



MYCOBACTERIAL SPECIES AS CASE- STUDY OF COMPARATIVE GENOME ANALYSIS

F. ZAKHAM^{1,2,3}, L. BELAYACHI³, D. USSERY⁴, M. AKRIM², A. BENJOUAD³,
R. EL AOUAD² AND M. M. ENNAJI¹ ✉

1 Laboratory of Virology and Hygiene & Microbiology, University Hassan II Mohammedia-Casablanca School of Science and Techniques.. BP 146, Mohammedia, 20650, Morocco.

2 National Institute of Hygiene, Laboratory of Molecular Biology, Rabat, Morocco.

3 Laboratory of Immunology and Biochemistry, School of Science. University Mohammed V- Rabat. Morocco.

4 Centre for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark.

Abstract

The genus *Mycobacterium* represents more than 120 species including important pathogens of human and cause major public health problems and illnesses. Further, with more than 100 genome sequences from this genus, comparative genome analysis can provide new insights for better understanding the evolutionary events of these species and improving drugs, vaccines, and diagnostics tools for controlling Mycobacterial diseases. In this present study we aim to outline a comparative genome analysis of fourteen Mycobacterial genomes: *M. avium* subsp. *paratuberculosis* K-10, *M. bovis* AF2122/97, *M. bovis* BCG str. Pasteur 1173P2, *M. leprae* Br4923, *M. marinum* M, *M. sp.* KMS, *M. sp.* MCS, *M. tuberculosis* CDC1551, *M. tuberculosis* F11, *M. tuberculosis* H37Ra, *M. tuberculosis* H37Rv, *M. tuberculosis* KZN 1435, *M. ulcerans* Agy99, and *M. vanbaalenii* PYR-1. For this purpose a comparison has been done based on their length of genomes, GC content, number of genes in different data bases (Genbank, Refseq, and Prodigal). The BLAST matrix of these genomes has been figured to give a lot of information about the similarity between species in a simple scheme. As a result of multiple genome analysis, the pan and core genome have been defined for twelve Mycobacterial species. We have also introduced the genome atlas of the reference strain *M. tuberculosis* H37Rv which can give a good overview of this genome. And for examining the phylogenetic relationships among these bacteria, a phylogenetic tree has been constructed from 16S rRNA gene for tuberculosis and non tuberculosis Mycobacteria to understand the evolutionary events of these species.

Key words: *Mycobacterium*, Bioinformatics, comparative genomics.

Article information's

Received on November 30, 2010

Accepted on February 8, 2011

Corresponding author

Professor Moulay Mustapha ENNAJI

Laboratory of Virology and Hygiene & Microbiology,
University Hassan II .Mohammedia- Casablanca. School of
Sciences and Techniques. BP 146, Mohammedia, 20650,
Morocco.

Fax: 212 5 23 31 53 53

E-mail: m.ennaji@yahoo.fr

Abbreviations: AIDS: Acquired Immunodeficiency Syndrome; DNA: Deoxyribonucleic Acid; M: *Mycobacterium*; rRNA: ribosomal Ribonucleic Acid, BLAST: Basic Local Alignment Search Tool; NCBI: National Center for Biotechnology Information.

INTRODUCTION

The Mycobacteria have been classified into the family Mycobacteriaceae within the order Actinomycetales based on similarities in morphological, physiological, and biochemical characters (1). This genus represents more than 120 species (16) including important pathogens of human and cause major public health problems and illnesses, such as tuberculosis, leprosy, and Buruli ulcer, and emerging diseases induced by atypical Mycobacteria which infect patients with AIDS or other immunocompromised individuals (2).

The huge amount of DNA sequence information has enriched the development of comparative genomics which can be exploited in the field of infectious diseases research (25),

additionally there are many tools available in different web sites to facilitate the comparison of different microbial genomes (19).

Further, with more than 100 Mycobacterial genome sequences, comparative genome analysis provides exceptional new insights into the biology of these worldwide important pathogens.

