Journal of Bacteriology and Parasitology

Akinola et al., J Bacteriol Parasitol 2013, 4:4
DOI: 10.4172/2155-9597.1000173

Research Article Open Access

A Systems Level Comparison of *Mycobacterium tuberculosis*, *Mycobacterium leprae* and *Mycobacterium smegmatis* Based on Functional Interaction Network Analysis

Richard O Akinola, Gaston K Mazandu and Nicola J Mulder*

Computational Biology Group/Institute of Infectious Disease and Molecular Medicine, Department of Clinical Laboratory Sciences, Faculty of Health Sciences, University of Cape Town, South Africa

Abstract

Mycobacterium leprae is a pathogenic bacteria that causes leprosy, a disease which affects mainly the skin, peripherical nerves, eyes and mucosa of the upper respiratory tract. Despite significant progress recorded in the last few years to stop this disease through a Multi-Drug Therapy (MDT) strategy, every year there are new reported cases of the disease. According to the World Health Organization, there were 192,242 new cases at the beginning of 2011. *Mycobacterium leprae* cannot be cultured in the laboratory but can be grown in mouse foot pads and more recently in nine banded armadillos because of its susceptibility to leprosy. Its highly reduced genome makes it an interesting species as a model for reductive evolution within the mycobacterial genus; it shares the same ancestor with *Mycobacterium tuberculosis* (MTB). A functional network for MTB was generated previously and extensive computational analyses were conducted to reveal the biological organization of the organism on the basis of the network's topological properties. Here, we use genomic sequences and functional data from public databases to build protein functional networks for another slow grower, *Mycobacterium leprae* (MLP) and the fast growing non-pathogenic *Mycobacterium smegmatis* (MSM). Together with the MTB network, this provides an opportunity for comparison of three mycobacteria with different sized genomes. In this paper, we use network centrality measures to systematically compare MTB, MLP and MSM to quantify differences between these organisms at the systems biology level and to study network biology and evolution.

Keywords: Biological networks; Mycobacterial pathogens; *Mycobacterium tuberculosis*; *Mycobacterium leprae*; *Mycobacterium smegmatis*; Rewiring rate

Introduction

Leprosy is a chronic dermatological [1] and malignant human neurological disease [2]. It is caused by the pathogen *Mycobacterium leprae* (MLP) which has similar characteristics to *Mycobacterium tuberculosis* (MTB). Various attempts to culture MLP have failed because, of all known bacteria, it has the longest doubling time [2]. MLP is an acid-fast, rod shaped bacillus [3] and has a doubling time of ~14 days [2] and is host specific. It has been grown in mouse foot pads and more recently in nine banded armadillos because of their susceptibility to leprosy. According to the WHO [4], leprosy has been classified into two types based on how they smear the skin, paucibacillary (PB) and multibacillary (MB). Leprosy affects mainly the skin, peripheral nerves, the eyes and mucosa of the upper respiratory tract [3]. However, through Multidrug Therapy, there has been a reduction in the number of reported cases of the disease, from the 228 474 new cases in 2010 to 192 246 cases at the beginning of 2011 [5]. Its highly reduced genome makes it an interesting species as a model for reductive evolution within a genus.

According to Monot et al. [1], there are seven strains of *Mycobacterium leprae* that are well characterised, namely: India2, Thai53, TN, Africa, NHDP63, NHDP98 and Br4923. The first three strains are of SNP type 1, the fourth of SNP type 2, the fifth and sixth i.e., NHDP63 and NHDP98 are of SNP type 3, while the last one is of SNP 4 [1]. The MLP strain TN is from Tamil Nadu, the Thai53 strain is from Thailand, the NHDP strains are from the United States and the Br4923 strain is from Brazil. So far, complete genome sequences have been obtained for TN and Br4923. In this article, we use the MLP strain TN.

Tuberculosis is caused by MTB and is one of the 'most dangerous' infectious diseases [6]; it claimed about 1.8 million victims in 2008 and there were estimates of 9.4 million new cases that year (3.6 million of whom are women), including 1.4 million cases among people living with Human Immunodeficiency Virus (HIV) or Acquired Immunodeficiency Syndrome (AIDS) according to the World Health Organization (WHO). The MTB bacillus is slow growing and has a complex cell wall.

Mycobacterium smegmatis (MSM) is an aerobic, fast growing, non-pathogenic mycobacterium which has many common features with pathogenic mycobacteria [7]. It has the potential to adapt to microaerobiosis by changing from active growth to dormant or latent states. It can be dormant in conditions of low oxygen concentrations and can survive for more than 650 days in the absence of carbon, nitrogen and phosphorus. MSM is particularly useful in understanding the cellular processes that are important to pathogenic mycobacteria like MLP, MTB and *M. avium* subsp. *paratuberculosis* [8]. This is one of the major reasons why we are including this mycobacterium in the present study. For the purpose of our comparisons, we will use the MC^2 155 strain of MSM.

The genome sizes of MLP, MTB and MSM are 3,268,203; 4,411,532

***Corresponding author:** Nicola J Mulder, Computational Biology Group/Institute of Infectious Disease and Molecular Medicine, Department of Clinical Laboratory Sciences, Faculty of Health Sciences, University of Cape Town, Anzio Road, 7925 Observatory, South Africa, Tel: +27 21 406 6058; E-mail: nicola.mulder@uct.ac.za

Received July 01, 2013; **Accepted** August 06, 2013; **Published** August 12, 2013

Citation: Akinola RO, Mazandu GK, Mulder NJ (2013) A Systems Level Comparison of *Mycobacterium tuberculosis, Mycobacterium leprae* and *Mycobacterium smegmatis* Based on Functional Interaction Network Analysis. J Bacteriol Parasitol 4: 173. doi:10.4172/2155-9597.1000173

and 6,988,209 base pairs, respectively (see, for example [2] and http://mycobrowser.epfl.ch/smegmalist.html, accessed on 15 September, 2012). This implies that the genome of MLP is approximately 1.4 Mb smaller than MTB and less than half the size of MSM. In addition, the G+C content of MLP is 57% which is lower than other mycobacterial genomes. However, as pointed out in Eiglmeier et al. [9], genomes of organisms that have suffered reductive evolution are usually richer in A+T content. MSM has a doubling period of three to four hours in culture and forms colonies in 3-4 days, MTB has a doubling time of twenty to twenty-four hours and forms a colony in an agar medium in three to four weeks [10], while, as stated earlier, MLP doubles in approximately fourteen days. Therefore, these three organisms represent three different genome sizes as well as different growth rates within one genus.

Although MTB and MLP share a common ancestor, MLP is an obligate intracellular parasite while MTB is a facultative intracellular parasite [11]. Youm and Saier [11] compared the clinical CDC1551 strain of MTB to the TN strain of MLP. The genome of the MTB strain encodes 4189 proteins and the MLP strain 1605 proteins. This reduction in the MLP is attributed to reductive evolution with many genes having become pseudogenes. They defined a pseudogene as an inactivated gene that no longer produces functional proteins. In comparison to other mycobacterial species, two main consequences were proposed for the reduction in the genome of MLP [12]: the presence of few proteins belonging to the PE and PPE functional category and traces belonging to insertion sequences and bacteriophages. As shown in Table S1, the number of proteins in the MTB genome belonging to the PE and PPE family is roughly fifteen times that of MLP, and while 82 proteins in MTB are insertion sequences or derived from bacteriophages there are only two in MLP. In addition, the presence of pseudogenes in MLP and the corresponding absence thereof in MTB accounts for some of the phenotypic differences between the two pathogens.

