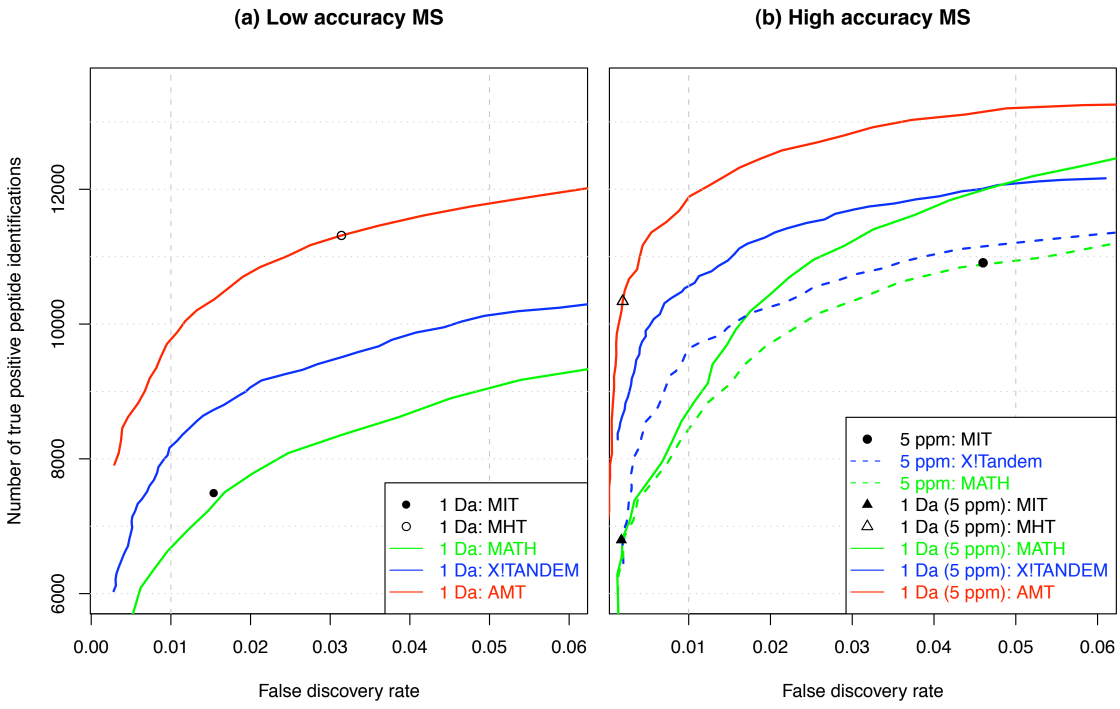


Figure 2.9: MIT, MHT, MATH, X!Tandem and AMT comparison for low and high accuracy mass tolerance settings. A 1 Da search (a), a 5 ppm search (b, dashed lines) and a 1 Da search with subsequent peptide mass accuracy filtering at 5 ppm (b, solid lines) were performed. The estimated number of TPs was determined as a function of the FDRs, represented in the receiver operator curve, enabling the user to choose where the best trade-off between sensitivity (TPs) and specificity (FDR).

TP with an inherent 1.5% FDR. MIT variation for these mass tolerance settings effectively acts as a global cut-off, hence MATH also identified a similar number at the same FDR. MATH however allows the user to freely select the target FDR, and at a 5% FDR it identified about 20% more TP peptides than at 1.5% FDR. X!Tandem empirical scoring outperformed both MIT (13% more TP at the same FDR of 1.5%) and MATH (between 10-15% more TP over the whole range of FDRs). The most striking observation is the MHT performance, identifying 11,315 TPs at the inherent FDR of 3.1%, improving correct identifications by 18% and 35% over X!Tandem and MATH at the same FDR. The AMT extends application of the MHT over the whole range of FDRs, improving the TP assignments by 18%, 39% and 42% over X!Tandem, MIT and MATH at 1.5% FDR, and by 16% and 30% over

X!Tandem and MATH at 5% FDR.

For the analysis of high accuracy data I have evaluated two strategies (I) searching high accuracy data at stringent mass tolerance settings (5 ppm) followed by peptide score thresholding (Figure 2.9b, dashed lines), and (II) searching high accuracy data at a relaxed mass window (1 Da) with subsequent peptide mass accuracy filtering (5 ppm) followed by peptide score thresholding (Figure 2.9b, solid lines).

(I) Using direct high mass accuracy searching at 5 ppm MMD setting, the number of expected true peptide identifications was similar, approximately 11,000, for MIT, X!Tandem and MATH at around 4.5% FDR. However, X!Tandem performed better for lower FDRs, e.g. at 1% X!Tandem identified about 1,000 more TPs than MATH. MHT was not assessed at these mass tolerance settings since it was absent for 40% of queries.

(II) The alternative mass filtering approach returned very conservative FDRs below 0.2% and identified 6,798 and 10,338 TP hits for the MIT and MHT respectively. Mass filtered X!Tandem results identified approximately 25% more peptides than the MIT and 18% less TP hits than with the MHT, at the corresponding FDRs. By relaxing the E-values of X!Tandem, 10,611 TP at 1% FDR and 12,100 TP at 5% FDR were identified. Using MATH, 6,821 TP assignments were made at the 0.2% FDR, which is again similar to MIT and significantly worse than X!Tandem or MHT. At a 1% FDR about 18% fewer identifications were made using MATH as compared to X!Tandem, while they performed similarly at 5% FDR. Significantly, the AMT identified 11,893 TP assignments at 1%, outperforming both MATH and X!Tandem by 35% and 12% at the same FDR.

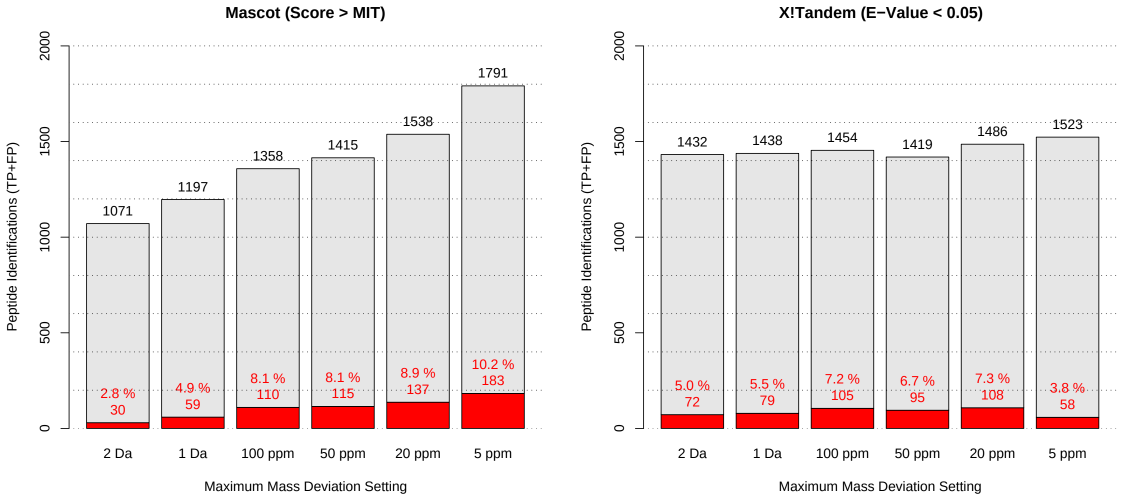
Compared to the direct 5 ppm search strategy in (I), the mass accuracy filter approach in (II) was generally more sensitive, e.g. MATH and X!Tandem with mass filtering identified about 8-9% more TPs at 5% FDR than without mass filtering. The improvement of performance with X!Tandem can be seen throughout the whole range of FDRs, whereas for MATH sensitivity is only gained above a 1% FDR. By far the most sensitive approach at any given FDR was provided by mass accuracy filtering combined with the AMT. Against a direct 5 ppm search using MIT, MATH and X!Tandem, about 18-20% more TPs at a FDR of 4.6% were made, which corresponds to approximately 1,500 more unique peptides identifications.

In summary, application of MIT or MHT always results in a fixed pre-imposed FDR, while X!Tandem together with a target/decoy database search enables FDR adjustment using an appropriate E-value cut-off. MATH and AMT implement this target/decoy FDR estimation and directly deliver the defined FDRs. For low accuracy MS, MHT performed best at a fixed FDR, whilst this performance was extended to the whole FDR range by AMT. X!Tandem was significantly less sensitive than AMT, and MATH together with the MIT were the least sensitive thresholds. For direct high mass accuracy searching, MIT, MATH and X!Tandem performance was very similar and overall sensitivity improved over the low accuracy search. Exploiting high mass accuracy via mass filtering was the most sensitive search strategy at the corresponding FDRs. For this approach, AMT significantly outperformed X!Tandem, followed by MATH and MIT.

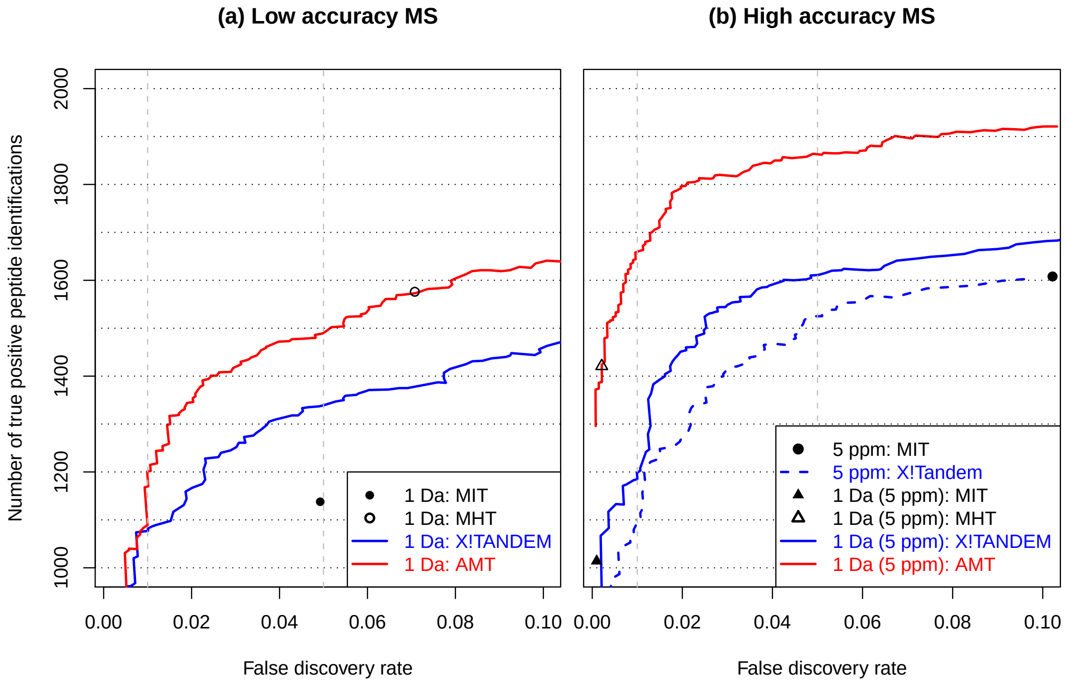
2.3.7 Validation with independent dataset

To validate the findings and the AMT performance, a standard mixture of 48 proteins (sample 2) was analysed in the same way as sample 1. First, data were searched against a 50 ppm peptide mass tolerance to identify any systematic mass error (Figure 2.4b), which was corrected (-5 ppm) subsequently.

Next, data were searched at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm peptide mass tolerances and the FDRs were determined accordingly (figure 2.10a). The same FDR trends as for sample 1 (figure 2.5) were observed: using the MIT resulted in the FDR being dependent on the search parameters used, rising from 2.8% to over 10% when the mass tolerance window was narrowed from 2 Da to 5 ppm. However, X!Tandem was shown to be quite robust again, indicating little



(a) Comparative evaluation of Mascot and X!Tandem performance and FDR robustness for sample 2. Compare with figure 2.6.



(b) Comparison of the performance of X!Tandem and Mascot using the MIT, MHT and AMT for sample 2. Compare with figure 2.9. The vertical dashed lines correspond to commonly used 1% and 5% FDR values.

Figure 2.10: Validation of results on an independent protein standard dataset.

dependance of the search space on the scoring scheme.

ROC curves were compiled to enable comparison (Figure 2.10b) of the AMT and the standard Mascot thresholds as well as the X!Tandem performance. Again, the MHT was shown to be significantly more sensitive than the MIT, but the AMT scoring method clearly outperformed the MIT and MHT as well as X!Tandem, validating the findings of sample 1.

2.4 Conclusion

In this chapter I have investigated how MMD settings affect peptide identification using Mascot and X!Tandem and presented an alternative search strategy and an Adjusted Mascot Threshold (AMT) to enable sensitive identification of high accuracy data with Mascot.

I have demonstrated the correlation between the MIT and search space, which is for example affected by MMD settings. I have shown that the MIT can be very conservative for MMD settings commonly used for ion trap data, leading to very specific identifications at the expense of sensitivity, while it tends to become more optimistic for stringent MMD settings used for high accuracy data. The MHT was found to be significantly more sensitive for ion trap data, but is not comprehensively applicable to very stringent MMD settings commonly used for high accuracy data. However, the actual FDRs for both MIT and MHT are pre-imposed and deviate from the theoretically defined rate. Furthermore, my results indicate that X!Tandem is more robust than the MIT and MHT when faced with MMD changes and is equally applicable to both low and high accuracy MS data with a sensitivity that was better than using the MIT but worse than using the MHT.

I also investigated the use of mass accuracy filtering as the sole discriminator between correct and incorrect peptide assignments. Mass accuracy filtering served as a highly sensitive discriminator with limited specificity and sub-ppm mass errors would be needed for more specific identifications. Alternatively, a two-step filtering strategy can be employed. I first searched the data at relaxed MMD settings, followed by applying mass accuracy filtering. The results demonstrate that combining peptide mass accuracy filtering with the MHT serves as a very strong discriminator, efficiently eliminating incorrect peptide assignments, although sensitivity was limited. To regain sensitivity I propose an Adjusted Mascot Threshold (AMT) that allows the user to freely select the best trade-off between sensitivity and specificity by having full control over the actual FDR. The AMT can easily be applied on top of any Mascot search where target/decoy searching is amenable. It is independent of search parameters affecting the search space and is expected to adjust with MS/MS data quality. AMT outperforms MIT and MHT, as well as MATH and X!Tandem for both low and high accuracy MS data.

Chapter 3

Accurate and sensitive peptide identification with Mascot Percolator

3.1 Introduction

With the advent of high accuracy instrumentation, it was anticipated that peptide identification specificity would improve, since peptide mass accuracy in the region of a few ppm reduces the search space by orders of magnitude (Elias and Gygi, 2007; Zubarev and Mann, 2007; Zubarev *et al.*, 1996). However, in chapter 2 I have evaluated the performance of Mascot and demonstrated that this is not necessarily the case. I have shown that the Mascot Identity Threshold (MIT) was anti-conservative (low specificity, but high sensitivity) for stringent peptide mass tolerance settings (small search space) and conversely very conservative (high specificity, but low sensitivity) for relaxed parameter settings. Mascot also reports an empirical Mascot Homology Threshold (MHT) at which a Mascot score can be considered a significant outlier from the score distribution of all peptide matches to a given spectrum. Overall, the MHT was shown to be more sensitive than the MIT, but is only reported for peptide-spectrum matches (PSMs) where sufficient peptide candidates are scored, e.g. at relaxed search parameter settings. These findings led me to implement the Adjusted Mascot Homology Threshold (AMT), utilising the MHT at relaxed search parameters that, combined with a peptide mass deviation filter (AMT/mass-filter) on mass error recalibrated data, was shown to be a sensitive Mascot scoring method for high accuracy data (see chapter 2).

However, a limitation of the AMT/mass-filtering strategy is that it requires a fixed mass tolerance filter in order to subsequently determine a score threshold that maintains a predefined FDR. A more flexible implementation would be to use both features, the score cut-off and the mass deviation, in combination for discrimination of correct and incorrect PSMs. This can be achieved using the iterative machine learning method called Percolator (Käll *et al.*, 2007). See section 1.1.2.3 for more background information concerning Percolator.

Although Percolator was originally designed for Sequest use only, the availability of a standard input format enables the use of Percolator as a generic machine learning algorithm where target/decoy data are available. I have therefore developed a Mascot extension ("Mascot Percolator") that extracts and computes relevant features from the Mascot search results, trains Percolator, applies the resulting classifier to each PSM and writes a result file. I firstly assessed the AMT/mass-filtering approach with Mascot Percolator, but also extended this method with more features directly available from Mascot search results, such as Mascot scoring information and peptide properties. Moreover, an extended feature set comprising information not directly accessible from Mascot search results, including ion matching statistics and intensity information, was explored. I have evaluated the performance of Mascot Percolator with high precursor mass accuracy LC-MS/MS datasets, but also benchmarked it with the low mass accuracy LC-MS/MS dataset used in the original Percolator publication. In a final assessment, I validated the q-value accuracy reported by Percolator with a protein standard dataset acquired on a range of instruments. Mascot Percolator is freely available at http://www.sanger.ac.uk/Software/analysis/MascotPercolator/ including databases, peak lists and results as presented in this chapter.

Parts of this chapter were published in the Journal of Proteome Research (Brosch *et al.*, 2009) by the author of this thesis (Markus Brosch) and my supervisors (Tim Hubbard, Jyoti Choudhary) as well as by Lu Yu who run the mass spectrometry experiments (specifically indicated in the relevant section).

Moreover, in collaboration with John Cottrell and David Creasy (Matrix Science, London) my method presented in this chapter is currently implemented into the official Mascot 2.3 software release (http://www.matrixscience.com/workshop_2009.html), and will be readily applicable by the proteomics community without the need of any third party software.

3.2 Methods

3.2.1 Datasets and experimental methods

- Dataset 1: LTQ-FT (Thermo Fisher Scientific) dataset from a nuclear protein extract of murine embryonic stem cells. Data and methods were described in detail and used in chapter 2.

- Dataset 2: Käll *et al.* (2007) provided us with their Yeast (*Saccharomyces cerevisiae*) dataset acquired on an LTQ (Thermo Fisher Scientific).

- Dataset 3: LTQ-FT dataset from a standard protein set comprising 48 human proteins (Universal Proteomics Standard Set UPS-1, Sigma). Data and methods were described in detail in chapter 2. In addition, the same sample was also acquired on a LTQ, LTQ-FT Ultra and Q-Tof Premier (Waters) by Lu Yu (Team 17, Wellcome Trust Sanger Institute), providing me a comprehensive set of protein standard data.

3.2.2 MS/MS database searching

Dataset 1: Peak lists of (38,058 spectra) were searched with Mascot 2.2 using the following parameters: enzyme = trypsin (allowing for cleavage before proline (Rodriguez *et al.*, 2007)); maximum missed cleavages = 2; variable modifications = carbamidomethylation of cysteine, oxidation of methionine; product mass tolerance = 0.5 Da. The International Protein Index (IPI) database version 337 (Mus musculus) was used as a protein sequence database. Common external contaminants from cRAP (a maintained list of contaminants, laboratory proteins and protein standards provided through the Global Proteome Machine Organisation, `http://www.thegpm.org/crap/index.html`) were appended (see 1.1.1.2). The compounded database contained 51,355 sequences and 23,635,027 residues. For FDR assessment, a separate decoy database was generated from the protein sequence database using the "decoy.pl" Perl script provided by Matrix Science. This script randomises each entry, but retains the average amino acid composition and length of the entries. Data was searched at 100 ppm peptide mass tolerance to evaluate data mass accuracy. After a correction of a systematic mass deviation of 3 ppm(Brosch *et al.*, 2008), 90% and 99% of all PSMs with a Mascot score greater than 30 fell within a ±5 ppm and ±20 ppm mass window respectively. For the most stringent mass tolerance settings, where Mascot thresholds are most sensitive, the data was searched at 20 ppm. Moreover, data was also searched at 500 ppm peptide mass tolerance to enable mass accuracy filtering combined with the adjusted MHT (Adjusted Mascot Threshold, AMT (Brosch *et al.*, 2008), see chapter 2). The mass deviation filter was set to 5 ppm, which was shown to be the most effective filter setting in combination with the AMT (figure 3.1).

Dataset 2: Peak lists of (35,236 spectra) were searched with Mascot 2.2. against the same target and decoy databases that were used by Käll *et al.* (2007). The following parameters were used: enzyme = trypsin; maximum missed cleavages = 2; fixed modification = carbamidomethylation of cysteine; peptide mass tolerance settings = 3 Da; product mass tolerance = 0.5 Da.

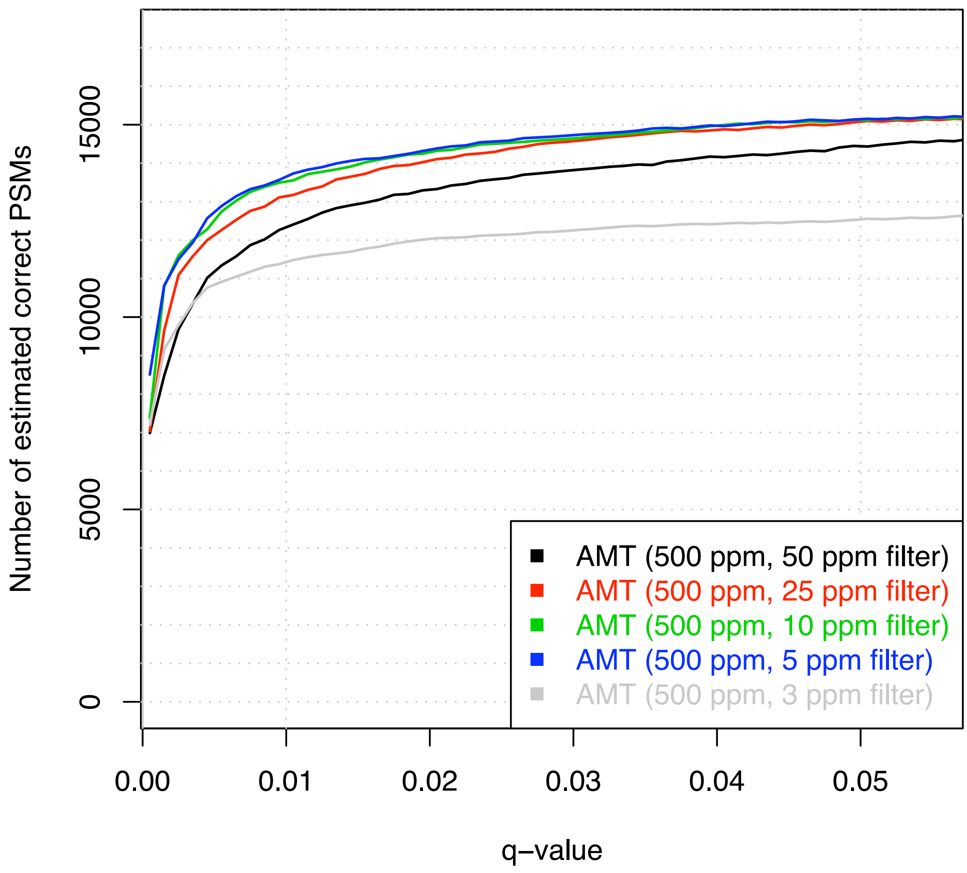


Figure 3.1: The performance of the Adjusted Mascot Threshold (AMT) was evaluated using mass deviation filter settings of 50, 25, 10, 5 and 3 ppm: for each, the number of estimated correct PSMs was determined across a range of q-values. These results show the trade-off between improving specificity with more stringent mass tolerance filters and conversely excluding potentially correct PSMs when the filters become too stringent. For this dataset the best mass filter was found to be 5 ppm.

Dataset 3: Peak lists (spectra count LTQ: 43,710, LTQ-FT: 45,289, LTQ-FT Ultra: 18,285, Q-Tof: 1206) were searched with Mascot 2.2 against human IPI (June 2007, 68,322 sequences, 28,806,780 residues) including common external contaminants from cRAP. Parameters used: enzyme = trypsin; maximum missed cleavages = 2; variable modifications = carbamidomethylation of cysteine, oxidation of methionine and deamidation of aspargine and glutamine; peptide/product mass tolerance = LTQ: 0.9 / 0.5 Da, LTQ-FT: 20, 50, 200 ppm / 0.5 Da, LTQ-FT Ultra: 10 ppm / 0.5 Da, Q-Tof: 30 ppm / 0.2 Da; 5 randomised versions (decoy.pl) of the sequence database were generated and searched individually under the same conditions.

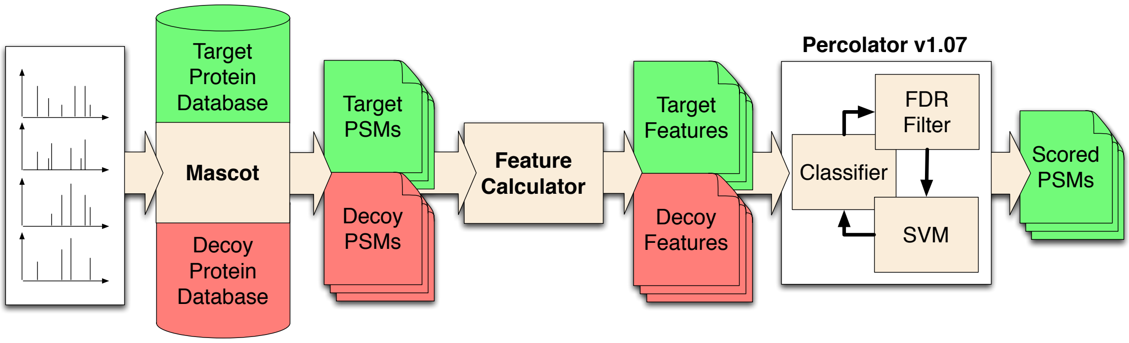


Figure 3.2: Illustration of the Mascot Percolator workflow.

3.2.3 Mascot Percolator implementation

Mascot Percolator was implemented with the Java programming language, ensuring platform independent operation. It utilizes the Mascot Java parser library provided by Matrix Science (`http://www.matrixscience.com/msparser.html`) and uses the generic interface to Percolator (Washington University, `http://noble.gs.washington.edu/proj/percolator/`). The latest Percolator version 1.12 using default parameters was used for this study, which should be taken into account when comparing results of this study to the original publication of Percolator (Käll *et al.*, 2007), where version 1.01 was used. Results in this chapter are based on Mascot Percolator version 1.09.

The Mascot Percolator implementation performs the following operations for each run: it reads the Mascot results files, computes the scoring features as introduced in the results and discussion section and uses these for the Percolator training as described in section 1.1.2.3. In a last step, the result file of Percolator and the input files are merged to combine peptide, protein and scoring information (figure 3.2).

3.2.4 Data analysis

Receiver Operating Characteristics (ROCs) for Mascot Percolator were generated by varying the q-value cutoffs and reporting the corresponding number of estimated true positives. The MIT, MHT and AMT were used as a reference for comparison.

When no MHT was reported, the MIT was used instead, which is the default behaviour of Mascot. ROCs for the MIT and MHT were generated by varying the Mascot significance threshold p (default 0.05) between 1×10^{-5} to 1×10^{1}, the latter representing the maximum allowed. Percolator factors the percentage of target PSMs that are incorrect (π_0) into the q-value calculation (see section 1.1.2.3). For consistency, the q-value calculations of MIT, MHT and AMT also take this factor into account and were determined using "Qvality": 0.55 (dataset 1), 0.5 (dataset 2), 0.77 (dataset 3).

3.3 Results and Discussion

3.3.1 Peptide mass accuracy features

Dataset 1 is representative of a large high mass accuracy proteomics experiment. For this dataset I previously showed that the AMT/mass-filtering method was the most sensitive Mascot scoring method available (see chapter 2). Therefore, the data were searched at 500 ppm peptide mass tolerance, filtered to 5 ppm (figure 3.1) and AMT thresholding was applied, resulting in 13,668 estimated true positive peptide identifications at a q-value of 1.0%. In comparison, the MIT and MHT at the same q-value only identified 10,385 and 12,338 true positives at the most restrictive (see method section) peptide mass tolerance setting of 20 ppm (figure 3.3, AMT, MIT, MHT).

A more flexible implementation would be to use both features, the score cut-off and the mass deviation, in combination for improved discrimination of correct and incorrect PSMs, for example accepting PSMs with slightly larger mass deviation given the PSM scores are highly significant.