This analysis gives accurate genomic information in respect of the genetic differences between these species, which are useful for better understanding of their evolution, pathogenesis mechanism, basis for virulence, and consequently to deal with the scientific priorities of better drugs, vaccines, and diagnostics tools for Mycobacterial diseases (6, 9). In this study we aim to outline a comparison between fourteen Mycobacterial genomes, based on the results of bioinformatics methods (Bio Linux) in comparing genome sequences which permits the study of more complex evolutionary events.

Most of these genomes are belonging to pathogenic Mycobacteria (*M. avium subsp. paratuberculosis* K-10, *M. bovis* AF2122/9, *M. leprae* Br4923, *M. tuberculosis* CDC1551, *M. tuberculosis* F11, *M. tuberculosis* H37Ra, *M. tuberculosis* H37Rv, *M. tuberculosis* KZN 1435, and *M. ulcerans*) and the attenuated vaccine strain *M. bovis* BCG str. Pasteur 1173P2, with some of the free living Mycobacterial genomes (*M. marinum* M, *M. sp.* KMS, *M. sp.* MCS, and *M. vanbaalenii* PYR-1) to give an idea about the gene deletion on the term of evolution for those pathogenic strains (5).

We have also introduced the genome atlas of the reference strain *M. tuberculosis* H37Rv which can give a good overview of this genome (11, 19).

As a result of multiple genome analysis, the pan and core genome will be defined for some of Mycobacterial species, which is extremely useful to investigate the true diversity within and between bacterial species (15, 19).

And for examining the phylogenetic relationships among these bacteria, a phylogenetic tree has been constructed from 16S rRNA gene, for tuberculosis and non tuberculosis Mycobacteria to understand the evolutionary events of these species, as stated by other authors (4, 16).

MATERIALS AND METHODS

The bio Linux system issued from Linux, and a lot of additional programs that are often used within bioinformatics and computational molecular biology, organized to use easily that made Bio Linux as the efficient way to deal with and to handle large amounts of biological data.

Linux is an operative system similar to UNIX, which can be considered as a programming language by using the command interpreter Shell (bash) that can give us the opportunity to control executing programs and collect their results into files; furthermore this system has been used in this study to figure the BLAST matrix and the Pan and Core-genome curve.

The main source of information is (<http://www.cbs.dtu.dk/services/GenomeAtlas>) (22) which is a web based user interface linked to the Genome Atlas Database, can be accessed easily via this web site, in fact there are seven different structural atlases presented (such as Base Atlas, Structure Atlas, and Repeat Atlas, etc).

The Genome Atlas gives us a rapid overview of a sequenced bacterial genome in a simple scheme (10, 19). A number of parameters are calculated for the DNA double helix based on the nucleotide sequence (11). These parameters belong to three categories: repeats, structural parameters, and the base composition.

Web sites providing access to the main DNA sequence databases used in this paper are:

- <http://www.ncbi.nlm.nih.gov/> GenBank at NCBI.
- <http://www.ncbi.nlm.nih.gov/> Ref Seq at NCBI.
- <http://prodigal.ornl.gov/>(Prokaryotic Dynamic Programming Gene finding Algorithm).

RESULTS AND DISCUSSION

Genome length

As can be observed for the fourteen Mycobacterial genomes in Table 1, even for the different strains of the same species, there can be a considerable size variation. And this is known in bacteria as Ussery and Hallin, 2004 (17), mentioned in their study of the genomes of different strains of *Prochlorococcus marinus*. The smallest mycobacterial genome in Table1 is the genome of *M. leprae* 3.268Mb, the infectious agent of leprosy (an intracellular bacteria), which can be explained by an extensive genome downsizing or genome decay occurred during the evolution of *M. leprae* (5).

On other hand the *M. marinum* has a genome size of 6.65983 (chromosome of 6.63683 Mb plus plasmid of 0.023Mb), a free living bacteria which need to have a more extensive adaptation potential, reflected by a larger genome, in contrast of the pathogenic bacteria which can obtain their needs from the host (20).

GC content and GC skew

The GC content is an important tool to classify bacteria and it is considered to be an indicator of some physical properties of bacterial genome.

As shown in Table 1, from the data which were obtained in this study Mycobacterial

Table1. Summary of the mycobacterial genomes discussed in this paper (genome length, GC content, and No of genes in different data bases).

