2. REVIEW OF LITERATURE

2.1 Etiology of TB

TB is an ancient airborne contagious disease caused by *M. tuberculosis* has been the major cause of morbidity and mortality among the infectious diseases. The social and economic consequences of TB had a profound effect on its existence. The disease disproportionately affects the poor persons in both developing and developed countries. Mycobacteria are basically saprophytes that have evolved from different environmental conditions and further developed their pathogenicity. However, it is interesting to study how saprophytic mycobacteria have shifted as an obligate pathogen. The two main mycobacterial species that affect humans are *M. tuberculosis* and *Mycobacterium leprae* (*M. leprae*) that cause TB and leprosy respectively.

2.2 Historical background of TB

Mankind has been aware of TB for several thousands of years. TB has also been known as consumption, phthisis, scrofula, Pott’s disease and white plague. *M. prototuberculosis* was reported to be a progenitor for the emergence of *M. tuberculosis* complex. TB is thought to have evolved as a progenitor from Africa as early as 70,000 years ago (Comas et al., 2013). TB has been prevalent in Egypt as far as 5,000 years ago. Identification of TB from preserved bones and skeletal abnormalities like Pott’s deformities confirms its existence (Morse, Brothwell, & Ucko, 1964). Recent advancements in molecular techniques have identified the existence of tubercle bacilli as early as 3 million years ago which suggest that they may have infected early hominids at that time (Gutierrez et al., 2005).

In India, TB is noted as a very ancient disease. Indian literature mentioned its existence from around 1500 BCE as consumption. The disease is attributed to excessive fatigue, worries, hunger, pregnancy and chest wounds. It is also reported that the king of
Brahmanas was the first victim of this disease, referred as Rajayakshma or King’s disease. The history of TB was changed in March 1882, when Robert Koch demonstrated that the bacilli, *M. tuberculosis* is responsible for TB to the Berlin Physiological society ("[Die Atiologie der Tuberculose. Facsimile of the original contribution by Robert Koch in "Berliner Klinische Wochenschrift" 10 April 1882]," 1982). Later, in 1905 he was awarded with the Nobel Prize in medicine for his contributions in elucidation of TB.

### 2.3 Epidemiology of TB

In 2014, WHO estimated 9.6 million new TB cases and 1.5 million TB deaths (0.4 million HIV-positive people) (Fig 1). It also identified 22 high TB burden countries, of which 58% belongs to South-East Asia and Western Pacific regions. The South-East Asia region alone accounts for 38% of the global TB burden in terms of incidence ("World Health Organisation. Global Tuberculosis report 2015. WHO/HTM/TB/2015. Geneva, Switzerland," 2015). In 2013, it was estimated that about 3.4 million new cases of TB and about 440,000 people died because of TB in Bangladesh, India, Indonesia, Myanmar and Thailand. India is one among the high burden countries in the world by contributing 24% of the estimated global incident cases in 2013. Despite the low MDR-TB prevalence, India ranks first among the 27 MDR-TB high burden countries by contributing 21% of all MDR-TB cases estimated among notified cases and 3.3% of new MDR-TB cases. It is also estimated that 5.95% (95% CI: 5.93% - 5.97%) of people among 2.4 million Indians are positive for HIV-TB ("World Health Organisation. Global Tuberculosis report 2015. WHO/HTM/TB/2015. Geneva, Switzerland," 2015).

Molecular epidemiology is the field of epidemiology for investigating and monitoring the distribution of disease at the molecular level across the populations. This is achieved
by identifying the monomeric repetitive sequences in the *M. tuberculosis* genome. The two types of repetitive units, which are present in *M. tuberculosis* are interspersed repeats (direct repeats and insertion sequences-like repeats) and tandem repeats (head-to-tail direct uninterrupted repeats). Based on these repeats the three major tools, IS6110-RFLP, spoligotyping and MIRU-VNTR typing were developed to analyze the transmission dynamics (Mathema, Kurepina, Bifani, & Kreiswirth, 2006; Narayanan, 2004).

### 2.4 *Mycobacterium tuberculosis*

Mycobacteria are a genus of Actinobacteria and belong to the family of *Mycobacteriaceae*. The *Mycobacterium* genus comprises of more than 120 different species. *M. tuberculosis* complex (MTBC) consist of the following closely related organisms *M. africanum, M. bovis, M. microti, M. canetti and M. pinnipedii*. *M. tuberculosis* is a gram positive, acid-fast, aerobic, non-sporing and non-motile bacteria. They form slightly curved or straight rods with a measurement of about 0.2-0.6 µm by 1.0-10 µm. They are slow growers with a doubling time of approximately 15 to 20 h and they form rough off-white colonies on Lowenstein-Jensen medium. The bacterium has the tendency to undergo dormancy by shutting down the metabolic pathways and remains quiescent until it receives nutrients.

The *M. tuberculosis* genome comprises of a circular chromosome of about 4,411,529 bp with 3,959 coding genes and a high genomic G+C content of about 59-66% (NCBI, 2007). The genome also contains 16 copies of insertion sequence IS6110 and six copies of IS1081 (Philipp et al., 1996). The *M. tuberculosis* genes were classified into 10 functional categories namely lipid metabolism, information pathways, cell wall and cell processes, stable RNAs, IS elements and bacteriophages, regulatory proteins, PE and
PPE proteins, virulence/detoxification/adaptation, intermediary metabolism and respiration, conserved hypotheticals and the remaining comprises of proteins with unknown function (Smith, 2003). Previous studies with various databases have identified precise functions for ~40% and similarity for 44% of proteins, however the remaining 16% showed no known proteins, which could comprise the *M. tuberculosis* specific functions (Cole et al., 1998). The cell wall of *M. tuberculosis* is very unusual with an additional layer of unusual lipids and polysaccharides beyond the peptidoglycan layer (Bloom, Flynn, McDonough, Kress, & Chan, 1994). This unusual cell wall comprises of mycolic acids, mycocerosic acid, phenolthiocerol, lipoarabinomannan and arabinogalactan is biosynthesized by multiple pathways that confers resistance to the bacteria against antibiotics (Cole et al., 1998).

*M. tuberculosis* is genetically diverse throughout the world, which results in significant phenotypical differences between the clinical isolates. The major lineages of *M. tuberculosis* strains are Beijing (alternatively termed as East-Asian), Haarlem, T, East African-Indian (EAI or Indo-Oceanic) and Latin American-Mediterranean (LAM) (Gagneux et al., 2006). This diversity leads to a significant difference in immunogenicity and virulence. However, recent studies showed that there is no difference in drug resistance rate between the lineages (Gagneux, 2013). Hence, it is necessary to take all these into consideration during the vaccine designing process against TB.

### 2.5 Diagnosis for TB

Control of TB depends on rapid and accurate diagnosis to prevent drug mismanagement. TB diagnosis starts with a complete medical evaluation that includes a medical history, physical examination and microbiological examination for identifying the *M. tuberculosis* bacteria from the clinical specimens of the patient. The typical
symptom of active pulmonary TB is prolonged cough for more than three weeks, chest pain, low grade remittent fever, chills, night sweats, appetite loss, weight loss, easy fatigability and production of sputum.

The diagnosis of latent TB infection is a positive reaction to the tuberculin skin test (TST) and interferon-gamma release assay (IGRA). However, a recent policy by WHO and the European Centre for Disease Prevention and Control guidelines discourages the use of IGRA test in preference to TST in areas where TB is highly endemic ("European Centre for Disease Prevention and Control; Use of interferon-gamma release assays in support of TB diagnosis. ", 2011; "World Health Organization. Use of tuberculosis interferon- gamma release assays (IGRAs) in Low- and Middle-Income Countries: Policy Statement," 2011).