This can be achieved with a machine learning algorithm such as Percolator using features relevant to the AMT/mass-filtering strategy. Accordingly, the following features were calculated from the 500 ppm Mascot target and decoy searches and

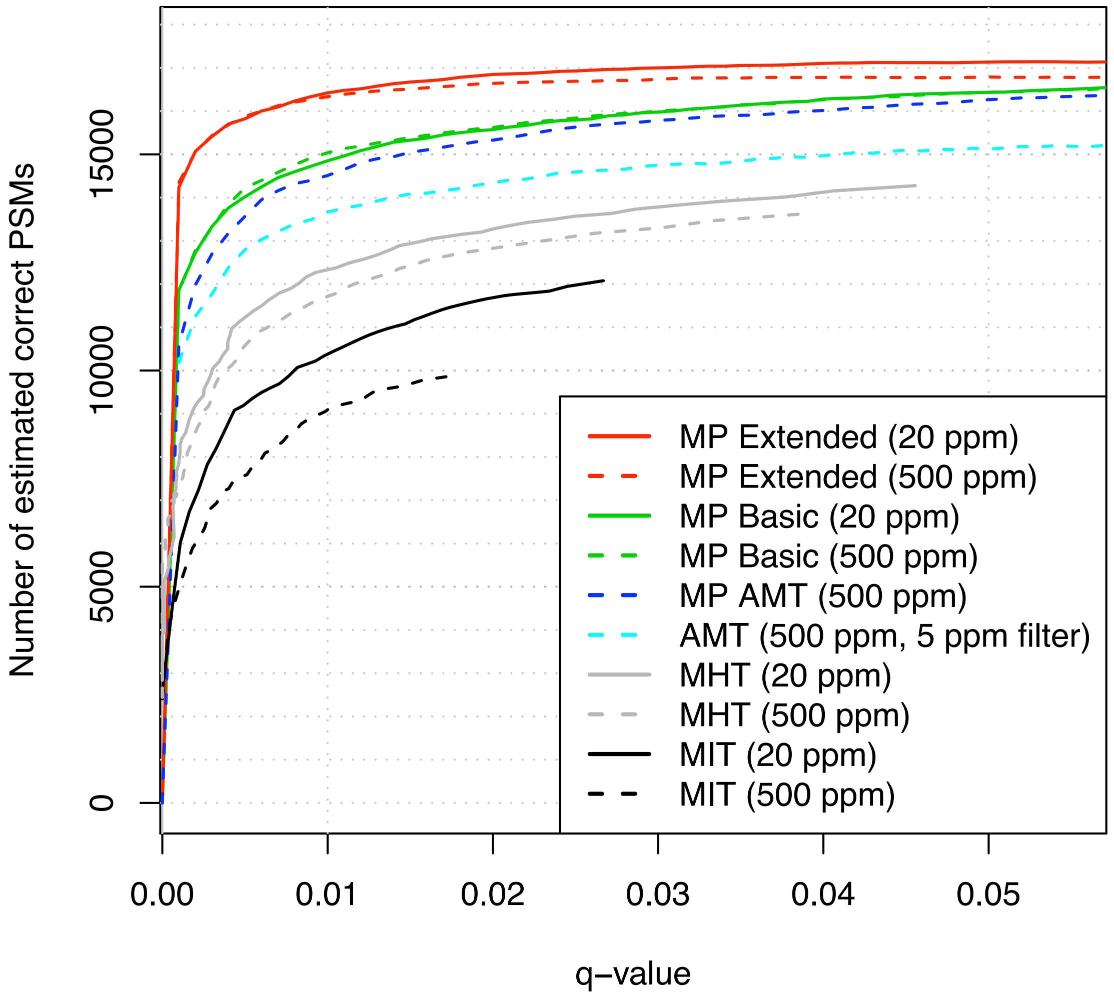


Figure 3.3: For the 20 ppm Mascot search, the basic and extended Mascot Percolator (MP), the Mascot Identity Threshold (MIT) and the Mascot Homology Threshold (MHT) performance was determined as a function of q-value cut-offs ranging from 0 to 0.06. Moreover, the performance of the mass-filtering (5 ppm) strategy together with the Adjusted Mascot Threshold (AMT), the emulated Percolator AMT method (MP AMT), the MIT and MHT are shown for the 500 ppm Mascot search. Note: if no MHT was reported, the MIT was used (default Mascot behaviour).

were used for Percolator training: MHT minus Mascot score, deviation of theoretical and observed peptide mass, and the absolute value of the mass deviation.

Mascot Percolator identified a total of 14,512 estimated true positive PSMs at a 1.0% q-value (figure 3.3, MP AMT), clearly outperforming the AMT/mass-filtering approach by 6.2%. When Mascot Percolator was compared to the Mascot thresholds, it identified 40% (37%) and 18% (17%) more true positive (unique) peptides than the MIT and MHT, respectively, significantly improving performance upon both Mascot thresholds.

These results demonstrate that the combined use of the score threshold and the mass deviation features as a discriminator provide better performance than the AMT/mass-filtering strategy. It should be noted that the used features tackle systematic mass errors and random mass errors separately, therefore simplifying the usability since post-processing to remove systematic mass shifts is not required. These promising results motivated the assessment of more comprehensive feature sets.

3.3.2 Mascot Percolator using extended feature sets

In addition to the mass deviation features described previously, features that can be directly extracted from the Mascot search results were added, defining the “basic feature set”(table 3.1, feature 1-9).

The idea behind the additional chosen features: the native Mascot score is known to correlate well with the quality of a PSM (feature 1); the difference of the Mascot score between two non-isobaric peptide hits indicates the level of ambiguity between two competing matches (feature 2); the number of missed tryptic cleavages or variable modifications of a peptide may be indicative of whether the PSM matches the properties of the rest of the dataset (feature 8 and 9 respectively). Feature 3-4 are not expected to provide discrimination power by themselves, but they may correlate with other features and thereby improve discrimination.

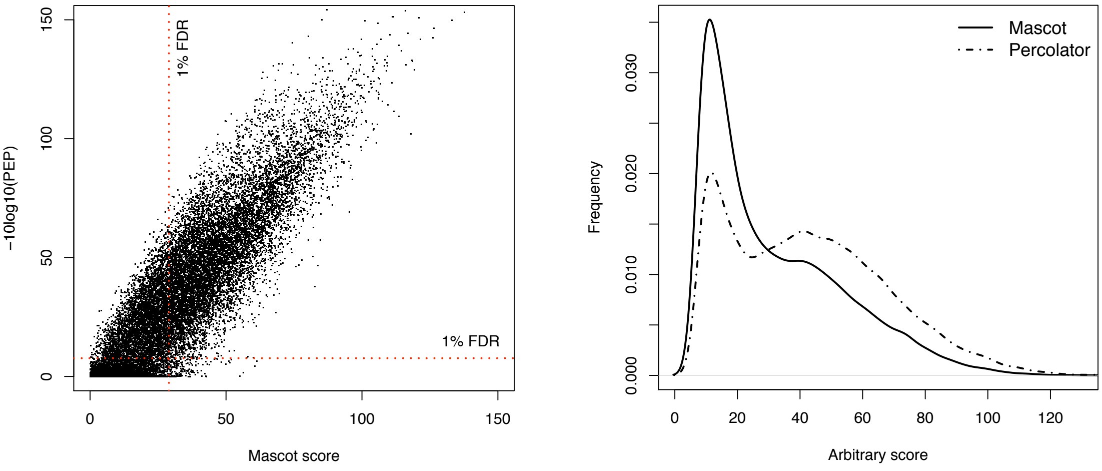
Moreover, an “extended feature set”(table 3.1, feature 1-17) that required rematching the experimental spectra against the theoretical spectra was considered. The idea was to include fragment ion matching statistics, not readily available from the Mascot results: a higher total (matched) ion intensity can be indicative of better spectrum and peptide-spectrum match quality (feature 10-12 respectively); fragment mass error statistics is widely used to manually validate PSMs (feature

13-14); the longest consecutive ion series as well as the fraction of ions matched is another commonly used feature for manual validation (feature 15, 16) and lastly, the fraction of matched ion intensity relative to the total ion intensity may further aid discrimination. Importantly, features 15-17 are computed for each ion series separately, e.g. *b* and *y* series, doubly charged *b* and *y* series as well as the *b* and *y* series including derivatives such as neutral losses of ammonia or water, enabling Mascot Percolator to learn ion series preferences from the dataset at hand.

Using the target/decoy Mascot search results for subsequent Percolator training with the basic and extended feature set, the peptide identification performance improved by 2.5% and 13%, respectively, as compared to the Mascot Percolator performance using only the AMT/mass-filtering features (figure 3.3). Since about the same number of identifications were made for the 500 ppm and 20 ppm search, the basic and extended feature sets appear to effectively substitute the necessity for strong mass accuracy discriminators.

Table 3.1: Features 1-9 represent the "basic feature set" and features 1-17 represent the "extended feature set" as used in Mascot Percolator.

Feature No.	Short description
1	Mascot score
2	Mascot score of current peptide hit relative to 2nd best hit rank
3	Calculated monoisotopic mass of the identified peptide
4	Charge (1 to n)
5	Calculated minus observed peptide mass (in Dalton and ppm)
6	Feature No. 5, corrected for isotope error
7	Absolute value of feature No. 5
8	Number of missed tryptic cleavages
9	Number of variable modifications
10	Total ion intensity
11	Total ion intensity of matched ions
12	Relative ion intensity (Feature No. 11 / Feature No. 12)
13	Median of delta mass of fragment ions (in Dalton and ppm)
14	Interquartile range of delta mass of fragment ions (in Dalton and ppm)
15	Longest consecutive ion series (per ion series)
16	Fraction of ions matched (per ion series)
17	Relative ion intensity matched (per ion series)



(a) Mascot score plotted against the Mascot Percolator Posterior Error Probability score (log transformed). The 1% score cut-offs are indicated for each dimension.

(b) Mascot score overlaid with the Mascot Percolator Posterior Error Probability score (log transformed and scaled, for better visualisation only).

Figure 3.4: Both figures enable the visual comparison of the raw Mascot scores against the Mascot Percolator (posterior error probability) scores. Figure (a) highlights all the extra PSMs with weak Mascot score that Mascot Percolator accepted based on its discrimination power using all 17 features. Figure (b) compares the score distributions and the improved bi-modal distribution of the Percolator scores indicates the improved discrimination power.

Therefore, Mascot Percolator with features that include Mascot scoring and peptide features as well as ion matching statistics, identified more than 58% (52%) and 33% (29%) more true positive (unique) peptides than the MIT and MHT respectively at a 1.0% q-value with a standard 20 ppm search (figure 3.3), clearly demonstrating the enhanced discrimination power when using an ensembl of features (figure 3.4). These improvements translate into 15% and 6% more protein identifications over the MIT and MHT, respectively. Overall, these results are a significant improvement over all current Mascot scoring methods, including AMT, and eliminate the need to search high accuracy data at relaxed mass tolerances to improve sensitivity.

3.3.3 Mascot Percolator applied to low mass accuracy data

The following evaluation is concerned with dataset 2, a yeast sample acquired on a LTQ instrument that was used for the evaluation of Sequest Percolator. To enable

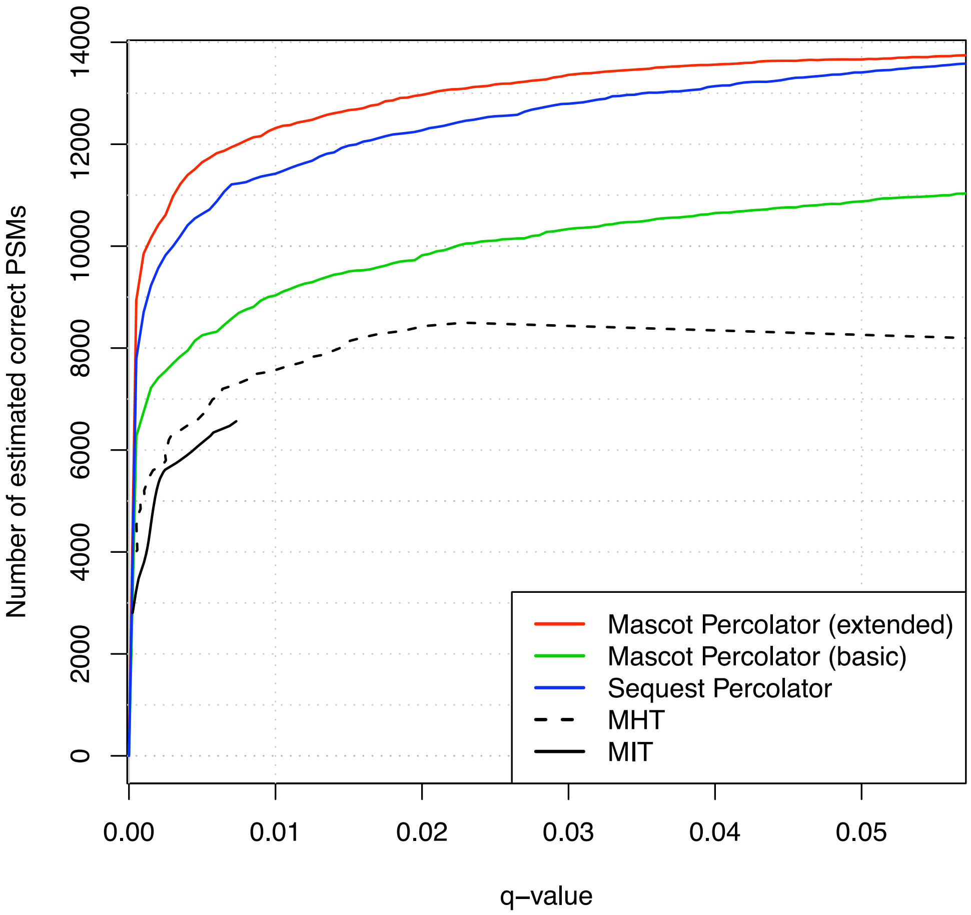


Figure 3.5: The number of estimated correct PSMs were determined for each q-value cut-off for the basic and extended Mascot Percolator (MP) runs, the Adjusted Mascot Threshold (AMT), the Mascot Identity Threshold (MIT) and Mascot Homology Threshold (MHT) as well as for the Sequest Percolator.

comparison of Mascot Percolator and Sequest Percolator, the subsequent experiments were therefore not only based on the same data, but also on the same target/decoy databases and search parameters as described by Käll *et al.* (2007), with the only exception being the trypsin specificity parameter.

Using the MIT and MHT, 6,379 and 7,541 true positive identifications (figure 3.5, MIT, MHT) were made at a q-value of 0.7% and 1.0%, respectively (the Mascot significance threshold is limited to 0.1, corresponding to a q-value of 0.7%). Using the basic feature set with Mascot Percolator improved sensitivity over MIT and MHT by more than 41% and 19%, respectively, at a 1.0% q-value (figure 3.5, MP basic). Sensitivity was further boosted by more than 50% when the extended feature set was applied (figure 3.5, MP extended). Compared to the MIT and MHT, this relates to a (unique) peptide identification gain of 93% (82%) and 63% (55%), respectively, at the standard 1.0% q-value. Overall, these results further support the performance advantages of Mascot Percolator over the default MIT and MHT.

Moreover, the difference in performance of Mascot Percolator between the basic and extended feature set was significantly more prominent than it was with dataset 1, highlighting that feature contribution can vary substantially for different datasets and demonstrating the dynamic and adaptive property of the Percolator algorithm (Käll *et al.* 2007, supplement 2). It could be speculated that low accuracy data benefit from more discriminating features, while high accuracy data almost reaches the maximum sensitivity with the basic feature set due to the more restrictive search parameters and known charge states.

Käll *et al.* (2007) identified trypsin-specificity as a strong discriminating feature and consequently they searched without enzyme specificity in their study. However, this practice is significantly more CPU intensive due to the larger search space. Search times in Mascot are one order of magnitude slower when semi-trypsin is specified instead of trypsin, and two orders of magnitude slower when no enzyme specificity is defined instead of trypsin (`http://www.matrixscience.com/pdf/2006WKSHP1.pdf`). Therefore, Mascot Percolator does not make use of any enzyme specificity related features, yet improves upon the Sequest Percolator sensitivity with the extended feature set by about 8% at a 1% q-value (figure 3.5).

3.3.4 Validation with standard protein datasets

The robustness and precision of the q-value was validated in the supplemental material of the original Percolator publication (Käll *et al.*, 2007). The employed target/decoy search strategy for q-value estimation is a widely accepted approach, but various methods exist for generating the decoy databases (see section 1.1.2.3). Therefore, I extended this evaluation by assessing the accuracy of the q-value as

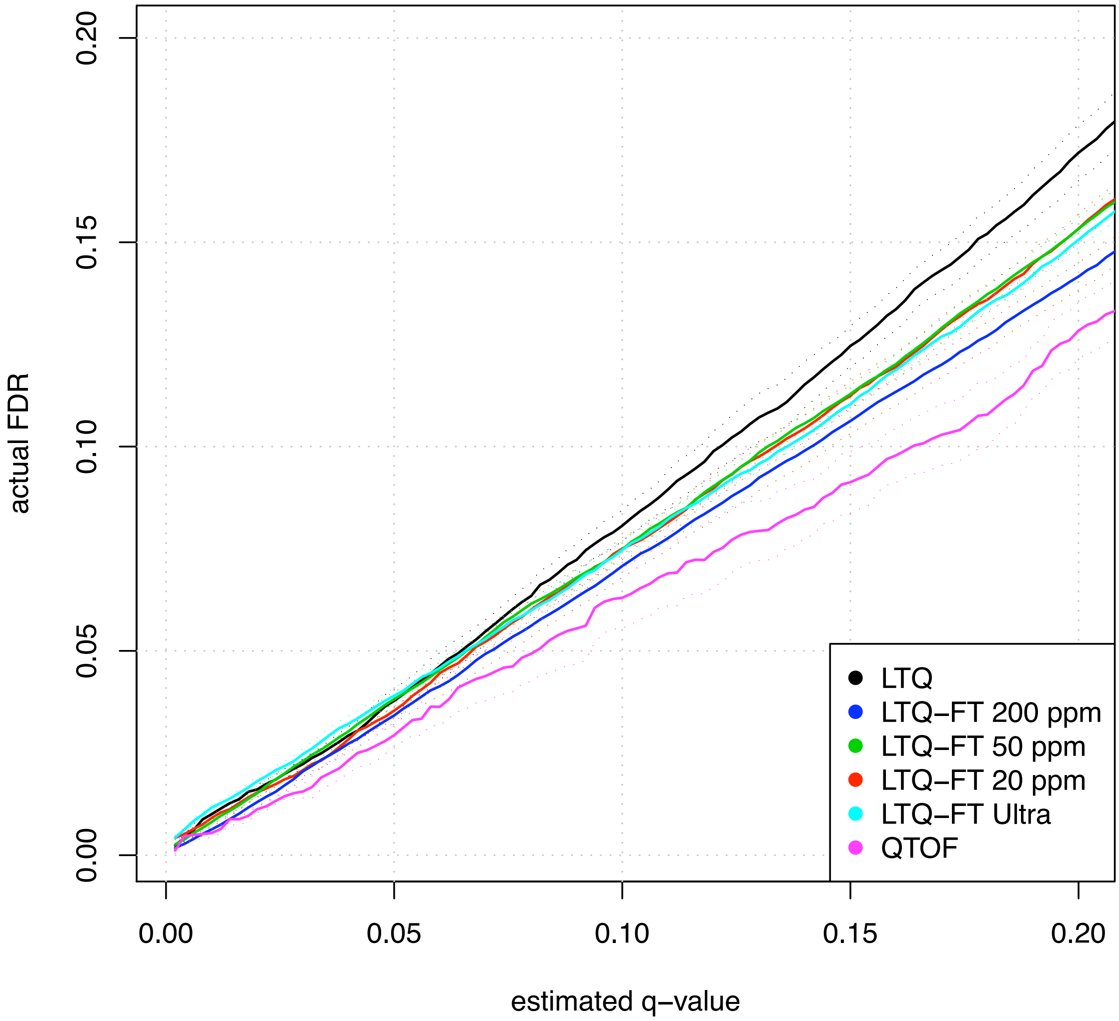


Figure 3.6: The estimated q-values were plotted against the false discovery rates as reported by the protein standard datasets for the extended and the basic Mascot Percolator runs. The dotted lines represent the standard error.

a result of the Matrix Science decoy.pl script (see methods) with a set of protein standard datasets (dataset 3). Five Mascot searches were performed and analysed with Mascot Percolator for each data, using the same target but independently generated random databases. This enabled computation of the standard error for the q-value calculations. For every estimated q-value, the corresponding observed FDR was determined by counting the incorrect PSMs that did not match the expected protein sequences.

It was found that q-value estimates were in very good agreement with the results

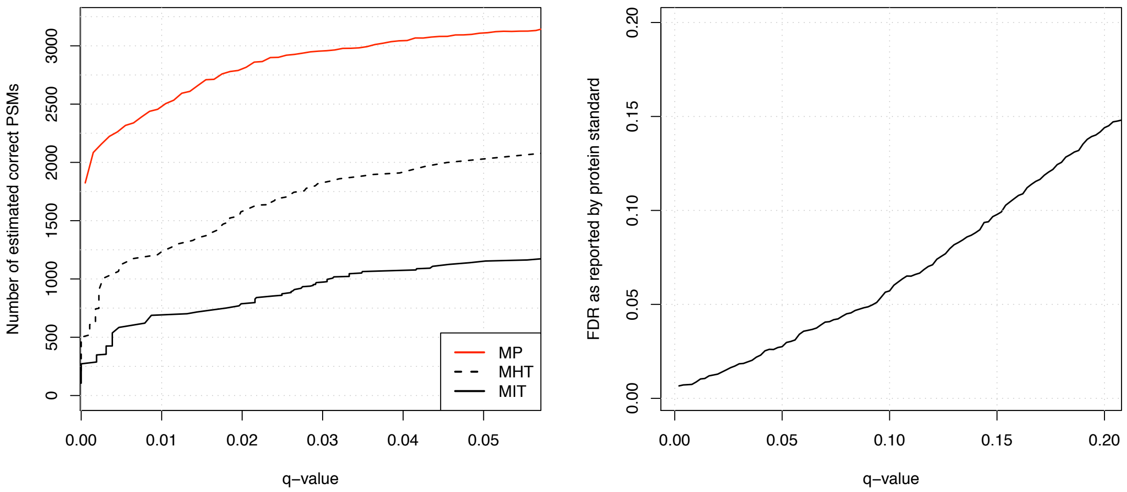


Figure 3.7: Performance of the Mascot Percolator, MIT and MHT were compared for a no enzyme search of the LTQ-FT protein standard dataset (left). Estimated q-values were also plotted against the FDR as reported by this protein standard dataset (right).

obtained by the expected protein sequences (figure 3.6), while the standard errors were negligible, particularly in the low FDR region. This implies that the gain in sensitivity with Mascot Percolator is limited to valid sequences within the expected error rates. These results demonstrate that none of the chosen features introduced any bias towards severe under- or overestimation of the q-values and that these can be seen as accurate and reliable estimates of the real error rates for a variety of analytical platforms. This is a significant improvement over the standard Mascot results using the MIT or MHT, for which I have previously shown that the actual FDR can differ by several fold from the expected FDR (Brosch *et al.*, 2008).

Moreover, the LTQ-FT dataset was also used for a more demanding no-enzyme search. The Mascot scoring scheme as well as the Mascot Percolator were evaluated with the protein standard dataset that was searched in Mascot without any enzyme constraints. Mascot Percolator identified 265% and 96% more peptides than using the MIT or MHT respectively, at a q-value of 1%. It should be noted that none of the features I use discriminate by enzyme specificity as pointed out earlier. Estimated q-values were validated against this protein standard dataset and showed good accuracy, indicating that the identifications are limited to valid sequences within the expected error rates.

Overall these results demonstrate that Mascot Percolator can also be applied to more challenging conditions than standard tryptic searches, where the search space is increased by several orders of magnitude (`http://www.matrixscience.com/help/search_field_help.html`), such as searches without any enzyme constrains or excessive variable modification settings.

3.3.5 Mascot Percolator applied to a pool of 73 datasets

About 10 million tandem MS spectra from various sources were post-processed with Mascot Percolator in chapter 5 (see section 5.2.1 for details). This data enabled the evaluation of Mascot Percolator on a large scale. For each of the 73 datasets the increase in peptide identifications with Mascot Percolator over MHT, MIT and Mascot was determined, allowing to compute the median and interquartile range for each comparison: the median improvement at a 1% q-value were 54%, 109% and 99%, with an interquartile range (IQR) of 39%, 84% and 69% when the number of PSMs of Mascot Percolator were compared with MHT, MIT and the Mascot score respectively.

In a next step, the same data were searched against a database that was supplemented with gene predictions, resulting in about a 10-fold search space increase (see section 5.2.2 for details). Mascot Percolator identified 16% fewer peptides (median value) at a 1% q-value when compared with the searches against the smaller database, while the difference in performance between Mascot Percolator and the MIT, MHT and Mascot score also changed: the median improvement of peptide identifications with Mascot Percolator at a 1% q-value over the MHT, MIT and Mascot score were 65%, 197% and 155% respectively. While the improvement over the MHT was almost constant, the change over the MIT and raw score was considerable, resulting in about half the number of peptide identifications when compared to the search against the

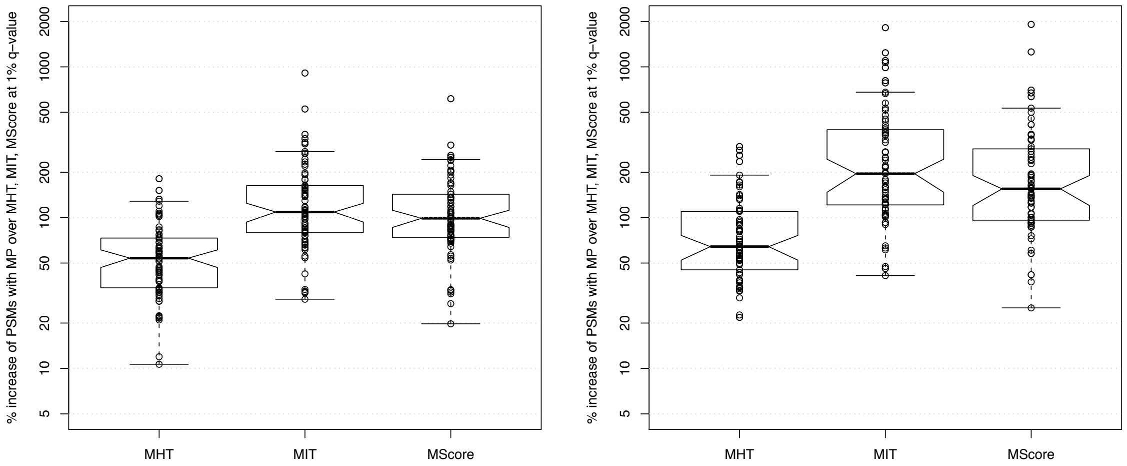


Figure 3.8: 73 datasets were processed with Mascot Percolator and the number of peptide identifications were compared to the identification rates of Mascot Identity and Homology thresholds as well as the raw Mascot score (left).
The same data was also searched against a 10-fold increased search space, followed by the same evaluation (right).

smaller database, further supporting the findings of chapter 2.

Overall this demonstrates on a large scale that Mascot Percolator shows a robust improvement over the native Mascot scoring, including a significant less severe drop in performance when search space inflates.

3.4 Mascot Percolator availability

Standalone package

Mascot Percolator was designed as a command line program to run either as a stand-alone application or as a component that can be embedded into existing data processing pipelines, allowing for streamlining data and automation.

An example of executing the program follows for illustration: "java -cp MascotPercolator.jar cli.MascotPercolator -target 11026 -decoy 11027 -out 11026-11027 -newDat". This command line triggers Mascot Percolator to parse the Mascot search results from the files that are associated with the provided Mascot job identifiers (11026, 11027) to subsequently calculate the features discussed above for the subsequent Percolator run. Results and logging files are written into files prefixed with "11026-11027". Moreover, the flag "newDat" directs Mascot Percolator to write a new Mascot results file (*.dat) that can be opened by the Mascot server just as if it was a standard Mascot result file. The differences however are as follows: the Mascot scores are replaced with the $-10log_{10}$ of the Posterior Error Probability (PEP); the expect value that is calculated on the Mascot results page directly corresponds to the PEP; the accepted FDR can be changed by setting the Probability values on the Mascot results page accordingly. A summary for illustration is shown in figure 3.9.

Mascot Percolator is available at `http://www.sanger.ac.uk/Software/analysis/MascotPercolator/`, where I also documented the more advanced command line options.