Genome	Genome length	Gene No Genbank	Gene No Refseq	Gene No Prodigal	GC content	Accession No
<i>Mycobacterium avium subsp. paratuberculosis K-10</i>	4.82978	4350	4350	4350	69.0%	AE016958
<i>Mycobacterium bovis AF2122/97</i>	4.34549		3918	3918	65.0%	BX248333
<i>Mycobacterium bovis BCG str. Pasteur 1173P2</i>	4.37452	3949	3949	3949	65.0%	AM408590
<i>Mycobacterium leprae Br4923</i>	3.26807	1604			57.0%	FM211192
<i>Mycobacterium marinum</i>			5452	5452		
Chromosome	6.63683				65.0%	CP000854
Plasmid pMM23	0.023				67.0%	CP000895
<i>Mycobacterium sp. KMS</i>		5975	5975	5975		
Chromosome	5.73723				68.0%	CP000518
Plasmid pMKMS02	0.22				66.0%	CP000520
Plasmid pMKMS01	0.3				65.0%	CP000519
<i>Mycobacterium sp. MCS</i>		5615	5615	5615		
Chromosome	5.70545				68.0%	CP000384
Plasmid1	0.215075				66.0%	CP000385
<i>Mycobacterium tuberculosis CDC1551</i>	4.40384	4189	4189	4189	65.0%	AE000516
<i>Mycobacterium tuberculosis F11</i>	4.42443	3941	3941	3941	65.0%	CP000717
<i>Mycobacterium tuberculosis H37Ra</i>	4.4	4034	4034	4034	65.0%	CP000611
<i>Mycobacterium tuberculosis H37Rv</i>	4.4	3991	3988	3988	65.0%	AL123456
<i>Mycobacterium tuberculosis KZN 1435</i>	4.4	4060	4059	4059	65.0%	CP001658
<i>Mycobacterium ulcerans Agy99</i>		4241	4241	4241		
Chromosome	5.63161				65.0%	CP000325
Plasmid pMUM001	0.17				62.0%	BX649209
<i>Mycobacterium vanbaalenii PYR-1</i>	6.5	5979	5979	5979	67.0%	CP000511

species, are GC rich bacteria and it appears to be normal due to the large genome size of these species as stated in different references (14, 19, 20, and 21).

For most bacterial groups, there is a trend for G's to be biased towards the replication leading strand and a bias for C's towards the other strand, and generally there is a tendency of GC rich region toward the origin of replication and a trend of AT rich region around the replication terminus as it can melt easily (18, 19).

Number of genes in different data bases

After comparing the genes number in different data bases: GenBank, ResSeq, and Prodigal, we approximately obtained the same genes number of each species in different data bases, except in the cases of *M. tuberculosis H37Rv* (3991 in GenBank and 3988 in both other databases), and *M. tuberculosis KZN 1435* (4060 in GenBank and 4059 in both other databases).

The different genes number, basically depend on the source of the data, in Genban

genes are reported by the authors around the world, and are submitted to the NCBI data repository, but the RefSeq genes are derived from the primary GenBank and have been curated by RefSeq and This means that there can be differences between the two databases, although GeneBank and RefSeq are data bases at NCBI (23).

But prodigal is an automated gene prediction program which can analyze an entire microbial genome in 30 seconds or less (<http://prodigal.ornl.gov/>) (24).

Genome Atlas

The Genome Atlas of the reference strain chromosome, *M. tuberculosis H37Rv* Fig 1 shows, three outer circles which are representing DNA structural properties: *intrinsic DNA curvature* in the outermost followed by *stacking energy* and *position preference*.

DNA curvature and base-stacking measure structural properties of DNA alone, whereas position preference measures the ability of the DNA helix to be bent by proteins (19, 22).

In Fig 1 it is also clear that the genes are more or less randomly distributed between the leading and lagging strands, blue on the positive strand (Coding sequence proteins CDs+) and red on the negative strand (non Coding sequence proteins CDs-) as re-annotated by Campaus *et al* in 2002 (3) for this genome.

