A Bioinformatic study on the feasibility of a cross-species proteomics analyses of mycobacteria

Elinambinina Rajaonarifara

University of Cape Town

Supervised by: Professor Jonathan Blackburn

Co-supervisor: Professor Nicola Mulder

30 January 2013

Submitted in partial fulfillment of an Msc at the University of Cape Town

Abstract

Database search approaches via Mass Spectrometry (MS) are fairly accurate methods for protein identification making use of the proteins database. However, many species, especially those within mycobacterial species are still not annotated and do not have protein sequences in any known database sources. Although *de novo* peptide sequencing approaches have been introduced to overcome that issue, their success requires high quality data. Accordingly, extending proteomic database search methods to include non-annotated mycobacteria is of great interest for a more expanded and accurate result.

The first part of our study involves analysis of the proportion of identical *in silico* tryptic peptides shared between different mycobacterial organisms relative to their distance in phylogeny. This aims to evaluate the use of the closest annotated species' protein database for an MS analysis of non-annotated species. The result of this first part highlights the utility of a cross-species proteomic analysis for mycobacterial species within a phylogenetic distance less than 0.3 to each other. The second part involves the use of a six frame translation database obtained from the genome sequence for proteogenomic annotation. This allows identification of potential novel proteins from species with incomplete databases and may also be applied to non-annotated species. Applied to *Mycobacterium avium*, this methodology allowed the identification of 81 extra proteins not previously reported in the existing database of *M. avium*.

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Elinambinina Rajaonarifara 30 January 2013

Contents

1. Introduction

1.1 Motivation

Mycobacterium is a genus of *Actinobacteria* within the order *Mycobacteriales* in the family *Mycobacteriaceae* [1]. It comprises pathogens known to cause serious diseases such as Leprosy and Tuberculosis in animals and in humans. Tuberculosis especially remains a major world health problem [2]. Although BCG (Bacillus Calmette-Guerin) is used as a vaccine, its effectiveness in human varies widely especially in adults and the result has been less than desired [3]. Moreover, the increased rate of drug resistance undermines drug efficacy towards the treatment of tuberculosis [4]. The study of proteome complement expressed in cell is useful and may contain important information regarding the functioning and activity of the disease pathogen since the proteins are actually representative of the phenotype. Therefore accurate protein identification from sample mixture is a useful step forward towards development of new biomarkers, potential drug targets and vaccines.

Shotgun proteomics via mass spectrometry (MS) has emerged as the most commonly used analytical method for such proteomic analysis. In this technology, computational methods such as database search approaches are used to analyse experimental mass spectra by matching them against theoretical mass spectra derived from database proteins [5–7]. These methods are fairly accurate but limited to the identification of peptides from known genomes, excluding those from non-annotated organisms such as *Mycobacterium kansasii* and *Mycobacterium shottsi*. *De novo* peptide sequencing approaches [8,9] which extract amino acid sequences directly from the experimental spectra have been introduced to overcome this problem but their success requires high quality data with high signal to noise ratios, high spectral resolu-

tion and mass accuracy. To date, *de novo* sequencing of peptides in complex mixtures therefore remains challenging.

In this thesis, we investigated two approaches towards the identification of proteins from mycobacterial species which are non-annotated or have incomplete databases. We study the feasibility of a cross-species analysis within mycobacterial species to evaluate the use of the closest annotated species database to carry out database search-based analysis of spectra generated from MS analysis of non-annotated organisms. We then study the use of proteogenomic analyses both for mycobacterial species with incomplete databases as well as sequenced but non-annotated mycobacterial species by creating an automated six frame translation database from the genome sequence.

1.2 Analytical technique

Mass spectrometry (MS) is one of the most commonly used analytical technique for protein identification. The process of an MS analysis can be divided into three fundamental parts comprising sample preparation, mass spectrometry phase and data analysis. Tandem mass spectrometry or MS/MS involves multiple steps of mass spectrometry for more accurate characterization. Figure 1.1 illustrates the workflow of an MS/MS analysis.

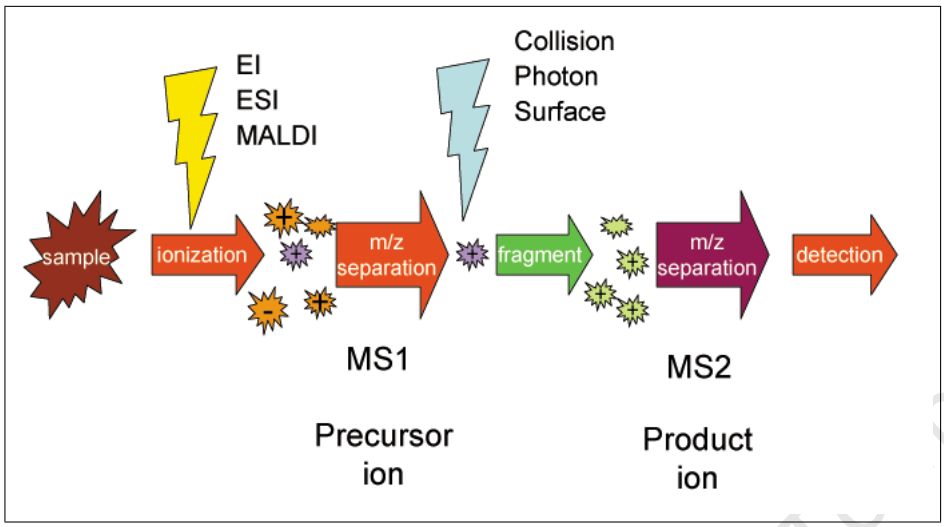

Figure 1.1: Tandem mass spectrometry [10].

During the sample preparation, complex mixtures are fractionated to reduce complexity and to remove the highly abundant component of the proteome. Proteins can be separated according to their size, hydrophobicity, charge, isoelectric point or affinity. A number of methods can be used to fractionate sample mixture including Isoelectric focusing [11–13] in which proteins are separated based on their isoelectric point, Liquid Chromatography [14] or SDS-PAGE [15]. Enzymes such as trypsin are added to digest the proteins. Tryptic digests should then be purified so that molecules that interfere with ionisation/detection such as salts, detergents and chaotropes are eliminated. Digested peptides are then loaded onto the mass spectrometer to be analysed. This process is shown in Figure 1.2.

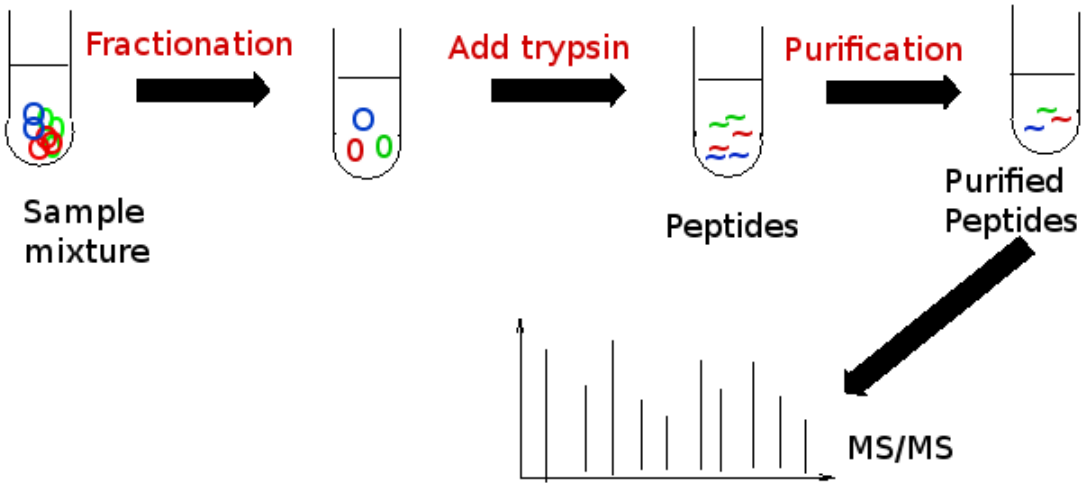

Figure 1.2: Sample preparation.

In the mass spectrometry phase, samples are ionised and parent ions are selectively fragmented by collision induced dissociation for an MS/MS analysis. Fragmented ions are extracted and are separated according to their mass-to-charge ratio(m/z). The separated ions are then detected and the signal is sent to a data system where the m/z ratios are stored together with their relative abundance for presentation in the format of a m/z spectrum.

Once the spectra are generated, computational methods are needed to extract the correct amino acid sequence from the spectra: this is the post-mass spectrometry phase and consists of interpreting the displayed spectra from the mass spectrometry to get information about the protein present in the sample. Some existing methods for computational analysis of MS/MS data are briefly reviewed in Chapter Two.

1.3 Thesis outline

This report is comprised of five Chapters. In Chapter 2, we present some commonly used methods for protein identification. Examples of database search approaches are described in Section 2.1 and some *de novo* sequencing algorithms are explained in Section 2.2. The advantages and disadvantages of these existing method are discussed in Section 2.3. The feasibility of cross-species analysis is studied in Chapter 3. A brief overview is presented in Section 3.1. We describe our methodology for the cross-species analysis along with the result obtained throughout this part in Section 3.2. Section 3.3 introduces the conclusion of how far a cross-species analysis holds true. Chapter 4 advances the proteogenomic analysis of *M. avium* and *M. kansasii* using six frame translation. Section 4.1 shows a brief overview of Chapter 4. We present our methodology for the proteogenomic analysis in Section 4.2 and Section 4.3 provides the result from the proteins identification. We summarize our work and present the result obtained throughout the study in Chapter 5.

2. Existing methods to interpret MS/MS data

Reliable interpretation of the MS/MS spectra is critical for providing confidence in the protein identification. Several methods have been developed to address this task including database searches, which make use of an *in silico* database to match theoretical MS/MS spectra with the experimental MS/MS spectra. The limitation of database search methods for identification of peptides within an unknown genome can be solved in principle by the use of *de novo* sequencing algorithms which extract sequence information directly from the MS/MS data without the need for a sequence database. We present in this chapter a brief review of major existing database search and *de novo* sequencing approaches and discuss their advantages and limitations.

2.1 Database search methods

Database search algorithms involve the scanning of all known peptide MS/MS spectral space to find the best match to experimentally observed MS/MS spectra. Peptides that have mass matching with the experimental peptide mass within some tolerance are assigned to be candidates for Peptide Spectral Matching(PSM). Their sequences in the database are converted into hypothetical MS/MS spectra to enable the comparison. The experimental spectrum is then matched with the theoretical one obtained from the candidate peptides database and a score is assigned for each PSM. The database search method's workflow is illustrated in Figure 2.1. The most popular database searches engines include Sequest [5], Mascot [6] and X! tandem [7].

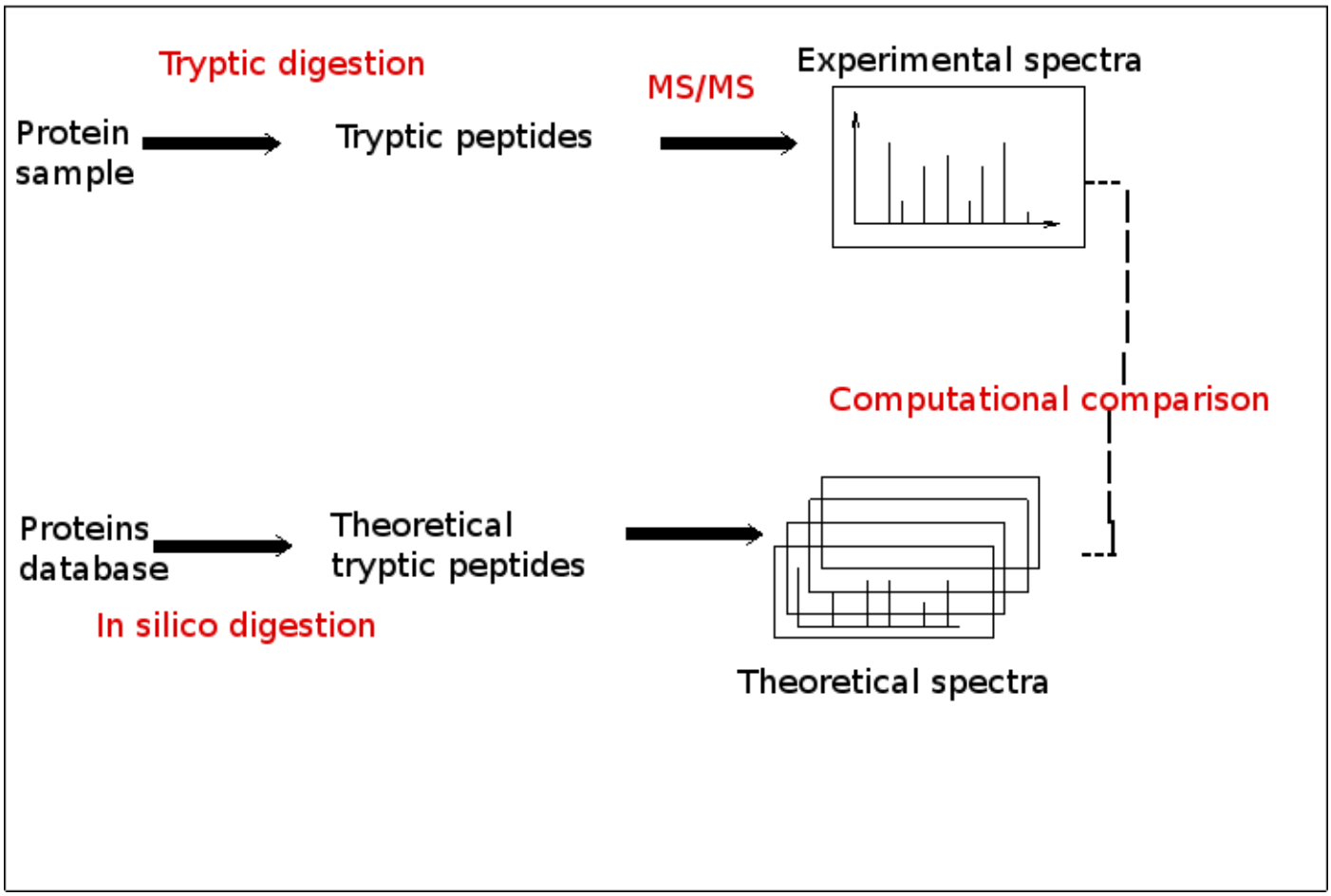

Figure 2.1: Database search method workflow.

