Advances in Life Science and Technology
ISSN 2224-7181 (Paper) ISSN 2225-062X (Online)
Vol 5, 2012

www.iiste.org

Comparative Computational Analysis of Mycobacterium Species by using Different Techniques in Study

Nitesh Chandra Mishra*, Kiran Singh
Department of Bio-Informatics, Maulana Azad National Institute of Technology, Bhopal, M.P., India
*E-mail of the Corresponding author: researchnm.kiran@gmail.com, nmishra.nit@gmail.com

Abstract

Mycobacterium tuberculosis (MTB) is a pathogenic bacteria species in the genus Mycobacterium and the causative agent of most cases of tuberculosis. It is spread through the air when people who have an active MTB infection cough, sneeze, or otherwise transmit their saliva through the air. Most infections in humans result in an asymptomatic, latent infection, and about one in ten latent infections eventually progresses to active disease, which, if left untreated, kills more than 50% of its victims. Mycobacterium tuberculosis is a member of the genus 'tuberculosis' in which contains various other mycobacterium species also. These species within a gene must have some similarity in them. In spite of this similarity only mycobacterium tuberculosis cause the tuberculosis disease, the remaining does not. This signifies that mycobacterium tuberculosis must be having some specific genes or proteins which are uniquely present only in it and not in the other species. This fact is used in this research and blast program is executed recursively for the comparison between these mycobacterium species.

Keywords: BLAST, Mycobacterium Tuberculosis, Nontuberculous mycobacterium group, EEA1, KEGG

1. Introduction

Tuberculosis, MTB or TB is a common and in some cases infectious disease caused by various strains of mycobacterium, usually Mycobacterium tuberculosis in humans. Tuberculosis usually attacks the lungs but can also affect other parts of the body. It is spread through the air when people who have an active MTB infection cough, sneeze, or otherwise transmit their saliva through the air. Most infections in humans result in an asymptomatic, latent infection, and about one in ten latent infections eventually progresses to active disease, which, if left untreated, kills more than 50% of its victims.

The classic symptoms are a chronic cough with blood-tinged sputum, fever, night sweats, and weight loss (the last giving rise to the formerly prevalent colloquial term "consumption"). Infection of other organs causes a wide range of symptoms. Diagnosis relies on radiology (commonly chest X-rays), a tuberculin skin test, blood tests, as well as microscopic examination and microbiological culture of bodily fluids. Treatment is difficult and requires long courses of multiple antibiotics. Contacts are also screened and treated if necessary. Antibiotic resistance is a growing problem in (extensively) multi-drug-resistant tuberculosis. Prevention relies on screening programs and vaccination, usually with Bacillus Calmette-Guérin vaccine.

One third of the world's population is thought to be infected with M. tuberculosis, and new infections occur at a rate of about one per second. The proportion of people who become sick with tuberculosis each year is stable or falling worldwide but, because of population growth, the absolute number of new cases is still increasing. In 2007 there were an estimated 13.7 million chronic active cases, 9.3 million new cases, and 1.8 million deaths, mostly in developing countries. In addition, more people in the developed world contract tuberculosis because their immune systems are more likely to be compromised due to higher exposure to immunosuppressive drugs, substance abuse, or AIDS. The distribution of tuberculosis is not uniform across the globe; about 80% of the population in many Asian and African countries test positive in tuberculin tests, while only 5–10% of the US population test positive.

The cause of TB, Mycobacterium tuberculosis (MTB), is a small aerobic non-motile bacillus. High lipid content of this pathogen accounts for many of its unique clinical characteristics. It divides every 16 to 20 hours, an extremely slow rate compared with other bacteria, which usually divide in less than an hour. (For example, one of the fastest-growing bacteria is a strain of E. coli that can divide roughly every 20 minutes.) Since MTB has a cell wall but lacks a phospholipid outer membrane, it is classified as a Gram-positive bacterium. However, if a Gram stain is performed, MTB either stains very weakly Gram-positive or does not retain dye as a result of the high lipid & mycolic acid content of its cell wall. MTB can withstand weak disinfectants and survive in a dry state for weeks. In nature, the bacterium can grow only within the cells of a host organism, but M. tuberculosis can be cultured in vitro.