The two other circles of the genome show the presence of two kinds of repeats:

The global direct repeats, which are relatively common for this micro organism, the colours are scaled such that they vary from white for less than 50% to darkly colored for more than 75% identity. The next circle shows the global inverted repeats, on the same scale, we noted that many of the inverted repeats match the direct repeats; simply because, the sequence is repeated both in the forward and reverse directions. Mycobacteria have relatively many global direct repeats but few global inverted repeats (19).

In *M. tuberculosis H37Rv* chromosome there is a Bias of G's towards the leading replication strand and the A's are biased towards the replication lagging strand of the genome, which clearly shows the DNA replication origin near the 12 o'clock position.

Finally, the deviation of AT content from the chromosomal average percentage AT is plotted, ranging from 40% to 60% AT, with 50% AT in the middle; we remarked in the genome atlas of this organism that this germ is GC-rich (blue).

As showed in Fig 2, the *Base skew*, towards the leading and lagging strand influences the structure of the DNA, the AT and GC skew are pointing in apposite directions with a clear bias of G's towards the leading replication strand, and the A's towards the lagging strand.

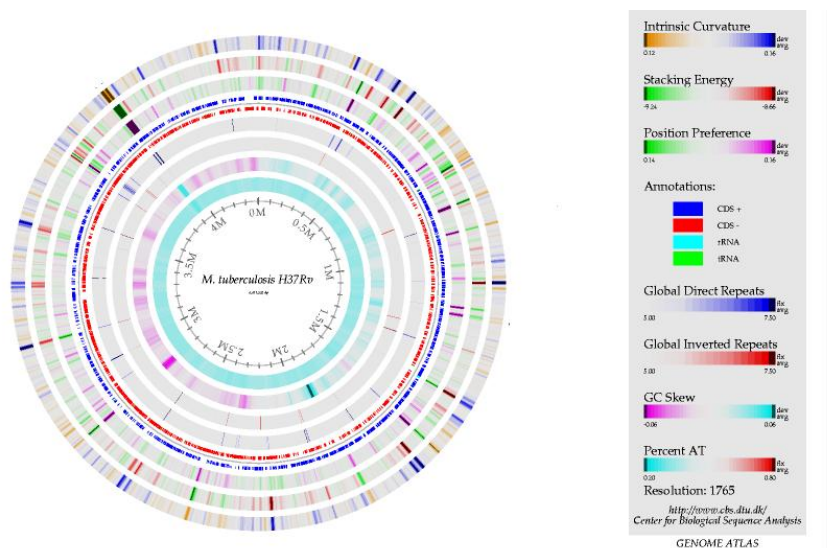


Figure 1. Genome Atlas of chromosome of *M. tuberculosis H37Rv*. The outer three lanes represent physical properties of the DNA (intrinsic curvature, stacking energy and position preference). Following the two lanes with annotated genes for the positive and negative strand, two lanes show the presence of repeats, and the last two lanes are taken from a Base Atlas (<http://www.cbs.dtu.dk/services/GenomeAtlas>).

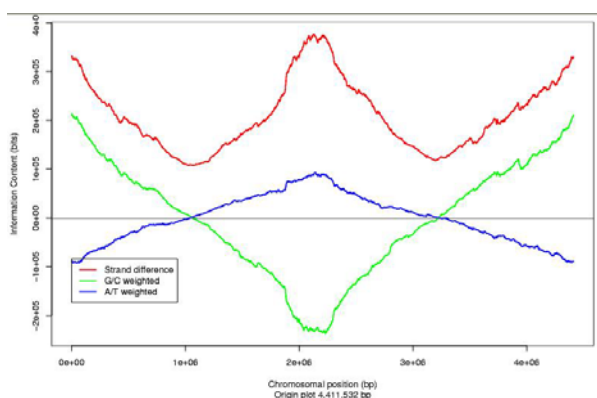


Figure 2. The *M. tuberculosis H37Rv* genome at the displays a clear bias of G's towards the leading replication strand, and the A's towards the lagging strand (<http://www.cbs.dtu.dk/services/GenomeAtlas>).

