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

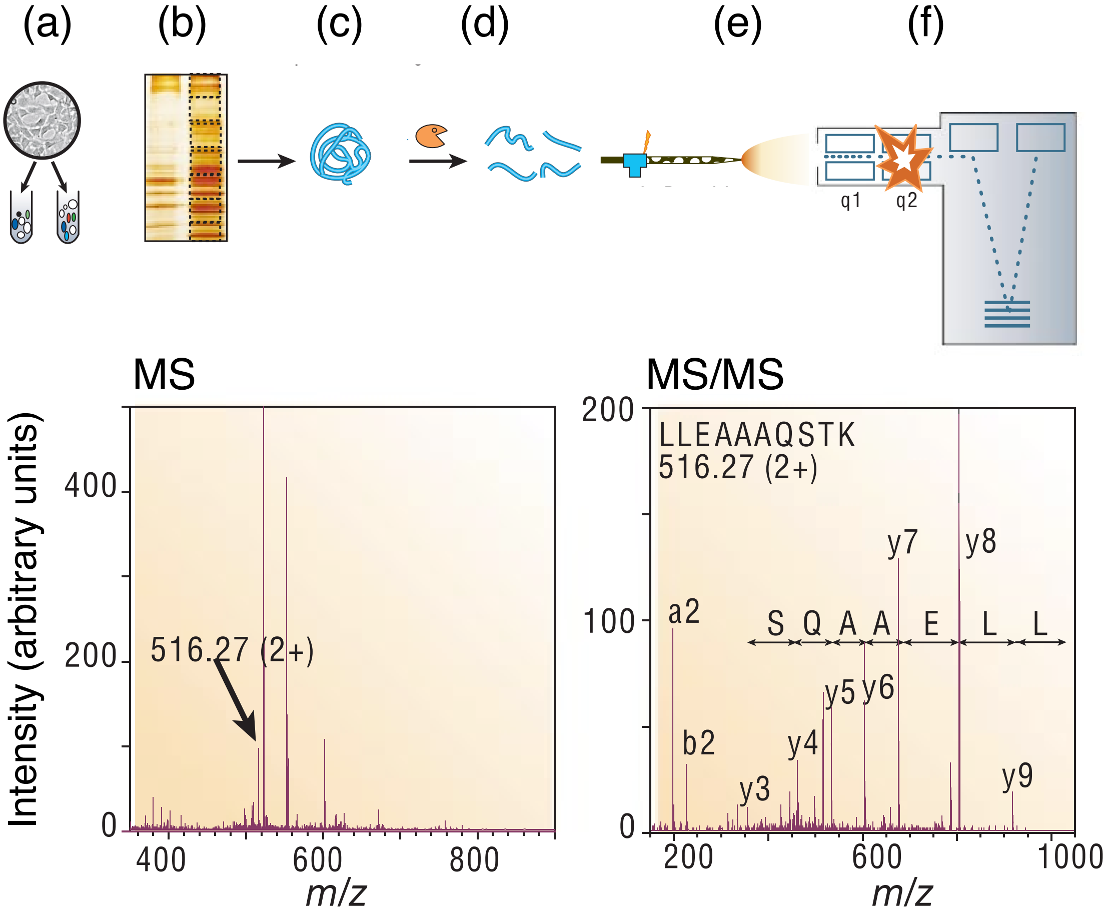


Figure 1.1: Schematic of a generic bottom-up proteomics MS experiment. (a) Sample preparation and fractionation, (b) protein separation via gel-electrophoresis, (c) protein extraction, (d) enzymatic protein digestion, (e) separation of peptides in one or multiple steps of liquid chromatography, followed 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