Rapid diagnosis of sputum sample by acid fast staining (Ziehl-Neelsen staining) remains the frequently used test for TB detection. However, the sensitivity of sputum smear microscopy highly varies from ~35% to ~70% due to high rate of HIV-TB co-infection (Corbett et al., 2003; Steingart et al., 2006). Hence, culturing the *M. tuberculosis* from clinical specimens using Lownstein-Jensen solid media acts as a gold standard technique for identifying the TB bacteria and also facilitates drug susceptibility testing (DST). However, the major drawback of solid media culturing is that the diagnosis takes as long as 8 weeks to declare the negative results. In 1980s, new semi-automated and automated liquid culture systems such as BACTEC-TB460 radiometric systems were introduced to detect *M. tuberculosis* with a turnaround time of about 10 days (Palomino, 2012). Then, the non-radiometric liquid techniques such as Mycobacteria growth indicator tube (MGIT) and BacT/ALERT (BioMe’rieux) with DST testing emerged (Ardito, Posteraro, Sanguinetti, Zanetti, & Fadda, 2001).
Processing of clinical TB samples has undergone remarkable improvements in the last two decades. Use of molecular tests has improved the TB diagnosis for rapid detection and mutation detection associated with anti-tuberculosis drugs (Balasingham, Davidsen, Szpinda, Frye, & Tonjum, 2009). GeneXpert, integrated nucleic acid amplification based TB diagnostic test (NAAT) is an integrated automated platform where the identification of *M. tuberculosis* and its resistance to rifampicin drug was accessed in less than 120 minutes (Ling, Flores, Riley, & Pai, 2008). The serology tests were also established for TB diagnosis like other infectious diseases (HIV, Syphilis and viral hepatitis). The recent availability of next generation sequencing (NGS) technique started revolutionizing the infectious disease diagnosis. The deep sequencing methodology in NGS is advantageous in TB diagnosis, as it provides the pattern for both known and unknown drug resistance. Hence, the NGS based routine TB diagnosis in clinical laboratories is expected in the near future (Lecuit & Eloit, 2014).

### 2.6 Vaccine for TB

Developing an active vaccine against pathogens depends on their cellular immune response. The currently available vaccine, *Mycobacterium bovis* BCG is the most widely used, but it does not show protection against adolescents and latent TB individuals (Andersen, 2007). Also, its efficacy against pulmonary TB varies geographically. Thus, there is an urgent need for an optimal vaccine design, which should take the following under consideration, (1) prevent primary infection with good memory response in adults and latent TB individuals, (2) enhance chemotherapy for drug resistant TB individuals and (3) effective against HIV-TB co-infected individuals.

MTBVAC is the first live-attenuated vaccine against *M. tuberculosis* that enters the phase I clinical evaluation by fulfilling the Geneva consensus safety requirements (Kamath et al., 2005). The MTBVAC that carries two deletions in *phoP* and *faD26*
genes has shown comparable safety in animal models. Recently, the MTBVAC *erp* mutant vaccine strain demonstrated its utility as a potential TB vaccine for use in high-risk immune suppression populations (Solans et al., 2014).

Subunit vaccines are non-live vaccines, which can be administered to the host regardless of immune competence to boost the immunological response. Subunit vaccines in TB are recombinant proteins that are administered with proper adjuvants. In addition, most subunit vaccines use either replication deficient vectors or non-living vectors to protect the HIV co-infected individuals. Some of the antigens that are used as subunit vaccines are ESAT-6, Ag85 complex and TB10.4 (da Costa, Walker, & Bonavia, 2015). Currently there are 15 vaccine candidates in various stages of the clinical trials (Fig 2).

### 2.7 Treatment for TB

The success rate of TB depends on the effective drug management. WHO recommends DOTS strategy to reduce the risk of the development of drug resistance. TB is treated with four first line drugs such as isoniazid (INH), rifampicin (RIF), ethambutol (EMB) and pyrazinamide (PZA) under DOTS strategy for first two months, then INH and RIF alone for further four months (Central TB Division, 2010). The second line drugs are used only to those individuals who become resistant to first line drugs. List of second line drugs that are used are Aminoglycosides (Kanamycin, Amikacin), polypeptides (Capreomycin) and Streptomycin.

The drug resistant strains are classified into four types (1) Monoresistant – Resistant to a single drug; (2) MDR – Resistant to INH and RIF; (3) XDR – This category includes MDR strains which are additionally resistant to any fluoroquinolones; (4) TDR – This category includes TB strains that are resistant to all first and second line TB drugs. The
recommended regimen for MDR-TB treatment includes four drugs such as fluoroquinolone, aminoglycoside, any one susceptible first line drug and an addition of one group 4 drugs (Cycloserine, para-aminosalicylic acid, terizidone and prothionamide or ethionamide) with treatment duration of 12-18 months (Zumla et al., 2015). However, research is going on to decrease the treatment period to 4 months instead of 6 months by including moxifloxacin. There are also various studies going on to identify a new effective drug against TB. The list of TB drugs that are in the developmental pipeline are mentioned in Fig 3.

### 2.8 Mycobacterium smegmatis

*M. smegmatis* is a gram positive, aerobic, fast growing acid-fast bacterial species classified under the genus *Mycobacterium*. They are 3.0-5.0 µm long bacilli shaped and appear as creamy white colonies with fine wrinkles in nutrient rich agar plates. The genome is 6,988,209 nucleotides long with high G+C content (67%) containing 6,689 coding genes. Out of which, 2000 genes are homologues to *M. tuberculosis*. Like *M. tuberculosis*, its cell wall structure is also composed of unusual mycolic acids and fatty acids (King, 2003). *M. smegmatis* is a non-pathogenic bacteria, which showed a poor pathogenicity in macrophage and mice models of infections (Jordao, Bleck, Mayorga, Griffiths, & Anes, 2008; Tyagi & Sharma, 2002). Owing to the conservation of sigma factors, two-component systems, dormancy regulons, heat shock proteins and stress response between *M. smegmatis* and *M. tuberculosis* suggests the system could serve as a model (Tyagi & Sharma, 2002). Also, the doubling time of *M. smegmatis* is as low as 2-3 h, makes it as a best model organism to study *M. tuberculosis*. It is noted that the first line TB drug, INH that inhibits *M. tuberculosis* also inhibits *M. smegmatis*. 

12
2.9 Gene regulation

Gene regulation is a cellular process that controls the genetic code of a gene to direct protein synthesis. These gene expression systems are important molecular tools in controlling the bacterial genes during infection are also useful in studying the gene functions and validation of drug targets (DeVito et al., 2002; Mnaimneh et al., 2004). Gene regulation works in two different approaches, (1) constitutive expression – where the molecular switch is always turned on regardless of environmental conditions to express the proteins constitutively at low level and (2) inducible expression – where the gene regulation is turned on and off with respect to the organisms need. The gene regulation is a complex process that involves multiple gene control elements. Some of the gene control elements are, transcriptional start site (TSS), promoter, operator, enhancers, silencers and transcription factors (TFs) (positive and negative regulators) (Raibaud & Schwartz, 1984). TSS is a site in an operon, where the initiation of transcription takes place. Promoter is a nucleotide sequence that controls a target gene to be turned on or off by binding the positive or negative regulators respectively. Operator is also a nucleotide sequence that a repressor protein binds. Enhancers are gene activators and silencers are gene repressors. Genes that code for amino acid sequences are known as structural genes and genes that code for operon regulations (regulators) are known as functional genes (Goldstein & Doi, 1995; Riethoven, 2010).

The TFs generally have two domains, DNA-binding domain and an effector domain, where the former binds to the DNA sequences and the latter regulates the activation or repression of their targeted genes. The TFs form homodimers or higher order oligomers and binds to the upstream sequences of the gene promoter or operator regions containing “-35” (TTGACA) and “-10” (TATAAT) consensus sequences (Goldstein &
Doi, 1995; Kobayashi et al., 2000). Generally the TFs bind to a 6-10 bp DNA binding motif and thereby regulate the gene expression through activators or repressors. In negative regulation, the regulatory repressor protein binds to the operator region and thereby blocks the expression of downstream genes. In positive regulation, the regulatory activator protein removes the repressor protein from the operator region and the inducer molecules binds to the regulatory activator protein and modify its conformation so that it can bind to the inducible promoter region and activates transcription.