Blast Matrix

BLAST matrix is one of the important methods of comparing bacterial genomes (19), which plots the number of hits in a given set of proteomes against each other. In this matrix, the similarity between fourteen Mycobacterial genomes have been introduced with inter and intra-species comparison Fig.3. The intra species comparison of *M. tuberculosis* strains (*M. tuberculosis CDC1551*, *M. tuberculosis F11*, *M. tuberculosis H37Ra*, *M. tuberculosis H37Rv*, *M. tuberculosis KZN 1435*) represents a good example of similarity among intra-species, which ranged between; 96.1% -99.2 (which is darker green, indicative of a higher degree of similarity within these strains).

A high similarity between these strains and other *M. tuberculosis* complex members (*M. bovis AF2122/97*, and *M. bovis BCG str. Pasteur 1173P2*) has been observed, which ranged between (93.0%- 94.8%). In contrast with other Mycobacterial species, this similarity has been decreased especially in the free living Mycobacteria (indicated by pale colours).

Pan and core Genome

The core genome and a pan-genome are hypothetical combinations of genes that describe the full genetic collection of an investigated population as in the pan-genome (15), or the hypothetical set of genes that will always be present in the investigated population as the core genome.

Here we have defined the pan-genome and core genome of 12 Mycobacterial genomes of (*M. bovis AF2122/97*, *M. bovis BCG str. Pasteur 1173P2*, *M. leprae Br4923* *M. marinum M.*, *M. sp. KMS*, *M. sp. MCS*, *M. tuberculosis CDC1551*, *M. tuberculosis F11*, *M. tuberculosis H37Ra*, *M.*

tuberculosis H37Rv, *M. tuberculosis KZN 1435*, *M. ulcerans Agy99*, and *M. vanbaalenii- PYR1*, which represent more than 12 500 genes for the pan genome shown in Fig.4., which are larger three times than are present in the reference strain *M. tuberculosis H37Rv*.

Some of these genes are only found in one or a few genomes, whereas others are present in most or all of the sequenced isolates. However, if we look at the pan-genome of an individual species, for instance, *M. bovis AF2122/97* which infects a very wide variety of mammalian species including humans, or *M. bovis BCG str. Pasteur 1173P2* which is an attenuated variant of *M. bovis* (7), their species' pan-genome would contain approximately the same genes of their typical genomes.

As expected, there is a jump when adding a new species, in our case *M. leprae* which is an intracellular bacterium induces the leprosy in humans.

The core genome of a single *Mycobacterium sp* contains approximately 550 genes. Thus, approximately a 1/7of the genes are indispensable genes. Since the core genome covers all genes conserved between all (sequenced) members of a species and will also contain all genes that are essential for all life forms, such as genes coding for transcription, translation, replication, and essential metabolism proteins (19).

One thing to note is that the pan-genome of *M. bovis AF2122/97*, and *M. bovis BCG str.* is similar to the core genome which is remarked by the two adherent curves in this part of the graph. Another remarkable point that the five *M. tuberculosis* strains have the same core and pan-genome presented by the straight curve above each strain which is in correlation with the results of the BLAST Matrix and the phylogentic analysis(as we will see later) on the term of similarity between these strains. A significant jump in the pan-genome is introduced when moving to the last two genomes of *M. ulcerans Agy99*, and *M. vanbaalenii- PYR1*, although the core genome is still the same number of genes in both of them.

Phylogenic relation

Phylogenetic trees are often constructed from 16S rRNA genes. These are genes that code for a 16S rRNA that is found within the small subunit of the ribosome.

The 16S rRNA gene sequence is about 1,550 bp long and is composed of both variable and conserved regions. The gene is large enough,