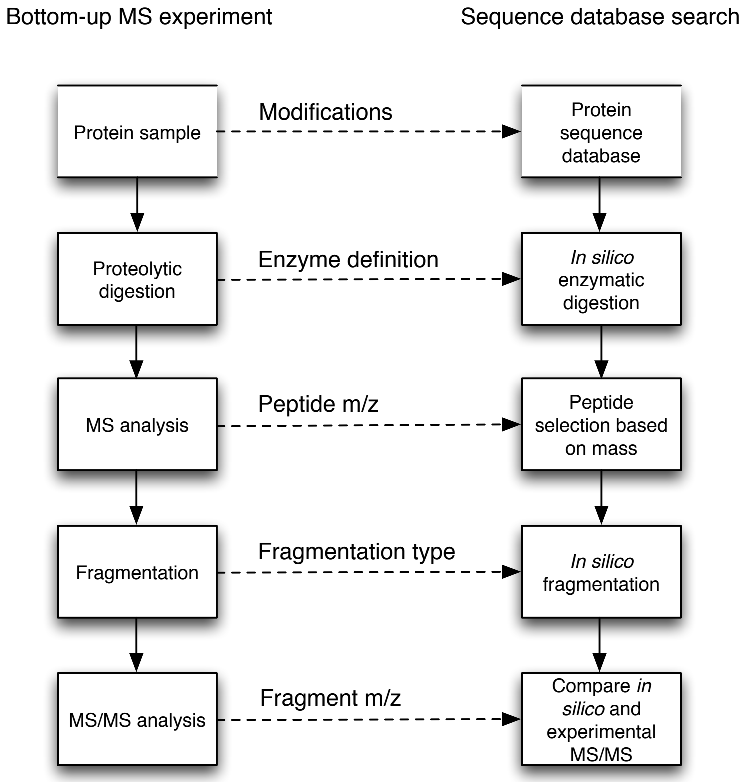


Figure 1.2: Concept of sequence database searching resembles a generic bottom-up MS experiment, as for each stage of the experiment, an *in silico* equivalent component is available.

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.

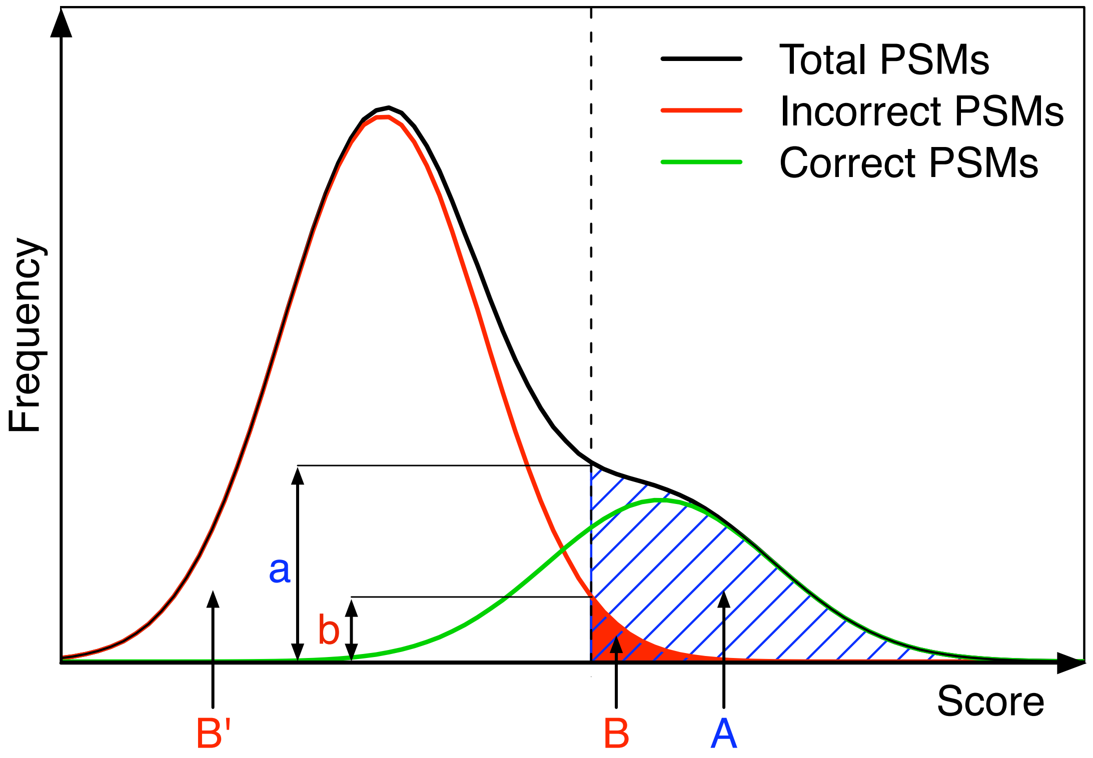
Often only the top hit is considered for further investigation and some search engines like X!Tandem (Craig and Beavis, 2004) exclusively report that very best match. However, not all these identifications are correct. Sorting all top hit PSMs (absolute ranking) according to their score enables the selective 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.