The M. tuberculosis complex includes four other TB-causing mycobacterium: M. bovis, M. africanum, M. canetti and M. microti. M. africanum is not widespread, but in parts of Africa it is a significant cause of tuberculosis. M. bovis was once a common cause of tuberculosis, but the introduction of pasteurized milk has largely eliminated this as a public health problem in developed countries. M. canetti is rare and seems to be limited to Africa, although a few cases have been seen in African emigrants. M. microti is mostly seen in immunodeficient people, although it is possible that the prevalence of this pathogen has been underestimated.

Other known pathogenic mycobacterium includes Mycobacterium leprae, Mycobacterium marinum, Mycobacterium avium and M. kansasii. The last two are part of the nontuberculous mycobacterium (NTM) group. Nontuberculous mycobacterium cause neither TB nor leprosy, but they do cause pulmonary diseases resembling TB.

Mycobacterium tuberculosis (MTB) is a pathogenic bacteria species in the genus Mycobacterium and the causative agent of most cases of tuberculosis. First discovered in 1882 by Robert Koch, M. tuberculosis has an unusual, waxy coating on the cell surface (primarily mycolic acid), which makes the cells impervious to Gram staining so acid-fast detection techniques are used instead. The physiology of M. tuberculosis is highly aerobic and requires high levels of oxygen. Primarily a pathogen of the mammalian respiratory system, MTB infects the lungs. The most frequently used diagnostic methods for TB are the tuberculin skin test, acid-fast stain, and chest radiographs. The M. tuberculosis genome was sequenced in 1998.

M. tuberculosis requires oxygen to grow. It does not retain any bacteriological stain due to high lipid content in its wall, and thus is neither Gram positive nor Gram negative; hence Ziehl-Neelsen staining, or acid-fast staining, is used. While mycobacterium do not seem to fit the Gram-positive category from an empirical standpoint (i.e., they do not retain the crystal violet stain), they are classified as acid-fast Gram-positive bacteria due to their lack of an outer cell membrane. M. tuberculosis divides every 15–20 hours, which is extremely slow compared to other bacteria, which tend to have division times measured in minutes (Escherichia coli can divide roughly every 20 minutes). It is a small bacillus that can withstand weak disinfectants and can survive in a dry state for weeks. Its unusual cell wall, rich in lipids (e.g., mycolic acid), is likely responsible for this resistance and is a key virulence factor. When in the lungs, M. tuberculosis is taken up by alveolar macrophages, but they are unable to digest the bacterium. Its cell wall prevents the fusion of the phagosome with a lysosome. Specifically, M. tuberculosis blocks the bridging molecule, early endosomal autoantigen 1 (EEA1); however, this blockade does not prevent fusion of vesicles filled with nutrients. Consequently, the bacteria multiply unchecked within the macrophage. The bacteria also carried the UreC gene, which prevents acidification of the phagosome. The bacteria also evade macrophage-killing by neutralizing reactive nitrogen intermediates. The ability to construct M. tuberculosis mutants and test individual gene products for specific functions has significantly advanced our understanding of the pathogenesis and virulence factors of M. tuberculosis. Many secreted and exported proteins are known to be important in pathogenesis.

M. tuberculosis comes from the genus Mycobacterium, which is composed of approximately 100 recognized and proposed species. The most familiar of the species are Mycobacterium tuberculosis and Mycobacterium leprae (leprosy). M. tuberculosis appears to be genetically diverse, which results in significant phenotypic differences between clinical isolates. M. tuberculosis exhibits a biogeographic population structure and different strain lineages are associated with different geographic regions. Phenotypic studies suggest this strain variation never has implications for the development of new diagnostics and vaccines. Microevolutionary variation affects the relative fitness and transmission dynamics of antibiotic-resistant strains. Mycobacterium outbreaks are often caused by hypervirulent strains of M. tuberculosis. In laboratory experiments, these clinical isolates elicit unusual immunopathology, and may be either hyperinflammatory or hypoinflammatory. Studies have shown the majority of hypervirulent mutants have deletions in their cell wall modifying enzymes or regulators that respond to environmental stimuli. Studies of these mutants have indicated the mechanisms that enable M. tuberculosis to mask its full pathogenic potential, inducing a granuloma that provides a protective niche and enables the bacilli to sustain a long-term, persistent infection.