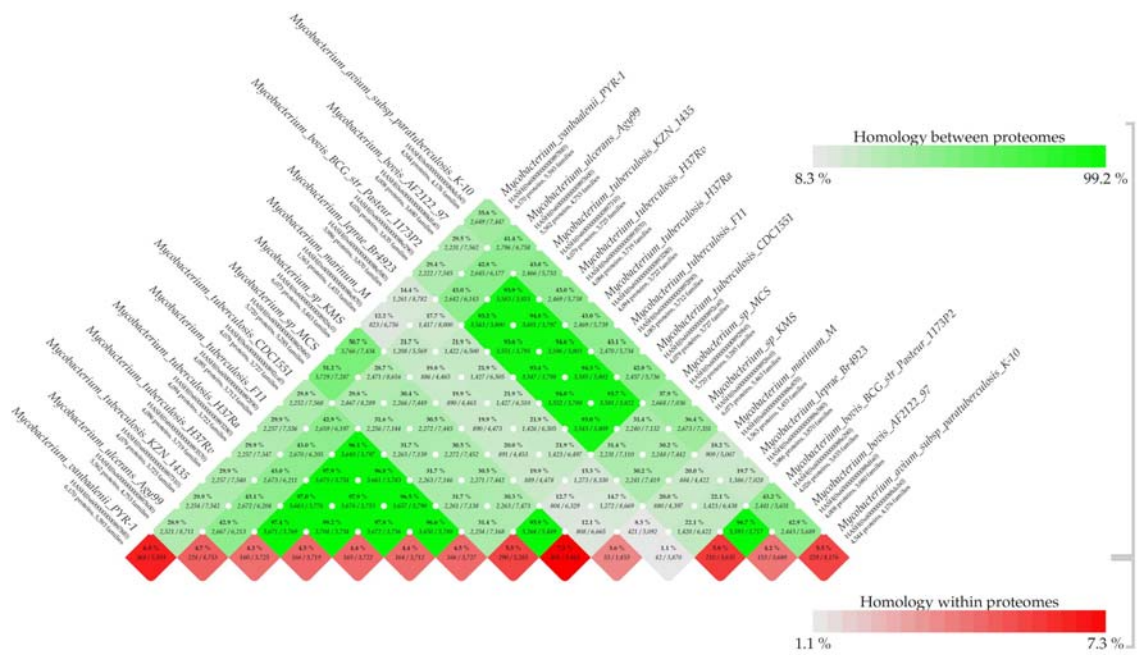


Figure 3. The blast matrix of fourteen mycobacterial genomes.

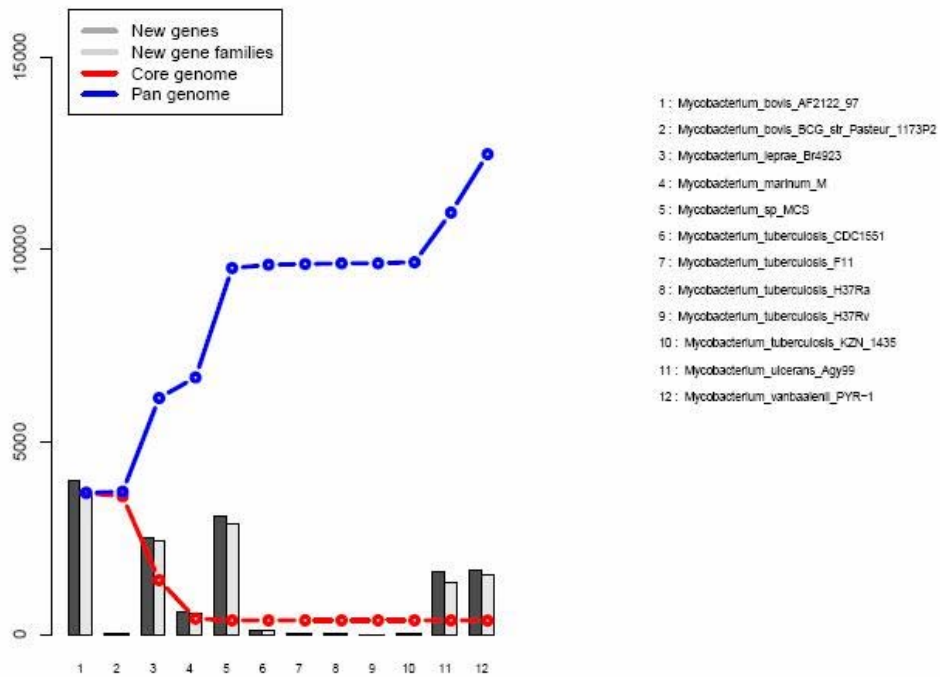


Figure 4. The pan-genome (blue line) and core genome (red line) for Mycobacterium. The number of discovered novel genes (dark bars) and novel gene families (light-grey bars) are also shown for each added genome.

with sufficient inter specific polymorphisms of 16S rRNA gene, to provide distinguishing and statistically applicable measurements (4). Whole-genome analysis has been tried, but it is quite difficult because the genomes have different sizes; however, it has been observed that the trees based on whole-genomic analysis and the 16S rRNA gene trees are similar (8) and for this reason these trees are used to investigate long-distance evolutionary relationships.