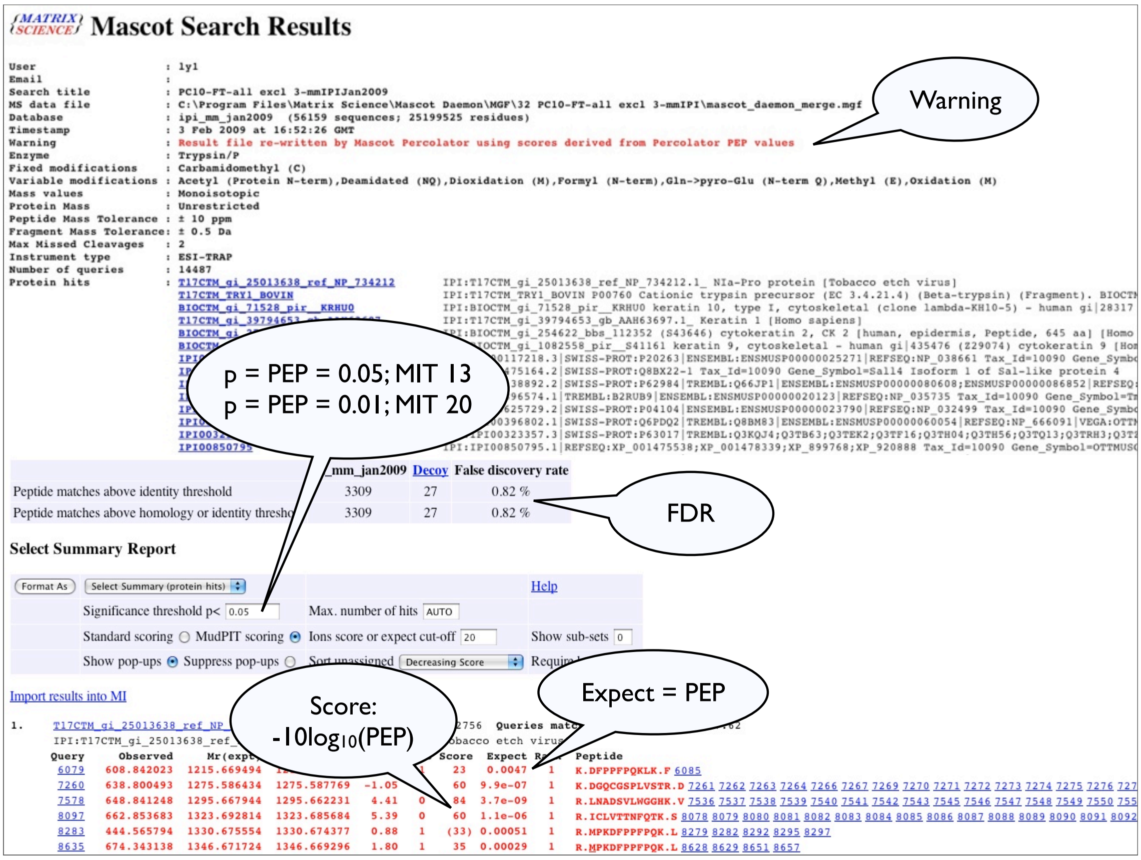
Distributed package

I was confronted with the problem to post-process the search results of 146 Mascot searches comprising a total of about 20 million spectra (2×10 million, see chapter 5). Moreover, I had to process these data as quickly as possibly and ultimately wanted to make use of our Mascot compute farm. This farm does not have a "Load Sharing Facility[1]" and hence there was a need to develop a distributed version of Mascot Percolator, which is now used by default in our lab.

The system is based on several components: a database server was implemented that runs independently of the system and logs every action of the distributed Mascot Percolator system. A Mascot Percolator Server was developed that after starting up connects to the database and triggers the status page for the intranet to be updated and listens for Mascot Percolator Nodes (figure 3.10). These nodes can be run on either Unix or Windows computers and automatically connect to the server. A script as well as a web-interface[2] were implemented that enable the submission of jobs in

[1] A commercial computer software job scheduler that can be used to execute batch jobs on networked Unix and Windows systems.

[2] The web-based submission interface was developed by Parthiban Vijayarangakannan, Team



Mascot Search Results

```
User                    : ly1
Email                   :
Search title            : PC10-FT-all excl 3-mmIPIJan2009
MS data file            : C:\Program Files\Matrix Science\Mascot Daemon\MGF\32 PC10-FT-all excl 3-mmIPI\mascot_daemon_merge.mgf
Database                : ipi_mm_jan2009  (56159 sequences; 25199525 residues)
Timestamp               : 3 Feb 2009 at 16:52:26 GMT
Warning                 : Result file re-written by Mascot Percolator using scores derived from Percolator PEP values
Enzyme                  : Trypsin/P
Fixed modifications     : Carbamidomethyl (C)
Variable modifications  : Acetyl (Protein N-term),Deamidated (NQ),Dioxidation (M),Formyl (N-term),Gln->pyro-Glu (N-term Q),Methyl (E),Oxidation (M)
Mass values             : Monoisotopic
Protein Mass            : Unrestricted
Peptide Mass Tolerance  : ± 10 ppm
Fragment Mass Tolerance: ± 0.5 Da
Max Missed Cleavages    : 2
Instrument type         : ESI-TRAP
Number of queries       : 14487
Protein hits            : T17CTM_gi_25013638_ref_NP_734212       IPI:T17CTM_gi_25013638_ref_NP_734212.1_ NIa-Pro protein [Tobacco etch virus]
                          T17CTM_TRY1_BOVIN                      IPI:T17CTM_TRY1_BOVIN P00760 Cationic trypsin precursor (EC 3.4.21.4) (Beta-trypsin) (Fragment). BIOCTM
                          BIOCTM_gi_71528_pir__KRHU0             IPI:BIOCTM_gi_71528_pir__KRHU0 keratin 10, type I, cytoskeletal (clone lambda-KH10-5) - human gi|28317
                          T17CTM_gi_39794653                     IPI:T17CTM_gi_39794653_gb_AAH63697.1_ Keratin 1 [Homo sapiens]
                          BIOCTM_gi                              IPI:BIOCTM_gi_254622_bbs_112352 (S43646) cytokeratin 2, CK 2 [human, epidermis, Peptide, 645 aa] [Homo
                          BIOCTM                                  IOCTM_gi_1082558_pir__S41161 keratin 9, cytoskeletal - human gi|435476 (Z29074) cytokeratin 9 [Hom
                          IPI                                    00117218.3|SWISS-PROT:P20263|ENSEMBL:ENSMUSP00000025271|REFSEQ:NP_038661 Tax_Id=10090 Gene_Symbo
                          IP                                     475164.2|SWISS-PROT:Q8BX22-1 Tax_Id=10090 Gene_Symbol=Sall4 Isoform 1 of Sal-like protein 4
                          I                                      38892.2|SWISS-PROT:P62984|TREMBL:Q66JP1|ENSEMBL:ENSMUSP00000080608;ENSMUSP00000086852|REFSEQ:
                          I                                      96574.1|TREMBL:B2RUB9|ENSEMBL:ENSMUSP00000020123|REFSEQ:NP_035735 Tax_Id=10090 Gene_Symbol=Tn
                          IP                                     625729.2|SWISS-PROT:P04104|ENSEMBL:ENSMUSP00000023790|REFSEQ:NP_032499 Tax_Id=10090 Gene_Symbo
                          IPI0                                   00396802.1|SWISS-PROT:Q6PDQ2|TREMBL:Q8BM83|ENSEMBL:ENSMUSP00000060054|REFSEQ:NP_666091|VEGA:OTTM
                          IPI003                                 IPI00323357.3|SWISS-PROT:P63017|TREMBL:Q3KQJ4;Q3TB63;Q3TEK2;Q3TF16;Q3TH04;Q3TH56;Q3TQ13;Q3TRH3;Q3T
                          IPI00850795                            IPI:IPI00850795.1|REFSEQ:XP_001475538;XP_001478339;XP_899768;XP_920888 Tax_Id=10090 Gene_Symbol=OTTMUSG
```

	_mm_jan2009	Decoy	False discovery rate
Peptide matches above identity threshold	3309	27	0.82 %
Peptide matches above homology or identity thresho	3309	27	0.82 %

Select Summary Report

Format As	Select Summary (protein hits)		Help
	Significance threshold p< 0.05	Max. number of hits AUTO	
	Standard scoring ○ MudPIT scoring ●	Ions score or expect cut-off 20	Show sub-sets 0
	Show pop-ups ● Suppress pop-ups ○	Sort unassigned Decreasing Score	Require

Import results into MI

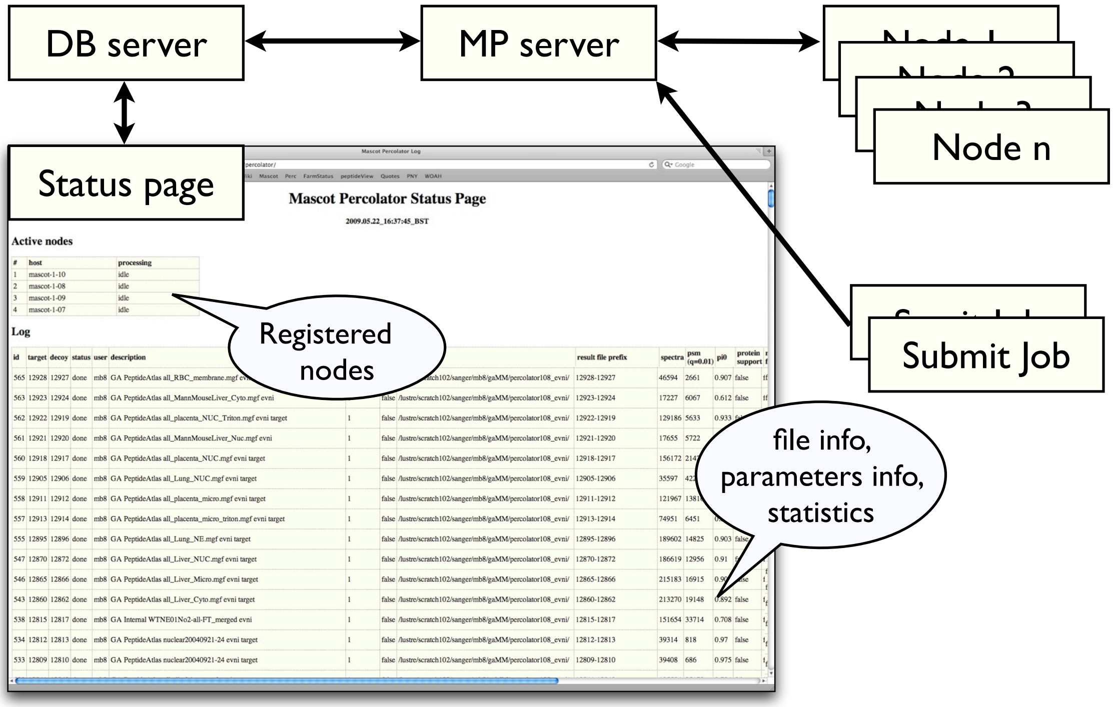
```
1.     T17CTM_gi_25013638_ref_NP_                        2756   Queries matc                 .62
       IPI:T17CTM_gi_25013638_ref_                       obacco etch virus
  Query     Observed     Mr(expt)                             Score  Expect  Ra     Peptide
  6079    608.842023   1215.669494  12                    1    23   0.0047    1   K.DFPPFPQKLK.F 6085
  7260    638.800493   1275.586434  1275.587769  -1.05         60   9.9e-07   1   K.DGQCGSPLVSTR.D 7261 7262 7263 7264 7266 7267 7269 7270 7271 7272 7273 7274 7275 7276 727
  7578    648.841248   1295.667944  1295.662231   4.41    0    84   3.7e-09   1   R.LNADSVLWGGHK.V 7536 7537 7538 7539 7540 7541 7542 7543 7545 7546 7547 7548 7549 7550 755
  8097    662.853683   1323.692814  1323.685684   5.39    0    60   1.1e-06   1   R.ICLVTTNFQTK.S 8078 8079 8080 8081 8082 8083 8084 8085 8086 8087 8088 8089 8090 8091 8092
  8283    444.565794   1330.675554  1330.674377   0.88    1   (33)  0.00051   1   R.MPKDFPPFPQK.L 8279 8282 8292 8295 8297
  8635    674.343138   1346.671724  1346.669296   1.80    1    35   0.00029   1   R.MPKDFPPFPQK.L 8628 8629 8651 8657
```

Figure 3.9: Screenshot of a Mascot results page that was generated by Mascot Percolator. The results are basically identical to a standard Mascot results page, can be opened by the Mascot server, but the scoring values are derived from the Mascot Percolator run. A warning at the top of the page states this very clearly to avoid confusion.

batches. The server then schedules the jobs and distributes these onto the different available nodes. When jobs complete, the status page updates and the result files can be browsed. The system has some more advanced options that are documented at `http://www.sanger.ac.uk/Software/analysis/MascotPercolator/`.

In conclusion this distributed system enabled me to process the large datasets reliably and efficiently in a fraction of the time when compared to the stand alone version of Mascot Percolator.

17, Wellcome Trust Sanger Institute



Mascot Percolator Status Page

2009.05.22_16:37:45_BST

Active nodes

#	host	processing
1	mascot-1-10	idle
2	mascot-1-08	idle
3	mascot-1-09	idle
4	mascot-1-07	idle

Log

id	target	decoy	status	user	description				result file prefix	spectra	psm (q=0.01)	pi0	protein support
565	12928	12927	done	mb8	GA PeptideAtlas all_RBC_membrane.mgf ev			tch102/sanger/mb8/gaMM/percolator108_evni/	12928-12927	46594	2661	0.907	false
563	12923	12924	done	mb8	GA PeptideAtlas all_MannMouseLiver_Cyto.mgf evni		false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12923-12924	17227	6067	0.612	false
562	12922	12919	done	mb8	GA PeptideAtlas all_placenta_NUC_Triton.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12922-12919	129186	5633	0.933	
561	12921	12920	done	mb8	GA PeptideAtlas all_MannMouseLiver_Nuc.mgf evni	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12921-12920	17655	5722		
560	12918	12917	done	mb8	GA PeptideAtlas all_placenta_NUC.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12918-12917	156172	214		
559	12905	12906	done	mb8	GA PeptideAtlas all_Lung_NUC.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12905-12906	35597	422		
558	12911	12912	done	mb8	GA PeptideAtlas all_placenta_micro.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12911-12912	121967	1381		
557	12913	12914	done	mb8	GA PeptideAtlas all_placenta_micro_triton.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12913-12914	74951	6451		
555	12895	12896	done	mb8	GA PeptideAtlas all_Lung_NE.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12895-12896	189602	14825	0.903	false
547	12870	12872	done	mb8	GA PeptideAtlas all_Liver_NUC.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12870-12872	186619	12956	0.91	
546	12865	12866	done	mb8	GA PeptideAtlas all_Liver_Micro.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12865-12866	215183	16915	0.90	
543	12860	12862	done	mb8	GA PeptideAtlas all_Liver_Cyto.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12860-12862	213270	19148	0.892	false
538	12815	12817	done	mb8	GA Internal WTNE01No2-all-FT_merged evni	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12815-12817	151654	33714	0.708	false
534	12812	12813	done	mb8	GA PeptideAtlas nuclear20040921-24 evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12812-12813	39314	818	0.97	false
533	12809	12810	done	mb8	GA PeptideAtlas nuclear20040921-24 evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12809-12810	39408	686	0.975	false

Figure 3.10: Schematic of the Distributed Mascot Percolator package and screenshot of the status webpage.

Official Mascot Percolator support by Matrix Science

In collaboration with John Cottrell, David Creasy (Matrix Science, London) and Lukas Käll (University of Stockholm), Mascot Percolator is currently implemented into the official Mascot release 2.3 (see `http://www.matrixscience.com/workshop_2009.html`). Features will be pre-computed for every Mascot search, cutting compute time and allowing streamlined processing through Percolator, without the need of a user to access the command line. This will ultimately expand the user group of Mascot Percolator significantly.

3.5 Conclusion

The Percolator machine learning algorithm was recently introduced to rescore Sequest results and demonstrated significantly improved sensitivity for peptide and protein identification. Percolator learns a classifier independently for each dataset, thereby adapting to inherent variations between different datasets, such as changing analytical protocols or instrumentation.

In this work, I have developed and evaluated Mascot Percolator, a software package that interfaces Mascot with Percolator. It automatically extracts and computes relevant features from target/decoy Mascot search results, trains Percolator, applies the resulting classifier to each PSM and writes a result file. Mascot Percolator has been developed as a command line tool and can be readily integrated into existing pipelines or be used as a stand-alone application. A large number of features that are relevant to the quality of a PSM, such as Mascot scores, parent and fragment mass accuracy, peptide as well as ion matching statistics, amongst others, were incorporated.

I have demonstrated that Mascot Percolator substantially outperforms previous Mascot scoring methods for high and low mass accuracy data and applied it to a large ensembl of 73 datasets, where up to 65% and 197% more peptides than the MIT and MHT were identified with Mascot Percolator at a 1% q-value (median values). This demonstrates the improved discrimination potential achieved when several factors that define the quality of a PSM are used collectively for scoring instead of only one metric. Furthermore, I have shown that the estimated q-values are in very good agreement with the actual FDRs and represent a significant improvement in accuracy as compared to the Mascot thresholds. Lastly, Percolator calculates both significance measures, the q-value and the posterior error probability. The latter is particularly important for my genome annotation efforts, where the significance of every peptide identification should be known.

Chapter 4

Development of a proteogenomics pipeline

4.1 Introduction

In chapter 3 the development and evaluation of Mascot Percolator was discussed, a powerful peptide scoring scheme with an implementation that can be automated and run in batch-mode, providing high-throughput capability. This system delivers sensitive peptide identification with accurate significance measures and thereby sets the foundation for a reliable proteogenomic pipeline. Mascot Percolator results can be written into a tab delimited text file or exported into the proprietary Mascot results file format. To use these data for genome annotation purposes, a proteogenomic pipeline is required that stores these data, maps the peptide identifications to the genome and enables comprehensive data analysis.

The currently available proteogenomics pipelines, Peptide Atlas (Desiere *et al.*, 2005) and GAPP (Shadforth *et al.*, 2006), which were described in detail in section 1.3, were found to be not suitable since these systems are highly specialised and the code bases are not in the public domain. The envisaged system should integrate peptide identifications available from Mascot Percolator, enable validation and

Chapter 4

Development of a proteogenomics pipeline

4.1 Introduction

In chapter 3 the development and evaluation of Mascot Percolator was discussed, a powerful peptide scoring scheme with an implementation that can be automated and run in batch-mode, providing high-throughput capability. This system delivers sensitive peptide identification with accurate significance measures and thereby sets the foundation for a reliable proteogenomic pipeline. Mascot Percolator results can be written into a tab delimited text file or exported into the proprietary Mascot results file format. To use these data for genome annotation purposes, a proteogenomic pipeline is required that stores these data, maps the peptide identifications to the genome and enables comprehensive data analysis.

The currently available proteogenomics pipelines, Peptide Atlas (Desiere *et al.*, 2005) and GAPP (Shadforth *et al.*, 2006), which were described in detail in section 1.3, were found to be not suitable since these systems are highly specialised and the code bases are not in the public domain. The envisaged system should integrate peptide identifications available from Mascot Percolator, enable validation and refinement of Ensembl (Hubbard *et al.*, 2002) and Vega (Ashurst *et al.*, 2005) annotation, and provide a modular implementation together with a small code base for easy maintenance. Moreover, the pipeline should automatically and readily map the available peptide identifications onto the genome and provide comprehensive means to analyse and visualise these data. This chapter discusses the pipeline development, design and its individual components. Chapter 5 applies this pipeline in a proteogenomics pilot study.

Parts of this chapter will be published together with the next chapter by the author of this thesis (Markus Brosch) and my thesis supervisors (Tim Hubbard, Jyoti Choudhary).

4.2 Pipeline design and development

Figure 4.1 illustrates the pipeline design with its components. At the core of the system is a relational database "GenoMS-DB", which integrates all *in silico* digested peptides, each of which is associated with its genomic context. This approach offers several advantages:

- Non-redundant peptide-level FASTA files can be constructed, enabling more efficient Mascot searches.
- Gene level FASTA files can be constructed for gene centric viewing of Mascot results. Optionally, a peptide list that comprises peptides unique to a gene or gene isoform supports targeted proteomics experiment.
- Peptides identified with Mascot and processed with Mascot Percolator can be flagged and linked in the database with experimental and scoring information.
- The database provides readily available peptide-genome mapping, allowing immediate and direct mapping of peptides onto the genome.

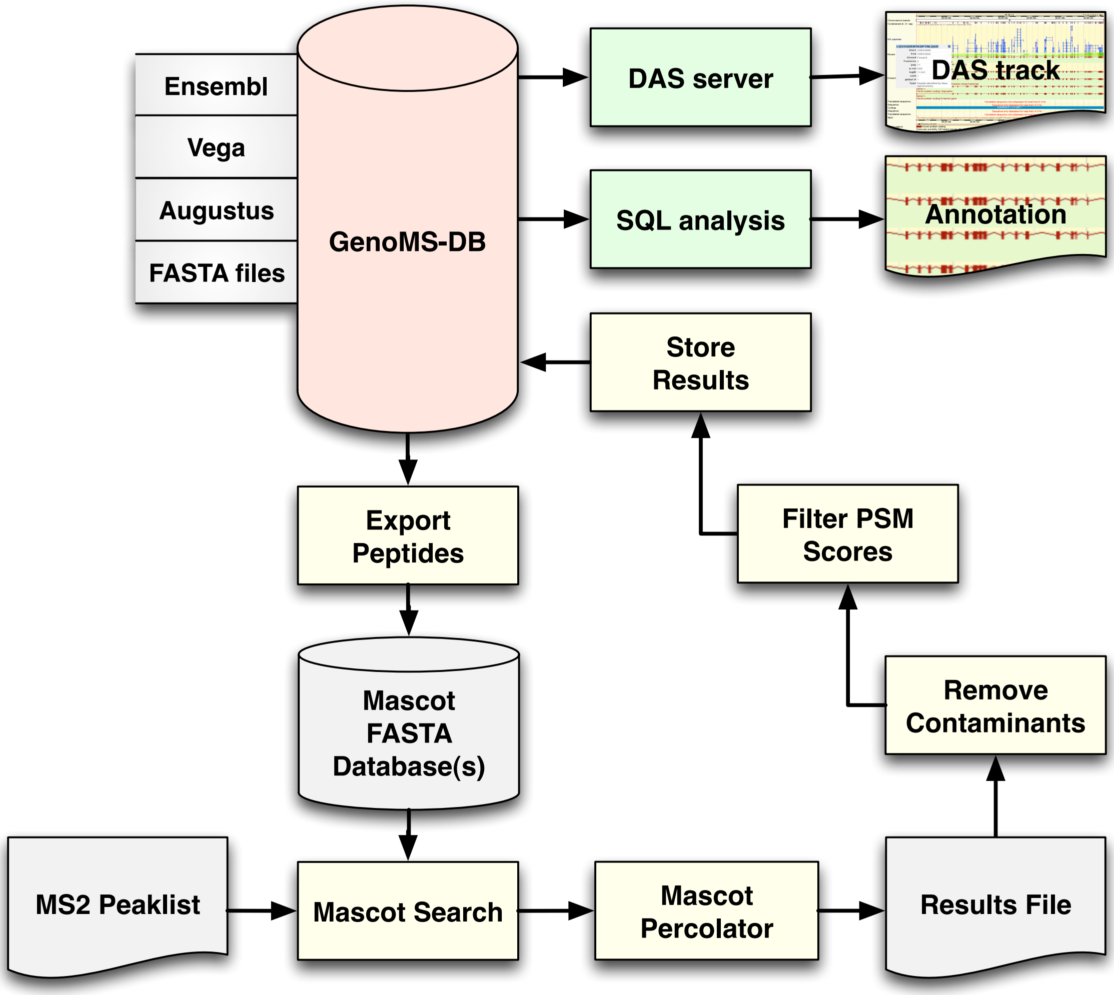


Figure 4.1: Schematic overview of the proteogenomics pipeline. The database at the core of the system, "GenoMS-DB", is built by integrating all peptides that are derived from an *in silico* digestion of available data sources. These can comprise Ensembl, Vega or Augustus gene predictions. Each peptide derived from these data-sources is associated with its genomic locus and context, such as gene, transcript, exon or splice site information. Peptides from FASTA protein databases can optionally be integrated, but lack genome mapping.
GenoMS-DB is then used to export a set of all non-redundant *in silico* digested peptides, which are used by Mascot as a search database. Tandem MS spectra are searched with Mascot and post processed with Mascot Percolator. This is followed by removing common contaminant sequences and low scoring peptide-spectrum matches (PSMs) from the results, prior to storing the remaining identifications into GenoMS-DB database. This integration of peptide-genome mapping together with peptide identifications enables streamlined analysis with standard SQL or visualisation via a DAS feature server.

- Comprehensive data analysis can be performed with fast and efficient Structured Query Language (SQL), a standard language for accessing and querying databases.
- The availability of the complete peptide-genome mapping enables theoretical studies such as genome coverage with a specific set of peptide.

The core of the system is written in Perl, comprising more than 2,000 lines of code, extensively relying on the Ensembl Perl API (Stabenau *et al.*, 2004), without which the codebase would have been significantly larger. This API provided all the core functionality required to establish the proteome-genome relationship, such as the coordinate conversions between translated gene products and the underlying genomic sequences. Therefore, none of the involved steps required any sequence alignment tools. The next few sections of this chapter briefly describe individual components.

4.2.1 Genome annotation data sources and integration

Section 1.2.5 discussed the Ensembl and Vega projects in detail, which together with Augustus gene predictions build the annotation data basis for this pipeline. Details and parameters are discussed in depth in the pilot study of chapter 5.

Conveniently, the Ensembl core API can be configured with the `Bio::EnsEMBL::Registry` module to handle either Ensembl or Vega data sources with the same platform. Moreover, the API module extension `Bio::EnsEMBL::Analysis::Runnable::Finished::Augustus` enables full API functionality for Augustus gene predictions (Stanke and Waack, 2003), which are otherwise only available as GFF files (`http://www.sanger.ac.uk/Software/formats/GFF/`). Lastly, standard text-based FASTA sequences can be integrated as supplemental data, such as a selected laboratory or contaminant protein sequences, but genome association is not possible.

During the database build process, the system performs the following simplified steps: (a) for each chromosome, get all protein coding genes; (b) for each gene, get all protein coding transcripts; (c) for each transcript, obtain the protein sequence; (d) determine all enzymatic cleavage sites within the protein sequence; (e) calculate the individual peptide start and end coordinates within the translation; (f) perform *in silico* digestion according to the available cleavage sites, allowing for the defined number of missed cleavage sites and sequence length constraints; (g) for each peptide, calculate the genomic coordinates, accounting for multiple loci if the peptide spans one or multiple splice sites; (h) store these loci in GenoMS-DB for each peptide, along with gene, transcript, exon and splice site information; (i) optionally account for coding SNPs (Schandorff *et al.*, 2007) and N-terminal methionine excision (Frottin *et al.*, 2006) by generating alternative peptide variants. In the current form, known post translational modifications or cleavages are not accounted for. It should be noted that the organism, enzyme settings, missed cleavages, peptide minimum and maximum length are user defined parameters, with the following default values: *mus musculus*, trypsin (cleavage after Arginine and Lysine), 2, 6, 50 respectively.

4.2.2 Database design

The database is an integral part of this pipeline and its relational model (schema) is illustrated in figure 4.2. The database is populated with one or multiple data sources, which were discussed in the previous section. Once built, the database is only used for querying with the only exception of some “PeptideSequence” table attributes, which are related to Mascot Percolator results integration (see section below).

GenoMS-DB was designed to allow the user to construct simple and fast SQL queries. This is done by selective denormalisation (Shin and Sanders, 2006) and choosing the peptide-genome mapping information as a central element, which is represented by the `PeptideMapping` table. This peptide-centric orientation, together with the involved denormalisation, led me to the development of a new schema instead of adapting the existing Ensembl database schema (Stabenau *et al.*, 2004).

