2003

Characterization of protein secretion in Mycobacterium leprae using phoA fusions in Escherichia coli and Mycobacterium smegmatis

Marina Noemi Torrero
Louisiana State University and Agricultural and Mechanical College, mtorre3@lsu.edu

CHARACTERIZATION OF PROTEIN SECRETION IN *MYCOBACTERIUM LEPRAE* USING PHOA FUSIONS IN *ESCHERICHIA COLI* AND *MYCOBACTERIUM SMEGMATIS*

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Interdepartmental Program in Veterinary Medical
Sciences through the Department of Pathobiological Sciences

by
Marina Noemi Torrero
B.S., University of Buenos Aires, 1993
M.S., Louisiana State University, 1998
December 2003

ACKNOWLEDGMENTS

I would like to thank my major professor, Dr. Thomas Gillis for his support and guidance during this project. I am also thankful to Dr. Gillis for giving me the opportunity to attend various meetings, where I met important people in the field of microbiology.

I would like to extend my gratitude to my committee members, Dr. Sue Bartlett, Dr. Horohov, Dr. Krahenbuhl and Dr. O'Reilly for their advice and contributions to this project.

Special thanks go out to Sandra Robinson, Naoko Robbins, Greg McCormick and Vilma Tulagan for their technical assistance.

Finally, I wish to express my gratitude to my parents, Oscar and Hilda Torrero, my husband Luis Moscovich and my daughter Tammy Moscovich because their love and support gave me the strength to work hard on this project with a positive attitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Complete sequencing and annotation of the *M. leprae* genome has provided new information related to proteins constituting its hypothetical proteome. Since *M. leprae* can not be grown *in vitro*, novel approaches are needed to determine which proteins are expressed during infection and whether these proteins are related to pathogenesis. Secreted proteins represent a distinct group of protein with respect to their structure and function, contribution to virulence and are of particular importance for vaccine development because they are often immunogenic and have the potential to be recognized early in infection. The objectives of this study were: 1) to identify putatively secreted proteins of *M. leprae* based on protein sequences homologies with known MT secreted proteins; 2) to apply bioinformatic tools designed to assess proteins for secretion, to proteins selected in objective 1 with the goal of improving the likelihood that selected proteins are secreted by *M. leprae*, 3) to validate secretion of selected ML proteins through genetic cloning of predicted secreted ML protein genes using surrogate host bacteria, *E. coli* and *M. smegmatis*.

Bioinformatics identified 24 proteins with high probability for secretion in *M. leprae*. Fifteen of 24 ML genes showed more than 50% amino acid homology with their *M. tuberculosis* counterparts and were studied for gene expression and secretion. mRNA analysis identified transcripts for all Sec-dependent pathway proteins of 15 genes predicted to be secreted in *M. leprae*. PhoA fusion studies in *E. coli* showed that 5 of 6 (83%) ML proteins (ML0091, ML0097, ML0620, ML1811 and ML1812) were secreted in *E. coli* and 2 of 7 (29%) proteins (ML0715 and ML2569) were secreted in *M. smegmatis*. Only lipoproteins were secreted in *M. smegmatis* suggesting the importance

of mycobacterial-related characteristics for secretion of ML lipoproteins. These results suggest that bioinformatic tools are reliable predictors for identifying secreted proteins in *M. leprae* and support the hypothesis that Sec-dependent secretion exists in *M. leprae*.

INTRODUCTION

Secreted proteins represent a distinct group of proteins with respect to their structure and function and contribution to virulence. They are of particular importance for vaccine development because they are often immunogenic and have the potential to be recognized early in infection. Little is known regarding mycobacterial protein secretion. Elucidation of protein secretion in *M. leprae* could provide new insights into virulence factors of *M. leprae* and provide a source of proteins with potential for vaccine development.

A requisite step in protein secretion is protein translocation across the cytoplasmic membrane. This step is common to proteins that are released to the extracellular space or remain associated with the cell wall. The Sec-dependent pathway translocates precursor proteins containing N-terminal signal sequences across the cytoplasmic membrane (Oliver and Beckwith, 1981). Four different experimental approaches have been used to study *M. tuberculosis* secreted proteins. One is the analysis of culture filtrates that contains as many as 200 proteins (Sonnenberg and Belisle, 1997). This procedure has been used to study cultivable mycobacteria, such as *M. tuberculosis* and *M. bovis*, however, it cannot be applied to *M. leprae* since the bacteria must be harvested from infected host tissues where secreted proteins are lost during isolation of the bacilli.

A second approach, comparative genomics, can be used to study secretion in *M. leprae* by identifying gene sequences for *M. tuberculosis* or other mycobacteria secreted proteins and using these proteins to identify homologs in the *M. leprae* genome. A third approach for studying secretion utilizes bioinformatics tools which consists of computer algorithms capable of predicting various properties of proteins. Location (e.g., cytoplasmic, cell membrane, secreted) of proteins have been predicted accurately using

this approach (Wiker *et al*, 2000). While comparative genomics and bioinformatics are powerful analytical tools, results must be authenticated in biological systems. Genetic approaches have been used for authentication of secretion by making gene fusions with reporter genes encoding enzymes that become active upon translocation across the cell membrane (e.g. alkaline phosphatase, Hoffman and Wright, 1985).

The following chapters describe studies designed to characterize secreted proteins in *M. leprae* using comparative genomics, bioinformatics (Chapter 2) and authentication of secretion by phoA-ML gene fusions (Chapter 3).

Chapter 2 reports the use of a computer strategy to predict *M.leprae* secreted proteins taking advantage of the recently released *M. leprae* genome (Cole *et al*, 2001). Two different programs were used to identify proteins having secretory signal peptides but lacking additional membrane attachment domains: SignalP was used to predict the presence and location of signal peptide cleavage sites in amino acid sequences and Transmembrane Hidden Markov Model (TMHMM) was used to predict the location and orientation of transmembrane helices in protein sequences.

Chapter 3 describes the results from gene transcription studies of 15 ML proteins predicted to be secreted using SignalP and TMHMM. The mRNA was purified from ML grown in nude mice and gene expression was monitored by RT-PCR. Each gene was cloned into 2 separate reporter plasmids in an attempt to demonstrate protein secretion in either *E. coli* or *M. smegmatis*.

The hypothesis of this study is "*M. leprae* produces secreted proteins and SignalP/TMHMM are reliable bioinformatics tools to identify secreted proteins with special Sec-dependent characteristics.

The results of this work should help improve the understanding of Sec-dependent secretion in *M. leprae* and provides a listing of ML proteins with potential for development as vaccines or as diagnostic reagents capable of adding to current strategies for controlling leprosy.

CHAPTER 1

LITERATURE REVIEW

1.1. Leprosy

Gerhard Hansen was the first to recognize that leprosy was caused by a bacterium, *M. leprae,* and in so doing identified the first human bacterial pathogen (Hansen, 1874). In 1960, Shepard described the limited multiplication of *M. leprae* which occurs when *M. leprae* are injected into the footpads of immunocompetent mice. This model made it possible for the first time to screen drugs for anti-leprosy activity, diagnose drug-resistant leprosy and study certain aspects of immunity to *M. leprae*, including evaluation of leprosy vaccines. Although *M. leprae* has not been cultured in routine laboratory media, in the late 1960s armadillos, nude mice, and immunosuppressed rats were shown to yield high numbers of *M. leprae* following infection. This has greatly increased the availability of *M. leprae* and permitted investigations into the genetics of *M. leprae* and the pathogenesis of leprosy.

1.1.1. Epidemiology

One-hundred three countries reported new leprosy cases for 2003 totalling 524,000 globally. The global leprosy prevalence for 2002 was 612,000 (WHO). The global prevalence of leprosy has been significantly reduced from the estimated 10-15 million cases reported a decade ago as a result of the World Health Organization program for global elimination of leprosy as a public health problem, based on the worldwide implementation of multidrug therapy (MDT). The implementation of MDT had a significant effect on reducing the prevalence of leprosy, but the incidence of the disease has remained steady for the last 10-15 years at a rate of around 500,000-600,000 new cases per year. These data suggest that to truly eradicate leprosy other forms of

intervention such as tools for early diagnosis and prevention by vaccination may be necessary.

Physical deformities in leprosy, which are permanent and often progressive, result in both reduced opportunities for patients and, in areas where leprosy is highly endemic, economic loss to the community. The consequences of leprosy related deformities and disabilities are much more pronounced than other disabilities due to the added effects of social disability resulting from the stigma attached to the disease.

Unlike many other communicable diseases, there is considerable difficulty in identifying the three reference points that are involved in the transmission of leprosy. The identification of the point of onset of infection is the most important and difficult of the problems in the study of transmission. The human being is the only known reservoir of infection in leprosy, with the exception that naturally occurring disease with organisms indistinguishable from *M. leprae* has also been detected among wild armadillos in parts of the southern United States and Brazil. Up to 5% of armadillos in Louisiana have been found to have clinical disease, with about 20% having serological evidence of *M. leprae* infection (Truman *et al*, 1986). In addition, primates (Meyers *et al*, 1991) and chimpanzees (Leininger *et al*, 1980) naturally infected with *M. leprae* have been identified.

1.1.2. Clinical Disease

Clinical disease is categorized into two polar forms, tuberculoid (TT) and lepromatous (LL) leprosy and variations between the two polar forms are referred to as borderline leprosy.

TT exhibits one or a few circumscribed skin lesions containing rare demonstrable bacilli. Aspects of cellular immunity appear to be active in the pathogenesis of TT leprosy as the

characteristic histological picture is one of well formed granulomas. LL is the disseminated form of the disease and the patient lacks demonstrable cell-mediated immunity (CMI) against *M. leprae*, resulting in disseminated growth of the bacilli. Nerve involvement in early LL is less severe than TT disease due in part to lack of CMI, but as the disease progresses, deformities may occur resulting from nerve invasion by massive numbers of bacilli. The majority of leprosy cases fall into the borderline classification. The major divisions of borderline disease are borderline tuberculoid (BT) and borderline lepromatous (BL). The gross appearance of leprosy along the disease spectrum from BT to BL generally reflects more numerous and disseminated lesions. The bacteria within lesions become more abundant and histopathological evidence of CMI diminishes as the clinical spectrum moves from BT to BL. Because borderline patients have significant levels of bacilli in their tissues and an active, although limited, immunological capacity to recognize *M. leprae*, they suffer from the greatest degree of nerve damage and related deformities (Gillis and Krahenbuhl, 1998).

The natural course of leprosy often involves complications referred to as leprosy reactions. Reactions occur due to invasion of tissue by *M. leprae* and resultant immunological activity. The term reaction is used to describe the appearance of symptoms and signs of acute inflammation in lesions of patients with leprosy.

There are two types of reactions:

- Type 1, associated with cell-mediated hypersensitivity, occurs in patients with tuberculoid and borderline leprosy. Untreated, type 1 reactions may last for months or years, and can relapse unexpectadely. Neuritis is the most important consequence of a type 1 reaction and it may occur together with skin changes or independently, possibly reflecting hypersensitivity to different antigens of *M. leprae*.

- Type 2 reactions are thought to be associated with immune complexes and are also known as erythema nodosum leprosum (ENL). Type 2 reactions occur in patients with multibacillary disease and cause acute inflammation in any organ or tissue where *M. leprae* are found. ENL is seen primarily in BL and LL patients who produce large amounts of antibody to *M. leprae* (Hastings, 1994).

1.1.3. Immunology

In TT leprosy a strong cellular immune response to *M. leprae* is associated with limited disease and few, if any, demonstrable bacilli within granulomas located in the dermis and peripheral nerves. By contrast, in LL leprosy the absence of a cellular immune response leads to the uncontrolled proliferation of leprosy bacilli, extensive clinical lesions and a strong antibody response to mycobacterial antigens. In the borderline forms of leprosy progressive reduction in the cellular immune response and delayed- type hypersensitivity to *M. leprae* from BT to BL disease is accompanied by more frequent skin and nerve lesions, a greater bacillary load and increasing antibody levels (Waters, 1992).

Activation of the cellular immune response to *M. leprae* is dependent on the initial interaction of T cells and antigen presenting cells (APC). Immunohistochemical studies have established the predominance of $CD4^+$ over $CD8^+$ T cells in lesions of TT leprosy patients at a ratio 2:1, while in BL/LL lesions there are far fewer lymphocytes and similar proportions (1:1) of $CD4^+$ and $CD8^+$ T cells (Mehra and Modlin, 1989).

In tuberculoid leprosy, the granuloma is organized in distinct immunological microenviroments. The core of the granuloma is composed of well-differentiated macrophages and epithelioid cells diffusely infiltrated with $CD4^+$ T lymphocytes. The predominant localization of $CD8^+$T cells at the periphery of the granulomas suggests a

role in localizing the granulomatous reaction (Hastings, 1994). By contrast, the granuloma of lepromatous leprosy does not show this compartmentalization of CD4 and CD8 lymphocytes. CD4 and CD8 cells are distributed evenly throughout the granuloma and at an equal ratio. This arrangement of CD4 and CD8 T cells may reflect a relatively inefficient host response, allowing virtually uninhibited bacillary proliferation. The lack of a surrounding mantle of CD8 T cells may also facilitate the dissemination of the granulomatous response.

One study of cytokine mRNA production in leprosy lesions using reverse transcriptase polymerase chain reaction (RT PCR) demonstrated a predominance of IL-2 and IFN-γ transcripts in TT lesions, while IL-4, IL-5 and IL-10 were more apparent in LL lesions (Yamamura *et al*, 1991). These results are in agreement with the Th1/Th2 paradigm of Mosmann and Coffman, 1989, and help explain the immunological spectrum underlying clinical leprosy. In addition, Th1 cells play a central part in the activation of cellular immune mechanisms whereas Th2 cells are of crucial importance for humoral immunity.

Cells of the monocyte/macrophage lineage play various roles in the parasite/host relationship in leprosy, serving as cellular habitat for *M. leprae*, activators of T cells as antigen-presenting cells (APC), and effector cells in destruction of the bacillus. Other cells such as dendritic cells, Langerhans cells, B cells and endothelial cells can act as APC, but macrophages are uniquely efficient because they are capable not only of endocytosis of soluble antigens, but they can also phagocytose particulate antigens and present them for stimulation of CMI in a major histocompatibility complex (MHC) class II restricted context. By secretion of T cell amplifying cytokines such as IL-1 and tumor necrosis factor alpha (TNFα), macrophages can up-regulate CMI.

The immunological understanding of clinical leprosy continues to be dominated by the unusual specific anergy to *M. leprae* in LL. This minority of leprosy patients are unable to mount and efficient CMI to *M. leprae.* It is not known whether this specific immunodeficiency is predetermined, e.g., by genetic factors, by the kind and extent of exposure to other mycobacteria prior to infection with *M. leprae*, or is induced by *M. leprae* itself (Hastings, 1994). Based on the current understanding of the immunological events observed in leprosy a vaccine capable of inducing a Th1-type response to *M. leprae* in tissues may be capable of preventing the initial infection in TT and borderline leprosy, but may not in LL disease.

1.2. Potential Significance of ML Secreted Proteins

Proteins released by mycobacteria to the extracellular environment have been the focus of research directed at identifying antigens that induce protective immunity and that may be involved in pathogenesis. Living mycobacteria are significantly more effective as inducers of protective immunity than dead bacilli (Bloch and Segal, 1955) and several studies have shown that the majority of bacterial virulence factors are extracellular proteins (Finlay and Falkow, 1997, and Miller and Cossart, 1999).. These observations underscore the importance of understanding secretion in *M. leprae* with the intent of identifying secreted proteins with potential for vaccine development and that may play a role in the pathogenesis of leprosy.

Secreted proteins of *M. tuberculosis* are believed to be responsible in part for the efficacy of the live vaccine, *M. bovis* BCG, used against tuberculosis and leprosy. Shepard tried different mycobacterial preparations to measure protection in mice when challenged with *M. leprae*. The protection provided by BCG was markedly reduced when the vaccine was heated at 60ºC for 30 minutes killing the bacteria (Shepard *et al*, 1978).

A study in Malawi showed that a second BCG vaccination reduced the risk of leprosy by about 50% among individuals who had received BCG vaccination in the past. There was no evidence that addition of killed *M. leprae* enhanced the protection against leprosy afforded by BCG alone (Fine, 1996) suggesting that addition of static structural antigenic components of *M. leprae* did not improve the vaccine potency.

Recognition of mycobacterial secreted proteins by the immune system may lead to early detection of infected macrophages and control of the disease. Subunit vaccines based on mixtures of culture filtrate proteins from *M. tuberculosis* have resulted in protective immunity in several studies using animal models of tuberculosis (Andersen P, 1994; Hubbard *et al*, 1992; Pal and Horwitz, 1992; Roberts *et al*, 1995). Only a few of these proteins, for example, Ag85 complex and ESAT-6, have been isolated and characterized at present. Since *M. leprae* encodes homologs of these and other *M. tuberculosis* secreted proteins, the *M. leprae* secreted proteins represent an unexplored resource with great potential for developing new vaccines with therapeutic and prophylactic potential.

1.3. Protein Secretion Mechanisms in Bacteria

While exported proteins are those proteins primarily associated with the cell wall and possibly released into the culture medium over time, secreted proteins are soluble proteins that are released into the culture supernatant and not associated with the bacteria (Wiker *et al*, 1999). Prokaryotes have a number of pathways dedicated to the process of protein secretion. In general, these organisms translocate the majority of their secreted proteins via the Sec pathway. There are at least 6 different pathways for protein secretion known in Gram-negative bacteria as reviewed by Thanassi and Hultgren, 2000. Four of these pathways release proteins with cleavable amino-terminal signal sequences to the

extracellular space in a two-step process that requires the Sec pathway for translocation across the inner membrane. The four pathways are autotransporters, chaperone/usher, type II and type IV secretion. The other pathways are Sec-independent and capable of exporting substrates in one step to the extracellular space. These pathways are type I and type III secretion. An alternate secretion mechanism, the twin-arginine translocation (Tat) pathway, was originally identified in chloroplasts and has recently been found in bacteria and Archae (Santini *et al*, 1998). Tat signal peptides are similar to Sec signal peptides, but they contain a highly conserved twin-arginine motif. *Mycobacterium* species possess the genes required to translocate proteins by both the Sec and twin-arginine pathways.

Various approaches have been used to study secretion and secreted proteins from bacteria. Analysis of *in vitro* culture filtrates showed that *M. tuberculosis* produces as many as 200 putatively secreted proteins (Sonnenberg and Belisle, 1997). Whereas this approach has been used successfully for cultivable bacteria, it is inappropriate for similar studies with the leprosy bacillus since *M. leprae* must be harvested from infected tissues resulting in the loss of secreted proteins from the purified bacilli.

A second approach for studying secreted proteins utilizes comparative genomics and is readily applicable to studying *M. leprae* secreted proteins. For example, genes known to encode secreted proteins in *M. tuberculosis* can be used to search the genome of *M. leprae* for sequences with strong DNA sequence similarities. If these regions of similarity appear to encode genes with potential for secretion, then the *M. leprae* genes can be cloned and studied further to verify transcription, translation and secretion.

Bioinformatics can also be used to support comparative genomic studies by providing tools for predicting properties of proteins, including cellular location. Computer algorithms have been developed that predict whether a protein is located

within the cytoplasm, the cell membrane or exported from the cell. Proteins predicted for secretion by either comparative genomics or bioinformatics approaches require validation in biological systems. Genetic tools can help bridge the gap between *in silico* predictors and *in situ* authentication. A particularly useful genetic tool for this purpose utilizes gene fusion technology in which an unknown gene, suspected of being secreted, is fused to a reporter gene (e.g., alkaline phosphatase) which acts to signal the secretion event upon translocation. For mycobacterial genes, including *M. leprae*, validation of secretion can be accomplished by cloning *M. leprae*-reporter fusion genes into *E.* coli or *M. smegmatis*.

