Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/IJMYCO

Comparative proteomic analysis of *Mycobacterium tuberculosis* strain $H_{37}Rv$ versus $H_{37}Ra$

Lingaraja Jena [a], *Supriya Kashikar* [b], *Satish Kumar* [a], *Bhaskar C. Harinath* [a,*]

[a] *Bioinformatics Centre, Biochemistry & JBTDRC, MGIMS, Sevagram, Maharashtra, India*
[b] *GeNext Genomics Pvt. Ltd., Nagpur, Maharashtra, India*

ARTICLE INFO

Article history:
Received 7 September 2013
Received in revised form
11 October 2013
Accepted 14 October 2013
Available online 1 November 2013

Keywords:
MTB
Proteome
Virulence
Comparative proteomics
Bioinformatics

ABSTRACT

Background: *Mycobacterium tuberculosis* (MTB) $H_{37}Ra$ is an attenuated tubercle bacillus closely related to the virulent type strain MTB $H_{37}Rv$. In spite of extensive study, variation in virulence between the MTB $H_{37}Rv$ and MTB $H_{37}Ra$ strains is still to be understood. The difference in protein expression or structure due to mutation may probably be an important factor for the virulence property of MTB $H_{37}Rv$ strain.

*Methods:*In this study, a whole proteome comparison between these two strains was carried out using bioinformatics approaches to elucidate differences in their protein sequences.

*Results:*On comparison of whole proteome using NCBI standalone BLAST program between these two strains, 3759 identical proteins in both the strains out of 4003 proteins were revealed in MTB $H_{37}Rv$ and 4034 proteins were revealed in MTB $H_{37}Ra$; 244 proteins of MTB $H_{37}Rv$ and 260 proteins of MTB $H_{37}Ra$ were found to be non-identical. A total of 172 proteins were identified with mutations (Insertions/deletions/substitutions) in MTB $H_{37}Ra$ while 53 proteins of MTB $H_{37}Rv$ and 85 proteins of MTB $H_{37}Ra$ were found to be distinct. Among 244 non-identical proteins, 19 proteins were reported to have an important biological function; In this study, mutation was shown in these proteins of MTB $H_{37}Ra$.

*Conclusion:*This study reports the protein differences with mutations between MTB $H_{37}Rv$ and $H_{37}Ra$, which may help in better understanding the pathogenesis and virulence properties of MTB $H_{37}Rv$.

Introduction

Tuberculosis (TB) is a complex disease caused by *Mycobacterium tuberculosis* (MTB), which has evolved with highly successful mechanisms to equivocate host defenses and existing classes of antibiotics. Decades after the discovery of MTB, TB remains a major cause of morbidity and mortality in many developing countries. One third of the World's population is considered to be infected with MTB, with 8.7 million new patients and 1.4 million deaths in the year 2011, including 1 million deaths among HIV-negative and 430,000 HIV-positive individuals [1]. Multi-drug-resistant strains of this pathogen, emerging in association with HIV, have added a frightening dimension to the problem [2]. Outbreaks of extensively drug-resistant (XDR) TB have also been an increasing threat in certain regions around the world [3]. Despite abundant research on MTB diagnostics, vaccinations and treatments, this disease poses a considerable risk in many developed countries. MTB is very virulent, but there has been no simple answer found yet for what makes MTB so virulent. Historically, MTB $H_{37}Ra$ is the avirulent counterpart of the virulent strain MTB $H_{37}Rv$, and both strains were

* Corresponding author. Address: JB Tropical Disease Research Centre, Mahatma Gandhi Institute of Medical Sciences, Sevagram 442 102, Wardha, Maharashtra, India. Tel.: +91 7152 284341/355/262/303; fax: +91 7152 284038.
E-mail addresses: bc_harinath@yahoo.com, info@jbtdrc.org (B.C. Harinath).
2212-5531/$ - see front matter
http://dx.doi.org/10.1016/j.ijmyco.2013.10.004

derived from their virulent parent strain H37 discovered in 1935 by William Steenken through a process of aging and dissociation from *in vitro* culture [4]. These strains are phenotypically and genotypically different from each other, but the virulence power is different among these strains, which could probably be owing to a difference in protein expression. Owing to the advancement of Bioinformatics and the genome sequencing project, whole genome and proteome sequences of both the strains are available in the public domain. Current evidence suggests that as a species, MTB exhibits very little genomic sequence diversity [5,6]. MTB $H_{37}Rv$ is virulent and susceptible to most of the anti-tuberculous drugs, while MTB $H_{37}Ra$ is an avirulent strain and the MTB KZN (KwaZulu-Natal, South Africa) strain is resistant to different drugs like isoniazid, rifampicin, kanamycin, ofloxacin, ethambutol, pyrazinamide, etc. [7]. This may be due to a genetic mutation resulting in the generation of mutated proteins. Therefore, there is a need for genomic as well as proteomic analysis among different strains of MTB to understand the variation among them.