The genome of the H37Rv strain was published in 1998. Its size is 4 million base pairs, with 3959 genes; 40% of these genes have had their function characterized, with possible function postulated for another 44%. Within the genome are also 6 pseudogenes.The genome contains 250 genes involved in fatty acid metabolism, with 39 of these involved in the polyketide metabolism generating the waxy coat. Such large numbers of conserved genes show the evolutionary importance of the waxy coat to pathogen survival. About 10% of the coding capacity is taken up by two clustered gene families that encode acidic, glycine-rich proteins. These proteins have a conserved N-terminal motif, deletion of which impairs growth in macrophages and granulomas. Nine noncoding sRNAs have been characterized in M. tuberculosis, with a further 56 predicted only an estimated 10% of people infected with M. tuberculosis ever develop the disease, and many of those have the disease only for the first few years following infection, even though the bacillus may lie dormant in the body for decades.

The symptoms that patients infected with M. tuberculosis may experience are usually absent until the disease has become more complicated. It may take many months from the time the infection initially gets into the lungs until symptoms develop. Cough is however the first symptom of the infection with M. tuberculosis. The initial symptoms, including loss of appetite, fever, productive cough and loss of energy or loss of weight or night sweats, are not specific and might be easily attributed to another condition. Primary pulmonary tuberculosis is the first stage of the condition, and it may cause fever, dry cough and some abnormalities that may be noticed on a chest X-ray. In most cases, though, primary infections tend to cause no symptoms that people do not overcome. This condition resolves itself, although it returns in more than half of the cases.

Tuberculosis causing lung disease may result in tuberculous pleuritis, a condition that may cause symptoms such as chest pain, nonproductive cough and fever. Moreover, infection with M. tuberculosis can spread to other parts of the body, especially in patients with a weakened immune system. This condition is referred to as miliary tuberculosis, and people contacting it may experience fever, weight loss, weakness and a poor appetite. In more rare cases, miliary tuberculosis can cause coughing and difficulty breathing.

Only an estimated 10% of people infected with M. tuberculosis ever develop the disease, and many of those have the disease only for the first few years following infection, even though the bacillus may lie dormant in the body for decades. The symptoms that patients infected with M. tuberculosis may experience are usually absent until the disease has become more complicated. It may take many months from the time the infection initially gets into the lungs until symptoms develop. Cough is however the first symptom of the infection with M. tuberculosis. The initial symptoms, including loss of appetite, fever, productive cough and loss of energy or loss of weight or night sweats, are not specific and might be easily attributed to another condition. Primary pulmonary tuberculosis is the first stage of the condition, and it may cause fever, dry cough and some abnormalities that may be noticed on a chest X-ray. In most cases, though, primary infections tend to cause no symptoms that people do not overcome. This condition resolves itself, although it returns in more than half of the cases. Tuberculosis causing lung disease may result in tuberculous pleuritis, a condition that may cause symptoms such as chest pain, nonproductive cough and fever. Moreover, infection with M. tuberculosis can spread to other parts of the body, especially in patients with a weakened immune system. This condition is referred to as miliary tuberculosis, and people contacting it may experience fever, weight loss, weakness and a poor appetite. In more rare cases, miliary tuberculosis can cause coughing and difficulty breathing.

2. Method

The present study was conducted at the department of Bioinformatics, Maulana Azad National Institute of Technology, Madhya Pradesh, India. Institutional ethical board approved the study and the informed consent was obtained from all the subjects before the commencement.

2.1. Database used for comparative study of 11 different M. tuberculosis species (table.1)

KEGG (**Kyoto Encyclopedia of Genes and Genomes**) is a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. The PATHWAY database records networks of molecular interactions in the cells, and variants of them specific to particular organisms.