From Fig 5 we found that the species of *M. vanbaalenii*- *PYR1* and *M. sp MCS* (their genomes length are 6.5, 5.9, respectively), which are environmental micro organisms, able to metabolize a wide range of polycyclic aromatic hydrocarbons (12), are separated by a complete outlier. The horizontal line at the top (in this case, 0.005) is used to provide a rough measure of genetic distance.

In contrast of the pathogenic Mycobacteria which represent the other outlier, as we mentioned in the first part of this paper, a lot of these bacteria had undergone to genome reduction due to their parasitic life style which offer their needs from the host (20).

The strains of *M. tuberculosis* MTB (*M. tuberculosis CDC1551*, *M. tuberculosis H37Ra*, *M. tuberculosis H37Rv*, *M. tuberculosis KZN 1435*), are presented by a separated clade, which indicate, that these bacteria are sharing the same ancestor in their evolutionary events. For the other pathogenic non tuberculosis mycobacteria (*M. leprae Br4923*, *sp. KMS*, *M. sp. MCS*, *M. ulcerans Agy99*), which are displayed by different clades, representing the relatedness between these species and the members of *MTB* as proved by different studies on comparative genome analysis, for example Li et al in 2005(13) identified more than 3,000 genes of *M. avium* subspecies *paratuberculosis* strain K-10,(the causative agent of Johne's disease in cattle) are homologs to the human pathogen *M. tuberculosis*. Even for *M. leprae* (5) which suffered of an extensive genome downsizing.

Concluding Remarks

With this wealthness of knowledge of bacterial genomes, there's a special need for tools of comparison, visualization and analyzing these huge amount of sequence information.

Especially in the case of pathogenic micro organisms which represent a real threat of humanity such as *M. tuberculosis*, and *M. leprae* the causative agents of tuberculosis and leprosy, respectively. Indeed there is an important

necessity for improving techniques of diagnosis and controlling these diseases.

The software of bioinformatics and computational sequence analysis are potential and powerful tools to view and to analyse sequence data generated from databases.

Furthermore the comparative genomic tool allows direct, and interactive, comparisons of multiple genomes/sequences. This enables us to exploit the growing number of genomes from closely related organisms to look at genome architecture and evolution.

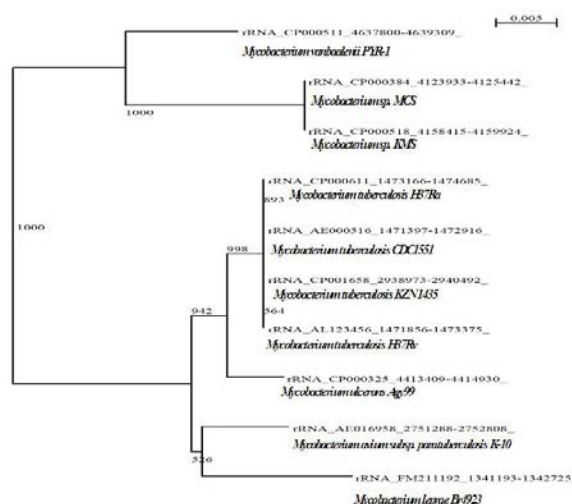


Figure 5. Phylogenetic tree based on 16S rRNA of 10 mycobacterial species constructed by a neighbor-joining algorithm.

Acknowledgments - This study was held at the National Centre of Scientific Research & Techniques (CNRST), in Rabat, Morocco, in collaboration with School of Science and Techniques- Mohammedia (Laboratory of Virology and Hygiene & Microbiology), School of Science-Rabat, and National Institute of Hygiene-Rabat.

The authors thank Pr. Peter Dawyndt and Karen Langensen for their generous assistance with informatics. Special thanks to Pr. Mohammed Amar the head of Microbiology department and Pr. EL Mustafa El Fahime the head of Molecular Biology department in the CNRST.