1.3.1. Signal Sequence Dependent Mechanisms for Translocation

- Sec Pathway

The Sec pathway translocates precursor proteins containing N-terminal signal sequences across the cytoplasmic membrane. The work of Oliver and Beckwith, 1981, established that the translocation of precursor proteins across the cytoplasmic membrane of *E. coli* requires the products of at least six genes. These genes, *secA, secB, secD, secE, secF*, and *secY*, encode interacting core-components of a complex translocation apparatus that consists of soluble and membrane integrated proteins. SecA is an ATPase that performs a central role in bacterial protein secretion, as it is the molecular motor that drives translocation. During the last few years, genes encoding SecA homologs from many different Gram-positive organisms, have been cloned (Klein *et al*, 1995; Blanco *et al*, 1996; Braunstein *et al*, 2001; Limia *et al*, 2001). The SecA protein binds the SecYEG complex to form the translocase, which mediates the ATP translocation of precursor proteins across the membrane. Proteins are secreted through an aqueous channel formed by the SecY, SecE and SecG polypeptides. SecY is the largest subunit of the membrane domain of the protein translocase. SecE is a small integral membrane protein that is

essential for translocation and viability. SecE of Gram-positive bacteria are considerably smaller than the *E. coli* SecE. Gram-positive SecE sequence is homologous to the carboxy-terminal part of the *E.coli* SecE, which corresponds to the highly conserved cytosolic loop region and the third transmembrane segment. Another protein, band 1 or SecG, copurifies and coimmunoprecipitates with SecY and SecE forming a three-component complex as described by Brundage, *et al*, 1990. Sequence analysis showed that Gram-positive SecG have weak but significant sequence similarity with the *E. coli* SecG. SecG proteins from Gram-positive bacteria tend to be shorter than their Gram-negative counterparts. Unlike SecY and SecE, SecG is not essential for viability and protein translocation in *E. coli*. SecD and SecF are needed for the late stages of translocation. SecB is a chaperone protein that forms a complex with precursor proteins but does not catalyze folding or unfolding. SecB binds to the carboxy-terminus of Sec-A and is essential only in rapidly growing cells. This region of SecA is highly conserved among many Gram-negative and Gram-positive bacteria; however, no homologs with sequence similarity to SecB are present in any of the Gram-positive organisms of which the genome sequence has been completed. (van Wely *et al*, 2001).

- Twin Arginine Translocation Pathway

The twin-arginine (Tat) pathway translocates redox proteins across the cytoplasmic membrane in *E. coli* (Sargent, *et al*., 1998) and has recently been shown to secrete virulence factors from *Psuedomonas aeruginosa* (Dilks, *et al*., 2003). Even though secreted peptides exhibit a highly conserved twin-arginine motif, the Tat pathway is not universally conserved. Signal sequences for proteins in the twin-arginine pathway are longer than those found in proteins secreted by Sec-dependent mechanisms and

contain two arginines. The twin arginine pathway differs from the Sec pathway in that proteins are secreted in a folded conformation.

The Tat pathway is Sec-independent as multiple Sec factors have been shown to play no role in the process. The *E. coli* Tat pathway contains 4 genes (*tatA, tatB, tatC and tatD*) organized in an operon with an unassociated fifth gene, *tatE*. Tat B and C have essential roles in translocation, disruption of these genes leads to a complete block in the export of a subset of periplasmic proteins (Sargent, *et al*, 1998). Disruption of *tatD,* however, has no detectable effect on secretion suggesting TatD is not involved in the export process. The Tat complex has been purified from *E. coli* and was shown to contain only TatA, TatB and TatC (Bolhuis, *et al*, 2001).

1.3.2. Amino-terminal Signal Peptides

Signal peptides consist of short amino acid sequences which, after protein delivery to the correct subcellular compartment, are frequently removed by specialized signal peptidases. Three distinct domains of signal peptides can be recognized: The first domain is the amino-terminal N-domain containing positively charged residues (e.g., arginine or lysine). The positively charged N-domain is thought to interact with the translocation machinery and negatively charged phospholipids in the lipid bilayer of the cytoplasmic membrane during translocation. The second domain is the H-domain which is composed of hydrophobic residues that appear to adopt an α-helical conformation in the membrane. Helix-breaking glycine or proline residues are frequently present in the middle of this hydrophobic core. Helix-breaking residues found at the end of the H-domain are thought to facilitate cleavage by a specific signal peptidase. The C-domain, following the H-domain, contains the cleavage site for signal peptidase. The (-3, -1) rule states that the residues at positions -3 and -1 (relative to the cleavage site) must be small

and neutral for cleavage to occur correctly; this area is often referred to as an Ala-X-Ala cleavage site. Alternatively, the C-domain may contain Gly or Ser in position -1 and Val, Ser or Thr in position -3 (von Heijne, 1983).

The signal peptidase removes the signal peptide from the mature part of the secreted protein during or shortly after translocation. The mature part of the protein is released from the membrane and can fold into its native conformation. Finally the signal peptide is degraded by signal peptide peptidases and removed from the membrane (Tjalsma *et al*, 2000).

Four major classes of amino-terminal signal peptides can be distinguished in *B. subtilis* on the basis of the signal peptidase recognition sequence. The first class is composed of typical signal peptides which are present in preproteins that are cleaved by one of the various type I signal peptidases. Although most proteins having such a signal seem to be secreted into the extracellular environment, some of them are retained in the cell wall. A subgroup of these signal peptides contain a so called twin-arginine motif (RR-motif), which directs proteins into a distinct translocation pathway known as the Tat pathway. The second major class of signal peptides is present in prelipoproteins, which are cleaved by the lipoprotein-specific type II signal peptidase. Lipoproteins have a well-conserved lipobox that contains a cysteine residue that is lipid modified by the diacylglyceryl transferase prior to cleavage by signal peptidase II. After translocation across the cytoplasmic membrane, exported lipid-modified proteins remain anchored to the membrane by their amino-terminal lipid-modified cysteine residue. The third major class is formed by signal peptides of prepilin-like proteins which in *B. subtilis*, are cleaved by the prepilin-specific signal peptidase ComC. The recognition sequence for the prepilin signal peptidase is localized between the N- and H-domains, leaving the H-

domain attached to the mature pilin after cleavage. Finally, the fourth major class of signal peptides is found on ribosomally synthesized bacteriocins and pheromones that are exported by ABC transporters. These signal peptides lack a hydrophobic H-domain and are removed from the mature protein by a subunit of the ABC transporter that is responsible for the export of a particular bacteriocin or pheromone (Tjalsma *et al*, 2000).

1.3.3.Autotransporters

The autotransporter secretion pathway is a terminal branch of the Sec pathway that exports proteins with diverse functions, including proteases, toxins, adhesins and invasins. A typical autotransporter contains three domains as described by Henderson *et al*, 1998: an amino-terminal signal sequence for secretion across the inner membrane by the Sec pathway, an internal passenger or functional domain, and a carboxy-terminal β-domain. The β-domain inserts into the outer membrane to form what is predicted to be a β-barrel pore structure, similar to the bacterial porins, through which the passenger domain passes to the cell surface. A linker region connecting the passenger and β-domains is also essential for export and may guide the passenger region through the β-domain channel. Once secreted, the passenger domain is either retained on the bacterial surface or released into the environment by proteolysis.

Neisseria secretes IgA protease, an enzyme that cleaves antibodies on mucosal surfaces by the autotransporter secretion pathway. IgA protease is synthesized as a pre-proenzyme with an N-terminal signal peptide that initiates the precursor into the Sec pathway. After cleavage of the signal peptide by signal peptidase, the proenzyme resides in the bacterial periplasm. The C-terminal β-domain of IgA protease assumes a β-barrel structure that inserts into the outer membrane and functions as an autotransporter for the N-terminal domain. Once the N-terminal protease domain is exposed on the bacterial

surface, it cleaves the proenzyme at the junction between the N-terminal and the C-terminal domain. The cleaved N-terminal domain of the proenzyme is released from the bacterial surface and acts as virulence factor.

1.3.4. Chaperone/usher Pathway

The chaperone/usher pathway is a branch of the Sec pathway for secretion of a broad range of adhesive virulence structures on the Gram-negative bacterial surface (Thanassi *et al*, 1998). Secretion across the outer membrane by this pathway requires only 2 components: a periplasmic chaperone and an outer membrane protein termed an usher.

The assembly of P and type I pili, expressed by uropathogenic *E. coli,* is an example of this pathway (Roberts *et al*, 1994). These pili consist of a thin, flexible tip fibrillum connected to a rigid, helical rod. Following export across the inner membrane via the Sec pathway, pilus subunits interact with the periplasmic chaperone via a conserved carboxy-terminal motif present on each of the pilus subunits. The chaperone facilitates release of pilus subunits into the periplasm (Jones *et al*, 1997). Interaction with the usher protein in the outer membrane triggers chaperone dissociation from the subunit, allowing incorporation of the subunit into the pilus fiber. The usher provides a translocation channel through the outer membrane for secretion of the pilus.

1.3.5. Type I Secretion

ATP binding cassette (ABC) transporters, found in eukaryotes and prokaryotes, constitute a large superfamily of multi-subunit permeases that transport different molecules (ions, amino acids, peptides, antibiotics, polysaccharides, proteins, etc) across biological membranes. ABC transporters are classified as importers and exporters depending on the direction of translocation of their substrate. The type I pathway is Sec

independent and secretes proteins directly from the cytoplasm across the outer membrane. Substrates of this pathway lack a cleavable amino-terminal signal sequence. Instead, the substrates possess a carboxy-terminal amino acid secretion signal. The type I export apparatus consists of two hydrophobic membrane spanning domains (MSDs) associated with two cytoplasmic nucleotide binding domains (NBDs). The NBDs of ABC transporters bind ATP and couple ATP hydrolysis to the transport process. The MSDs consist of four to eight transmembrane α-helices forming a channel allowing the translocation of the substrate through the membrane (Braibant *et al*, 2000).

An example of a type I ABC exporter is the secretion of hemolysin (HlyA) by pathogenic *E. coli*. HlyA is a lipid-modified polypeptide with a domain that is composed of 11-17 nine-amino-acid repeats. The repeat domains bind calcium and are thought to interact with host cells, triggering HlyA insertion into the plasma membrane and leakage of the cytoplasmic contents of target cells. TolC, the outer membrane protein for hemolysin export, assembles as a trimeric complex in the outer membrane as a porin-like β-barrel membrane domain with a carboxy-terminal hydrophilic region that extends into the periplasm. TolC appears to be involved in several *E. coli* transport processes such as the efflux of antibiotics, heavy metal ions, detergents and solvent (Thanabalu *et al*, 1998). The *E. coli* genome encodes 57 ABC transporters, of which 44 are importers and 13 are exporters. Gram-positive organisms also employ ABC transporters The *B. subtilis* genome encodes 78 ABC transporters, of which 38 are importers and 40 are exporters**.**

1.3.6. Type II Secretion

The type II secretory pathway represents a third terminal branch of the Sec pathway. This pathway is responsible for secretion of extracellular enzymes and toxins by a wide variety of Gram-negative bacteria. Secretion across the outer membrane by the

type II branch requires 12 to 16 accessory proteins (Pugsley *et al*, 1997). The type II pathway includes two outer membrane components: GspD, an integral protein and GspS, a small lipoprotein required in at least some type II systems for proper targeting and insertion of GspD in the outer membrane. Type II substrates cross the inner membrane via the Sec pathway followed by signal-sequence cleavage and protein folding in the periplasm. The GspD forms a complex with the GspS lipoprotein and serves as a gated channel for secretion of substrates to the cell surface. Most of components of the type II pathway are associated with the inner membrane (GspC, GspF, GspM, GspL and GspO). There are 5 pseudopilins (GspG, GspH, GspI, GspJ and GspK) each with a pilin signal peptide that is cleaved by prepilin peptidase (GspO) during Sec-mediated translocation across the plasma membrane. GspE is a cytoplasmic protein that localizes to the inner membrane via interaction with GspL, contains a conserved ATP-binding motif and has autokinase activity. GspE may regulate secretion or energize the secretion process (Lee and Schneewind, 2001). The type II secretion machinery is responsible for secretion of proteins across the outer membrane and assembly or retraction of type IV pili.

1.3.7. Type III Secretion

Type III secretion is activated by bacterial contact with host cells and is capable of translocating antihost factors into the cytosol of target eukaryotic cells. Type III secretion is Sec independent, may take place without a periplasmic intermediate, and requires about 20 secretion components that assemble into a large structure that spans bacterial membranes as well as the host cell membrane (Hueck, 1998). Secretion of *Yersinia* outer proteins (Yops) by *Yersinia* spp. represents the typical type III export pathway. There are 13 Yop proteins of *Yersinia* species (YopB, D, E, H, M, N, O, P, Q, R, T, LcrV and Q) that do not share common peptide sequences and are secreted into the

extracellular medium without post-translational modification or cleavage of the polypeptide chain. The genes encoding Yops are located on a virulence plasmid. A temperature shift to 37ºC, which occurs when bacteria enter the human body, induces expression of the type III genes and assembly of the secretion machinery (Cornelis, 1998).Type III secretion is highly regulated with the first signal appearing to reside in the mRNA which may target the RNA-ribosome complex to the type III machinery for translation and secretion. The second Yop secretion signal serves as the binding site for cytoplasmic chaperones termed Syc proteins and may target Yops to the type III machinery for translocation into host cells. Most models for the mechanism of translocation across the host cell membrane described a pore formation by YopB and YopD proteins. Microscopic detection have revealed that YopE, H, M, O, P and T are injected into eukaryotic cells; YopB, D and R are secreted into the extracellular space; and YopQ is associated with the bacterial envelope (Cheng and Schneewind, 2000).

1.3.8. Type IV Secretion

Type IV transporters mobilize proteins and DNA either from bacteria to bacteria or from bacteria to eukaryotic cells. The most common example of type IV secretion pathway is the *virB* system of *Agrobacterium tumefaciens* that exports T-DNA across the bacterial membranes and into plant cells, where the T-DNA integrates into the plant genome (Zambryski, 1988).The *virB* locus consists of 11 genes, 10 of which (*virB2* to *virB11*) are critical for DNA transfer. Although *virB1* is not essential, deletion of this gene attenuates virulence and leads to a lower efficiency of DNA transfer (Berger and Christie, 1994). The DNA transport system described for *Agrobacterium tumefaciens* is very similar to a toxin transport system of *Bordetella pertussis*. This type IV pathway consists of nine proteins required for the secretion of pertussis toxin across bacterial

membranes. In general, type IV transporters contain 2 proteins (virB4 and virB11) with nucleotide-binding motifs, these proteins might serve to signal the opening of a gate or channel via kinase activity, or act as molecular chaperones in the assembly of the transporter. Presently, very little is known about the series of events that occur during the transport process.

1.4. Protein Secretion Mechanisms in Mycobacteria

Mycobacterial secreted proteins are translocated across the cytoplasmic membrane, transported across the entire cell envelope and released into the extracellular space. Exported proteins are those translocated across the cytoplasmic membrane and either totally or partially associated with the cell envelope.

There is currently no evidence of autosecretion, type II, III or IV secretion system in *M. tuberculosis* as described in *E. coli* and other Gram-negative bacteria. The elucidation of the genome sequence of *M. tuberculosis* demonstrated that this organism possesses all the genes required for Sec-dependent translocation of proteins; however, it is not known how proteins are further exported through the cell wall of *M. tuberculosis*. The cellular envelope of *Mycobacterium* species possesses characteristics of the cell walls of both Gram-negative and Gram-positive bacteria. The pseudo-bilayer formed by the cell wall-attached mycolic acids and outer layer lipids is similar to the outer membrane of Gram-negative bacteria, whereas the arabinogalactan covalently attached to the peptidoglycan is similar to the cell wall teichoic acids of some Gram-positive bacteria (Brennan and Nikaido, 1995). Therefore, the mycobacterial cellular envelope with mycolic acids and free lipids represents a formidable obstacle to protein secretion. Some of the possible mechanisms that might be active to translocate this barrier are ABC transporters, secretin-like proteins and porins (Braunstein and Belisle, 2000). Many

putative ABC transporters are encoded by the *M. tuberculosis* genome; however, since ABC transporters can import and export a wide range of molecules, it is difficult to predict the transported substrate. Braibant, *et al*, 2000, reported that the genes encoding the ABC transporters occupy about 2.5% of the genome of *M. tuberculosis*. Contrary to *E. coli* but similar to *B. subtilis*, there is a relatively equal representation of exporting systems versus importing systems in *M. tuberculosis* (16 importers versus 21 exporters). Many of those exporting systems have been implicated in the export of anti-bacterial drugs. In addition, there is one exporting system specific to *M. tuberculosis* and *A. tumefaciens* that could be required for virulence and bacterial attachment to host cells (Draper, 1998).

Several *M. tuberculosis* proteins identified in the cell wall or culture filtrate are known to possess typical N-terminal signal sequences for translocation via the Sec-dependent pathway. The first evidence of a Sec-dependent pathway in mycobacteria was the recognition of N-terminal signal peptides in the predicted amino acid sequences of known *M. tuberculosis* exported and secreted proteins. The second indication that this pathway exists in *M. tuberculosis* came from the identification of homologous secretion factors (*secA, secD, secE, secF* and *secY*) encoded by the *M. tuberculosis* genome sequence (Cole *et al*, 1998).

Wiker *et al*, 2000, studied 28 *M. tuberculosis* secreted proteins and found the median signal peptide length was 32 residues, the median n-region length was 9 amino acids, the median h-region length was 16 amino-acids and the median c-region length was 7 amino acids. In terms of overall amino acid composition, the n-region of *M. tuberculosis* signal peptides is characterized by high arginine content (22%) and a relative lower amount of lysine (6%). The arginine-rich n-regions are most likely explained by

the high G+C content. Analysis of the preferred amino acids in the positions around the cleavage site showed that the mycobacterial signal peptides generally fulfill the (-3,-1) rule, with alanine in 19 of 28 signal peptides in position -3 and alanine in 23 of 28 in the position -1.

The only Sec factor apparently lacking in *M. tuberculosis* is SecB, but this protein has been found only in Gram-negative organisms.

Mycobacterium species possess all the genes required for Sec-dependent translocation of proteins; however, *M. tuberculosis* encodes two homologs of SecA. SecA1 and SecA2 contain ATP binding motifs but they are only 34% identical to each other at the amino acid level. This sequence difference between the two SecA proteins may reflect specific roles played by each in protein translocation (Braunstein *et al*, 2001). Analysis of SecA nucleotide sequence data from different bacterial species has revealed a high degree of conservation within the amino-terminus of the predicted protein sequence, as well as a similarly conserved ATP binding motif. There are some differences in the predicted carboxy-terminal regions, consistent with the likelihood that this portion of the SecA protein interacts with the cytoplasmic membrane during the translocation process. Owens *et al*, 2002, were able to build a chimeric form of SecA containing the amino-terminal end from *M. tuberculosis* and the carboxy-terminal end from *E. coli* to complement the defective SecA protein in *E. coli*. The full-length *M. tuberculosis secA* was unable to compensate for the temperature-sensitive (ts) defect, whereas the *E. coli* mutant that had been transformed with the plasmid containing the chimeric *secA* gene was able to grow at 42ºC. This experiment confirms that the N-terminal end of *M. tuberculosis* SecA contains appropriate ATP binding sites and possesses sufficient ATPase activity to compensate for the temperature-sensitive defect of the *E. coli* mutant.

There is evidence of mechanisms to transport proteins with recognizable signal sequences across the cytoplasmic membrane, however, it is unclear what determines the extracellular release or retention in the cell wall (Braunstein and Belisle, 2000). The twin-arginine pathway is Sec-independent and it is responsible for the translocation of folded proteins across the cytoplasmic membrane. The possible presence of a twin-arginine translocation system in *M. tuberculosis* is supported by the observation of homologs to components of the twin-arginine system in *E. coli* (TatA, TatB, TatC and TatD). Recently, Dilks *et al*, 2003, designed a program to predict the presence of Tat substrates in bacterial genomes. TATFIND identified 31 open reading frames in *M. tuberculosis* and 9 in *M. leprae,* most of them are hypothetical proteins.

1.5. General Design of Studies to Identify *M. leprae* Secreted Proteins.

The standard approach for studying secretion in cultivable bacteria is to analyze cell-free culture filtrates containing truly secreted proteins. This approach has successfully identified approximately 200 secreted proteins in *M. tuberculosis* but has been shown to contain cytoplasmic proteins that leaked into the culture filtrate through normal bacterial lysis in culture (Sonnenberg and Belisle, 1997). While this approach has been important in defining major groups of secreted proteins from *M. tuberculosis* and other cultivable mycobacteria, it cannot be applied to studying ML proteins since secreted proteins from *M. leprae* are lost during isolation of the bacilli from infected animal tissues.

Comparative genomics has been another successful approach used to study secretion. By identifying gene homologs in bacterial genomes known to be involved in secretion pathways and identifying particular secretory peptide motifs in genes it is possible to predict with good accuracy the destination of a selected protein. I have used

this approach in my dissertation research by selecting known and predicted secreted proteins of *M. tuberculosis* as a screen for homologous proteins in *M. leprae*. The ML proteins selected by this approach were then reanalyzed for secretion motifs using bioinformatic tools and finally examined for secretion following cloning into either *E. coli* or *M. smegmatis*. Secretion of ML proteins was validated by observing the location (cytoplasmic vs. extracellular) of a reporter protein (alkaline phosphatase) to which ML genes were fused.