Gómez-Valero et al. [13] defined reductive evolution as the process by which genes and their corresponding functions are lost, resulting in the downsizing of the genome. Three reasons based on changes in lifestyle were given why an organism may have reductive evolution: a desire to 'move' from a free living to a host-associated or intracellular life, when the organism restricts itself from multiple to specific hosts and from multiple to specific host tissues. By analyzing the distribution of gene-loss along the ancestral genome, it was shown that the genome downsizing in MLP was as a result of gene by gene inactivation and not inactivation in blocks or large chunks, before a gradual nucleotide loss [13]. In addition, they classified ancestral genes in the MLP genome into three categories: retained, absent/deleted and pseudogenized. Genes belonging to the 'absent' category have either diverged so much that they cannot be recognized or were totally deleted, while those in the pseudogenized category have sufficient levels of nucleotide similarity with MTB. It was also reported that 1537 genes have been lost from the ancestor to MLP, of which, 1129 are pseudogenes. In this work, we use functional genomic data from public databases to generate functional interaction networks for slow growers: MLP and MTB, and the fast growing non-pathogenic MSM. We used network centrality measures to make the following comparisons: MTB versus MLP, MTB versus MSM and MLP versus MSM in order to determine the impact of genome size on network evolution.

Materials and Methods

We downloaded datasets containing Uniprot protein accession numbers (ids) and gene names from UniProt Consortium [14] http://www.uniprot.org, accessed on 29 June, 2012) for the three mycobacterial organisms: *Mycobacterium leprae*, *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. These protein accession numbers were then used to extract protein protein interactions data from STRING [15,16]. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a database containing predicted and known Protein-Protein Interactions (PPI). These functional protein-protein associations are derived from conserved genomic neighbourhood, gene fusion, imports from database (knowledge), phylogenetic co-occurrence, high-throughput experiments and text mining. STRING web resources and databases can be accessed from http://string-db.org/(accessed on 30 June, 2012). All the data from these different sources are integrated into a single network, before computing the combined confidence score for all PPI's.

In addition, we derived other interactions from sequence similarity and signatures (shared domains), microarray data (co-expression), Protein Data Bank (PDB) [17,18] and MINT [19], DIP [20] and Intact http://www.ebi.ac.uk/intact/(accessed on 6 October 2012) data. PPI data from MINT, DIP and Intact were used to predict interologs in MLP and MSM based on the premise that orthologs of interacting proteins should themselves interact. Ortholog data were downloaded from Biomart (http://www.ebi.ac.uk/uniprot/biomart/, accessed on 12 August 2012). DOMINE is a database containing known and predicted protein domain interactions. The Domain-Domain Interactions (DDI) are inferred from Protein Data Bank (PDB) entries and those interactions from PFAM domain definitions predicted by thirteen different methodologies. We extracted DDI's with PFAM ids from the DOMINE website (http://domine.utdallas.edu/cgi-bin/Domine, accessed on 17 October, 2012), neglecting self interactions to avoid loops. With the aid of the data containing PFAM ids and their corresponding InterPro ids, we converted those interactions from DDI into their interPro equivalents, before changing them to Uniprot-Uniprot protein interaction ids. InterPro data was downloaded from the interPro website for both MLP and MSM. We assigned a uniform score of 0.85 for all these interactions. We also used an information-theory based technique proposed by Mazandu and Mulder [21] to derive PPI's from protein sequence similarity and signatures as well as shared domains. In line with Mazandu et al. [22], the microarray data for MTB were downloaded from the Standford Microarray Database (SMD), at http:smd.stanford.edu/ (accessed on 28 October, 2011) and NCBI Gene Expression Omnibus (GEO) (see, http://www.ncbi.nlm.nih.gov/geo/ (accessed on 12 September, 2012).

Out of the seven experiments used in analyzing the microarray data for the MTB network, two with file ids 15569 and 15575 were downloaded from SMD and the remaining five; GSM219305, GSM219324, GSM219694, GSM219695, GSM219696 from GEO [22]. For MSM, we retrieved 23 experiments from the Array Express database: four with GEO accessions: GSM743320, GSM743321, GSM743322, and GSM743323. The remaining 19 ids are GSM748761, GSM748762, GSM748763, to GSM748779. We then employed a random partial least squares regression approach described by Mazandu et al. [22] to generate functional association scores between pairs of interacting proteins. However, for MLP, we downloaded only four experiments contained in the GSE17191 series matrix from GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc .cgi?acc=GSE17191, accessed on 12 August, 2012). This limited number of microarray experiments prevented us from using the same technique used for MSM and MTB so we decided to calculate the correlation instead. The Pearson correlation coefficient was calculated to find co-expressed genes and we inferred interactions between genes for which the correlation coefficient was exactly one.

After calculating the confidence score for each functional association protein pair, we computed the combined confidence score $C_{(p,q)}$ for interacting proteins *p* and *q* using the formula [15]

$$C_{(p,q)} = 1 - \prod_{s=1}^{n} (1 - c^{s}_{(p,q)}), \quad (1)$$

where *n* is the total number of PPI data sources and $c^{s}_{(p,q)}$ is the confidence score of a functional association between *p* and *q* predicted using the type of data source *s*. In all the three networks, *n*=11. Next, we define some network centrality measures that we used to characterize the proteins in each of the networks and in determining the most central protein.

Let $P = \{p_1, p_2, p_3, ..., p_m\}$ be an ordered set of all proteins and $Q = \{(p_1, p_2), (p_3, p_4), ..., (p_{m-1}, p_m)\}$ be the set of all interacting proteins in the network *G*. It is conventional to define a network *G* as a graph *G*=(*P*, *Q*) [23]. Then, we define the entries b_{ij} for *i*, *j*=1, 2, 3, · · ·,*m* of the [24] *m* by *m* adjacency matrix of *G* as

$$b_{(p_i,p_j)} = \begin{cases} 1 & \text{if } (p_i, p_j) \text{ are functional interactions in G} \\ 0 & \text{Otherwise.} \end{cases}$$

We assume that the networks under consideration are simple, meaning there are no loops and no multiple functional protein-protein interactions. Hence, $b_{ii=}0$. Most importantly, the adjacency matrix is symmetric for undirected networks. Next, let $d(p_1, p_2)$ be the distance between protein p_1 and p_2. For all possible paths in a network *G* from p_1 to p_2, $d(p_1, p_2)$ is the length of the shortest path from p_1 to p_2. If there is no path between any two proteins, then the distance between them is infinite. The average shortest path length of a network is defined as [25].

$$c = \sum_{p_i, p_j \in P} \frac{d(p_i, p_j)}{m(m-1)}, \quad (2)$$

where $d(p_1, p_2)$ is the shortest path length from p_1 to p_2. The density of a network is the ratio of the number of edges to the number of possible edges in the network [25].

When two proteins p_1 and p_2 are functionally linked together by an edge, we say they are adjacent to each other. Several approaches were used to define the degree centrality of a protein p_i. However, Nieminen [26,27], gave a mathematical formula for computing it as:

$$D(p_i) = \sum_{i=1}^{m} u(p_i, p_j), \qquad \text{for } i=1,2,3,\cdots m,$$

where $u(p_i, p_j)$ is the Kronecker-Delta function. $D(p_i)$ is large if p_i is functionally connected to other proteins in the network, meaning that such a protein partakes in a high number of biological interactions in the organism. If $D(p_i)$=0, then it means p_i does not interact with other proteins and it 'may' not be an important protein neccessary for the survival of the organism. From the above formula, it means that the maximum $D(p_i)$ is *m*–1. The degree centrality of a protein gives an indication of its communicability in a network [28].

