CHAPTER 6

Discussion

The major highlight of the study has been the use of high-resolution and high mass accuracy mass spectrometry-based proteomics for deep proteomic analysis and accurate re-annotation of the protein coding genes in Candida tropicalis. Multipronged fractionation methods in combination with state-of-the-art next generation mass spectrometry were utilized to achieve comprehensive proteomic coverage of the opportunist pathogen, Candida tropicalis. To our knowledge, this is the first high throughput proteomic profiling carried out on C. tropicalis. The expressed proteome was identified using high resolution and high accurate peptide spectral data resulted in validation of several predicted and hypothetical proteins. The uses of high-end mass spectrometers have resulted in a more confident identification which is the most critical factor in any proteomic study. The important aspect has been acquisition of data in high resolution mode at both precursor ion mass and fragment ion mass during LC-MS/MS runs. These critical points have made data of this study extremely valuable for the research community. High accuracy spectral identification has been achieved by observing sub-ppm level of mass error in the entire dataset.

As mentioned earlier, the first aim of this study was to identify the basal level protein expression using high resolution mass spectrometry. We did not intend to study the proteome of C. tropicalis under any stress conditions or drug response. Label-free quantitation using iBAQ allowed us to estimate protein abundance across the identified proteome. The identified proteins represent a wide variety of cellular pathways and biological processes including protein synthesis, glycolysis, gluconeogenesis, and protein trafficking. The most abundant proteins identified in this study belonged to energy metabolism pathways such as tricarboxylic acid (TCA) and glycolysis. Another important aspect of this study is confirmation predicted proteins. ~80% of C. tropicalis proteins are annotated as “hypothetical protein” or “predicted protein”. Here, we were able to provide protein expression evidence for ~1,800 of these hypothetical or predicted proteins. These proteins could be studied further in order to ascertain function for these proteins.

The high-resolution and high mass accuracy mass spectrometric data has been further utilized in identification of novel protein coding regions. This has resulted in an increased accuracy of
genome annotation in an otherwise predicted annotation. The proteogenomic analysis performed in the study has led to the identification of identified 86 novel genes, 12 novel exons and corrected 49 computationally predicted gene models. Extensive bioinformatics and manual evaluation further enhanced the confidence of this dataset. As described, multiple search algorithms were deployed in order to increase the protein sequence coverage. Stringent false discovery rates were calculated in order to achieve confident peptide identifications. The occurrence of false positives cannot be completely ruled out from this kind of an analysis. The widely utilised approach of using reverse or decoy database search results for identifying false hits has been applied in the current study. Decoy database-based elimination of false positives has its own limitations and care should be taken while accepting peptide hits with low ion score or low number of PSMs.

The *C. tropicalis* genome sequence has been available since 2009. However, there is hardly any change in the available protein database. The predicted or hypothetical protein entries have not been modified due to lack of experimental evidence. With this study, we anticipate that necessary changes in the protein database would occur. The identification of several novel protein coding genes and refinements in existing predicted gene structures in this study also needs to be incorporated into the available protein database.

**Future perspectives**

We have identified a total of 1,513 hypothetical and 285 predicted proteins in this study and thus provided translational evidence for these otherwise computationally predicted protein coding genes. The functional biology of these hypothetical and predicted protein coding genes needs to be performed, both computationally as well as experimentally. Further, the role of identified secreted proteins in the pathophysiology of Candidiasis needs to be examined. As it is an established fact that secreted proteins play a role in the pathobiology of disease, the probability of these proteins playing a vital role in disease manifestation and/or progression is very high. Additionally, this study of basal level expression may act as a reference proteome map while studying the proteome of this opportunistic pathogen under various conditions such as yeast to hyphal transition, change in pH conditions, and change in media composition.
In addition to proteomic analysis, using a proteogenomics approach, we have proposed novel protein-coding genes and gene refinement of existing protein-coding genes. These proposed novel protein-coding gene models contain a number of well characterized domains. Also, one of the novel proteins identified was found to be secreted in nature. Functional analysis of these novel proteins need to be carried out to decipher the role they may play in Candidiasis. With a consistent increase in the number of Candidiasis cases due to *C. tropicalis* infection and development of azole resistance by this species, identification of new and improved modes of therapy is of prime importance.

This study has helped me gain insights into the complicated process of genome annotation. Taking into consideration both the advantages and shortcomings of existing methods for genome annotation, it is clear that proteogenomic approach has to be paired with computational methods for a better and more accurate annotation of new and existing genomes.