Synopsis

Introduction

There has been a lot of effort towards understanding and tackling *Mycobacterium tuberculosis* using tools of ‘molecular biology’. Knowledge of genome and accurate genome annotation is necessary for modern biological investigations. Constant revisions are made to the genome annotation of *Mycobacterium* over the years. Yet, the presence of many coding sequences which are missed in annotation is predicted. In the present study, proteogenomic analysis on *M. tuberculosis* was carried out which led to the identification of many novel protein coding genes were identified. The complete genome for *M. tuberculosis* was sequenced for H37Rv strain in 1998 by Cole et al., which was followed by that of CDC1551 strain by Fleischmann *et al.* in 2002. Analysis of the genome sequence had provided many insights into biology of the pathogen. Total 3,924 genes were predicted from the first draft of the genome. In the re-annotation of the H37Rv genome, the gene number was changed to 3,995. Since then, genome annotation is continuously re-evaluated based on bioinformatics as well as experimental data from the public domain. As of April 2012, the TubercuList database consisted of 4,015 annotated genes in the genome of *M. tuberculosis* H37Rv.

Present study was carried out with the following objectives-

1) In depth mass spectrometry-based proteomic analysis of *M. tuberculosis* H37Rv, and

2) Genome annotation of *M. tuberculosis* H37Rv using proteogenomic approach

Methods and research outcome

Two types of samples from *M. tuberculosis* were analyzed -1) Whole cell lysate and 2) Culture filtrate. In data-dependent mass spectrometry analysis most abundant proteins from the sample are repeatedly analyzed and identified. Because of this, abundant proteins have a masking effect on less abundant proteins. Extensive fractionation at both protein and peptide
level is a key to overcome this masking effect. H37Rv strain *M. tuberculosis* was grown in liquid media for up to 5 weeks. Culture filtrate was prepared by filtering the liquid media culture through 0.22 µm membrane filters. Whole cell lysate was pre-fractionated by three different techniques prior to LC-MS/MS analysis- SDS-PAGE, strong cation exchange (SCX) chromatography and isoelectric focusing (IEF/OFFGEL). Culture filtrate was fractionated only by SDS-PAGE. LC-MS/MS analysis was carried out for total 123 fractions. For the purpose of genome annotation it is important that peptide identification is highly confident. Therefore, the LTQ-Orbitrap Velos mass spectrometer was chosen for the analysis of samples as it has very high resolution power and high speed of analysis. Because of this, considerable depth in proteome coverage was achieved while obtaining high quality data. Around 1,800,000 MS/MS spectra were obtained as a result of mass spectrometry analysis.

Use of more than one MS/MS search algorithms can lead to maximization of identifications. Three search algorithms were used for peptide identification – Mascot, Sequest and MassWiz. From the 3,998 annotated proteins of *M. tuberculosis* H37Rv, 3,167 proteins were identified amounting to ~80% of the total proteome (number of protein sequences was revised to 4,003 in NCBI RefSeq database after the findings of the study were published). One hundred and eighty proteins were uniquely identified from the culture filtrate analysis. The second objective of the study was to carry out proteogenomic analysis using the peptide data to improve the genome annotation of *M. tuberculosis* H37Rv. To identify novel protein coding regions in the genome of *M. tuberculosis* H37Rv, mass spectrometry data was searched against hypothetical protein database generated by genome translated in six frames. Peptides identified from this genome search were mapped to protein database and those
peptides which did not map to protein database were listed separately. Every peptide spectral match for these novel peptides was manually validated. Genome annotation refinement was carried out by combining good quality novel peptides, alternative gene prediction models (or ORF finding tools) and orthologous protein sequences from other *M. tuberculosis* strains and related organisms. The instances of genome annotation refinement from this study can be categorized as – 1) Identification of novel protein coding gene, 2) Change in the structure of a gene, and 3) Identification of alternative translational start sites. Forty one novel protein coding genes were identified in the H37Rv genome. About ninety instances were found where gene structure change could be proposed. Finally, analysis of acetylated or unmodified N-terminal peptides led to the identification of 33 alternative translational start sites and confirmation of 727 existing translational start sites.

An attempt is made to assign broad functional category to the novel proteins identified in this study. Two approaches were used for this - sequence homology and genomic context.

**Conclusions**

Comprehensive proteomic profile of H37Rv strain of *M. tuberculosis* was obtained by the identification of 80% of the known proteome of the organism. Proteogenomic analysis using mass spectrometry generated proteomic data led to the identification of 41 genes not annotated previously. Gene structure changes and alternative translational start sites for some of the genes were also proposed based on this analysis. This approach of using mass spectrometry-based proteomic data to annotate protein coding genes in the genome can prove to be an essential method complementary to computational methods for annotating both newly sequenced genomes as well as genome sequences which have been available for many years such as that of *M. tuberculosis*. 

vi