2.9.1 Inducible expression systems for mycobacteria

Studying the gene regulation is not only important for understanding the disease pathogenesis of the bacteria, however they are also useful in designing the molecular tools to study the gene functions. Studies on molecular genetics of mycobacteria started back in 1979 with the identification of plasmids from *M. avium* (Crawford & Bates, 1979). Later, several plasmids have been identified from other mycobacterial species such as *M. scrofulaceum*, *M. chelonae* and *M. fortuitum* (Labidi, Martin, Guibourdenche, & Riou, 1984; Meissner & Falkinham, 1984). Then, in 1990 the first origin of replication, pAL5000 for *E. coli*-mycobacterium shuttle vectors emerged (Ranes, Rauzier, Lagranderie, Gheorghiu, & Gicquel, 1990). Further, the use of shuttle vectors such as pMV261 and pMF were developed with the help of a constitutively expressing promoter, *hsp60* (Fan et al., 2009; Stover et al., 1991). Since the *hsp60* is a weak constitutive promoter, demand for an inducible promoter was much needed. Hence, the currently available inducible promoters such as isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible T7 promoter, isovaleronitrile inducible pNIT-1, pristinamycin inducible pMY696, tetracyclin inducible Tet and acetamide inducible
systems were used as *E. coli*-mycobacterium shuttle vectors to express the *M. tuberculosis* genes (Parikh et al., 2013). Other than acetamide inducible system, all the other highly inducible promoters are isolated from non-mycobacterial species.

The transcription and translation mechanism in mycobacteria is different from other bacteria. Analysis of cumulated codon usage tables constructed for MTBC and *M. leprae* showed that out of 6 available codons for serine amino acid, the 3 G/C codons, UCC, UCG and AGC account for 88% of coding the serine residues and interestingly, the AGA/AGG codon for arginine has rarely been used in *M. tuberculosis* genes. Similarly, the GGC codon accounts for about half of all the occurrences of glycine. In contrast, the GGA and GGG have been classified as rare codons for *E. coli* (Ikemura, 1985; Sharp & Li, 1986).

The codon usage of *M. tuberculosis* makes it difficult to express several mycobacterial genes using other bacterial inducible systems (Andersson & Sharp, 1996). In mycobacteria, the GTG is often used as a start codon instead of the usual ATG. In addition, the upstream regions of mycobacterial genes have a high G+C content as compared with *E. coli*. Because of this high degree of codon bias, the use of non-mycobacterial systems sometimes leads to the production of non-native recombinant proteins (Leotta, Spratt, Kong, & Triccas, 2015). Henceforth, the only available mycobacterium specific inducible promoter, acetamide inducible system is the only choice for overcoming all these issues. Molecular characterization of many mycobacterial genes and assessment of drug susceptibility and sensitivity for mycobacterial pathogens were hampered by the lack of good homologous expression vectors (Parish & Stoker, 1999). Acetamidase promoter of *M. smegmatis* has been used as an inducible system for expressing *M. tuberculosis* proteins for drug screening and
vaccine studies (Daugelat et al., 2003; Raghunand, Bishai, & Chen, 2006; Triccas, Parish, Britton, & Gicquel, 1998). The disadvantage of currently available acetamide inducible promoter system is that it carries a large 1.4 kb operator, which lies 1.5 kb upstream of the promoter (Parish, Mahenthiralingam, Draper, Davis, & Colston, 1997). Also, the current acetamide inducible system is not suitable for some reasons such as (1) the higher molecular weight proteins could not be expressed and (2) leaky expression of the promoter during non-inducing conditions. Hence, isolating the respective inducible promoter with tight repression activity in the absence of inducer can efficiently use the acetamide promoter system. This can be achieved only by characterizing the gene regulatory components of the acetamidase operon such as positive regulator, negative regulator, promoter and operator.

2.9.2 Mycobacterial promoters

Mycobacteria have a complex architecture for promoters when compared to E. coli. A number of E. coli and mycobacterial promoters have been examined in detail and the transcription start site for most of the genes have been mapped, which revealed that both promoter architectures resembled that of σ70 type (Kieser, Moss, Dale, & Hopwood, 1986). Most of the mycobacterial promoter components resemble the classical bacterial promoter components, but there are lot of additions and exceptions that are unique to the expression of mycobacterial genes. A collective list of mycobacterial promoter architecture illustrates its various novel components (Newton-Foot & Gey van Pittius, 2013).

The transcription of mycobacterial promoters begins with all the four bases (A, T, G and C) as their transcription start point. Whereas, the other bacteria like E. coli initiates their transcription with only purine bases (A or G) (Bashyam, Kaushal, Dasgupta, & Tyagi, 1996). The majority of mycobacterial -10 consensus sequences resemble that of
other bacteria. There are some exceptions in mycobacteria, where the promoter transcription starts at the third base instead of the usual 5-9 base downstream to the -10 regions (Jain, Sujatha, Ojha, & Chatterji, 2005). Most of the mycobacterial promoters also have an extended -10 region characterized by TGN motif immediately upstream of the -10 region (A. Kumar et al., 1993). The lower priority for transcription, -35 consensus sequences have a variety of DNA motifs in mycobacteria and they did not resemble the classical TTGACA motif (Bashyam et al., 1996). The probable mycobacterial consensus sequences for -35 region were T (92%), T (53%), G (62%), A (44%), C (60%), G (31%)/A (26%) and -10 region are T (68%), A (76%), T (41%), A (36%)/G (26%), A (34%)/C (28%), T (76%) (Mulder, Zappe, & Steyn, 1997). As there is no homolog for -35 consensus was maintained between mycobacteria, the greater promoter activity was observed with TTGCGA DNA motif (Agarwal & Tyagi, 2006). Hence, these differential residues contribute to the promoter activity.

The common universal start codon for transcription is ATG. In mycobacteria, due to high G+C content, along with ATG, GTG also often occur as start codon. The possibilities of other bases as start codons are also not avoided. Because, in *M. tuberculosis*, the TTG confers 4.7% of start codons and ATG, GTG confers 60.8% and 33.5% respectively. Some examples of start codons in *M. tuberculosis* are ATC, GCA, TGG, GGA, GGC, CCC, ATA, TCT, GAT, CTT, ATT, TCC, GAC, GTC and CCA (Newton-Foot & Gey van Pittius, 2013). The promoter activity is also contributed by the macroscopic structure of promoter DNA. The *E. coli* $\sigma^{70}$ type mycobacterial promoters exhibit medium to high curvature profiles suggesting they are strong promoters and non *E. coli* $\sigma^{70}$ type mycobacterial promoters that exhibit low curvature profiles had weak promoter activity (Kalate, Kulkarni, & Nagaraja, 2002).
Mycobacteria are also unique in expressing the proteins with multiple promoters, whereas the other bacterial systems express proteins from single promoter. These multiple promoters are regulated by differential induction of TFs to vary the strength of their promoter levels from basal to peak expression, which depends on the environmental conditions. Some of the multiple promoter regulations in *M. tuberculosis* are *recA* and DNA gyrase promoters and in *M. smegmatis* is acetamidase promoter (Gopaul, Brooks, Prost, & Davis, 2003; Narayanan, Selvakumar, Aarati, Vasan, & Narayanan, 2000; Unniraman, Chatterji, & Nagaraja, 2002).

### 2.9.3 Acetamidase operon

The amide hydrolase activity of cell free extract of *M. smegmatis* and subsequently, the presence of a highly inducible enzyme that enables the organism to utilize several amide compounds as sole carbon source was first demonstrated by Draper in 1967 and the enzyme was named as acetamidase (Draper, 1967). It is expressed at basal level in non-induced conditions or when bacteria grown in nutrient rich medium. This enzyme can be induced hundred fold in the presence of simple chemical inducers like acetamide that resulted in an increase of acetamidase protein concentration up to 10% of total proteins (Mahenthiralingam, Draper, Davis, & Colston, 1993; Parish et al., 1997). Recent identification of acetamide and formamide near the center of the Milky Way galaxy suggests the importance of amides in the formation of universe (Hollis, 2006).