$$FPR = B/(B'+B) \qquad FDR = B/A = (\sum_{i=1}^{A} PEP_i)/A \qquad PEP = b/a$$

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

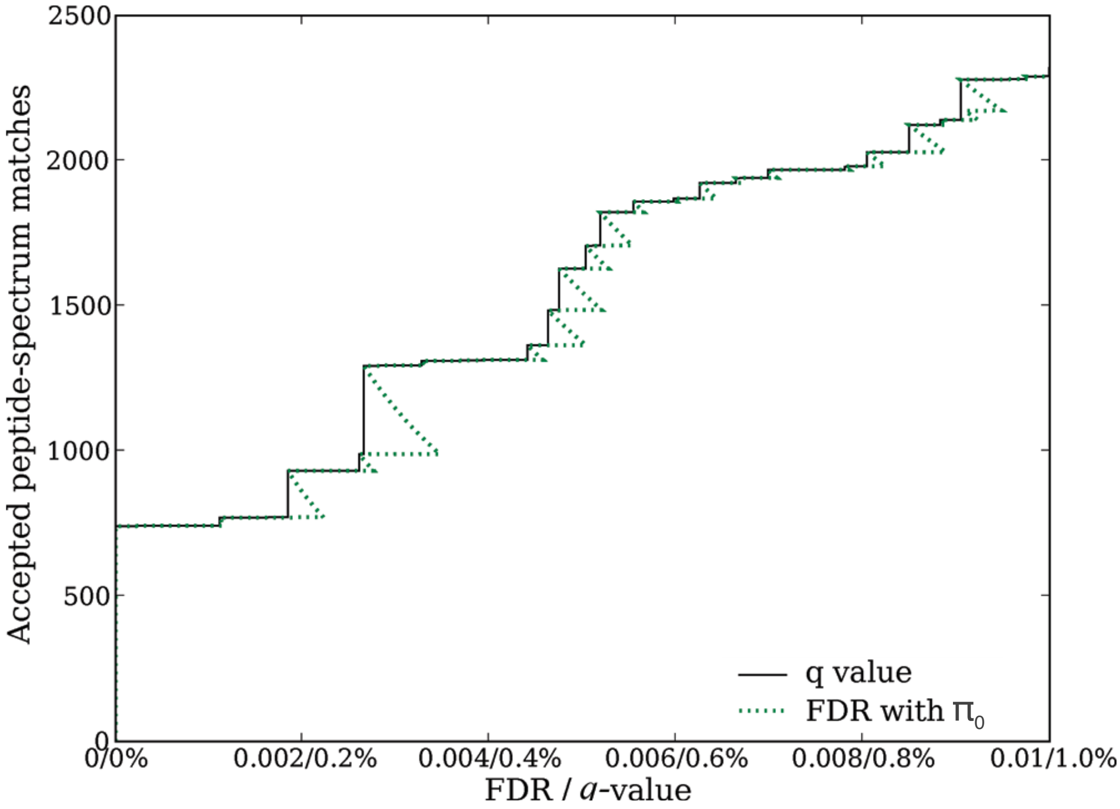


Figure 1.4: FDR compared with q-value: two or more different scores may lead to the same FDR, whereas the q-value is defined as the minimal FDR threshold at which a PSM is accepted, allowing to associate every PSM score with a specific q-value. Adapted from Käll *et al.* (2008a), figure 4b.

