Comparative Functional Genomics Studies for Understanding the Hypothetical Proteins in *Mycobacterium tuberculosis* KZN 1435

Swapnil Sanmukh
Ecosystem Division, National Environmental Engineering Research Institute (NEERI), Nehru Marg, Nagpur-440020, Maharashtra (India)

Sumita Goswami
Ecosystem Division, National Environmental Engineering Research Institute (NEERI), Nehru Marg, Nagpur-440020, Maharashtra (India)

Sandhya Swaminathan
National Environmental Engineering Research Institute (NEERI), CSIR-Complex, Chennai-600113

Waman Paunikar
Ecosystem Division, National Environmental Engineering Research Institute (NEERI), Nehru Marg, Nagpur-440020, Maharashtra (India)

ABSTRACT

The prediction for the unknown proteins from *Mycobacterium tuberculosis* KZN 1435 were carried out for characterization of the proteins in their respective families. In *Mycobacterium tuberculosis* KZN 1435 out of 1560 genes for hypothetical proteins, functions were predicted for 1221 hypothetical protein whereas, structures for 803 unknown proteins were revealed. The Bioinformatics web tools like CDD-BLAST, INTERPROSCAN, PFAM and COGs were used for the prediction of functions in the proteins by searching protein databases for the presence of conserved domains; whereas, tertiary structures were constructed using PS^2 Server-Protein Structure Prediction server. This study was helpful in understanding functional characteristics of hypothetical proteins in *Mycobacterium tuberculosis* KZN 1435 as well as their role in the life cycle of the bacterium.

Keywords

Unknown proteins; Bioinformatics web tools, protein databases, tertiary structures, functional characteristics.

1. INTRODUCTION

Mycobacterium tuberculosis (MTB) is one of the most pathogenic bacterial species in the genus *Mycobacterium* and the causative agent of most cases of tuberculosis (TB). The resent studies carried out for three strains of *M. tuberculosis* isolated from patients in KwaZulu-Natal, South Africa have been sequenced [9]. These three strains are reported to have a range of important drug resistance phenotypes spanning fully drug-sensitive (DS) to multiple drug resistant (MDR), to extensively drug resistant (XDR). XDR TB is extremely difficult to treat. Recently, a high mortality rate for patients infected with XDR TB was reported in the KwaZulu-Natal region [10]. We carried out bioinformatics studies for multiple drug resistant (MDR) *M. tuberculosis* (KZN 1435) out of three sequenced strains of M. tuberculosis XDRi (KZN 605), MDR (KZN 1435), and DS (KZN 4207) strains from the KZN region of South Africa by Murray et al.[9].

The Computational biology (in-silico studies) is one of the newly emerging technologies which assist us to carry out comparative genomic studies to predict the functionality within the uncharacterized sequences using the different strategies of comparative genomics. The online bioinformatics tools and servers have ability for searching databases by choosing standard parameters for revealing the function of a particular gene (protein). Bioinformatics can help us to determine the presence of the enzymatic conserved domain/s in the sequences and may assist in the categorizing protein into specific family. The online automated servers are also available which can predict the three dimensional structures for protein sequences by using the strategy of aligning target sequences with orthologous sequences by virtue of sequence homology using best scored template of orthologous family member.

Bioinformatics web tools like CDD-BLAST,INTERPROSCAN, PFAM and COGs can search the orthologous sequence in biological sequence databases for the target sequence, while assist in classification of target sequence in particular family [8, 13-20]. We can predict 3-D structure of such proteins by using Protein Structure Prediction Server (PS2 server) [7, 21].

The present paper reports the comparative genomic studies for understanding the structural and functional properties of hypothetical proteins present in *Mycobacterium tuberculosis* KZN 1435 which will prove to be helpful for identifying novel enzymes and protein candidates with possible applications in the near future. Moreover, they will be helpful in developing drug against *Mycobacterium tuberculosis* and for understanding the evolutionary relationship with other Mycobacterium species.