Acetamide is an organic aliphatic amide with the formula CH$_3$CONH$_2$. Acetamide is an intermediate product in the enzymatic hydrolysis of acetonitrile (DiGeronimo & Antoine, 1976). Various microorganisms have the capacity to hydrolyze and/or utilize acetamide as a source of carbon and/or nitrogen. While the acetate molecule leads to tricarboxylic acid-cycle (TCA) intermediates, the NH$_3^+$ liberated is taken into the cells and used in the urea cycle. Many bacteria including *M. tuberculosis*, *P. aeruginosa* and
fungi expresses amidase. Amidase operon of *P. aeruginosa* is well characterized and shares similarities with the acetamidase operon of *M. smegmatis* (Norman, Poh, Pearl, O'Hara, & Drew, 2000; S. A. Wilson, Wachira, Norman, Pearl, & Drew, 1996). In *M. smegmatis*, the acetamidase operon genes are organized as *amiADSE* and *amiC* is divergently transcribed (Narayanan et al., 2000; Parish et al., 1997). Although there are similarities in the DNA/protein sequence and their function, the regulation of amidase operon in *P. aeruginosa* is entirely different compared to acetamidase operon of *M. smegmatis*. *P. aeruginosa* amidase operon has *amiEBCRS* genes in which *amiE, amiC* and *amiS* share considerable functional and structural identity and similarity with their counterparts in *M. smegmatis* operon (Parish et al., 1997). The acetamidase operon of *M. smegmatis* is regulated by four structural genes *amiC, amiA, amiD, amiS* and four promoters viz., Pc, P1, P2 and P3 (Mahenthiralingam et al., 1993).

### 2.10 Protein interactions

The availability of human genome project has increased more complications on the human proteome. Though the genome stores all the genetic information, the proteome occupies the larger portion than the genome in a cell. Proteins are complex molecules that regulate the life processes. The classical paradigm of molecular biology describes how DNA makes RNA makes proteins. Proteins differ among themselves by varying the 20 amino acids in their polypeptide chain. They substitute side chain amino acids uniquely to converse different physical, chemical, biological and structural properties.

In 1838, Berzelius proposed the term “Protein” in a letter to Dutch chemist, Mulder. Soon after that the enzyme activities of proteins as catalysts were discovered in 1833. The very first PPI was identified between the enzyme trypsin and its inhibitor in 1906. Later in 20th century, investigations of more and more physical interactions between the
proteins were uncovered and their importance was characterized (Braun & Gingras, 2012).

Proteins fold in three dimensions with the hierarchy from primary to quaternary structure through variety of bondings among them. The primary structure of proteins represents the amino acids sequence or peptide. The secondary structure of protein is formed by hydrogen bonding between amino groups and carboxyl groups, the α–helix and β–sheet. The disulfide bridges, salt bridges and ionic interactions between the polypeptide form the tertiary structure of proteins. Finally, the quaternary structure of a protein refers the macromolecules with multiple polypeptide subunits. Molecular chaperones or heat shock proteins prevent aggregation of several proteins in the cytoplasm by properly folding them during their macromolecule assembly. In mycobacteria, chaperoning proteins like GroEL, GroES and HspX are helpful in proper folding (D. B. Young & Garbe, 1991). Proteins (polypeptides) are also prone to be modified or processed during their synthesis from the ribosomes (co-translational) or after their synthesis (post-translational) (Prabakaran, Lippens, Steen, & Gunawardena, 2012; Saraogi & Shan, 2014). Prokaryotes and eukaryotes use several co-translational translocation pathways to transport the proteins to the cell membrane or endoplasmic reticulum. However, proteins use post-translational modifications (PTMs) to functionally regulate the diverse range of cellular processes. The major PTMs include phosphorylation, acetylation, pupylation, methylation, carboxylation, proteolysis, glycosylation, ubiquitination, nitrosylation, lipidation, phosphoform modifications, ribosylation, deamination and non-enzymatic PTMs of cysteines plays a key role in functional proteomics (Cain, Solis, & Cordwell, 2014).
Protein interactions play a fundamental biological role in all the living organisms. The majority of proteins do not function in isolation (Berggard, Linse, & James, 2007). However, they execute their functions by various physical interactions such as protein-DNA interaction, protein-RNA interaction, protein-protein interaction, protein-ligand interaction, protein-lipid interaction and protein-metabolite interaction. These interactions vary in many aspects based on the cellular necessity. But the real physiological importance of these interactions like who, what, where, when and why needs to be answered.

2.10.1 Protein-DNA interactions

Proteins interact with nucleic acids to regulate the biological processes in the cell. Understanding how the proteins recognize their DNA binding sites in a long strand of sequence containing A, T, G and C is important to study their regulations in disease conditions. Proteins preferentially bind to the major groove of the DNA, however in some cases it binds to minor groove, since the major groove is wider and can accommodate large motifs (Bewley, Gronenborn, & Clore, 1998). It was believed that the proteins recognize the short DNA sequences, hydrogen bond donors, acceptors and nonpolar groups in the DNA using the complementary donor and acceptor amino acid side chain residues in the protein (Seeman, Rosenberg, & Rich, 1976; Viswamitra et al., 1978).

Proteins act as TFs and interact with DNA or RNA to regulate gene expressions. These TFs are DNA binding proteins that regulate several cellular processes such as replication, transcription, translation, DNA repair/recombination and development (Hegarat, Francois, & Praseuth, 2008). TFs interact with DNA both specifically and
non-specifically. Specific protein-DNA interactions are strong interactions, which use α-helices to bind to the specific sequences using hydrogen bonding, ionic interactions, hydrophobic, van der Waals and salt bridges. The common DNA binding motifs observed in prokaryotes and eukaryotes are helix-turn-helix (HTH) motif, zinc finger motif and leucine zipper motif. In case of non-specific interactions, the protein does not prefer the sequence conservation and interacts with the sugar-phosphate backbone of DNA (Luscombe & Thornton, 2002).

2.10.1.1 Methods to study protein-DNA interaction

The protein-DNA interactions can be studied by high-throughput *in vitro* and *in vivo* techniques. Though various protein-DNA interaction techniques with good specificity like electrophoretic mobility shift assay (EMSA), DNase I footprinting assay, chromatin immunoprecipitation (ChIP), X-ray crystallography, protein binding microarrays (PBM)s and one-hybrid systems are available, they are highly laborious and time consuming (Stormo & Zhao, 2010). However, few other techniques with moderate sensitivity like mechanically induced trapping of molecular interactions (MITOMI), Bind-n-seq, SELEX/SELEX-seq, microarray evaluation of genomic aptamers by shift, cognate site identifier (CSI), HT sequencing fluorescent ligand interaction profiling (HiTS-FLIP) are also available to facilitate the protein-DNA interactions (Siggers & Gordan, 2014).

2.10.2 Protein-protein interactions

Protein-protein interactions (PPIs) refer to the physical contact between the active site regions of two or more proteins. The protein interactions occur in many ways like strong or weak and they exist as stable or transient. The two subunits of same proteins
interact together to form a homodimer and two subunits of different proteins interact together to form heterodimer. The homodimers are also called as oligomers, which have non-covalent bonding to form as dimers, trimers and tetramers. Non-covalent bonding also occurs in heterodimers when two different proteins interact together to execute a common cellular function. The interface amino acid residues between the two proteins determine the type of interaction. However, these “hot spots” residues in a protein were structurally conserved, whereas the other residues are not evolutionarily conserved (Ma, Elkayam, Wolfson, & Nussinov, 2003).

2.10.2.1 Methods to study PPIs

The major drawback in characterizing the PPIs depends on the type of method that is used to study the interactions. Though multitudes of interaction identification techniques are available, each of them has their own strength and weakness. Hence, combinations of techniques are needed to characterize the PPIs. The protein interactions can be identified by various in silico, in vivo and in vitro techniques (Rao, Srinivas, Sujini, & Kumar, 2014).