Many tools have also been developed for the complete determination of the genome sequence of a huge number of bacteria, but still, their proteomes remain relatively poorly defined. In the post-genomic era, proteomics is a rapidly growing field of research that is becoming increasingly important, because it deals with the study of proteins involved in carcinogenesis as well as a novel biomarker discovery for clinical use, such as screening, diagnosis, prognosis, detection of recurrent disease, etc. [8]. While a genome remains unchanged to a larger extent, the proteins in any particular cell change dramatically as genes are turned on or off in response to the environment.

Comparative genomic analysis of MTB $H_{37}Ra$ versus $H_{37}Rv$ by Zheng et al. revealed the genetic basis of virulence among these two strains [9]. However, proteomics is still nascent and requires extensive study. Since it is proteins that are directly involved in both normal and disease-related biochemical processes, a more comprehensive understanding of disease may be achieved by looking directly into the proteins within a disease cell or tissue [10]. Proteomics has much promise in novel drug discovery by targeting proteins of pathogenic organisms causing different diseases in the host, whereas comparative proteomics is very significant in studying the proteomic variations among different pathogens.

Identification of the virulence factors of MTB is a fundamental goal if new vaccines and anti-mycobacterial drugs against this pathogen are to be developed. A single amino acid mutation in protein sequence may cause alteration in the protein structure and function that may account for virulence and drug resistance properties of pathogenic organisms. Therefore, the development of an *in silico* technology to study the proteomic variations of different strains of genetically intractable pathogens such as MTB will enhance the analysis of virulence and drug resistance properties and significantly advance the understanding of the mechanisms of disease. In this study, the proteomic variations in these two strains (MTB $H_{37}Rv$ and $H_{37}Ra$) were determined, and the proteins that had undergone mutations (insertions/deletions/substitutions) were identified in the same variations. The findings of the present study provide a unique platform for the discovery of proteomic variation in other strains/species of *Mycobacterium* as well as the discovery and development of TB drugs, vaccines, biomarkers, etc.

Materials and methods

Dataset preparation

The dataset was prepared by retrieving the whole proteome of MTB $H_{37}Rv$ (NCBI RefSeq: NC_000962.2) and $H_{37}Ra$ (NCBI RefSeq: NC_009525.1) from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/). The protein sequence data ($H_{37}Rv$ and $H_{37}Ra$ proteome) were formatted using in-house developed PERL Script for quick analysis. NCBI Standalone BLAST-2.2.26 (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/) was used to perform protein BLAST [11] between the proteome of MTB $H_{37}Rv$ against MTB $H_{37}Ra$ and vice versa to discover protein variation and duplication.

Proteome comparison and database designing

Whole proteome comparison of these two organisms was done using PERL script and standalone BLAST. The output of the BLAST result was parsed and stored in MS SQL relational database tables using in-house developed PERL script. While parsing BLAST output results, percentage identities, positivities, number of gaps, identical residues, bits, bits score, e-value, query length, subject length, query sequence, subject sequence, consensus sequence, etc., of the first hit obtained were taken into consideration for each protein comparison.

Data retrieval and analysis

Different SQL queries were written to retrieve comparison data from MS SQL database tables. Sub-cellular localization (integral membrane, cytoplasmic, secretory and membrane attached by lipid anchor) of the selected MTB protein with variations were predicted by TBpred Prediction server [12].

Results

Genomic features and proteomic comparison of MTB $H_{37}Rv$ and its avirulent counterpart $H_{37}Ra$

MTB $H_{37}Rv$ contains a single circular chromosome of 4,411,532 bp with an average G + C content of 65.6% (NCBI RefSeq: NC_000962.2), which is 8445 bp smaller than the MTB $H_{37}Ra$ genome (NCBI RefSeq: NC_009525.1 and 4,419,977 bp length). A total of 4003 protein-coding sequences (CDS) are identified amongst 4062 genes in the $H_{37}Rv$ genome, while there are 4084 genes with 4034 protein-coding sequences in the genome of MTB $H_{37}Ra$ (Table 1). The proteomic comparison of MTB $H_{37}Rv$ and MTB $H_{37}Ra$ in this study revealed 3759 identical proteins between these two strains, while a reverse comparison, i.e., the proteome of MTB $H_{37}Ra$ against MTB $H_{37}Rv$, revealed 3774 identical proteins. Upon further analysis of this difference in number, it was found that there were 16 multiple copies of one protein, i.e., IS6110 transposase in MTB $H_{37}Ra$ while there were only four in MTB $H_{37}Rv$ of the same protein,