2. METHODOLOGY

2.1 Sequence Retrieval

The whole genome sequences for *Mycobacterium tuberculosis* KZN 1435 was retrieved from the KEGG database (http://www.genome.jp/kegg/). [9]

2.2 Functional Annotation and Categorization

The hypothetical proteins from all *Mycobacterium tuberculosis* KZN 1435 were screened for the presence of

conserved domains using the web-tools. The four bioinformatics web tools CDD-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) [2, 3, 12] INTERPROSCAN (http://www.abi.ac.uk/interpro) [22] Pfam (http://www.pfam.sanger.ac.uk/) [1] and COGs (http://www.ncbi.nih, gov/cog)[11] were used, which shows the ability to search the defined conserved domains in the sequences and assist in the classification of proteins in appropriate family. The function prediction web tools have shown variable results depending upon the information available in databases, when searched for the conserved domains in the submitted protein sequences under study.
Hypothetical proteins analyzed by the function prediction web tools have shown variable results depending upon the information available in databases, when searched for the conserved domains in hypothetical sequences. The results proved to be helpful in further categorizing the proteins into their respective family.

2.3 Protein Structure Prediction

Online PS2 Protein Structure Prediction Server was used to generate 3D-structures of hypothetical proteins (http://www.ps2.life.nctu.edu.tw/) [2, 6, 7, 12]. The server accepts the protein (query) sequences in FASTA format and uses the strategies of Pair-wise and multiple alignments to generate resultant proteins 3D structures, which are constructed using structural positioning information of atomic coordinates for known template in PDB format using best scored alignment data. Where the selection of template was based on the same conserved domain detected in the functional annotations and which must be available in the structure alignment for modeling purpose.

3. RESULTS AND DISCUSSION

The comparative genomic studies for characterizing 1560 genes of *Mycobacterium tuberculosis* KZN 1435 by using sequence similarity search with close orthologous family members available in various protein databases using the web tools were carried out. The online-automated PS2server was used for the prediction of 3-D structure of screened hypothetical proteins. The results obtained after analysis of proteins by using web tools for classification of 1221 hypothetical proteins into particular protein family based on conserved domain available in the sequence and the predicted three dimensional structures for 803 proteins using best scored orthologous template are represented in Table 1. The 3-D structures built by using best scoring templates are represented in the order as Template ID, Identity, Score and E-value in structure column of respective *Mycobacterium tuberculosis* KZN 1435 specific gene in Table 1. (http://research.ijcaonline.org/volume60/number1/table.pdf)

4. CONCLUSION

These in-silico studies have sorted many functionally important hypothetical proteins of *Mycobacterium tuberculosis* KZN 1435 applying the parameters of pair-wise and multiple sequence alignment tools along with structure prediction tools, which suggest that many probable functional uncharacterized proteins are available in the *Mycobacterium tuberculosis* KZN 1435. These studies for characterization of unknown proteins of *Mycobacterium tuberculosis* KZN 1435 were carried out for verifying the structure and functions of the gene products. We were able to predict and categorized 1221 proteins functionally and 803 proteins structurally from 1560 hypothetical protein sequences screened from the *Mycobacterium tuberculosis* KZN 1435. This predicted functions and three dimensional structures may assist in establishing their role in life cycle of *Mycobacterium tuberculosis* KZN 1435 which are still unclear and can be used in future for the further understanding of pathogenicity and evolutionary development of *Mycobacterium* species and its life cycle [4,5]. This computationally generated data can also be used for developing drugs through drug designing by computational docking studies.

5. ACKNOWLEDGEMENTS

W. P. wants to thanks S. S. (Ph.D. Research Scholar) & S. G. (Trainee) for carrying out extensive bioinformatics analysis, critical editing, referencing and preparation of Manuscript. The authors also want to thanks Dr. Asha Juwarkar (Head, Ecosystem Division, NEERI, Nagpur-440020, Maharashtra, India) and Dr. Sandhya Swaminathan (Principal Scientist & Head, National Environmental Engineering Research Institute (NEERI), CSIR-Complex, Chennai-600113) for support and suggestions for carrying out this work.

6. REFERENCES

[1] Murray M, Pillay M, Borowsky ML., Young SK, Zeng Q, Koehrsen M, Alvarado L, Berlin AM, Borenstein D, Chen Z, Engels R, Freedman E , Gellesch M, Goldberg J, Griggs A, Gujja S, Heiman DI., Hepburn,T.A., Howarth,C., Jen,D., Larson,L., Lewis,B., Mehta,T., Park,D., Pearson M, Roberts A, Saif S, Shea TD, Shenoy N, Sisk P, Stolte C, Sykes SN., Walk T, White J, Yandava C., Haas B, Nusbaum C, Galagan J and Birren B. The Genome Sequence of *Mycobacterium tuberculosis* strain KZN 1435 (Unpublished)

[2] Neel RG, Anthony MA, Willem S, Robert P, Thiloshini G, Umesh L, Kimberly Z, Jason A, Gerald F. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. The Lancet, 368, 1575 - 1580, (2006).