The computational analysis of PPIs identifies the interaction networks in the mathematical form consisting of nodes and edges (Wagner, 2003). Molecular docking is a modeling based in silico technique that predicts the near native three-dimensional protein-protein interaction structures and ranks them using the algorithm specific scoring functions. Molecular docking coupled with alanine scanning mutagenesis is useful in identifying the residue level interaction between the two proteins (Gray et al., 2003; Kortemme, Kim, & Baker, 2004).
The *in vitro* PPIs are helpful in studying both stable and transient interactions. Several methods to analyze *in vitro* biochemical PPIs are pull down assay, far-western blotting (FWB), co-immunoprecipitation, crosslinking protein interaction analysis, label transfer protein interaction analysis, X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and a newer technique, protein array based surface enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI TOF MS) (Howell, Winstone, Coorssen, & Turner, 2006; Phizicky & Fields, 1995; Tong et al., 2001). Several biophysical techniques employed are isothermal titration calorimetry, fluorescence anisotrophy, dual polarization interferometry, surface plasmon resonance (SPR), static light scattering and circular dichroism (Stynen, Tournu, Tavernier, & Van Dijck, 2012).

The identified PPIs should be validated by an *in vivo* assay to know their existence inside the cell. The only available *in vivo* protein interaction technique is two-hybrid system. The bacterial two-hybrid system detects the interaction by utilization of transcriptional activation protein from unfused to the fused state (Singh, Mai, Kumar, & Steyn, 2006). However, the other widely used techniques are co-localization, atomic force microscopy, electron microscopy, bioluminescence resonance energy transfer (BRET), fluorescence resonance energy transfer (FRET), bimolecular fluorescence complementation (BiFC), requires extensive instrumentation (Lleres, Swift, & Lamond, 2007; Shoemaker & Panchenko, 2007).

### 2.11 Genetic manipulations

Genes play an important role in determining the function of the bacteria. Functional genomics is the study of each and every individual gene in the genome of living organisms. They are important in understanding the complex relationship between
genotype and phenotype by focusing on gene transcription, translation and PPI. However, these are achieved by genetic manipulations (gene modifications) of the organism’s genome by various techniques such as “knockout”, “knock in” and “knock down”. While gene knock out involves the complete deletion of a target gene and gene knock down involves the partial inhibition of the target gene, the gene knock in experiments involve the targeted insertion of a particular gene into the genome of the bacteria. In 2007, Mario Capecchi, Martin Evans and Oliver Smithies were awarded the Nobel Prize for their invention of genetic modifications in embryonic stem cells. In mycobacteria, genetic engineering has been commonly used for more than a decade to isolate the attenuated strains and identify the virulence factors. The whole genome screening of *M. tuberculosis* with transposon mutagenesis identifies that 600 genes out of 4000 (~15%) are essential for cell growth (Sassetti, Boyd, & Rubin, 2003).

2.11.1 Genetic tools for *M. tuberculosis* genome

Genetic manipulation systems are problematic in *M. tuberculosis* as compared to other bacteria because of several factors like pathogenesis, slow growth rate and poor DNA uptake mechanisms (van Kessel, Marinelli, & Hatfull, 2008). The commonly used stable genome editing techniques in mycobacteria are transposon mutagenesis, allelic exchange substrate and antisense RNA mediated depletion (Balasubramanian et al., 1996; Kaur, Agarwal, & Datta, 2009; McAdam et al., 2002; Pelicic et al., 1997). The other genome editing techniques such as site directed mutagenesis and signature tagged mutagenesis are also used accordingly (Camacho, Ensergueix, Perez, Gicquel, & Guilhot, 1999; Saint-Joanis et al., 1999).

2.11.1.1 Transposon mutagenesis

Transposon mutagenesis is one of the widely used techniques to understand the virulence factors of the bacteria. This technique mutates the non-essential genes
randomly in the bacterial chromosomes using transposons. In mycobacteria, the very first insertion sequence (IS) IS1096 was isolated from *M. smegmatis* and delivered in *M. bovis* BCG using suicide vector delivery system (McAdam et al., 1995). However, the suicide vector delivery systems were not suitable for mutant library constructions because of their low transposon frequencies. The present *M. tuberculosis* transposon library that was created with Tn5370 has showed 1474 total insertions, which disturbed 351 genes in the genome (McAdam et al., 2002). The other available transposition systems are Tn552, Himar1 and MycoMarT7 (Lamrabet & Drancourt, 2012). Though it is easy to isolate mutants through transposon mutagenesis, their inaccuracy in transposition, low frequency of transposition in living systems and difficulty in mutant screening has made the system to be less operational.

### 2.11.1.2 Allelic exchange

Gene knockout or allelic exchange is a successful technique to mutate the genes in *M. tuberculosis*. In *M. tuberculosis*, this allelic exchange technique is widely used for both short (< 5 kb) and long linear (40-50 kb) sequences (Balasubramanian et al., 1996; Piddington et al., 2001). This technique has shown low integration of genes in eukaryotes, however because of slow growth, *M. tuberculosis* possess a high level of illegitimate recombinations accompanied by a decreased level of homologous recombination (Kalpana, Bloom, & Jacobs, 1991). This technique involves homologous recombination of the chromosomal target gene with the inactive target copy from the vector. This homologous recombination was accomplished earlier by suicide vector systems. However, due to its low efficiency, this allelic exchange is accomplished by phage based plasmid delivery system of *sacB* gene with counter selection markers hygromycin or kanamycin resistance cassette and growth on sucrose media (Sirakova,
The other allelic exchange method available for isolating unmarked mutants in *M. tuberculosis* is pMIL/pGOAL system. Allelic exchange is much more superior as compared to transposon based mutagenesis, because they undergo targeted gene deletion with double cross over homologous recombination. Finally, the screening of knock out mutants was based on antibiotic resistance, colony PCR, southern blotting and western blotting using both allelic exchange and transposon mutagenesis methods.

### 2.11.1.3 Antisense mediated targeting

Gene knock down or antisense RNA targets the sequence specific chromosomal mRNA leading to the reduced expression of the target gene (Brantl, 2002; Good, 2003). This plasmid based antisense inhibition of mRNA molecules was first reported in *E. coli* in 1984 (Pestka, Daugherty, Jung, Hotta, & Pestka, 1984). Since, the knockout systems totally inactivates the target genes, they are often not suitable for studying the essential genes in the living organisms. Hence, the alternate methods to study the essential genes are RNA silencing in eukaryotes and antisense gene knockdown in prokaryotes. This antisense mRNA targeting method is less laborious as compared to gene knock out methods. This technique involves the cloning of antisense sequence of a target gene under the inducible promoter. The transcript level is maintained according to the inducer concentration. Generally, the extra chromosomal plasmid synthesized by antisense mRNA molecules anneal to a complementary mRNA, which results in the degradation of duplex RNA, hybrid RNA/DNA duplex or target RNA of antisense compound itself and inhibits the protein synthesis. This binding of sense and antisense RNA molecules occur in cytosol of the cell and the complete degradation of mRNA occurs in 2-3 minutes (Fig 4). This strategy has been applied in different bacteria such as *M. tuberculosis* (*ftsZ*) (Kaur et al., 2009), *M. smegmatis* (*hisD*) (Parish & Stoker,
2.12 Signal transduction

*M. tuberculosis* is a very successful pathogen, which adapts to the myriad stress conditions during infection. Upon infecting the host, the tubercle bacterium is exposed to several host induced stresses such as reactive oxygen species (ROS) production, reactive nitrogen species (RNS) production, low pH, hypoxia, starvation for essential nutrients and anti-microbial agents. To overcome these stresses, the bacterium activates different TFs upon receiving extracellular signals through various signal transduction pathways. Signal transduction involves sensing the signal from the transmembrane and converting it into a response or output that alters the cellular physiology. Sensor is the domain of the protein that senses the signal and transducer is the domain of the protein or molecule that converts the signal into a response. *M. tuberculosis* signal transduction is conducted by multiple families such as two-component regulatory systems (TCSs), STPKs and tyrosine protein kinases/phosphatases (Av-Gay & Everett, 2000; Bach, Wong, & Av-Gay, 2009; Stock, Robinson, & Goudreau, 2000).