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.

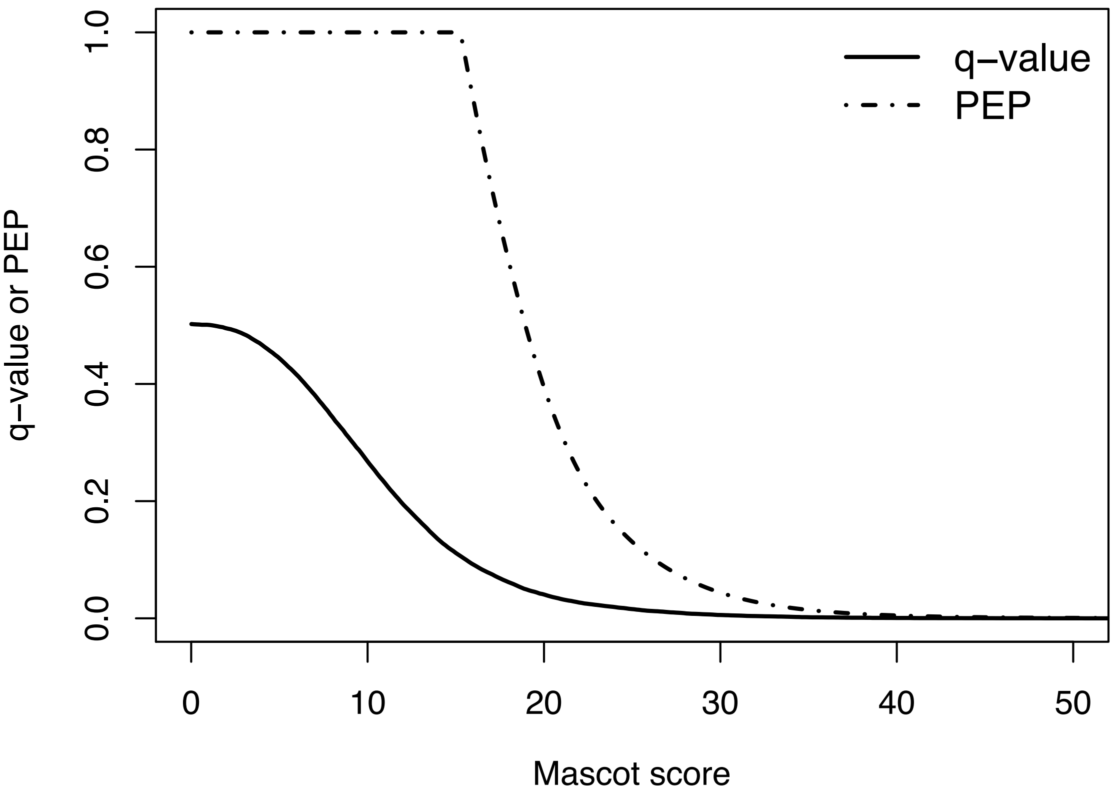


Figure 1.5: Mascot PSM scores were transformed into q-values and posterior error probabilities (PEP) using Qvality (see section 1.1.2.3). A score cut-off of 30 demonstrates the fundamental difference of the two significance measures: the q-value would have reported about 0.5% of all the PSMs as incorrect 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