For the purposes of the studies outlined in my dissertation research I have not attempted to distinguish between secreted and exported proteins. Rather, I have used bioinformatic tools to predict those proteins of *M. leprae* that may be secreted via the Sec-pathway and that may or may not remain associated with the bacterial cell wall. Because *M. leprae* cannot be grown *in vitro*, it is impossible to validate the purely extracellular location of proteins produced by *M. leprae* in culture. Therefore, protein secretion in my studies was validated by genetic cloning of predicted secreted ML protein genes into plasmid vectors containing reporter genes. The ML-reporter gene fusions were transformed into either *E. coli* or *M. smegmatis* to evaluate protein secretion in culture.

1.5.1.Comparative Genomics: *M. tuberculosis* vs. *M. leprae*

The complete genome sequence of *M. leprae* contains 3,268,203 base pairs (bp), and has an average G+C content of 57.8% (Cole *et al*, 2001). These values are much lower than those reported for the *M. tuberculosis* genome, which comprises ~4,000 genes, 4,411,532 bp and 65.6% G+C (Cole *et al*, 1998). The apparent reason for the large discrepancy in gene number has to do with *M. leprae*'s large number of pseudogenes. The distribution of the 1,116 pseudogenes in *M. leprae* is essentially

random and preliminary studies have suggested that some pseudogenes may produce either partial or complete transcripts (Williams *et al*, 2003). Of the 1605 genes predicted in *M. leprae* there are 1440 which are also found in *M. tuberculosis* and 165 genes that have no orthologue in *M. tuberculosis* (Cole *et al*, 2001). The *M. tuberculosis* genome was the first to be sequenced and revealed an organism with broad metabolic potential; most anabolic pathways are present and it has many genes involved in lipid synthesis and metabolism. By contrast, the *M. leprae* genome appears highly degenerate as defined by deficiencies in genes for recombination, appears to have fewer lipolytic genes than *M. tuberculosis*, lacks iron scavenging siderophores and is deficient in energy production. The major protein families in *M. leprae* are involved in lipid or polyketide metabolism, modification and synthesis of cell envelope components (methyl-transferases, glycosyltransferases), transport processes (ABC transporters, MmpL proteins), or in gene regulation (TetR, WhiB, two component system response regulators). While most of the *M. leprae* genes have orthologues in *M. tuberculosis,* there are several that appear to be unique and may have novel activities. These include hypothetical proteins, inorganic pyrophosphatase, prolyl-tRNA synthetase, uridine phosphorylase, adenylate cyclase, cytochrome P450 (Cole *et al*, 2001).

Examination of the genome sequence of the leprosy bacillus provides possible explanations for some of *M. leprae*'s unique properties. *M. leprae* has the longest doubling time of any known bacteria and cannot be grown in culture, both properties of which are potentially due to an extreme case of gene loss through reductive evolution. Less than half of the genome contains functional genes, with many pseudogenes showing functional counterparts in *Mycobacterium tuberculosis*. Cole *et al*, 2001, made the assumption that the genomes of *M. leprae* and *M. tuberculosis* were once topologically

equivalent and aproximately 4.4 Mb in size, as is the case for many other slow-growing mycobacteria. Extensive downsizing must have occurred during evolution of the leprosy bacillus since its genome is less than 75% of the size of that of *M. tuberculosis*. Since diverging from the last common mycobacterial ancestor, the leprosy bacillus may have lost over 2000 genes, and reductive evolution may have defined the minimal gene set for a pathogenic mycobacterium. Based on these data it is reasonable to expect a reduced number of secreted proteins from *M. leprae* as compared to *M. tuberculosis*. For my study I used a total of 204 *M. tuberculosis* protein sequences previously characterized as secreted to search for homologs in the *M. leprae* genome.

1.5.2.Bioinformatics

Bioinformatics is the field of study in which computer algorithms are used to predict properties of proteins. I utilized bioinformatics tools to predict protein location (e.g. cytoplasmic, cell membrane, secreted) of *M. leprae* proteins. The first tool used predicted signal peptides which involves two tasks: 1) given that a sequence has a signal peptide, locate the cleavage site and, 2) discriminate between secretory proteins with signal peptides and non-secretory proteins. Nielsen *et al*, 1997, have developed a combined neural network approach to the recognition of signal peptides and their cleavage sites, using one network to recognize the cleavage site and another network to distinguish between signal peptides and non-signal peptides. It is called SignalP. Nielsen *et al* used a list of known secreted proteins taken from SWISS-PROT and divided them into prokaryotic and eukaryotic entries. The prokaryotic data sets were further divided into Gram-positive eubacteria and Gram-negative eubacteria.

The output from the signal peptide/non-signal peptide networks, the S score, can be interpreted as an estimate of the probability of the amino acid position belonging to a

signal peptide, while the output from the cleavage site/non-cleavage site networks, the C-score, can be interpreted as an estimate of the probability of the amino acid position being the first in the mature protein. If there are several C-score peaks of comparable strength, the true cleavage site may often be found by inspecting the S-score curve in order to see which of the C-score peaks coincides best with the transition from the signal peptide to the non-signal peptide region. In order to formalize this and improve the prediction, Nielsen *et al* defined Y as the geometric average of the C-scores and a smoothed derivative of the S-score. They found the best discriminator between signal peptides and non-secretory proteins to be the average of the S-score in the predicted signal peptide region. If this value, the mean S-score, is greater than 0.5, they predict the sequence in question to be a signal peptide.

Signal anchors often have sites similar to signal peptide cleavage sites after their hydrophobic (transmembrane) region; therefore, a prediction method can easily be expected to mistake signal anchors for peptides. A similar method to predict the location and orientation of transmembrane alpha helices based on a hidden Markov model (TMHMM) was designed by Sonnhammer *et al*, 1998. A model for a transmembrane orientation consists of 3 states: one for inside loops, one for transmembrane regions and one for outside loops. Each state has an associated probability distribution over the 20 amino acids characterizing the variability of amino acids in the region it models. An analysis of the performance of different programs for the prediction of transmembrane regions in proteins by Moller *et al*, 2001, showed that TMHMM is currently the best performing transmembrane prediction program. Eighty-five percent of proteins analyzed with TMHMM have their membrane spanning regions correctly predicted. When the

polarity of their integration into the membrane is included, the number of correct predictions is reduced to 70%.

Gomez *et al*, 2000, used two algorithms, SignalP and SPScan to predict the location of 3,924 *M. tuberculosis* proteins. Cutoff values for the computer predictions were chosen on the basis of scores assigned to nine known secreted proteins of *M. tuberculosis* that contain a signal peptide. 208 proteins that scored above the cutoff with both programs were analyzed with TMpred and PrositeScan and sorted in 3 groups: 52 proteins were classified as most likely secreted, 25 proteins as transmembrane proteins and 16 as lipoproteins. Ten proteins were chosen at random for cloning into a phoA fusion vector and screened for alkaline phosphatase activity. The computer predictions were confirmed in 90% (9 of 10) of the putatively secreted proteins tested by *E. coli phoA* gene fusion methods.

A disadvantage of the computer-based approach is that it is limited to only those proteins secreted via the general secretory pathway. While many secreted antigens of *Mycobacterium tuberculosis* fall into this category, the presence in culture filtrates of proteins lacking secretory signal peptides suggest the existence of other mechanisms of protein secretion in *M. tuberculosis*.

In this study, I used a bioinformatic strategy to predict *M. leprae* secreted proteins. Two different programs (SignalP and TMHMM) were used to identify proteins having secretory signal peptides and appropriate membrane attachment domain for secretion.

1.5.3.Validation of Protein Secretion

Proteins released by *Mycobacterium tuberculosis* into the extracellular environment have been the focus of investigation to identify antigens that may induce

protective immunity or elicit immune responses of diagnostic value. *M. tuberculosis* culture filtrates have been used to identify mycobacterial proteins that induce protection against *M. tuberculosis* infection in animal models (Collins *et al*, 1988; Hubbard *et al*, 1992; Freer *et al*, 1998; Weldingh *et al*, 1998). Culture filtrates have provided a rich array of secreted proteins for study from *M. tuberculosis* and *M. bovis*, however, because *M. leprae* cannot be grown *in vitro* similar culture fractions are not available for study. Andersen *et al*, 1991, classified the proteins present in culture filtrates of *M. tuberculosis* into three major groups: extracellular proteins that accumulate in large quantities in the medium but are present only in trace amounts in the intact bacilli, secreted proteins that are gradually released during growth of the bacilli and cytoplasmic proteins released from dead bacteria during the late logarithmic growth phase. Wiker *et al*, 1991, developed an index for individual mycobacterial antigens expressing the ratio between the amount of each antigen in culture fluid and in sonicate preparations. There was good agreement between signal sequences and measured index.

More recently, genetic approaches have been applied to identify secreted proteins of *M. tuberculosis.* These studies involve screening libraries composed of *M. tuberculosis*-reporter gene fusions. By constructing protein fusions with these reporters and assaying for activity in colonies growing on plates, the location (intracellular vs. extracellular) of a given protein can be determined. Hoffman and Wright, 1985, constructed a fusion vector encoding a protein with a modified form of the *phoA* gene, an easily assayable enzyme normally located in the periplasm of *E. coli* K-12. PhoA can be detected even at low levels in bacterial colonies by use of chromogenic substrate 5-bromo-4-chloro-3-indolyl phosphate (XP); however, its activity is absolutely dependent upon secretion from the cytoplasm. Hoffman and Wright altered the *phoA* gene,

removing its promoter region, ribosome binding site and the complete signal sequence encoding region, and showed that the *phoA* gene can function as a reporter for secreted proteins in *E. coli*. PhoA is also known to function as a reporter for exported and secreted proteins in *M. smegmatis* (Timm *et al*, 1994). Lim *et al*, 1995, screened *M. tuberculosis* DNA-*phoA* fusion libraries respectively by using a *phoA* reporter shuttle plasmid pJEM11 which replicates in *E. coli* and *M. smegmatis*. They found 12 different inserts allowing PhoA expression. Some of the known *M. tuberculosis* proteins identified were the 19kDa lipoprotein, the 28kDa (erp) protein and an enzyme implicated in the biosynthetic pathway of fatty acids. Carroll *et al*, 2000, used the same approach to study *M. avium* secreted proteins. They isolated 100 PhoA recombinants and 15 of these were sequenced, most of them exhibited high degree of homology with known *M. tuberculosis* and *M. leprae* sequences corresponding to phosphate permeases, cutinases, glycosyltransferases, multicopper oxidases and putative invasins. Wiker *et al*, 2000, used computational algorithms in combination with gene fusions to identify secreted proteins, membrane proteins and lipoproteins of *M. tuberculosis*. The computer-algorithms were used to predict the subset of *M. tuberculosis* genes that encode exported proteins. Of the 34 genes identified by the *phoA* method, 22 were classified to encode potential soluble secreted proteins. Among these were some known antigens (fbpB and C, mpt53, mpt64, mtb12) as well as 14 novel secreted proteins. Six of the remaining 12 genes were predicted to encode membrane lipoproteins and an additional six to encode integral membrane proteins.

In this study, I used a genetic approach to validate secretion of proteins predicted for secretion by constructing alkaline phosphatase-ML gene fusions. Both *E. coli* and *M. smegmatis* were selected as surrogate host to validate secretion of *M. leprae* proteins

because the Sec dependent pathway is similar to *M. tuberculosis* and *M. leprae* and these two organisms were used before to study secretion in *M. tuberculosis*.

CHAPTER 2

BIOINFORMATICS

2.1.Introduction

Bioinformatics is a newly emerging interdisciplinary research area which may be defined as the interface between biological and computational sciences. One of the sub-disciplines of bioinformatics is the development of new algorithms and statistics to assess relationships among members of large data sets. As a consequence of the large amount of data produced in the field of molecular biology, most of the current bioinformatics projects address structural and functional aspects of genes and proteins. With the increasing number of bacterial genomes being sequenced demands are being made upon workers to develop new methods to define proteomes of these organisms.

One group of proteins of great importance is the secreted proteins, given their dominant immunogenicity and role in pathogenesis (Closs *et al*, 1980; Harboe and Nagai, 1984; Wiker *et al*, 1986). The large majority of these proteins possess an amino-terminal signal sequence that mediates their membrane translocation via the Sec-dependent general export pathway. Following translocation, cleavage of the signal peptide by a signal peptidase releases the mature protein, provided there are not additional membrane-spanning segments (Murphy and Beckwith, 1987). Three distinct regions comprise the N-terminal signal sequence; the charged N-terminus (n-region), the hydrophobic core (h-region), and the C-terminal cleavage domain (c-region) (von Heijne, 1985). Secreted proteins that follow the Sec-dependent general export pathway can be predicted using computer algorithms.

Analysis of *Mycobacterium tuberculosis* proteins resulted in the observation that the culture filtrate of *in vitro*-grown bacilli contains the majority of dominant antigens

(Abou-Zeid *et al*, 1988; Collins *et al*, 1988; Hubbard *et al*, 1992). This procedure cannot be applied to *M. leprae* because the bacilli do not grow in artificial media and secreted proteins are lost during isolation of bacilli from tissues. To search for secreted proteins of *M. leprae*, we have compared known and putatively secreted protein sequences of *M. tuberculosis* with the *M. leprae* genome. Gomez, *et al*, 2000, identified 52 *M. tuberculosis* proteins by computer-based analysis as most likely secreted. They used SignalP and SPScan to predict potential signal peptides, and TMpred to predict the presence of membrane-anchoring sequences. This approach identified novel secreted proteins that can be characterized for their potential for immunological diagnosis of tuberculosis or vaccine design. In this study, I used bioinformatic tools to predict *M.leprae* secreted proteins. Two different programs were used to identify proteins having secretory signal peptides but lacking additional membrane attachment domains:

- SignalP, Nielsen, *et al*, 1997, predicts the presence and location of signal peptide cleavage sites in amino acid sequences. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of Neural Networks and Hidden Markov models.
- Transmembrane Helix Markov model (TMHMM), Sonnhammer, *et al*, 1998, predicts the location and orientation of transmembrane helices (TMH) in protein sequences. The number and location of TMH predicts whether a protein is secreted or anchored in the membrane.

2.2.Materials and Methods

2.2.1.Genome Database

M. tuberculosis and *M. leprae* DNA and protein sequences were obtained from the Sanger Centre and the National Center for Biotechnology Information (NCBI).

2.2.2.Amino Acid Comparisons of *M. leprae* and *M. tuberculosis* Proteins

Comparisons between *M. tuberculosis* and *M. leprae* completed genomes were restricted to amino acid sequences and were performed using BLASTP (protein vs. protein) server from the Sanger Centre. A total of 204 *M. tuberculosis* protein sequences defined as putatively secreted (Sanger Centre) or previously characterized as secreted (Gomez *et al*, 2000 and Wiker *et al*, 2000) were used to search for homologs in the *M. leprae* genome. Each *M. tuberculosis* known or putatively secreted protein served as a query sequence against the entire database of *M. leprae.* Final selection of ML proteins was restricted to proteins showing more than 40% identity with an *M. tuberculosis* protein.

2.2.3.Computer-based Algorithms

SignalP (www.cbs.dtu.dk/services/SignalP-2.0)

The SignalP World Wide Web server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks. SignalP V2.0 comprises two signal peptide prediction methods, SignalP-NN (based on neural networks, corresponding to SignalP V1.1) and SignalP-HMM (based on hidden Markov models). The Neural Networks are based on "learning" by adjusting the weights in the network until its performance on the training set is acceptable. For example, known secreted proteins are used as database to train the neural network to recognize unknown secreted proteins.

The Hidden Markov Models use general statistical modeling technique for linear problems like sequences or time series and have been widely used in speech recognition

applications for 20 years. An HMM for proteins consists of a number of states that are connected by transition probabilities. Associated with each state is a distribution over the 20 amino acids. The SignalP server will return 3 different scores between 0 and 1 for each position in the sequence: C, S and Y. In addition, the maximal Y-score, maximal S-score and mean S-score values are given for the entire sequence. If the mean S-score is equal or greater than 0.5, SignalP predicts the sequence in question to be a signal peptide.

- C-score (raw cleavage site score): The output score from networks trained to recognize cleavage sites vs. other sequence positions.
- S-score (signal peptide score): The output score from networks trained to recognize signal peptide vs. non-signal peptide positions.
- Y-score (combined cleavage site score): The prediction of cleavage site location is optimized by observing where the C-score is high and the S-score changes from high to a low value. The Y-score formalizes this by combining the height of the C-score with the slope of the S-score.

Up to 70 amino acid long N-terminal sequences were analyzed in my dataset using the program trained for Gram-positive organisms. If a sequence was predicted to have a signal peptide, the cleavage site was predicted to be immediately before the position with the maximal Y-score. Discrimination between signal peptides and non-secretory proteins was done by using the mean value of the S-score, averaged from position 1 to the most likely cleavage site.

TMHMM (www.cbs.dtu.dk/services/TMHMM-2.0)

Transmembrane Hidden Markov Model (TMHMM) is a method to predict the location and orientation of transmembrane helices in protein sequences, it is based on a hidden Markov model. The basic principle is to define a set of states, each corresponding

to a region or specific site in the proteins. In the simplest case, a model for a transmembrane protein may consist of three states: one for inside loops, one for transmembrane regions and one for outside loops. Positively charged residues are predominantly found in loops on the cytoplasmic side. Most transmembrane α helices are encoded by a long stretch of hydrophobic residues, that are suitable for hydrophobic interactions with lipids. Many signal peptides also have one hydrophobic region. The number and location of transmembrane domains predicts whether the protein is secreted or anchored in the membrane. The algorithm provides information about the number of transmembrane domains along the amino acid sequence and also predicts an extracellular or intracellular location of the protein. For example, there are 3 main locations of a residue: in the transmembrane helix core (in the hydrophobic tail region of the membrane), in the transmembrane helix caps (in head region of the membrane), and in loops. A typical secreted protein will show an N-terminal small intracellular portion followed by a transmembrane helix (20 to 40 amino acid long) and the remaining sequence located extracellularly. Some sequences have no transmembrane helix and they are located either extracellularly or intracellularly.

2.2.4.Comparisons between *M. tuberculosis* and *M. leprae* Sequences

Pairwise alignments of amino acid sequences from signal sequences and mature proteins were performed using Omiga™ 2.0 (Oxford Molecular Ltd., www.gcg.com). Identification of N-region (positively-charged amino acids), H-region (hydrophobic amino acids) and C-region (signal peptidase recognition site) was based on SignalP. Transmembrane domains were predicted using TMHMM.

2.3.Results

A total of 204 putatively secreted or known to be secreted proteins of *M. tuberculosis* were used to identify homologous amino acid sequences in *M. leprae*. The BLASTP analysis showed that 52 of the 204 *M. tuberculosis* proteins had no match to proteins in *M. leprae*. Thirty-eight *M. tuberculosis* proteins were represented by pseudogenes or inactive reading frames in the *M. leprae* genome and 114 *M. tuberculosis* proteins showed amino acid identity of greater than or equal to 40% with proteins in *M. leprae*. These 114 *M. leprae* proteins were analyzed to predict secretion with 2 computer algorithms: SignalP to predict the presence and location of signal sequences, and TMHMM to predict the location and orientation of transmembrane helices in protein sequences. Analysis with SignalP identified 32 proteins with a mean S score ≥ 0.5 (Table 1). These 32 protein sequences were analyzed with TMHMM resulting in the selection of 24 sequences designated with high probability of being secreted based on the presence of 1 or zero transmembrane domains in proper orientation with the predicted signal sequence (Table 1, gene name in bold letters). The remaining 8 sequences were predicted to be either transmembrane or intracellular proteins (Table 1). Nineteen proteins (ML0091, ML0097, ML0098, ML0620, ML0885, ML1339, ML1417, ML1633, ML1811, ML2028, ML2055, ML2331, ML2380, ML2450, ML2522, ML2569A, ML2591, ML2598 and ML2659) from a total of 24 putatively secreted proteins had a typical N-terminal intracellular sequence followed by a transmembrane domain and good cleavage site prediction. The remaining five sequences (ML0715, ML1812, ML1923, ML2274 and ML2569) showed 0 transmembrane domains and were predicted by TMHMM to be secreted based on amino acid composition (Table 1) as well.