2.1.1 Sequest

Sequest was developed by Eng et al [5] originally in 1994 as one of the earliest database search methods. It was developed to interpret mass spectra from experimental result using a protein database by converting the protein sequences to a predicted MS/MS spectra for tryptic peptides and then by comparing them with the experimental spectra. Although various versions of Sequest exist currently, the basic algorithm remains approximately the same. It can be described as comprising four steps.

The first step consists of the tandem mass spectra data reduction. The mass-to-charge ratio of the fragment ion is rounded to the nearest integer, noise filtering is processed by eliminating all but the 200 most abundant ions and then the remaining ions are renormalized by 100. The abundances of fragment ion within ± 1 u of each other are equalized to the higher value.

In the second step, the experimental spectra are matched to a proteolitic peptide sequence in the database according to their molecular weight. *In silico* generated tryptic peptide sequences are retrieved and scanned to find the best combination of amino acids that closely match the mass of the experimental peptide, the amino acid masses are summed until the *in silico* peptide mass falls within a defined mass tolerance from the experimentally observed mass. The masses of the fragment ions were obtained using Equation 2.1:

$$b_n = \sum a_n + 1 \qquad\qquad y_n = MW - \sum a_n \tag{2.1}$$

where a_n is the mass of amino acids, b_n is the type $b-ion$ and y_n is the type $y-ion$. Equation 2.1 equates to the mass-to-charge ratio value for fragment ions in the charge +1 state that is considered by the Sequest scoring routine.

Step three involves the preliminary scoring method. Each *in silico* peptide se-

quence is scored depending on the number of predicted fragment ions denoted by n_i that match ions observed in the experimental spectrum within a defined mass tolerance, their abundances i_m, the continuity of an ion series, the presence or absence of an ammonium ion and the total number of predicted sequence ions n_t. The score is then given in Equation 2.2:

$$S_p = \frac{(\sum i_m) n_i (1+\beta)(1+\rho)}{n_t} \tag{2.2}$$

where β represents the type-b and y ions continuity and ρ represents the presence or absence of an ammonium ion and their respective amino acids in the predicted sequence. This is a preliminary score and the top 500 amino acid sequences obtained using this score are then analysed in the 4^{th} step using the cross correlation analysis.

The fourth step is the most important for scoring in Sequest. It entails the comparison of each of the 500 preliminary PSMs to the experimental spectra. Their spectra are constructed in the following way. Values that represent mass-to-charge ratio of type-b ion or type-y ion are assigned a magnitude of 50.0. A magnitude of 25 is assigned to mass-to-charge ratio within ± 1 u of type- b ion or type- y ion. Mass-to-charge ratio of type-a ion and mass-to-charge ratio within ± 1 u are assigned a magnitude of 10. The experimental spectra represented as x and the reconstructed spectra represented as y from the database are then compared using the cross correlation formula R_τ (Equation 2.3) and the final score is denoted by $Xcorr$ as calculated in Equation 2.4 and normalized to 1:

$$R_\tau = \sum_{i=0}^{n-1} x[i]y[i+\tau] \tag{2.3}$$

$$XCorr = R_0 - \frac{\left(\sum_{\tau=-75}^{\tau=75} R_\tau\right)}{151} \tag{2.4}$$

where τ is a displacement value. The $Xcorr$ measure is an absolute measure of spectral quality and closeness of fit to the model spectrum. To distinguish correct

identification from false positives, the difference between the normalized cross correlations of the first and the second best amino acid sequences are used.

2.1.2 Mascot

The Mascot search engine algorithm is based on probability MOWSE scoring. The ions score is $-10 \times Log(P)$ where P is the probability that the match occurs randomly [6]. The proteins score is then derived from the ion score as a non-probabilistic basis for ranking protein hits. The Mascot method interface is shown in Figure 2.2 where the user can choose the database to search the dataset against, and select the enzyme to be used for digestion. The modification can also be set as 'fixed' or 'variable' as required. It also allows the user to select peptide and MS/MS tolerance as well as the desired peptide charge.

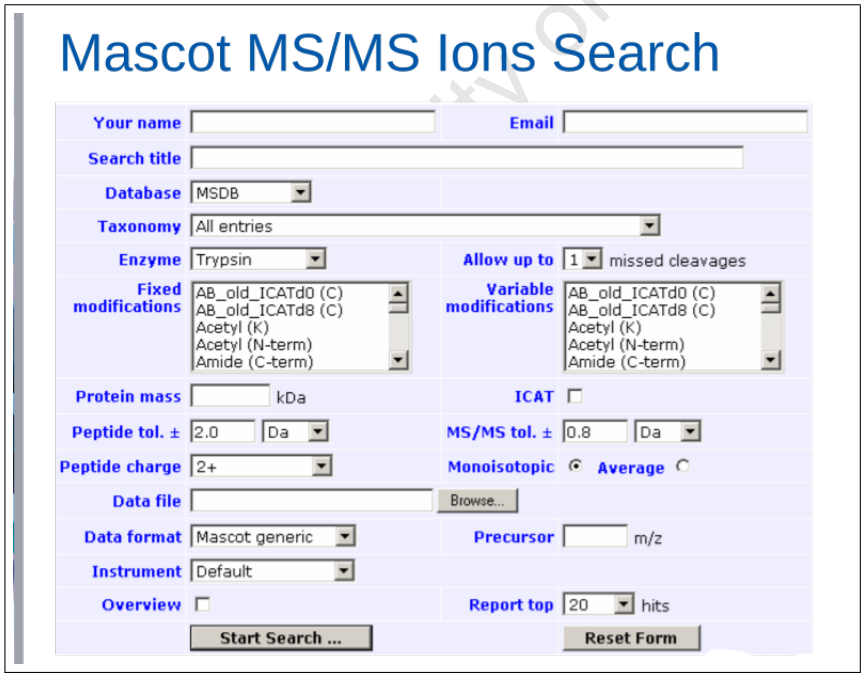

Figure 2.2: Mascot search engine interface [6].

The report from a Mascot search result is set to represent peptides with score higher than the significance threshold which depends on the size of the experimental MS/MS dataset and database used. For each identified protein, information about all of the peptides identified can be obtained by clicking on its accession number. The protein view is also included for the top one, two or three proteins where the percentage of the sequence coverage is given and the protein sequence is displayed with the experimentally observed peptides identified in bold red .

2.1.3 X!Tandem

X! Tandem differs from Sequest in the way the scoring function is performed. X! Tandem considers only b and y type ions. Only peaks that match to the hypothetical spectra are used in this model [7]. The preliminary score used in X! Tandem is defined in Equation 2.5:

$$y/bScore = \left(\sum_{0}^{n} I_i \times P_i \right) \tag{2.5}$$

where I_i represents the peaks intensities while P_i takes the value 1 if the peak was predicted and 0 otherwise.

This preliminary score is multiplied by the factorial of the number of b ions and y ions which gives the $Hyperscore$ as defined in Equation 2.6:

$$HyperScore = y/bScore \times N_b! \times N_y! \tag{2.6}$$

X! tandem assumes that the amino acid sequence with the best score is the only possible correct match and to evaluate the correctness of the first best score, it looks at the distribution of lower scoring hits.

2.2 *De novo* peptide sequencing

De novo peptide sequencing involves extracting amino acid sequence information directly from the MS/MS without prior knowledge of any protein database. With an idealized process of fragmentation in the mass spectrometer, a peptide would be cleaved at random between every two consecutive amino acids and a single charge would be retained on only the N-terminal fragment [16]. This would enable the determination of the peptide sequences by simply converting the mass differences of consecutive ions in the spectrum to the corresponding amino acids. However in practice, the fragmentation process in mass spectrometer is far from ideal [16]. Therefore, a scoring function is used to evaluate the match between the candidate peptide and the given experimental spectrum in a *de novo* sequencing approach. A typical *de novo* sequencing workflow is illustrated in Figure 2.3. Here we describe the algorithm of two *de novo* sequencing methods namely Pepnovo [8] and PEAKS [9].

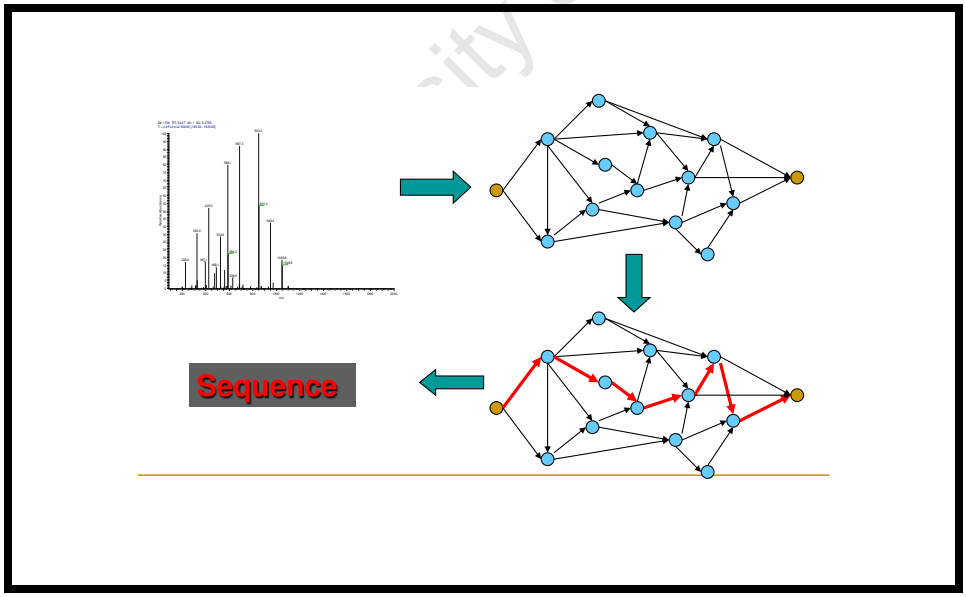

Figure 2.3: *De novo* sequencing algorithm.

2.2.1 Pepnovo

Like most *de novo* peptide sequencing approaches, Pepnovo [8] uses a spectrum graph created from the MS/MS spectra to assign the peptide sequences. The nodes are cleavage sites having mass m and score $Score(m, S)$ where S represents the spectrum. The score for each node is computed as a function of the probability $P_{CID}(\overrightarrow{I}|M, S)$ of detecting an observed set of fragment intensities $\overrightarrow{I}$ given that mass m is a cleavage site in the peptide that created S and the probability $P_{RAND}(\overrightarrow{I}|M, S)$ that the peaks in the spectrum are caused by a random event. This score is given by the formula 2.7:

$$Score(m, S) = log(\frac{P_{CID}(\overrightarrow{I}|M, S)}{P_{RAND}(\overrightarrow{I}|M, S)}) \tag{2.7}$$

which means that a peak intensities $\overrightarrow{I}$ is more likely to be caused by a cleavage event if the score is positive and it is more likely to be caused by a random event if the score is negative.

To compute the probability $P_{CID}(\overrightarrow{I}|M, S)$, Frank and Pevzner [8] used a network diagram as shown in Figure 2.4. $V\{b, y, ...\}$ denote the vertexes in the network excluding the pos(m), N-aa and C-aa; π denotes the v's parents in the graph; $\overrightarrow{\pi}(v)$ denotes the set of values assigned to the vertexes $\pi(v)$. $P_{CID}(I_v = i|\overrightarrow{\pi} = i_1, i_2, ...)$ is the probability of detecting the intensity i at fragment ion v given the intensities detected at its parents. The probability $P_{CID}(\overrightarrow{I}|M, S)$ can be decomposed in the function of the $P_{CID}(I_v = i|\overrightarrow{\pi}, m, S)$ as shown in the relation 2.8 since each vertex v is independent of other vertexes in the network graph given that the values of its parents are known.

$$P_{CID}(\overrightarrow{I}|M, S) \prod_{v \in V} P_{CID}(I_v = i|\overrightarrow{\pi}, m, S) \tag{2.8}$$

The probability of randomly observing a peak with a given intensity in the spec-

trum is computed using the fact that each peak is distributed independently of the others and $P_{RAND}(\overrightarrow{I|m,S})$ are the products of the probabilities of seeing the individual peaks. This is given in Equation 2.9.

$$P_{RAND}(\overrightarrow{I}|M,S) = \prod_{i=1}^{k} P_{RAND}(i_I|n_{i_1}, n_{i_2}, ..., n_{i_d}) \tag{2.9}$$

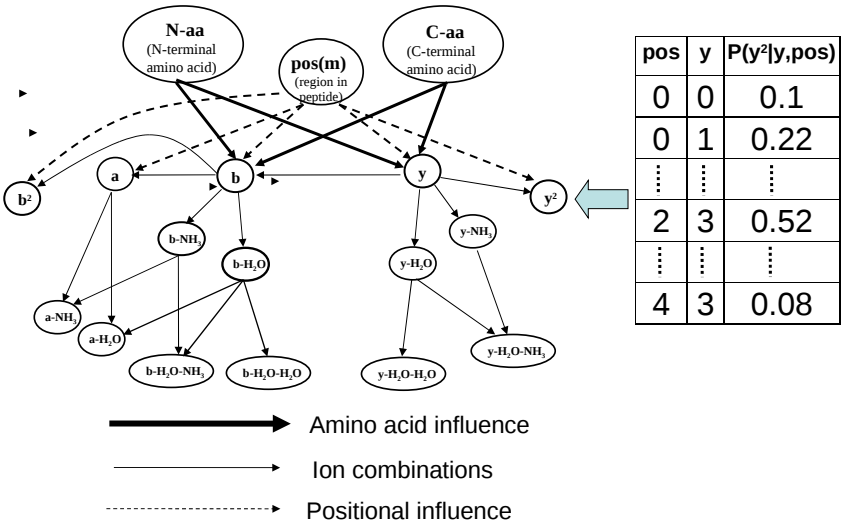

Figure 2.4: Probabilistic network for the fragmentation model of tryptic peptides.