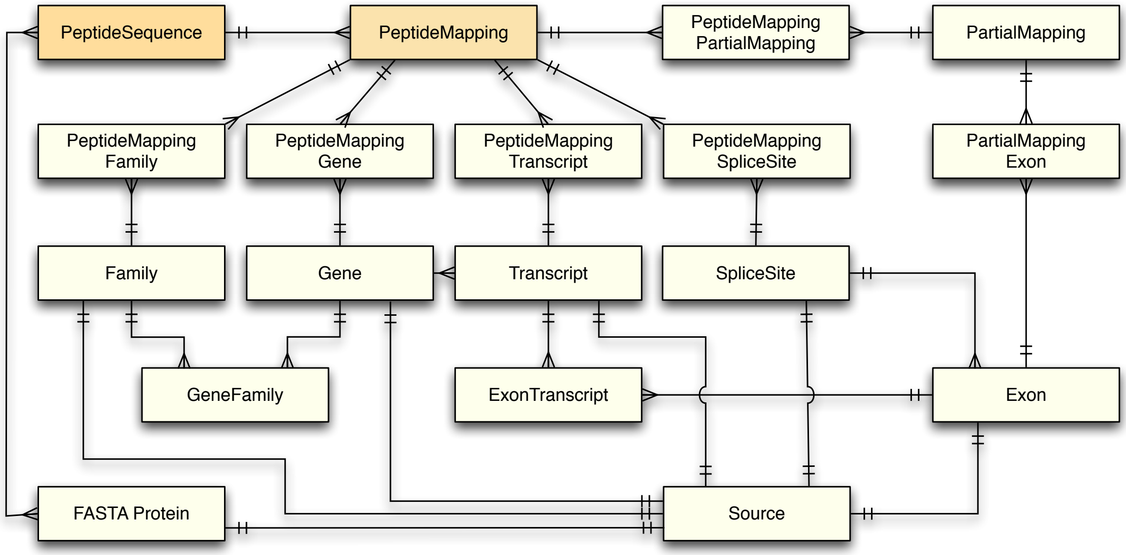


Figure 4.2: GenoMS-DB database schema. The `PeptideSeqeunce` together with the `PeptideMapping` table are the central elements in the schema (highlighted in orange). The former comprises peptide properties alongside peptide identification scores where available. The latter provides the genomic mapping coordinates for every peptide in the database. Note that this design is not fully normalised to provide optimal performance and ease of use. Yet, data integrity is guaranteed by the carefully controlled data integration process. Later use of the database only updates specific attributes of the `PeptideSequence` table that cannot lead to inconsistent data. The notation used in this schema is a simplified Crow's Foot notation.

In the following section the design of GenoMS-DB is discussed in more detail.

The `PeptideSequence` table stores all peptide related information, such as peptide sequence, the number of missed cleavages or sequence length. The attribute `seqKQLI` stores a sequence version with all Lysine (K) and Glutamine (Q) residues replaced with "1" and all Leucine (L) and Isoleucine (I) replaced with "2". Sequences that only differ in either K/Q or L/I residues cannot be differentiated in low resolution fragmentation spectra. Therefore, the uniqueness of every peptide within the proteome accounting for these ambiguities is tested in a post database build process. The results are stored as an integer number in the `ambigEnsVega` and `ambiguity` attributes, relating to the number of genomic loci these substituted peptides match within the tested space. The former attribute confines this space to Ensembl and Vega annotation only, whereas the latter also accounts for a much larger search space, such as predicted sequences generated by Augustus.

It should be noted that this approach is a simplification of the potential complexity involved in peptide-spectrum matching. For example, a modified residue such as deamidated Asparagine could result in a similar mass as an unmodified amino acid such as aspartic acid, which differs by only 3.6×10^{-5} Dalton. The selection of the wrong monoisotopic peak or the occurrence of a single-nucleotide polymorphism can lead to similar artefacts, hence sound MS data processing and database searching requires careful selection of parameters to avoid these caveats. Finally, the `PeptideSequence` table also stores identification information if the peptide was identified with Mascot and scored with Mascot Percolator. A separate table could have been directly linked with this table, but as pointed out earlier, the design is partly denormalised to minimise unnecessary table joins and yet provide a robust database schema.

To link peptide sequences with their genomic context, the `PeptideSequence` table has a one-to-many relationship with the `PeptideMapping` table. The latter represents the central proteogenomic element in the database, comprising attributes such as `start`, `end`, `chromosome` and `strand`. If a peptide spans one or multiple splice sites, the peptide maps partially to different genomic locations, which are stored in the `PartialMapping` table. If no splice site is spanned, the information of the `PeptideMapping` table and the `PartialMapping` table are redundant. If multiple alternative gene products give rise to multiple distinct peptide sequences, which have the same genomic start and end coordinate but differ in the partial genomic mapping due to variation in splicing, an alternative `PeptideMapping` entity is created in the table for each distinct case. Hence, a `PeptideMapping` entity maps to one and only one `PeptideSequence` entity.

`PartialMapping` entities are linked to the underlying `Exon`, which in turn links to related `SpliceSite` and `Transcript` entities. A `SpliceSite` entity maps to two distinct `Exon` entities, one representing the donor and one the acceptor exon. Since a transcript contains one or many exons and one exonic region can be part of multiple transcripts, the `Exon` table is linked to the `Transcript` table by a many-to-many relationship. Genes can give rise to multiple alternative transcripts, but a transcript always belongs to one gene, hence the relationship between the `gene` and `Transcript` table is one-to-many. To complete the genomic relationships, each `Gene` entity can belong to multiple `GeneFamily` entities and *vice versa.* The `Gene`, `Transcript` and `Exon` table store only the most relevant information, such as the chromosomal coordinates, identifiers, annotation status or short descriptions.

The `PeptideMapping` table has a many-to-many relationship with the `GeneFamily`, `Gene`, `Transcript` and `SpliceSite` tables. This redundant data representation minimises table joins to optimise performance and enables simplified query building.

The following example should clarify the rationale behind this design: to identify all splice sites that are associated with a set of peptides, the `PeptideSequence` table is joined with the table holding the peptide-genome mapping information, which in turn is joined with the table that maps to these requested splice sites. A fully normalised design would have required significantly more table joins, complicating the SQL query and slowing down the execution performance.

Lastly, the `PeptideSequence` table also has a many-to-many relationship with the `FastaProtein` table. This table is only populated if FASTA sequence information is integrated into the database, whereby no genomic mapping is available. This can be useful for protein contaminants or laboratory proteins as well as protein databases which are to be compared with Ensembl, Vega or Augustus. The `Source` table has a one-to-one relationship with `GeneFamily`, `Gene`, `Transcript`, `SpliceSite` and `FastaProtein` table. It comprises the version and type of the source database, which was used to build GenoMS-DB.

Overall the partly denormalised database represents a powerful tool to analyse proteogenomic data in an effective and efficient way. GenoMS-DB is individually built for genome annotation resources that are to be validated and refined with proteomics data, such as a specific Ensembl or Vega builds.

4.2.3 Mascot search database construction

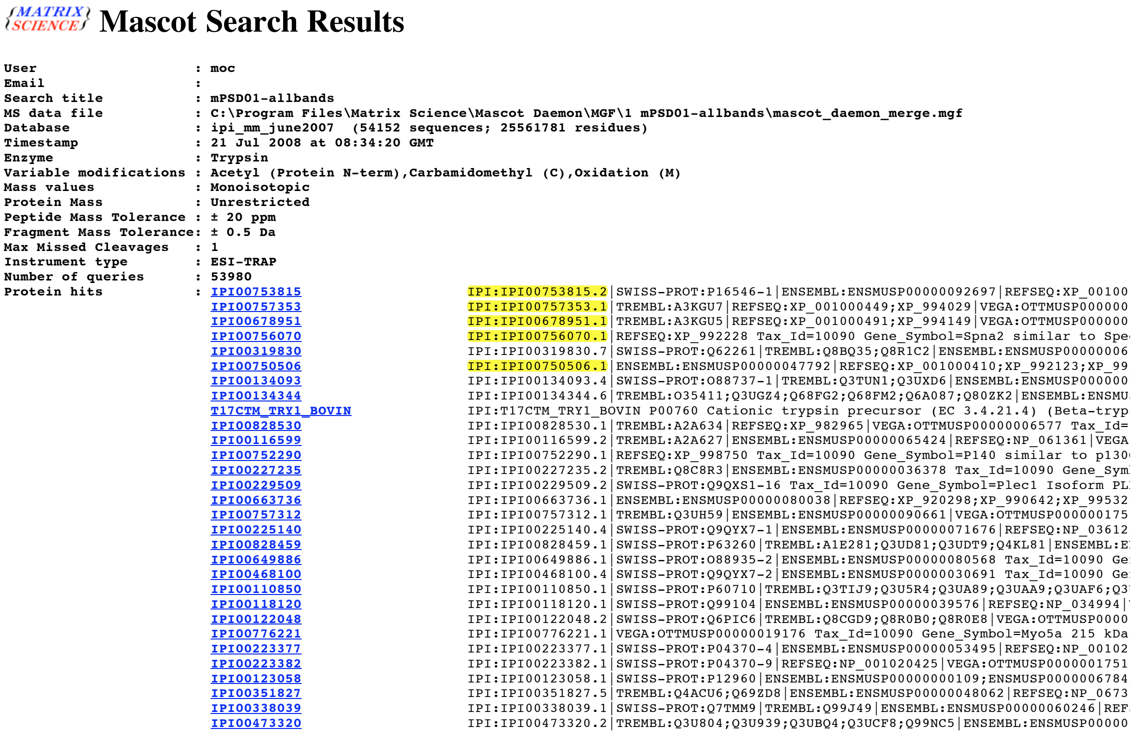
The search "database" of Mascot is simply a FASTA text file, where individual protein sequence entries are concatenated with the protein identifier as a delimiter, with alternative protein isoforms being handled as an individual protein entry.

Ensembl annotates alternative isoforms conservatively and as shown in chapter 5, of the almost 23 thousand protein coding Ensembl genes in mouse, only 8,877 had multiple protein coding isoforms annotated (Ensembl 54). However, a small subset of 1,542 protein coding Ensembl genes were predicted to code for 13,664 transcripts. Recently, Wang *et al.* (2008) has shown that more than 90% of human genes are expected to code for alternative isoforms. It is anticipated that manual annotation and improved automated annotation systems together with increased availability of expression data from different cell types or tissues will further increase the number of known alternative gene products.

Protein variants typically share most peptides (see section 5.3.2.4), which leads to significant peptide redundancy in these text based FASTA files. Moreover, when gene finding algorithms predict tens of alternative isoforms for a gene, a compact representation eliminating the inherent peptide redundancy is required.

Many database search tools, including Mascot, do not remove this redundancy since they sequentially cross correlate the *in silico* digested peptides (personal communication, John Cottrell, Matrix Science, London). Therefore search times scale linearly with database size.

To enable a compact representation, Martens *et al.* (2005b) proposed to digest the proteome into a peptide centric database that can be filtered and indexed to remove redundancy. For this, peptides are concatenated by an artificial residue that is used as a spacer element to separate individual peptides. The Mascot search enzyme settings need to be set accordingly to cleave at this artificial spacer element (see



{MATRIX SCIENCE} Mascot Search Results

```
User                    : moc
Email                   :
Search title            : mPSD01-allbands
MS data file            : C:\Program Files\Matrix Science\Mascot Daemon\MGF\1 mPSD01-allbands\mascot_daemon_merge.mgf
Database                : ipi_mm_june2007  (54152 sequences; 25561781 residues)
Timestamp               : 21 Jul 2008 at 08:34:20 GMT
Enzyme                  : Trypsin
Variable modifications  : Acetyl (Protein N-term),Carbamidomethyl (C),Oxidation (M)
Mass values             : Monoisotopic
Protein Mass            : Unrestricted
Peptide Mass Tolerance  : ± 20 ppm
Fragment Mass Tolerance: ± 0.5 Da
Max Missed Cleavages    : 1
Instrument type         : ESI-TRAP
Number of queries       : 53980
Protein hits            : IPI00753815                  IPI:IPI00753815.2|SWISS-PROT:P16546-1|ENSEMBL:ENSMUSP00000092697|REFSEQ:XP_00100
                          IPI00757353                  IPI:IPI00757353.1|TREMBL:A3KGU7|REFSEQ:XP_001000449;XP_994029|VEGA:OTTMUSP000000
                          IPI00678951                  IPI:IPI00678951.1|TREMBL:A3KGU5|REFSEQ:XP_001000491;XP_994149|VEGA:OTTMUSP000000
                          IPI00756070                  IPI:IPI00756070.1|REFSEQ:XP_992228 Tax_Id=10090 Gene_Symbol=Spna2 similar to Spe
                          IPI00319830                  IPI:IPI00319830.7|SWISS-PROT:Q62261|TREMBL:Q8BQ35;Q8R1C2|ENSEMBL:ENSMUSP00000006
                          IPI00750506                  IPI:IPI00750506.1|ENSEMBL:ENSMUSP00000047792|REFSEQ:XP_001000410;XP_992123;XP_99
                          IPI00134093                  IPI:IPI00134093.4|SWISS-PROT:O88737-1|TREMBL:Q3TUN1;Q3UXD6|ENSEMBL:ENSMUSP000000
                          IPI00134344                  IPI:IPI00134344.6|TREMBL:O35411;Q3UGZ4;Q68FG2;Q68FM2;Q6A087;Q80ZK2|ENSEMBL:ENSMU
                          T17CTM_TRY1_BOVIN            IPI:T17CTM_TRY1_BOVIN P00760 Cationic trypsin precursor (EC 3.4.21.4) (Beta-tryp
                          IPI00828530                  IPI:IPI00828530.1|TREMBL:A2A634|REFSEQ:XP_982965|VEGA:OTTMUSP0000006577 Tax_Id=
                          IPI00116599                  IPI:IPI00116599.2|TREMBL:A2A627|ENSEMBL:ENSMUSP00000065424|REFSEQ:NP_061361|VEGA
                          IPI00752290                  IPI:IPI00752290.1|REFSEQ:XP_998750 Tax_Id=10090 Gene_Symbol=P140 similar to p130
                          IPI00227235                  IPI:IPI00227235.2|TREMBL:Q8C8R3|ENSEMBL:ENSMUSP00000036378 Tax_Id=10090 Gene_Sym
                          IPI00229509                  IPI:IPI00229509.2|SWISS-PROT:Q9QXS1-16 Tax_Id=10090 Gene_Symbol=Plec1 Isoform PL
                          IPI00663736                  IPI:IPI00663736.1|ENSEMBL:ENSMUSP00000080038|REFSEQ:XP_920298;XP_990642;XP_99532
                          IPI00757312                  IPI:IPI00757312.1|TREMBL:Q3UH59|ENSEMBL:ENSMUSP00000090661|VEGA:OTTMUSP000000175
                          IPI00225140                  IPI:IPI00225140.4|SWISS-PROT:Q9QYX7-1|ENSEMBL:ENSMUSP00000071676|REFSEQ:NP_03612
                          IPI00828459                  IPI:IPI00828459.1|SWISS-PROT:P63260|TREMBL:A1E281;Q3UD81;Q3UDT9;Q4KL81|ENSEMBL:E
                          IPI00649886                  IPI:IPI00649886.1|SWISS-PROT:O88935-2|ENSEMBL:ENSMUSP00000080568 Tax_Id=10090 Ge
                          IPI00468100                  IPI:IPI00468100.4|SWISS-PROT:Q9QYX7-2|ENSEMBL:ENSMUSP00000030691 Tax_Id=10090 Ge
                          IPI00110850                  IPI:IPI00110850.1|SWISS-PROT:P60710|TREMBL:Q3TIJ9;Q3U5R4;Q3UA89;Q3UAA9;Q3UAF6;Q3
                          IPI00118120                  IPI:IPI00118120.1|SWISS-PROT:Q99104|ENSEMBL:ENSMUSP00000039576|REFSEQ:NP_034994|
                          IPI00122048                  IPI:IPI00122048.2|SWISS-PROT:Q6PIC6|TREMBL:Q8CGD9;Q8R0B0;Q8R0E8|VEGA:OTTMUSP0000
                          IPI00776221                  IPI:IPI00776221.1|VEGA:OTTMUSP00000019176 Tax_Id=10090 Gene_Symbol=Myo5a 215 kDa
                          IPI00223377                  IPI:IPI00223377.1|SWISS-PROT:P04370-4|ENSEMBL:ENSMUSP00000053495|REFSEQ:NP_00102
                          IPI00223382                  IPI:IPI00223382.1|SWISS-PROT:P04370-9|REFSEQ:NP_001020425|VEGA:OTTMUSP000001751
                          IPI00123058                  IPI:IPI00123058.1|SWISS-PROT:P12960|ENSEMBL:ENSMUSP00000000109;ENSMUSP000006784
                          IPI00351827                  IPI:IPI00351827.5|TREMBL:Q4ACU6;Q69ZD8|ENSEMBL:ENSMUSP00000048062|REFSEQ:NP_0673
                          IPI00338039                  IPI:IPI00338039.1|SWISS-PROT:Q7TMM9|TREMBL:Q99J49|ENSEMBL:ENSMUSP00000060246|REF
                          IPI00473320                  IPI:IPI00473320.2|TREMBL:Q3U804;Q3U939;Q3UBQ4;Q3UCF8;Q99NC5|ENSEMBL:ENSMUSP00000
```

Figure 4.3: Screenshot of a typical Mascot search result page. In this example, five of the top six protein entries belong to the same gene *spectrin alpha 2*, indicated in yellow.

section 5.2.2). In the following section I present two alternative peptide-level search databases, both of which can be directly derived from the GenoMS-DB database.

Gene centric peptide level database

The first approach concatenates all peptides from GenoMS-DB on a per gene basis in a non-redundant manner, thereby compressing the database size and reducing the database search times, since multiple occurrences of a peptide in alternative gene products are collapsed into one gene entry. This "gene-centric" search database can be useful to simplify the analysis of a complex sample, when there is no need to distinguish the individual isoforms.

To contrast this approach with the classic protein database search, figure 4.3 shows a typical Mascot search result page with a list of protein hits that were identified from a standard protein database (IPI). Browsing this list manually to analyse the dataset can be cumbersome, especially when multiple protein isoforms belong to the same underlying gene. In the illustrated example, five of the top six proteins belong to the same gene *spectrin alpha 2*. On the other hand, the search result against the gene-centric database introduced above only returned the gene of interest instead of the five individual entries, which can be useful for users who directly browse the search results with Mascot.

An alternative strategy would be to use GenoMS-DB to selectively export only peptides that map uniquely to one locus. This would enable a targeted proteomics experiment with the focus of gene identification. Even more complex scenarios could be designed, such as the selection of peptides that enable discrimination of protein isoforms.

Strict peptide level database

For my work however, no protein nor gene information was required from the Mascot search results, since the genomic context is established by integrating these peptide identifications into GenoMS-DB, which is discussed in section 4.2.4. Hence, this second approach simply concatenates all peptides stored in the database in a non-redundant manner, where 1000 peptides at a time are binned and concatenated into one FASTA entry, separated only by a spacer element (e.g. artificial residue “J”). The FASTA header is set to the number of the entry as a text string. By default, there is no selection for specific peptide parameters during the FASTA file build process, but optionally sequence length or number of missed cleavages could be restricted. By default only the database build parameters (see 4.2.1) limit the peptides available in the resulting FASTA file. Given a peptide is only stored once in the whole FASTA file, this method further reduces peptide-level redundancy that is caused by peptides that are present in multiple genes. This database type is used in the pilot study conducted in chapter 5.

Listing 4.1: A simple SQL query example that selects the list of Ensembl protein coding genes from GenoMS-DB with the corresponding number of identified peptides per gene that exceed a PEP score of 20 and match a unique genomic locus within the Ensembl annotation.

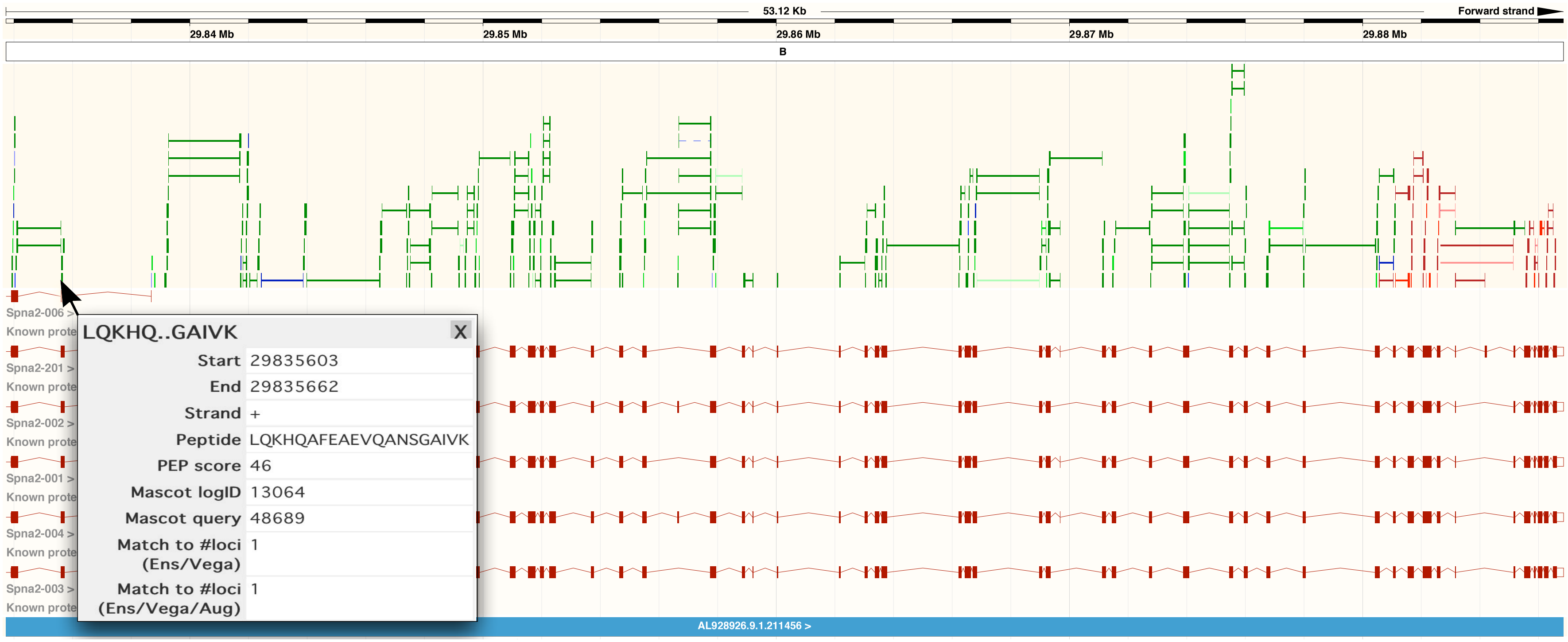
```
select geneID, count(distinct peptideSequenceID)
  from PeptideSequence
  inner join PeptideMapping using(peptideSequenceID)
  inner join PeptideMapping_Gene using(peptideMappingID)
  inner join Gene using(geneID)
  inner join Source using(sourceID)
  where PeptideSequence.score > 20
  and PeptideSequence.ambigEnsVega = 1
  and Source.db = "ensembl"
  and Gene.biotype = "protein_coding"
  group by geneID
```

4.2.4 Results integration

After searching tandem MS data with Mascot against the peptide level database described above, search results are post-processed by Mascot Percolator and stored in a specified folder. A Perl script, which can be executed on a regular basis with a job scheduler such as “Cron” on Unix-like operating systems, scans this results folder and processes the Mascot Percolator result files in the following way: (a) firstly peptides that match to user defined contaminants or laboratory proteins are filtered out from the search results; (b) all remaining peptide identifications that exceed user defined scoring criteria are then persisted in the PeptideSequence table of GenoMS-DB (figure 4.2) by updating the relevant attributes. Currently the system only keeps the best scoring peptide identification in GenoMS-DB.

4.2.5 SQL analysis and DAS server implementation

After results integration, GenoMS-DB can be leveraged for large scale proteogenomics analysis, employing standard SQL queries. For each analysis, a custom SQL query statement can be designed and executed, providing efficient means to an otherwise complex manual analysis process. In listing 4.1 a simple SQL query example is



LQKHQ..GAIVK X

Start	29835603
End	29835662
Strand	+
Peptide	LQKHQAFEAEVQANSGAIVK
PEP score	46
Mascot logID	13064
Mascot query	48689
Match to #loci (Ens/Vega)	1
Match to #loci (Ens/Vega/Aug)	1

Figure 4.4: The Ensembl browser representing the *spectrin alpha 2* locus. The page shows the Ensembl annotation tracks (red tracks on lower half) as well as a peptide identifications (upper half), which are integrated into Ensembl via the DAS server that queries GenoMS-DB and renders the peptide mapping accordingly. Peptide features can be selected and meta-information can be displayed (see illustrated information window), such as the exact genomic mapping coordinates, scoring information, number of genomic loci the peptide matches perfectly or the Mascot logging identifier and the spectrum query number, enabling to track back into the original Mascot results. The type of colour of these peptide features depends on the uniqueness of the peptide within the genome: unique within Ensembl/Vega/Augustus (green), unique within Ensembl/Vega (blue), ambiguous and multiple matches (red). The brightness of the peptide features is correlated to the Mascot Percolator Posterior Error Probability (PEP) score: $5 \times 10^{-2} \leq PEP < 1 \times 10^{-2}$ (light), $1 \times 10^{-2} \leq PEP < 1 \times 10^{-3}$ (medium), $PEP \geq 1 \times 10^{-3}$ (dark). Although these data are also available through the information window, the Vega annotation pipeline cannot show meta-data and hence colour coding was a way to provide uniqueness and scoring information to annotation curators. This gene is also used as an example in figure 4.3 and is further discussed in section 4.2.3

Listing 4.2: A more complex SQL query example that selects the list of Augustus genes that were validated by identified peptides that exceed a PEP score of 30 and map to an unique genomic locus and not match either the Ensembl or the Vega annotation in order to only select truly novel coding regions. The resulting list provides all gene details, the matching peptide sequences and the genomic peptide mapping information.

```
select G.*, S.seq, P.*
  from PeptideSequence as S
  inner join PeptideMapping as P using(sequenceID)
  inner join PeptideMapping_Gene using(peptideMappingID)
  inner join Gene as G using(geneID)
  inner join Source using(sourceID)
  where PEP > 30
  and db = "augustus"
  and ambiguity = 1
  and not exists (
    select * from PeptideSequence
      inner join PeptideMapping using(sequenceID)
      inner join PeptideMapping_Gene using(peptideMappingID)
      inner join Gene using(geneID)
      inner join Source using(sourceID)
      where PEP > 30
      and db in (’ensembl’, ’vega’)
      and S.seq = seq
  )
```

provided that demonstrates how the list of Ensembl genes with the corresponding peptide matches can be selected. Listing 4.2 is a more complex query that demonstrates the simplicity with which relatively complex questions can be answered.

SQL queries were also used for the development of a stand-alone proteogenomic distributed annotation server (DAS) (Dowell *et al.*, 2001) that is accessing the integrated data of GenoMS-DB. When the DAS server receives a request for a specific genomic region, all peptides together with their associated genomic mapping data are selected within the defined region and returned as DAS features. The required SQL statements were implemented into ProServer, an extendable Perl based DAS feature server (Finn *et al.*, 2007). Peptide features provided through the DAS server can then be attached to genome browsers, such as Ensembl (figure 4.4). Features can be supplemented with meta-information that are required for automatic or manual genome annotation: peptide mapping details, uniqueness of peptide within the genome, scoring details or original Mascot spectrum ID.

4.3 Conclusion

In this work I have developed a proteogenomic pipeline that enables efficient and effective large scale genome wide data analysis. It leverages the power of a relational database, which is at the core of the system. My database design allows high performance analysis with easy to construct SQL statements. GenoMS-DB, integrates all relevant information for subsequent proteogenomic analysis. This database accepts annotation data from Ensembl, Vega or Augustus, as well as supplemental data from FASTA databases. Therefore, proteins or protein isoforms not present in these databases cannot be identified with this pipeline, which is generally true for any database search algorithm (see section 1.1.1). These data are digested *in silico* and stored in GenoMS-DB together with their genomic context. The genomic mapping coordinates are calculated, enabling the ad-hoc mapping of millions of peptides with GenoMS-DB, since alignment tools to map peptide sequences against the genome are not required. Integrated peptides can be exported to non-redundant peptide collections, which can in turn be used by Mascot as efficient search databases. Results from Mascot Percolator can also be stored in GenoMS-DB. This complete integration enables proteogenomic analysis with standard SQL. Even complex questions can be formulated in a few lines of SQL code, whereby analysis is fully automated, avoiding any manual intervention. Large scale studies can be carried out since genome mapping is readily available through GenoMS-DB. This also allows the analysis of theoretical peptide collections, such as the whole proteome. The next chapter tests and uses this pipeline in a pilot study.