2.12.1 Two component systems

TCSs that enable the living organisms to regulate gene expression in response to several environmental stresses are widely distributed among prokaryotes and eukaryotes (Cole et al., 1998). The *M. tuberculosis* genome encodes 11 complete TCSs and several orphan kinases and regulators that are involved in regulation of virulence, resistance and persistence following infection (Hoch, 2000). They regulate the TFs by interacting with their histidine kinase (HK) and a response regulator (RR). HKs are membrane
proteins that sense the environmental stimulus and transmit the signal with its kinase activity. RRs are cytosolic proteins, which get phosphorylated on its aspartate residue by HKs. RRs are two domain proteins, which receive the phosphorylation through receiver domain and further activate the target genes through effector domain. However, the mechanisms by which the sensory kinases establish the signal transduction process remain unclear. Although all the TCSs are directly or indirectly involved in virulence of the bacterium, one specific TCS mtrA-mtrB was found to be essential for cellular growth (Zahrt & Deretic, 2000). The other characterized TCSs are involved in several functions including mprA-mprB for maintenance of persistence (Zahrt & Deretic, 2001), prrA-prrB for intramacrophage growth (Ewann, Locht, & Supply, 2004) and devR-devS for hypoxic response regulation (Park et al., 2003).

2.12.2 Serine/Threonine protein kinase

Unlike TCSs, STPKs are less widely distributed among the bacteria. The M. tuberculosis STPKs are described as “eukaryotic-like” protein kinases based on their sequence homology to eukaryotic STPKs (Av-Gay & Everett, 2000). The STPK phosphorylation and tyrosine phosphorylation have been identified in several prokaryotes such as Streptomyces, Anabaena, Myxococcus, Yersinia pseudotuberculosis and P. aeruginosa (Chaba, Raje, & Chakraborti, 2002; Peirs, De Wit, Braibant, Huygen, & Content, 1997). M. tuberculosis genome encodes 11 STPKs (PknA to PknL, except C) and a Ser/Thr specific phosphatase in which two STPKs, PknG and PknK do not have an extracellular transmembrane domain in their structure. Along with STPKs, it also has a protein tyrosine kinase, PtkA and a pair of protein tyrosine phosphatases, PtpA and PtpB (Bach et al., 2009). In contrast to TCSs, protein kinases signal with one protein through extracellular transmembrane sensor domain and an intracellular kinase domain. These protein kinases get autophosphorylated by ATPase activity on their
serine, threonine or tyrosine residues. The activated kinases additionally phosphorylate their substrate proteins through phosphorylation by adding phosphate groups. Further, their specific phosphatase proteins dephosphorylate the kinase proteins. Protein kinases are shown to regulate different processes such as development, cell wall processes, stress responses, pathogenicity and metabolic processes (Av-Gay & Everett, 2000; J. Chao et al., 2010). However, in recent years a wide range of substrate proteins of STPKs were identified (J. Chao et al., 2010) (Fig 5).

*pknA* and *pknB* are adjacent genes that reside in the same operon are known to be essential for the cell growth. These two kinase proteins act as master regulators in regulating the cell wall biogenesis of *M. tuberculosis*. Structural analysis of extracellular region of PknB protein showed the presence of four penicillin binding protein and serine/threonine kinase associated (PASTA) domains (Barthe, Mukamolova, Roumestand, & Cohen-Gonsaud, 2010). These PASTA domains are shown to bind extracellular muropeptides, by which they resuscitate the dormant *M. tuberculosis* (Mir et al., 2011). This was supported by the downregulation of *pknA* and *pknB* during nutrient starvation conditions (Betts, Lukey, Robb, McAdam, & Duncan, 2002). PknA and PknB are important in maintaining the cellular morphology of the bacteria. Their differential expression leads to marked effects on cell shape including branching, elongation and incomplete septation (Kang et al., 2005). In addition to cell wall synthesis, PknA and PknB have been shown to regulate several additional cellular processes including lipid synthesis and transcriptional regulation (Larsen et al., 2002).

*pknD* is a non-essential gene which co-transcribes with a phosphate transport system regulated gene *pstS*. Knockdown of *pknD* showed compromising survival in a phosphate deficient growth medium suggesting the role of PknD in phosphate uptake (Vanzembergh et al., 2010). Structural analysis of PknD showed the presence of six
imperfect tandem repeats homologous to β-propeller structures in its carboxy-terminal (Springer, 1998). Other than phosphate uptake, \textit{pknD} was also shown to be important for central nervous system TB invasion (Be, Bishai, & Jain, 2012).

\textit{pknE} was classified under the family of integral membrane proteins and cytosolic kinases due to the presence of adjacent transporter genes (Narayan et al., 2007). Structural analysis of PknE revealed the presence of CXXC motif in its C-terminal responsible for redox activity (Gay, Ng, & Alber, 2006). PknE phosphorylates type II fatty acid synthesis system which suggests its role in lipid biogenesis (Molle, Brown, Besra, Cozzone, & Kremer, 2006). In addition, PknE was also shown to regulate apoptosis upon \textit{M. tuberculosis} infection with reduced intracellular survival (D. Kumar & Narayanan, 2012).

\textit{pknF} is present in an operon adjacent to ATP-binding cassette (ABC) transporter gene that contains two forkhead associated (FHA) domains suggesting the possible role of PknF in transport system (Av-Gay & Everett, 2000). This is supported by the physical interaction and phosphorylation of PknF and the neighbor transporter gene Rv1747 (Curry et al., 2005). PknF phosphorylates Rv1747 at two sites, Thr150 and Thr208 that is between the two FHA domains of this protein (Spivey et al., 2011).

\textit{pknG} is encoded in an operon containing a glutamine binding lipoprotein gene and a conserved membrane protein of unknown function (Camus, Pryor, Medigue, & Cole, 2002). PknG is a soluble protein without a transmembrane region in its structure. It is a two-domain protein with a rubredoxin redox-sensing domain in the N-terminal and tetratricopeptide repeat domain in its C-terminal. However, they execute their kinase
activity with the help of amino terminal tetratricopeptide repeat domain (Tiwari et al., 2009). *pknG* is an essential gene, deletion of it results in decreased expression of *de novo* glutamine synthesis. Also, PknG was identified as a substrate for FHA domain containing protein GarA, which is responsible for central carbon and nitrogen metabolism (Ventura et al., 2013). Additionally the role of *pknG* in *M. tuberculosis* virulence was also noted by its capacity to modulate phagosome-lysosome fusion (Walburger et al., 2004).

*pknH* is encoded in the downstream of an operon containing the ABC-transporter type transcriptional regulator *embR*. *embR* along with *embA* and *embB* genes encoding cell wall arabinosyltransferases are one of the targets of first line TB drug, ethambutol (Belanger et al., 1996). Structural studies with transmembrane region of PknH protein identify the presence of two disulfide bonds in its active site with a mixture of hydrophobic and polar residues (Cavazos, Prigozhin, & Alber, 2012). The physical proximity between these two genes identifies them as their substrates by phosphorylating the EmbR in its FHA domain (Molle et al., 2003). Deletions of *pknH* showed its role in regulating the cell envelop glycolipids of *M. tuberculosis* (Gomez-Velasco et al., 2013). Also, PknH modulates the dormancy response by phosphorylating the dormancy regulon DosR (J. D. Chao et al., 2010).

*pknI* is a part of cell division gene cluster that includes *dacB2*, which encodes D-alanyl-D-alanyl hydrolase, D-amino acid aminohydrolase and signal recognition pathway components *ffh* and *ftsY* (Av-Gay & Everett, 2000) (Fig 6). PknI shows a distinct phenomenon, when compared with other mycobacterial kinases. Cellular fractionation revealed that PknI localized predominantly in the cytosolic fraction despite the presence
of a transmembrane domain (Singh, Singh, et al., 2006). PknI is the only protein kinase in mycobacteria, whose active site contains an asparagine residue instead of usual lysine in the catalytic loop. Loss of PknI results in hypervirulence phenotype in human THP-1 macrophage like cells and SCID mice model (Gopalaswamy et al., 2009).