2.2.2 PEAKS *de novo* peptide sequencing

The scoring for PEAKS *de novo* sequencing [9] is comprised of four steps involving preprocessing, candidate computation, refined scoring and global and positional confidence scoring. The first step involves noise filtering and peak centering as well as

deconvolution of the doubly and triply charged species to singly charged ions. The second step consists of computing the 10000 best sequences of all possible combination of amino acids for a given precursor ion mass. PEAKS *de novo* sequencing method scoring is based on penalty/reward score computing of each possible mass value instead of a spectrum graph drawing. The objective is to find a sequence such that its y and b ions maximize the total rewards at their mass values. In the third step, each of the 10000 sequences is re-evaluated by considering a stricter mass tolerance and a reward for immonium ions and internal cleavage ions. A recalibration of the data is performed to account for minor deviations in MS/MS data. In the last step, the confidence score is computed for the top scoring peptide sequences.

2.3 Advantage and disadvantage of these methods

Both database search and *de novo* sequencing methods have their strengths and weaknesses. Database search methods are very successful for identification of already known peptides. The major drawback for any database search method is its limitation to fully annotated organisms only. Although the number of sequenced organisms is increasing with the advent of next generation DNA (Deoxyribonucleic acid) sequencing, there are still many species that are not annotated or only poorly annotated. Other limitations of database search approaches include the lack of assignment of a large portion of the MS/MS data due to the incompleteness of the database or low quality MS/MS spectra.

Non-annotated organisms as well as species with incomplete proteomic databases therefore need other ways such as *de novo* sequencing to interpret spectra from MS/MS data. However, the use of *de novo* sequencing necessitates backbone cleavage between each pair of adjacent amino acids which only occurs with very high qual-

ity data. As a result, *de novo* sequencing methods are known to be accurate for identification of single amino acids but still lack accuracy when identifying peptide sequences.

Therefore, we discuss in the next chapter the feasibility of extending database search methods to non-annotated species by the use of its closest sequenced and annotated species database. In particular, we describe a methodology based on cross-species analysis within mycobacterial species for proteomic spectrum matching.

3. Cross-species analysis

3.1 Overview

We have developed a methodology to study the feasibility of a cross-species proteomic analysis within mycobacteria. This method aims to evaluate the use of the closest species database to carry out database search methods to assign spectra generated from MS/MS analysis of non-annotated organisms. The feasibility of the cross-species analysis in our study is evaluated according to the closeness of pairs of mycobacterial species in phylogeny proportionally to the number of identical tryptic peptides shared between them. The closeness of the species is measured using the phylogenetic distance calculated from the 16S ribosomal RNA (16S rRNA) sequences.

We show in this project that the divergence between mycobacterial species in terms of identical peptides is directly related to their phylogenetic distance: the farther they are in phylogeny, the lower the number of peptides they share. This relationship enables the estimation of the number of peptides shared between non-annotated and annotated species knowing their distance separation in phylogeny. The result gives us insight into how far apart in phylogeny actinobacteria can be before we lose the ability to use the proteome of a sequenced organism to interpret spectra from a non-sequenced one. We focused our comparison on the number of tryptic peptides because typically the protein identification is based on PSM of tryptic peptides.

3.2 Methodology and Results

Two members of the *Mycobacterium avium* complexes (MAC) *i.e M. avium* (Av) strain 104 and *M. paratuberculosis* (Ptb), 4 species within the *Mycobacterium tuberculosis* complex (MTBC) *i.e M. bovis* (Bov), *M. bovis* BCG Pasteur (BCG), *M. tuberculosis* H37Rv (HRv) and *M. tuberculosis* KZN (KZN) and 4 additional mycobacterial species including *M. leprae* (Lep), *M. ulcerans* (Ulc), *M. marinum* (Mar) and *M. smegmatis* (Smeg) were analysed as the reference annotated species. Protein databases utilised, including the proteome files, and paralog and ortholog files for each species were obtained from Ensembl [17]. The non-annotated species used here to test our methodology were *M. kansasii* (Kans) and *M. shottsii* (Shot).

To evaluate the similarity between two species, the number of shared tryptic peptides were determined taking into account that true identical peptides should be from homologous proteins. Homologous proteins are proteins that are derived from the same ancestor. They are referred as "orthologs" when they are from different species and arose from speciation and "paralogs" if they result from duplication in the same species. Any other shared peptides are considered to be false positive. The closeness of two species, including non-annotated ones, was then calculated using their 16S rRNA sequences. Thereafter, the feasibility of the cross-species analysis of two species was measured by the proportionality of their peptides' similarity and their closeness in phylogeny.

3.2.1 Comparison in terms of identical peptides

Peptide filtering

Proteins from each proteome fasta file were digested *in silico* with the enzyme Trypsin using the software DBToolkit [18]. No miscleavages were allowed. The enzyme trypsin cuts after every Lysine (K) or Arginine (R) unless it is followed by a Proline (P). To make our comparison meaningful, a probability search was performed to determine in our subsequent analyses the shortest amino acid to be considered such that a peptide match between any two species was unlikely to be a chance event. The probability was computed for peptides of a given length starting with monopeptides and increasing in length until the probability of the peptide's random occurrence in any given proteome file is smaller than 0.05.

Given the restriction that for any tryptic peptide, the C-terminal residue must be R or K, for monopeptides, there are only two possibilities, either it is R or K. Therefore, the probability of having R is $\frac{1}{2}$ and likewise for K if there was only one monopeptide. For dipeptides, the possibilities are: XR or XK where X can be any amino acid but R, K. So the probability of getting any one specific dipeptide is: $\frac{1}{18}\times\frac{1}{2}$ if there was only one dipeptide. Pursuing this strategy, we have the probability of getting one specific peptide of 5 amino acid as follows: $\frac{1}{18}\times\frac{1}{18}\times\frac{1}{18}\times\frac{1}{18}\times\frac{1}{2}$ if there was only one peptide of length 5. If N is the number of peptides of length k in the organism, the likelihood of finding any one specific peptide of length k in that organism is then the following:

$$Pr = pr_k \times \left(1 + \sum_{n=1}^{N} (1 - pr_k)^n\right) \tag{3.1}$$

where pr_k denotes the previously calculated probability representing the likelihood of finding a specific peptide of length k if there was only one peptide of that length. Since

N is different for each organism, we used the mean of the N of all organisms used in this study for each k and we got, $Pr(monopeptide) = 0.99$, $Pr(dipeptide) = 0.89$, $Pr(pentapeptide) = 0.03$, which means that the probability that a specific peptide of length 5 occurs by chance is 0.03. The result from these probability computations leads us to the conclusion that the shortest peptide that can be considered as a non random match within any given mycobacterial proteome file is a pentapeptide.

Peptide comparison

As previously stated, the database of the organisms used in this study were collected from Ensembl [17]. We removed redundancy of the proteins within each species *i.e* proteins that have exactly the same sequence were reported only once. We then carried out *in silico* tryptic digests on the non-redundant proteome files for each organism and only peptides with greater than or equal to five amino acids were counted according to the probability search described above. We then counted the number of all shared peptides between each pair of species and then we further computed those shared between homologous proteins. We assumed that identical peptide between species could either genuinely be shared by ortholog proteins or by a protein and the paralog of its ortholog. Other shared peptides were then considered to be false positives.

The distribution of the number of identical peptides shared between ortholog proteins in any pairwise comparison of mycobacteria is shown in Figure 3.1 which reveals that an average ortholog protein pair shares 9 peptides and the highest peptide count shared between two ortholog proteins is 208. However, there are some proteins not reported to be orthologs yet they share as many peptides as two typical ortholog proteins. An example of one such protein is shown in Figure 3.2 for the protein

"Malate synthase" from *M. avium* and *M. bovis* BCG. These two proteins share 14 identical peptides, exceeding the number of shared peptide between most of ortholog proteins, yet are not annotated as orthologs. The alignment from CLUSTAL W [19] of these proteins is shown in Figure 3.3.

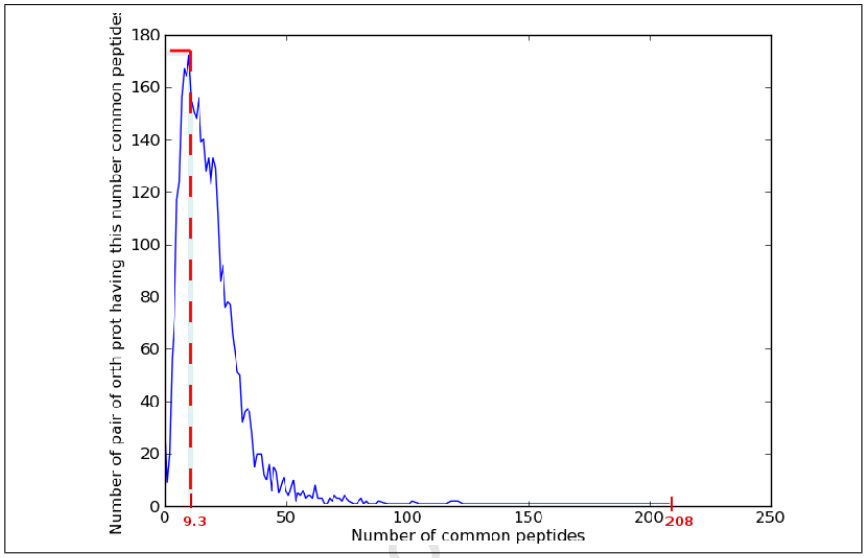

Figure 3.1: Distribution of peptides shared between ortholog proteins. The X-axis shows the numbers of identical peptides shared between pair of proteins that are ortholog. The Y-axis shows the number $\overline{y}$ (as mean for all pair of species) where y is the number of ortholog protein that share x identical peptides for each species-pair.

.

Figure 3.2: Non-ortholog proteins from *M. avium* and *M. bovis* BCG respectively sharing 14 identical peptides. Highlighted peptides indicate common peptides between the proteins.

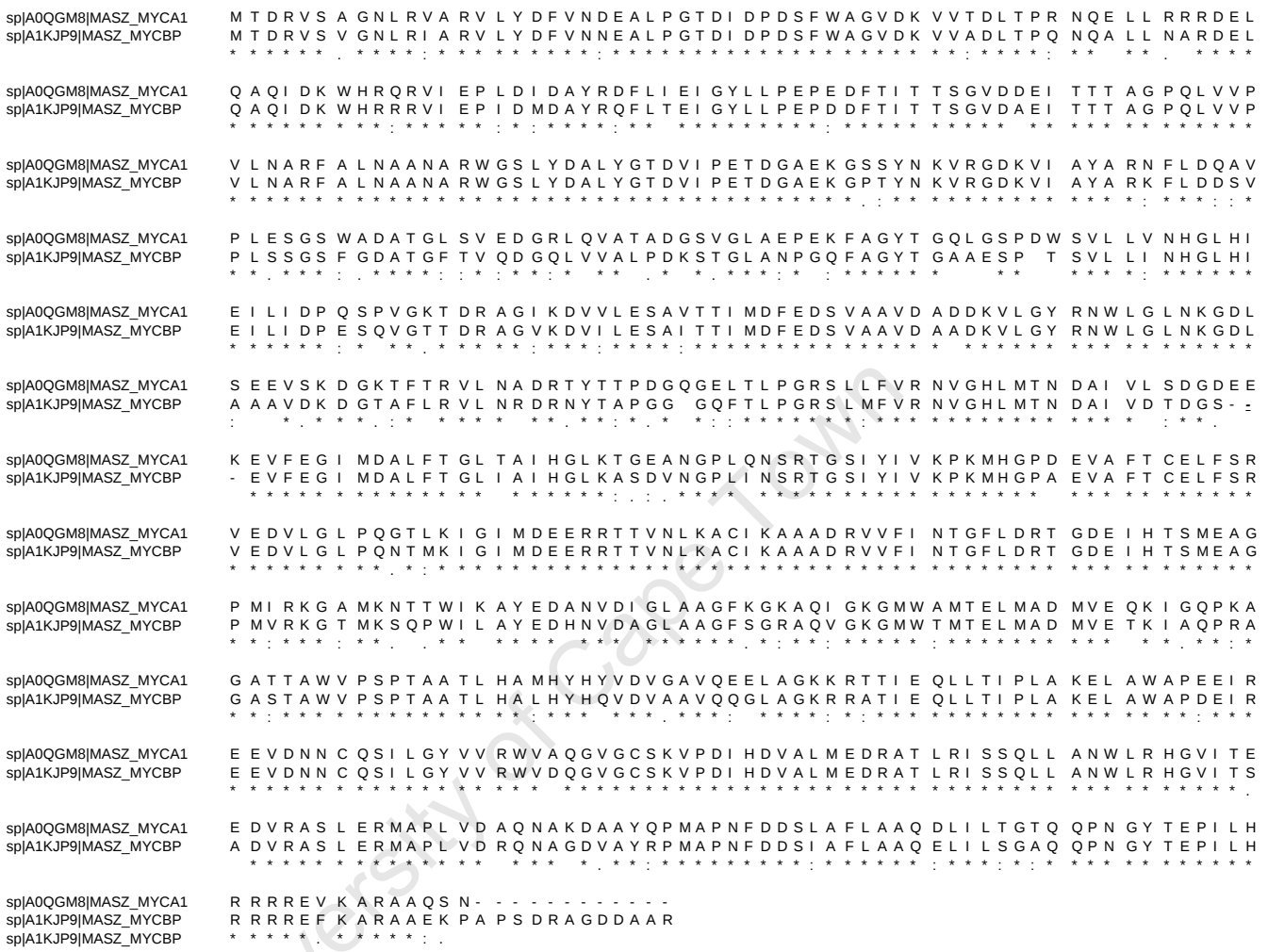

Figure 3.3: Alignment of the two non-ortholog proteins from *M. avium* and *M. bovis*.