Chapter 5

Refining annotation of the mouse genome using mass spectrometry

5.1 Introduction

This chapter applies the work of the previous chapters in the form of a pilot study to validate and extend genome annotation for *Mus musculus* as available through Ensembl and Vega. In section 1.2 the current strategies of genome annotation, including Ensembl and Vega, were discussed in detail and a brief introduction to the field of proteogenomics was provided in section 1.3.

In this work I build upon these efforts and apply a two stage search strategy with the aim of validating and refining mouse (Waterston *et al.*, 2002) genome annotation for the first time. MS data, obtained from the Peptide Atlas project (Desiere *et al.*, 2006) and generated in-house, was first searched against a peptide centric non-redundant superset of Ensembl, Vega and IPI (Kersey *et al.*, 2004) that was generated with GenoMS-DB (see chapter 4). IPI, commonly used as a standard protein database for MS proteomics, was included for completeness. It is expected that these databases comprise most of the proteome and due to the limited search space, peptide identification sensitivity is maintained at a high level. In a second stage, I have incorporated protein predictions from Augustus that significantly inflate search space, but enable refinement of existing gene annotations. These data were then used to validate existing Ensembl and Vega gene models at the gene, exon and splice-boundary level. Interestingly, I show evidence of alternatively translated protein variants and discuss the implications of not detecting any translational evidence for transcripts that are tagged with nonsense mediated decay (NMD) (Maquat, 2005), which are discussed in section 5.3.2.5. Furthermore, I highlight the value of proteogenomics to refine gene structures: significant peptide identifications were made outside annotated coding regions as well as within annotated pseudogenes. Novel exons or exon boundaries, as well as a set of novel genes that are not annotated in existing databases, were also identified. Lastly, the pre-computation of genome mapping for all peptides, as provided through GenoMS-DB, enabled me to assess for the first time not only the value of proteogenomics for observed peptides, but also offer a perspective of what could theoretically be achieved with this approach.

Parts of this chapter will be published by the author of this thesis (Markus Brosch), my supervisors (Tim Hubbard, Jyoti Choudhary), Lu Yu and Mark Collins who run the mass spectrometry experiments and Jennifer Harrow and co-workers who will further investigate the results in collaboration with the HAVANA team at the Wellcome Trust Sanger Institute.

5.2 Methods

5.2.1 Tandem mass spectrometry data

This pilot study is based on 10,465,149 tandem MS spectra, where 729,583 spectra were obtained from in-house experiments on nuclear protein extracts of murine embryonic stem cells and murine brain membrane fractions. These experiments were performed by Lu Yu and Mark Collins and the experimental procedures follow the methods described in section 2.2 (sample 1).

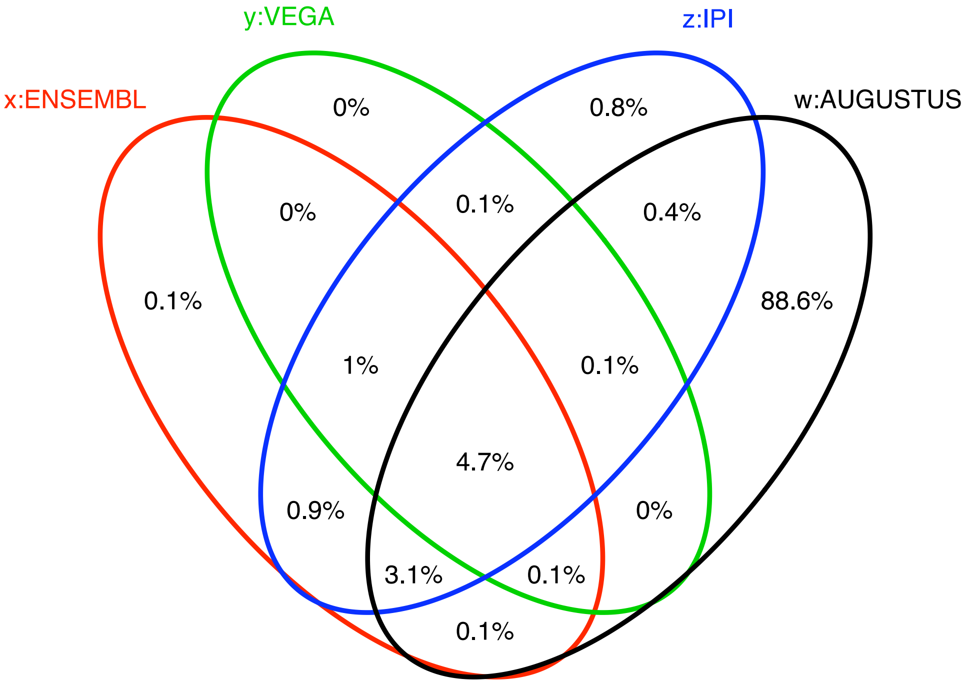
9,735,566 spectra were provided by Eric Deutsch and Zhi Sun (Institute for Systems Biology, Seattle, US) as Mascot mgf peaklist files. These data were selected from the *Mus Musculus* Peptide Atlas data repository (unpublished, Feb. 2009 data snapshot, `http://www.peptideatlas.org/repository/`). Data were not associated with any publication records, but short descriptions suggested sampling across various tissues of mouse such as brain, liver, lung, heart, kidney, testes and placenta.

5.2.2 Search database construction

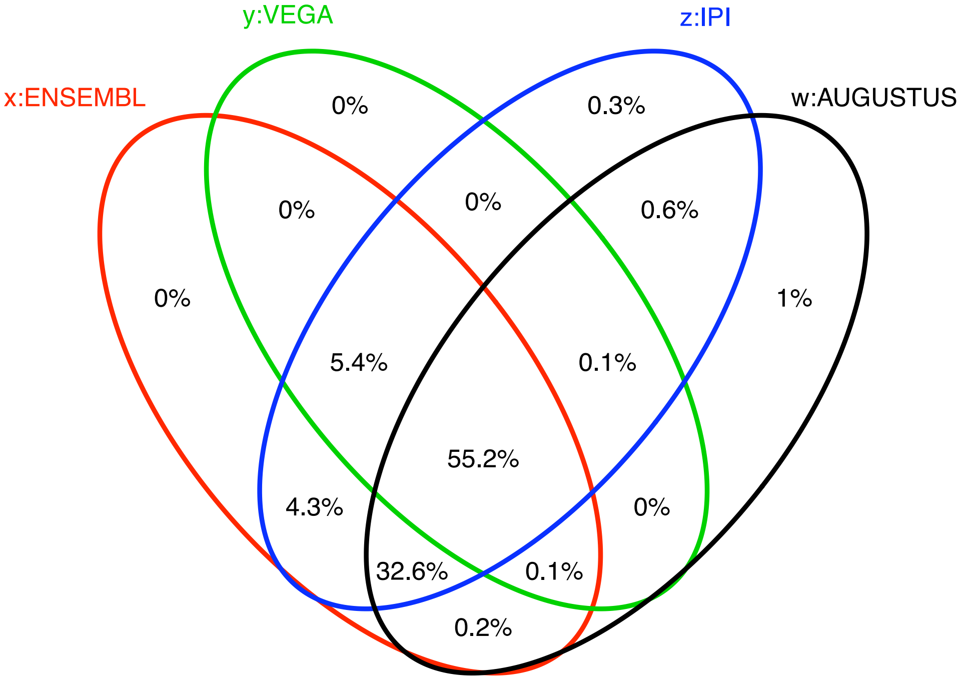
All gene products from Ensembl (mouse, release 54) and Vega (mouse, release 35, December 2008) as well as all protein entries from the IPI database (mouse, v3.55) were tryptically digested *in silico* (cutting after arginine and lysine), allowing up to two missed tryptic cleavages. Protein N-terminal methionine excision by Methionine aminopeptidase (Frottin *et al.*, 2006) was considered and therefore the N-terminus peptide of a protein was staggered. In addition, all potential 2,690 mouse NMD products (internal data release, February 2009) and common external contaminants from cRAP (a maintained list of contaminants and laboratory proteins provided through the Global Proteome Machine Organisation, `http://www.thegpm.org/crap/index.html`) were appended and processed in the same way. In total, 3,276,592 distinct tryptic peptides where generated and integrated into GenoMS-DB together with the corresponding genomic context (see chapter 4). Figure 5.1a illustrates the peptide distribution between the different data sources.

The search database (FASTA flatfile) for Mascot was built by concatenating all tryptic peptides in a non-redundant manner, thereby eliminating multiple occurrences of a peptide in alternative gene products as described in section 4.2.3. The artificial residue “J” was introduced as a spacer element to separate individual peptides, similar to the method described by Schandorff *et al.* (2007).

A second search database was constructed that extends the former database by *ab initio* Augustus (version 2.0.3) gene predictions, resulting in 28,742,036 distinct pep-



(a) All *in silico* digested peptides



(b) All identified peptides ($PEP \leqslant 0.01$, filtered)

Figure 5.1: Four-way Venn diagram of all distinct fully tryptic peptides from Ensembl, Vega, IPI and Augustus.

tides. For this, DNA sequences (*Mus Musculus*, NCBI37) for each chromosome were downloaded from the ensembl data resource (`ftp://ftp.ensembl.org/pub/current_fasta/mus_musculus/dna/`) and Augustus was run on all chromosome sequences, each of which was chopped into 50 Mb slices, overlapping by 2.5 Mb. The Augustus release provided a script (`join_aug_pred.pl`) to re-assemble predictions from individual slices and those that spanned the slice boundaries. The resulting file in GFF format was processed and converted into tryptic peptides in the same manner as described above and imported into GenoMS-DB.

In total, three individual Augustus runs were performed: (a) standard mode, (b) over-prediction mode and (c) single exon gene mode. The standard mode (a) used the recommended default parameters that provide similar performance as other gene prediction tools (Guigo and Reese, 2005; Stanke *et al.*, 2006). In mode (b) Augustus was run with parameter settings that provide maximum sensitivity and also allowing for shorter gene predictions. When Augustus is used directly for genome annotation purposes without any subsequent validation, false positive predictions are generally unwanted and a trade-off between sensitivity and accuracy needs to be made. However, here the aim was to minimize false negatives and thereby maximize sensitivity. The false positive Augustus gene predictions are controlled in the MS peptide-spectra correlation stage with stringent and robust scoring, essentially acting as a validator for this large set of potential genes. Lastly, in mode (c) Augustus was optimized to predict single exon containing genes, which are known to be difficult to annotate. The detailed parameters for these customised runs (b) and (c) were as follows (provided by the author of Augustus, Mario Stanke, personal communication, November 2008):

b) The parameter `/Constant/min_coding_len` in the configuration file `config/-species/human/human_parameters.cfg` was set to 50. The Augustus program parameters were set to: `--sample=1000 --maxtracks=10 --minexonintron-prob=0 --minmeanexonintronprob=0 --alternativesfromsampling=true`.

c) Same parameters as in (b) were used, but additionally the following changes were made in the configuration file `config/model/trans_shadow_partial.pbl`: single exon (final intergenic region) was set to: `1 0 10` and reverse single exon (intergenic region) was set to `24 0 10`.

For both FASTA databases, corresponding decoy databases were constructed for significance assessment (see section 1.1.2.3). However, the default Mascot decoy method was not sufficient: Mascot randomizes each protein sequence (FASTA entry), while retaining the average amino acid composition and length. This does not suffice for the FASTA entries that were artificially constructed in this work, where fully tryptic peptides are concatenated with a spacer residue “J”. The chance of obtaining an arginine (R) or lysine (K) residue immediately before “J” when the decoy script is applied, would be approximately 10% (2 in 20 residues), meaning that the decoy database would be significantly depleted in “real” potential decoy matches. I therefore implemented a Perl script that shuffles each unique peptide entry individually by maintaining the tryptic cleavage site, instead of shuffling a whole FASTA entry. After each randomization round, it is tested whether the peptide was either produced before or exists as a natural peptide in the target database. In both of these cases, the randomization process is continued until a new random fully tryptic peptide was determined. Overall this process maintains the trypticity, the amino acid composition, the peptide length distribution as well as the number of peptides in an entry.

5.2.3 Data processing and database searching with Mascot

In-house LTQ-FT and LTQ-FT Ultra (Thermo Fisher Scientific) generated MS raw data files were processed to peak lists with BioWorks (version 3.2 and 3.3, Thermo Fisher Scientific). Processing parameters were identical to those used in section 2.2.3.

All MS peaklist data (in-house and PeptideAltas) were searched with Mascot and post processed with Mascot Percolator. For this, each peaklist file was searched against both target and decoy databases using an enzyme setting that is compatible with the custom made peptide centric search databases; therefore the artificial amino acid “J” was introduced under the mascot config file that defines the amino acid masses. “J” was set to a mass that does not correspond to a naturally occurring amino acid (300 Da). The enzyme was set to cut at the N- and C-terminal of the peptide, thereby only fully tryptic peptides that were separated by “J” were searched with Mascot. For in-house data, parameters were identical to 2.2.4 with the parent mass tolerance set to 20 ppm. Peptide Atlas data was searched with the parameters supplemented with the data file.

5.2.4 Post processing with Mascot Percolator and results integration

Mascot search results were post-processed with Mascot Percolator (1.09, default settings) using Percolator version 1.12. Each peptide-spectrum match was assigned a q-value and a posterior error probability. All sequences that either had a posterior error probability greater than 0.05, or matched any entry of the contaminants protein list, were filtered out. The remaining peptide identifications were integrated into GenoMS-DB (see section 4.2.4).

The Distributed Annotation Server (Dowell *et al.*, 2001) (DAS), which is build on top of GenoMS-DB (see section 4.2.5), provides access to the results of this pilot study. Meta-information for each peptide is provided in the form of scoring statistics (q-value, $-10 \times log_{10}$ transformed posterior error probability), genomic uniqueness of the peptide within the Ensembl/Vega and Ensembl/Vega/Augustus annotation, Mascot search log ID and spectrum ID. The DAS data source can be accessed at `http://das.sanger.ac.uk/das/ms_das/` and can be readily integrated into genome browsers that allow embedding external DAS sources.

5.3 Results and Discussion

5.3.1 Peptide identification and genome mapping

It is expected that the superset of the Ensembl, Vega and IPI database represents most of the mouse proteome and was therefore used for the first pass Mascot search. After post-processing with Mascot Percolator, in total 1,491,410 and 1,772,159 peptides were identified at a q-value (a more advanced notion of the false discovery rate, see section 1.1.2.2) of 1% and 5%, respectively. Applying a maximum allowed probability of 1% and 5% of an individual peptide match to be incorrect (posterior error probability, PEP), 1,124,724 and 1,358,323 peptides were identified, corresponding to a q-value of less than 0.14% and 0.59%, respectively.

When data was searched against the database that was supplemented with the Augustus predictions (see methods), 16% fewer identifications (1,253,074 and 1,490,020 at a q-value of 1% and 5%) were made due to the search space inflation of almost one order of magnitude (figure 5.1a), which increases the chance of incorrectly identifying peptides and hence more restrictive scoring was required in order to maintain the q-value (this is discussed in detail in chapter 2). 967,131 and 1,171,060 peptides were identified with a maximum PEP of 1% and 5%, corresponding to q-values of 0.12% and 0.57%, respectively. Interestingly, 88.1% of the distinct peptide identifications ($PEP \leqslant 1\%$) overlapped between Ensembl and the Augustus predictions (figure 5.1b), suggesting good sensitivity for the chosen Augustus configuration.

For subsequent analyses, only the best PEP and q-value score for each peptide sequence was considered (PEP significance threshold $\leqslant$ 5%), resulting in 95,606 distinct peptide identifications, 3,260 of which matched common contaminants. Since all fragment ion (MS/MS) data were generated with collision induced fragmentation (Biemann, 1988; Roepstorff and Fohlman, 1984) and analyzed with a low resolution instrument, Leucine/Isoleucine as well as Lysine/Glutamine sequence isoforms could not be differentiated due to identical or similar residue mass. Therefore, sequences

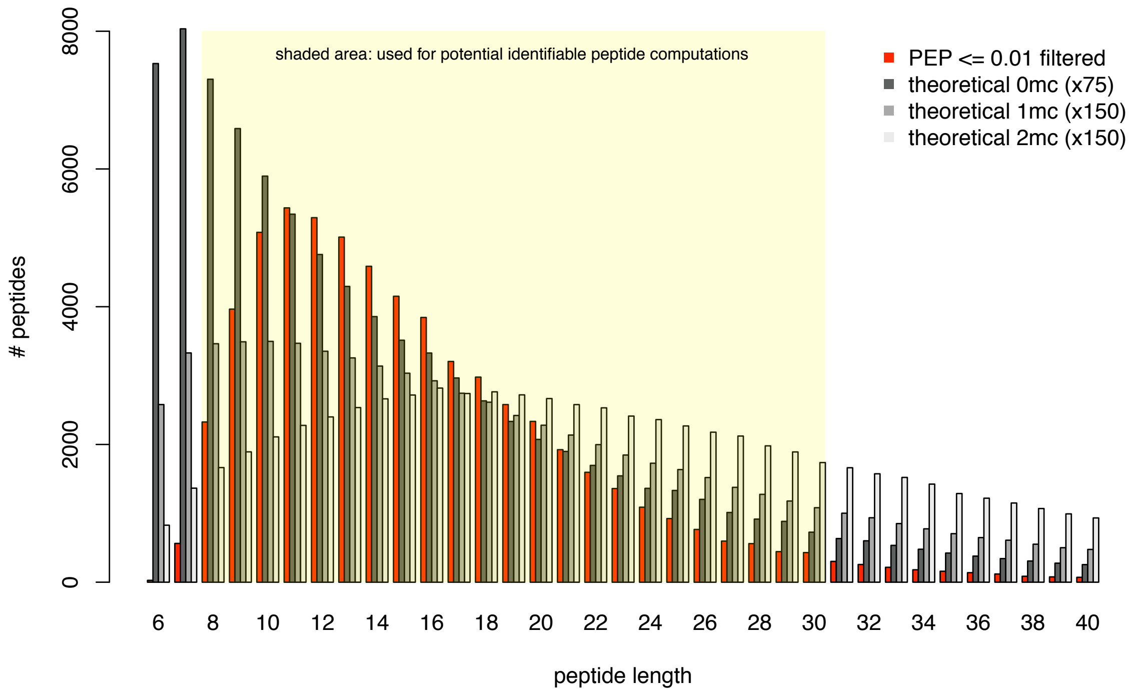


Figure 5.2: Peptide length distribution of identified peptides that passed the filtering criteria (red) and of the potential identifiable peptides as derived from the protein digest (black/grey). The number of peptides for the latter are scaled down by a factor of 75 for peptides without a missed cleavages (0mc) site (black) and by 150 for peptides with one (1mc, mid-grey) or two (2mc, light-grey) missed cleavages respectively. When theoretical genome annotation coverage was computed in this work, only peptides ranging from 8 to 30 residues were considered (shaded area).

that have an isoform in any of these residues were filtered out (1,159 cases). In total, 83% (76,029) of the remaining peptides mapped unambiguously to one genomic locus. Since only fully tryptic peptides were considered, it was further tested whether a semi-tryptic form of the peptide sequence mapped elsewhere in the genome (758 cases). As a last measure, it was evaluated whether peptides with one residue substitution or indel could be identified elsewhere in the genome (6,685 cases, preferentially short peptide identifications), since coding SNPs were not considered in this study. Therefore, a total of 68,586 distinct peptides built the basis for subsequent genome annotation. However, peptide-spectrum matches with a PEP between 1-5% were exclusively used as supplementing peptides and only peptide identifications with a PEP of 1% or better (58,574 cases) were used as a primary annotation data source.

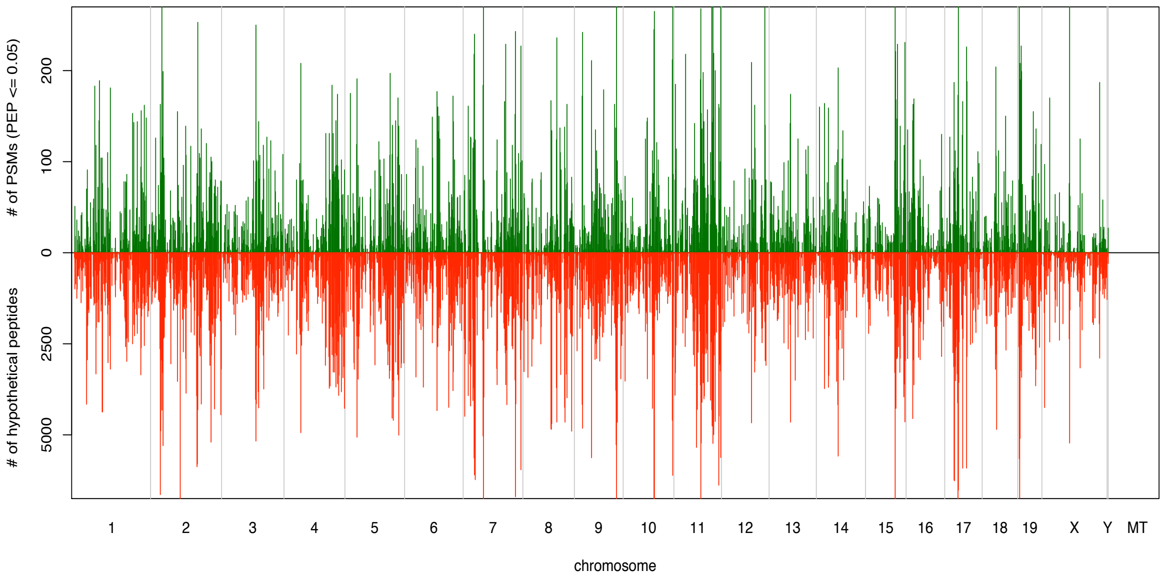
This meant that the chance of a wrong peptide identification would be 1% in the worst-case scenario, corresponding to a false discovery rate of less than 0.14%. Most proteogenomics research studies to date have used a false discovery rate of 1% to 5%, but I have taken a conservative approach to avoid the propagation of erroneous identifications into genome annotation pipelines.

5.3.2 Ensembl/Vega annotation validation

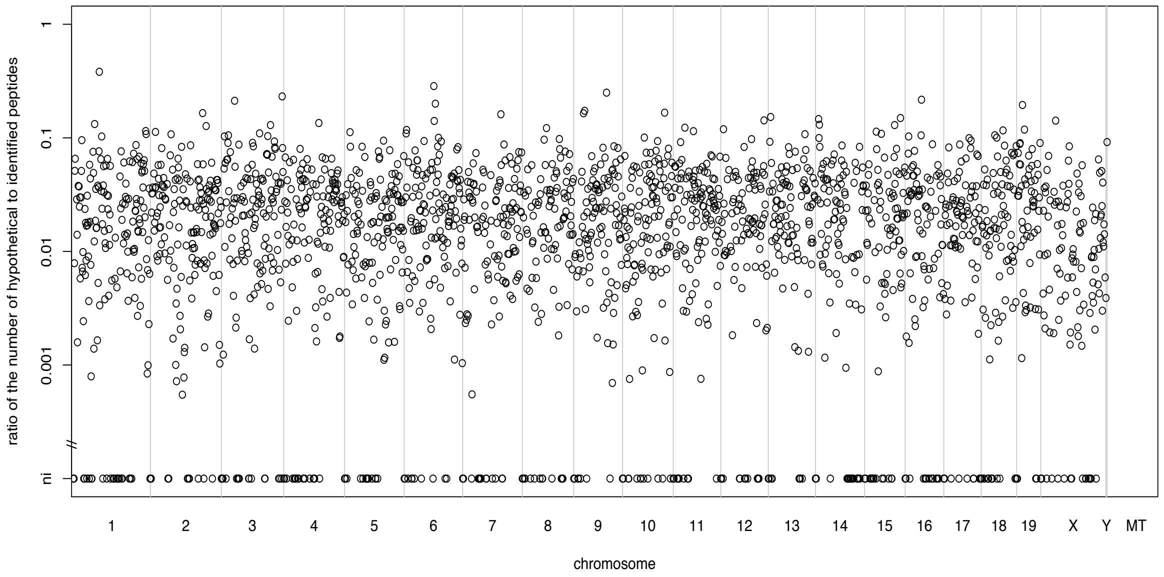
98.1% of all identified peptides ($PEP \leqslant 1\%$) matched the Ensembl/Vega database with only 1.9% attributed solely to IPI and Augustus (figure 5.1b). Therefore I focus first on confirming Ensembl/Vega annotation at the level of gene translation and structure.

5.3.2.1 Genome coverage

Figure 5.3a shows the distribution of fully tryptic peptides across the genome. Each chromosome was binned into 1Mb blocks and the number of potentially identifiable (all *in silico* digested peptides), as well as the number of identified peptides were calculated to evaluate genome coverage. Gene density varies across mouse chromosomes (Waterston *et al.*, 2002) and each gene contains peptides over a range of three orders of magnitude (see next section), the number of identifiable peptides per 1Mb block is also highly variable (nil to 12,910 peptides, median 715). The number of identified peptides (median 10) per 1Mb block is not only dependent on the number of identifiable peptides, but most notably on the expression level of the gene products, which determines the number of peptides that can be sampled by the MS instrument (Ishihama *et al.*, 2005; Lu *et al.*, 2007). The ratio between identified and identifiable peptides varied by more than two orders of magnitude (figure 5.3b). These results indicate that there was no mistake or bias in the data processing towards certain chromosomes and a more in-depth analysis can be conducted.



(a) Peptide counts of all potential identifiable peptides (red) and all peptides that have been identified (green) are plotted for each chromosome. Note the different y-axis scale.



(b) Relative peptide identification rate as defined by all peptides that have been identified versus all potential identifiable peptides.

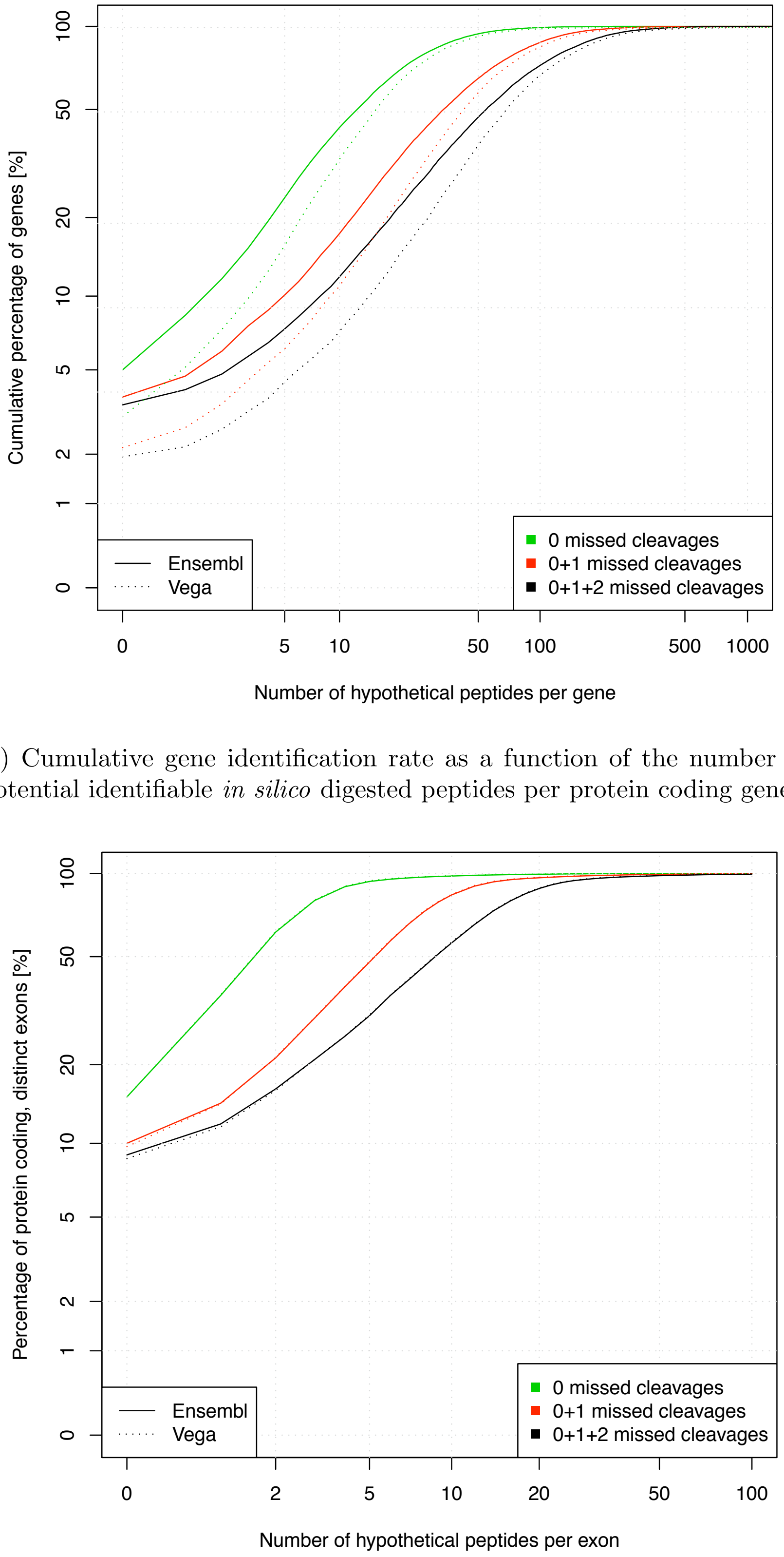
Figure 5.3: Chromosomes were binned into 1Mb blocks and absolute (a) and relative (b) peptide counts were evaluated, allowing the evaluation of peptide coverage at a genome scale.