KEGG connects known information on molecular interaction networks, such as pathways and complexes (this is the Pathway Database), information about genes and proteins generated by genome projects (including the gene database) and information about biochemical compounds and reactions (including compound and reaction databases). These databases are different networks, known as the protein network, and the chemical universe respectively. There are efforts in progress to add to the knowledge of KEGG, including information regarding ortholog clusters in the KO (KEGG Orthology) database.

2.2 Tools & Software's used for comparative study of 11 different M. tuberculosis species

2.2.1 BLAST Tool

BLAST, or **B**asic **L**ocal **A**lignment **S**earch **T**ool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. Different types of BLASTs are available according to the query sequences. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence. The BLAST program was designed by Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman, and Webb Miller at the NIH and was published in the Journal of Molecular Biology in 1990.

2.2.2 BlastClust Tool

BLASTClust is a program within the standalone BLAST package used to cluster either protein or nucleotide sequences. The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster. In the case of proteins, the blastp algorithm is used to compute the pairwise matches; in the case of nucleotide sequences, the Megablast algorithm is used. In the simplest case, BLASTClust takes as input a file containing catenated FASTA-format sequences, each with a unique identifier at the start of the definition line. BLASTClust formats the input sequence to produce a temporary BLAST database, performs the clustering, and removes the database at completion. Hence, there is no need to run formatdb in advance to use BLASTClust. The output of BLASTClust consists of a file, one cluster to a line, of sequence identifiers separated by spaces. The clusters are sorted from the largest cluster to the smallest.

BLASTClust accepts a number of parameters that can be used to control the stringency of clustering including thresholds for score density, percent identity, and alignment length. The BLASTClust program has a number of applications, the simplest of which is to create a non-redundant set of sequences from a source database. As an example, one might have a library of a few thousand short nucleotide sequence reads and wish to replace these with a non-redundant set. To produce the non-redundant set, one might use:

```
Blastclust -i infile -o outfile -p F -L .9 -b T -S 95
```

The sequences in "infile" will be clustered and the results will be written to "outfile". The input sequences are identified as nucleotide (-p F); "-p T", or protein, is the default. To register a pairwise match two sequences will need to be 95% identical (-S 95) over an area covering 90% of the length (-L .9) of each sequence (-b T). Using "-b F" instead of "-b T" would enforce the alignment length threshold on only one member of a sequence pair. The parameter "S", used here to specify the percent identity, can also be used to specify, instead, a "score density." The latter is equivalent to the BLAST score divided by the alignment length. If "S" is given as a number between 0 and 3, it is interpreted as a score density threshold; otherwise it is interpreted as a percent_identity_threshold.

To create a stringent non-redundant protein sequence set, use the following command line:

```
Blastclust -i infile -o outfile -p T -L 1 -b T -S 100
```

In this case, only sequences which are identical will be clustered together. The “blastclust.txt” file in the standalone BLAST package details the full range of BLASTClust parameters.

2.2.3 Downloading Genomes & Proteomes

Genome and proteomes of the 11 organisms is downloaded from the data repository KEGG (Kyoto Encyclopedia of Genes and Genomes)

Mtu (Mycobacterium tuberculosis H37Rv)
Mle (Mycobacterium Leprae)
Mbo (Mycobacterium Bovis)
Msm (Mycobacterium Smegmatis)
Mmi (Mycobacterium Marinum)
Mpa (Mycobacterium avian Para tuberculosis)
Mgi (Mycobacterium Flavescens (gilvum))
Mva (Mycobacterium Vanbaalenii)
Mkm (Mycobacterium sp KMS)
Mjl (Mycobacterium sp JLS)
Mmc (Mycobacterium sp MCS)

2.2.4 Installation of BLAST

The blast+ archive downloaded above contains a built-in installer. Accepting the license agreement after double-clicking, the installer will prompt for an installation directory. The default installation directory in this test case is "C:\blast-2.2.23+". Clicking the "Install" button, the installer will create a "doc" subdirectory containing a comprehensive user manual in pdf format, an "uninstaller" for future removal of the installation, and a "bin" subdirectory where the following BLAST programs and accessory utilities are kept.