Betweenness is a structural property of communication in a network. According to Freeman [28], this means that a protein in a biological network is central if it falls on the shortest path between connecting pairs of other proteins. Thus, other proteins in the network depend on such a protein because it could withhold or distort information during transmission. The greater the betweenness of a protein [28] the greater its influence on the flow of information and importance in the biological processes of an organism. Mathematically, the normalized betweenness of a protein p_1 in a network is defined as [21],

$$B(p_1) = \frac{1}{(m^2 - 3m + 2)} \sum_{(x,y) \in P_{\bar{1}}} \frac{\sigma_{(x,y)}(p_1)}{\sigma_{(x,y)}},$$

Where $\sigma_{(x,y)} = \sigma_{(y,x)}$ is the number of shortest paths from protein *x* to *y* with $\sigma_{(x,x)} = 1$, $\sigma_{(x,y)}(p_1)$ represents the number of shortest paths from *x* to *y* with p_1 as an inner protein [29], and

$$P_{p_1} = \{(x, y) \in P \times P : x \neq y \neq p_1\}.$$

The eigenvector centrality $e(p_i)$ of a protein p_i is a positive multiple [23,30] of the sum of the adjacent centralities

$$\sum_{j=1}^{m} a_{ij} e(p_j) = \lambda e(p_i),$$

for all i. This can be expressed as $Ae = \lambda e$, where *e* is the eigenvector of the adjacency matrix *A* corresponding to the eigenvalue *l*. According to Cvetkovi et al. [31], the eigenvector chosen as the eigenvector centrality must have all positive entries. Among all the eigenvectors corresponding to different eigenvalues *l*, only the one corresponding to the eigenvalue of the largest modulus should be the eigenvector centrality. The eigenvector centrality of a protein gives an indication of how connected the protein is to other well connected proteins in the network.

The closeness of a protein in a network is a measure of the degree to which it is close to other proteins on average and it can also be defined as the reciprocal of the average distance to other proteins. The closeness and betweenness centralities are anchored on the fact that information flows [29] along the shortest paths in a network and does not split. The closeness centrality of a central protein is usually high as it has a shorter distance to other proteins on average [21]. Let $\mathbb{R}^+$ be the set of all positive real numbers, the normalized closeness $C : P \mapsto \mathbb{R}^+$ of a protein is

$$C = \frac{(m-1)}{\sum_{p_1 \neq p_2} d(p_1, p_2)},$$

where $d(p_1, p_2)$ is as defined in (2) and it is the length of the shortest path between p_1 and p_2.

Comparing two biological networks using evolution rewiring

Given any two networks, the approach used by Shou et al. [32] in measuring the evolutionary rewiring rate of biological networks was to name one as the reference network and the other as the compared network. For example, we take MTB as the reference network and MLP as the compared one. Firstly, all orthologous nodes from both networks are identified. This is then followed by the identification of three sets of nodes: Common nodes (CN), Lost Nodes (LN) and Gained Nodes (GN). Common nodes are nodes that have orthologous counterparts in both networks. Loss nodes are nodes present in the MTB network but with an absence of the orthologous counterpart in the MLP network, while gained nodes are nodes present in the compared network that do not have orthologous counterparts in the reference network. Three types of edges were distinguished as: gained edges from gained nodes, lost edges from lost nodes, and common edges from common nodes. TE is the total number of possible edges if both networks were fully connected. TE is the total number of rewired edges obtained by counting all interolog edges in both networks.

The ratio RE/TE was defined as the percentage of edge change in both networks. This ratio will be used later to measure edge changes in the different types of biological networks we want to compare. The formula [32]

$$\text{Rewiring rate} = \frac{RE}{TE \times \text{Time divergence}}, \quad (3)$$

was then used to calculate the evolutionary rewiring rate of two different organisms, where time divergence is the estimated evolutionary divergence time (measured in Mys) between the two organisms. The rewiring rate is measured as the number of rewired edges per edge per Mys. Network identity is defined as [32] the ratio of the number of common edges between orthologous nodes present in both networks to the total number of edges in both networks times 100%.

Results and Discussions

We generated three functional interaction networks for MTB, MLP and MSM by integrating data from eleven different evidence types. In Table 1, we present a comparison of the number of interactions and confidence scores from each of the data sources. We classified the confidence scores into three confidence levels viz-a-viz, low confidence scores are those scores less than 0.3 but not equal to zero (score<0.3), medium confidence scores range from 0.3 to 0.7 ($0.3 \leq$ score ≤ 0.7), while high confidence scores are scores strictly greater than 0.7 but less than or equal to 1 ($0.7<$score ≤ 1). In most cases, the combined score is higher than the individual sub-scores [15] and the confidence increases when a protein-protein interaction is predicted by many data sources or when interaction data are integrated from many evidence types. An understanding of the biological organization of an organism from its PPI network can play a crucial role in vaccine or drug target discovery by highlighting important proteins. Network centrality measures can be used to locate central proteins that play important roles in the biological processes and molecular functions of the organism.

In the next section, we present and compare the functional PPI networks for MTB, MLP and MSM.

Functional Protein-Protein Interaction Networks for the three Mycobacteria

A description of the MTB network has been given previously by Mazandu and Mulder [21], the only difference between the MTB network presented in this work and the previous one, is that the number of functional interactions has increased from 58098 to 59919, because of the addition of PPIs predicted from interologs and PDB. In this paper, we follow the approach used by Mazandu and Mulder [21] in generating biological networks by taking those interactions with medium and high confidence scores. In addition, all the three networks included functional associations with a low confidence score if the interactions were predicted by two or more data sources. However, the number of functional interactions with low confidence scores is small in all three networks compared to those predicted by medium and high confidence scores. Table 2 summarizes important structural properties of the three networks under consideration.

The numbers of proteins in the MLP and MSM networks are 1412 and 4953, while the numbers of protein protein functional interactions are 27042 and 66543, respectively. This shows that the number of proteins and functional association pairs in MSM and MTB are roughly three times that of MLP, though they share a common ancestor. The 1412 proteins in the MLP network corresponds to 87.9% of the complete proteome obtained from the Uniprot database [14,33,34], while the 4953 proteins in MSM amounts to 74.5% of the complete proteome. From Mazandu and Mulder [21] and Table 2, there are 201 hubs, in the MTB network, while MLP and MSM, have 103 and 755 hubs, respectively. Degree based hubs are proteins with a high degree and structural hubs are those proteins that are able to disconnect the network [21]. Here we are referring to structural hubs. The high number of hubs in the MSM network may simply be a reflection of the larger genome and network.

The average path length is computed by finding the mean over all shortest paths between all pairs of proteins in the network [21]. While the MLP network has an average shortest path length of approximately 3 which can be seen in Figure S1(b), the MTB and MSM networks have average shortest path length of approximately 4, as shown in Figures S1(a) and S1(c), respectively. These figures show the probability distributions of their shortest path lengths. The computed average path length for each of the organisms is of the order log $|P|$ in magnitude. Using the same argument as in Mazandu and Mulder [21], this means that each of the networks exhibit the 'small world property' [35,36]. These values give an indication of information spread in their respective networks independent of the number of proteins. Out of the three organisms, MSM has the highest number of connected components at 166 compared to the 23 for MTB and 19 for MLP.