\( pknJ \) is present in an operon containing several transposon genes, conserved hypothetical proteins and dipeptidase protein. However, little is known about the role of PknJ in \( M.\) \( \text{tuberculosis} \). \( In\) \( \text{vitro} \) phosphorylation assays showed its autophosphorylation activity on three Thr residues (Jang et al., 2010).

PknK is a high molecular weight protein with a kinase domain in N-terminal and an ATP-dependent transcriptional regulator homologue in C-terminal. Within the C-terminal, it has a P-loop ATP binding motif characteristic AAA+ ATPase, a PDZ domain and a single tetratricopeptide repeat sequence (Marchler-Bauer et al., 2011). PknK was shown to phosphorylate VirS, a transcriptional regulator and thereby regulate the VirS activity (P. Kumar et al., 2009). Deletion of \( pknK \) showed increased resistance to acidic, oxidative and hypoxic stress conditions which suggests its role in translation in relation to growth and environmental conditions (Malhotra, Okon, & Clark-Curtiss, 2012).

\( pknL \) is also encoded by genes such as transposase, putative DNA binding protein, biosynthetic enzymes for aromatic amino acids, mannan and lipid biosynthesis. PknL autophosphorylates on two Thr residues in its activation loop and also phosphorylate its adjacent transcriptional factor gene Rv2175c (Canova et al., 2008). However, little is known about the function of PknL in \( M.\) \( \text{tuberculosis} \).
2.13 Protein export pathways

Proteins that are synthesized in the cytoplasm are transported to the cell membrane or outside the membrane (secretory proteins) using essential protein export systems, which is a fundamental process to maintain the proper cellular function. The membrane proteins are secreted with a unique N-terminal signal peptide sequence, that is recognized by the protein export system components and translocate to their respective locations in the membrane. *M. tuberculosis* largely depends on these protein export systems to transport the antigenic proteins, thereby maintaining the virulence of the bacteria. *M. tuberculosis* encodes several bacterially conserved protein transport pathways such as general secretion (Sec) pathway, twin-arginine translocation (Tat) pathway, ESAT-6 secretion system (Esx) pathway and signal recognition particle (SRP) pathway (Feltcher, Sullivan, & Braunstein, 2010). However, the bacterium uses Sec and Tat pathways to transport their majority of the proteins as compared to Esx and SRP pathways. The SRP pathway differs from the other three pathways by transporting the proteins during translation and is also termed as co-translational targeting.

2.13.1 Secretory pathway

Understanding protein export systems in building the cell envelope and secreting the proteins to the extracellular environment was recognized in early 1970s (Beckwith & Silhavy, 1983). The Sec pathway is a post-translational process that specifically exports the unfolded proteins. The *E. coli* Sec pathway was well established by studying their essential secretion components through various genetic and biochemical studies (Danese & Silhavy, 1998). The bacterial Sec translocase pathway is present in cytoplasmic membrane and it consists of an ATP dependent motor protein SecA along with three integral membrane proteins SecY, SecE and SecG (de Keyzer, van der Does,
The other components that assist in Sec transport are SecD, SecF and YajC. The Sec translocase pathway transports the newly synthesized proteins with an N-terminal signal peptide sequence in their unfolded state by SecA and transports them through the transmembrane channel formed by SecYEG protein complex and delivers to the extracytoplasmic environment with the help of ATP and proton motive-forces (PMFs). Other than N-terminal signal peptide proteins, the Sec translocon also transports lipoproteins that contain lipobox motif in the C-terminal region of the signal peptide. Once the proteins are transported to their respective locations, the signal peptide is removed by the peptidases such as Type I signal peptidase, LepB and lipoprotein Type II signal peptidase, LspA (Paetzel, Karla, Strynadka, & Dalbey, 2002).

*M. tuberculosis* Sec pathway components are highly conserved like their counterparts in other bacteria. However, it contains two SecA components as compared to other bacteria. But the precise role of having two Sec components is interesting. SecA1 is an essential component responsible for cargo transport whereas SecA2 is an accessory non-essential component of SecA (Braunstein, Brown, Kurtz, & Jacobs, 2001). Both SecA1 and SecA2 are not functionally redundant. Though both the components are ATPases with ATP binding activity, the SecA1 is present in both the cytosol and membrane fractions, the SecA2 is predominantly present only in the cytosol of the bacteria (Rigel et al., 2009). Interestingly, SecA2 recognize the newly synthesized substrate proteins without an active signal peptide sequence in *M. tuberculosis* and proteins with signal peptide sequence in *M. smegmatis* (Feltcher et al., 2010). Hence, identifying the specific proteins that are transported by SecA1 and SecA2 components is important to better understand their process in protein export.
2.13.2 Twin arginine pathway

Tat is a Sec independent pathway, which is predominantly present in cytosol and uses two important arginine residues was discovered in early 1990s (Chaddock et al., 1995; Cline, Ettinger, & Theg, 1992). Unlike Sec pathway, Tat pathway is not present in all the bacteria and they transport only prefolded proteins to the cell membrane. The Tat translocon consists of two or three core membrane components such as TatA and TatC or TatA, TatB and TatC (Natale, Bruser, & Driessen, 2008). These Tat components form a TatBC complex in the cytoplasmic membrane and bind to the prefolded proteins that contain signal peptide sequences with their cofactors before transporting them to the membrane. After the protein transport to the membrane, the signal peptide is removed by Type I signal peptidase or Type II signal peptidases (Gimenez, Dilks, & Pohlschroder, 2007; Luke, Handford, Palmer, & Sargent, 2009). It is noted that the Tat translocon substrates are also transported by Sec pathway when the substrate proteins are fused with Sec signal peptide sequences (Ize, Stanley, Buchanan, & Palmer, 2003).

Mycobacteria also show conserved Tat components like other bacterial species. These Tat components tatA, tatB and tatC are essential for cellular growth in *M. tuberculosis*, however they are not essential in *M. smegmatis* (Posey, Shinnick, & Quinn, 2006; Saint-Joanis et al., 2006). An *in silico* study identified a total of 108 proteins with Tat signal peptides in *M. tuberculosis* suggesting they are the substrates for Tat dependent protein export (McDonough et al., 2008). Some of these predicted substrates are classified under pathogenesis or essential physiologic process group of proteins.

2.13.3 ESX export pathway

ESX secretion system is the 6 kDa early secreted antigenic target ESAT-6 based protein export system which was first discovered in *M. tuberculosis*. The bacterium has 5 ESX export systems (ESX-1 to ESX-5) that transport antigenic proteins and are classified as
Type VII secretion system (Abdallah et al., 2007). They transport antigenic proteins that are not transported by Sec and Tat pathways. The ESX systems are composed of two proteins that comprise the secretion machinery. However, ESX-1 system is best characterized as compared to other ESX systems.

ESX-1 system is encoded by two proteins viz., EsxA (ESAT-6) and EsxB (CFP-10) are secreted as a heterodimer (Brodin et al., 2005). ESAT-6 and CFP-10 form a heterodimer and the complex is needed for CFP10 secretion. Also, this ESAT-6/CFP-10 complex attracts the secretory or C-terminal signal peptide containing proteins and translocates them to the cell membrane with the help of ATP hydrolysis. For this, the two cytoplasmic AAA ATPases EccA1 and EccCb1 proteins act as ATP donors for the ESX-1 system (Luthra, Mahmood, Arora, & Ramachandran, 2008). However, the mechanism behind how ESX-1 system transports the substrate proteins to the outer membrane is not yet known. Proteins that are secreted by ESX-1 system are EspR, EspA, EspB and EspC (Feltcher et al., 2010). The ESX-2, ESX-4 and ESX-5 systems are not essential for in vitro growth or virulence of the M. tuberculosis (Sassetti et al., 2003). Both these systems are not yet studied in mycobacterial species. However, ESX-3 is an essential transport system, which is shown to be important for iron-mediated acquisition (Serafini, Boldrin, Palu, & Manganelli, 2009; Siegrist et al., 2009). Hence, complete understanding of the proteins that are transported by this pathway will be useful for designing potential anti-TB targets.