blastdbcheck.exe blastdbcmd.exe blastdb_aliastool.exe
blastn.exe blastp.exe blastx.exe
blast_formatter.exe convert2blastmask.exe dustmasker.exe
legacy_blast.pl makeblastdb.exe makembindex.exe
psiblast.exe rpsblast.exe rpstblastn.exe
segmasker.exe tblastn.exe tblastx.exe update_blastdb.pl
windowmasker.exe

There is another subdirectory which is created during the installation of BLAST “data”. This directory contains some files like

asn2ff.prt   blosum45 blosum62 blosum80 bstdt.val ecnum_ambiguous ecnum_specificUniVec.nhr
sgmlbb.ent

These are supporting files required by the BLAST program to run successfully. So these files are copied and shifted into the subdirectory “bin”.

2.2.5 Output file of BlastClust

The output file generated by BLASTclust program on its execution contains cluters of sequences.
This output file ‘sequences_1.0_100_complete.ssv’ is filtered using MS Access Database management tool.
After using MS Access the following type of data is obtained:
Unique genes in every strain seperately
Total common clusters

2.2.6 Executing BLAST with composite Proteome file

For composite protein file (all_proteomes.txt) exactly same procedure is applied as that of composite genome file except for the blastclust command which is

```
blastclust -a 1 -i all_genomes.txt –p T   -o sequences_1.0_100_complete.ssv –S 100 -L 1.0
```

The parellel execution of BLASTclust for genome data is done to validate the final results we get from proteome analysis.

3. Results

3.1 Proteome Results

Starting from the set of values (100% identity and 100% sequence coverage) the set of values was lowered down upto (1% identity and 1% sequence coverage). (Table 2)
According to comparative analysis we are concerned with lowering down the number of unique proteins in m. tuberculosis so only relevant results are shown in form of tables while the whole results are presented (Fig.1) graphically.
The sizes of respective proteomes 11 different species of Mycobacterium are:
Mtu (Mycobacterium tuberculosis H37Rv) =3988
Mle (Mycobacterium Leprae) =1605
Mbo (Mycobacterium Bovis) =3918
Msm (Mycobacterium Smegmatis) =6716
Mmi (Mycobacterium Marinum) =5452
Mpa (Mycobacterium avian Para tuberculosis)=4350
Mgi (Mycobacterium Flavescens (gilvum)) =5579
Mva (Mycobacterium Vanbaalenii) =5979
Mkm (Mycobacterium sp KMS) =5975
Mjl (Mycobacterium sp JLS) =5739
Mmc (Mycobacterium sp MCS) =5615
Total = 54916
On further decrease in sequence coverage, no decrease in the number of unique proteins in mycobacterium tuberculosis was noticed.

3.2 Unique proteins in Mycobacterium tuberculosis

Mtu: Rv1515c
Mtu: Rv1509
Mtu: Rv1507A
Mtu: Rv2645
Mtu: Rv2658c
Mtu: Rv2653c
Mtu: Rv2654c
Mtu: Rv3599c

3.3 Genome results

Starting from the set of values (100% identity and 100% sequence coverage) the set of values was lowered down upto (1% identity and 1% sequence coverage). According to comparative analysis we are concerned with lowering down the number of unique genes in m. tuberculosis so only relevant results are shown in form of table 3.

According to comparative analysis 23 unique genes obtained are those genes which have no similarity with any other genes in the remaining 10 mycobacterium strains i.e. they are specific to mtu only.