Table 1: *M. leprae* proteins selected for secretion by SignalP and TMHMM

Gene name	[1]MT Homolog*	[2]Signal peptidase motif	[3]C site	[4]SignalP-mean S	[5]Predicted TMs	[6]Location
ML0041	Rv3883c	Ala-Leu-Ala	21-22	0.92	1 (417-439)	Transmembrane
ML0091	*erp*	Ala-Ile-Ala	22-23	0.95	1 (9-31)	Secreted
ML0097	*fbpA*	Ala-Glu-Ala	42-43	0.69	1 (13-35)	Secreted
ML0098	*fbpC1*	Ala-Lys-Ala	36-37	0.88	1 (7-29)	Secreted
ML0175	Rv0982	Ile-Ser-Ala	49-50	0.96	2 (30-52 and 162-84)	Transmembrane
ML0486	Rv2588c	Ala-Ser-Arg	23-24	0.74	1 (4-21)	Intracellular
ML0575	Rv0849	Cys-Leu-Ala	39-40	0.75	2 (5-27 and 42-64)	Transmembrane
ML0620	*mtb12*	Ala-Pro-Ala	52-53	0.5	1 (5-27)	Secreted
ML0715	*lpqC*	Val-Ser-Ala	26-27	0.73	0	Secreted
ML0885	Rv2190c	Ala-Met-Ala	37-38	0.79	1 (12-34)	Secreted
ML1010	*bcpB*	His-Gly-Ala	24-25	0.5	0	Intracellular
ML1214	Rv1566c	Ala-Tyr-Ala	26-27	0.76	1 (4-26)	Intracellular
ML1339	Rv2672	Ser-Gly-Ala	33-34	0.81	1 (13-35)	Secreted
ML1417	Rv2289	Ala-Glu-Ala	25-26	0.83	1 (5-23)	Secreted
ML1537	Rv1797	Trp-Gln-Ala	39-40	0.8	2 (7-29 and 44-66)	Transmembrane
ML1633	Rv2223c	Val-Arg-Val	56-57	0.56	1 (30-52)	Secreted
ML1811	Rv1478	Ala-Thr-Ala	31-32	0.85	1 (7-29)	Secreted
ML1812	Rv1477	Ala-Thr-Ala	39-40	0.91	0	Secreted
ML1923	*lpqF*	Ala-His-Ser	37-38	0.76	0	Secreted
ML2028	*fbpB*	Ala-Gly-Gly	32-33	0.74	1 (13-35)	Secreted
ML2055	*modD*	Ala-Ala-Ala	37-38	0.63	1 (13-35)	Secreted
ML2274	Rv0559c	Ala-Leu-Ala	25-26	0.78	0	Secreted
ML2331	Rv3717	Ala-Val-Ala	31-32	0.89	1 (13-32)	Secreted
ML2380	Rv0455c	Ala-Val-Ala	30-31	0.84	1 (7-29)	Secreted
ML2450	Rv0479c	Ile-Gly-Ala	26-27	0.76	1 (7-29)	Secreted

(Table continued)

ML2522	Rv0309	Ala-Gly-Ala	30-31	0.95	1 (5-27)	Secreted
ML2569	Rv0237	Ala-Gln-Ala	27-28	0.76	0	Secreted
ML2569A	Rv0236A	Val-Gln-Gln	32-33	0.77	1 (7-29)	Secreted
ML2591	*mce1C*	Leu-Phe-Ala	38-39	0.56	1 (13-35)	Secreted
ML2598	Rv0178	Ala-Gly-Ala	39-40	0.64	1 (20-42)	Secreted
ML2659	Rv0125	Gly-Ser-Ala	32-33	0.87	1 (13-35)	Secreted
ML2664	Rv0116c	Ser-Val-Ala	28-29	0.69	2 (7-29 and 39-58)	Transmembrane

1 Gene names for proteins from MT that have 50% or greater amino acid identity with matched ML protein
2 The signal peptidase motif represents the site recognized by the signal peptidase. Ala-X-Ala is the conserved motif for signal peptidase.
3 C site represents the amino acid position at which the signal peptidase cleavage site was predicted
4 This value represents the mean of signal peptide scores.
5 TM represents number of transmembrane domains for each protein with amino acid position given in parenthesis.
6 Predicted location of ML proteins based on analysis by SignalP and TMHMM algorithms

A group of 8 proteins were identified as most likely not secreted based on physical characteristics including amino acid composition and transmembrane domains. ML1010 (no transmembrane domain), ML0486 and 1214 (1 N-terminal transmembrane domain each) were composed of amino acid seen primarily in proteins located intracellularly. Analysis of ML0041 showed 1 C-terminal transmembrane domain with the remainder of the protein predicted to be located extracellularly. Four sequences (ML0175, ML0575, ML1537 and ML2664) had 2 transmembrane domains and, therefore, were classified as putative membrane proteins.

The Ala-X-Ala motif for the signal peptidase was conserved only in 15 of 24 sequences that were predicted to be secreted by SignalP and TMHMM (Table 1). Gly or Ser in position -1 was found in 2 sequences and Val or Ser in position -3 was found in 5 sequences as alternative signal peptidase recognition sites (von Heijne, 1983). Only 1 sequence (ML2569A) from a total of 24 had an amino acid different from Ala, Gly or Ser in position -1 and 3 sequences (ML2450, ML2591 and ML2659) had an amino acid different from Ala, Val, Ser or Thr in position -3. The signal sequences ranged in size from 21 to 56 amino acids (Table 1).

A comparison between each of the 24 *M. leprae* proteins and their *M. tuberculosis* counterparts was performed using the amino acid alignment program from Omiga 2.0. Fifteen *M. leprae* proteins showed 50% or greater amino acid homology for the mature protein (not including signal sequence) with the corresponding *M. tuberculosis* sequences (Table 2). Fourteen of 15 alignments showed that the amino acid identity between mature proteins from *M. leprae* and *M. tuberculosis* homologs was considerably higher than the amino acid identity between signal sequences (Table 2).

Table 2: Comparison of the signal sequences and the mature proteins from *Mycobacterium leprae* and *Mycobacterim tuberculosis* predicted secreted proteins.

Gene name	[1]Length	[2]MT homolog	Length	[3]Signal seq. (%Id/%Sim)		[4]Mature prot. (%Id /%Sim)	
ML0091	236	*erp*	284	73	86	55	67
ML0097	330	*fbpA*	338	63	74	86	92
ML0098	301	*fbpC1*	299	44	69	82	89
ML0620	167	*mtb12*	168	40	50	58	70
ML0715	304	*lpqC*	304	58	62	73	81
ML0885	374	2190c	385	27	43	60	72
ML1811	241	1478	241	55	68	80	85
ML1812	479	1477	472	72	79	79	87
ML1923	454	*lpqF*	452	44	47	82	84
ML2028	327	*fbpB*	325	67	77	85	91
ML2274	112	559c	112	64	76	71	82
ML2331	256	3717	241	58	74	77	84
ML2380	153	455c	148	57	63	68	76
ML2569	387	237	388	63	74	74	82
ML2591	519	*mce1C*	515	71	82	73	83

1 Length of protein in amino acids.
2 MT protein that has 50% or greater amino acid identity with matched ML protein.
3 Comparison alignments between ML and MT signal sequences. Percentages of amino acid identities and similarities are indicated for each comparison.
4 Comparison alignments between ML and MT mature proteins (signal sequences removed). Percentages of identities and similarities are indicated for each comparison.

Signal sequences from *M. leprae* and *M. tuberculosis* were aligned to compare the length and percentage of arginine in the N-region. Six out of 15 pairs comparisons showed identical arginine content in the N-region for *M. leprae* and *M. tuberculosis* signal sequences (Figure 2). Two signal sequences (ML0620 and ML2274) showed no arginine in the signal sequence as did their MT homologs. The remaining ML signal sequences ranged in arginine content from 10% to 37.5%. The mean of the arginine content in the

N-region was 22% for *M. tuberculosis* sequences and 19% for *M. leprae* sequences (Figure 2).

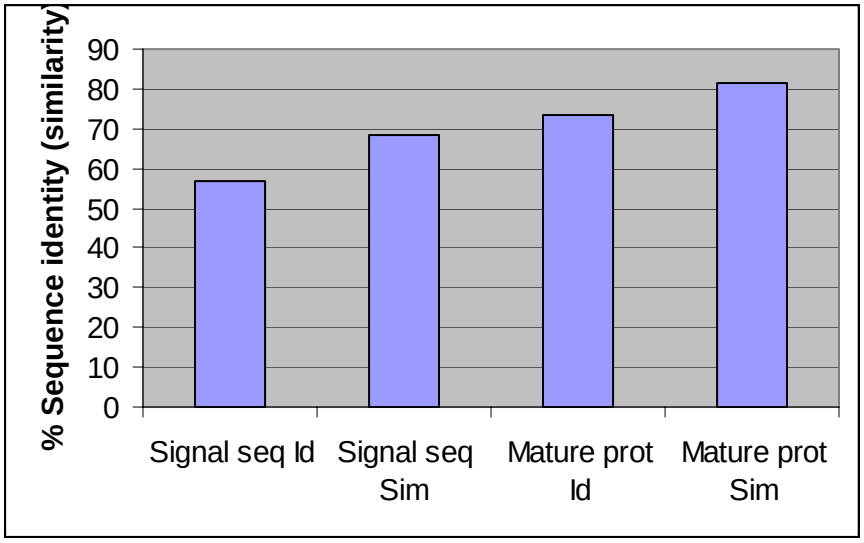


Figure 1: Comparison of amino acid sequence identity and similarity between ML and MT proteins.

	Signal sequence	Percentage of arginine (R) in the N-region
ML0091	MPNRRRCKLSTAISTVATLAIA	37.5%
Erp	MPNRRRRKLSTAMSAVAALAVA	50%
ML0097	MKFVDRFRGAVAGMLRRLVVEAMGVALLSALIGVVGSAPAEA	23.5%
fbpA	MQLVDRVRGAVTGMSRRLVVGAVGAALVSGLVGAVGGTATAGA	23.5%
ML0098	MRGLSAVVRVLCVAALAVGVFAAAVLLAGTAGNAKA	22%
fbpC1	MKGRSALLRALWIAALSFGLGGVAVAAEPTAKA	22%
ML0620	MTMKSIATYAALAIIGAAV	0%
Rv2376c	MKMVKSIAAGLTAAAAIGAAA	0%

Figure 2: Comparison alignments of 15 signal sequences of predicted secreted proteins in *M. leprae* and *M. tuberculosis*.
N-region in blue, hydrophobic region in black, cleavage site in red and lipoprotein motif in green. Identification of N-region (positively-charged amino acids), H-region (hydrophobic amino acids) and C-region (signal peptidase recognition site) was based on TMHMM prediction. Color printer needed to reproduce figure accurately (Figure continued).

ML0715	MNVARWLASVVLAVCLAGCVGRQVSA	17%
lpqC	MPWARMLSLIVLMVCLAGCGGDQLLA	17%
ML0885	MRLGCKHPVVRLIAHFVVGTLVGFAVLSRFFVATAMA	13%
Rv2190c	MRLDQRWLIARVIMRSAIGFFASFTVSSGVLAANVLA	27%
ML1811	MRHKNFRLINLAGLTAMVAGLIIVVTIPATA	20%
Rv1478	MRHTRFHPIKLAWITAVVAGLMVGVATPADA	20%
ML1812	MKRPRRGSVSRPTARFVRPAIPSLVSAALLVSLPVLATA	33%
Rv1477	MRRNRRGSPARPAARFVRPAIPSALSVALLVCTPGLATA	40%
ML1923	MPQPARRTNQRPPRHRTVALAATAALVVTMAPGCAHS	31%
lpqF	MGPARLHNRRAGRRMLALSAAAALIVALASGCSSAPTPSANA	36%
ML2028	MIDVSGKIRAWGRWLLVGAAATLPSLISLAGG	14%
fbpB	MTDVSRKIRAWGRRLMIGTAAAVVLPGLVGLA	29%
ML2274	MKGTGLAANVAMAAAATVLAAPALA	0%
Rv0559c	MKGTKLAVVVGMTVAAVSLAAPAQA	0%
ML2331	MNTRVSLRIGFRMVVGLLVAALTTITPTAVA	25%
Rv3717	MIVGVLVAAATPIISSASATPANIAGMVVFI	0%
ML2380	MSRLSTSLCKGAVFLVFGIIPVAFPTTAVA	10%
Rv0455c	MSRLSSILRAGAAFLVLGIAAATFPQSAAA	20%
ML2569	MAFPRTLIVLAAASALVVTCGHDVAQA	17%
lpqI	MAFPRTLAILAAAAALVVACSHGG	17%
ML2591	MRTLELPNRLRSGLIGVLVVLLIIGVGQSFTSVPILFA	27%
mce1C	MRTLEPPNRMRIGLMGIVVALLVVAVGQSFTSVPMLFA	27%

2.4.Discussion

204 *M. tuberculosis* known secreted and putatively secreted proteins were used to search for homologous *M. leprae* proteins using bioinformatic tools. Fifty-two *M. tuberculosis* proteins did not show any matches with the *M. leprae* genome while 38 *M. tuberculosis* proteins had homologies with pseudogenes in the *M. leprae* genome. The large number of pseudogenes was not unexpected based on the previously reported relatively low number of functional genes and the high number of pseudogenes in *M. leprae* (Cole *et al*, 2001) (Table 3).

Table 3: Comparison of *M. leprae* and *M. tuberculosis* genome features.

Feature	***M. leprae***	***M. tuberculosis***
Genome size (bp)	3,268,203	4,411,532
G+C%	57.79	65.61
Protein-coding genes	1,604	3,959
Pseudogenes	1,116	6
Predicted secreted proteins	24	52*
G+C% secreted prot. genes	60.5	66

*Gomez et al, 2000.

One hundred and fourteen *M. leprae* proteins were identified using the *M. tuberculosis* protein homolog search. Analysis with SignalP and TMHMM identified 24 *M. leprae* proteins meeting the criteria for secretion. Five of these genes encode authentic secreted proteins in *M. tuberculosis* (*fbpA, B and C*, *erp* and *mtb12*) supporting the validity of this approach for selecting secreted proteins from *M. leprae*.

TMHMM is currently one of the best performing transmembrane prediction programs and is especially good at reliably distinguishing between soluble intracellular and transmembrane proteins (Moller *et al*, 2001). Eighty-five percent of biochemically characterized membrane proteins analyzed with TMHMM by Moller *et al*, 2001, were predicted correctly. The number of predicted secreted proteins for *M. tuberculosis* and *M. leprae* (Table 3) resulting from the computer-based approach appears low compared to the 200 plus *M. tuberculosis* proteins found in culture filtrates. However, the low number of predicted ML secreted proteins does correlates with the fact that the algorithms used in this study detect only those proteins secreted via the Sec pathway. While many secreted antigens of *M. tuberculosis* are translocated via the Sec pathway, the presence in culture filtrates of proteins lacking secretory signal peptides (ex.; ESAT-6, *sodA*, *glnA* and *katG*) suggests the existence of other mechanisms of protein secretion in *Mycobacterium spp*. For example, ESAT-6 secreted antigen, which is present in *M. tuberculosis*, virulent *M. bovis* and has been recently identified in *M. leprae* (Spencer *et*

al, 2002), is a member of a family of 100-amino acid proteins with sequence similarity focused on a central WXG motif. Members of this family have been identified in several Gram-positive pathogens (e.g.; *S. aureus* and *B. anthracis*) and may be components of a surface-located secretion apparatus. By analogy with Gram-negative mechanisms of secretion, this apparatus might be assembled only under specific conditions (e.g., contact with host cell) (Pallen, 2002). Since *M. leprae* has homologs of ESAT-6, *sodA* and *glnA,* this secretory system may also be functional in *M. leprae* adding to its potential number of total secreted proteins.

Four of the 24 *M. leprae* proteins predicted to be secreted using SignalP and TMHMM were previously described *M. tuberculosis* and *M. leprae* proteins: ML0097 fibronectin-binding protein (*fbpA*), ML2028 (*fbpB*) and ML0098 (*fbpC*); and ML2591 (*mce1C*), one of the genes of the mammalian cell entry operon. The fibronectin-binding proteins (*fbpA*, *fbpB* and *fbpC*) stimulate the uptake of mycobacterial bacilli by human macrophages by interacting with the gelatin binding site of human fibronectin (Abou-Zeid *et al*, 1988) and also have cell wall mycolyltransferase activity (Belisle, *et al*, 1997).

Arruda *et al*, 1993, described a DNA fragment of *M. tuberculosis* that conferred to a non-pathogenic *E. coli* strain the ability to gain entry into mammalian cells and to survive inside macrophages. The mammalian cell entry (*mce*) gene was termed *mce1*. Analysis of the *M. tuberculosis* genome revealed four copies of *mce*, situated in operons of 8 genes each.

The *erp* gene known as exported repetitive protein was identified in *M. bovis* culture supernatants by Western blot (Bigi et al, 1995). Erp is a secreted antigen from *M. tuberculosis* required for virulence (Berthet et al, 1998). Together the fibronectin-binding proteins (*fbpA, fbpB and fbpC*), *mce1C* and *erp* represent secreted virulence factors of *M.*

tuberculosis and *M. bovis*. My study shows that they are also predicted to be secreted in *M. leprae,* however, further studies will be necessary to confirm whether they represent virulence factors in *M. leprae*.

The generally accepted Ala-X-Ala motif for signal peptidase was conserved in 15 of 24 ML sequences that were predicted to be secreted by SignalP and TMHMM (Table 1). Gly or Ser in position -1 was found in 2 sequences and Val or Ser in position -3 was found in 5 sequences as alternative signal peptidase recognition sites (Saleh, *et al*, 2001). Only 1 sequence (ML2569A) from a total of 24 had an amino acid different from Ala, Gly or Ser in position -1, and 3 sequences (ML2450, ML2591 and ML2659) had an amino acid different from Ala, Val, Ser or Thr in position -3 (Table 1). These results are in agreement with Wiker, *et al*, 2000, who showed that *M. leprae* signal peptides have similar cleavage site motifs to those found in *M. tuberculosis* signal peptides. Four sequences from a total of 24 do not fit the model described by Saleh *et al*, 2001. It is possible that signal peptidases from *M. leprae* may accept amino acids not present in the conventional model described for cleavage site motif.

Fifteen *M. leprae* proteins showed 50% or greater homology for the mature protein (not including signal sequence) with the corresponding *M. tuberculosis* sequences meeting my requirement for further analysis (Table 2). Fourteen of these alignments showed that the homology between mature proteins is considerably higher than the homology between signal sequences. The mean identity between ML and MT homologs for mature proteins was 73.5% and only 57% for signal sequences. These data correlate with the fact that signal peptides from Gram-positive bacteria are considerably longer and more variable than those from other organisms (Nielsen *et al*, 1997). Harboe and Wiker, 1999, searched for secreted proteins from *M. leprae* by comparing DNA sequences from

M. tuberculosis and *M. leprae* prior to the completion of the ML genome. They identified only 10 sequences (*fbpA, fbpB, fbpC, mpt32, mpt51, erp, mtc28, mtb12*, Rv3354 and Rv0526) with high homology between *M. tuberculosis and M. leprae*. My study showed that 5 of 10 sequences predicted to be secreted by Harboe and Wiker (*fbpA, fbpB, fbpC, erp* and *mtb12*) plus 19 other proteins were predicted to be secreted using the criteria set for secretion. Harboe and Wiker's analysis was performed prior to the completion of the *M leprae* genome leaving their analysis incomplete. In addition, they used an older version of SignalP and no algorithm to predict the presence of transmembrane domains. Nevertheless, their conclusions and my results are in agreement. For example, they reported that the genomic organization of genes for secreted proteins was similar in *M. leprae* and *M. tuberculosis* and that amino acid homologies between ML and MT secreted proteins was found to be higher for the mature polypeptide chains than for the corresponding signal peptides.

My study showed that differences between amino acid identities of ML and MT signal sequences were 15.4 %. The same comparison for ML and MT mature proteins was 8.13 %. These data are consistent with the concept that similar amino acids can substitute for various signal sequence motifs (positively-charged, hydrophobic, etc.) Similarly, mature proteins may allow fewer amino acid substitutions because of stringent functional and structural requirements needed for structural or enzymatic activity.

Six out of 15 pairs of amino acid sequences showed the same percentage of arginine at the N-region for *M. leprae* and *M. tuberculosis* proteins (Figure 1). The arginine rich N-regions are most likely explained by the high G+C content found in mycobacterial genomes and related preferred codon usage. It is interesting to note that even though *M. leprae* has a much reduced overall genome G+C content (57.8%)

compared to *M. tuberculosis* (65.6%), several secreted proteins identified in this study had G+C% ratios in the order of 60% (Table 3). One possible explanation for the difference between ML G+C content and MT G+C content might be that the ML genome suffered a considerable loss of G+C in pseudogenes during reductive evolution. This assumption agrees with the fact that the G+C content of functional ML genes is higher than the G+C content of the overall ML genome (Table 3).

In summary, 204 *M. tuberculosis* known and predicted putatively secreted proteins were used to search for homologous *M. leprae* proteins. The computer-based analysis showed that 24 *M. leprae* proteins were predicted to be secreted. These data suggest that Sec-dependent secretion is operative in *M. leprae* and is closely related to commonly described Sec-dependent secretion in other Gram-positive microorganisms. The cleavage-site motif for the signal peptidase was conserved in 20 of 24 ML proteins studied and showed that *M. leprae* signal sequences were very similar to *M. tuberculosis* signal sequences.

