Int J Pharm Bio Sci 2012 July; 3(3): (B) 856 - 862

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

PREDICTION OF SUBCELLULAR LOCALIZATION AND FUNCTION OF HYPOTHETICAL PROTEINS OF MYCOBACTERIUM TUBERCULOSIS H37RA STRAIN

LAKSHMI PILLAI* AND USHA CHAUHAN

Department Of Mathematics MANIT, BHOPAL, M.P, INDIA

ABSTRACT

The Tuberculosis is the classical human mycobacterial disease, caused by *Mycobacterium Tuberculosis. Mycobacterium tuberculosis* is a facultative intracellular pathogen that has evolved the ability to survive and multiply within human macrophages. These bacteria comprise of significant proteins, which involve in the pathogenesis and regulation of cell activity. Thus there arises the need to understand various parameters of these proteins for prediction of their functionality. The computational approaches for prediction of their classes are fast and economical therefore can be used to complement the existing wet lab techniques. Realizing their importance, in this paper an attempt has been made for the *insilico* prediction of protein subcellular localization and major functions. As in the case of *Mycobacterium,* proteins are often involved in extensive interactions at various subcellular localizations in cell. Total one thousand four hundred and thirty-two hypothetical proteins of *M. tuberculosis* were predicted for four locations viz cytoplasmic, integral membrane, secretory protein and protein attached to membrane by Lipid anchor in the subcellular localization. And also major functions like virulence factors, information molecule, cellular process and metabolism molecule were predicted. Such predictions provide a method to annotate *Mycobacterium* proteomes with subcellular localization and functional information rapidly. And they have widespread applications in function of proteins in the host cell and in designing the tuberculosis drugs.

Keywords: Sub cellular localization, pathogenicity, chemotherapeutic drugs, virulence factors, hemolytic molecules, secretary proteins.

LAKSHMI PILLAI
Department of Mathematics MANIT, BHOPAL, M.P, INDIA

Corresponding author

INTRODUCTION

Mycobacterium tuberculosis continues to be the major infectious cause of human death in developing countries and has reemerged in industrialized countries. Tuberculosis is a global problem and its suffering ranges from less than 10 per 100,000 in North America, 100-300 per 100,000 in Asia and Western Russia to over 300 per 100,000 in Southern and Central Africa. In every 15 seconds there is one death from tuberculosis (2 million deaths per year) and 8 million people develop tuberculosis annually, without treatment up to 60% people infected will be dying. Its major rationales were poverty, lack of healthy living conditions and adequate medical care. Tuberculosis continues to affect about 30% world's population, mainly in developing countries, despite existence of chemotherapeutic drugs and widespread use of the *Mycobacterium bovis* BCG vaccine. Effective chemotherapeutic treatment is expensive, takes long time and not available to people in various parts of world. The situation is further complicated by appearance of multidrug-resistant strains. BCG vaccination efficacy is also controversial, as it is not succeeded to protect adults against pulmonary tuberculosis [6].

Various methods had been developed for predicting sub cellular location of eukaryotic, prokaryotic (Gram-negative and Gram-positive bacteria) but only one method has been developed for mycobacterium protein. In this analysis, an attempt is made to predict sub cellular location of *Mycobacterium* proteins. This group of organism is well known for its pathogenicity. After BCG developed in 1921, till date we do not have any promising vaccine against tuberculosis. Furthermore, several new pharmaceutical targets unravel to combat against the multi-drug resistant strains of *Mycobacterium*. One of the key features of Gene ontology is cellular localization, which gives important information about a protein [10]. Earlier, cellular localization of *M. tuberculosis* is based on *in vitro* assay like ELISA, western blotting and *in situ*. Seven novel antigens of *M. tuberculosis*, previously identified based on its reactivity to sera from patients with tuberculosis, were characterized. One protein identity was localized in membranes and two were cytosolic, while two others, which had a high proline contents, were tightly associated with the cell wall one protein was secreted.

Thus, it is important to predict subcellular localization of protein in pathogenic organism like *Mycobacterium*. Generally, existing methods of subcellular localization developed for eukaryotic proteins like TSSub, LOCSVMPSI [14], ESLpred, Euk-Ploc [9]. As various tools were available for prediction of subcellular localization of prokaryotic proteins viz. PSORTb, PSLpred [7], a model named TBpred [15] has been developed for predicting four subcellular locations of mycobacterium proteins, namely cytoplasmic, integral membrane, secretary and membrane- attached proteins. In the present study, online server for prediction of four subcellular locations of Mycobacterial protein has been used.