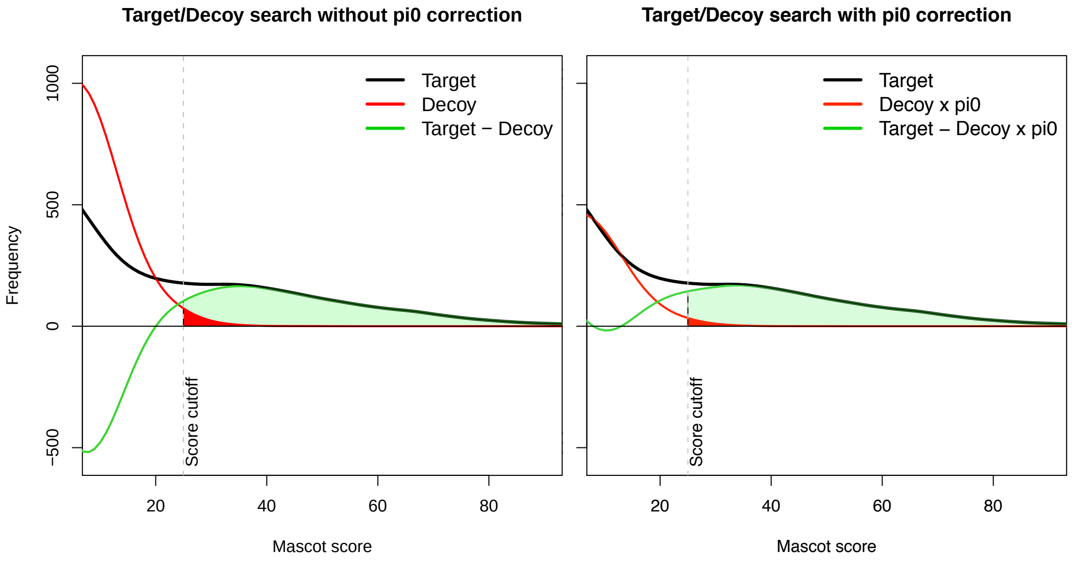


Figure 1.6: Score distributions of a target and decoy search with and without accounting for π_0 (pi0, percentage of target PSMs that are incorrect). Generally, the target score distribution (black) is a mixture of correct (green) and incorrect (red) peptide-spectrum matches, while the decoy matches 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.

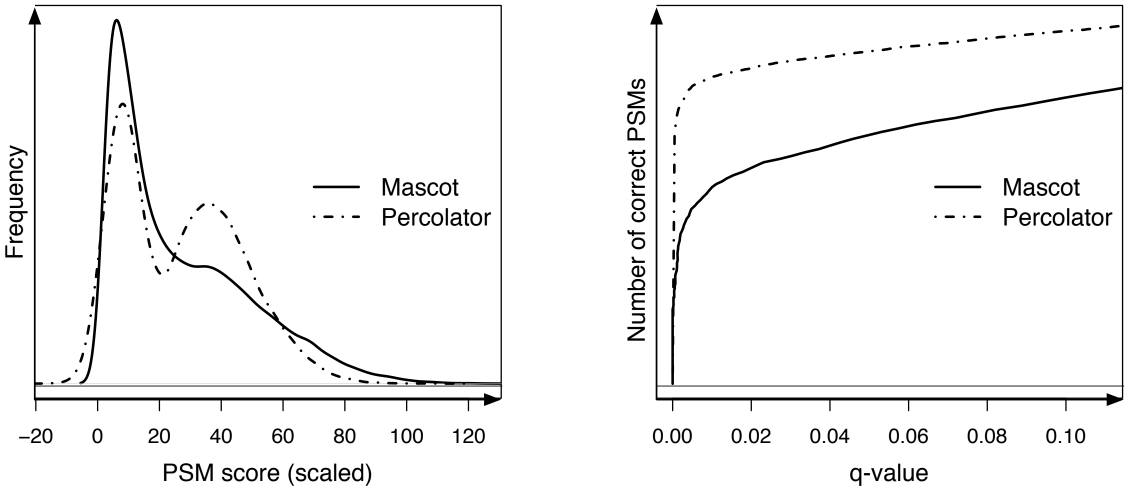


Figure 1.7: Distributions of Mascot and Percolator scores were generated from a high accuracy LTQ-FT Ultra dataset (left). This illustrates the bi-modal nature of PSM matching scores as simulated in figure 1.3 and further demonstrates the discrimination performance improvement between correct and incorrect 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