Furthermore, based on the degree of each protein in the three networks, results of our computation show that the distribution of the degree approximates a power-law, that is, for each protein degree

Type of evidence	Low Confidence			Medium Confidence			High Confidence		
	MTB	MLP	MSM	MTB	MLP	MSM	MTB	MLP	MSM
Genomic neighbourhood	1163	417	402	6972	1698	1661	4731	1237	836
Gene fusion	337	26	86	52	14	16	99	27	6
Co-occurence	1033	197	316	5862	276	932	1461	120	369
Experiments	220	185	116	170	224	236	133	510	377
Database knowledge	3	24	19	970	31	33	2002	216	430
Text mining	1174	279	171	722	498	224	93	95	38
Shared domain	0	0	0	20915	0	42805	17792	6070	6478
Sequence similarity	8524	777	0	1345	48	921	77	9	244
Interologs	0	0	0	0	0	0	1701	1600	34
PDB	0	0	0	5082	0	0	864	4683	5487
Coexpression	6538	0	0	225	0	3559	4	12856	4523
Combined Score	6844	145	55	30142	1655	48848	29776	25904	18527

Table 1: Data source and confidence range (low confidence: scores less than 0.3; medium confidence: scores from 0.3 to 0.7; high confidence: scores greater than 0.7) of the functional networks for MLP, MTB and MSM.

k, $P(k) = k^{-\beta}$. We show that each of the networks depicts a scale-free topology. The degree exponents for each of the mycobacterial species are $\beta \sim 3.48$, $\beta \sim 3.41$ and $\beta \sim 3.41$ for the MTB (Figure S2(a)),MSM(Figure S2(c)) and MLP (Figure S2(b)) networks, respectively.

Locating the most central protein

One of the major problems in network analysis is which criteria should be used to identify the most central protein in a network. However, as stated earlier, the higher the betweenness of a protein, the greater the influence on the flow of information and importance in the biological processes of an organism. By ordering proteins based on betweenness, the MTB network in Mazandu and Mulder [21] was analyzed for the variations in the betweeness metric in terms of protein category by showing that proteins with high degree centralities that are positioned in the centre of the network are more likely to locate other proteins in a connected component faster than structural hubs. From the three computed networks, we used the betweenness centrality metric to obtain the following most central proteins: Q8VKQ9 in the MTB network, which does not have an ortholog in the other two networks; A0R5I7 in the MSM network, which is not orthologous to any protein in the MTB and MLP networks, and P57993 in the MLP network, which is an ortholog of P65728 in the MTB network and A0QQK3 in the MSM network. Table 3 shows some important properties of these central proteins based on network centrality measures. Apart from the most central MTB protein, which belongs to the PE family, the most central proteins in MSM and MLP (and its orthologs) are Serine-threonine protein kinases.

Next, for each network, we calculated the shortest path length, d of all proteins from Q8VKQ9, A0R5I7 and P57993, classified them into d=1, 2, 3, · · ·, 9 from the protein and n, the total number of proteins in each set. For example, from Figure S3 (supplementary material), d=1 represents proteins which have a shortest path length one from Q8VKQ9 and n=156 proteins belong to that category, in the same vein, d=2 means those proteins with distance two from Q8VKQ9 and n=987 etc. Out of the 4136 proteins in the MTB network, only 53 have no path to this protein. A similar diagram showing the distribution of shortest path lengths of other proteins from A0R5I7 for the MSM network and P57993 for the MLP network are illustrated in Figures S4 and S5 (supplementary material), respectively. Thirty-six (out of 1412) proteins have no path to P57993 in MLP and 409 out of the 4953 proteins have no path to A0R5I7 in MSM.

For different protein sets at each distance from the central protein for the three organisms under consideration, we used their Gene Ontology (GO) annotations to carry out GO term over-representation analysis using Blast2GO [37]. Blast2GO enrichment analysis tool produces a statistical significance [38] analysis of each GO term in a given protein set using Fisher's exact test to find over or under represented functional labels between two protein sets. The proteins in the sets under consideration were used as the query sets and the remaining proteins in the network constitute the reference-set. We considered functions with p-values less than 0.05 to be significant and computed the adjusted p-values using the BONFERRONI correction [39]. The results are shown in Tables S2-S10 (supplementary material). From Table S2, the result shows that proteins annotated to cellular metabolic process were the most over-represented for proteins sets with d=1 and 2 in MLP. In Table S3, the GO term corresponding to 'integral to membrane' is over-represented in d=1, oxidoreductase and acyl-CoA dehydrogenase activity in d=2, and proteins involved in biosynthetic processes are over-represented in d=3 for MTB. Similarly, in Tables S4-S5, the GO terms cellular metabolic process and protein metabolic process were over-represented in MSM for d=1 and both nucleotide and ATP binding were over-represented for d=2. Furthermore, we subdivided the total proteins in each mycobacterial organism into different sets based on network centrality measures and used Blast2GO's [37] Fisher's Exact test to find over represented GO terms in each of the sets. The results are as tabulated in Tables S2-S10 (supplementary material). Table S6 shows for high degree proteins the GO-term 'translation', high closeness, the GO-term 'nucleotide binding' and high eigenvector centrality the GO-term 'ribonucleoprotein complex' are significantly over-represented in the MLP network. From tables S7-S8, hubs having GO-term 'cobalin binding' and high degree proteins (degree >100) have GO-term 'regulation of transcription, DNA-dependent' significantly over-represented in MSM. In the same vein, as shown in Tables S9-S10 for the MTB network, protein sets that are hubs, high degree, high betweenness, high closeness and eigenvector have the following respective GO-terms: 'transposase activity', 'oxidoreductase activity', 'binding', 'ACP phosphopantetheine', and 'oxidoreductase activity' significantly over-represented.

Comparing important proteins in the MTB and MLP networks

We compare important proteins in the MTB and MLP networks using the approaches presented in Mazandu and Mulder [21] and Shou et al. [32] to understand the biological processes to which these proteins are involved. These proteins possess certain topological properties such as having high betweenness, closeness and eigenvector centralities, which make them important in the functionality of the network.

The center of gravity of a network is defined as the set of proteins that maximizes the closeness measure to any other protein in the network [21]. A protein in a functional network belongs to the gravity centre, if its closeness centrality measure is strictly greater than the reciprocal of the average shortest path length [21]. This value corresponds to 1/3.62739 or 0.27568 in MTB and 1/3.16955 or 0.31550 in the MLP network. In using the betweenness centrality to determine important proteins, we considered those proteins in which their betweenness is greater than the total number of shortest paths; obtained by multiplying the average shortest path length by the total number of proteins in the functional network. We then combined these criteria with the requirement that the eigenvector centrality should be greater than 10−5. We obtained a set of 355 and 116 proteins which have a high centre of gravity and thus may be potentially interesting as drug targets in the MTB and MLP networks, respectively [40-43]. Interestingly, proteins belonging to the intermediary metabolism and respiration functional class are the most represented in these lists of potential drug targets as shown in Table S1 and Figure 1. This is followed by proteins belonging to the unknown classes for MTB and information pathways for MLP. We obtained the functional classes from Tuberculist (http://genolist.pasteur.fr/Tuberculist, accessed 28 October, 2011) & Leproma (http://genolist.pasteur.fr/Leproma, accessed 28 September, 2012). Among these potential drug targets, we extracted those proteins with high closeness which are classified as central proteins, and influential proteins, which are those with high eigenvector centralities. 241 and 69 are central and influential targets, respectively in the MTB network, while 95 and 37 are central and influential targets, respectively in MLP.