Along with the fact that neither of these proteins has any other ortholog, this high amount of peptides similarity between these two sequences suggests that they may be missed in the orthologs classification. We consider peptides shared between such pairs as false false positive since they are grouped with the false positives but they should actually be true positive.

In Table 3.1, the similarity of pairwise mycobacterial species is represented in term of identical peptides. The table represents the name of the species (Species), the total number of tryptic peptides ($\geq$ 5 mers) in the species (Peptides), the name of the other species compared to the species in the first column (Other Species), the number of known ortholog proteins between the two species compared (Orth Prot), the number of identical shared peptides between otholog proteins in these species (Shared Pep), the false positive (FP, which is the number of peptide shared between two non-ortholog proteins) and the false false positive (FFP, which is the number of peptides shared between two proteins that may have been missed as being annotated as ortholog).

Table 3.1: Number of identical peptides shared between mycobacterial species.

Species	Peptides	Other Species	Orth Prot	Shared Pep	FP	FFP
M. avium	95956	*M. bovis*	2677	11172	942	101
		M. bovis BCG	2671	11163	965	90
		Mtb H37Rv	2701	11256	952	101
		Mtb KZN	2709	11240	965	99
		M. leprae	1381	5938	423	66
		M. marinum	3354	13132	1207	102
		M. smegmatis	3184	8166	1632	50
		M. ulcerans	2897	10947	918	108
		M. paratuberculosis	4005	67748	791	35
M. bovis	75469	*M. bovis* BCG	3837	73952	114	13
		Mtb H37Rv	3825	72864	201	44
		Mtb KZN	3776	71738	251	49
		M. leprae	1419	6638	333	73
		M. marinum	3127	22644	879	81
		M. smegmatis	2576	6691	1375	107
		M. ulcerans	2755	19873	671	113
M. bovis BCG	75307	*Mtb* H37Rv	3774	72547	72274	273
		Mtb KZN	3750	71521	327	68
		M. leprae	1413	6638	330	163
		M. marinum	3103	24082	868	223
		M. smegmatis	2568	6684	1373	160
		M. ulcerans	2747	19877	660	199

Species	Peptides	Other Species	Orth Prot	Shared Pep	FP	FFP
Mtb H37Rv	76514	*Mtb* KZN	3847	74238	62	6
		M. leprae	1419	6647	345	73
		M. marinum	3158	25110	887	81
		M. smegmatis	2600	6728	1392	99
		M. ulcerans	2771	19266	675	16
Mtb KZN	77032	*M. leprae*	1421	6624	342	71
		M. marinum	3173	25430	908	79
		M. smegmatis	2611	6714	1396	105
		M. ulcerans	2783	19240	684	16
M. leprae	32431	*M. marinum*	1448	6369	368	28
		M. smegmatis	1363	3915	563	54
		M. ulcerans	1403	5915	292	28
M. marinum	108933	*M. smegmatis*	3331	7872	1683	99
		M. ulcerans	3820	60082	171	42
M. smegmatis	123455	*M. ulcerans*	2789	6722	1265	79
M. ulcerans	81470					

This comparison in terms of identical peptides reveals that mycobacterial species within MTBC share about 98% identical peptides while those within MAC share 96%. The farthest removed mycobacterial pair studied here includes *M. smegmatis* and *M. avium* with 9% of their peptides in common. The number of false positives is higher when the similarity is lower which may relate to a poor ability to predict ortholog proteins between more distant species. We compute in the next section the closeness of two species according to their phylogenetic distance including the sequenced but

non-annotated species namely *M. kansasii* and *M. shottsii*.

3.2.2 Comparison in phylogeny

Phylogenetic distance

An accurate way of measuring the evolutionary distance between two mycobacterial species comprises the comparison of their 16S rRNA sequences. A reason for this is that the 16S rRNA is highly conserved in almost all bacteria [20]; it is about 1550 nucleotides in length which is large enough to provide necessary information about the evolution [21]. Furthermore, the function of 16S rRNA has not changed over time which suggests that the variable region can be used for measurement of the time of evolution [21].

The phylogenetic distance between two sequences is defined as the percentage of nucleotides in one sequence that are different from those in another taking into account the mutation that could have occurred considering the time rate. We measure the distance between species using their aligned 16S rRNA sequences from the Ribosomal database project (RDP) [22]. The distance used is the Jukes Cantor distance [23], calculated from the Formula 3.2.

$$K = -\frac{3}{4}\log(1 - \frac{4}{3}D) \tag{3.2}$$

where D is the proportion of nucleotides which differs between the two sequences. Table 3.2 lists the value of the distance between each mycobacterial species pair considered in this study.

Table 3.2: Phylogenetic distance computed using the 16S rRNA.

	Smeg	Ulc	Lep	Kans	Ptb	Bov	HRv	Shot	Av	BCG	KZN	Mar
Smeg	0	3.2	3.3	2.9	3.0	2.6	2.6	2.8	2.9	2.6	2.6	2.9
Ulc		0	1.7	1.1	1.3	0.7	0.7	0.4	1.1	0.7	0.7	0.2
Lep			0	1.2	1.2	1.1	1.1	1.3	1.2	1.1	1.1	1.3
Kans				0	0.8	0.8	0.8	0.8	0.5	0.8	0.8	0.8
Ptb					0	0.7	0.7	0.8	0.1	0.7	0.7	1.0
Bov						0	0.1	0.3	0.8	0.1	0.1	0.4
HRv							0	0.3	0.8	0.1	0.1	0.4
Shot								0	0.8	0.3	0.3	0.1
Av									0	0.8	0.8	0.8
BCG										0	0.1	0.4
KZN											0	0.4
Mar												0

Phylogenetic tree

Using the aligned 16S rRNA sequences from RDP, the tree of the 13 mycobacterial species studied here was constructed to view the closeness of the species. The tree was built using the Quicktree [24] program from the Mobyle portal web site [25] to get the newick format and then we used biopython [26] to display the tree in Figure 3.4. The Quicktree software uses a Neighbor-joining method for the construction of the tree.

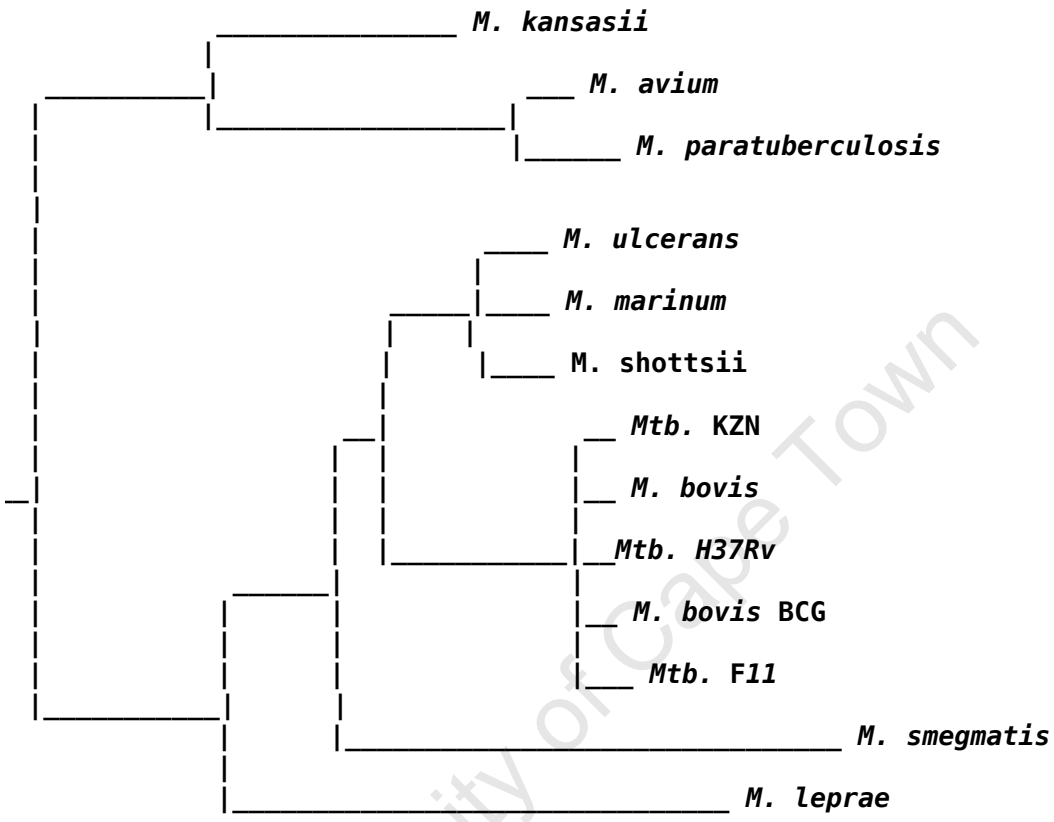

Figure 3.4: Phylogenetic tree of mycobacterial species using distance matrix from Mobyle portal. An outgroup was not included in this tree because we are interested in quantitative analysis of sequence divergence not in construction of a strict phylogenetic tree per se.

It can be seen from Table 3.2 and Figure 3.4 that the closest species to *M. kansasii* and *M. shottsii* that have been sequenced are respectively *M. avium* and *M. marinum* with phylogenetic distance 0.59 and 0.1 respectively. In order to estimate the number of peptides shared between these species, the result showing the proportionality between the number of peptides shared and the phylogenetic distance is presented in Subsection 3.2.3. This also enables us to determine the critical limit in phylogenetic distance outside of which this proportionality fails, and therefore the maximum phylogenetic distance between annotated and non-annotated organism before we lose the ability to use the former to assign MS/MS spectra from the latter.

3.2.3 Relationship between phylogenetic distance and peptides shared

The correlation between phylogenetic distance and the number of peptides shared identically between two annotated species is shown in Figure 3.5. This relationship allows the estimation of the number of identical peptides that will be shared between a sequenced and annotated species and non-annotated one given their distance in phylogeny.

The comparison of mycobacterial species was represented starting with two highly similar organisms, within the MTBC (*Mtb* H37Rv and *Mtb* KZN) whose distance is almost 0 in the phylogeny and ending with two species far away from each other (*M. avium* and *M. smegmatis*) with distance 2.9. It shows that the number of identical peptides shared by two species and their distance in phylogeny are related as expected, with the number of peptides shared decreasing from 98% to 8% and the distance in phylogeny increasing from 0 to 3.0.

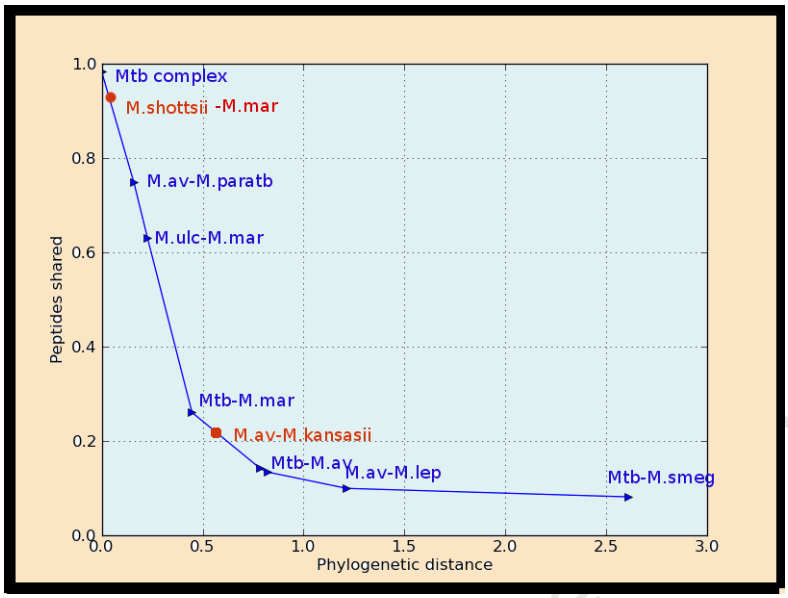

Figure 3.5: Relationship between the phylogenetic distance and the percentage identical shared peptides.

To confirm the feasibility of a cross-species proteomic analysis that includes non-annotated species, we looked at the number of false positives among the identical peptide between each pair compared. Figure 3.6 shows the distribution of false positives found in our study. For species that are significantly similar to each other, the number of false positives can be less than 2% of all their shared peptides. The significant increase in apparent false positives with increasing phylogenetic distance is interesting and may simply reflect increasingly inaccurate identification of true orthologs between species.

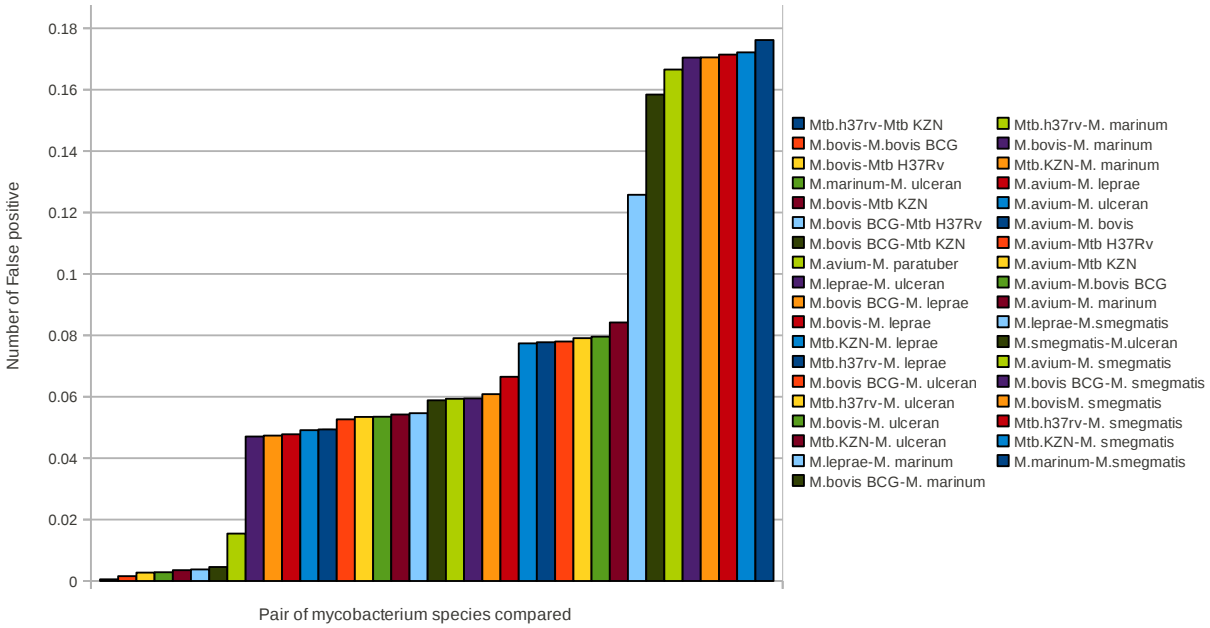

Figure 3.6: False positive rate distribution between all species-pairs compared. The false positive rate here is the proportion of the number of identical peptides shared between two non-ortholog proteins over all identical peptides.

