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.