The sizes of respective genomes are:

Mtu (Mycobacterium tuberculosis H37Rv) =4047
Mle (Mycobacterium Leprae) =2770
Mbo (Mycobacterium Bovis) =4001
Msm (Mycobacterium Smegmatis) =6938
Mmi (Mycobacterium Marinum) =5570
Mpa (Mycobacterium avian Para tuberculosis) =4399
Mgi (Mycobacterium Flavescens (gilvum)) =5669
Mva (Mycobacterium Vanbaalenii) =6139
Mkm (Mycobacterium sp KMS) =6079
Mjl (Mycobacterium sp JLS) =5845
Mmc (Mycobacterium sp MCS) =5698
Total = 54916

4. Discussion

11 Mycobacterium strains were taken for the analysis
Mtu (Mycobacterium tuberculosis H37Rv)
Mle (Mycobacterium Leprae)
Mbo (Mycobacterium Bovis)
Msm (Mycobacterium Smegmatis)
Mmi (Mycobacterium Marinum)
Mpa (Mycobacterium avian Para tuberculosis)
Mgi (Mycobacterium Flavescens (gilvum))
Mva (Mycobacterium Vanbaalenii)
Mkm (Mycobacterium sp KMS)
Mjl (Mycobacterium sp JLS)
Mmc (Mycobacterium sp MCS)

- On executing blastclust on the whole proteome data, finally 8 unique proteins in the mycobacterium tuberculosis were present i.e. the proteins that were specific in the mtu strain only.
- On executing blastclust on the whole genome data, finally 23 unique genes in the mycobacterium tuberculosis were presents i.e. the genes that were specific in the mtu strain only.

Behalf on above studty we can easily idntify unique strains of 11 different Mycobacterium species which only found Mycobacterium tuberculosis species.

5. Conclusion

The 8 unique proteins obtained in the comparative proteome analysis and 23 unique genes are obtained in comparative genome analysis. These 8 unique proteins were also present in the list of proteins which are obtained from expression of the 23 unique genes. This validation check gives an important support to the analysis that the work done is valid and has given accurate results. Out of these 8 unique proteins 5 are hypothetical proteins and 3 proteins are prophage proteins. So one or more unique proteins obtained must be having role in the specific behavior of mtu strain which is responsible for causing the tuberculosis disease. This gives a strong support to the DRUG DEVELOPMENT and can be very beneficial in the drug discovery area because these proteins can directly or indirectly act as sensitive drug targets for a drug for tuberculosis.

6. Acknowledgment

We are grateful and thank our subjects for their cooperation to conduct this study and making it a good success. We thank the department of bioinformatics computational biology lab, MANIT, Bhopal, M.P., India for supporting us to conduct such a study.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (October 1990). "Basic local alignment search tool". J Mol Biol 215 (3): 403–410. doi: 10.1006/jmbi.1990.9999. PMID 2231712

Centers for Disease Control and Prevention (CDC), Division of Tuberculosis Elimination. Core Curriculum on Tuberculosis: What the Clinician Should Know. 4th edition (2000). Updated August 2003.

Cole ST, Brosch R, Parkhill J, et al. (June 1998). "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence". Nature 393 (6685): 537–44. doi: 10.1038/31159. PMID 9634230

Camus JC, Pryor MJ, Médigue C, Cole ST (October 2002). "Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv". Microbiology (Reading, Engl.) 148 (Pt 10): 2967–73. PMID 12368430. http://mic.sgmjournals.org/cgi/pmidlookup?view=long&pmid=12368430

Casey, R. M. (2005). "BLAST Sequences Aid in Genomics and Proteomics". Business Intelligence Network. http://www.b-eye-network.com/view/1730

Flynn, Laurie (August, 21 1989). "Microsoft Waits on SQL Front Ends". InfoWorld: p. 109. http://books.google.pl/books?id=rjAEAAAAMBAJ.

http://www.australianprescriber.com/magazine/33/1/12/18/

http://www.sanger.ac.uk/Projects/M_tuberculosis/. Retrieved 2008-11-16.

http://jasoncartermd.com/resources/pdf/Latent%20TB%20Infection.pdf. , which cites Dolin, PJ; Raviglione, MC; Kochi, A (1994). "Global tuberculosis incidence and mortality during 1990–2000". Bull World Health Organ 72 (2): 213–20. PMC 2486541. PMID 8205640.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=2486541

"Introduction to importing and exporting data". Microsoft. http://office.microsoft.com/en-gb/access-help/introduction-to-importing-and-exporting-data-HA101790599.aspx. Retrieved 15 October 2010.