3.2.4 Peptide conservation in mycobacterial species

Figure 3.5 shows expectedly that the fraction of peptides shared between two species decreased rapidly as their phylogenetic distance ranges from 0 to 1. However, unexpectedly this relationship tends to approach a constant value of 0.08 when the phylogenetic distance is above 1. This suggests that there is a strong conservation of certain peptides that are shared identically across pairwise mycobacterial species, even when the distance in phylogeny between mycobacteria is high. This conservation was seen to be across all mycobacterial species studied here, among them *M. leprae*

which is the species with the smallest proteome of mycobacteria containing only 1600 proteins. The conserved peptides are about 2600 in number and found in approximately 700 proteins in each species. A GO (Gene Ontology) enrichment analysis of the proteins conserved between *M. avium* and *M. bovis* are presented in Tables 3.3 and 3.4. Information about the GO enrichment of the conserved proteins was obtained using the StRAnGER [27] software and the GO annotation downloaded from integr8 [28] of each species was used as the background database. The StRAnGER software takes as input a group of protein IDs to be analysed and a background database of GO annotation and gives as output information about the GO enrichment of the group. It uses three choices of statistical test (hypergeometric test, Fisher exact test or χ^2 test) for the identification of enriched ontological terms. Here we used a p-value cut off of 0.01 and the Fisher exact test where the probability for a term T_i to be over-represented is given by Equation 3.3.

$$P(z,n,t,x) = \frac{\begin{pmatrix} z+x \\ z \end{pmatrix} \begin{pmatrix} t-z+n-x \\ t-z \end{pmatrix}}{\begin{pmatrix} n \\ t \end{pmatrix}} \tag{3.3}$$

n denotes the number of genes in the microarray/reference list, x the number of genes in the array/reference list associated with the term T_i, t the number of genes in the significant list and z the number of genes in the significant list annotated to the term T_i.

The result shows the significant enrichment of certain catalytic and binding activity for molecular function, and transport and metabolic process for biological process as seen in Tables 3.3, 3.4. This conservation indicates an underlying basic similarity of a subset of proteins from all mycobacteria, which represents about half of the whole proteome of *M. leprae*.

Table 3.3: GO enrichment analysis of the conserved proteins from *M. avium* and *M. bovis*

	M. avium		*M. bovis*	
	Go term	p-value	Go term	p-value
BP	GO:0045449	$4.84462003847e^{-08}$	GO:0045449	$3.33713057188e^{-11}$
	GO:0055085	$1.39227926965e^{-08}$	GO:0055085	$2.3442359165e^{-11}$
	GO:0055114	$1.19112497643e^{-11}$	GO:0055114	$3.84110521168e^{-11}$
MF	GO:0016787	$2.19974938984e^{-11}$	GO:0016787	$6.74162947689e^{-11}$
	GO:0016853	$4.95378182919e^{-11}$	GO:0016829	$6.38449293433e^{-11}$
	GO:0016740	$6.08113559508e^{-12}$	GO:0016740	$9.1432417193e^{-11}$
	GO:0016747	$0.00354735937991e^{-11}$		
	GO:0016874	$3.16768833386e^{-11}$	GO:0016874	$1.98052685363e^{-12}$
	GO:0046872	$3.77875508661e^{-12}$	GO:0046872	$2.57529553238e^{-11}$
CC	GO:0016012	$5.80362775615e^{-08}$	GO:0016012	$2.36374150031e^{-07}$
			GO:0030529	$1.79722903226e^{-11}$

Table 3.4: Description of the GO terms significantly enriched in the conserved proteins from *M. avium* and *M. bovis*

	Go term	Description
BP	GO:0045449	Regulation of cellular transcription
	GO:0055085	Transmembrane transport
	GO:0055114	Oxydation reduction
MF	GO:0016787	Hydrolase activity
	GO:0016853	Isomerase activity
	GO:0016829	Lyase activity
	GO:0016740	Transferase activity
	GO:0016747	Transferase activity, transferring acyle group other than amino-acyl groups
	GO:0016874	Ligase activity
	GO:0046872	Metal ion binding
CC	GO:0016012	Integral to membrane
	GO:0030529	Ribonucleoprotein complex

3.2.5 Validation of the cross-species analysis

The relationship between the distance in phylogeny of pairs of mycobacterium species and the number of their common tryptic peptides enables the determination of a phylogenetic distance limit for the use of a cross-species analysis. Species within the same family such as MTBC or MAC are seen to have significant similarity and with false positive rates of less than 2% of the total number of common peptides. The corresponding phylogenetic distance between these species are less than 0.3.

Moreover, as the distance increases from this number, the number of tryptic peptides shared decreases rapidly and less than 50% similarity is observed which will complicate the cross-species proteomic comparisons, at least at the peptide level.

Application to *M. avium*, *M. bovis* BCG and *Mtb* H37Rv

For the validation of our analysis, we have searched *M. avium*, *M. bovis* BCG and *Mtb* H37Rv MS/MS data using their respective database fasta file from Ensembl [17] first and then re-searched each using the *Mtb* H37Rv database. The methodology for protein identification used is described in Chapter 4, Subsection 4.2.2. The result is shown in Table 3.5

Table 3.5: Result obtained from a cross-species analysis of *M. avium*, *M. bovis* BCG and *Mtb* H37Rv.

Species	# of identified protein using *Mtb* H37Rv file	# of identified proteins using their own file	% of the # of identified proteins using *Mtb* H37Rv	Phylogenetic distance from *Mtb* H37Rv
Mtb H37Rv	1100	1100	100%	0.0
M. bovis BCG	1723	1780	96.80%	0.1
M. avium	221	1500	14.73%	0.8

The result shows that 1723 proteins have been identified from *M. bovis* BCG using the *Mtb* H37Rv database while 1780 are found using its own database. The number of protein identified from using *Mtb* H37Rv database is then about 96.80% of the number of protein identified using the *M. bovis* BCG fasta file which is not far from the theoretical result of 98% similarity between these two species. From *M. avium*,

the experimental result of 221 proteins using the *Mtb* H37Rv database represents approximately 14.73% of the whole protein identified using its own database which also agrees with the result from the theoretical study of 14% similarity. Interestingly, if we take the inverse of these percentages they correlate with the genetic distances: $\frac{1}{96.8} = 0.01$ and the genetic distance is 0.1, while $\frac{1}{14.73} = 0.07$ and the distance is 0.8.

Possible application to *M. kansasii* and *M. shottsi*

Knowing that the closest species to *M. kansasii* and *M. shottsi* are *M. avium* and *M. marinum* respectively, we plotted them in the Figure 3.5 based on their phylogenetic distance. This enabled us to estimate the percentage identical peptides between *M. marinum* and *M. shottsii* to be 98% but only 21% for *M. avium* and *M. kansasii*. This means that using the *M. marinum* database to identify spectra from *M. shottsii*, we can expect 98% true positives from the result while if we use the *M. avium* database for *M. kansasii*, we can only expect to assign $\frac{1}{5^{th}}$ of the spectra that are true matches to *M. kansasii* peptides; clearly this will be limiting for any proteomic analysis of *M. kansasii*, suggesting that an alternative bioinformatic approach is likely to be necessary. An experimental mass spectrometry protein identification was produced for *M. kansasii* using the *M. avium* database and the result confirms this theory, where we identified only 275 proteins in *M. kansasii* while about 1500 ($\sim$ 18%) proteins were identified from *M. avium* mass spectra.

3.3 Summary

Data from the study of a cross-species *in silico* analysis reveals the relationship between the tryptic peptides shared identically between two species and the phylogenetic distance between the species. The relationship enables the evaluation of how far apart actinobacteria can be in phylogeny before we lose the ability to use the proteome of an annotated organism to interpret spectra from a non-annotated one. It also shows that there is a strong conservation of about 2600 tryptic peptides accross all mycobacterial species; these are present in 700 proteins which represents half of all proteins in the mycobacterium species with the smallest proteome(*M. leprae*), and perhaps suggests that these 700 proteins are the core, basal, essential machinery of mycobacteria, so further examination of these may be warranted from a drug target or vaccine target perspective.

Our data suggests that a cross-species analysis can be applied to species within phylogenetic distance less than 0.3 such as those within the same family, MTBC or MAC. It also suggests that proteomic analysis of non-annotated species is possible, providing their distance from a sequenced and annotated organism is less than 0.3. For instance, a cross-species preotomic analysis can be applied to *M. shottsii* using the *M. marinum* database but for *M. kansasii*, the use of a cross-species analysis will not enable the identification of more than 20% of the expected identified proteins from its sample. The next part of our study describes another way of identifying proteins from sequenced but non-annotated species like *M. kansasii* as well as those with incomplete annotation by creating an open reading frame (ORF) database from a six frame translation of the genome.

4. Proteogenomic analyses

4.1 Overview

Interpretation of MS/MS spectra remains a big challenge in the proteome area. Generally, despite huge advancement in mass spectrometry equipment in recent years, still typically a huge part of the whole data from MS/MS is not inferred to peptides. One reason for such inefficiency may be the absence or incompleteness of the genome annotation that gives rise to the protein database for the species being studied. Here we present a proteogenomic analysis of *M. avium* -representing a well sequenced species in mycobacteria- and *M. kansasii* representing a poorly or non-annotated species. This analysis consists of searching the MS/MS data against the genome sequence database using an automated six frame translation and comparing the result with those obtained from use of the existing protein database.

MS/MS analysis of the *M. avium* and *M. kansasii* samples were analysed using their respective six frame translation database. In *M. avium*, we were able to identify 1594 proteins including 81 that had not previously been annotated in the Ensembl [17] proteome file. Following the same experiment with *M. kansasii* using its six frame translation database, we were able to identify 160 proteins from the *M. kansasii* mass spectra, which is twice the number of *M. kansasii* proteins currently present in the Uniprot database [29].

4.2 Methodology

4.2.1 Six frame translation

There are six possible reading frames for a genome sequence. In any one reading frame, open reading frames can exist between stop codons with each triplet of nucleotides encoding an amino acid that can be predicted using the bacterial codon table. Each open reading frame typically has a start codon. The possible start codons for bacteria are: ATG (coded as Methionine M), GTG (coded as Valine V), and sometimes, TTG and CTG (coded as Leucine L). All start codons are read as M during the translation. In bacteria, the start codon is usually preceded by a ribosome binding site, but these remain difficult to identify in an automated scanner due to the sequence variability of ribosome binding site. Here we therefore decided to adopt a simpler approach, simply defining an ORF as having a start codon and a stop codon and accepting that this will over predict the total number of ORFs in any given bacterial genome. We wrote a python script to model the translation of the *M. avium* and *M. kansasii* genomes requiring minimum peptide size of 20 amino acids to conform with the shortest peptides in the Ensembl database and using ATG, GTG and TTG as start codons. This procedure discovered some DNA sequence ambiguities denoted most commonly by N (which can be any of the nucleotides), and less often D, H, R, K, Y, M and S (codes described at [30]) in the *M. kansasii* genome sequences. So, each triplet with one or more ambiguities was translated as X such that the specific tryptic peptides containing these ambiguous amino acids will be missed during the database matching but the remainder of the pepides from that protein could still be matched. New annotations were inferred from the translated sequences, showing the species ID, the frame in which it was read, and the start and end positions of the potential gene in the genome. For *M. avium* the Ensembl acces-

sion number was added at the end of this annotation for previously known protein sequences in order to differentiate them from the new ORFs. Where there were some differences in length between newly annotated sequences and Ensembl's, the longer sequence was used in each case in order to get as much information as possible from it. The resultant six frame translation ORF databases were then used for assignment of MS/MS data for proteogenomic annotations.

4.2.2 Protein Identification

Sample preparation

In general, the mycobacterial cultures were grown until mid to late log phase OD $0.6 - 0.9$ and were harvested by centrifugation. Cell lysis was carried out by boiling the cell pellet in 3% SDS buffer for 30min. Whole cell lysates were prepared from the strains *M. avium* and *M. kansasii* initially in Middlebrook 7H11 media, and thereafter in Sautons media. Clarified cell lysate was obtained by centrifugation at 10000g for 15 minutes to remove cell debris. Protein extracts from the culture filtrate and intracellular lysate were concentrated on 3kDa Molecular weight cut off filters. For each protein sample, 40μg was separated on a 12% SDS PAGE gel prior to further analysis. Culture filtrate and intracellular protein preparations were analysed separately.

Each gel lane was cut into 5 pieces resulting in 10 pieces per organism. All gel pieces were cut into smaller cubes and washed twice with water followed by 50% (v/v) acetonitrile for 10min. The acetonitrile was replaced with 50mM ammonium bicarbonate and incubated for 10min, and repeated two more times. All the gel pieces were then incubated in 100% acetonitrile until they turned white, after which the gel pieces were dried *in vacuo*. Proteins were reduced with 10mM DTT for 1h at 57°C.

This was followed by brief washing steps of ammonium bicarbonate followed by 50% acetonitrile before proteins were alkylated with 55mM iodoacetamide for 1h in the dark.

Following alkylation the gel pieces were washed with ammonium bicarbonate for 10 min followed by 50% acetonitrile for 20min, before being dried *in vacuo*. The gel pieces were digested with 100μl of a 10ng/μl trypsin solution at 37°C overnight. The resulting peptides were extracted twice with 70% acetonitrile in 0.1% formic acid for 30min, and then dried and stored at -20°C. Dried peptides were dissolved in 5% acetonitrile in 0.1% formic acid and 10μl injections were made for nano-LC chromatography.