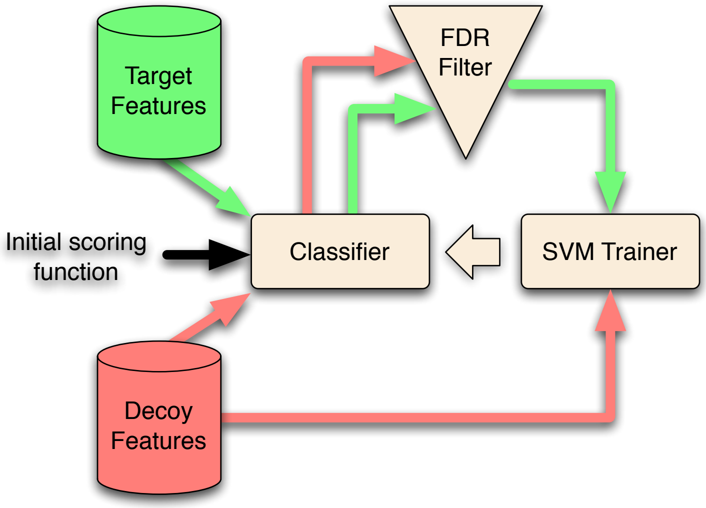


Figure 1.8: Schematic of the iterative learning process as implemented by Percolator