REFERENCES

1. Baron, S., In: *Medical Microbiology*, Galveston (TX): University of Texas. Medical Branch. 1996. NCBI bookshelf.
2. Brooks, G. F., Butel, J. S., Morse, S. A., In: *Jawetz, Melnick, Adelberg's Medical Microbiology*, 21st edition, Lange Medical book. 1998, PP. 279-283.
3. Camus, J.C., Pryor, M. J., Medigue, C. and Cole, S. T., Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*. 2002, **148**: 2967-2973.
4. Clarridge III, J. E., Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. 2004, *Clin Microbiol revidw*, 840-862.

5. Cole, S.T., Eiglmeier, K. , Parkhill, J, James K. D., Thomson, N. R., Wheeler, P. R., Honore , N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R. M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J. , Moule, S., Murphy, L., Oliver, K., Quail, M. A., Rajandream, M.-A., Rutherford, K. M., Rutter, S., Seeger, K., Simon, S. , Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J. R. and Barrell, B. G., Massive gene decay in the leprosy bacillus. *Nature*. 2001, **409**: 1007-1011.
6. Cole, S. T., Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J*. 2002a, **20 (Suppl 36)**: 78–86.
7. Cole, S. T., Comparative and functional genomics of the *M. tuberculosis* complex. *Microbiology*. 2002 b, **148**:2919–2928.
8. Eisen, A. J., Assessing evolutionary relationships among microbes from whole-genome analysis. *Current Opinion in Microbiol*. 2000, **3**:475–480.
9. Fraser, C. M., Eisen, J. A. and Salzberg, S. L., Microbial genome sequencing. *Nature*. 2000, **406**: 799-803.
10. Hallin, P. F. and Ussery, D. W., CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics*. 2004, **20**:3682–3686.
11. Jensen, L.J., Friis, C., and Ussery, D.W., Three views of microbial genomes. *Res Microbiol*. 1999, **150**: 773–777.
12. Kim, S. J., Kweon, O., Freeman, J. P., et al, Molecular Cloning and Expression of Genes Encoding a Novel Dioxygenase Involved in Low- and High-Molecular-Weight Polycyclic Aromatic Hydrocarbon Degradation in *M. vanbaalenii* PYR-1. *Applied & Environmental Microbiol*. 2006, **72**: 1045–1054.
13. Li L, Bannantine J P., Zhang et al, 2005. The complete genome sequence of *M. avium* subspecies paratuberculosis. *PNAS*, **102**: 12344–12349.
14. Mann, S., Chen, Y. P., Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics*. 2010, **95**:7–15.
15. Read, T. D. and Ussery, D. W., Opening the pan-genomics box. *Current Opinion in Microbiol*. 2006, **9**:496–498.
16. Tortoli, E., The new mycobacteria: an update. *FEMS Immunol Med Microbiol*. 2006, **48**: 159–178.
17. Ussery, D. W. and Hallin, P. F., Genome Update: length distributions of sequenced prokaryotic genomes. *Microbiol Comment*. 2004a, **150**: 513-516.
18. Ussery, D. W. and Hallin, P. F., Genome Update: AT content in sequenced prokaryotic genomes. *Microbiol Comment*. 2004b, **150**:749-752.
19. Ussery, D. W., Borini, S., Wassenaar, T. M., In: Computing for Comparative Microbial Genomics: Bioinformatics for Microbiologists (Computational series). London, Verlag: Springer. 2009.
20. Wassenaar T.M, Bohlin J, Binnewies T.T and Ussery D.W., Genome Comparison of Bacterial Pathogens. *Microbial Pathogenomics*. 2009, **6**: 1–20.
21. Wassenaar T. M, Binnewies T. T, Hallin P. F, Ussery D. W. 2010 Tools for Comparison of Bacterial Genomes. Handbook of Hydrocarbon and Lipid Microbiology. Springer-Verlag Berlin Heidelberg.
22. www.cbs.dtu.dk/services/GenomeAtlas.
23. www.ncbi.nlm.nih.gov.
24. www.prodigal.ornl.gov.
25. Zhang R, Zhang C-T. The impact of comparative genomics on infectious disease research. *Microbes and Infection*. 2006, **8**: 1613-1622.