All mycobacterial growth, protein preparation and tryptic peptide preparation steps were carried out by Julian Peters in this Laboratory.

Mass spectrometry

All experiments were performed on a Thermo Scientific EASY-nLC II connected to a LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Bremen, Germany) equipped with a nano-electrospray source. For liquid chromatography, separation was performed on an EASY-Column (2cm, ID 100μm, 5μm, C18) pre-column followed by a EASY-column (10cm, ID 75μm, 3μm, $C18$) column with a flow rate of 300nl/min. The gradient used was from $5-15$%B in 5min, $15-35$% B in 90min, $35-60$% B in 10min, $60-80$% B in 5min and kept at 80% B for 10min. Solvent A was 100% water in 0.1% formic acid, and solvent B was 100% acetonitrile in 0.1% formic acid. MS/MS data was acquired from the Orbitrap Velos in top 20 CID mode by Dr Salomie Smit (University of Stellenbosch).

Data analysis

Raw MS/MS data files were converted to MS2 format prior to analysis using database search algorithms. The CRUX [31] database search engine was used as the computational program to interpret the MS/MS spectra. CRUX basically utilizes similar algorithms as Sequest for the peptide matching scoring but has decreased peptide candidate retrieval time and better back-end statistics to estimate the false discovery rates. CRUX uses on-the-fly decoy databases to evaluate the statistical significance of a Peptide Spectral Match (PSM). The decoy database is generated by shuffling the target peptides ensuring that each decoy peptide has the same amino acid composition and total mass as the corresponding target peptide [31]. The False Discovery Rate (FDR) is then estimated based upon the decoy PSMs and a q-value is reported along with each PSM. The FDR is defined as the expected value of the number of false positive features over the number of all expected features. And the q-value is estimated by Equation 4.1 [32].

$$q(p_i) = \min_{t \geqslant p_i} \widehat{FDR}(t) \tag{4.1}$$

The FDR was set to 1% in our experience. Carbamidomethylation was chosen as a fixed modification whilst oxidation of methionine residues was set as a variable modification. Three missed cleavages were allowed and initial peptide mass tolerance was set at 10ppm whilst fragment mass tolerance was set to 0.5 Da.

4.3 Results

4.3.1 *Mycobacterium avium*

Mycobacterium avium is included in the *MAC* group which is the most commonly found group of non-tuberculosis mycobacterial in human samples [33]. *M. avium* is a well annotated manually curated (as distinct from automated annotation-only)" organism containing 5475491 nucleotides and 5120 predicted protein-coding genes where 5040 are non-redundant set and 143 are known as pseudogenes. Pseudogenes can be defined as genomic DNA sequences similar to normal genes but predicted to be non-functional; they are regarded as defunct descendant of functional genes [34]; in most cases the pseudogenes were annotated as such because of frameshifts in the genes. The predicted proteome of *M. avium* can be found at different sources such as Ensembl, Uniprot and Integr8 [28]. However, the proteomic studies of *M. avium* often result in many un-interpreted spectra when matching them against the *M. avium* proteome database. This suggests perhaps that some true proteins are not present in the predicted proteome database. Hence, the use of the six frame translation database may allows new PSMs and more inferred proteins as a result of potential new ORFs. The genome sequence of *M. avium* used for the six frame translation was obtained from the ENA (European Nucleotide Archive) database [35].

We were able to identify 1594 *M. avium* proteins in total from the MS/MS data searched against the six frame translation database. Comparing this result with that using the Ensembl proteome database of *M. avium*, 81 new proteins were identified from the MS/MS data using the six frame translation. To verify the significance of these proteins we did two analyses: (i) first looking at the position of the proteins in the *M. avium* chromosome (Coordinate search) and then (ii) searching them against

the Uniprot database (BLAST search).

Coordinate search

We made use of DAS (Distributed Annotation System) technology [36] in Ensembl for the coordinate search of the 81 newly identified proteins in the *M. avium* Ensembl chromosome and to locate the peptides that have been identified by the MS/MS spectra to predict these proteins. This was performed by creating a new DAS track for the MS peptides and the six frame translation which allows the visualisation of the proteins and peptides (that uniquely matched these proteins) in the *M. avium* chromosome in the Ensembl viewer according to their coordinates in the genome sequence, an example is shown in Figure 4.1. These peptides are unique to these protein

A)

>CP000479.1_3_76239_76730
MAGDTTITVVGNLTADPELRFTPSGAAVANFTVASTPRIYDRQSGEWKDGEALFLRCNIW
REAAENVAESLTRGSRVIVTGRLKQRSFETREGEKRTVVEVEVDEIGPSLRYATAKVNKA
SRSGGGGGGFGGGSRQQSAPASSAPADDPWGSASASGSFGWRR

B)

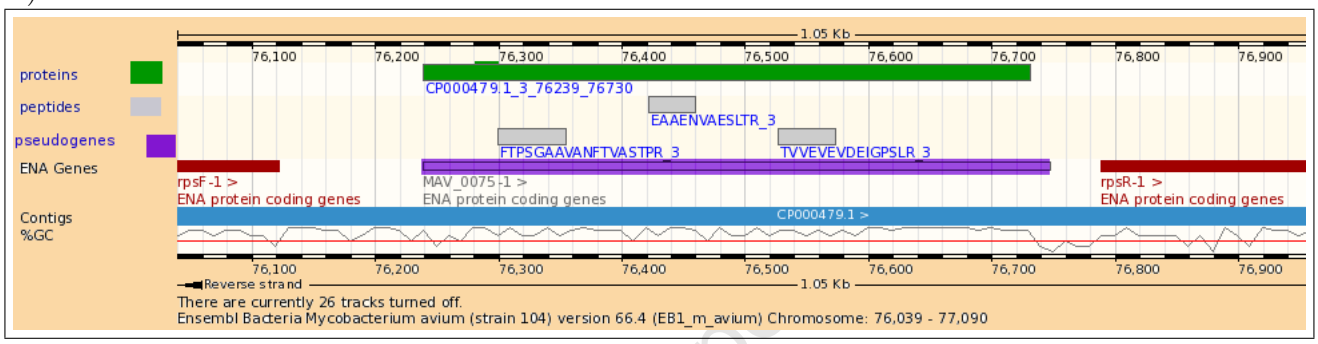

Figure 4.1: A) Sequence of an identified protein reported in Ensembl as a pseudogene, with peptides identified from the MS/MS analysis highlighted. B) The protein displayed on the Ensembl chromsome of M. avium with the identified peptides (grey) and protein predicted from the 6-frame translation (green).

From the coordinate search in the *M. avium* chromosome, we found that 34 of the 81 new proteins overlap with known pseudogenes in the *M. avium* chromosome. Figure 4.1 shows an example of an identified protein previously known as a pseudogene where 3 unique peptides have been identified from MS/MS to predict this protein. Given the number of observed peptides, there is therefore strong evidence that the current assignment of this chromosomal segment as a pseudogene is incorrect and is most likely due to DNA sequencing errors, although sequence differences between the *M. avium* isolates used here and in the genome sequencing cannot yet be excluded. Some of the newly identified non-pseudogenes lie on genes read in another strand as shown for example in Figure 4.2 where our prediction is a sequence on the forward

strand overlapping with an annotated protein on the reverse strand. This new protein has been identified by 3 peptides from the MS/MS analysis. Other newly identified proteins overlap with part of two genes as shown in Figure 4.3, where the newly identified protein and the coded proteins it overlaps are read in different frames. Interestingly, yet other proteins were found in blank regions of the chromosome not overlapping any existing proteins in *M. avium*, as illustrated in Figure 4.4.

A)

>CP000479.1_1_5024065_5024457
MALTEEDTAVNFNPSHSGPSGGAFRAPTPAAGPSGDAAPTERLTSIRQPGGPRVPSGPPANQA
GRTQRTRRTVDLPAATHRALDIWQREAADRLGVARVTGQEVLTALIDQLLVDPKLTAQITRA
IKERR

B)

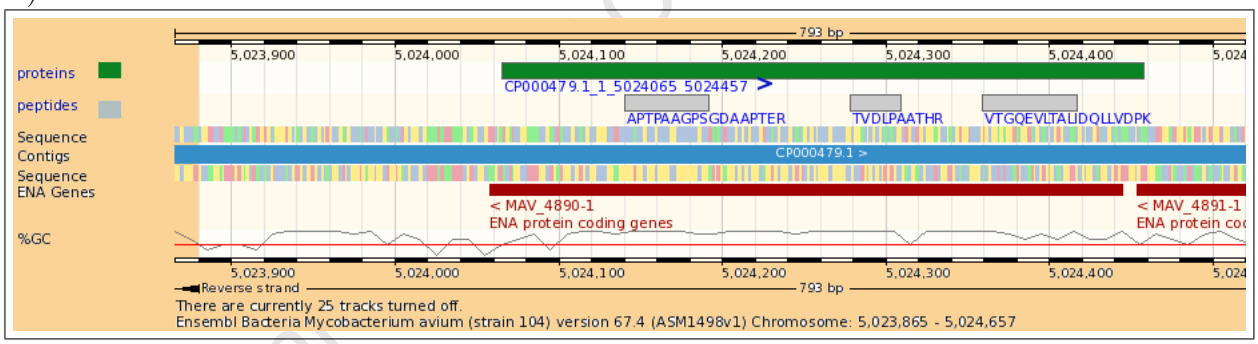

Figure 4.2: A) Sequence of an identified protein on the forward strand overlapping with a protein on the reverse strand with 3 peptides identified from the MS/MS spectra matching the former. B) The protein displayed on the Ensembl chromosome of *M. avium*.

A)

>CP000479.1_3_4375743_4376018
MRCAETTRTSVGTSNSASAAAASAITDQSLSLPMITATRGASVTLIDAAPVRVGPIRPSTARG
TASAPPARRAAAARPARPCRRRGSRRGS

B)

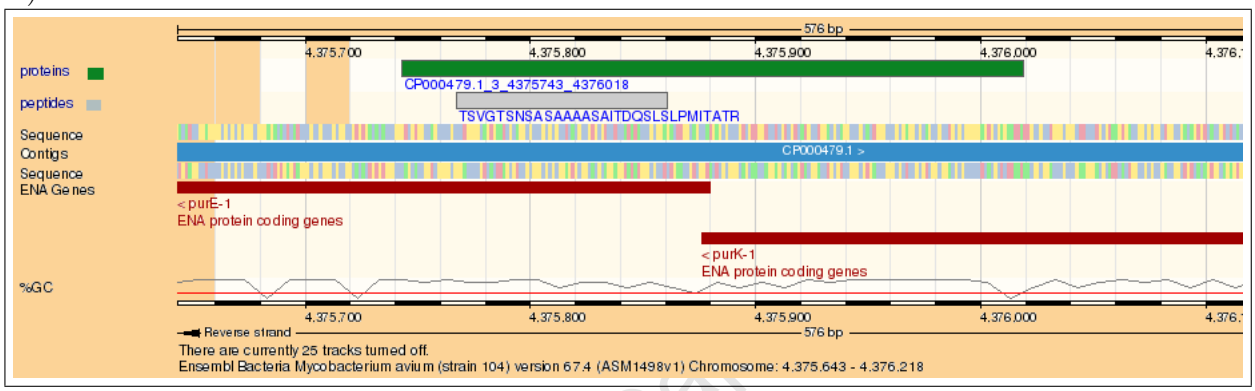

Figure 4.3: A) Sequence of an identified protein overlapping part of two proteins with one unique peptide identified from the MS/MS spectra. B) Identified protein displayed in the Ensembl chromosome of *M. avium*. The newly identified protein is not read in the same frame as any of the two proteins it overlaps.

A)

>CP000479.1_6_1853344_1853598
MEDICHECGFDQSRTPPRTVAEALPPVARAIGDGIRAISDDELRRRPTPAVWSLLEYVGHLRE
SMAFHRWLVVLTTDVRYGVVW

B)

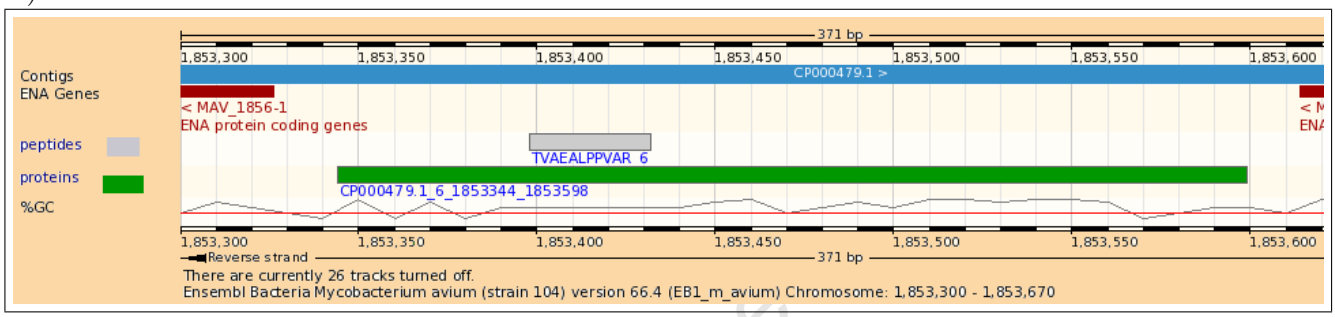

Figure 4.4: A) Sequence of an identified protein situated in a blank region in the chromosome with one unique peptide identified from the MS/MS spectra. B) Identified protein shown on the chromosome, lying between two known proteins.

BLAST search

BLAST (Basic Local Alignment Search Tool) [37] is a set of similarity search programs designed to run sequences against the available sequence databases. It can be used for DNA as well as protein sequences. We used a BLAST search here to retrieve the possible functions of the 81 newly predicted proteins from their similarity with other known proteins. The 81 proteins were searched against the whole Uniprot database from the FTP server using the blastp (BLAST for proteins) tool of the BLAST version 2.2.25+. Significant matches were determined based on a similarity of at least 50% and E-value of less than 10.