We used the technique described in Shou et al. [32] to identify a total of 2859 proteins in the MTB network without a corresponding ortholog in the MLP network and 135 proteins in the MLP network

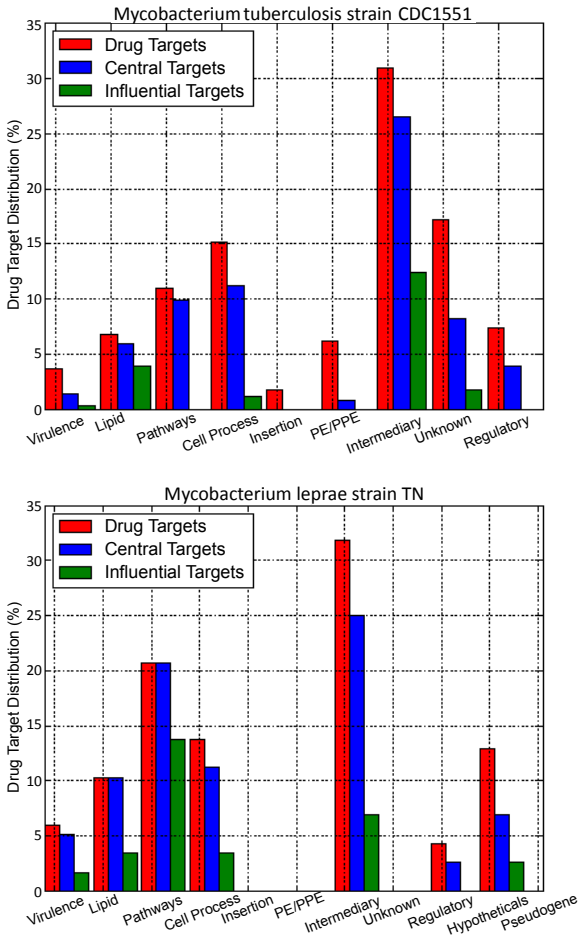

Figure 1: Comparing functional classes for drug, central, and influential targets in MTB and MLP.

without corresponding orthologs in the MTB network. In total, 1277 proteins have orthologous counterparts in both networks as shown in Table 4, only five pairs are both hubs (Table S10). A close look at Table S10 shows that these proteins have high betweenness and closeness. The protein O07727 (Probable D-amino-acid oxidase) in the MTB network is a drug, central and influential target and should be examined further as a potential drug target.

Comparing important proteins in the MTB and MSM networks

Due to the unavailability of curated functional classes for MSM at the time of writing, we were unable to make the same kind of comparison as with MLP. However, we identified 2148 distinct proteins belonging to the MTB network without corresponding orthologs in the MSM network. Similarly, 2965 proteins in the MSM network have no corresponding orthologs in the MTB network. 1988 proteins have orthologous counterparts in both networks (Table 5). Furthermore, we found five orthologous protein pairs in the networks that are both hubs (Table S11). Using the same approach as outlined previously, we identified 294 potential drug targets in the MSM network and by choosing proteins with closeness greater than 0.27, we found 184 proteins as central targets. As defined earlier, influential targets are proteins top ranked by their eigenvector centralities, with values greater than 0.07. We identified 16 influential target proteins in MSM.

Comparing important proteins in the MSM and MLP networks

As shown in Table 5, out of the 1412 proteins in the MLP proteome, only 342 have no orthologous counterpart in the MSM network. 3883 proteins in MSM have no corresponding ortholog in MLP. Sixteen out of the 1070 orthologs present in both networks are both hubs (Table S12).

Table 5 summarizes the three comparisons discussed so far. A common edge is an edge in which both protein pairs are corresponding orthologs in both networks and are interologs. From the second to last column, we include the total number of common edges to the two organisms being compared. 3693 functional interactions are common to the MTB and MLP networks, 2284 edges are common to the MSM and MTB networks, while 1901 are common to MLP and MSM.

From the three networks, we sought those proteins which have orthologous counterparts in all three organisms and found a total of 1001 proteins (Figure 2), 260 additional proteins have orthologs in the MLP and MTB networks, 46 proteins have orthologs in MLP and MSM alone, and 983 orthologous proteins are present in just the MTB and MSM networks *etc*. All three networks have 297 common edges. Based on the classification of proteins as drug, central and influential targets

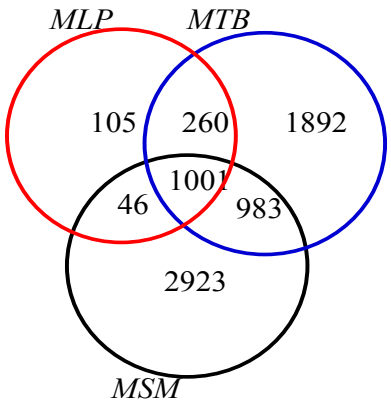

Figure 2: Venn diagram showing the orthologs shared and number of unique proteins in the three organisms.

Parameters	Values		
	Mycobacterium tuberculosis	*Mycobacterium leprae*	*Mycobacterium smegmatis*
Number of proteins (Nodes)	4136	1412	4953
Number of functional interactions (Edges)	59919	20742	66543
Number of hubs	201	103	755
Density	0.007	0.0208	0.0054
Average degree	28	29	26
Average shortest path length	3.62739	3.16955	4.2224
Number of connected components	23	19	166
% of Nodes in largest component	98.7%	97.5%	91.7%

Table 2: Comparing network parameters and values in the MTB, MLP and MSM networks.

and using the 1001 orthologs in their intersection, we found eight drug and one influential target overlaps among the three organisms.

Subnetworks of orthologs only

From the original three networks, we removed those proteins (and interactions or edges involving those proteins) that are not among the 1001 set of common proteins. We are now left with three subnetworks each consisting of 1001 proteins. The network properties of the three subnetworks are compared in Table 6. The last row (singletons) shows the number of proteins in the subnetworks without neighbours or proteins, and thus with degree zero. We then determined the number of shared edges for the orthologs and used this to calculate network identity for these subnetworks. The last column of Table 7 shows that the MLP and MTB subnetworks are more similar than the MTB versus MSM and MLP versus MSM subnetworks.

As an example, from Table S10, we chose the Q7D903 protein from the MTB network on the grounds that it has a high betweenness, it is a hub and has orthologs in both MLP and MSM networks. As shown in Figure 5, this protein has 37 neighbours while its corresponding orthologs Q9CD64 and A0R3F9 have 17 and 34 neighbours in MLP and MSM networks, respectively (Figures 6 and 7). However, out of the 37 neighbours, Figure 8 shows that only seven of them have orthologs in both MLP and MSM. In the same vein, out of the 17 neighbours of Q9CD64, 13 have orthologs in both MTB and MSM core subnetworks. Finally, the third network in Figure 8) shows that only seven proteins out of the 34 neighbours of A0R3F9 have orthologs in both MLP and MTB subnetworks. Among the three proteins in Figure 8, only six of their respective neighbours are orthologs of each other. The functional classes of the neighbour nodes are not conserved but then neither are those of the central ortholog proteins. This may be due to differences in assigning functional classes for different organisms and the fact that for MSM we had to use GO terms, since there was no functional class label available. We used PINV http://biosual.cbio.uct.ac.za/biosual/tests /pinv/pinv.html (accessed on 15 June, 2013) to generate the figure showing the interactions these proteins are involved in.

Evolutionary differences between the three mycobacterial species

In line with Shou et al. [32], for each pair of networks compared, we took sub-samples of their edges from 100% to 1%. We subjected both networks to random attacks by removing 50 nodes and all edges involving those nodes; repeated the simulations 10 times before calculating 95% confidence intervals on the resulting numbers. Since the MTB and MSM networks have more edges and nodes than the MLP network, we decided to remove 200 nodes rather than 50. The results in Tables S13-S15 show that as more nodes and their edges were removed from the networks, the percentage of edge change (defined previously) decreases. Since the three organisms have a common ancestor, and considering that they diverged approximately 2000 million years ago and computed rewiring rates for MTB and MLP networks and MLP and MSM networks.

In line with Shou et al. [32], we define bottlenecks as proteins within the top 10% ranked by betweenness and hubs as proteins within the top 10% ranked by degree. The choice of 10% is due to the small number of proteins in the MLP network. We grouped proteins in each network into Bottleneck Hubs (BH), Non-Hub-Bottlenecks (NH-B), Non-Bottleneck Hubs (NB-H) and Non-Hub Non-Bottlenecks (NH-NB). Bottleneck hubs are proteins within the top 10% ranked by betweenness and degree, non hubs non bottlenecks are neither within the top 10% ranked by betweenness nor degree. Non-hub bottlenecks are not within the top 10% ranked by degree but are top 10% ranked by betweenness, while non-bottleneck hubs are the converse of non-hub bottlenecks. Figures 3 and 4 show that Bottleneck hubs rewire faster than non-hubs non bottlenecks. For the MSM network, it can be observed from Figure 4 that no proteins belong to the Non-hub non-bottlenecks category. These results agree with those in Shou et al. [32].