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

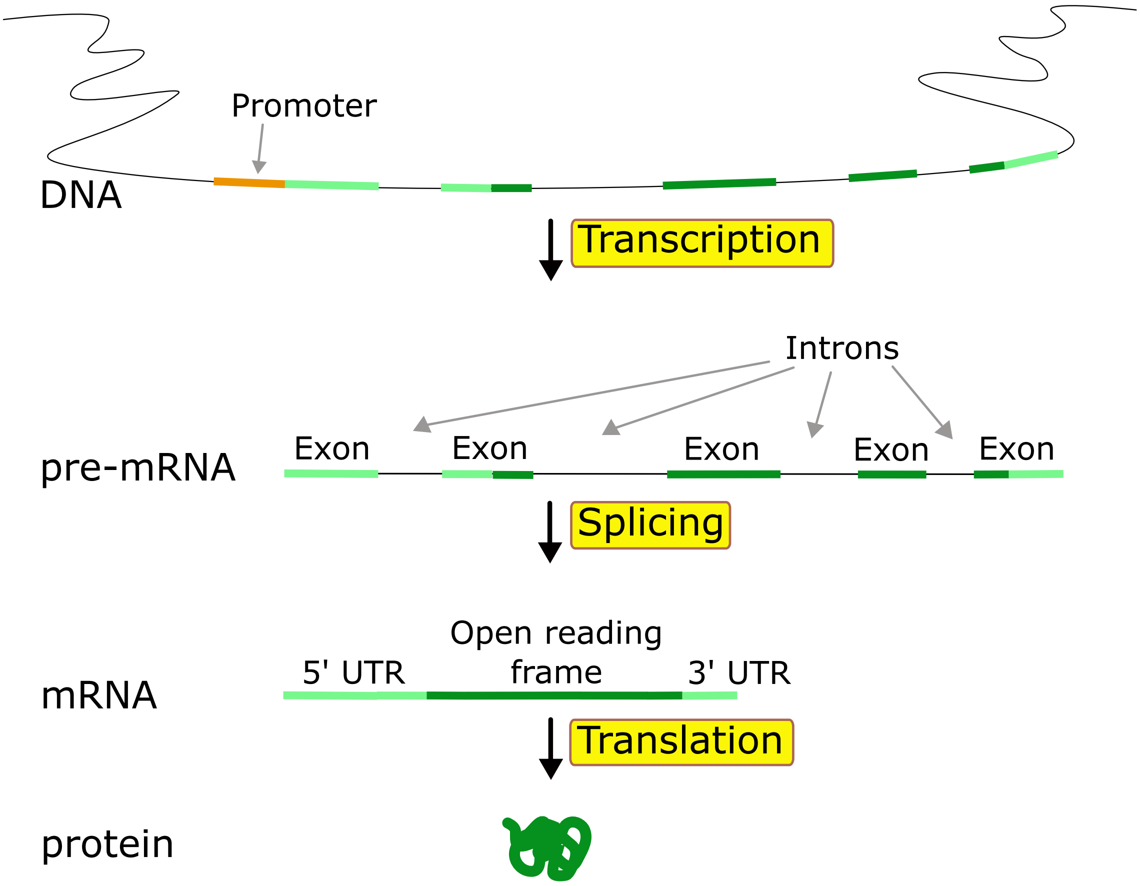


Figure 1.9: Illustration of gene transcription and translation according to the standard model. The figure was adapted from Wikipedia (`http://en.wikipedia.org/wiki/File:Gene2-plain.svg`)

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.

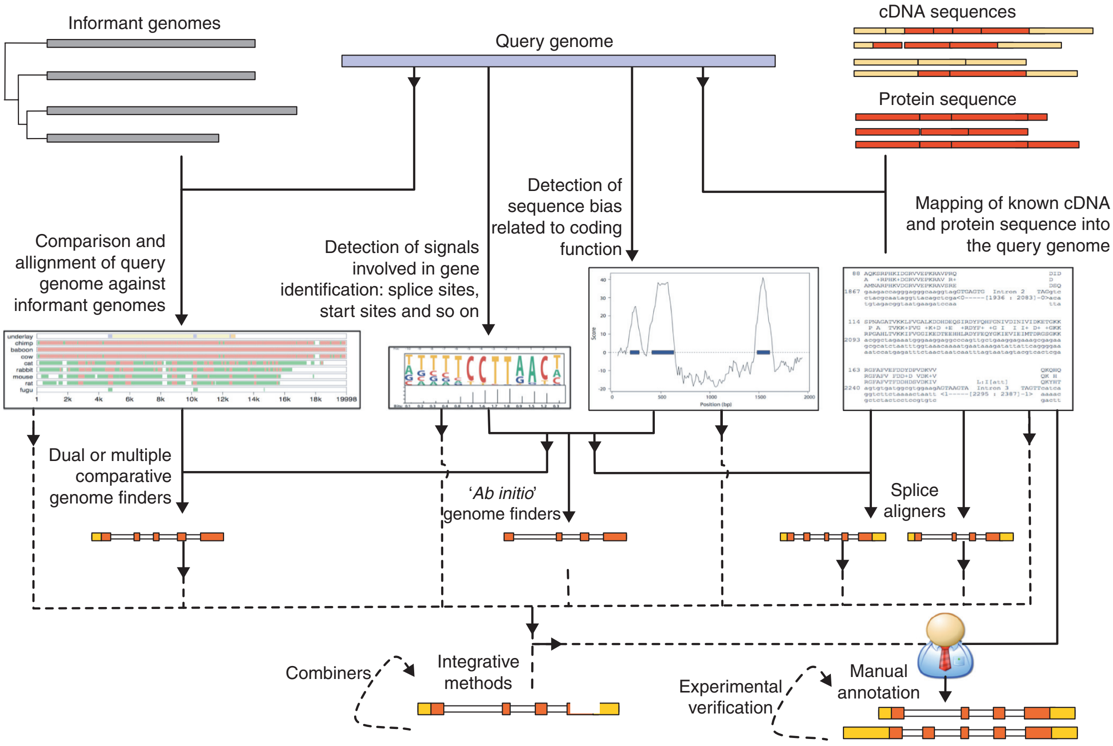

Informant genomes
Query genome
cDNA sequences
Protein sequence
Comparison and allignment of query genome against informant genomes
Detection of signals involved in gene identification: splice sites, start sites and so on
Detection of sequence bias related to coding function
Mapping of known cDNA and protein sequence into the query genome
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