Jasmer RM, Nahid P, Hopewell PC (2002). "Clinical practice. Latent tuberculosis infection". N. Engl. J. Med. 347 (23): 1860–6. doi:10.1056/NEJMcp021045. PMID 12466511.

Konstantinos, a (2010). "Testing for tuberculosis". Australian Prescriber 33 (1): 12–18.

"Mycobacterium tuberculosis". Sanger Institute. 2007-03-29.

Mount, D. W. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Press. ISBN 978-087969712-9. http://www.bioinformaticsonline.org/

Oehmen, C.; Nieplocha, J. (August 2006). "ScalaBLAST: A scalable implementation of BLAST for high-performance data-intensive bioinformatics analysis". IEEE Transactions on Parallel & Distributed Systems 17 (8): 740–749.

Tuberculosis Symptoms From eMedicineHealth. Author: George Schiffman, MD, FCCP. Last Editorial Review: 1/15/2009

"Tuberculosis (TB) Symptoms". NHS Choices Tuberculosis. National Health Service (NHS) UK. http://www.nhs.uk/Conditions/Tuberculosis/Pages/Symptoms.aspx. Retrieved 17 April 2011.

WHO Tuberculosis Factsheet, World Health Organization, March 2010

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

ftp://ftp.genome.jp/pub/kegg/genes/organisms/

Tables

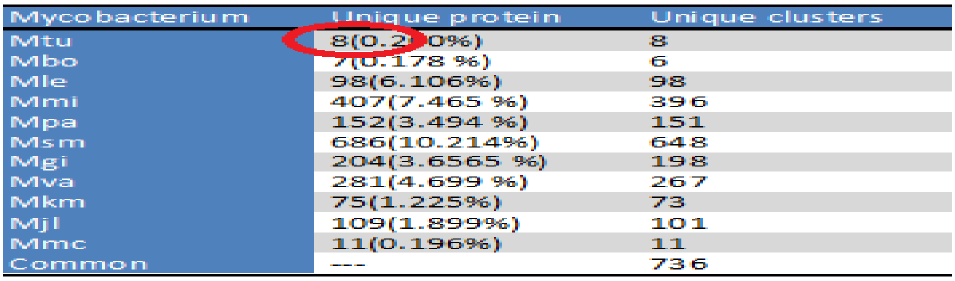
S.No.	ORGANISM	Short form	Number of genes present	Number of protein present
1.	M.tuberculosis	Mtu	4047	3988
2.	M.bovis	Mbo	4001	3918
3.	M.laprae	Mle	2770	1605
4.	M.smegmatis	Msm	6938	6716
5.	M.marinum	Mmi	5570	5452
6.	M.avium_paratuberculosis	Mpa	4399	4305
7.	M.gilvum	Mgi	5669	5579
8.	M.vanbaalenii	Mva	6136	5979
9.	Mycobacterium_JLS	Mjl	5845	5739
10.	Mycobacterium_MCS	Mmc	5698	5615
11.	Mycobacterium_KMS	Mkm	6079	5975

Table 1: Comparative study of Genome & Proteome genes of 11 different species of Mycobacterium

Results for 20% sequence coverage

Mycobacterium	Unique protein	Unique clusters
Mtu	33(0.827%)	33
Mbo	51(1.301%)	50
Mle	141(8.785%)	141
Mmi	477(8.749%)	463
Mpa	244(5.609%)	241
Msm	810(12.061%)	722
Mgi	276(4.947%)	270
Mva	332(5.552%)	320
Mkm	85(1.4225%)	83
Mjl	141(2.456%)	133
Mmc	18(0.321%)	18
Common	---	898

Results for 10% sequence coverage



Mycobacterium	Unique protein	Unique clusters
Mtu	8(0.2[illegible]0%)	8
Mbo	7(0.178 %)	6
Mle	98(6.106%)	98
Mmi	407(7.465 %)	396
Mpa	152(3.494 %)	151
Msm	686(10.214%)	648
Mgi	204(3.6565 %)	198
Mva	281(4.699 %)	267
Mkm	75(1.225%)	73
Mjl	109(1.899%)	101
Mmc	11(0.196%)	11
Common	---	736