Fifteen *M. leprae* proteins showed 50% or greater homology with their corresponding *M. tuberculosis* gene sequences and were selected for gene expression studies (Chapter 3). Comparison alignments between *M. tuberculosis* and *M. leprae* protein sequences showed that homologies between mature proteins are considerable higher than homologies between signal sequences. Analysis of amino acid composition of the N-region for both *M. leprae* and *M. tuberculosis* showed a high arginine content which is consistent with the high G+C content of mycobacterial genes. In terms of amino acid composition, my analysis showed that signal sequences from *M. leprae* were similar to *M. tuberculosis* signal sequences.

CHAPTER 3

M. LEPRAE GENE EXPRESSION AND SECRETION IN *E. COLI* AND *M. SMEGMATIS*

3.1.Introduction

Secreted proteins represent a distinct group of proteins with respect to their structure, function and contribution to virulence. They are of particular importance for vaccine development because they are often immunogenic and have the potential to be recognized early in infection. A requisite step in protein secretion is protein translocation across the cytoplasmic membrane. This step is common to proteins that are released to the extracellular space or remain associated with the cell wall. The Sec-dependent pathway translocates precursor proteins containing N-terminal signal sequences across the cytoplasmic membrane (Oliver and Beckwith, 1981).

Secretion systems in mycobacteria are important for the proper localization of structural and metabolic components as well as virulence factors. However, little is known about this aspect of mycobacterial physiology. In particular it is not known how proteins escape from the cell wall of mycobacteria or how proteins lacking N-terminal signal sequences are exported. The first evidence of a Sec-dependent pathway in mycobacteria was the recognition of N-terminal signal peptides in the predicted amino acid sequences of known *M. tuberculosis* and *M. leprae* secreted proteins (Harboe and Wiker, 1999; Gomez et al, 2000). The second indication that this pathway existed in *M. tuberculosis* and *M. leprae* came from the identification of genes (Cole, *et al*, 1998) with sequence homology to known proteins involved in Sec-dependent secretion in *E. coli* (Braunstein, *et al*, 2001).

The completion of the *M. leprae* genome sequence has provided investigators an opportunity to use bioinformatics tools to predict gene function by comparing translated protein sequences of *M. leprae* with those of other bacteria. Comparative genome analysis

has provided new information about the Sec-dependent pathway in *M. tuberculosis and M. leprae* including the presence of two SecA homologues (SecA1 and SecA2) (Braunstein *et al*, 2001). Comparative genomics coupled with other bioinformatics tools, such as SignalP and TMHMM can predict protein location in the bacterial cell and may aid approaches designed to identify proteins with diagnostic or vaccine potential. As powerful as these tools are they can only make predictions based on characteristics previously identified in proteins with predetermined functions such as secretion. New unknown proteins may meet some but not all of the critical characteristics sought and, therefore, may be excluded from analysis. Alternatively, a protein may have unique characteristics not recognized by a particular algorithm and, therefore, may not be secreted *in vivo* even thought it is predicted for secretion in silico. Therefore, experimental validation is critical when using bioinformatics tools for identifying proteins with potential for secretion.

Once various genes have been identified using bioinformatics tools it is important to verify that the gene is expressed during intracellular growth. Identifying *M. leprae* genes expressed during infection helps focus attention on bacterial factors necessary for survival and growth within the human host, thereby increasing our understanding of the host-pathogen interaction. Gene expression has been studied in *M. tuberculosis* during infection by reverse transcription-polymerase chain reaction (RT-PCR) and microarray analysis (Mariani *et al*, 2000; Triccas and Gicquel, 2000). Virulent mycobacteria must adapt to adverse conditions encountered during the infectious process. This means, at a mechanistic level, that the amount of some bacterial proteins must be increased in response to the changing environment and the amount of others must be lowered. The inability to culture *M. leprae* represented a major obstacle to obtaining sufficient quantities of mRNA to perform global gene expression analysis using conventional approaches. Just recently, new techniques

to isolate highly viable bacteria from infected foot pad tissue have evolved that make possible gene transcription analysis in *M. leprae* using gene specific RT-PCR based analysis of mRNA (Truman and Krahenbuhl, 2001).

The most direct approach for studying secretion is to identify expressed proteins in the extracellular space experimentally. For example, secretion can be studied in cultivable bacteria by analyzing cell-free culture filtrates containing exported proteins. While this approach has successfully identified major groups of secreted proteins from *M. tuberculosis*, it cannot be applied to studying ML proteins since *M.leprae* cannot be grown *in vitro*. Secreted proteins from *M.leprae* surely exist and are exported to the surrounding tissues during infection; however, these proteins are lost when *M.leprae* is isolated from infected tissues. Hibridization *in situ* with antibodies is another approach for studying secretion in intracellular bacteria because it can be done in tissue. A new approach for studying secretion is to take advantage of the properties of enzymes (e.g. alkaline phosphatase and β-lactamase) that require translocation across the cell membrane to be detected. These genes can be fused to unknown genes and act as a reporter in biological systems. For example the alkaline phosphatase (*phoA)* gene lacks a promoter, ribosome binding site (RBS) and export leader sequence. Fusion of *phoA* to an *M. leprae* gene supplying the missing upstream components will result in secreted proteins which when expressed will have detectable PhoA activity. Identification of reporter activity in bacterial colonies implies that the cloned fragment supplies all three functions and, therefore, contains a portion of a gene encoding a secreted protein. Alkaline phosphatase fusions have been used successfully to identify *M. tuberculosis* (Lim *et al*, 1995) and *M. avium* exported proteins (Carroll *et al*, 2000) in a surrogate organism, *M. smegmatis*. A similar approach using a β-lactamase gene and an upstream tac

promoter was used to identify *M. tuberculosis* exported proteins in *E. coli* (Chubb *et al*, 1998).

This chapter describes the results from gene transcription studies of 15 ML proteins predicted to be secreted using SignalP/TMHMM (Chapter 2). The mRNA was purified from *M. leprae* grown in nude mice and gene transcription was monitored by RT-PCR. Each gene was cloned into 2 separate reporter plasmids in an attempt to demonstrate proteins secretion in either *E. coli* or *M. smegmatis*.

3.2.Materials and Methods

3.2.1.Bacterial Strains and Culture Conditions

M. leprae T-53 was isolated from a lesion of an untreated lepromatous leprosy patient in Thailand in 1982 and maintained in serial-passage in the hind foot pads of athymic nude mice (Hsd:Athymic Nude, Harlan, Indianapolis, IN) at the Laboratory Research Branch of the National Hansen's Disease Programs, Baton Rouge, LA.

E. coli XL-1 Blue supercompetent cells (Stratagene, La Jolla, CA) were cultured in LB Lennox agar or broth (Life Technologies, Rockville, Maryland) supplemented with antibiotics (ampicillin, 100 ug/ml or kanamycin, 50 ug/ml, Sigma-Aldrich Co., St Louis, MO). *M. smegmatis* mc^2 155 (ATCC, Rockville, Maryland) was cultured in Middlebrook 7H9 broth supplemented with Tween 80 (Difco Laboratories, Detroit, MI) for electroporation and Middlebrook ADC (Becton, Dickinson and Company, Sparks, MD) or Luria Agar Miller's modified (Alpha Biosciences, Baltimore, MD) supplemented with kanamycin 50ug/ml. Both media contained 5-Bromo-4-Chloro-3-Indolyl Phosphate (BCIP) at 40 ug/ml (Sigma-Aldrich Co.) for PhoA screening.

3.2.2.Plasmids

pQUANTagen (Qbiogene, Carlsbad, CA) is a fusion vector that contains a mutated version of the *E. coli* alkaline phosphatase gene under the transcriptional control of the synthetic tac promoter and the lacI repressor (Figure 3). The alkaline phosphatase gene is out of frame in the vector and the reading frame is restored upon correct insertion of a cloned DNA fragment. A BamHI-BglII site containing the sequence for the signal peptide for PhoA in pQUANTagen that directs the fusion protein to the periplasm in pQUANTagen was removed to study secretion and is referred to as pQUANTagen-sphoA. The plasmid also carries the β-lactamase gene conferring ampicillin resistance for selective growth in *E. coli*.

pJEM11 was a gift from Denis Portnoi, Pasteur Institute, France. It has a *phoA* reporter shuttle plasmid, origins of replication (ori) for *E. coli* and mycobacteria and a multiple cloning site upstream of the *phoA* gene. The selectable marker is a kanamycin resistance gene (Km). The truncated *phoA* gene is devoid of a promoter, start codon and signal sequence. The expression and exportation of PhoA depends on translational fusion with amino termini of other proteins encoding the necessary sequences.

3.2.3.Purification of *M. leprae*

Bacteria were harvested from foot pad tissue using a modification of a previously described protocol by Truman and Krahenbuhl, 2001. Briefly, the hind feet were soaked in ethanol for 1 min and the granulomatous foot pad tissue was removed, minced to a uniform consistency with curved scissors and homogenized thoroughly for 1 min in a sterile tissue grinder (Fisher/Pyrex, Houston, TX) containing 5 ml of cold Middlebrook 7H12 medium (Difco/Becton Dickinson, Sparks, MD). Excess tissue was removed by centrifugation at 200 x g for 1 min at 4° C and the bacteria remaining in the supernatant fluid were pelleted at 10,000 x g for 10 min at 4° C. The pellet was resuspended in 10.5 ml Middlebrook 7H9 broth

(Difco/Becton Dickinson) containing ampicillin (50ug/ml) and held for 3 hours at 37° C. An aliquot was removed for acid-fast staining and counting (BBL7 TB Ziehl-Neelsen Kit, Becton Dickinson Microbiology Systems). All *M. leprae* preparations were analyzed and determined to be free of microbial contaminants by culturing the final *M. leprae* suspension on a variety of media including: blood agar, Lowenstein-Jensen, thioglycolate broth and trypticase soy broth (Becton Dickinson Microbiology Systems, Cockeysville, MD) for up to 2 weeks.

3.2.4.Purification of *M. leprae* and *M. smegmatis* RNA

RNA was purified from 2 x 10^{10} *M. leprae* T-53 or *M. smegmatis* mc^2 155 using a modification of a previously described protocol by Hellyer *et al*, 1999. Aliquots containing 2 x 10^9 *M. leprae* or *M. smegmatis* were transferred to sterile RNAase-free 2 ml microcentrifuge tubes (Sarstedt, Newton, NC) and pelleted by centrifugation at 10,000 x g for 10 min at 4° C. Bacterial pellets were resuspended in 80 µl of DEPC-H_2O (Sigma-Aldrich Chemical Co. St. Louis, MO). TRIzol® Reagent (Invitrogen, Carlsbad, CA) (920 µl) was added to each tube to maintain the RNA integrity while disrupting cells and dissolving cell organelles; the contents were transferred to a FastRNA® Blue tube (FastRNA™Kit-Blue, Qbiogene, Carlsbad, CA). Bacterial cells were lysed and total RNA was extracted by homogenization in a FastPrep® FP120 Instrument (Qbiogene) for 45 sec at a speed setting of 6.5. Tubes were cooled for 5 min, then this procedure was repeated and the tubes were held on ice for 5 min. A 200 µl aliquot of chloroform/isoamyl alcohol (CIA) (24:1) (Sigma Aldrich) was added to each tube and tubes were mixed by vortex for 1 min. The glass matrix was pelleted by centrifugation at 700 x g for 5 min and the top, aqueous layer was transferred to a new 1.5 ml microcentrifuge tube. Cleanascite HC (CPG Biotech, Lincoln Park, NJ) (100 µl) was added to the each tube and the tubes were mixed on a rocking platform for 10 min.

The Cleanascite was pelleted by centrifugation at 10,000 x g for 1 min and the supernatant fluids were added to fresh 1.5 ml tubes containing 500 µl CIA and vortex mixed for 10 sec. The phases were separated by centrifugation at 10,000 x g for 2 min and the top aqueous phase was transferred to a 1.5 ml microcentrifuge tube. GlycoBlue™ co-precipitant, (Ambion, Inc., Austin, TX) (1 µl), 1/10 volume of 5M NH_4OAc and an equal volume of cold isopropanol were added. The tubes were mixed and the RNA was precipitated at –80° C overnight. RNA was collected by centrifugation at 10,000 x g for 10 min at 4° C and the pellets were washed in 500 µl of salt/ethanol wash solution (FastRNA™ Kit-Blue), air-dried for 10 min and dissolved in 50 µl of DEPC-treated water (FastRNA™ Kit-Blue). RNA was pooled and DNA was removed from these preparations using the DNA-*free*™ kit (Ambion, Austin, TX). Briefly, 50 µl aliquots of RNA were treated with 2 units of DNase (Ambion) at 37ºC for 1 hr, the reaction was stopped with Dnase inactivation reagent (Ambion) and centrifuged for 1 minute at 10,000 rpm to pellet the reagent. Purified RNA was transferred to a new 1.5 ml tube and stored at -80°C**.**

3.2.5.Reverse Transcription of *M. leprae* RNA

Total RNA was converted to cDNA with random hexamer primers and MMLV reverse transcriptase (Advantage RT-for-PCR Kit, BD Biosciences, Clontech, Palo Alto, CA), according to manufacturer's recommendations. Briefly, primers and RNA were incubated at 70ºC for 2 minutes followed by 5 minutes at 4ºC. The master mix containing the reverse transcriptase was added and the reaction was incubated at 44ºC for 1 hr, heated at 94ºC for 5 minutes to stop cDNA synthesis and to destroy any DNase activity; and resuspended in a final volume of 100 µl. A control for monitoring genomic DNA contamination in the RNA extract consisted of RNA incubated with the reverse transcription reagents as described above excluding the reverse transcriptase (RT-). cDNA was also made

from BALB/c mouse spleen total RNA (BD Biosciences, Clontech) and human peripheral blood mononuclear cell RNA. These cDNAs were used as specificity controls for *M. leprae* genes by PCR assays.

3.2.6.PCR Amplification

M. leprae T-53 (10^7 AFB) were disrupted by 3 cycles of freezing and thawing at -70ºC for 15 minutes and 95ºC for 5 minutes. DNA in the disrupted bacterial suspension was used as template to amplify 15 ML predicted secreted genes that were found to have > 50% amino acid identity with *M. tuberculosis* proteins (Table 2). PCR primers and amplification protocols were designed for *M. leprae* genes by acquiring gene sequences from the *M. leprae* genome database at the Sanger Centre (www.sanger.ac.uk) and using Omiga™ 2.0 Primer Design software (Oxford Molecular Ltd, Madison, WI). PCR assays were initially characterized for specificity using 1 ng *M. leprae* T-53 DNA and mouse spleen cDNA. PCR fragments were separated by gel electrophoresis on 2% NuSieve GTG-SeaKem GTG (1:1) agarose gels (BioWhitaker, Rockland, ME) in TAE buffer (4 M Tris-acetate, pH 8.0, 1 mM EDTA). Ethidium bromide-stained gels were visualized by UV transillumination and photographed using a GelDoc®2000 Instrument (Bio-Rad Systems, Hercules, CA). The amplicons were purified using QIAquick PCR columns (QIAGEN, Valencia, CA) and DNA sequences of PCR fragments were obtained by automated DNA sequencing on an Applied Biosystems sequencer (GeneLab, SVM, LSU, Baton Rouge, LA). Transcription analysis was performed by PCR (40 cycles) using cDNA from *M. leprae* as template and gene specific primers for amplification. PCR conditions were denaturation at 94º C for 30 seconds, annealing for 1 minute at Tº (temperature determined by Omiga 2.0 Primer Design software) and primer extension at 72º C for 2 minutes.

3.2.7.Cloning ML Genes into pQUANTagen

M. leprae T-53 DNA was used as template for initial gene specific PCR amplification. PCR conditions were 40 cycles of denaturation at 94º C for 30 seconds, annealing for 1 minute temperature determined by Omiga, 1st PCR, Table 4 and primer extension at 72º C for 2 minutes. The PCR products were separated and purified as described above. A 2nd PCR, used to amplify the gene of interest was performed on the purified amplicon from the 1st PCR as template using primers containing restriction enzyme sites for cloning into pQUANTagen (Figure 3). PCR conditions were the same as those used for the 1st PCR with the annealing temperature set at 60º C for all of the genes amplified in the 2nd PCR. Fifteen *M. leprae* genes encoding predicted secreted proteins were amplified from the start codon (ATG or GTG) to a few bases before the stop codon for cloning into pQUANTagen. Amplification primers included two different restriction sites for unidirectional cloning except when using BamHI with BglII (Table 4). PCR products were purified and digested with appropriate restriction enzymes (New England Biolabs, Beverly, MA) for cloning (Table 4) and separated by gel electrophoresis on 0.8% NuSieve GTG-SeaKem GTG (1:1) agarose gels. Restricted pQUANTagen vector and PCR fragments were purified from agarose gels using QIA quick Gel Extraction kit, Qiagen. Ligation was performed at a molar ratio of 2:1 (insert:vector) with 1unit of T4 ligase (Invitrogen, Carlsbad, CA) at 14ºC overnight. Recombinant plasmid DNA was introduced into *E. coli* XL-1 blue supercompetent cells by transformation (Stratagene, La Jolla, CA). Briefly, 100ul of cells were placed in prechilled 15 ml tubes for each transformation. β-mercaptoethanol (Stratagene) 1.7ul was added and cells were held on ice for 10 minutes. The ligation mixture (10ul) was added to each tube and then held on ice for 30 minutes. The tubes were incubated at 42ºC for 45 seconds (heat pulse) and then held on ice for 2 minutes. SOC medium

(Qbiogene, Carlsbad, CA) (0.9 ml) was added to each tube followed by incubation at 37ºC for 1 hour with shaking at 225 rpm. The transformation mixture was plated on Luria agar, supplemented with ampicillin and BCIP. ML gene-pQUANTagen recombinant plasmids were screened for *phoA* expression in *E. coli* XL-1 blue cells. Blue colonies were recovered at day 2 and restreaked to obtain single-colony isolates. Transformants were screened for carriage of pQUANTagen recombinant derivatives with ML gene insert by PCR of frozen and thawed (3X) bacterial lysates. Primers for PCR detection of ML insert (Figure 2) were designed using Omiga™ 2.0. DNA sequences of PCR fragments were obtained by automated DNA sequencing on an Applied Biosystems sequencer (GeneLab, SVM, LSU, Baton Rouge, LA).

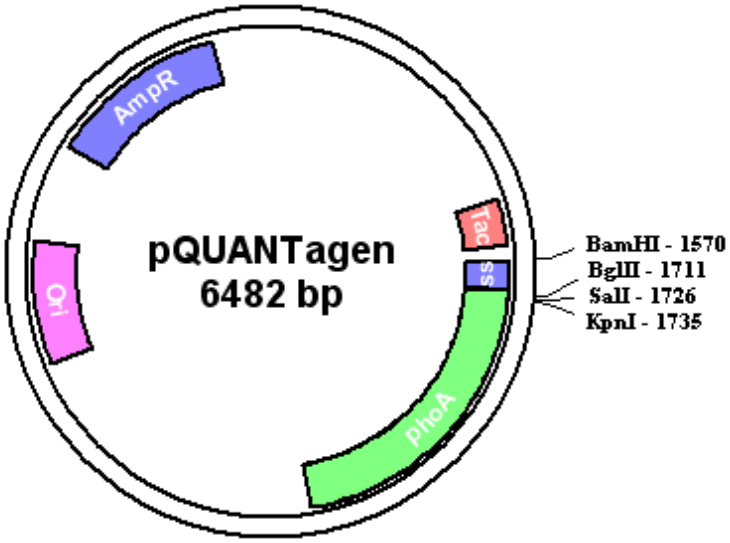


Figure 3: Map of pQUANTagen.The PhoA signal sequence (s*phoA*) corresponding to BamHI (nucleotide position 1570)-BglII (nucleotide position 1711) was removed from the original vector before cloning to create pQUANTagen (- s*phoA)*. In some cases the plasmid was cut with BsrGI (nucleotide position 1585) when ML genes that showed BamHI sites within the gene. The plasmid was cut with KpnI (nucleotide position 1735) or SalI (nucleotide position 1726) when ML genes showed BglII sites in the restriction map of the gene (Table 4).The *tac* promoter is located upstream of the *phoA* gene. Primers for DNA sequencing were located at nucleotide position 1525 (forward primer: 5'-CCAAGCTTACTCCCCATCCCC-3') and nucleotide position 1824 (reverse primer: 5'-CAGTCTGATCACCCGTTAAAC-3').