5.3.2.2 Verification of gene translation

Figure 5.4a shows the cumulative percentage of genes that could be validated theoretically by tryptic peptides that map uniquely to a genomic locus and comprise between eight to 30 amino acids. These are the default peptide parameters for all theoretical considerations in the remainder of this work (peptide length distribution is illustrated in figure 5.2) and no predictions were made about proteotypic peptides (Fusaro *et al.*, 2009; Mallick *et al.*, 2007). Interestingly, when nil, one and two missed cleavages were allowed, 5.0%, 3.8% and 3.5% protein coding Ensembl gene products contain no tryptic peptides and 43.0%, 17.4% and 11.9% contain only ten or fewer peptides respectively. This could potentially limit the chances of gene validation, given that not all peptides are amenable for MS analysis (Fusaro *et al.*, 2009).

Nevertheless, a significant number of 7,221 (4,463) protein coding Ensembl (Vega) genes could be validated with peptide identifications that mapped uniquely to one gene, corresponding to 31.6% (36.7%) of all protein coding genes. However, peptide coverage was limited, with only 7.9% (9.0%) of the genes being validated by more than ten peptides and 0.08% (0.09%) by more than 100 peptides (figure 5.6a).

In order to further study the relationship between identified and potentially identifiable peptides, it was tested whether a linear model could be fitted (figure 5.5). A perfect fit would mean that the MS instrument would sample more peptides from gene products with more potential peptides. However, it was found that there is no correlation ($R^2 = 0.10$) and this is consistent with the above statement that peptide sampling is mainly determined by relative protein abundance. Moreover, genes that are only expressed in specific tissues would not be identified if the tissue of interest was not analyzed. For example, *Obscurin* (ENSMUSG00000061462) is a muscle protein and is amongst the top ten genes with most potentially identifiable peptides (1192) and yet none of the peptides were identified, suggesting that it was either expressed at very low abundance or not at all in the tissues or cell lines that were analyzed (see also `http://tinyurl.com/Obscurin`). In contrast, *Plectin-1*



(a) Cumulative gene identification rate as a function of the number of potential identifiable *in silico* digested peptides per protein coding gene.

(b) Cumulative exon identification rate as a function of the number of potential identifiable *in silico* digested peptides per protein coding exon.

Figure 5.4: Theoretical gene and exon validation rate. Note: considered peptides where fully tryptic, ranged from 8-30 residues and were unique to a genomic locus.

(ENSMUSG00000022565), a cytoskeletal protein that is more widely expressed, has a similar number of potential peptides (1447), but has the highest number of identified peptides (280). Other genes amongst the top five genes with highest peptide coverage (> 170 distinct peptides) include: *Spectrin alpha chain 2 (Spna2) brain* (ENSMUSG00000057738), *Bassoon (Bsn)* (ENSMUSG00000032589), *Cytoplasmic dynein 1 heavy chain 1 (Dync1h1)* (ENSMUSG00000018707) and *Spectrin alpha chain brain 1 (Spna1)* (ENSMUSG00000020315). All of these proteins are very large (275-533 kDa) and therefore smaller proteins at the same expression level would always result in lower peptide coverage.

It is important to note that the consideration of missed cleavages makes a significant difference. Allowing missed cleavages results in better gene coverage, which can be explained by the fact that peptides with missed cleavages tend to be longer. Trypsin is a very specific enzyme, but is not always 100% efficient. In fact, 31.7% and 9.9% of all identified peptides in this study have one and two missed tryptic cleavage sites respectively and only 58.4% have no missed cleavage sites, hence about 90% of the peptides have none or one missed tryptic cleavage site.

Overall I show that proteomics MS data is of significant value for confirming genes, some of which could be validated with extensive peptide coverage. Considering that currently eukaryotic proteomes are far from being saturated (de Godoy *et al.*, 2006), gene validation coverage of proteogenomics data will further increase as improved methods and instrumentation allow for deeper proteome sequencing, theoretically enabling validation of considerable portion of the genes.

5.3.2.3 Gene structure validation

A similar analysis at the exon level, using the same peptide properties as before, revealed that 15.1%, 10.0% and 9.0% of all Ensembl protein coding exons do not contain detectable peptides when nil, one or two missed cleavages are allowed, respectively. In addition, 93.6%, 47.8%, and 30.4% of the protein coding Ensembl

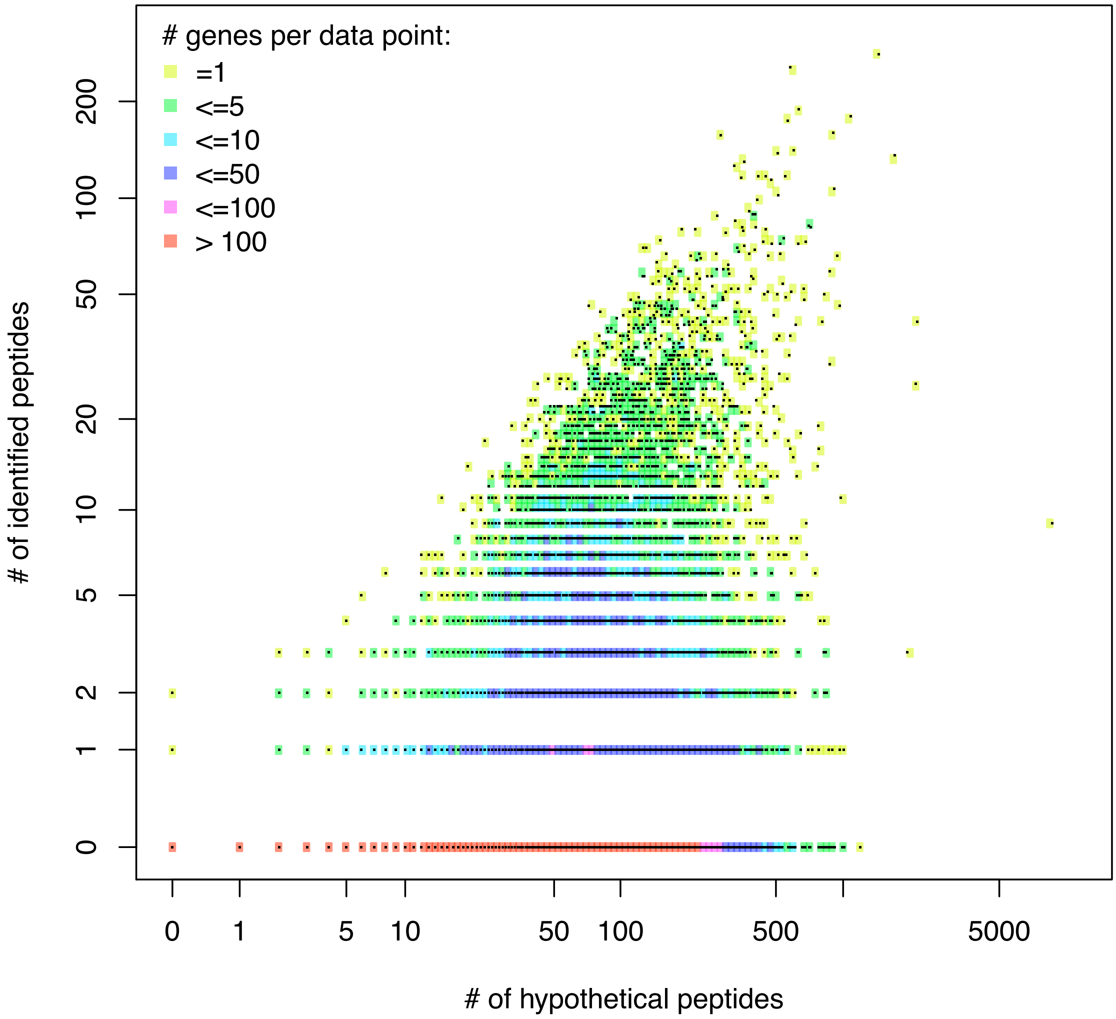
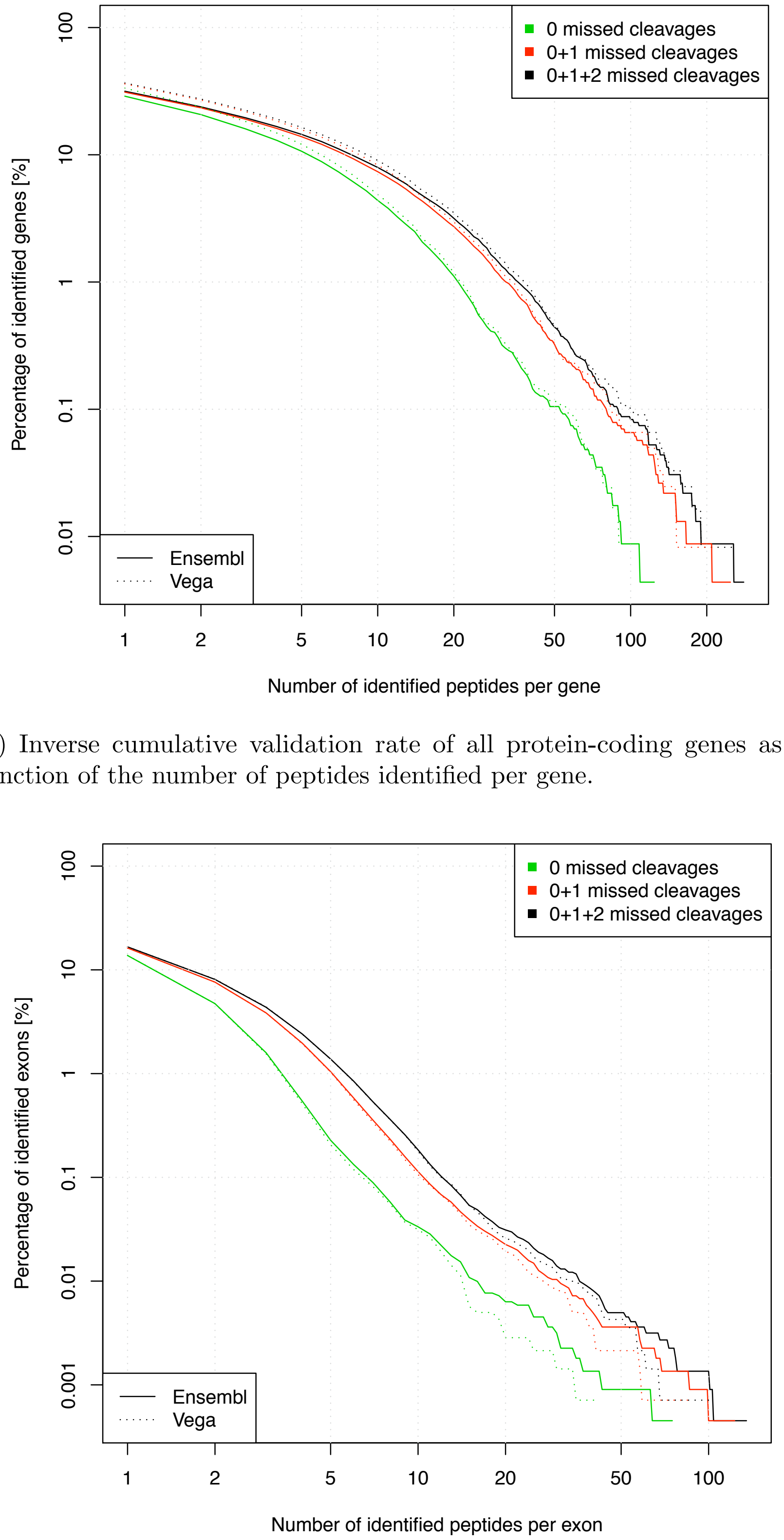


Figure 5.5: Correlation analysis between the number of identified peptides and the number of potential identifiable peptides per gene. Since many data points have the same x-y values, the number of overlaying data points (genes) is encoded with the color gradient (available from the legend).

exons contained five or fewer peptides, respectively (figure 5.4b). The lower peptide coverage compared with complete genes can be explained by the fact that the average protein coding exon count per gene in mouse is around 9.7.

A total of 16.7% of the total 222,378 Ensembl protein coding exons could be validated by peptide identifications. About 8.0% and 1.4% of Ensembl exons were validated by at least two and five peptides, respectively (figure 5.6b). Validation rates for Vega were insignificantly different.



(a) Inverse cumulative validation rate of all protein-coding genes as a function of the number of peptides identified per gene.

(b) Inverse cumulative validation rate of all protein-coding exons as a function of the number of peptides identified per exon.

Figure 5.6: Observed gene and exon validation rate using identified peptides.

A more difficult challenge is to validate annotation of introns, since this requires a fully tryptic and unique peptide spanning splice boundaries to be identified. Defining the accurate splice donor and acceptor sites is not trivial and a peptide spanning these sites not only validates them, but implicitly also validates the joined exons and thereby significantly contributes to gene structure validation.

Of the 202,205 (131,336) introns in Ensembl (Vega) that span a protein-coding splice boundary, up to 70.9% and 86.2% could theoretically be confirmed by peptides, allowing for one or two missed cleavages, respectively. However, when only peptides without missed cleavages are considered, the theoretical validation rate drops to 46%.

Using the subset of identified peptides that span a splice site, a total of 14,426 (9,347) Ensembl (Vega) introns could be confirmed, corresponding to 7.1% of all splice sites that join protein coding exons in both Ensembl and Vega, 1.3% of which were validated with two or more distinct peptides.

Clearly, the value of translational evidence is indispensable for independent gene structure validation. Notably up to 91.0% of all protein coding exons and 86.2% of all introns could theoretically be confirmed with peptides obtained in typical proteomics experiments. Applying the peptides identified in this study, 16.7% of all exons and 7.1% of all introns could be confirmed, highlighting that with relatively moderate efforts a significant proportion of gene structures can be validated.

5.3.2.4 Evidence of alternative translation

Until recently, only limited evidence was available of alternatively expressed transcripts at the protein level (Tress *et al.*, 2008). The detection of these variants by standard MS proteomics experiments is hindered by the fact that the majority of protein sequence is shared between transcripts, differing only in small parts of the translation products. Validation of alternative translation requires identification of at least one “signature” peptide for each protein isoform. 8,877 (40%) protein coding Ensembl genes code for alternative products, but only 16,664 transcripts from 1,542 genes could theoretically be discriminated by 168,726 “signature” peptides. For example, *Catenin delta-1* (ENSMUSG00000034101) has 25 alternative transcripts annotated as coding, but only nine “signature” peptides could theoretically distinguish the alternative translation of three protein isoforms.

Nevertheless, protein evidence for alternatively translated genes from tryptic digests was shown recently; Tanner *et al.* (2007) found evidence for 16 human genes, Castellana *et al.* (2008) found evidence for 47 Arabidopsis genes and Tress *et al.* (2008) identified 130 drosophila genes that express at least two protein isoforms. Here, a total of 370 peptides enabled discrimination of 112 transcripts in 53 genes, corresponding to 3.4% of all protein coding genes with multiple isoforms that can be discriminated by a peptide. *UDP-glucuronosyltransferase 1-2 Precursor* (ENSMUSG00000054545), which has 12 alternative coding transcripts within one locus, is unusual as all variants have an alternative 5’ exon spliced to a common set of downstream constant exons. These variable first exons confer diverse functional mRNAs with different tissue specific expression profiles (Zhang *et al.*, 2004). Figure 5.7 shows an overview of this complex locus with evidence for expression of five alternative protein isoforms from 27 “signature” peptides. Other examples with evidence for three alternative gene products include: *ankyrin 2 brain isoform 2* (ENSMUSG00000032826), *Synaptotagmin-7* (ENSMUSG00000024743) and *Core histone macro-H2A*.1 (ENSMUSG00000015937). Two alternative isoforms were validated for each of the remaining 49 genes.

Even though the overall rate of peptide identifications that could be attributed to alternative protein isoforms is low in proteogenomic studies due to only few available “signature” peptides that are unique to one isoform, these results demonstrate evidence for the presence of alternative splice variants *in vivo*. It would be interesting to follow-up this study with a more sensitive hypothesis driven targeted proteomics approach (Anderson *et al.*, 2009; Arnott *et al.*, 2002), in which the mass spectrometer is directed to scan specifically for “signature” peptides of individual protein isoforms.

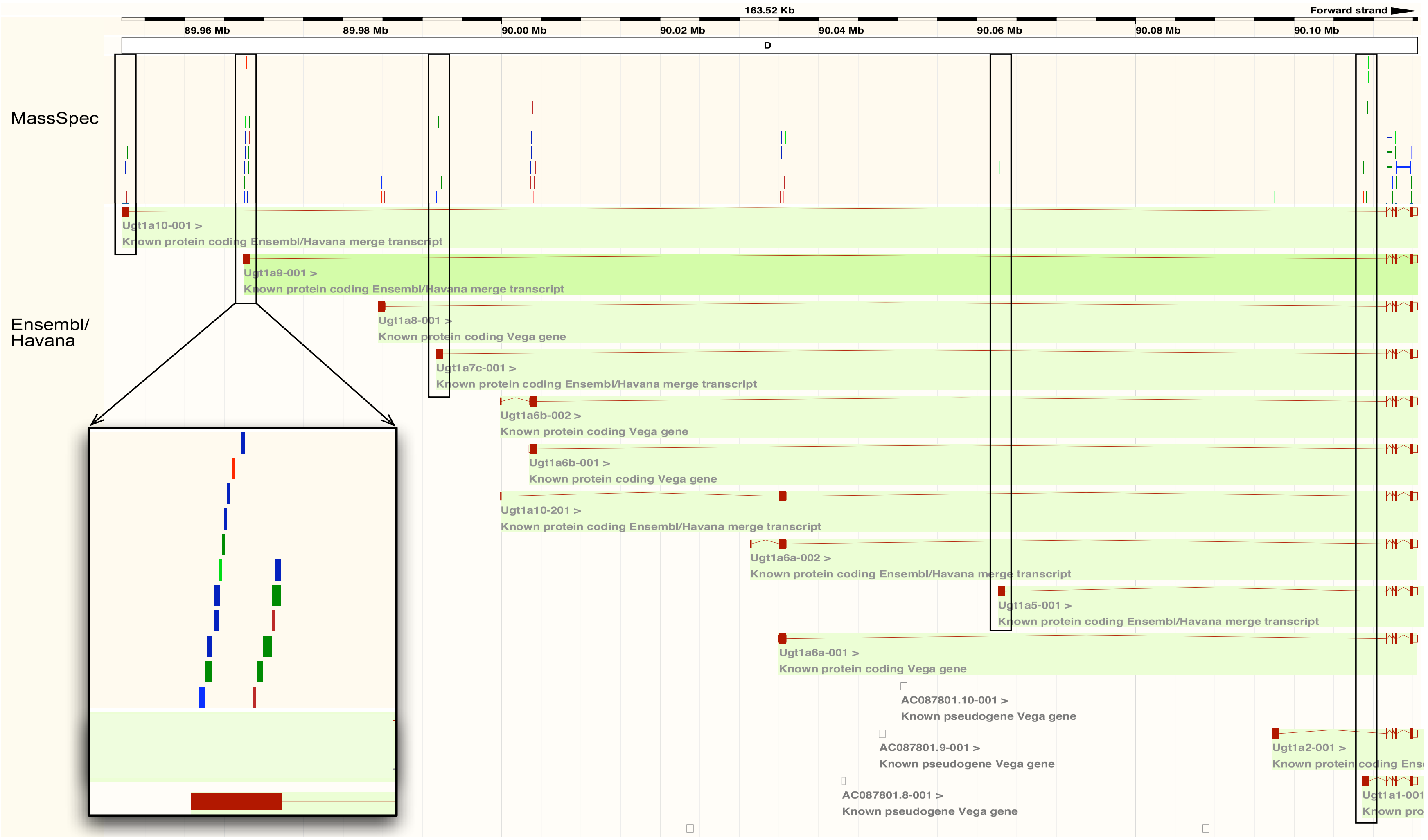


Figure 5.7: Peptide evidence that is specific to five alternative gene products of *UGT1A2*. Each isoform was identified with a set of signature peptides that were unique to only one variant (highlighted in figure, one example magnified). This assumes complete annotation since novel gene variants may have shared peptides with existing gene variants. Peptide colour codes, see figure 4.4.

5.3.2.5 Nonsense mediated decay

Nonsense mediated decay (NMD) is a translational-coupled mechanism that eliminates mRNAs containing premature translation-termination codons (PTCs) (Brogna and Wen, 2009) and is estimated to effect 75-90% of human genes (McGlincy and Smith, 2008). The exact mechanism of how NMD occurs in mammals is still under debate (Brogna and Wen, 2009). Some known proteins e.g. NRAS have transcripts that appear to escape the NMD since they contain a PTC but still a functional protein appears to be produced. Since the Vega database contains annotation of transcripts predicted to be subject to NMD, I used the MS data to test whether any of the NMD transcripts actually produced a detectable translated protein. The search database contained 2,690 NMD transcripts, which would allow identification. However, only 1,704 NMD transcripts could theoretically be validated by 9,202 potential "signature" peptides. Interestingly, I have not been able to identify any "signature" peptides that would suggest the translation of NMD transcripts. Using Fisher's exact test, this result is significantly different (p-value 5×10^{-9}) from what would be expected by chance (20 peptides) when compared with the conservative peptide identification rate that could be attributed to alternative transcripts. This reinforces the theory that transcripts flagged with NMD indeed undergo degradation with a short half life. On the other hand, these proteins may not be expressed at all or at a very low level, hindering detection by the MS instrument.

5.3.3 Gene model correction

Peptide identifications are also of great value for correcting gene structures, only limited by the fact that the protein sequence needs to be in the search database to be identified in the first place. For this reason, the search database was supplemented with Augustus predictions, containing about ten-fold the number of peptides compared to Ensembl (see methods and figure 5.1a). Moreover, peptides derived from

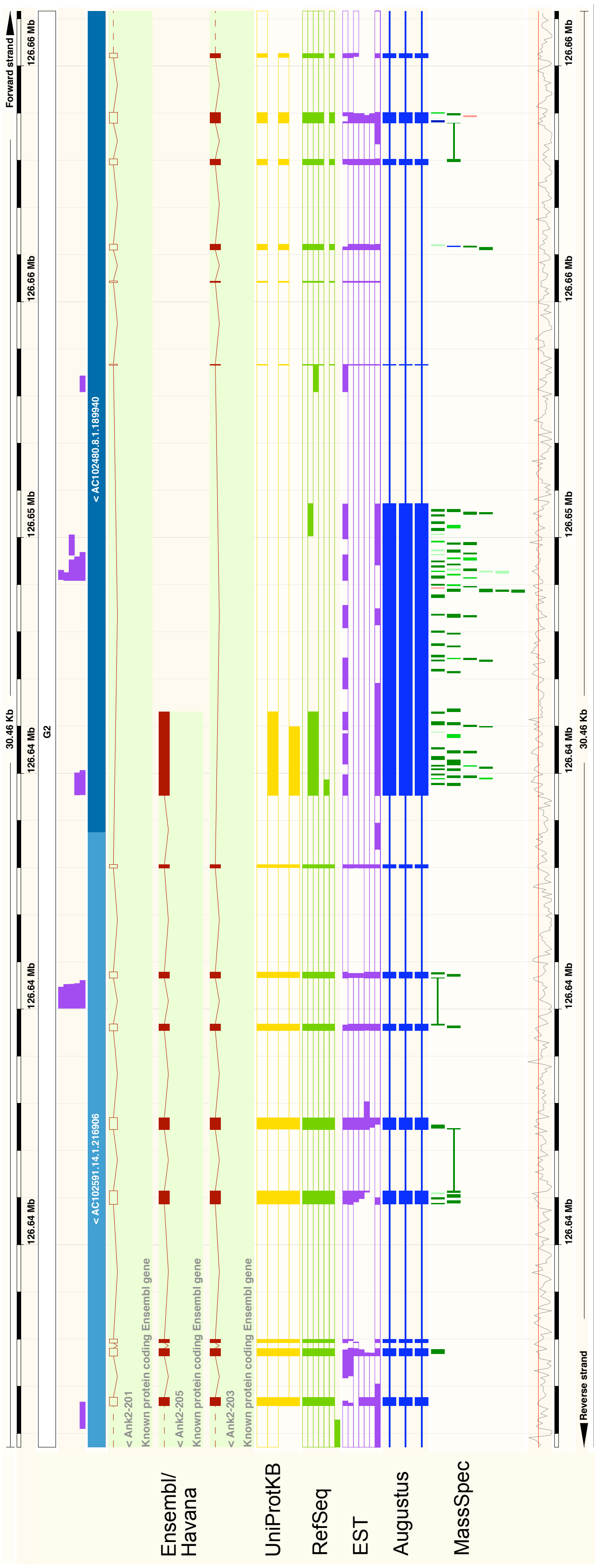


Figure 5.8: Gene *Ankyrin 2* (brain isoform 2) is annotated in Ensembl with three alternative transcripts (red). The Augustus runs predicted a large exon, not present in Ensembl/Havana, that was validated by multiple peptide identifications (“MassSpec” track), suggesting either incorrect Ensembl annotation or a new alternative isoform. This was further supported by transcriptional evidence (RefSeq and EST track).

the IPI database could be indicative of differences between the genomic annotation and the protein database. Therefore the superset of Ensembl, Vega, Augustus, IPI and the cRAP contaminants database was used to search the MS data with Mascot for a second round for extended analysis.

1.9% of all peptide identifications matched neither Ensembl nor Vega, but were present in the IPI database or Augustus gene predictions, indicating a significant number of identifications that contribute to gene structure refinements or novel genes (figure 5.1b). These identifications do not fall into the expected number of incorrect identifications (0.12% false discovery rate at the chosen 1% PEP threshold) and were therefore further investigated.

5.3.3.1 Gene model refinements

Predicting the correct gene structure and defining the exact donor and acceptor splice site remains one of the most difficult problems in genome annotation. Using peptide data that was searched against the Augustus database, a total of 168 intron refinements could be made, which include the correction of splice donor and acceptor sites, skipping exons, as well as the introduction or refinement of novel exons.

Figure 5.8 shows one example where Augustus predicted an exon extension that was not annotated in any of the Ensembl/Vega transcripts but was validated by 52 distinct peptide identifications. This clearly suggests that either the existing annotation was incorrect or a novel isoform was found. This example demonstrates the power of searching tandem-MS data against an over-predicted genome to detected flaws or missing annotation.

Refinements identified in this work will be manually investigated by the HAVANA team in-house and future Vega releases will have validated refinements incorporated.