**2.13.4 Signal recognition particle pathway**

Unlike Sec and Tat pathways, SRP pathway operates in co-translational fashion. However, the posttranslational mechanism of SRP targeting is not excluded. Protein targeting by SRP is universally conserved among all the living organisms. In
prokaryotes and eukaryotes, SRP pathway transports a group of hydrophobic N-terminal signal peptide containing proteins to the plasma membrane and endoplasmic reticulum respectively. The universal SRP pathway consists of two proteins and an RNA molecule. The components are fifty four homolog or SRP, SRP receptor (SR) and an RNA molecule.

In eukaryotes, SRP pathway comprises of six proteins (SRP72, SRP68, SRP54, SRP19, SRP14 and SRP9) and a 7SL RNA molecule. SRP consists of two domains such as Alu domain and S domain (Gundelfinger, Krause, Melli, & Dobberstein, 1983). Alu domain comprises the SRP9 and SRP14, whereas the S domain comprises the SRP72, SRP68, SRP54 and SRP19 (Weichenrieder, Wild, Strub, & Cusack, 2000). The assembly of SRP constitutes the binding of SRP components with SRP RNA and transport to the cytoplasm. Further, binding of SRP54 to the above complex completes its assembly (Grosshans, Deinert, Hurt, & Simos, 2001). The SRP receptor that is present in the membrane comprises of two proteins, SRα and SRβ (Tajima, Lauffer, Rath, & Walter, 1986). Both SRα and SRβ are GTP binding proteins in which SRα is an essential component and SRβ is a nonessential component. In eukaryotes, SRP forms a core that includes SRP54, SRP RNA and SRα and further transports the substrate proteins to the endoplasmic reticulum.

In bacteria, *E. coli* SRP system is well studied is shown to be essential for cellular growth. *E. coli* SRP system consists of eukaryotic SRP54 homologue Ffh (fifty four homolog), SR receptor FtsY and a 4.5S RNA molecule. The *M. tuberculosis* genome also encodes SRP pathway that has SRP, SR and 4.5S RNA molecule. Both these SRP and SR are GTP binding proteins that activate themselves mutually so that no external GTP activators are needed (Legate, Falcone, & Andrews, 2000). Transportation of substrates by SRP pathway is a multi-step process in which the SRP (Ffh) binds with domain IV of the 4.5S RNA molecule. Next, this Ffh-4.5sRNA complex binds to the
substrate proteins from ribosomes that contain specific signal peptide sequence. Now, the SRP substrate complex binds with SR (FtsY) in the membrane and delivers the substrate protein into the cell membrane. Later, the SRP/SR complex gets dissociated due to their mutual GTP hydrolysis activity and the next round of SRP transport takes place (Connolly, Rapiejko, & Gilmore, 1991) (Fig 7).

Both the Ffh and FtsY are multidomain proteins with GTPase domain. Ffh has three domains with N-terminal NG domain and C-terminal M domain. N terminal comprises of four $\alpha$ helices, G domain comprises the GTP binding site and M domain contains the SRP RNA and signal peptide-binding site. FtsY also has three domains with N-terminal acidic domain and the Ffh homolog NG domain responsible for GTP binding and substrate binding respectively (Keenan, Freymann, Walter, & Stroud, 1998; Rosendal, Wild, Montoya, & Sinning, 2003). The $M.\ tuberculosi$s SRP is less studied as compared to other bacterial SRP systems. Since SRP components are potential drug targets, studying their physiological importance and identifying their substrate proteins in $M.\ tuberculosi$s system would be helpful in designing the anti-TB drugs.

### 2.14 Research problems

Proteins are master regulators of the living cells. Though genes code them, proteins functionally regulate the biological processes. The amino acid sequences in the protein describe its function in the cell. However, most of the proteins are responsible for executing more than one function. They communicate and activate their substrates through protein interactions and execute their functions. For executing a single process, proteins form a complex signaling network. Nevertheless, the cellular machineries have multiple pathways to execute a single function. If one pathway shuts down, the bacterium uses alternate pathways and delivers the cellular needs. Additionally, expressions of some proteins have interconnected mechanisms in which, protein 1 depends on protein 2, which in turn depends on protein 3. Hence, understanding this
complex network is necessary to unravel the proteins functions. For achieving this, a large number of very efficient techniques are needed to visualize the interactions. Until recently, our knowledge about complex networks of protein interactions was very limited. Intense research resulted in identification of interactome map in several model organisms such as *Saccharomyces cerevisiae* and *Drosophila melanogaster*. However there is no available interactome map for mycobacterial species. Identifying the complex signaling in mycobacteria will be very useful in designing a novel anti-TB drug against this dreadful infectious disease. In addition, protein interaction studies are also needed to identify the secondary drug targets when the primary drug target protein becomes dysfunctional.

Proteins are dynamic in nature; they interact with biomolecules in different fashions. They often interact with DNA, RNA, proteins and metabolites based on their structural fold. Also the interactions are stable and transient. For example, the protein phosphorylations are transient interactions that stay for less time as compared to stable physical interactions. Hence, a variety of investigations are needed to pick up the interaction map. Investigations of protein interactions often allow one to identify the function of an unknown protein based on their substrate protein function. In some cases, the protein interactions bring up a totally different function to them when they are placed in a different environment. Proteins are shown to play different roles in biological processes such as transcription, translation, cell cycle control, signal transduction, protein export mechanisms and several regulatory processes.

This thesis addresses the identification of various functional protein-DNA and protein-protein interactions in mycobacteria. We specifically studied the involvement of protein interactions in transcription control, signal transduction and protein export systems of mycobacteria. We studied the gene regulation of acetamidase operon of *M. smegmatis*
by identifying the potential protein-DNA and protein-protein interactions involved in regulation of the operon. We also studied the protein-protein interactions in the PknI cell division gene cluster of *M. tuberculosis*. In PknI gene cluster, we identified the protein interactions for PknI, Ffh and FtsY proteins and further functionally characterized their interacting proteins.

### 2.15 Thesis organization

This thesis stands at the intersection of two areas namely, protein-DNA and protein-protein interactions of mycobacteria. The two main research problems, which are the main focus in this thesis, have been introduced and briefly reviewed. Further, the identification and functional characterization of specific protein-DNA and protein-protein interactions were explained in five chapters. Finally, experimental outcome of the studies were summarized and directed for future investigations.

Chapter one provides a protein-DNA interaction study of acetamidase operon of *M. smegmatis*. The author describes the protein-DNA interaction between AmiA and P2 promoter region that leads to negative regulation of the acetamidase operon. The DNA residues that contributing to the AmiA interaction are also identified. This chapter was published in the Microbiological Research journal (Sundararaman, Palaniyandi, Venkatesan, & Narayanan, 2014).

Chapter two is the continuation of chapter one. Here, the author describes the involvement of protein-DNA and protein-protein interactions in positive regulation of acetamidase operon. The positive regulatory protein, AmiC interacts with the negative regulatory protein AmiA and removes AmiA from the P2 promoter region. Further, AmiC establishes its positive regulation by binding to the inducer acetamide as well as to the P1 and P3 promoter regions.
Chapter three provides a bioinformatics approach of protein-protein interactions for PknI of *M. tuberculosis*. The interacting proteins for PknI were identified using an *in vitro* experiment, far-western blotting. One of the interacting proteins, Rv2159c was studied using various bioinformatics tools for identifying the critical amino acid residues in Rv2159c that are responsible for the interaction. This chapter was published in the Journal of Molecular Graphics and Modelling (Venkatesan, Hassan, Palaniyandi, & Narayanan, 2015).

Chapter four involves the functional characterization of PknI and Rv2159c interaction in *M. tuberculosis*. The bioinformatics study that predicted the Rv2159c motif responsible for its interaction with PknI was validated using *in vitro* protein-protein interaction experiments. Further, functional role of the hypothetical protein Rv2159c was characterized using genetic, biochemical and proteomic approaches.

Chapter five provides the protein-protein interactions involved in the SRP pathway of *M. tuberculosis*. The physical interaction between the universally conserved SRP components, Ffh and FtsY was established using *in vitro* protein-protein interaction experiment. Further, the functional role of Ffh and FtsY proteins were studied using genetic, biochemical and proteomic approaches.