3.2.8.Cloning ML Genes into pJEM11

Fifteen *M. leprae* gene sequences were screened for putative promoter areas using the Gene and Functional Signal Finding Program (www.softberry.com) and 10 genes were selected for cloning into pJEM11 based on the presence of potential upstream elements for promoter activity. pJEM11 is a *phoA* reporter shuttle plasmid which replicates in *E. coli* and *M. smegmatis* and expresses kanamycin resistance in both genetic backgrounds (Figure 4). Primers containing BamHI restriction sites were designed for cloning into pJEM11 (3rd PCR, Table 4). PCR conditions were the same as used for the 1st PCR (Table 4), but the annealing temperature was 60º C for all genes amplified. PCR products were purified and digested with BamHI. pJEM11 was digested with BamHI and dephosphorylated with calf-intestine alkaline phosphatase (CIP). PCR products and pJEM11 were purified from agarose gels as described above. Recombinant plasmids were propagated in *E.coli* XL-1 blue. Plasmid DNA from colonies recovered on LB plus kanamycin plates was isolated using a QIAprep spin miniprep kit (Qiagen). Purified plasmid DNA was introduced into electrocompetent *M. smegmatis* mc^2 155 cells by electroporation using a Bio-Rad Gene Pulser (Hercules, CA) following a previously described protocol by Cirillo *et al*, 1993. Briefly, a culture (400ml) of *M. smegmatis* mc^2155 was grown at 37ºC for 2 days with constant shaking until reaching mid log growth (approximately 48 hrs.). The culture was held on ice for 1 hour and then centrifuged for 10 minutes at 5,000 x g at 4 ºC. The pellet was washed with cold 10% glycerol three times and the final bacterial pellet was resuspended in 1 ml of 10% glycerol.

One-hundred ul of this cell suspension and 50 ng of DNA were placed in 1.5 ml polypropylene tubes for each electroporation. The Gene Pulser apparatus was set at 2.5 kV, 25uF and 900 Ω. After delivering the pulse, the cell/DNA mixture was transferred to a culture tube containing 1 ml of 7H9 plus ADC and incubated at 37ºC with constant shaking

for 2 hours. Recombinant *M. smegmatis* colonies were screened for *phoA* expression on Luria agar containing kanamycin and BCIP. Blue colonies were recovered on days 7-14 and restreaked to obtain single-colony isolates and confirm phoA reporter activity. Transformants were screened for carriage of pJEM11 recombinant derivatives by PCR as described above with pQUANTagen recombinants. Primers in the multiple cloning site and *E. coli phoA* gene sequence were designed with Omiga™ 2.0. DNA sequence of ML genes were obtained by automated DNA sequencing on an Applied Biosystems sequencer (GeneLab, SVM, LSU, Baton Rouge, LA)

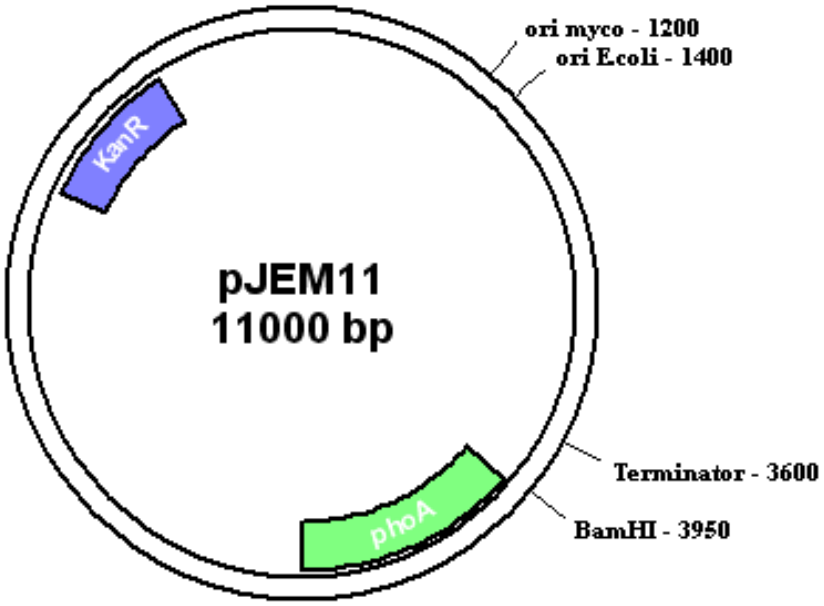


Figure 4: Map of pJEM11. pJEM11, a *phoA* reporter shuttle plasmid, has origins of replication (ori) for *E. coli* and mycobacteria and a multiple cloning site (MCS) upstream the phoA gene, represented by BamHI site. The selectable marker is the kanamycin resistance gene (Km). The truncated *phoA* gene is devoid of the promoter, start codon and signal sequence; the expression and exportation of PhoA depends on translational fusion with amino termini of other proteins encoding these sequences. The transcriptional terminator avoids transcription by read-through from plasmid sequences. Primers for DNA sequencing were located at the multiple cloning site (forward primer: 5'-CTAGTACTGGGCCCGCGG AT-3') and *E. coli phoA* gene (reverse primer: 5'-CCCCATCCCATCGCCAAT-3'

Table 4: Primers used for PCR cloning ML genes into plasmid vectors

Gene name	1st PCR from ML lysates			12nd PCR		23rd PCR
	AT	Ampl	Primers 5'→ 3'	Ampl	R.E.	Ampl*
ML0091	61	884	GGTCGCAATTACCTCACG GTCAGGTGACAGGTTCTTCG	692	BamHI-BglII	603
ML0097	60	1301	CAGGAATTTTAGACAGGCATCG AGCCCCTAAATACTGCTGAAGG	973	BamHI-BglII	813
ML0098	60	965	CTTGCATCGAATCATCG CCTCCCTAGCGAATGG	919	BamHI-BglII	640
ML0620	60	516	GATACGAAAGACCAGGAACAAGG CGACACGGAAACGTCAGC	418	BamHI-BglII	514
ML0715	62	1054	TTCAACCCTGACCGCACC GAAAAACTGCGCGATAACTTCC	901	BamHI-BglII	670
ML0885	58	1392	CTCTGTTGCAGGATGAACG ACCCGACGAATTCATCG	1073	BamHI-KpnI	865
ML1811	61	759	TAAACGAGGCTGAATATCAACG CGAATCAAGACTCGATGATCC	735	BsrGI-BglII	---
ML1812	65	1572	TGCAGTTTGTGACTTGCGTTTCC TATTCGATGTAGCGGACCACATACG	1349	BsrGI-SalI	950
ML1923	64	1219	TCGCTGCTGGAATTCGAGG CCCGATATACGGCCATTTGC	1200	BsrGI-KpnI	---
ML2028	61	1383	CGTAAGACAACCGCTGAGG GGGGTACAGCCATCAAGG	983	BamHI-BglII	750
ML2274	60	340	AAGTATATCATCCGGCTTATGAAGG GTTGCAGGATGCCTATTTGG	340	BamHI-BglII	---
ML2331	60	747	AACACACGAGTTAGCCTCCGTATCG GGCAAGGAAGCCCTCGACG	688	BamHI-BglII	---

(Table continued)

ML2380	60	512	GTTCCTGCTGGCTCGTAAGC TGGTACTGGTTGCAAACCTCG	455	BamHI-BglII	512
ML2569	60	1502	ACGAGTCCGCCAGCGTTTTTACC CCTTCTCTAGCCGGTCCAAGACTGC	969	BamHI-BglII	850
ML2591	65	1555	GCTGGAACTCCCCAATCG TACTCAATTCCGACTACCTCCTGC	1417	BsrGI-BglII	---

AT: Annealing temperature

* Length of amplicon of first, second and third PCR in base pairs (bp)

R.E: Restriction enzymes used for cloning into pQUANTagen.

---- Gene was not cloned in pJEM11 because there was no putative promoter region identified.

1 A second PCR was performed on the purified amplicons from the first PCR using primers containing appropriate restriction sites for cloning into pQUANTagen.

2 A third PCR was performed on the purified amplicons from the first PCR using primers containing BamHI sites at the 5' and 3' ends.

3.2.9.Detection of mRNA for ML Genes Cloned into *M. smegmatis*

Primers specific for each ML gene that was successfully cloned in pJEM11(ML0091, ML0097, ML0098, ML0715, ML2028, ML2380 and ML2569) were designed to amplify a fragment from each *M. smegmatis* transformant cDNA. mRNA expression analysis was performed using cDNA from *M. smegmatis* transformants and the same conditions for first PCR assay as described above.

Table 5: Primers and amplicon sizes for 7 ML genes cloned in *M. smegmatis* for mRNA transcript analysis.

Gene name	Amplicon (bp)	Primers* 5'→ 3'
ML0091 (*erp*)	300	ACCCGCGGCCAAACACTATG CAGGTGCCACCCAGGTTCAG
ML0097 (*fbpA*)	400	CCGGTGGAGTACCTTCAGGT GTGGTAGATCGCCAGCGTC
ML0098 (*fbpC*)	350	GCAATGGGCCGGGATATTCC CGGTCGGGGTGGAAAGCC
ML0715 (*lpqC*)	700	CGCCAGGCTCTCCTGTGG GCCCGCACCTTCGGCATAA
ML2028 (*fbpB*)	300	GACCGCAAGCGCGTTCTC CTTGTAGGTCGTGCAACCTGCC
ML2380	250	GGATTTTCCGATCCCCCGC CCAATGCGTGGCCATCTGCT
ML2569	430	GCCACGATGTTGCACAGGCT TTGTGACCACGGTCCAGCG

* Specific primers designed to amplify a portion of each gene. Annealing temperatures were 60ºC except for ML2028 which was 65ºC.

3.2.10.ML Protein Expression in *M. smegmatis* and *E. coli*

Late log cultures of recombinant *E. coli* and *M. smegmatis* were harvested by centrifugation at 6000 x g (4ºC) for 15 minutes. Culture supernatants were precipitated with iced-cold 100% ethanol (Sigma), the precipitate collected by centrifugation at 11,000 x g (4º C) for 30 minutes and the pellet was dried and resuspended in 100ul PBS. Bacterial pellets were washed 3 times with phosphate buffered saline, pH 7.4 (PBS) and resuspended in 0.5 ml of PBS. *M. smegmatis* samples were sonicated (3 cycles of 10 minutes at 50% power) using a microtip sonicator with an Ultrasonic Homogenizer 4710 (Cole-Parmer Instrument

Co., Chicago, IL). The bacterial sonicates were span at 11,000 x g (4º C) for 10 minutes and the supernatants were transferred to clean tubes. Total proteins of culture filtrates and bacterial sonicates were measured by BCA Protein Assay (Pierce, Rockford, IL) using BSA as a standard.

Concentrated culture supernatants and bacterial sonicates (200 μg) were analyzed by immunobloting after separation of proteins on 10% polyacrylamide gels containing sodium dodecylsulfate (SDS-PAGE). Sonicates and culture filtrates mixed (1:1) with sample buffer (62.5mM Tris-HCl, 10% glycerol, 2% SDS, 0.05% β-mercaptoethanol, 0.05% bromophenol blue) and loaded onto gels. SDS-PAGE gels were run for 2 hours at 25 mA and 2 hour at 50 mA in a Protean II xi cell (Bio-Rad, Hercules, CA) and blotted on Immobilon-P transfer membrane (Millipore, Bedford, MA) using a trans-blot cell (Bio-Rad) overnight at 12 V in Tris-glycine buffer containing 30% methanol (Sigma). The membrane was blocked with 3% BSA (Sigma) in PBS at room temperature for 45 minutes followed by incubation at room temperature for 2 hours with mouse anti-*E. coli* alkaline phosphatase monoclonal antibody (Chemicon International, Inc., Temecula, CA), diluted 1:2500 in 1% BSA in PBS. The membrane was washed 3 times with PBS containing 0.05% Tween 20 (Sigma), incubated at room temperature for 1 hour with horseradish peroxidase (HRP) conjugated goat anti-mouse immunoglobulins (Dako Corp., Carpenteria, CA) diluted 1:2000 in 1% BSA in PBS. The membrane was washed again as stated above and placed in HRP color development reagent (30 mg of 4-chloro-1-naphthol) (Bio-Rad) in 10 ml of methanol, 300 μl of 3% hydrogen peroxide (Medic, Jacksonville, FL) and 50 ml of Tris-buffered saline, pH 7.5 at room temperature for 10 minutes.

3.3.Results

3.3.1.Gene Expression of 15 Putatively Secreted Proteins in *M. leprae* during Infection in the Nude Mouse

Fifteen ML predicted secreted proteins that showed 50% amino acid homology with *M. tuberculosis* (Table 2) were selected for studying gene transcriptionion in *M. leprae* during intracellular growth in the nude mouse by RT-PCR. Amplicons produced by RT-PCR showed predicted DNA fragment sizes for each of the 15 genes tested. An example of one RT-PCR gene transcript analysis is shown in Figure 5.Both the cDNA (lane 2) and genomic DNA (lane 5) produced amplicons of the appropriate size (450 bp) for the ML0097 (*fbpA*) gene. DNA sequencing of the amplicon confirmed the existence of *fbpA* mRNA in *M. leprae* during infection of nude mice. Mouse cDNA and RT(-) ML samples were run as negative controls for each gene (Figure 4). These data confirm that all 15 *M. leprae* genes identified as putatively secreted proteins are transcribed during infection and, therefore, may be critical for maintenance of intracellular survival during infection.

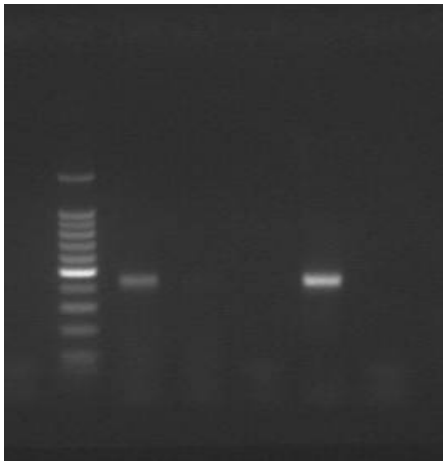


Figure 5: Agarose gel electrophoretic analysis of PCR products of ML0097 (*fbpA)* from ML genomic and cDNA derived from ML-infected nude mice. Lane 1 1kb ladder, lane 2 *M. leprae*, strain T53 cDNA, lane 3 normal mouse cDNA, lane 4 RT (-) and lane 5 *M. leprae* genomic DNA.

3.3.2.Cloning ML Genes into pQUANTagen

15 ML genes were amplified by PCR (Table 4) for cloning into pQuant (- sphoA), a vector that contains the alkaline phosphatase gene under the transcriptional control of the synthetic *tac* promoter. The signal peptide for *phoA* that directs the fusion protein through the membrane to the periplasm of *E. coli* was removed to study secretion of the 15 ML putatively secreted proteins (Figure 3) in *E. coli*. Ligation mixtures for each gene were transformed into *E. coli* XL-1 blue supercompetent cells and produced between 100-200 Amp^r transformant colonies. All ML gene transformants produced blue and white colonies on agar plates supplemented with the substrate for the alkaline phosphatase with the percentage of blue colonies per transformant ranging from 10% to 70%.

3.3.3.PCR Amplification and Sequence Analysis of *M. leprae* Inserts in pQUANTagen

For each ML gene *E. coli* transformation ten colonies were pooled and used to inoculate LB plus Amp media. Plasmid DNA from these cultures were purified, amplified by PCR with primers designed for sequencing in pQUANTagen and were analyzed on agarose gels. PCR products that matched gene insert size (Table 6, Figure 6) were sequenced starting 30-40 bases upstream of the ML gene start codon to authenticate gene orientation and reading frame. Blue colonies recovered from *E. coli* transformants ML0091, ML0097, ML0620, ML1811 and ML1812, and white colonies from one *E. coli* transformant ML2380 confirmed proper gene insert and alignment for expression (Table 6). While proper gene insert and alignment were found for ML0091, ML0097, ML1811 and ML1812 sequences were incomplete finishing before the stop codon for each gene. DNA sequencing from only two gene inserts ML0620 and ML2380 verified the presence of the complete gene inserts (Table 6). DNA from *E. coli* transformants ML0098, ML0715, ML0885, ML1923, ML2028, ML2274, ML2331, ML2569 and ML2591 each produced a PCR fragment of about 300bp

indicating the absence of a ML gene insert (Table 6 and Figure 6). An example of two *E. coli* transformants containing ML genes ML0091 and ML0097 is shown in Figure 6. *E. coli* transformants produced amplicons of the appropriate size for the ML0091 (992 bp) and ML0097 (1273 bp) genes and where no insert was detected showed a band at 300 bp.

Table 6: ML PhoA gene fusions in *E. coli*

Gene name	**Reporter in *E. coli***	**ML gene insert* (bp)**	**Sequence length** (bp)**
ML0091 (*erp*)	blue	992	620
ML0097 (*fbpA*)	blue	1273	510
ML0098 (*fbpC*)	[1]---	300	---
ML0620 (*mtb12*)	blue	718	Complete gene
ML0715 (*lpqC*)	---	300	---
ML0885	---	300	---
ML1811	blue	1035	620
ML1812	blue	1649	470
ML1923 (*lpqF*)	---	300	---
ML2028 (*fbpB*)	---	300	---
ML2274	---	300	---
ML2331	---	300	---
ML2380	[2]white	755	Complete gene
ML2569	---	300	---
ML2591(*mce1C*)	---	300	---

1 Low percentage of blue colonies all testing negative for gene inserts, all white colonies tested showed no gene insert.
2 Low percentage of blue colonies all tested negative for gene insert, some white colonies tested positive for gene insert.
* ML gene inserts amplified by PCR with pQUANTagen primers have amplicon sizes of 2nd PCR (Table 4) plus multiple cloning site from pQUANTagen (300bp). ML genes with no insert identified by PCR showed amplicon size of 300bp, the same as the multiple cloning site of pQUANTagen.
** DNA sequences of cloned genes from PCR products yielding predicted amplicon sizes were determined beginning 30-40 bases upstream of ML gene start codon and finishing after the ML gene stop codon (ML0620 and ML2380) or before the ML gene stop codon (ML0091, ML0097, ML1811 and ML1812).

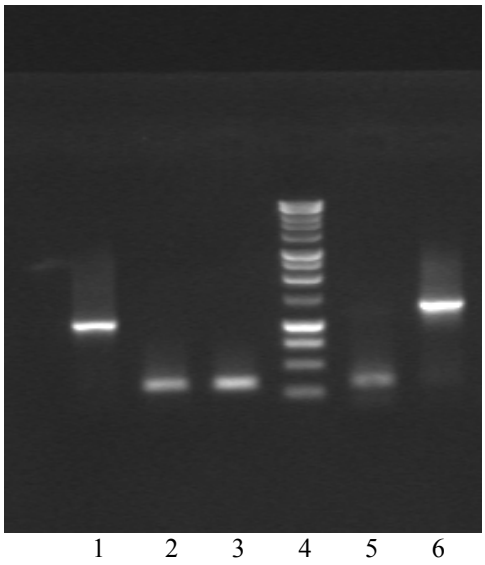


Figure 6: Agarose gel electrophoretic analysis of PCR products from *E. coli* transformants containing ML genes. Lane 1, ML0091 (992 bp), lanes 2 and 3, transformants without insert, lanes 4, 1kb ladder , lane 5 pQUANTagen plasmid, lane 6 ML0097 (1273 bp).

3.3.4.Cloning ML Putatively Secreted Protein Genes into pJEM11

Because pJEM11 contains a truncated phoA reporter gene without a promoter, start codon and signal sequence (Figure 4), 10 ML genes (ML0091, ML0097, ML0098, ML0620, ML0715, ML0885, ML1812, ML2028, ML2380 and ML2569) were selected for analysis in pJEM11 based on their putative possession of the required genetic elements. These 10 ML genes (Table 4) were amplified by PCR and cloned into pJEM11. Each ligation mixture was first transformed into *E. coli* XL-1 Blue supercompetent cells from which 50-100 Kan^r transformant colonies were recovered for each ML gene. For each cloned gene plasmid DNA was purified from a culture grown from a pool of colonies and introduced into mc^2155 by electroporation. Twenty to forty Kan^r *M. smegmatis* transformants colonies were recovered from Luria agar plates containing BCIP. Blue colonies, representing *M. smegmatis* transformants capable of exporting alkaline phosphatase, were identified from transformants ML0715 and ML2569 (Table 7 and Figure 7). White colonies, representing *M. smegmatis*

transformants without a gene insert or with a gene insert lacking all or some of the required genetic elements for transcription, translation or secretion, were identified from transformants ML0091, ML0097, ML0098, ML0620, ML0885, ML1812, ML2028 and ML2380 (Table 7).