Table 2: The marked section represents the unique proteins in the strain mycobacterium tuberculosis. These are the proteins which are not present in any of the remaining 10 strains i.e. they are specific to m. tuberculosis only (Acc. - to Proteome results)

Results for 20% sequence coverage

Mycobacterium	Unique protein	Unique clusters
Mtu	40(%)	35
Mbo	52(%)	52
Mle	143(%)	141
Mmi	479(%)	467
Mpa	247(%)	245
Msm	821(%)	730
Mgi	276(%)	271
Mva	335(%)	325
Mkm	91(%)	86
Mjl	145(%)	134
Mmc	29(%)	19
Common	---	919

Table for 10% sequence coverage

Mycobacterium	Unique protein	Unique clusters
Mtu	23(%)	15
Mbo	9(%)	11
Mle	102(%)	99
Mmi	412(%)	398
Mpa	161(%)	165
Msm	695(%)	654
Mgi	207(%)	203
Mva	289(%)	272
Mkm	77(%)	74
Mjl	111(%)	107
Mmc	15(%)	13
Common	---	741

Table 3: These 23 unique genes obtained are those genes which have no similarity with any other genes in the remaining 10 mycobacterium strains i.e. they are specific to mtu only. (Acc. to Genome results)

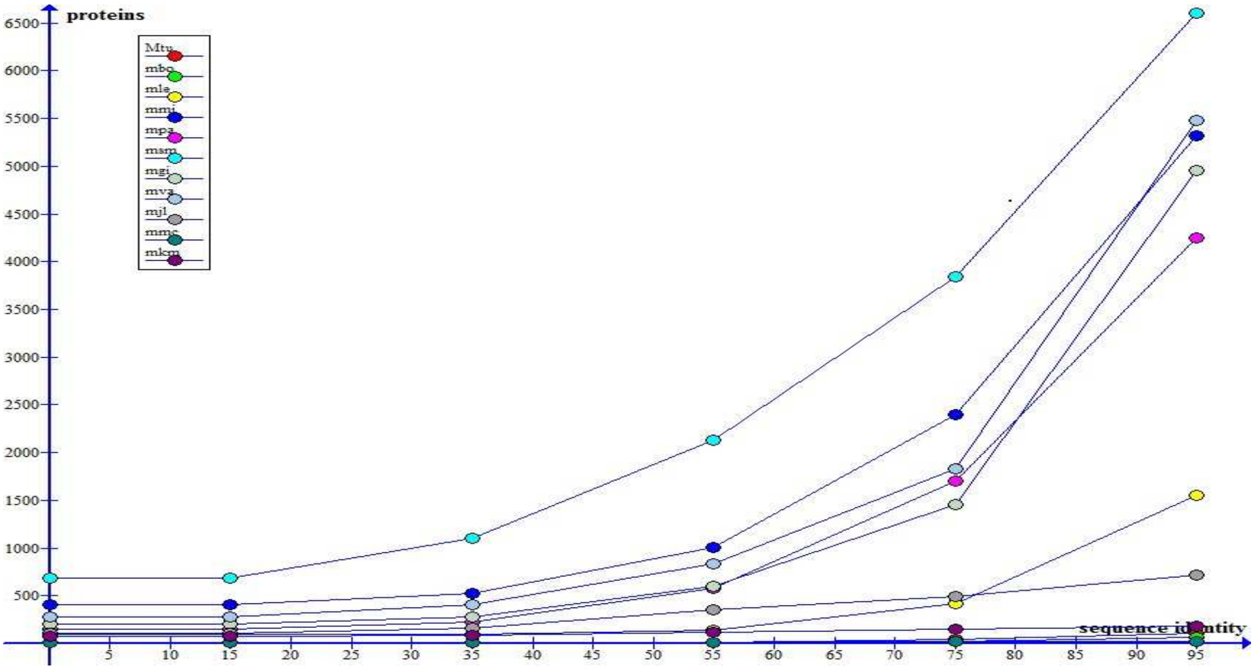


Figure 1: Graph shows the number of unique proteins in each mycobacterium strain against the sequence coverage parameter (% L value)