Conclusion

In this study, we have generated functional protein-protein interaction networks for *Mycobacterium leprae*, *Mycobacterium smegmatis* and used an updated *Mycobacterium tuberculosis* network. In addition, since the betweenness centrality is a measure of the flow of information in a network, we have identified central proteins in each of the three networks, and carried out an overrepresentation analysis of the GO terms in the three networks under different headings. We compared the three networks using the following three-way approach: slow grower (MTB) versus slow grower (MLP), fast grower MSM versus MLP, MTB versus MSM and using orthologs as described in Shou et al. [32], we identified 1001 orthologous proteins common to the three networks. We also computed the network identities of the compared networks and determined that MLP's network is more similar to the MTB network than the MSM network which makes sense as they are more closely related and both are slow growers. Furthermore, from the original three networks, we removed those proteins and functional interactions involving those proteins that are not among the 1001 set of proteins. Thus, obtaining three subnetworks each consisting of 1001 proteins. Based on the criteria outlined in Section 3-C, we determined eight overlapping drug targets proteins and one overlapping influential

Organism	Uniprot Acc	Description	Eigenvector	Betweenness	Closeness	Degree	Hubs
Mycobacterium tuberculosis	Q8VKQ9	PE family protein	1.65648e-03	117248.09	0.34124	156	N
Mycobacterium smegmatis	A0R5I7	Serine/threonine protein kinase	5.26173e-04	605562.88	0.31979	153	Y
Mycobacterium leprae	P57993	Probable serine/threonine protein kinase	3.05175e-02	36356.78	0.42189	125	N

Table 3: The most central proteins for the three mycobacterial organisms and their centrality measures.

Organism	Uniprot Acc.	Uniprot Description	shortest path length (*d*)								
			1	2	3	4	5	6	7	8	9
Mycobacterium tuberculosis	Q8VKQ9	PE family protein	156	987	2190	656	77	12	4	0	0
Mycobacterium leprae	P57993	Probable serine/threonine-protein kinase	125	820	328	87	11	4	0	0	0
Mycobacterium smegmatis	A0R5I7	Serine/threonine protein kinase	153	1573	1857	721	174	51	10	3	1

Table 4: The three most important proteins in the three networks and the shortest path lengths from each of them.

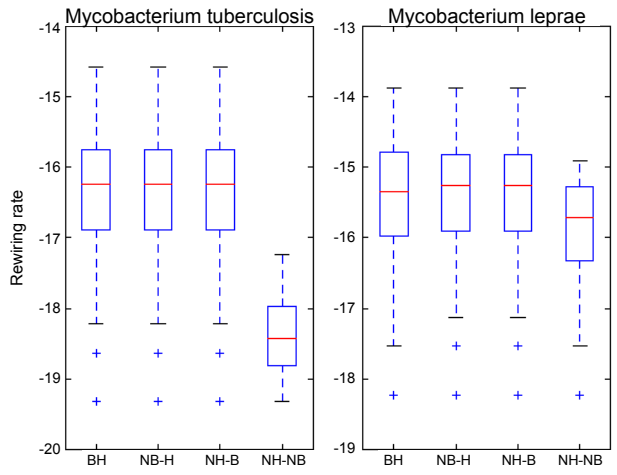

Figure 3: Rewiring rates for MTB and MLP networks. We used log scale on the rewiring axis.

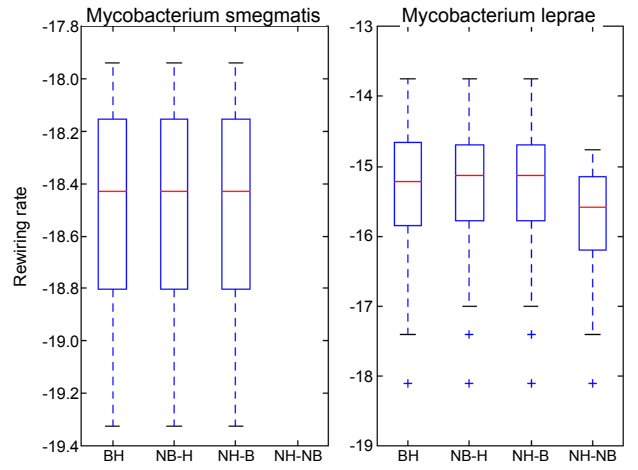

Figure 4: Rewiring rates for MSM and MLP networks.

A	B	Proteins in A only	Proteins in B only	Common Proteins	Common Edges	Network Identity
MLP	MTB	135	2859	1277	3693	4.5%
MSM	MTB	2965	2148	1988	2284	1.5%
MLP	MSM	342	3883	1070	1901	2.1%

Table 5: Number of ortholog proteins shared, common edges and network identity of the compared networks.

target protein among the 1001 proteins set. We then determined the number of shared edges for the orthologs and used this to calculate network identity for these subnetworks. Our result shows that the MLP and MTB subnetworks are more similar than the MTB versus MSM and MLP versus MSM subnetworks. One other interesting result that we found in this study is that among the three proteins in the three subnetworks shown in Figure 8, only six of their respective neighbours are orthologs of each other. Finally, by comparing the network rewiring rates of the compared organisms, we determined that bottleneck hubs rewire faster than non-hubs non-bottlenecks in line with Shou et al. [32].

Acknowledgements

The authors appreciate financial support received from the National Research Fundation (NRF) South Africa and the developers of open-source software. We also appreciate Kenneth Opap for useful discussions on PDB and most especially the Computational Biology Group of the Institute of Infectious Disease and Molecular Medicine here at the University of Cape Town.

Parameters	Values		
	Mycobacterium tuberculosis	*Mycobacterium leprae*	*Mycobacterium smegmatis*
Number of proteins (Nodes)	1001	1001	1001
Number of functional interactions (Edges)	9941	13670	5086
Number of hubs	38	60	160
Density	0.0198	0.0273	0.0101
Average degree (μ_d)	20	27	11
Average shortest path length	3.1055	3.0124	4.1655
Number of connected components	16	33	157
% of Nodes in largest component	98.5%	95.8%	79.2%
#(Degree-μ_d) < 0	628	598	773
#(Degree-μ_d) = 0	25	10	16
#(Degree-μ_d) > 0	976	393	212
Singletons	15	24	125

Table 6: Comparing network parameters and values in the MTB, MLP and MSM subnetworks. μ_d is the mean degree for each network. #(Degree-μ_d) < 0=628 means the total number of proteins such that the deviation from the mean degree (Degree-μ_d) is less than zero is 628.

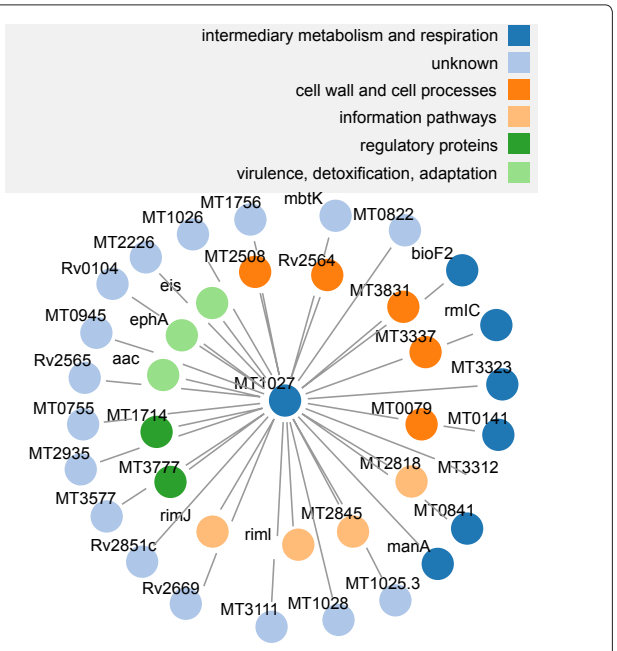

Figure 5: The structural hub protein MT1027 with Uniprot accession Q7D903, in the MTB network and its 37 neighbours. Nodes are coloured by functional classes. This is a drug target protein.