3.3.5.PCR Amplification and Sequence Analysis of *M. leprae* Inserts in pJEM11

DNA from each of ten colonies from *M. smegmatis* transformants for each ML gene was amplified by PCR with primers for sequencing in pJEM11 (Figure 4). PCR products from all clones except (ML0620, ML0885 and ML1812) matched predicted gene insert sizes (Table 7) and were sequenced starting 30-40 bases upstream of ML gene start codon and finishing from 400 to 800 bases before the ML gene stop codon. ML genes with correct orientation and reading frame were confirmed for the first 380 bp of *M. smegmatis* transformant ML0715 and the first 430 bp of *M. smegmatis* transformant ML2569. Also, ML gene inserts with correct orientation and reading frame were confirmed in 5 *M. smegmatis* transformant (white colonies) by sequencing the first 395bp (ML 0091), 430 bp (ML0097), 180 bp (ML0098), 210 bp (ML2028) and 200 bp (ML2380) (Table 7). Plasmid DNA from three *M. smegmatis* transformants yielding white colonies (ML0620, ML0885 and ML1812) produced PCR fragments of 250bp indicating the absence of a gene insert (Table 7). The analysis of promoter areas showed that 4 of 10 ML genes (Table 7) did not have identifiable RBS, -10 or -35 regions in the 200 bp segment analyzed suggesting that these 4 genes may be located within discrete operons.

Table 7: ML PhoA gene fusions in *M. smegmatis*.

Gene name	ML gene insert** (bp)	Promoter*	Reporter in *M. smegmatis*	Sequence length^(bp)
ML0091 (*erp*)	853	RBS	white	395
ML0097 (*fbpA*)	1063	RBS	white	430

(Table continued)

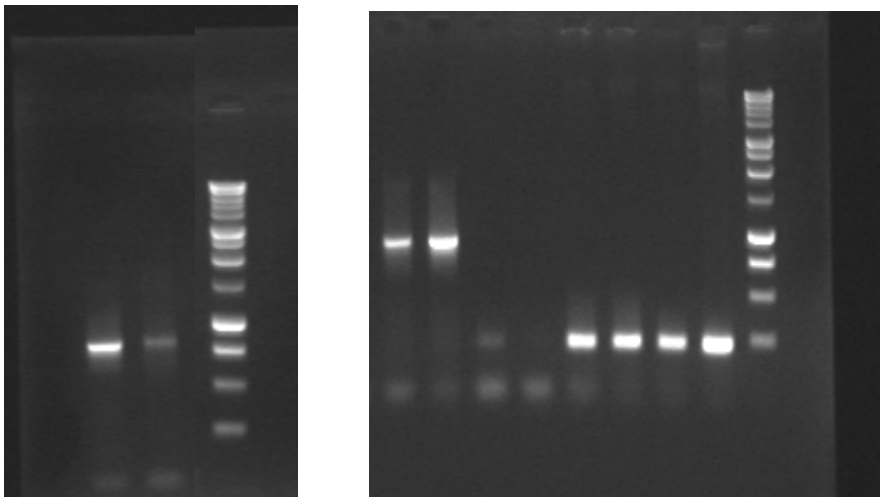
ML0098 (*fbpC*)	890	RBS, -10	white	180
ML0620 (*mtb12*)	250	NI[2]	white	---[1]
ML0715 (*lpqC*)	920	-10, -35	blue	380
ML0885	250	RBS, -10, -35	white	---
ML1812	250	NI	white	---
ML2028 (*fbpB*)	1000	-10, -35	white	210
ML2380	762	NI	white	200
ML2569	1100	NI	blue	430

1 No gene insert identified by PCR.
* RBS were identified by The Sanger Centre annotation of ML genome and –10 and –35 regions were identified using the Gene and Functional Signal Finding Program, SoftBerry (www.softberry.com).
** ML gene inserts amplified by PCR with pJEM11 primers have the size of amplicon sizes of 3rd PCR (Table 4) plus multiple cloning site from pJEM11 (250bp). ML genes with no insert identified by PCR showed amplicon sizes of 250bp, the same as the multiple cloning site of pJEM11.
^DNA sequences of cloned genes from PCR products yielding predicted amplicon sizes were determined beginning 30-40 bases upstream of ML gene start codon and finishing before the ML gene stop codon.
2 NI, nothing identified resembling a promoter or ribosome binding site

Agarose gel electrophoresis analysis of PCR fragment sizes for two pJEM11 cloned genes is shown in Figure 8. ML DNA from *M. smegmatis* transformants produced amplicons of the appropriate size for the ML0091 (853 bp) and ML0715 (920 bp) genes. Clones containing plasmid with no insert gave PCR fragments equivalent to 250 bp (Table 7).

pJEM11 ML2569

Figure 7: Colonies of *M.smegmatis* following pJEM11 transformation by electroporation. Transformants were plated in Luria agar with Kan and BCIP. Inset of plate ML2569 shows close-up of 2 blue colonies and a white colony.



1 2 3 4 5 6 7 8 9 10 11 12

Figure 8: Agarose gel electrophoretic analysis of PCR products from *M. smegmatis* transformed with pJEM11-ML gene fusions.
Lane 1 and 2, transformants ML0091, lanes 3 and 12 1kb ladder , lanes 4 and 5 transformants ML0715, lanes 6 and 7 transformants with no insert, lanes 8, 9, 10 and 11 pJEM11 plasmid.

3.3.6.Transcription of ML Genes Cloned into *M. smegmatis*

Because 5 *M. smegmatis* transformants contained a portion of a cloned ML gene but produced white colonies on culture, RT-PCR was employed to determine whether gene-specific mRNA was present in growing *M. smegmatis* cells. cDNA from total RNA was prepared using specific primers for each ML gene (ML0091, ML0097, ML0098, ML0715, ML2028, ML2380 and ML2569) cloned into *M. smegmatis*. mRNA from ML0715 was identified by PCR using a forward primer made from ML0715 gene sequence and a reverse primer made from *E. coli phoA* gene sequence (Table 5). Authentic ML gene sequences were confirmed by sequencing the PCR products.

An example of detecting mRNA from an *M. smegmatis* transformant containing ML0097 is shown in Figure 9. DNA and cDNA from *M. smegmatis* transformant ML0097 produced a 450 bp amplicon confirming transcription in *M. smegmatis* (Table 5). Negative controls RT (-) and pJEM11 plasmid DNA produced no discernible PCR product on gels.

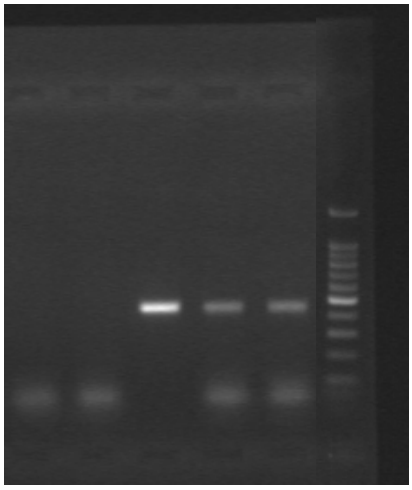


Figure 9: Agarose gel electrophoretic analysis of RT-PCR products derived from cDNA from *M. smegmatis* transformants. Lane 1, RT (-), lane 2 pJEM11, lane 3 ML0097 DNA amplified with specific primers. Lanes 4 and 5 *M. smegmatis* transformant ML0097 cDNA amplified with ML 0097 gene specific primers, respectively. Lane 6 contained 100 bp ladder.

3.3.7.Immunoblot Detection of PhoA Fusion Proteins in *M. smegmatis* and *E. coli*

Lysates prepared from recombinant *E. coli* clones (ML0091, ML0097, ML0620, ML1811 and ML1812) producing blue colonies on BCIP containing media were tested by immunoblotting for expression of ML-PhoA fusion proteins. Bacterial sonicates prepared from recombinant *M. smegmatis* clones (ML0715 and ML2569) producing blue colonies and recombinant *M. smegmatis* clones (ML0091, ML0097, ML0098, ML2028 and ML2380) producing white colonies on BCIP-containing media were also tested by immunoblotting for expression of ML-PhoA fusion proteins. A recombinant *E. coli* strain producing high levels of native PhoA was used as a positive control.

Immunoblots of *E. coli* transformants ML0620, ML0097 and ML0091 identified a band at approximately 50 kDa corresponding in size with native *E. coli* alkaline phosphatase (Fig 10). ML0620-*phoA* and ML0097-*phoA* fusions showed a second band at approximately 63 kDa and 84kDa, respectively (Fig 10). These protein bands corresponded to their respective predicted molecular weights for each ML-*phoA* fusion protein. ML0091-*phoA*

fusion showed two other bands at approximately 55 kDa and 90 kDa, neither of which corresponded to the predicted molecular weight for the authentic ML0091-*phoA* fusion of 73 kDa. Extracts from *E. coli* transformant ML0098, in which a cloned insert could not be identified by PCR (Table 6), showed no bands detectable by immunoblotting (Fig 10). Culture filtrates containing 200 ug of protein from cultures of all recombinant clones were tested by immunoblotting and showed no detectable bands under identical conditions used above (data not shown).

All attempts to identify fusion proteins in extracts from recombinant *M. smegmatis* clones were unsuccessful (data not shown). Both bacterial sonicates and culture filtrates showed negative profiles in immunoblots under identical conditions used above.

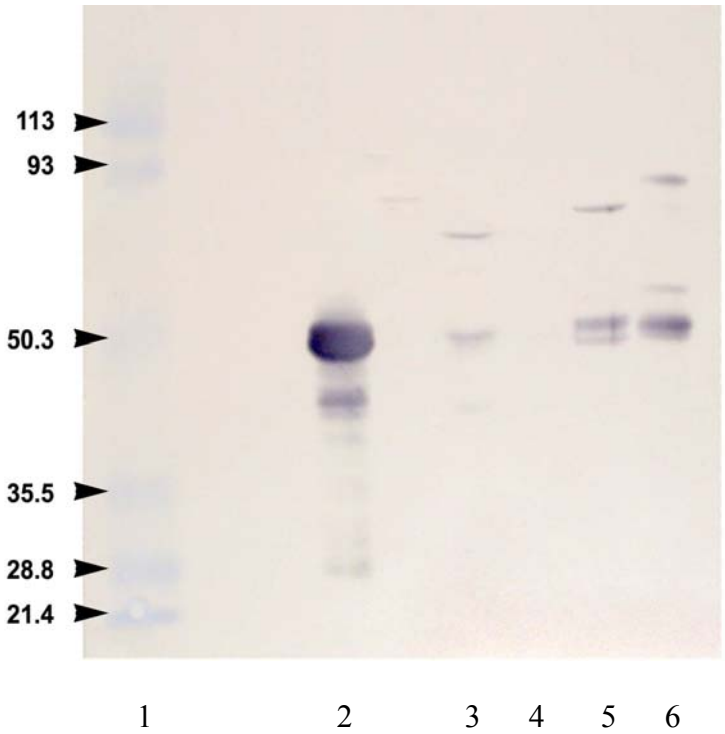


Figure 10: Western blot of *M. smegmatis* and *E. coli* ML transformants.
Lysates from *M. smegmatis* and *E. coli* transformants were electrophoretically separated on 10% SDS-PAGE, transferred to Immobilon-P membrane and reacted with 1:2,500 dilution of monoclonal anti-*phoA* Lane 1 MW marker, lane 2 is native *E. coli phoA*, lanes 3, 4, 5 and 6 are *E. coli* transformants ML0620, ML0098, ML0097 and ML0091 respectively.

3.4.Discussion

Characterization of exported proteins in *M. leprae* is important because it will provide a better understanding of the signals that direct proteins to the cell membrane, to the cell wall and to the bacteria's surrounding immediate exterior as soluble secreted proteins. Some of the exported proteins are involved in the building of the mycobacterial cell wall, such as the antigen 85 complex (fbpA, B and C), but the majority of exported proteins have not been functionally characterized. A comprehensive study of exported proteins in mycobacteria is needed to identify those proteins involved in cell envelope biogenesis, immunogenicity and virulence. In efforts to identify ML secreted proteins this work has explored the utility of bioinformatic tools for predicting secretion in *M. leprae* and has explored the usefulness of secretion vectors for the identification and validation of secretion of proteins in *E. coli* and *M. smegmatis*.

Fifteen ML proteins that showed greater than 50% amino acid sequence homology with a known or predicted *M. tuberculosis* secreted protein were selected for studying gene expression in *M. leprae* during intracellular growth in the nude mouse. All 15 putatively secreted protein genes were transcribed in *M. leprae* during infection suggesting that all 15 *M. leprae* genes are likely critical for maintenance of intracellular survival during infection. Seven of the *M. leprae* proteins identified in my study have been shown previously to be associated with virulence in *M. tuberculosis* (fbpA, B and C, Rv2190c similar to ML0885, Rv1477 similar toML1812, Rv1478 similar to ML1811 and mce1C), correlates with the observation of Finlay and Falkow, 1997, and Miller and Cossart, 1999, who showed that the majority of virulence factors in bacteria are extracytoplasmic proteins.

Williams *et al*, 2003, studied gene expression of over 200 ML genes during infection in nude mice. The study included genes involved in Sec-dependent secretion, DNA

replication, transcription, virulence factors, iron acquisition and numerous metabolic pathways. While only a small percentage of the 1600 potential genes of *M. leprae* were studied greater than 90 percent of the genes analyzed were transcribed. These data supported earlier speculation that since *M. leprae* has undergone a rather severe loss of genes due to a reductive evolutionary process, the remaining genes of *M. leprae* may define a minimal set of genes necessary for growth and survival of an intracellular mycobacterial pathogen. Accordingly, the relatively low number of potentially secreted proteins identified in my dissertation studies is in keeping with this interpretation. It is interesting to note that only 271 genes were necessary for survival of the free-living nonpathogen, *B. subtilis* (Ehrlich *et al*, 2003). Comparisons between minimal gene sets of free-living and pathogenic bacteria may help define genetic elements important in intracellular survival and help elucidate pathogenic mechanisms of mycobacteria and other intracellular pathogens.

PhoA fusion studies using a modified pQUANTagen without the phoA signal sequence, showed that 5 ML proteins (ML0091, *fbpA*, ML0620, ML1811 and ML1812) produced blue colonies and, therefore, were secreted in *E. coli*. Nine ML genes (*fbpC*, *lpqC*, ML0885, *lpqF*, *fbpB*, ML2274, ML2331, ML2569 and *mce1C*) were unable to be cloned in *E. coli*. A simple explanation for these results is that the 9 unclonable mycobacterial proteins were toxic for *E. coli*. Jobling, *et al*, 1997, reported that the production of wild type cholera toxin as a periplasmic protein was toxic for *E. coli*, but by replacing the native signal sequence with an *E. coli* signal sequence they were able to produced cholera toxin protein in high yield in *E. coli*. While the ML proteins were clearly not toxins like cholera toxin, many proteins are difficult to clone in *E. coli* for myriad reasons involving such as blockage of metabolic pathways to binding and disrupting membranes and others physiologically important structures. For example, *E. coli* may be unable to process mycobacterial

lipoproteins or other Sec-dependent motifs found in some ML signal peptides, leading to build up of these proteins eventual toxicity to *E. coli*.

PhoA fusion studies using pJEM11 showed that only 2 proteins (ML0715 and ML2569) produced blue colonies and, therefore, were secreted in *M. smegmatis*. Both proteins are thought to be lipoproteins based on the presence of a conserved motif. Lim, *et al*, 1995, used pJEM11 to construct an *M. tuberculosis* DNA library of fusions to the PhoA gene and identified a sequence corresponding to the exported 19kDa lipoprotein. Lipoproteins have been found in both Gram-negative and Gram-positive bacteria; however, differences in the conserved motif of lipoproteins or in the specificity of signal peptidase II to process lipoproteins may exist in *E. coli* making secretion impossible.

The low number of detectable ML proteins secreted by *M. smegmatis* could be related to the presence of mutations or deletions in the ML genes that produced an out of frame ML gene-phoA fusion. The sequence data for the 7 ML genes cloned in *M. smegmatis* is not complete, therefore, it is possible that all or some white colonies are *M. smegmatis* transformants carrying mutations of the ML genes resulting in defective phoA fusions. The expression of PhoA from pJEM11 in *M. smegmatis* depends on translational fusion with amino termini of the cloned ML proteins and the presence of a functional promoter. Of 10 ML genes that were selected for cloning into pJEM11, only 4 showed a putative promoter area (-35 and -10 sites) located 100 to 200 bases upstream from the starting codon. However, strict definition of mycobacterial promoters is still undefined and the potential for some of these genes to be within operons can not be ruled out.

Three ML proteins (ML0620, ML0885 and ML1812) were unable to be cloned in *M. smegmatis* and may be due to procedural issues. It is possible that some ML genes require specific environmental conditions (e.g. intracellular) for expression or for their product to be

exported. In addition, at the level of promoter recognition and transcription initiation controlled by the sigma factors of RNA polymerase, there may be significant differences between *M. leprae* and *M. smegmatis* blocking gene expression. Comparative genomic studies of *M. leprae* and *M. tuberculosis* have shown a significantly lower number of sigma factors in *M. leprae* (Cole *et al*, 2001).

The *E. coli* and *M. smegmatis* secretion vectors were available for this study and were used before to validate secretion in *M. tuberculosis*. *E. coli* vector showed more ML proteins secreted (5 out of 6) under a strong *E. coli* promoter. In addition, the alkaline phosphatase activity was detected in 2 days in Luria agar supplemented with the appropriate substrate. A disadvantage of using *E. coli* is the incompatibility to process post translational modifications (e.g. lipoproteins). *M. smegmatis* vector showed less ML proteins secreted (2 out of 7), and maybe due to weak ML promoters. The alkaline phosphatase activity was detected after 7-14 days in Luria agar supplemented with the appropriate substrate.

The Sec pathway for secretion is present in both *M.leprae* and *E. coli* but there appear to be some differences. For example, the absence of SecB and the presence of two SecA genes in *Mycobacterium spp* (Braunstein *et al*, 2001) could make the process of translocation slightly different. SecB acts as a chaperone and binds SecA, but it has been only found in Gram-negative bacteria. By comparison mycobacteria and some other Gram-positive bacteria code two SecA proteins. It is thought that the two proteins might be involved in exporting different subsets of proteins. In addition, the signal sequences for secretion described for Gram-negative and Gram-positive bacteria are not identical. Therefore, some of the mycobacterial signal sequences might not be recognized by the *E. coli* Sec-dependent pathway. Taken together these differences may explain why 9 of 15 ML proteins were not expressed in *E. coli*.

In order to study gene expression in *M. smegmatis* as a function of transcription, RT-PCR was performed. Transcripts were identified for 7 ML genes cloned in *M. smegmatis* (ML0091, ML0097, ML0098, ML0715, ML2028, ML2380 and ML2569). These data confirm that the required genetic elements for transcription are present in pJEM11.

White colonies recovered from *M. smegmatis* transformants (ML0091, ML0097, ML0098, ML2028 and ML2380 might represent mutants resulting in a defective *phoA* fusion or transformants lacking the required genetic elements for translation or secretion. In order to study protein expression of the 5 recombinant *M. smegmatis* white colonies immunoblot detection of phoA fusion proteins was performed. Under these conditions, ML recombinant protein was not observed in the blots. These results suggest that the *phoA* gene is not in frame with some or all the *M. smegmatis* white transformants due to mutations in ML genes. The *M. smegmatis* blue transformants showed alkaline phosphatase activity in Luria agar plates and therefore have a functional *phoA* gene, however, the transcription of *phoA* may be under a weak promoter. ML0091, ML0097 and ML0620 recombinant proteins from *E. coli* transformants were identified by immunoblot detection with monoclonal antibodies anti-PhoA, ML0620 complete sequence was confirmed in *E. coli*.

The experiments in this dissertation showed indirect evidence of Sec-dependent mechanisms for secretion in *M. leprae*. The possibility that some or all of these ML secreted proteins were translocated in E. coli or M. smegmatis by a different secretion pathway (Ex. Sec-independent pathway), cannot be ruled out.

In summary, the analysis of PhoA fusions validated 5 of 6 ML proteins (83%) in *E. coli* and 2 of 7 ML proteins (29%) in *M. smegmatis*. By comparison, Gomez *et al*, 2000, identified 52 *M. tuberculosis* predicted secreted proteins by computer-based analysis and 9 of 10 (90%) were confirmed using PhoA gene fusions in *E. coli*. These results suggest that

predictions for secretion in *M. leprae* using Signal P and TMHMM are useful in selecting Sec-dependent secreted proteins.

CONCLUSIONS

1) Bioinformatic Tools Predict Sec-dependent Protein Secretion in *M. leprae*

Bioinformatic tools trained to recognize secreted proteins were used to predict the existence of secreted proteins in *M. leprae*. Sequences for two hundred and four *M. tuberculosis* proteins either known to be or predicted to be secreted constituted the set of gene sequences from which *M. leprae* homologs were selected. Fifty-two MT proteins were not found in the *M. leprae* genome and 38 MT proteins were related to ML pseudogenes. Analysis of the remaining 114 homologs with SignalP and TMHMM predicted that 24 ML proteins display characteristics consistent with secretion via the sec-dependent protein secretion pathway.

M. leprae signal sequences were found to be very similar to other Gram-positive and *M. tuberculosis* signal sequences. Cleavage sites were conserved and the N-regions from ML signal sequences showed a high arginine content (19%) similar to that found in *M. tuberculosis* (22%) Sec-dependent secreted proteins.