The BLAST search result revealed that 34 of the 81 identified proteins have sig-

nificant match to another species, with 24 matching characterized proteins and 10 matching putative uncharacterised proteins. Although the whole Uniprot database was used for the BLAST search, highest similarity matches were made with proteins from mycobacterial species which is not surprising looking at the phylogenetic relatedness of mycobacterial species as shown in the previous Chapter. From the 81 proteins, 39 did not have significant match (a significant match being considered as greater than or equal to 50% similarity) with any proteins from other organisms and the remaining 8 did not have any hit at all, so these may all be considered as potential novel proteins.

Combined result

The summary of the combined result from BLAST coordinate searches on the chromosome is illustrated in Figure 4.5.

Interestingly, 23 of the newly identified proteins that are currently annotated as pseudogenes in the *M. avium* chromosome hit characterized proteins in other species with significant scores. The list of these 23 proteins is given in Table 4.1 with their top BLAST hit.

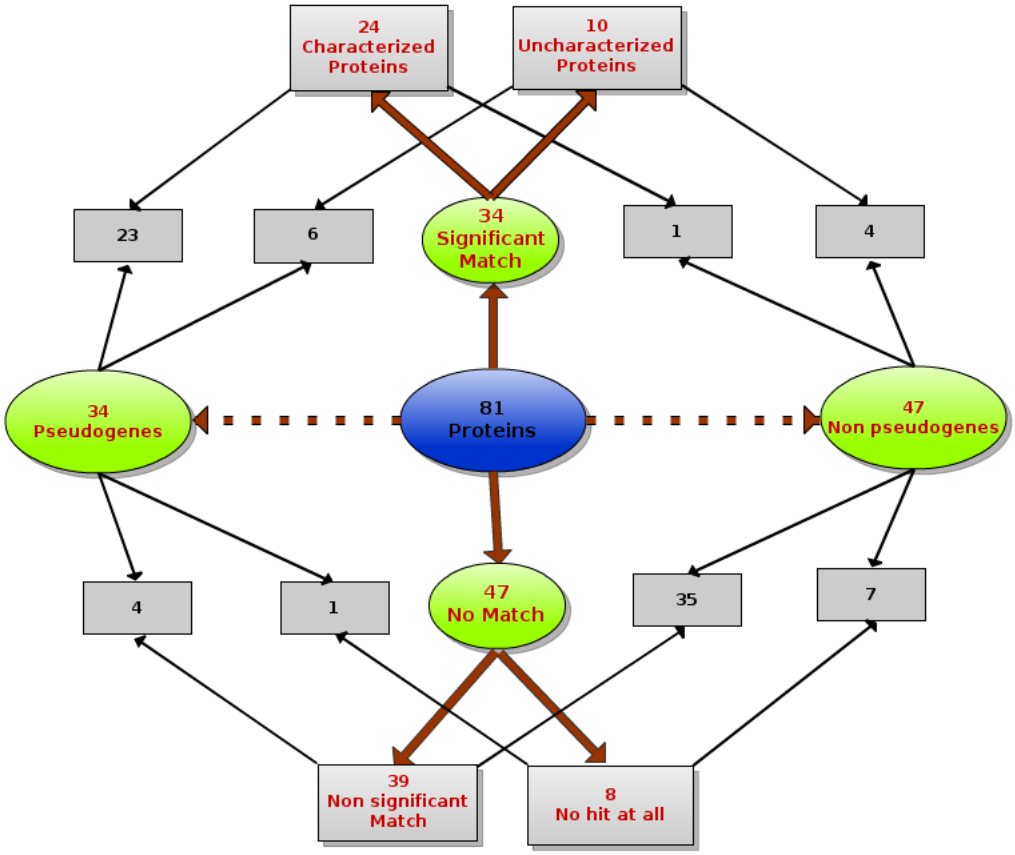

Figure 4.5: Summary of the result obtained from combining BLAST and coordinate searches. Of the 81 identified proteins, 34 overlap with known pseudogenes; 29 of these 34 match some known predicted protein in the mycobacterium databases, whilst 5 do not; these 5 are not pseudogenes because they are read in a different frame to the overlapping pseudogene. Of the remaining 47 identified proteins that do not overlap with pseudogenes, 42 do not show any significant matches to anything in the UniProt database; these are therefore interesting candidates for further study.

Table 4.1: List of proteins previously annotated as pseudogenes but which have significant matches with characterised proteins from other species and for which we have peptides evidence.

Protein ID	Top BLAST ID	Description
$CP000479.1\_1\_1464733\_1465263$	$Q73X40$:	Arginine-tRNA ligase
$CP000479.1\_1\_953548\_954963$	$Q742K6$:	Citrate synthase
$CP000479.1\_3\_1023762\_1024790$	$F7P3V7$:	Mg-chelatase subunit ChlI
$CP000479.1\_3\_76239\_76730$	$Q744V5$:	Single-stranded DNA-binding protein
$CP000479.1\_3\_1463610\_1464935$	$F7P506$:	Arginine-tRNA ligase
$CP000479.1\_3\_2424678\_2425244$	$F7P1Z1$:	Proteasome accessory factor PafA2
$CP000479.1\_4\_3868953\_3870401$	$Q73VQ8$:	AmiC
$CP000479.1\_6\_3968887\_3970191$	$Q73VI4$:	Glutamate-tRNA ligase
$CP000479.1\_4\_4330578\_4333205$	$F7P8Q8$:	Protein translocase subunit SecA 1
$CP000479.1\_6\_4350169\_4350969$	$F7P8P2$:	Phosphomannomutase
$CP000479.1\_2\_1024739\_1025155$	$Q9KII9$:	Rv0958-like protein
$CP000479.1\_5\_1344512\_1345201$	$A4KGH1$:	2-oxoglutarate dehydrogenase sucA
$CP000479.1\_5\_222098\_224686$	$P71486$:	Probable arabinosyltransferase B
$CP000479.1\_4\_4335282\_4336247$	$F7P8Q5$:	Sporulation/spore germination protein
$CP000479.1\_5\_2184665\_2186083$	$Q73YE8$:	AdhE2
$CP000479.1\_2\_1724666\_1725664$	$F7PCE7$:	Small-conductance mechanosensitive channe
$CP000479.1\_4\_964089\_964796$	$Q742J6$:	AccD3
$CP000479.1\_6\_3980053\_3980766$	$F7P9X3$:	Acetolactate synthase, large subunit
$CP000479.1\_5\_3978899\_3980185$	$Q73VH5$:	Acetolactate synthase
$CP000479.1\_2\_2423735\_2425210$	$F7P1Z1$:	Proteasome accessory factor PafA2
$CP000479.1\_6\_1344619\_1348272$	$Q73WX4$:	2-oxoglutarate decarboxylase
$CP000479.1\_3\_2802039\_2802992$	$F7PEJ9$:	Transcriptional regulator
$CP000479.1\_4\_844704\_845249$	$Q1B5H3$:	Transcriptional regulator, TetR family

Figure 4.1 shows an example of one such newly identified protein having a significant match with proteins in other species that are involved in the "single stranded DNA binding" pathway, yet it is currently annotated as a pseudogene in *M. avium*. From the 81 newly identified proteins, 5 overlap known pseudogenes in the *M. avium* chromosome but are not read in the same reading frame and they do not have any significant match in any other organisms. An example of such a protein is illustrated in Figure 4.6 which is read in the reverse strand while the pseudogene is in the forward strand.

A)

>CP000479.1_4_5339505_5340095
MASLWARRSASSSVHRAGTVIGRAASASHSADITEPMSPASGAAICWLLSNSDGSMSTWTN
LTPADHCGDSPCPRSQFSRAPTNSTASARPTASERAAATDCGCASGSSPLAIDIGRNGIPVHST
NRRISLSAWAYAAPLPSTINGRRALVSTSSARSSASGAGSWRGAGSTTRHSVPAAAEASMA
WPRTSPGMSR

B)

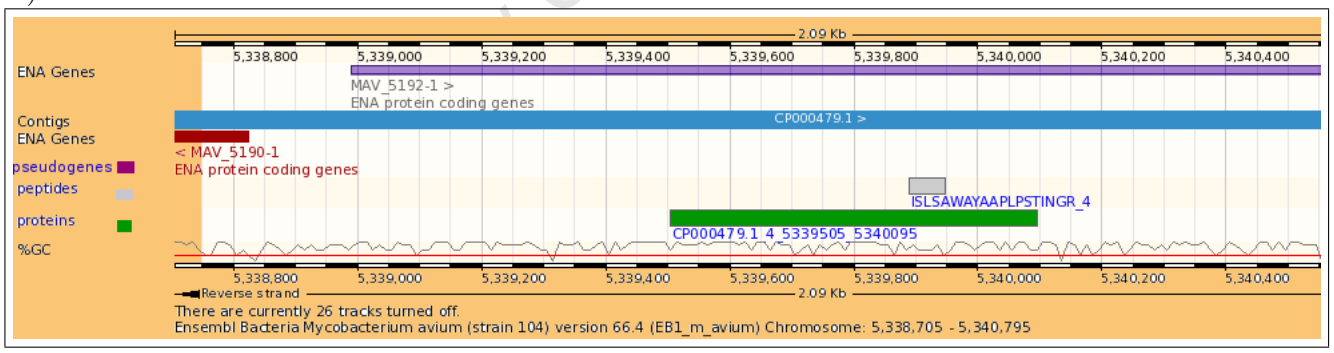

Figure 4.6: A) Sequence of an identified protein on the reverse strand overlapping with a pseudogene on the forward strand. It does not have any significant match to proteins from other species. B) Location of the protein on the *M. avium* chromosome.

Figure 4.5 shows that almost all of our newly identified proteins that match characterized proteins in a BLAST search (23 out of 24) are currently annotated

as pseudogenes in *M. avium*. Only one non-pseudogene overlapping a part of an existing gene in the chromosome matched a characterized protein, which is Q83YL6 in *Mycobacterium fortuitum*, functionally annotated as a 'Putative transposase'. The coordinate of this identified proteins on the *M. avium* chromosome is shown in Figure 4.7.

A)

>CP000479.1_4_1443957_1444451
MGVVRSHRPQSSPCRRRPGRRCPRGRSWSHPAPQNRQHPSPAGLPSTPTHLAPTNALALGR
SLDHVVVQHHRPQSTNNRDYLTTSPNRPNRSTQEKLDRPATTSCEHAAPPTEITVEPHQPG
QSTDRGLAAAVCRAMRLLRSVQPPCNSHPIGTRQNLTDGNTV

B)

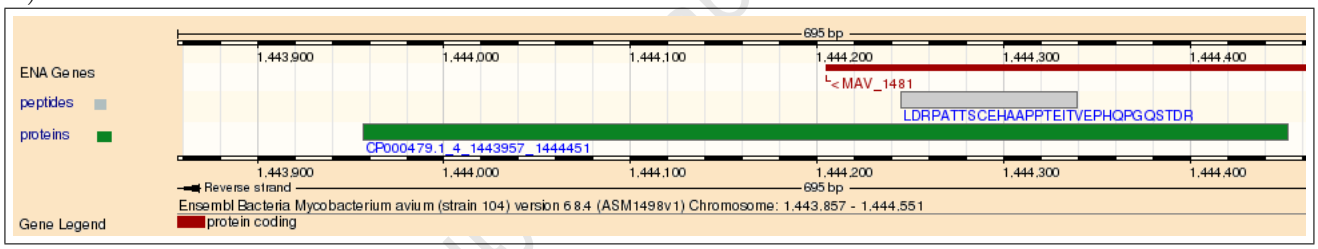

Figure 4.7: A) Sequence of the only identified protein overlapping with a non-pseudogene which hits a characterized protein in other species. B) Location of the protein on the *M. avium* chromosome; it is read in a different frame to the overlapping coded protein.

From the identified proteins overlapping non-pseudogenes, 4 matched hypothetical (i.e uncharacterised) proteins in other organisms and one of these is shown in Figure 4.4. Notably, of the 47 proteins overlapping non pseudogenes, 35 have no significant match (significance being greater than or equal to 50% similarity), (Figure 4.3 is one such example) and 7 proteins of this set did not have any hit in the BLAST search at all. An example of these 7 proteins is shown in Figure 4.8 which represents a very

short protein, with length 26 containing only 2 tryptic peptides of length 3 and 23 respectively but where the longer peptide has been identified by the MS/MS analysis. Only 5 non-pseudogenes overlapping proteins are present in other species of which 4 of them match hypothetical uncharacterised proteins.

A)

>CP000479.1_5_1333265_1333345
MAKRPPNCLVQSWAIAANVAPSVAVV

B)

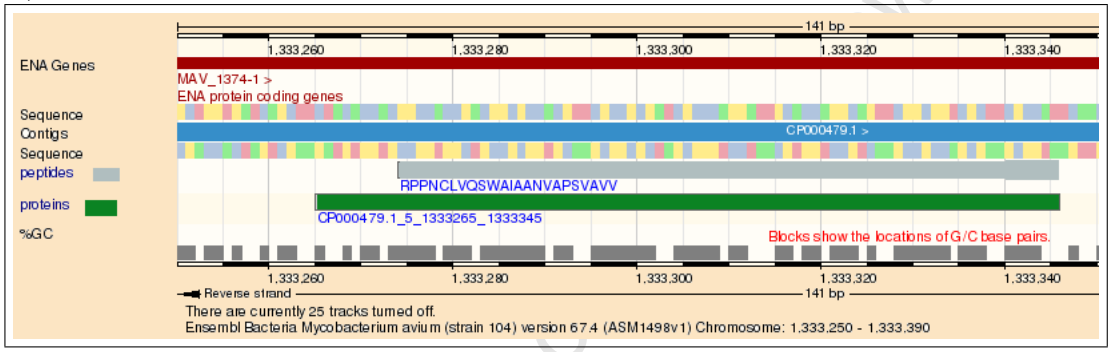

Figure 4.8:) Sequence of an identified short protein with no hit from other species. B) Location of this protein on the *M. avium* chromosome; it is read in different frame to the overlapping coded protein.