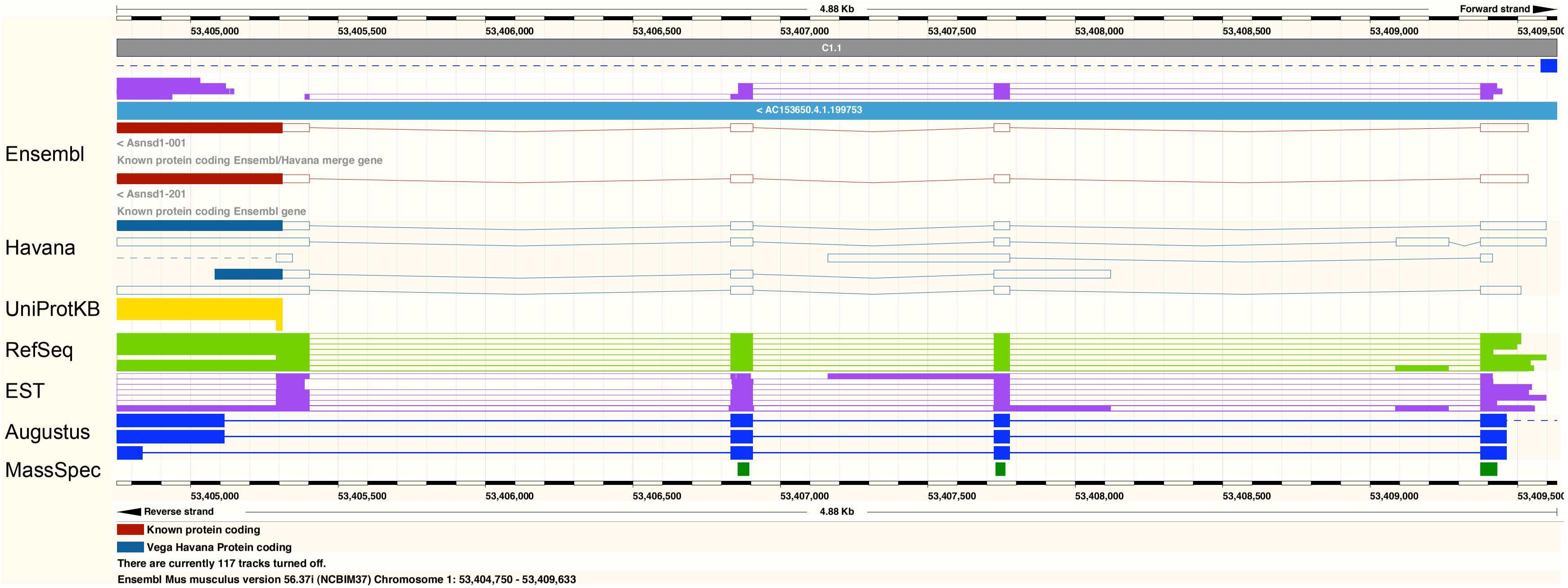


Figure 5.9: Three exons, part of the annotated 5' UTR of gene *Asnsd1* (ENSMUSG00000026095), were confirmed as coding by high confident peptide identifications ($PEP < 10^{-4.5}$), which are indicated in the "MassSpec" track. EST and cDNA evidence (EST and RefSeq track) support these exons, however, the existing protein evidence (UniProtKB track) suggests a translational start site downstream of these identified peptides, which may have led to the existing Ensembl/Havana annotation.

5.3.3.2 Translational evidence for annotated non-coding regions

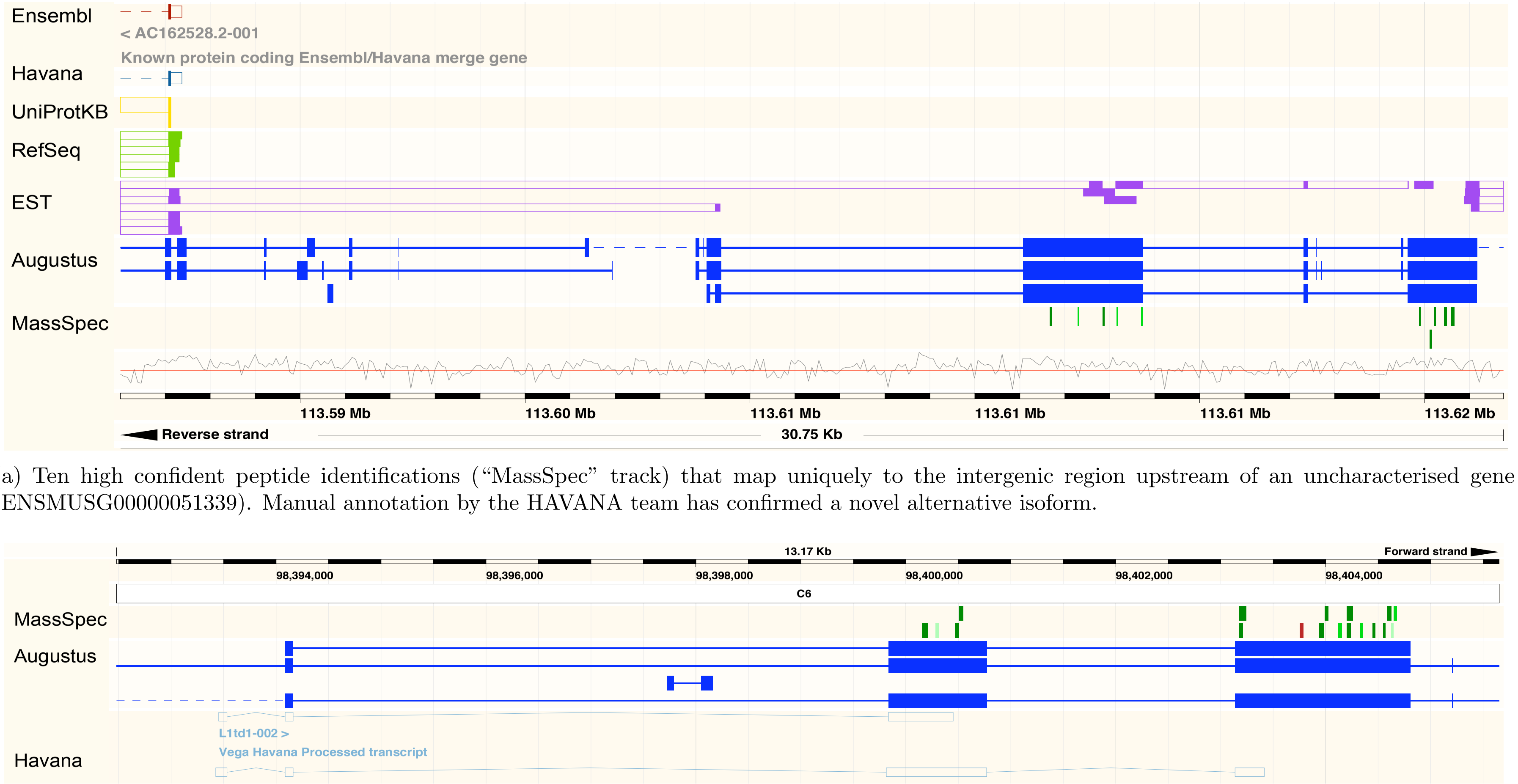
The accurate identification of the UTR and protein coding regions is another challenge in genome annotation. For example, cDNA sequences are often truncated and protein sequences from protein databases are not validated by mass spectrometry, which can lead to wrongly annotated UTRs or protein coding regions.

Data of this pilot study revealed translational evidence either within the UTR or in adjacent intergenic regions for 101 genes, suggesting incorrectly defined coding or gene boundaries. Of the 39 genes that were manually investigated, 85% had additional peptides matching upstream the 5' end.

Figure 5.9 illustrates one example where peptide identifications map uniquely to three exons in the 5' UTR of Ensembl/Vega gene *Asparagine synthetase domain-containing protein 1* (ENSMUSG00000026095), suggesting that either the UTR is incorrectly annotated or an alternative protein isoform exists.

Another example is illustrated in figure 5.10a, where ten peptides map to the intergenic region upstream of an uncharacterised gene (ENSMUSG00000051339). Gary Sounders from the HAVANA team investigated this region manually and built an *ab initio* gene model, which was supported by EST evidence and the ten identified peptides. EST *Em:BY593944.1* fused this novel upstream region with the existing annotation of ENSMUSG00000051339. The translation of an orthologous gene in human showed extensive sequence conservation, further supporting this novel variant.

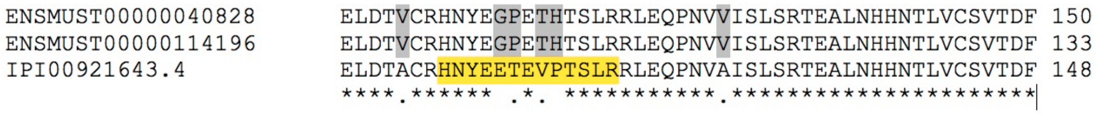
Moreover, pseudo and processed genes in Vega were predicted by Augustus to be protein coding. Strikingly, for 55 of these, translational evidence in the form of peptide identifications was found, suggesting incorrect Ensembl/Vega annotation. Figure 5.10b shows one example where gene *LINE-1 type transposase* (OTTMUST00000019654) was annotated as processed, but a significant number of peptide identifications clearly demonstrated translation. Similar proteogenomic findings of translated pseudogenes were recently demonstrated by Castellana *et al.* (2008) in *Arabidopsis thaliana* and by Merrihew *et al.* (2008) in *C. elegans.*



(a) Ten high confident peptide identifications ("MassSpec" track) that map uniquely to the intergenic region upstream of an uncharacterised gene (ENSMUSG00000051339). Manual annotation by the HAVANA team has confirmed a novel alternative isoform.

(b) The Vega gene OTTMUST00000019654 *(LINE-1 type transposase)* was annotated as a processed transcript without translation ("Havana" track). However, extensive peptide evidence ("MassSpec" track) that is unique to the locus was found, suggesting that there is at least one form of the gene that is translated. The blue tracks represent collapsed Augustus transcripts.

Figure 5.10: Examples of translational peptide evidence in annotated non-coding regions.



```
ENSMUST00000040828      ELDTVCRHNYEGPETHTSLRRLEQPNVVISLSRTEALNHHNTLVCSVTDF 150
ENSMUST00000114196      ELDTVCRHNYEGPETHTSLRRLEQPNVVISLSRTEALNHHNTLVCSVTDF 133
IPI00921643.4           ELDTACRHNYEETEVPTSLRRLEQPNVAISLSRTEALNHHNTLVCSVTDF 148
                        ****.****** .*. ***********.**********************
```

Figure 5.11: One example where the peptide sequence in Ensembl was different to the sequence in IPI (highlighted with grey). The identified peptide (indicated in yellow, $PEP = 1 \times 10^{-4}$) matches perfectly the IPI database but differs in four residues to the Ensembl translations.

5.3.3.3 Protein database derived peptide matches

Genotyping projects over the last number of years have populated large SNP databases, but although these are large, they are yet far from being complete, especially for the mouse genome. Insertion of SNPs into protein databases inflates the search space significantly since multiple variants of one peptide need to be searched, thereby reducing identification sensitivity. In a recent study, Tanner *et al.* (2007) searched a corpus of 18.5 million MS spectra (human) against a database incorporating coding SNPs, resulting in 1.2 million peptide identifications with only 0.02% (308) validated coding SNPs. I have decided to not include coding SNPs into the search databases, but currently there are large scale mouse sequencing efforts underway, potentially allowing strain specific search databases to be built in the future. Evaluation of their performance over a generic species-specific protein databases will be interesting to study.

However, since the search database included IPI protein sequences, some of which were not derived from genomic but from mRNA sequences, differences between IPI and the Ensembl/Vega protein sequences could be detected. 19 IPI proteins with peptide sequences not matching the reference genome were identified. In five cases, the sequence differences were caused by indels, with the remaining 14 cases caused by coding SNPs. Figure 5.11 shows one example where a peptide match was made against the IPI protein sequence, but the Ensembl/Vega reference sequence is different in four residues. This indicates that either the genome reference sequence was incorrect or four novel coding SNPs exist in this peptide.

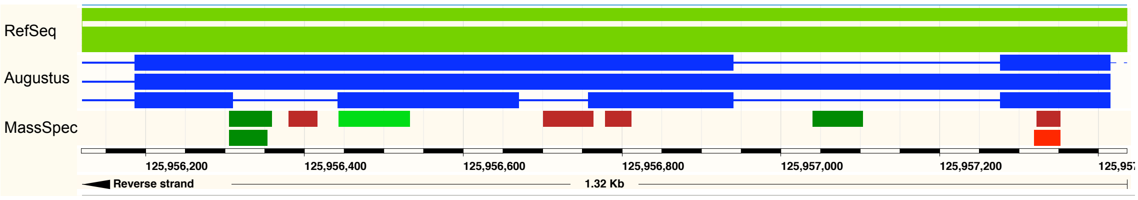


Figure 5.12: Four peptides match uniquely (green) to an intergenic region (“MassSpec” track). Together with full-length mouse cDNA evidence (”RefSeq” track) these data suggests a novel protein coding region.

5.3.3.4 Novel genes

Peptides matching to intergenic regions are of great interest to further complement the list of coding genes. The Ensembl genome annotation process is conservative (see section 1.2.5) and proteogenomic methods are ideally placed to identify such missing genes. The caveat is that the gene of interest must be present in the search database of Mascot to enable identification. As discussed above, the gene finding algorithm Augustus was employed to over-predict the genome. Peptides derived from these predictions and existing annotated protein coding sequences were used as a search database. Assuming that the Ensembl gene list is close to complete, the Augustus predictions contain 90% random sequence (figure 5.1a). Therefore, reliable and stringent peptide scoring together with subsequent filtering to exclude ambiguous matches are crucial to minimize and ultimately to exclude any false positive identifications. To reiterate, the worst peptide match considered in this study had a 1% probability to be incorrect, corresponding to a false discovery rate of less than 0.14%. For subsequent analysis, where peptides were not supported by any existing annotation, this was further constrained in that at least two peptides (one of which with a PEP of at least 0.01, the second of at least 0.05) had to be identified.

Using this approach, I propose 29 novel genes, supported by a total of 70 peptides. However, 12 of these genes have overlapping identifications with IPI protein entries, suggesting that the Ensembl/Vega annotation process missed these genes. The remaining 17 novel genes do not overlap with any known Ensembl/Vega genes or

IPI entries, but six are potentially an extension of known Ensembl/Vega genes, four at the 5' and two at the 3' region. For nine of these genes, Pfam-A domains (Sonnhammer *et al.*, 1997) could be detected with high significance, and based on this it is likely that most of these genes are RNA/DNA binding proteins. Some cDNA or EST evidence was observed for 50% of these novel regions, but vertebrate conservation was generally absent for all of these predictions (figure 5.12).

I have not further investigated an additional set of 50 potential new protein coding genes that were supported by only one peptide. Nevertheless, these peptides are strong matches, with PEP values ranging from 4×10^{-13} to 1×10^{-3}. Even though I am hesitant to identify these regions as novel genes only based on peptide identifications, these predictions together with the proposed new genes are available as a DAS annotation track for the HAVANA annotation team who currently investigate these cases manually and potentially demand additional experimental evidence to complete annotation in Vega.

5.4 Conclusion

Mass spectrometry has become the method of choice to identify peptides and infer proteins in a high-throughput manner and it is therefore a consequent development to incorporate these data into genome annotation pipelines as translational evidence. I have shown that, theoretically, peptide evidence could validate up to 96.5% of all protein coding genes, 91.0% of all protein coding exons and 86.2% of all exon-exon junctions.

However, the mouse proteome is far from being saturated by MS based peptide identifications. Even if every organ with all its regions, cell types and organelles could be isolated and analyzed, there would probably be a significant set of genes that would be missed because expression of these may be only activated under specific and transient cellular activation. There have not been systematic analyses at these levels of complexity, but if I compare studies from ten years ago with the latest achievements, it is clear that MS data becomes richer and more valuable for genome annotation every day.

Using the proteomics datasets readily available for this study, comprising about 10 million spectra, I could validate 31.6% of all protein coding genes, 16.7% of all protein coding exons and 7.1% of all exon-exon junctions in Ensembl, with similar numbers in Vega, significantly contributing to the validation of the mostly automatically annotated mouse genome. Interestingly, for 53 genes I have shown evidence of expression of alternatively spliced isoforms, yet I have also shown that MS data is not always sufficient to fully validate protein isoforms, since many share coding sequence and do not always allow the variants to be distinguished.

It is still not clear of whether transcripts that are flagged to undergo nonsense mediated decay could be translated into stable proteins. I have not detected a single peptide that was unique to NMD transcripts. This could be interpreted in two ways: either these transcripts indeed undergo degradation and cannot be detected or they are translated at very low abundance and were not sampled by the MS instrument for this reason.

Beyond validation, peptide identifications contributed to the identification of potential incorrect annotation. 129 regions were attributed to donor or acceptor splice site refinements or the introduction of novel exons, 101 genomic loci were identified that mapped outside the coding region of genes (mostly at the 5' region) and 55 pseudo- and processed genes were found to be coding. Lastly, I propose 29 protein coding genes, 12 of which are already present in IPI but not in Ensembl/Vega and 6 cases could be coding extensions of known genes.

Overall I have highlighted the possibilities and the limitations of the use of "bottom-up" proteomics for genome annotation and demonstrated the use of available MS data for incorporation into automatic genome annotation pipelines such as Ensembl as an additional layer of evidence.

Chapter 6

Concluding remarks

Despite significant efforts in annotating complex genomes such as mouse or human, accurate identification and structural elucidation of protein coding genes remains challenging. Current high-throughput and manually driven annotation methods rely largely on computational predictions and transcriptional evidence, such as full-length cDNA data. However, a lack of protein-level evidence leaves translation unverified in most cases.

Proteomic mass spectrometry (MS) is the method of choice for sequencing gene product fragments. This enables the validation of translation, the refinement of existing gene annotation, and the identification of novel protein coding regions. However, high-throughput application of proteomics data to genome annotation is hindered by the lack of suitable tools and methods to achieve automatic data processing and genome mapping at high accuracy and throughput. The work presented in this thesis attempts to address some of these issues.

The outcome of every proteomics MS/MS experiment is dependent on the reliability, sensitivity and specificity of the peptide identification procedure. This also underpins any proteogenomic analysis where proteomics data is applied to the field of genome annotation; incorrect peptide identifications would be propagated leading to incorrect annotation, which would subsequently be trusted incorrectly.

Therefore I initially evaluated the peptide identification software "Mascot" that is routinely used at the Wellcome Trust Sanger Institute and elsewhere as described in chapter 2. I have shown that the default Mascot scoring scheme deviates significantly from the expected error rates, due to sensitivity and specificity being correlated with search space. Counter intuitively the error rate was found to increase as the search space decreases. This is of significance when high accuracy MS instruments are used for proteomics experiments; here the search space can be orders of magnitude smaller than with standard instruments due to the afforded high mass accuracy. As a solution I proposed a novel "Adjusted Mascot Threshold" (AMT) that is based on false discovery rate estimates (Brosch *et al.*, 2008). AMT utilises the mass accuracy of recent state-of-the-art instruments by using peptide mass filtering as a first discriminator, which leverages the improved sensitivity of the method.

The limitation of this approach was that discrimination is solely based on mass accuracy and the adjusted score threshold. In the light of potentially large search databases used for detecting novel genes, it was desirable to further complement this approach with orthogonal scoring features that would aid discrimination between correct and incorrect peptide spectrum matches. This was achieved by utilising the machine learning algorithm "Percolator" (Käll *et al.*, 2007), as discussed in chapter 3. Percolator provided the framework to extend my AMT scoring scheme with a large set of scoring features, which led to the development of "Mascot Percolator" (MP). I showed that MP is the most sensitive Mascot scoring scheme available, providing reliable and robust significance measures, validated against standard protein datasets (Brosch *et al.*, 2009). MP is available as a standalone software package that can be run on top of any Mascot search where target/decoy searching is amenable. Moreover, the method is currently implemented into the official Mascot 2.3 release (`http://www.matrixscience.com/workshop_2009.html`), which will distribute MP to a large proteomics community. This system provides good sensitivity, an advanced notion of the global false discovery rate, and a peptide level scoring statistics (posterior error probability) that are calculated for each peptide spectrum match. This is important when peptide identifications are used for genome annotation; a probability measure can be attributed to each genome annotation that is based on a peptide identification. In future work Mascot Percolator could be extended to alternative fragmentation methods and alternative scoring features could be explored. I am confident that the widespread use of this method will increase research interest in the field of peptide scoring.

In chapter 4 I developed a genome annotation pipeline that closes the gap between high throughput peptide identification and scoring, as provided with Mascot and Mascot Percolator, and large scale genome annotation analysis. Most proteogenomics studies map peptides by alignment tools onto the genome, I presented a rather different approach, whereby the peptide-genome mapping is computed by utilising the application programming interfaces of the Ensembl pipeline. These mappings are stored in a comprehensive database which enables efficient and ad-hoc mapping of identified and predicted peptides to their genomic loci, each of which is associated with supplemental annotation information such as gene and transcript identifiers. The comprehensive database facilitates the export of compact non-redundant peptide level databases that can be used as Mascot search databases allowing for best possible performance. Considering the increased acceptance of targeted proteomic strategies in the proteomics community, it should be noted that the peptide export could facilitate these novel approaches by generating lists of signature peptides for individual genes or gene isoforms. The database enables the generation of automated genome annotation analysis reports and provides the data basis for a distributed annotation server (DAS) that can be integrated into existing genome annotation projects.

This proteogenomics pipeline was applied in a pilot study using a large mouse MS dataset in chapter 5. I showed where peptide identifications facilitated the validation and correction of existing annotation, such as re-defining the translated regions or splice boundaries. I also proposed a set of novel genes that were identified by the MS analysis pipeline with high confidence. Moreover, I demonstrated for the first time the value and level of coverage that can be achieved with proteogenomic analysis for validating genes and gene structures, while also highlighting the theoretical limitations of this technique. This was possible since for every *in silico* generated peptide the genomic mapping was readily available through the proteogenomics database. Detailed manual investigation of the refined and novel regions that were identified by MS are currently investigated by the HAVANA team at the Wellcome Trust Sanger Institute. Overall this study demonstrated the value of utilising proteomics data for genome annotation and it may be an interesting future direction to extend automated annotation pipelines such as Ensembl to complement cDNA evidence with high confident proteomics data.

Scaling up this pilot study to improve coverage should be an easy undertaking, only limited by available proteomics data. Nevertheless, the theoretical genome validation coverage, which was discussed in chapter 5, will be hard to achieve with current MS proteomics methods. The trade-off between sensitivity, dynamic range and throughput underpins current shotgun proteomics approaches. In addition, it is a significant challenge to analyse the complete proteome that covers every organ with all its regions, all cell types and organelles in various states. However, the incremental methodological and technological advancements have led to significant improvements in MS proteomics over the last two decades and with the ever increasing need for high performing proteomics applications, this trend is likely to continue.

References

Adams, M., *et al.* (1991), Complementary dna sequencing: expressed sequence tags and human genome project., *Science*, *252*(5013), 1651–1656. (page 24)

Aebersold, R., and M. Mann (2003), Mass spectrometry-based proteomics., *Nature*, *422*(6928), 198–207. (page 2, 3)

Alexandersson, M., S. Cawley, and L. Pachter (2003), Slam: cross-species gene finding and alignment with a generalized pair hidden markov model., *Genome Res*, *13*(3), 496–502. (page 25)

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman (1990), Basic local alignment search tool., *J Mol Biol*, *215*(3), 403–410. (page 30)

Anderson, D., W. Li, D. Payan, and W. Noble (2003), A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide ms/ms spectra and sequest scores., *J Proteome Res*, *2*(2), 137–146. (page 34)

Anderson, N., N. Anderson, T. Pearson, C. Borchers, A. Paulovich, S. Patterson, M. Gillette, R. Aebersold, and S. Carr (2009), A human proteome detection and quantitation project., *Mol Cell Proteomics*, *8*(5), 883–886. (page 115)

Arnott, D., A. Kishiyama, E. Luis, S. Ludlum, J. J. Marsters, and J. Stults (2002), Selective detection of membrane proteins without antibodies: a mass spectrometric version of the western blot., *Mol Cell Proteomics*, *1*(2), 148–156. (page 115)

Ashurst, J., *et al.* (2005), The vertebrate genome annotation (vega) database., *Nucleic Acids Res*, *33*(Database issue), D459–65. (page 28, 84)

Bairoch, A., and R. Apweiler (1997), The swiss-prot protein sequence data bank and its supplement trembl, *Nucleic Acids Research*, *25*(1), 31. (page 28)

Beausoleil, S., J. Villen, S. Gerber, J. Rush, and S. Gygi (2006), A probability-based approach for high-throughput protein phosphorylation analysis and site localization., *Nat Biotechnol*, *24*(10), 1285–1292. (page 37)

Ben-Hur, A., C. Ong, S. Sonnenburg, B. Scholkopf, and G. Ratsch (2008), Support vector machines and kernels for computational biology., *PLoS Comput Biol*, *4*(10), e1000,173. (page 19)

Benjamini, Y., and Y. Hochberg (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc., Ser B*, *57*(1), 289–300. (page 10)

Bern, M., Y. Cai, and D. Goldberg (2007), Lookup peaks: A hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry., *Anal Chem.* (page 5)

Bianco, L., J. Mead, and C. Bessant (2009), Comparison of novel decoy database designs for optimizing protein identification searches using abrf sprg2006 standard ms/ms datasets., *J Proteome Res.* (page 16)

Biemann, K. (1988), Contributions of mass spectrometry to peptide and protein structure., *Biomed Environ Mass Spectrom*, *16*(1-12), 99–111. (page 2, 37, 105)

Birney, E., and R. Durbin (1997), Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison., *Proc Int Conf Intell Syst Mol Biol*, *5*, 56–64. (page 24)

Birney, E., M. Clamp, and R. Durbin (2004), Genewise and genomewise., *Genome Res*, *14*(5), 988–995. (page 24)

Birney, E., *et al.* (2007), Identification and analysis of functional elements in 1% of the human genome by the encode pilot project., *Nature*, *447*(7146), 799–816. (page 23, 26, 29)

Bonferroni, C. (1935), Il calcolo delle assicurazioni su gruppi di teste, *Studi in Onore del Professore Salvatore Ortu Carboni*, *13*. (page 10)

Brent, M. (2005), Genome annotation past, present, and future: how to define an orf at each locus., *Genome Res*, *15*(12), 1777–1786. (page 23)

Brogna, S., and J. Wen (2009), Nonsense-mediated mrna decay (nmd) mechanisms., *Nat Struct Mol Biol*, *16*(2), 107–113. (page 117)

Brosch, M., S. Swamy, T. Hubbard, and J. Choudhary (2008), Comparison of mascot and x!tandem performance for low and high accuracy mass spectrometry and the development of an adjusted mascot threshold., *Mol Cell Proteomics*, *7*(5), 962–970. (page 39, 65, 76, 128)

Brosch, M., L. Yu, T. Hubbard, and J. Choudhary (2009), Accurate and sensitive peptide identification with mascot percolator., *J Proteome Res*, *8*(6), 3176–3181. (page 64, 128)

Brunner, E., *et al.* (2007), A high-quality catalog of the drosophila melanogaster proteome., *Nat Biotechnol*, *25*(5), 576–583. (page 31)

Bunger, M., B. Cargile, J. Sevinsky, E. Deyanova, N. Yates, R. Hendrickson, and J. J. Stephenson (2007), Detection and validation of non-synonymous coding snps from orthogonal analysis of shotgun proteomics data., *J Proteome Res*, *6*(6), 2331–2340. (page 31)

Burge, C., and S. Karlin (1997), Prediction of complete gene structures in human genomic dna., *J Mol Biol*, *268*(1), 78–94. (page 25)

Carninci, P., *et al.* (2005), The transcriptional landscape of the mammalian genome., *Science*, *309*(5740), 1559–1563. (page 21)

Castellana, N., S. Payne, Z. Shen, M. Stanke, V. Bafna, and S. Briggs (2008), Discovery and revision of arabidopsis genes by proteogenomics., *Proc Natl Acad Sci U S A*, *105*(52), 21,034–21,038. (page 31, 115, 121)

Chepanoske, C., B. Richardson, M. von Rechenberg, and J. Peltier (2005), Average peptide score: a useful parameter for identification of proteins derived from database searches of liquid chromatography/tandem mass spectrometry data., *Rapid Commun Mass Spectrom*, *19*(1), 9–14. (page 30)

Chernushevich, I., A. Loboda, and B. Thomson (2001), An introduction to quadrupole-time-of-flight mass spectrometry., *J Mass Spectrom*, *36*(8), 849–865. (page 4)

Choi, H., and A. Nesvizhskii (2008), Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics., *J Proteome Res*, *7*(1), 254–265. (page 18)

Choi, H., D. Ghosh, and A. Nesvizhskii (2008), Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling., *J Proteome Res*, *7*(1), 286–292. (page 18)

Choudhary, J., W. Blackstock, D. Creasy, and J. Cottrell (2001a), Matching peptide mass spectra to est and genomic dna databases., *Trends Biotechnol*, *19*(10 Suppl), S17–22. (page 29)

Choudhary, J., W. Blackstock, D. Creasy, and J. Cottrell (2001b), Interrogating the human genome using uninterpreted mass spectrometry data., *Proteomics*, *1*(5), 651–667. (page 29)

Clamp, M., B. Fry, M. Kamal, X. Xie, J. Cuff, M. Lin, M. Kellis, K. Lindblad-Toh, and E. Lander (2007), Distinguishing protein-coding and noncoding genes in the human genome., *Proc Natl Acad Sci U S A*. (page 21, 26)