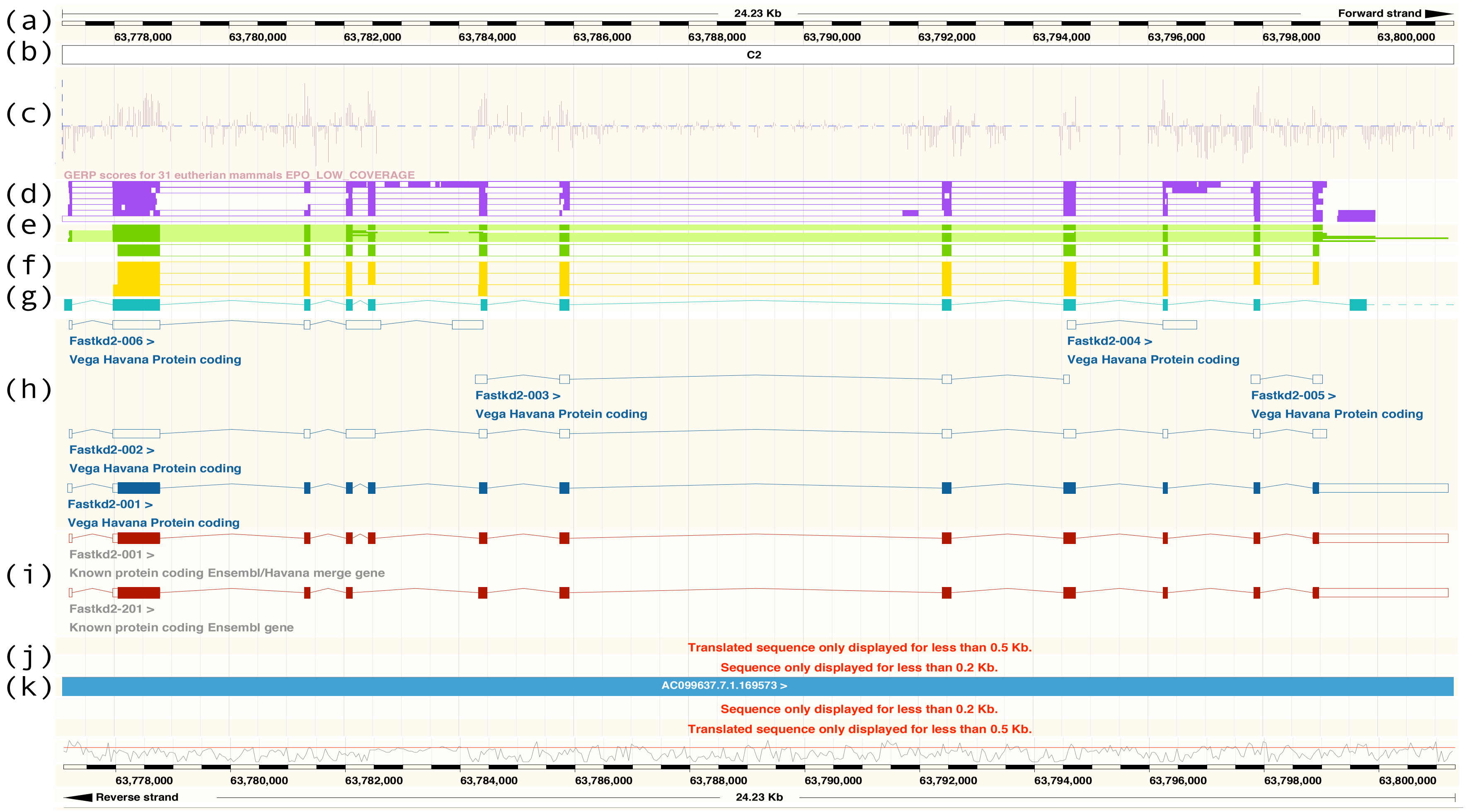

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
24.23 Kb
Forward strand
63,778,000
63,780,000
63,782,000
63,784,000
63,786,000
63,788,000
63,790,000
63,792,000
63,794,000
63,796,000
63,798,000
63,800,000
C2
GERP scores for 31 eutherian mammals EPO_LOW_COVERAGE
Fastkd2-006 >
Vega Havana Protein coding
Fastkd2-004 >
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.
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.