Overall, we found that most of the 81 extra proteins identified experimentally from the *M. avium* sample have evidence and significance to be real based on similarity to proteins annotated in other organisms. Although *M. avium* is thought to be well annotated, the result obtained here reveals that it is in fact incomplete and some ORFs previously known as pseudogenes are real. Possible functions of these proteins were assigned from their significant match with other known characterized proteins. Out of the 81 newly identified proteins here, 47 likely represent potential novel proteins because they have no significant match to any known protein in any

organism; these therefore warrant further study.

4.3.2 *Mycobacterium kansasii*

After *M. avium* complex organisms, *M. kansasii* infection is the second most common non-tuberculous opportunistic mycobacterial infection associated with AIDS. Unlike *M. avium*, *M. kansasii* has been poorly annotated as only 83 proteins are available in Uniprot corresponding to this organism. This species was discovered in 1953 and was sequenced in 2008 by Veyrier F. et.al [38] but the 5913 annotated proteins in Refseq from the Whole Genome Sequencing(WGS) [39] were obtained from automated annotation and interestingly have not been included yet in the Uniprot database. Other sources like Ensembl or ENA contain only the genome sequence of this species.

Using the six frame translation database of *M. kansasii* to search our MS/MS data, we identified only 160 proteins. Interestingly, we found a much higher number (660) of proteins by searching the MS/MS data against the *M. avium* six frame translation file. This observation caused us to question both the reliability of the genome sequence for *M. kansasii*, as well as the identity of the *M. kansasii* clinical isolate that gave rise to the experimental MS/MS data.

Comparison of the 16S rRNA sequence for the *M. kansasii* whole genome sequence to other *M. kansasii* 16S rRNA sequences obtained from the myRDP database [22] showed very high homology. Furthermore, searching the 83 *M. kansasii* proteins present in Uniprot against the automated translation of the WGS data showed high homology matches in all cases. Together, these lines of evidence appear to verify the WGS *M. kansasii* data.

At the present time, 16S rRNA sequencing of our *M. kansasii* isolate suggests

that in fact our isolate may not be *M. kansasii*. However, further work is needed to unravel this situation because the closest 16S rRNA match is to Peanibacillus species and not to any mycobacteria, yet a search of the MS/MS data against the whole Uniprot bacteria database returned more hits to *M. avium* than any other organism by some considerable margin.

It is an interesting twist that the proteogenomic analysis of the MS/MS data has demonstrated that our *M. kansasii* strain is in fact not M. kansasii; the discordant results of the 16s rRNA sequencing and the MS/MS data suggest that further experimental identification of this clinical isolate is warranted. Further discussion of the MS/MS data from our *M. kansasii* isolate is therefore beyond the scope of this thesis.

5. Conclusion

Summary

Two approaches have been described in this thesis to enable database searches of proteomic data for incomplete or non-annotated organisms. The first one consists of the use of the closest annotated species database and the second one is a proteogenomic analysis using an automated six frame translation database generated from the genomic data. The survey of this work and a future proposed workflow can be summarized in the Figure 5.1. For any species in mycobacteria, if it is annotated with a complete database, a normal database search method can be done using the existing database. If the species is not annotated, the feasibility of a cross-species analysis is tested by first computing the distance of the species with its closest sequenced organism and then by estimating the percentage of identical peptides between them. The cross-species analysis is feasible if the estimated number of identical peptides is high enough with an acceptable false positive rate. In the case that the cross-species analysis becomes non-significant for non-annotated species or if the species has been annotated but incompletely, a proteogenomic analysis can then be applied.

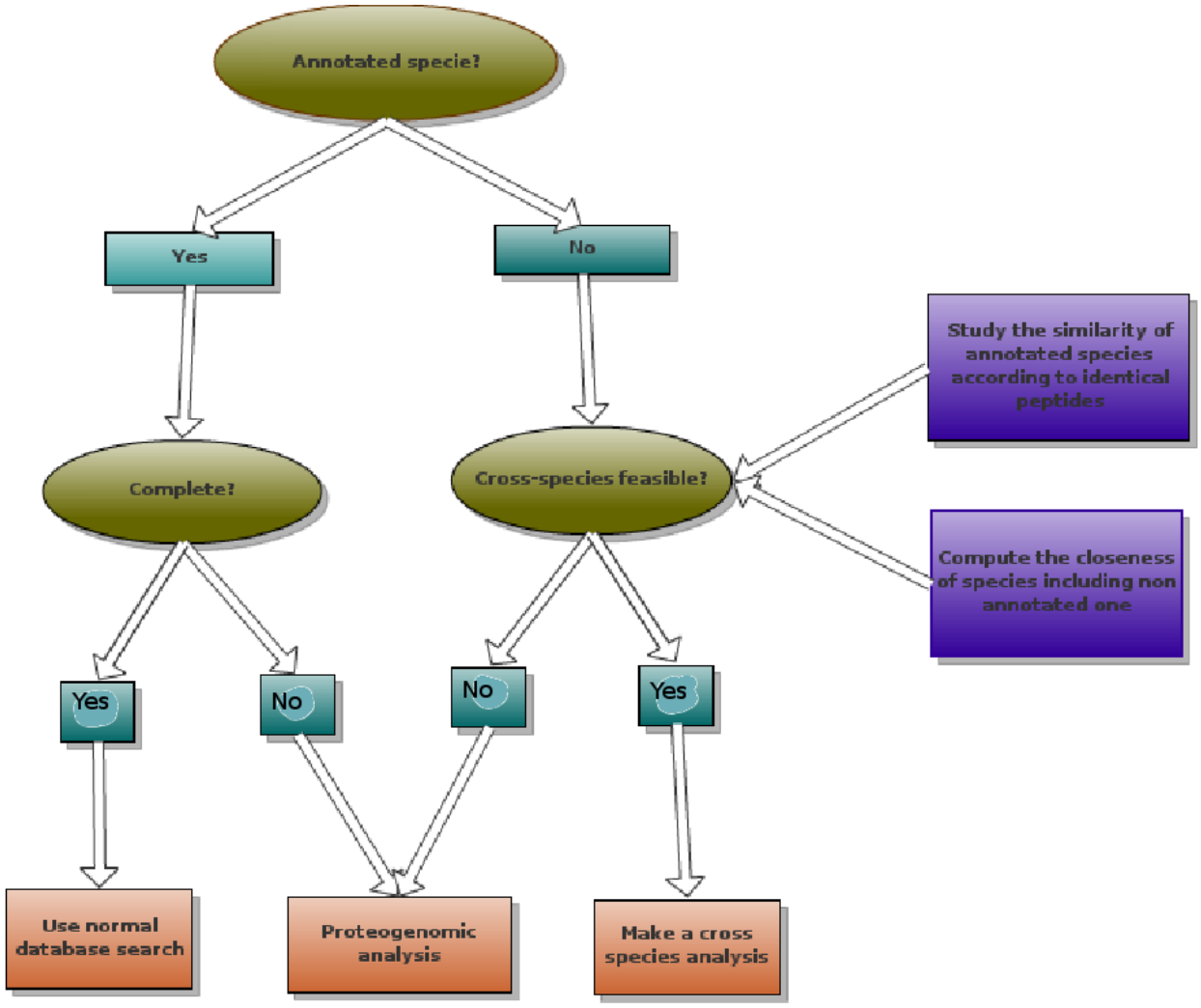

Figure 5.1: Workflow of this study, which can potentially be used for future analyses.

The study of the comparison of identical peptides between mycobacterial species revealed the strong conservation of 2600 peptides across all mycobacterial species representing around 700 proteins in each organism including about the half of the *M. leprae* proteome. We have shown that the cross species analysis is applicable for non-annotated organisms if its phylogenetic distance with a sequenced and annotated organism is small enough, such as those within MTBC or MAC. This approach

stops being viable when the phylogenetic distance to the nearest annotated organism increases, particularly beyond 0.3; in such circumstances, the use of proteogenomic analysis becomes the preferable approach.

The proteogenomic analysis of *M. avium* revealed the incompleteness of its database even though *M. avium* is known as a well annotated species among mycobacteria. We have shown that 81 proteins not included in the existing *M. avium* database have experimental evidence from the MS/MS spectra; 34 of these proteins were subsequently found to have significant match with proteins in other species that are involved in important pathways and 47 did not and so are likely to be novel proteins. There were 34 proteins previously annotated as pseudogenes in the *M. avium* chromosome that are shown here to have experimental evidence.

Future work

Our study here highlighted the importance of a cross-species analysis for mycobacterial species with phylogenetic distance less than 0.3 to each other. If we have non-annotated mycobacterial species having phylogenetic distance less than 0.3 with a well annotated species, then its closest species database can be used for an MS analysis. Our study was limited to mycobacterial species, all analysis were made using these species, but it can obviously be extended to and tested with other species outside the mycobacteria. Our future plan therefore consists of extending the cross-species analysis to other organisms to see if the distance limit remains the same and if we can generalize this pattern.

Acknowledgements

I would like to thank the following persons for the realisation of this project:

- My supervisor Professor Jonathan Blackburn ,
- My co-supervisor Professor Nicola Mulder,
- Putuma Gqamana,
- Julian Peters,
- Gustavo Adolfo Salazar Orejuela and
- All the Blackburn lab and Computational biology group members

References

[1] http://microbewiki.kenyon.edu/index.php/Mycobacterium.

[2] World Health Organization (WHO) Report. *Global tuberculosis control*. Geneva, 2011.

[3] http://www.healthieryou.com/tb.html.

[4] Alimuddin IZ Stephen DL. Tuberculosis. *The lancet*, 378:57–72, 2011.

[5] Jimmy KE, Ashley LM, and III John RY. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American society for mass spectrometry*, 5:976–989, 1994.

[6] http://cbio.ufs.ac.za/mascot.

[7] Robertson C and Ronald CB. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass spectrometry*, 17:2310–2316, 2003.

[8] Ari F and Pavel P. PepNovo: *De novo* Peptide Sequencing via Probabilistic Network. *Analytical Chemistry*, 77:964–973, 2005.

[9] Bin M, Kaizhong Z, Christopher H, Chengzhi L, Ming L, Amanda D, and Gilles L. Peaks: Powerful Software for Peptide *de novo* Sequencing by ms/ms. *Rapid Communication in Mass Spectrometry*, 17:2337–2342, 2003.

[10] http://en.wikipedia.org/wiki/file:MS MS.png.

[11] John DH and Paul KS. Gel electrophoresis of proteins and nucleic acids:II-Techniques and applications. *British Medical Journal*, 1989.

[12] Anthony TA. *Global tuberculosis control*. Oxford [Oxfordshire] : Clarendon Press ; New York : Oxford University Press, 1986.

[13] Reiner Westermeier. *Electrophoresis in Practice: A Guide to Methods and Applications of DNA and Protein Separations*. VCH press; Weinheim, 1997.

[14] http://en.wikipedia.org/wiki/Chromatography.

[15] http://en.wikipedia.org/wiki/Gel electrophoresis.

[16] Vlado DC, Theresa AA, Karl RC, James EV, and Pavel AP. *De novo* Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology*, 6:327–342, 1999.

[17] http://bacteria.ensembl.org/info/data/ftp/index.html.

[18] Lennart Martens, Jol Vandekerckhove, and Kris Gevaert. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics*, 21:3584–3585, 2005.

[19] Thompson JD, Higgins DG, and Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22:4673–4680, 1994.

[20] J. Michael Janda and Sharon L. Abbott. 16s rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of clinical Microbiology*, 45:2761–2764, 2007.

[21] Clarridge. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical microbiology reviews*, 17:840–862, 2004.

[22] http://rdp.cme.msu.edu.

[23] Jukes TH and Cantor CR. *Evolution of Protein Molecule.* in H.N Munro editor; Mammalian Protein Metabolism pp. 21-132; Academic Press; New York, 1969.

[24] Kevin H, Alex B, and Richard D. Quicktree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, 18:1546–1547, 2002.

[25] http://mobyle.pasteur.fr/cgi-bin/portal.py.

[26] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck, Kauff F, Wilczynski B, and de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422–1423, 2009.

[27] http://grissom.gr/stranger/.

[28] http://www.ebi.ac.uk/integr8/.

[29] http://www.uniprot.org/.

[30] http://www.dna.affrc.go.jp/misc/MPsrch/InfoIUPAC.html.

[31] Christopher YP, Aaron AK, Lukas K, Michael JM, and William SN. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *Journal of proteome research*, 7:3022–3023, 2008.

[32] John DS and Robert T. Statistical significance for genomewide studies. *Proceeding of the National Academy of Sciences of the United States of America*, 100:9440–9445, 2003.

[33] Wickremasinghe M, Ozerovitch LJ, Davies G, Wodehouse T, Chadwick MV, Abdallah S, Shah P, and Wilson R. Non-tuberculous mycobacteria in patients with bronchiectasis. *Thorax*, 60:1045–1051, 2005.

[34] http://www.pseudogene.org/background.php.

[35] http://www.ebi.ac.uk/ena/.

[36] http://www.ensembl.org/info/data/ensembl das.html.

[37] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic Alignment Search Tool. *Journal of Microbiology*, 215:403–410, 1990.

[38] http://www.ncbi.nlm.nih.gov/nuccore/NZCM000636.3.

[39] http://www.ncbi.nlm.nih.gov/genbank/wgs.

[40] Abigail Manson McGuir, Brian Weiner, Sang Tae Park, Ilan Wapinski, Sahadevan Raman, Gregory Dolganov, Matthew Peterson, Robert Riley, Jeremy Zucker, Thomas Abeel, Jared White, Peter Sisk, Christian Stolte, Mike Koehrsen, Robert T Yamamoto, Milena Iacobelli-Martinez, Matthew J Kidd, Andreia M Maer, Gary K Schoolnik, Aviv Regev, and James Galagan. Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of *mycobacterium tuberculosis* pathogenesis. *BMC Genomics*, 13:1471–2164, 2012.

[41] http://amigo.geneontology.org/cgi-bin/amigo/slimmer?session id=.