Many fewer secreted proteins were predicted in *M. leprae* (24) compared to *M. tuberculosis* (52). This discrepancy is also observed in the number of functional genes in *M. tuberculosis* (3959) compared to *M. leprae* (1604) and may be the result of a severely diminished genome in *M. leprae* due to evolutionary pressures. These data supported earlier speculation that since *M. leprae* has undergone a severe loss of genes due to a reductive evolutionary process, the remaining genes of *M. leprae* may define a minimal set of genes necessary for growth and survival of an intracellular mycobacterial pathogen.

2) All 15 ML Predicted Secreted Proteins Were Transcribed During Intracellular Growth in Nude Mice

Analysis of *M. leprae* growing in nude mice identified transcripts for all 15 putatively secreted protein genes of *M. leprae*. Seven of these *M. leprae* proteins have been associated

with potential virulence factors in *M. tuberculosis* and correlates with the observation that the majority of bacterial virulence factors are extracytoplasmic proteins.

3) Secretion of 83% of the ML Predicted Secreted Proteins Was Verified by PhoA Fusion Analyis in *E. coli* and Secretion of 29% of the ML Predicted Proteins Was Verified by PhoA Fusion Analysis in *M. smegmatis*

By comparison, Gomez *et al*, 2000, identified 52 *M. tuberculosis* predicted proteins using bioinformatic tools and 90% were verified by phoA fusion analysis in *E. coli.* These results suggest that predictions for secretion in *M. leprae* using bioinformatics tools are useful in selecting Sec-dependent secreted proteins. My data also suggest that validation of secretion in M. smegmatis can help extend coverage of potential secreted proteins to include lipoproteins of *M. leprae.*

Future studies will be directed to investigate the relevance of the two ML lipoproteins (ML0715 and ML2569) secreted in *M. smegmatis* and the 5 ML proteins (ML0091, ML0097, ML0620, ML1811, ML1812) secreted in *E. coli* as potential vaccines or diagnostic reagents for controlling leprosy. Because these ML proteins showed a relatively high amino acid identity with *M. tuberculosis*, they could be developed as vaccine candidates for both leprosy and tuberculosis. ML secreted proteins could be delivered as DNA vaccines using a strong eukaryotic promoter. In addition, these ML secreted proteins could be used to improve BCG vaccines by creating new over-expressing recombinant BCG strains.

These ML secreted proteins may not be good candidates for *M. leprae*-specific diagnostic reagents for leprosy because of the relatively high amino acid homology between the ML proteins and their *M. tuberculosis* homologs. However, it is possible that these secreted ML proteins and MT proteins have unique epitopes engendering immunological specificity for ML or MT.

Finally, some of the ML secreted proteins identified in this study have not been annotated and therefore, may have unique functional or structural capacities. Further studies into their function and immunogenicity could lead to a fuller understanding of the host immune response during infection and to the disease causing potential of *M. leprae*.

REFERENCES

Abou-Zeid C, Ratliff TL, Wiker HG, Harboe M, Bennedsen J and Rook GA, 1988. Characterization of fibronectin-binding antigens released by *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG. Infect Immun 56:3046-3051.

Abou-Zeid C, Smith I, Grange JM, Ratliff TL, Steele J and Rook GA, 1988. The secreted antigens of *Mycobacterium tuberculosis* and their relationship to those recognized by the available antibodies. J. Gen. Microbiol. 134:531-538.

Andersen P,1994. Effective vaccination of mice against Mycobacterium tuberculosis infection with a soluble mixture of secreted mycobacterial proteins. Infect. Immun. 62:2536-2544.

Andersen P, Askgaard D, Ljungqvist L, Bennedsen J and Heron I, 1991. Proteins released from *Mycobacterium tuberculosis* during growth. Infect Immun 59:1905-10.

Arruda S, Bomfim G, Knights R, Huima-Byron T, Riley LW, 1993. Cloning of an M. tuberculosis DNA fragment associated with entry and survival inside cells. Science 261:1454-57.

Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ and Besra GS, 1997. Role of the major antigen of *Mycobacterium tuberculosis* in the cell wall biogenesis. Science 276:1420-1422.

Berger BR and Christie PJ, 1994. Genetic complementation analysis of the *Agrobacterium tumefaciens* virB2 through virB11 are essential virulence genes. J Bacteriol 176:3646-3660.

Berthet FX, Langranderie M, Gounon P, 1998. Attenuation of virulence by disruption of the *Mycobacterium tuberculosis erp* gene. Science 282:759-62.

Bigi F, Alito A, Fisanotti JC, 1995. Characterization of a novel *Mycobacterium bovis* secreted antigen containing PGLTS repeats. Infect Immun 63:2581-2586.

Blanco J, Coque JJ and Martin JF, 1996. Characterization of the *secA* gene of *Streptomyces lividans* encoding a protein translocase which complements an *Escherichia coli* mutant defective in the ATP-ase activity of SecA. Gene 176:61-65.

Bloch H and Segal W, 1955. Viability and multiplication of vaccines in immunization against tuberculosis. Am Review of Tuberculosis and Pulmonary Diseases 71:228-248

Braibant M, Gilot P, Content J, 2000. FEMS Microbiology Reviews 24:449-467.

Braunstein M and Belisle JT, 2000. Molecular Genetics of Mycobacteria. ASM Press, Washington DC.

Braunstein M, Brown AM, Kurtz S and Jacobs WR, 2001. Two nonredundant SecA homologues function in Mycobacteria. J Bacteriol 183:6979-6990.

Brennan PJ and Nikaido H, 1995. The envelope of mycobacteria. Annu Rev Biochem 64:29-63.

Brundage L, Hendrick JP, Schiebel E, Driessen AJ and Wickner W, 1990. The purified E. coli integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. Cell 62:649-657.

Carroll JD, Wallace RC, Keane J, Remold HG, Arbeit RD, 2000. Identification of *M. avium* DNA sequences that encode exported proteins by using *phoA* gene fusions. Tub Lung Dis 80:117-130.

Cheng LW and Schneewind O, 2000. Type III machines of Gram-negative bacteria: delivering the goods. Trends Microbiol 8:214-220.

Chubb AJ, Woodman ZL, da Silva Tatley MP, Hoffman HJ, Scholle RR and Ehlers MR. 1998. Identification of *M. tuberculosis* signal sequences that direct the export of a leaderless *β*-lactamase gene product in *E. coli.*

Closs O, Harboe M, Axelsen NH, Bunch-Christensen K and Magnusson M, 1980. The antigens of *Mycobacterium bovis*, strain BCG, studied by crossed immunoelectrophoresis: a reference system. Scand J of Immunol. 12:249-263.

Cole ST *et al*, 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537-544.

Cole ST *et al*, 2001. Massive gene decay in the leprosy bacillus. Nature 409:1007-11.

Collins FM, Lamb JR and Young D, 1988. Biological activity of protein antigens isolated from *M. tuberculosis* culture filtrate. Infect Immun 56:1260-66.

Cornelis GR, 1998. The Yersinia deadly kiss. J Bacteriol 180:5495-5504.

Dilks K, Rose RW, Hartmann E and Pohlschroder M, 2003. Prokaryotic utilization of the twin-arginine translocation pathway: a genomic survey. J Bacteriol 185:1478-1483.

Draper P, 1998. The outer parts of the mycobacterial envelope as permeability barriers. Front Biosci 3:D1253-D1261.

Ehrlich SD, Kobayashi K, Albertini A, 2003. Essential *B. subtilis* genes. PNAS 100:4678-83.

Fine PEM, 1996. Randomised controlled trial of single BCG, repeated BCG or combined BCG and killed *M. leprae* vaccine for prevention of leprosy and tuberculosis in Malawi. Lancet 348:17-24.

Finlay BB and Falkow S, 1997. Common themes in microbial pathogenicity revisited. Microbiol. Mol.Biol.Rev. 61: 136-169.

Freer G, Florio W, Dalla Casa B, Castagna B, Maisetta G, Batoni G, Corsini V, Senesi S, Campa M, 1998. Characterization of antigens recognized by new monoclonal antibodies

raised against culture filtrate proteins of *M. bovis* bacillus Calmette-Guerin. FEMS Immunol Med Microbiol 20:129-138.

Gillis TP and Krahenbuhl JL, 1998. Global elimination of leprosy. Rev Med Microbiol 9:39-48.

Gomez M, Johnson S and Gennaro ML, 2000. Identification of secreted proteins of *M. tuberculosis* by a Bioinformatic approach. Infect Immun 68:2323-27.

Hansen GHA, 1874. Undersogelser angraaende spedalskhedens aasager. Norsk Magazin for Laegervidenskaben (Suppl) 4:1-88

Harboe M and Nagai S, 1984. MPB70, a unique antigen of *Mycobacterium bovis* BCG. Am Review of Resp Dis 129:444-452.

Harboe M and Wiker HG, 1999. Searching for secreted proteins of *Mycobacterium leprae*. Indian J Lepr 71:19-35.

Hasting RC, 1994. Leprosy. Longman Singapore Publishers.

Hellyer TJ, DesJardin LE, Teixeira L, Perkins MD, Cave MD and Eisenach KD, 1999. Detection of viable M. tuberculosis by reverse transcriptase-strand displacement amplification of mRNA. J. Clin Microbiol 37:518-523.

Henderson IR, Novarro-Garcia F, Nataro JP, 1998. The great escape: structure and function of the autotransporter proteins. Trends Microbiol. 6:370-378.

Hoffman CS and Wright A, 1985. Fusions of secreted proteins to alkaline phosphatase: An approach for studying protein secretion. Proc Natl Acad Sci USA 82:5107-11.

Hubbard RD, Flory CM and Collins FM, 1992. Immunization of mice with mycobacterial culture filtrate proteins. Clin Exp Immunol 87:94-98.

Hueck CJ, 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiol Molec Biol Rev 62:379-433.

Jobling MG, Palmer LM, Erbe JL, Holmes RK, 1997. Construction and characterization of versatile cloning vectors for efficient delivery of native foreign proteins to the periplasm of *E. coli*. Plasmid 38:158-73.

Jones CH, Danese PN, Pinker JS, Silhavy TJ, Hultgren SJ, 1997. The chaperone-assisted membrane release and folding pathway is sensed by two signal transduction systems. EMBO J 16:6394-6406.

Klein M, Meens J and Freudl R, 1995. Functional characterization of the *Staphylococcus carnosus* Sec A protein in *E. coli* and *B. subtilis secA* mutant strain. FEMS Microbiol. Lett. 131:271-77.

Lee VT and Schneewind O, 2001. Protein secretion and the pathogenesis of bacterial infections. Genes & Dev 15:1725-52.

Leininger JR, Donham KJ, Meyers WM,1980. Leprosy in a chimpanzee. Postmortem lesions. Int J Lepr other Mycobact Dis 48:414-421.

Lim EM, Rauzier J, Timm J, Torrea G, Murray A, Gicquel B and Portnoi D, 1995. Identification of M. tuberculosis DNA sequences encoding export proteins by using *phoA* gene fusions. J Bacteriol. 177:59-65.

Limia A, Sangari FJ, Wagner D and Bermudez LE, 2001. Characterization and expression of secA in *Mycobacterium avium*. FEMS Microbiol. Lett 197:151-157.

Mariani F, Cappelli G, Riccardi G, Colizzi V, 2000. Mycobacterium tuberculosis H37Rv comparative gene-expression analysis in synthetic medium and human macrophage. Gene 253:281-291.

Mehra V and Modlin RL, 1989. Th-0 lymphocytes in leprosy lesions. Curr Topics Microbiol Immunol. 155:97-109.

Meyers WM, Gormus BJ, Walsh GP, Baskin GB, Hubbard GB, 1991. Naturally acquired and experimental leprosy in nonhuman primates. Am J Trp Med Hyg 44:4, Pt 224-27.

Miller JF and Cossart P, 1999. Bacterial pathogenesis: before the post-genomic era. Curr Opin Microbiol. 2:15-17.

Moller S, Croning M and Apweiler R, 2001. Evaluation of methods for prediction of membrane spanning regions. Bioinformatics 17:646-653.

Mosmann TR, Coffman RL, 1989. TH1 and TH2 cells: different paterns of lymphokine secretion lead to different functional properties. Annu Rev Immunol 7:145-73.

Murphy CK and Beckwith J, 1987. Export of proteins to the cell envelope in *E. coli*. *E. coli* and *Salmonella typhimurium*: cellular and molecular biology (967-978). ASM, Washington DC.

Nielsen H, Engelbrecht J, Brunak S and Gunnar von Heijne, 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Prot Engineering 10:1-6.

Oliver DB and Beckwith J, 1981. *Escherichia coli* mutant pleiotropically defective in the export of secreted proteins. Cell 1981 Sep 25:765-72.

Owens MU, Swords WE, Schmidt MG, King CH, Quinn FD, 2002. Cloning, expression and functional characterization of the Mycobacterium tuberculosis secA gene. FEMS Microbiol Lett 211:133-141.

Pal PG and Horwitz MA, 1992. Immunization with extracellular proteins of Mycobacterium tuberculosis induces cell-mediated immune responses and substantial protective immunity in a guinea pig model of pulmonary tuberculosis. Infect. Immun. 60:4781-4792.

Pallen MJ, 2002. TheESAT-6/WXG100 superfamily and a new Gram-positive secretion system? Trends Microbiol 10:209-12.

Pugsley AP, Francetic O, Possot OM, Sauvonnet N, Hardie KR, 1997. Recent progress and future directions in studies of the main terminal branch of the general secretory pathway in Gram-negative bacteria. Gene 192:13-19.

Roberts AD, Sonnenberg MG, Ordway DJ, Furney SK, Brennan PJ, Belisle JT and Orme IM, 1995. Characteristics of protective immunity engendered by vaccination of mice with purified culture filtrate protein antigens of *Mycobacterium tuberculosis*. Immunol. 85:502-508.

Roberts JA, Marklund B, Ilver D, Haslam D, Kaack MB, Baskin G, Louis M, 1994. The Gal α (1-4) Gal-specific tip adhesin of Escherichia coli P-fimbriae is needed for pyelonephritis to occur in the normal urinary tract. Proc Natl Acad Sci USA 91:11889-11893.

Saleh MT, Fillon M, Brennan PJ, Belisle JT, 2001. Identification of putative exported/secreted proteins in prokaryotic proteomes. Gene 269:195-204.

Santini CL, Ize B, Chanal A, Muller M, Giordano G and Wu F, 1998. A novel sec-independent periplasmic protein translocation pathway in *E. coli*. EMBO J 17:101-12.

Sargent F, Bogsch EG, Stanley NR, Wexler M, Robinson C, Berks BC and Palmer T, 1998. Overlapping functions of components of a bacterial Sec-independent protein export pathway. EMBO J 17:3640-3650.

Settles AM, Yonetani A, Baron A, Bush DR, Cline K, Martienssen R, 1997. Sec-independent protein translocation by the maize Hcf106 protein. Science 278:1467-70.

Shepard C, 1960. The experimental disease that follows the injection of human leprosy bacilli into foot pads of mice. J Exp Med 112:445-454.

Shepard C, Walker L and Van Landingham R, 1978. Heat stability of *M. leprae* immunogenicity. Infect Immun 22:87-93.

Sonnenberg MG and Belisle JT, 1997. Definition of *Mycobacterium tuberculosis* culture filtrate proteins by two-dimensional polyacrylamide gel electrophoresis, N-terminal amino acid sequencing and electrospray mass spectrometry. Infect. Immun. 65:4515-4524.

Sonnhammer E, von Heijne G, Krogh A, 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Intel Syst for Mol Biol 175-182.

Spencer JS, Marques AM, Lima MC, Junqueira-Kipnis AP, Gregory BC, Truman RW and Brennan PJ, 2002. Antigenic specificity of the *Mycobacterium leprae* homologue of ESAT-6. Infect Immun 70:1010-13.

Thanabalu T, Koronakis E, Hughes C, Koronakis V, 1998. Substrate-induced assembly of a contiguous channel for protein export from *E. coli*: reversible bridging of an inner-membrane translocase to an outer membrane exit pore. EMBO J 17:6487-96.

Thanassi DG and Hultgren SJ, 2000. Multiple pathways allow protein secretion across the bacterial outer membrane. Curr Opin. Cell Biol. 12:420-430.

Thanassi DG, Saulino ET, Hultgren SJ, 1998. The chaperone/usher pathway: a major terminal branch of the general secretory pathway. Curr Opin Microbiol 1:223-231.

Timm J, Perilli M, Duez C, Trias J, Orefici G, Fattorini L, Amicosante G, Oratore A, Joris B, Frere JM, 1994. Transcription and expression analysis, using *lacZ* and *phoA* gene fusions of *M. fortuitum* beta-lactamase genes cloned from a natural isolate and a high level beta-lactamase producer. Mol Microbiol 12:491-504.

Tjalsma H, Bolhuis A, Jongbloed JD, Bron S and van Dijl JM, 2000. Signal peptide-dependent protein transport in *B. subtilis*: a genome-based survey of the secretome. Microbiol Mol Biol Rev 64:515-547.

Triccas JA and Gicquel B, 2000. Life on the inside: Probing *Mycobacterium tuberculosis* gene expression during infection. Immunol Cell Biol 78:311-317.

Truman RW, Sharman EJ, Hagstad HV, Hugh-Jones ME, Wolff A, Hasting RC, 1986. Evaluation of the origin of *M. leprae* infections in the wild armadillo, Dasypus novemcinctus. Am J Trop Med Hyg 35:588-593.

Truman RW, Krahenbuhl JL, 2001. Viable *M. leprae* as a research reagent. Int J Lepr other Mycobact Dis 69:1-12.

Van Wely KHM, Swaving J, Klein M, Freudl R and Driessen AJM, 2001. The carboxyl terminus of the *B. subtilis* SecA is dispensable for protein secretion and viability. Microbiol 146:2573-2581.

Von Heijne G, 1983. Patterns of amino acids near signal-sequence cleavage sites. Eur J Biochem 133:17-21.

Von Heijne G, 1985. Signal sequences, the limits of variation. J. Mol. Biol. 184:99-105.

Waters MFR, 1992. Royal Society of tropical medicine and hygiene meeting at Manson House. Leprosy 1962-92, 87:499-517.

Weldingh K, Rosenkrands I, Jacobsen S, Rasmussen PB, Elhay MJ and Andersen P, 1998. Two-dimensional electrophoresis for analysis of *M. tuberculosis* culture filtrate and purification and characterization of six novel proteins. Infect Immun 66:3492-3500.

Wiker HG, Harboe M and Lea T, 1986. Purification and characterization of two protein antigens from the heterogeneous BCG 85 complex in *Mycobacterium bovis* BCG. Int Arch of Allergy and App Immunol 81:298-306.

Wiker HG, Harboe M and Nagai S, 1991. A localization index for distinction between extracellular and intracellular antigens of *M. tuberculosis*. J Gen Microbiol 137:875-884.

Wiker HG, Michell SL, Hewinson RG, Spierings E, Nagai S and Harboe M, 1999. Cloning, expression and significance of MPT53 for identification of secreted proteins of *Mycobacterium tuberculosis*. Microb. Pathogen. 26:207-219.

Wiker HG, Wilson MA and Schoolnik GK, 2000. Extracytoplasmic proteins of *Mycobacterium tuberculosis* - mature secreted proteins often start with aspartic acid and proline. Microbiol 146:1525-1533.

Williams DL, Wheeler PR, Torrero MN, Truman RW, Yoder M, Morrison N, Bishai WR and Gillis TP, 2003. Transcriptionally active genes associated with intracellular survival of the leprosy bacillus. Mol. Mocrobiol. (submitted).

Yamamura M, Uyemura K, Deans RJ, 1991. Defining protective responses to pathogens: cytokine profiles in leprosy lesions. Science 254:277.

Zambryski P, 1988. Basic processes underlying *Agrobacterium*-mediated DNA transfer to plant cells. Annu Rev Genet 22:1-30.

VITA

Marina Noemi Torrero, the daughter of Oscar and Hilda Torrero, was born in Buenos Aires, Argentina, on October 30, 1967. In October, 1993, she obtained a bachelor of science degree in biology from the University of Buenos Aires. She worked at the School of Medicine of Buenos Aires doing research in bacteriology and molecular biology from 1990 to 1994. She came to United States in November, 1994, to join a research program studying anti-mycobacterial natural compounds at the National Hansen's Disease Programs at Louisiana State University, Baton Rouge, Louisiana. In January, 1997, she enrolled in the graduate program at Louisiana State University to pursue a master's degree in microbiology. Upon completion of the master's degree in 1998, she applied to Louisiana State University School of Veterinary Medicine to pursue a doctoral degree in pathobiological sciences. Her graduate work on secreted proteins from *Mycobacterium leprae* was guided by Dr. Thomas Gillis.