[3] Edward E, Gary LG., Osnat H, John M, John O, Roberto JP, Linda B, Delwood R., Andrew J H. Biological function made crystal clear- annotation of hypothetical proteins via structural genomics. Current Opinion in Biotechnology 11, 25-30, (2000).

[4] Swapnil GS, Waman NP, Tarun, KG. Study of Hypothetical Proteins in Salmonella Phages and Predicting their Structural and Functional Relationship CiiT International Journal of Biometrics and Bioinformatics. DOI: BB022011001, (2011).

[5] Swapnil GS, Dilip, B. M., Waman NP, Tarun, KG. Computational characterizations for structure and function of unclassified proteins in Ictalurus punctatus. CiiT International Journal of Artificial Intelligent Systems and Machine Learning DOI: AIML052011001, (2011).

[6] Swapnil GS, Waman NP. Study of hypothetical proteins in Shigella phages. CiiT International Journal of fuzzy Systems DOI: FS062011002, (2011).

[7] Swapnil GS, Waman NP, Dilip, BM., Tarun, KG. Functionality search in hypothetical proteins of Halobacterium salinarum. CiiT International Journal of fuzzy Systems DOI: FS062011001, (2011).

[8] Swapnil GS, Waman NP, Dilip BM., Tarun, KG. Insilico function prediction for hypothetical proteins in Vibrio parahaemolyticus Chromosome II. CiiT International Journal of Data Mining and Knowledge Engineering. DOI: DMKE052011003, (2011).

[9] Swapnil GS, Waman NP, Tarun, KG & Tapan, C. Structural & functional prediction of hypothetical Proteins In bacteriophages against halophilic bacteria - an in silico approach. Int J Pharm. Bio. Sci. Vol 2 (2), B61-B70, (2011)

[10] Swapnil GS, Waman NP, Tarun, KG. & Tapan, C. Structure and Function Predictions of Hypothetical Proteins in Vibrio Phages. International Journal of Biometrics and Bioinformatics. 4, 161-175, (2010).

[11] Swapnil GS, Waman NP, Tarun, KG. Computational approach for structure and functionality search for hypothetical proteins in Mycobacterium leprae CiiT International Journal of Data Mining and Knowledge Engineering DOI: DMKE032011014, (2011).

[12] Zafer, A., Yucel, A., Mark, B. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, BMC Bioinformatics ,7, 178, (2006).

[13] Altschul SF., Madden TL., Schaffer AA., Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389-402, (1997).

[14] Chih-Chieh C, Jenn-Kang H, Jinn-Moon Y $(PS)^2$: protein structure prediction server Nucl. Acids Res. 34, W152-W157, (2006).

[15] Alejandro AS, Aravind L, Thomas LM, Sergei S, John LS, Yuri IW, Eugene VK, Stephen F A. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29(14), 2994-3005, (2001).

[16] Aron MB, John BA, Myra KD, Carol DS, Noreen RG, Marc G, Luning H, Siqian H, David IH, John DJ, Zhaoxi K, Dmitri K, Christopher JL, Cynthia AL, Chunlei L, Fu L, Shennan L, Gabriele HM, Mikhail M, James SS., Narmada T, Roxanne AY., Jodie JY, Dachuan Z, Stephen HB. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Research, Vol. 35, D237–D240, (2006).

[17] Zdobnov, E. M., Rolf, A. Interproscan- an integration platform for the signatures recognition methods in InterPro. Bioinformatics 17,847-848, (2001).

[18] Alex B, Lachlan C., Richard D, Robert DF., Volker H, Sam GJ, Ajay K, Mhairi M, Simon M, Erik LLS., David JS., Corin Y, Sean RE. The Pfam families' database. Nucleic Acids Research, Vol. 32, D138-D141, (2004).

[19] Roman LT, Michael YG, Darren AN, Eugene VK. The COG database: a tool for genome –scale analysis of protein functions and evolution. Nucleic Acid Research. 28, 33-36, (2000).

[20] Cédric N, Desmond GH, Jaap H. T-coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205-217, (2000).

[21] Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H. Prophage genomics. Microbiol Mol Biol Rev 67: 238–276, 2003.

[22] Casjens S. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49: 277–300, (2003).