A	B	Edges in A only	Edges in B only	Common Proteins	Common Edges	Network Identity
MLP	MTB	13670	9941	1001	2820	11.9%
MSM	MTB	5086	9941	1001	656	4.3%
MLP	MSM	13670	5086	1001	1849	9.8%

Table 7: Number of common edges and network identity of the compared sub networks.

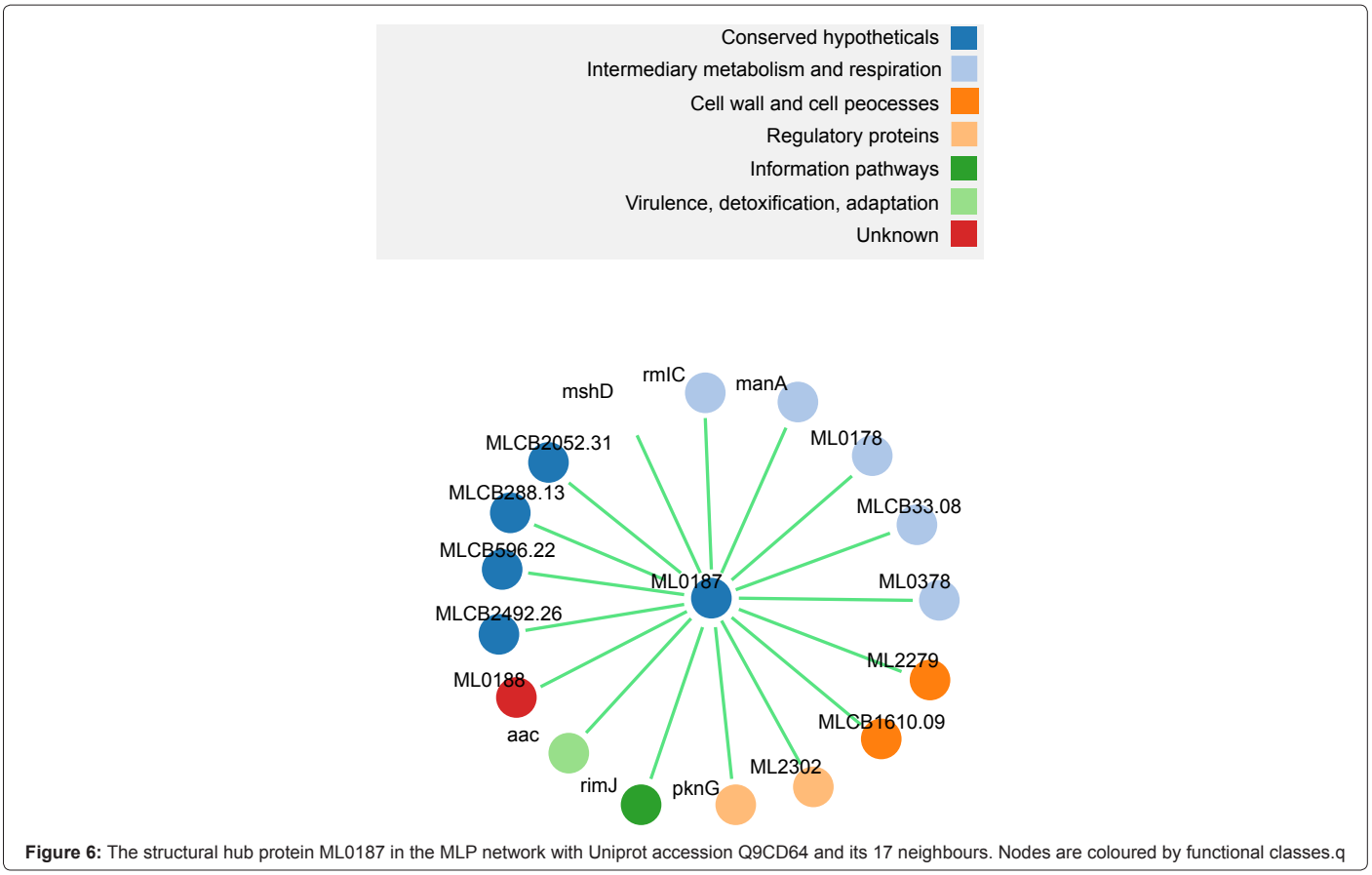

Figure 6: The structural hub protein ML0187 in the MLP network with Uniprot accession Q9CD64 and its 17 neighbours. Nodes are coloured by functional classes.q

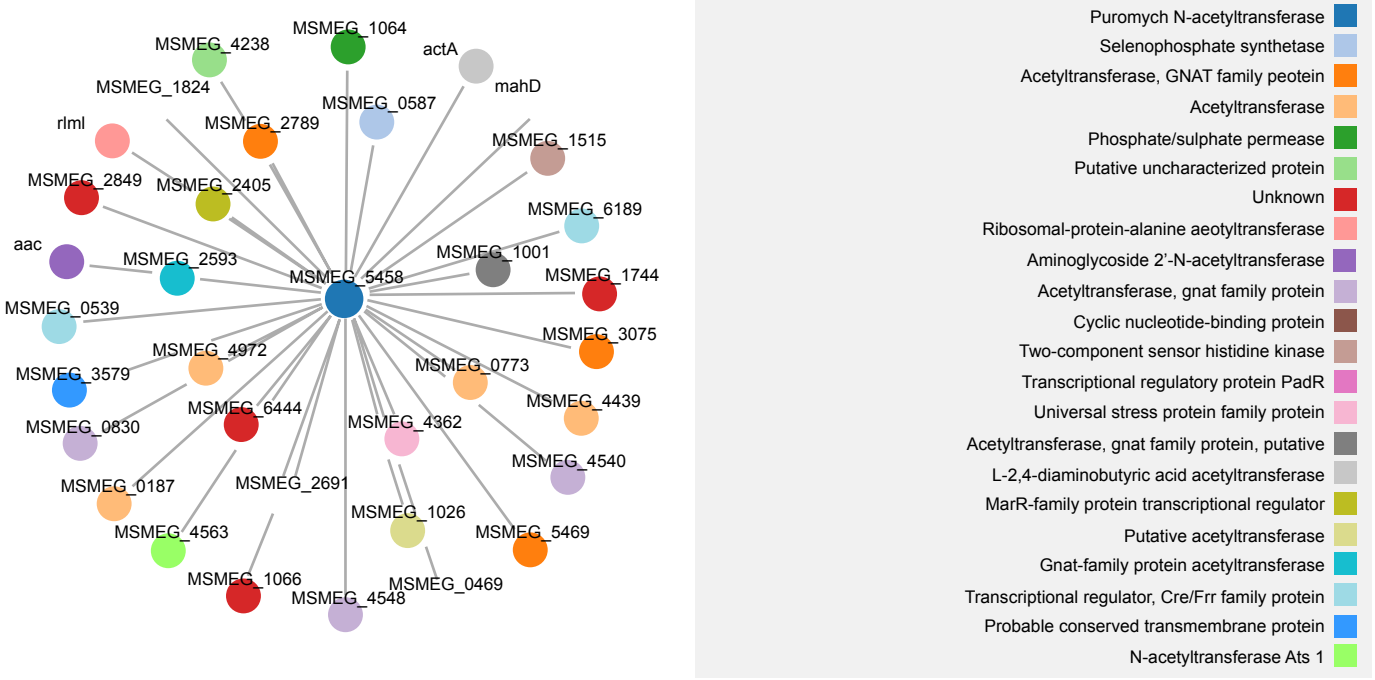

Figure 7: The structural hub protein MSMEG_5458 with Uniprot accession A0R3F9 in the MSM network and its 34 neighbours. Nodes are coloured by GO terms. This is a drug target protein.