Clauser, K., P. Baker, and A. Burlingame (1999), Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing ms or ms/ms and database searching., *Anal Chem*, *71*(14), 2871–2882. (page 4)

Claverie, J. (2005), Fewer genes, more noncoding rna., *Science*, *309*(5740), 1529–1530. (page 21, 23, 25, 29)

Colinge, J., A. Masselot, M. Giron, T. Dessingy, and J. Magnin (2003), Olav: towards high-throughput tandem mass spectrometry data identification., *Proteomics*, *3*(8), 1454–1463. (page 14)

Craig, R., and R. Beavis (2004), Tandem: matching proteins with tandem mass spectra., *Bioinformatics*, *20*(9), 1466–1467. (page 7, 34, 36)

Crick, F. (1958), On protein synthesis., *Symp Soc Exp Biol*, *12*, 138–163. (page 22)

Crick, F. (1970), Central dogma of molecular biology., *Nature*, *227*(5258), 561–563. (page 22)

Curwen, V., E. Eyras, T. Andrews, L. Clarke, E. Mongin, S. Searle, and M. Clamp (2004), The ensembl automatic gene annotation system., *Genome Res*, *14*(5), 942–950. (page 25)

Dancik, V., T. Addona, K. Clauser, J. Vath, and P. Pevzner (1999), De novo peptide sequencing via tandem mass spectrometry., *J Comput Biol*, *6*(3-4), 327–342. (page 5) de Godoy, L., J. Olsen, G. de Souza, G. Li, P. Mortensen, and M. Mann (2006), Status of complete proteome analysis by mass spectrometry: Silac labeled yeast as a model system., *Genome Biol*, *7*(6), R50. (page 2, 111)

Dempster, A., N. Laird, and D. Rubin (1977), Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. (page 18)

Desiere, F., *et al.* (2005), Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry., *Genome Biol*, *6*(1), R9. (page 29, 83)

Desiere, F., *et al.* (2006), The peptideatlas project., *Nucleic Acids Res*, *34*(Database issue), D655–8. (page 4, 29, 98)

Douglas, D., A. Frank, and D. Mao (2005), Linear ion traps in mass spectrometry., *Mass Spectrom Rev*, *24*(1), 1–29. (page 4)

Dowell, R., R. Jokerst, A. Day, S. Eddy, and L. Stein (2001), The distributed annotation system., *BMC Bioinformatics*, *2*, 7. (page 30, 96, 104)

Elias, J., and S. Gygi (2007), Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry., *Nat Methods*, *4*(3), 207–214. (page 16, 37, 62)

Elias, J., W. Haas, B. Faherty, and S. Gygi (2005), Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations., *Nat Methods*, *2*(9), 667–675. (page 4)

Eng, J., A. McCormack, and J. Yates (1994), An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *J. Am. Soc. Mass Spectrom*, *5*(11), 976–989. (page 6, 7)

Everley, P., C. Bakalarski, J. Elias, C. Waghorne, S. Beausoleil, S. Gerber, B. Faherty, B. Zetter, and S. Gygi (2006), Enhanced analysis of metastatic prostate cancer using stable isotopes and high mass accuracy instrumentation., *J Proteome Res*, *5*(5), 1224–1231. (page 37)

Fenyo, D., and R. Beavis (2003), A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes., *Anal Chem*, *75*(4), 768–774. (page 35, 36, 50)

Finn, R., J. Stalker, D. Jackson, E. Kulesha, J. Clements, and R. Pettett (2007), Proserver: a simple, extensible perl das server., *Bioinformatics*, *23*(12), 1568–1570. (page 96)

Fitzgibbon, M., Q. Li, and M. McIntosh (2007), Modes of inference for evaluating the confidence of peptide identifications, *Journal of Proteome Research Journal of Proteome Research J. Proteome Res.* (page 16)

Foster, L., C. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V. Mootha, and M. Mann (2006), A mammalian organelle map by protein correlation profiling., *Cell*, *125*(1), 187–199. (page 2)

Frank, A., and P. Pevzner (2005), Pepnovo: de novo peptide sequencing via probabilistic network modeling., *Anal Chem*, *77*(4), 964–973. (page 5)

Frank, A., M. Savitski, M. Nielsen, R. Zubarev, and P. Pevzner (2007), De novo peptide sequencing and identification with precision mass spectrometry., *J Proteome Res*, *6*(1), 114–123. (page 5)

Frottin, F., A. Martinez, P. Peynot, S. Mitra, R. Holz, C. Giglione, and T. Meinnel (2006), The proteomics of n-terminal methionine cleavage., *Mol Cell Proteomics*, *5*(12), 2336–2349. (page 87, 100)

Fullwood, M., C. Wei, E. Liu, and Y. Ruan (2009), Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses., *Genome Res*, *19*(4), 521–532. (page 28)

Furuno, M., T. Kasukawa, R. Saito, J. Adachi, H. Suzuki, R. Baldarelli, Y. Hayashizaki, and Y. Okazaki (2003), Cds annotation in full-length cdna sequence., *Genome Res*, *13*(6B), 1478–1487. (page 24)

Fusaro, V., D. Mani, J. Mesirov, and S. Carr (2009), Prediction of high-responding peptides for targeted protein assays by mass spectrometry., *Nat Biotechnol*, *27*(2), 190–198. (page 109)

Guigo, R., and M. Reese (2005), Egasp: collaboration through competition to find human genes., *Nat Methods*, *2*(8), 575–577. (page 102)

Guigo, R., *et al.* (2006), Egasp: the human encode genome annotation assessment project., *Genome Biol*, *7 Suppl 1*, S2.1–31. (page 26)

Gupta, N., *et al.* (2008), Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes., *Genome Res*, *18*(7), 1133–1142. (page 32)

Haas, W., B. Faherty, S. Gerber, J. Elias, S. Beausoleil, C. Bakalarski, X. Li, J. Villen, and S. Gygi (2006), Optimization and use of peptide mass measurement accuracy in shotgun proteomics., *Mol Cell Proteomics*, *5*(7), 1326–1337. (page 4)

Han, X., M. Jin, K. Breuker, and F. McLafferty (2006), Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons., *Science*, *314*(5796), 109–112. (page 3)

Harrow, J., A. Nagy, A. Reymond, T. Alioto, L. Patthy, S. Antonarakis, and R. Guigo (2009), Identifying protein-coding genes in genomic sequences., *Genome Biol*, *10*(1), 201. (page 24)

Harrow, J., *et al.* (2006), Gencode: producing a reference annotation for encode., *Genome Biol*, *7 Suppl 1*, S4.1–9. (page 26, 28)

Heeren, R., A. Kleinnijenhuis, L. McDonnell, and T. Mize (2004), A mini-review of mass spectrometry using high-performance fticr-ms methods., *Anal Bioanal Chem*, *378*(4), 1048–1058. (page 4)

Hsueh, H., J. Chen, and R. Kodell (2003), Comparison of methods for estimating the number of true null hypotheses in multiplicity testing., *Journal of biopharmaceutical statistics*, *13*(4), 675. (page 16)

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

Hubbard, T., *et al.* (2002), The ensembl genome database project., *Nucleic Acids Res*, *30*(1), 38–41. (page 25, 84)

Hubbard, T., *et al.* (2009), Ensembl 2009., *Nucleic Acids Res*, *37*(Database issue), D690–7. (page 25, 28)

Hunt, D., R. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A. Cox, E. Appella, and V. Engelhard (1992), Characterization of peptides bound to the class i mhc molecule hla-a2.1 by mass spectrometry., *Science*, *255*(5049), 1261–1263. (page 2)

Imanishi, T., *et al.* (2004), Integrative annotation of 21,037 human genes validated by full-length cdna clones., *PLoS Biol*, *2*(6), e162. (page 24)

International Human Genome Sequencing Consortium (2004), Finishing the euchromatic sequence of the human genome., *Nature*, *431*(7011), 931–945. (page 26)

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann (2005), Exponentially modified protein abundance index (empai) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein., *Mol Cell Proteomics*, *4*(9), 1265–1272. (page 107)

Jaffe, J., H. Berg, and G. Church (2004), Proteogenomic mapping as a complementary method to perform genome annotation., *Proteomics*, *4*(1), 59–77. (page 29)

Jin, J., and T. Cai (2006), Estimating the null and the proportion of non-null effects in large-scale multiple comparisons, *Arxiv preprint math.ST/0611108.* (page 16)

Jones, A., J. Siepen, S. Hubbard, and N. Paton (2009), Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines., *Proteomics*, *9*(5), 1220–1229. (page 34)

Jones, P., R. Cote, S. Cho, S. Klie, L. Martens, A. Quinn, D. Thorneycroft, and H. Hermjakob (2008), Pride: new developments and new datasets., *Nucleic Acids Res*, *36*(Database issue), D878–83. (page 4)

Käll, L., J. Canterbury, J. Weston, W. Noble, and M. MacCoss (2007), Semi-supervised learning for peptide identification from shotgun proteomics datasets., *Nat Methods*, *4*(11), 923–925. (page 18, 19, 20, 63, 64, 65, 67, 73, 74, 128)

Käll, L., J. Storey, M. MacCoss, and W. Noble (2008a), Assigning significance to peptides identified by tandem mass spectrometry using decoy databases., *J Proteome Res*, *7*(1), 29–34. (page 11, 16)

Käll, L., J. Storey, M. MacCoss, and W. Noble (2008b), Posterior error probabilities and false discovery rates: two sides of the same coin., *J Proteome Res*, *7*(1), 40–44. (page 11, 13, 20)

Käll, L., J. Storey, and W. Noble (2008c), Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry., *Bioinformatics*, *24*(16), i42–8. (page 13, 16, 20)

Kapp, E., *et al.* (2005), An evaluation, comparison, and accurate benchmarking of several publicly available ms/ms search algorithms: sensitivity and specificity analysis., *Proteomics*, *5*(13), 3475–3490. (page 6)

Keller, A., A. Nesvizhskii, E. Kolker, and R. Aebersold (2002), Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search., *Anal Chem*, *74*(20), 5383–5392. (page 13, 17)

Kersey, P., J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler (2004), The international protein index: an integrated database for proteomics experiments., *Proteomics*, *4*(7), 1985–1988. (page 29, 98)

Kersey, P., *et al.* (2009), Ensembl genomes: Extending ensembl across the taxonomic space., *Nucleic Acids Res.* (page 26)

Kim, S., N. Gupta, N. Bandeira, and P. Pevzner (2009), Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra., *Mol Cell Proteomics*, *8*(1), 53–69. (page 5)

Klammer, A., and M. MacCoss (2006), Effects of modified digestion schemes on the identification of proteins from complex mixtures., *J Proteome Res*, *5*(3), 695–700. (page 14)

Klimek, J., *et al.* (2007), The standard protein mix database: A diverse data set to assist in the production of improved peptide and protein identification software tools., *J Proteome Res.* (page 43)

Knowles, D., and A. McLysaght (2009), Recent de novo origin of human protein-coding genes., *Genome Res*, *19*(10), 1752–1759. (page 25)

Korf, I., P. Flicek, D. Duan, and M. Brent (2001), Integrating genomic homology into gene structure prediction., *Bioinformatics*, *17 Suppl 1*, S140–8. (page 25)

Kuhn, R., *et al.* (2009), The ucsc genome browser database: update 2009, *Nucleic acids research*, *37*(Database issue), D755. (page 28)

Kuster, B., P. Mortensen, J. Andersen, and M. Mann (2001), Mass spectrometry allows direct identification of proteins in large genomes., *Proteomics*, *1*(5), 641–650. (page 29)

Lander, E., *et al.* (2001), Initial sequencing and analysis of the human genome., *Nature*, *409*(6822), 860–921. (page 25, 26)

Liolios, K., I. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. Markowitz, and N. Kyrpides (2009), The genomes on line database (gold) in 2009: status of genomic and metagenomic projects and their associated metadata., *Nucleic Acids Res.* (page 22)

Lu, P., C. Vogel, R. Wang, X. Yao, and E. Marcotte (2007), Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation., *Nat Biotechnol*, *25*(1), 117–124. (page 107)

Mallick, P., *et al.* (2007), Computational prediction of proteotypic peptides for quantitative proteomics., *Nat Biotechnol*, *25*(1), 125–131. (page 109)

Mann, M., and M. Wilm (1994), Error-tolerant identification of peptides in sequence databases by peptide sequence tags., *Anal Chem*, *66*(24), 4390–4399. (page 5, 31)

Maquat, L. (2005), Nonsense-mediated mrna decay in mammals, *Journal of cell science*, *118(9)*(9), 1773–1776. (page 99)

Martens, L., H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove, and R. Apweiler (2005a), Pride: the proteomics identifications database., *Proteomics*, *5*(13), 3537–3545. (page 4)

Martens, L., J. Vandekerckhove, and K. Gevaert (2005b), Dbtoolkit: processing protein databases for peptide-centric proteomics., *Bioinformatics*, *21*(17), 3584–3585. (page 91)

McCormack, A., D. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. r. Yates (1997), Direct analysis and identification of proteins in mixtures by lc/ms/ms and database searching at the low-femtomole level., *Anal Chem*, *69*(4), 767–776. (page 2)

McGlincy, N., and C. Smith (2008), Alternative splicing resulting in nonsense-mediated mrna decay: what is the meaning of nonsense?, *Trends Biochem Sci*, *33*(8), 385–393. (page 117)

Meinshausen, N., and J. Rice (2006), Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses, *Annals of statistics*, *34*(1), 373. (page 16)

Merrihew, G., *et al.* (2008), Use of shotgun proteomics for the identification, confirmation, and correction of c. elegans gene annotations., *Genome Res*, *18*(10), 1660–1669. (page 121)

Moore, R., M. Young, and T. Lee (2002), Qscore: an algorithm for evaluating sequest database search results., *J Am Soc Mass Spectrom*, *13*(4), 378–386. (page 14)

Nagaraj, S., R. Gasser, and S. Ranganathan (2007), A hitchhiker's guide to expressed sequence tag (est) analysis., *Brief Bioinform*, *8*(1), 6–21. (page 24)

Nesvizhskii, A., and R. Aebersold (2004), Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms., *Drug Discov Today*, *9*(4), 173–181. (page 2, 30, 45)

Nesvizhskii, A., A. Keller, E. Kolker, and R. Aebersold (2003), A statistical model for identifying proteins by tandem mass spectrometry., *Anal Chem*, *75*(17), 4646–4658. (page 2, 30)

Nesvizhskii, A. I., O. Vitek, and R. Aebersold (2007), Analysis and validation of proteomic data generated by tandem mass spectrometry, *Nat Meth*, *4*(10), 787–797. (page 4)

Olsen, J., S. Ong, and M. Mann (2004), Trypsin cleaves exclusively c-terminal to arginine and lysine residues., *Mol Cell Proteomics*, *3*(6), 608–614. (page 5)

Parkinson, J., and M. Blaxter (2009), Expressed sequence tags: an overview., *Methods Mol Biol*, *533*, 1–12. (page 24)

Parks, B., L. Jiang, P. Thomas, C. Wenger, M. Roth, M. n. Boyne, P. Burke, K. Kwast, and N. Kelleher (2007), Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers., *Anal Chem*, *79*(21), 7984–7991. (page 3)

Parra, G., E. Blanco, and R. Guigo (2000), Geneid in drosophila., *Genome Res*, *10*(4), 511–515. (page 25)

Parra, G., P. Agarwal, J. Abril, T. Wiehe, J. Fickett, and R. Guigo (2003), Comparative gene prediction in human and mouse., *Genome Res*, *13*(1), 108–117. (page 25)

Pasa-Tolic, L., C. Masselon, R. Barry, Y. Shen, and R. Smith (2004), Proteomic analyses using an accurate mass and time tag strategy., *Biotechniques*, *37*(4), 621–4, 626–33, 636 passim. (page 39)

Patterson, S., and R. Aebersold (2003), Proteomics: the first decade and beyond., *Nat Genet*, *33 Suppl*, 311–323. (page 2)

Peng, J., J. Elias, C. Thoreen, L. Licklider, and S. Gygi (2003), Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (lc/lc-ms/ms) for large-scale protein analysis: the yeast proteome., *J Proteome Res*, *2*(1), 43–50. (page 4)

Pennisi, E. (2009), Dna sequencing. no genome left behind., *Science*, *326*(5954), 794–795. (page 22)

Perkins, D., D. Pappin, D. Creasy, and J. Cottrell (1999), Probability-based protein identification by searching sequence databases using mass spectrometry data., *Electrophoresis*, *20*(18), 3551–3567. (page 6, 7, 34)

Pitzer, E., A. Masselot, and J. Colinge (2007), Assessing peptide de novo sequencing algorithms performance on large and diverse data sets., *Proteomics.* (page 5)

Pruitt, K., T. Tatusova, and D. Maglott (2006), Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic acids research.* (page 28)

Pruitt, K., *et al.* (2009), The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes., *Genome Res*, *19*(7), 1316–1323. (page 28)

Resing, K., *et al.* (2004), Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics., *Anal Chem*, *76*(13), 3556–3568. (page 34)

Rodriguez, J., N. Gupta, R. D. Smith, and P. A. Pevzner (2007), Does trypsin cut before proline?, *Journal of Proteome Research Journal of Proteome Research J. Proteome Res.* (page 5, 65)

Roepstorff, P., and J. Fohlman (1984), Proposal for a common nomenclature for sequence ions in mass spectra of peptides., *Biomed Mass Spectrom*, *11(11)*(11), 601. (page 2, 37, 105)

Roth, M., B. Parks, J. Ferguson, M. n. Boyne, and N. Kelleher (2008), "proteotyping": population proteomics of human leukocytes using top down mass spectrometry., *Anal Chem*, *80*(8), 2857–2866. (page 3)

Rudnick, P., Y. Wang, E. Evans, C. Lee, and B. Balgley (2005), Large scale analysis of mascot results using a mass accuracy-based threshold (math) effectively improves data interpretation., *J Proteome Res*, *4*(4), 1353–1360. (page 37, 53)

Savitski, M., M. Nielsen, and R. Zubarev (2005), New data base-independent, sequence tag-based scoring of peptide ms/ms data validates mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of ms/ms techniques., *Mol Cell Proteomics*, *4*(8), 1180–1188. (page 37)

Schandorff, S., J. Olsen, J. Bunkenborg, B. Blagoev, Y. Zhang, J. Andersen, and M. Mann (2007), A mass spectrometry-friendly database for csnp identification., *Nat Methods*, *4*(6), 465–466. (page 31, 87, 100)

Sevinsky, J., B. Cargile, M. Bunger, F. Meng, N. Yates, R. Hendrickson, and J. J. Stephenson (2008), Whole genome searching with shotgun proteomic data: applications for genome annotation., *J Proteome Res*, *7*(1), 80–88. (page 31)

Shadforth, I., T. Dunkley, K. Lilley, D. Crowther, and C. Bessant (2005), Confident protein identification using the average peptide score method coupled with search-specific, ab initio thresholds., *Rapid Commun Mass Spectrom*, *19*(22), 3363–3368. (page 30)

Shadforth, I., W. Xu, D. Crowther, and C. Bessant (2006), Gapp: A fully automated software for the confident identification of human peptides from tandem mass spectra, *Journal of proteome research.* (page 30, 83)

Shaffer, J. (1995), Multiple hypothesis testing, *Annual Review of Psychology*, *46*(1), 561–584. (page 10)

Shevchenko, A., M. Wilm, O. Vorm, and M. Mann (1996), Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels., *Anal Chem*, *68*(5), 850–858. (page 2)

Shin, S., and G. Sanders (2006), Denormalization strategies for data retrieval from data warehouses, *Decision Support Systems*, *42*(1), 267–282. (page 87)

Simpson, R., L. Connolly, J. Eddes, J. Pereira, R. Moritz, and G. Reid (2000), Proteomic analysis of the human colon carcinoma cell line (lim 1215): development of a membrane protein database., *Electrophoresis*, *21*(9), 1707–1732. (page 2)

Slater, G., and E. Birney (2005), Automated generation of heuristics for biological sequence comparison., *BMC Bioinformatics*, *6*, 31. (page 24)

Sonnhammer, E., S. Eddy, and R. Durbin (1997), Pfam: a comprehensive database of protein domain families based on seed alignments., *Proteins*, *28*(3), 405–420. (page 125)

Stabenau, A., G. McVicker, C. Melsopp, G. Proctor, M. Clamp, and E. Birney (2004), The ensembl core software libraries, *Genome research*, *14*(5), 929. (page 86, 87)

Stanke, M., and S. Waack (2003), Gene prediction with a hidden markov model and a new intron submodel, *Bioinformatics-Oxford*, *19*(2), 215–225. (page 25, 86)

Stanke, M., A. Tzvetkova, and B. Morgenstern (2006), Augustus at egasp: using est, protein and genomic alignments for improved gene prediction in the human genome., *Genome Biol*, *7 Suppl 1*, S11.1–8. (page 102)

States, D., G. Omenn, T. Blackwell, D. Fermin, J. Eng, D. Speicher, and S. Hanash (2006), Challenges in deriving high-confidence protein identifications from data gathered by a hupo plasma proteome collaborative study., *Nat Biotechnol*, *24*(3), 333–338. (page 8)

Stein, L. (2001), Genome annotation: from sequence to biology., *Nat Rev Genet*, *2*(7), 493–503. (page 23)

Storey, J. (2002), A direct approach to false discovery rates, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*(3), 479–498. (page 16)

Storey, J., and R. Tibshirani (2003), Statistical significance for genomewide studies., *Proc Natl Acad Sci U S A*, *100*(16), 9440–9445. (page 10, 16)

Syka, J., *et al.* (2004), Novel linear quadrupole ion trap/ft mass spectrometer: performance characterization and use in the comparative analysis of histone h3 post-translational modifications., *J Proteome Res*, *3*(3), 621–626. (page 4)

Tabb, D., A. Saraf, and J. r. Yates (2003), Gutentag: high-throughput sequence tagging via an empirically derived fragmentation model., *Anal Chem*, *75*(23), 6415–6421. (page 5)

Tanner, S., H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P. Pevzner, and V. Bafna (2005), Inspect: identification of posttranslationally modified peptides from tandem mass spectra., *Anal Chem*, *77*(14), 4626–4639. (page 5, 31)

Tanner, S., Z. Shen, J. Ng, L. Florea, R. Guigo, S. Briggs, and V. Bafna (2007), Improving gene annotation using peptide mass spectrometry., *Genome Res.* (page 31, 115, 123)

Taylor, J., and R. Johnson (1997), Sequence database searches via de novo peptide sequencing by tandem mass spectrometry., *Rapid Commun Mass Spectrom*, *11*(9), 1067–1075. (page 5)

The ENCODE Project Consortium (2004), The encode (encyclopedia of dna elements) project, *Science*, *306*, 636–640. (page 26)

Tress, M., B. Bodenmiller, R. Aebersold, and A. Valencia (2008), Proteomics studies confirm the presence of alternative protein isoforms on a large scale., *Genome Biol*, *9*(11), R162. (page 31, 114, 115)

Ulintz, P., J. Zhu, Z. Qin, and P. Andrews (2006), Improved classification of mass spectrometry database search results using newer machine learning approaches., *Mol Cell Proteomics*, *5*(3), 497–509. (page 34)

Venter, J., *et al.* (2001), The sequence of the human genome., *Science*, *291*(5507), 1304–1351. (page 25, 26)

Wang, E., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. Kingsmore, G. Schroth, and C. Burge (2008), Alternative isoform regulation in human tissue transcriptomes., *Nature*, *456*(7221), 470–476. (page 29, 91)

Wang, Z., M. Gerstein, and M. Snyder (2009), Rna-seq: a revolutionary tool for transcriptomics., *Nat Rev Genet*, *10*(1), 57–63. (page 28)

Washburn, M., D. Wolters, and J. r. Yates (2001), Large-scale analysis of the yeast proteome by multidimensional protein identification technology., *Nat Biotechnol*, *19*(3), 242–247. (page 2)

Washietl, S., *et al.* (2007), Structured rnas in the encode selected regions of the human genome., *Genome Res*, *17*(6), 852–864. (page 21)

Waterston, R., *et al.* (2002), Initial sequencing and comparative analysis of the mouse genome., *Nature*, *420*(6915), 520–562. (page 98, 107)

Wilming, L., J. Gilbert, K. Howe, S. Trevanion, T. Hubbard, and J. Harrow (2008), The vertebrate genome annotation (vega) database., *Nucleic Acids Res*, *36*(Database issue), D753–60. (page 28)

Wolters, D., M. Washburn, and J. r. Yates (2001), An automated multidimensional protein identification technology for shotgun proteomics., *Anal Chem*, *73*(23), 5683–5690. (page 2)

Wright, J., D. Sugden, S. Francis-McIntyre, I. Riba-Garcia, S. Gaskell, I. Grigoriev, S. Baker, R. Beynon, and S. Hubbard (2009), Exploiting proteomic data for genome annotation and gene model validation in aspergillus niger., *BMC Genomics*, *10*, 61. (page 31)

Wu, C., *et al.* (2006), The universal protein resource (uniprot): an expanding universe of protein information, *Nucleic acids research*, *34*(Database Issue), D187. (page 28)

Yates, J. r., J. Eng, and A. McCormack (1995), Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases., *Anal Chem*, *67*(18), 3202–3210. (page 29)

Zhang, T., P. Haws, and Q. Wu (2004), Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation., *Genome Res*, *14*(1), 79–89. (page 115)

Zubarev, R. (2006), Protein primary structure using orthogonal fragmentation techniques in fourier transform mass spectrometry., *Expert Rev Proteomics*, *3*(2), 251–261. (page 4)

Zubarev, R., and M. Mann (2007), On the proper use of mass accuracy in proteomics., *Mol Cell Proteomics*, *6*(3), 377–381. (page 37, 45, 62)

Zubarev, R., P. Hakansson, and B. Sundqvist (1996), Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements, *Anal. Chem*, *68*(22), 4060–4063. (page 62)

Appendix A

Publications and presentations

A.1 Publications

- Brosch M, Swamy S, Hubbard T, Choudhary J: **Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold.** *Mol Cell Proteomics* 2008, **7:**962-970.

- Brosch M, Yu L, Hubbard T, Choudhary J: **Accurate and Sensitive Peptide Identification with Mascot Percolator.** *J Proteome Res* 2009, **8:**3176-3181.

- Brosch M, Choudhary J: **Scoring and validation of tandem MS peptide identification methods.** Proteome bioinformatics: Informatics for mass-spectrometry based protein science for the series “Methods of Molecular Biology”, chapter 4. *Humana Press* 2010.

- Brosch M, Harrow J, Saunders G, Frankish A, Collins M, Yu L, Choudhary J, Hubbard T: **Refining annotation of the mouse genome using mass spectrometry.** In preparation.

- Mattison J, Kool J, Uren A, Ridder J, Wessels L, Jonkers J, Bignell G, Butler

A, Rust A, Brosch M, Wilson C, Weyden L, Largaespada D, Stratton M, Futreal P, Lohuizen M, Berns A, Collier L, Hubbard T, Adams D: **Large-scale cross-species comparative oncogenomics identifies candidate oncogenes and tumour suppressor genes.** *Cancer Research* 2009.

- Freeman TC, Goldovsky L, Brosch M, van Dongen S, Maziere P, Grocock RJ, Freilich S, Thornton J, Enright AJ: **Construction, visualisation, and clustering of transcription networks from microarray expression data.** *PLoS Comput Biol* 2007, **3:**2032-2042.

A.2 Presentations

- Brosch M, Hubbard T, Choudhary J: **Development of an efficient proteogenomic annotation pipeline and its application to the Mouse genome**, International Mass Spectrometry Conference, Bremen (Germany), 2009 (poster).

- Brosch M: **Sensitive and accurate peptide identification with Mascot Percolator**, MatrixScience ASMS Workshop and User Meeting, Philadelphia (USA), 2009 (oral).

- Brosch M, Hubbard T, Choudhary J: **Development of a reliable and efficient genome annotation pipeline using proteomic mass spectrometry data**, ASMS Conference on Mass Spectrometry, Philadelphia (USA), 2009 (poster).

- Brosch M, Hubbard T, Choudhary J: **Mascot Percolator: improved peptide and protein identification**, ASMS Conference on Mass Spectrometry, Denver (USA), 2008 (poster).

- Brosch M, Hubbard T, Choudhary J: **An empirical Mascot scoring scheme**

for high accuracy mass spectrometry that enhances peptide identification, ASMS Conference on Mass Spectrometry, Indianapolis (USA), 2007 (poster).