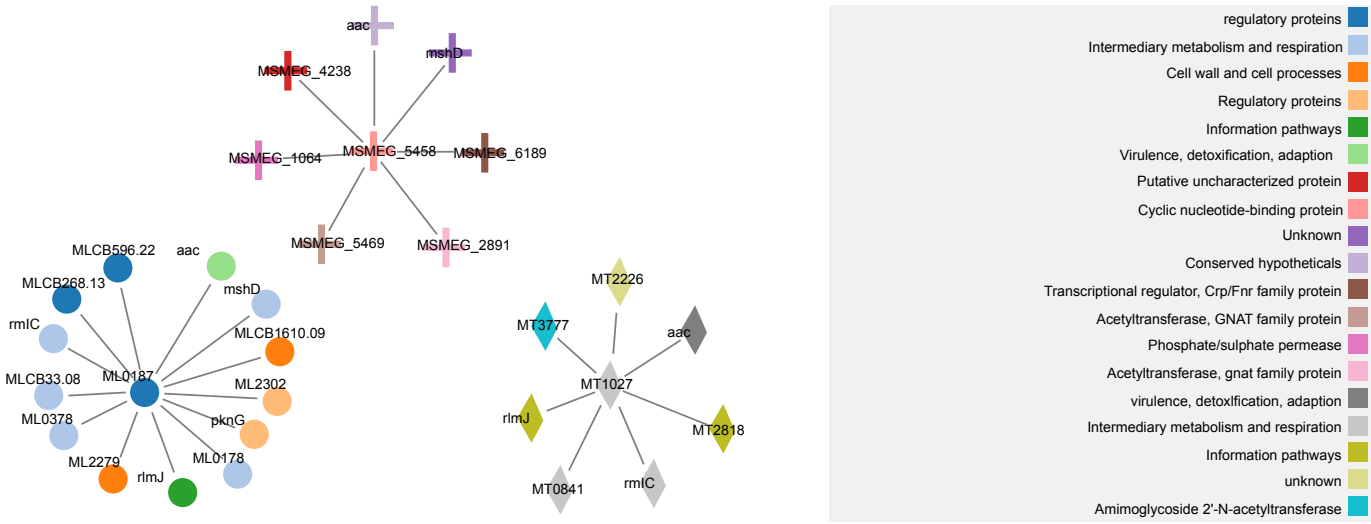

Figure 8: The three orthologous proteins Q7D903, Q9CD64, A0R3F9 in MTB, MLP, and MSM respectively and their neighbours. We coloured the nodes by functional classes and GO terms for MSM.

References

1. Monot M, Honoré N, Garnier T, Zidane N, Sherafi D, et al. (2009) Comparative genomic and phylogeographic analysis of Mycobacterium leprae. Nat Genet 41: 1282-1289.
2. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409: 1007-1011.
3. http://www.who.int/lep/leprosy/en/index.html
4. http://www.who.int/lep/classification/en/index.html
5. http://www.who.int/lep/situation/prevalence/en/index.html
6. Mazandu GK (2010) Data Integration for the Analysis of Un-characterized Proteins in Mycobacterium tuberculosis, University of Cape Town, South Africa.
7. Cordone A, Audrain B, Calabrese I, Euphrasie D, Reyrat JM (2011) Characterization of a Mycobacterium smegmatis uvrA mutant impaired in dormancy induced by hypoxia and low carbon concentration. BMC Microbiol 11: 231.
8. He Z, Buck DJ (2010) Cell wall proteome analysis of Mycobacterium smegmatis strain MC2 155. BMC Microbiology 10.
9. Eiglmeier K, Parkhill J, Honoré N, Garnier T, Tekaia F, et al. (2001) The decaying genome of Mycobacterium leprae. Lepr Rev 72: 387-398.
10. Merkov NL (2006) Glyoxylate Metabolism in Mycobacterium smeg-matis. Ph.D. thesis, The Rockefeller University, USA.
11. Youm J, Saier MH Jr (2012) Comparative analyses of transport proteins encoded within the genomes of Mycobacterium tuberculosis and Mycobacterium leprae. Biochim Biophys Acta 1818: 776-797.
12. Cole ST (1998) Comparative mycobacterial genomics. Curr Opin Microbiol 1: 567-571.
13. Gómez-Valero L, Rocha EP, Latorre A, Silva FJ (2007) Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction. Genome Res 17: 1178-1185.
14. UniProt Consortium (2009) The Universal Protein resource (Uniprot) in 2010, Nucleic Acids Res 38: D142–D148.
15. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33: D433-437.
16. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8 a global view on proteins and their functional inter-actions in 630 organisms, Nucleic Acids Research 37: D412-D416.
17. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res 39: D730-735.
18. Raghavachari B, Tasneem A, Przytycka TM, Jothi R (2008) DOMINE: a database of protein domain interactions. Nucleic Acids Res 36: D656-661.
19. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40: D857-D861.
20. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, et al. (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28: 289-291.
21. Mazandu GK, Mulder NJ (2011) Scoring protein relationships in functional interaction networks predicted from sequence data. PLoS One 6: e18607.
22. Mazandu GK, Opap K, Mulder NJ (2011) Contribution of microarray data to the advancement of knowledge on the My-cobacterium tuberculosis interactome: Use of the random partial least squares approach. Infect Genet Evol 11: 725-733.
23. Ruhnau B (2000) Eigenvector-centrality a node-centrality? Elsevier 22: 357-365.
24. http://cs.bme.hu/fcs/graphtheory.pdf
25. Hagberg A, Schult D, Swart P, (2012) NetworkX Release 1.7.
26. Nieminen UJ (1973) On the centrality in a directed graph, Social Science Research 2: 371–378.
27. Nieminen J (1974) On centrality in a graph. Scand J Psychol 15: 332–336.
28. Freeman LC (1978/79) Centrality in social networks conceptual clarification, Social Networks 1: 215–239.
29. Brandes U, Fleischer D, Centrality Measures Based on Current Flow. Lecture Notes in Computer Science 3404: 533-544.
30. Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Sociol 2: 113–120.
31. Cvetkovic DM, Doob M, Sachs H (1995) Spectra of graphs. (3rd Edn), Barth, Heidelberg Leipzig, Germany.
32. Shou C, Bhardwaj N, Lam HY, Yan KK, Kim PM, et al. (2011) Measuring the evolutionary rewiring of biological networks. PLoS Comput Biol 7: e1001050.
33. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115-119.
34. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10: 136.

35. Gursoy A, Keskin O, Nussinov R (2008) Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans 36: 1398-1403.

36. Mason O, Verwoerd M (2007) Graph theory and networks in Biology. IET Syst Biol 1: 89-119.

37. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

38. Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008: 619832.

39. Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310: 170.

40. Cole ST (2002) Comparative mycobacterial genomics as a tool for drug target and antigen discovery. Eur Respir J Suppl 36: S78-S86.

41. Parkash O, Singh BP (2012) Advances in Proteomics of Mycobacterium leprae. Scand J Immunol 75: 369-378.

42. Brosch R, Gordon SV, Eiglmeier K, Garnier T, Cole ST (2000) Comparative genomics of the leprosy and tubercle bacilli. Res Microbiol 151: 135-142.

43. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25: 288–289.