The method used here was an indirect method where attempts have been made to predict subcellular localization of proteins rather than function. The subcellular localization methods are based on observation that protein belongs to same compartment of protein has similar amino acid composition [2] and has similar functions.

A direct attempt has been made to predict major functions (virulence factors, information molecule, cellular process and metabolism molecule) of gram-negative bacterial proteins including virulence factors that cause pathogenicity to the host system. Most of the proteins in an organism involves in cellular process, metabolism and in information storage, remaining can be classified in virulence factors, which allow the germs to establish themselves in the host. Virulence factors include adhesions, toxins and hemolytic molecules. Identification of virulence factors is

crucial for drug development. So, the bacterial proteins were classified into four broad functional classes. The three broad functional classes were taken from COGs functional annotation. They are i) cellular process, which includes cell division, cell envelope biogenesis, cell motility and signal transduction molecule; ii) information storage and processing, in which transcription, translation and DNA replication and repair molecule is included; iii) metabolic includes energy production and carbohydrate, amino acid, nucleotide, lipid transport and metabolism.

The aim behind this study was to predict the sub cellular localization of putative proteins and major function of *M. tuberculosis* H37RA strain as they might be useful for targeting antimycobacterial drugs.

MATERIALS AND METHODS

Collection of sequences

The complete protein sequences of cell surface, lipid & fat metabolism, amino acid & purine biosynthesis genes, anaerobic respiration & oxidative stress, metal uptake of *Mycobacterium tuberculosis* H37RA were extracted from biological database National Centre for Biotechnology Information (NCBI) cited at http://www.ncbi.nlm.nih.gov

Prediction of sub cellular localization of proteins

The TBPred publically available online tool was used in this study. TBPred is a SVM based method, predicts 5 major subcellular localization (cyotplasm, inner-membrane, outer-membrane, extracellular, periplasm) of Mycobacterial proteins. This method includes various SVM modules based on different features of the proteins such as - Amino acid composition, Dipeptide Composition, and position specific scoring matrix (PSSM). The overall prediction accuracy of these SVM modules are 82.51, 80.39 and 86.62% respectively. Along with SVM other techniques like profile HMM and MEME/MAST motif based studies were also applied. Moreover a hybrid approach combining the pssm based SVM model and the MEME/MAST model has been incorporated. In this paper subcellular localization is predicted based on amino acid composition with an accuracy of 86%.

Prediction of major functions of proteins

VICMpred (http://www.imtech.res.in/raghava/vicmpred/) server was used to predict function of proteins from its amino acid sequence. The VICMpred server [3] uses SVM based method having patterns, amino acid and dipeptide composition of bacterial protein sequences and overall accuracy of this server is **70.75%**. The common gateway interface (CGI) script for VICMpred is written using PERL version 5.03. This server is installed on a Sun Server (420E) under a UNIX (Solaris 7) environment. It is important in drug and vaccine point of view to select virulence proteins from the pool of proteins or the proteome of an organism.

RESULTS AND DISCUSSION

In this study we have selected one thousand four hundred and thirty two putative proteins of *M. tuberculosis* H37RA and subcellular localization and functional classes were analyzed. An online TBPred server was used to predict protein localization within *Mycobacterium* tuberculosis bacteria or targeting the host. We investigate whole putative proteome of *Mycobacterium* and their specific subcellular location (Table1). An extent of utilization of human cellular localization mechanisms by bacterial proteins and that appropriate subcellular localization predictors can be used to predict bacterial protein localization within the host cell. This is a pathogenic strain of human. Therefore, we have selected secretary proteins [8], which is responsible for causing human disease. In this study, the subcellular localization of proteins within the *M. tuberculosis* was out of 1432 proteins (Table 2), the 1417 proteins was cytoplasmic, 13 integral membrane, 1 secretory and 1 protein attached to membrane

by Lipid. Also the functional classes (Table 3) of proteins were found i.e 498 proteins involved in cellular process, 728 in metabolism, 129 in information and storage, and 76 virulence factors.

Table 1
Functional classes and subcellular localization of putative proteins of Mycobacterium tuberculosis H37RA.

Acc id	amino acid based subcellular prediction	VICM Pred
ZP_02553214	CYTOPLASMIC PROTEIN	Metabolism molecule
ZP_02553212.1	INTEGRAL MEMBRANE PROTEIN	Metabolism Molecule
ZP_02553209.1	INTEGRAL MEMBRANE PROTEIN	Metabolism Molecule
ZP_02553208.1	CYTOPLASMIC PROTEIN	Cellular process
ZP_02553207.1	INTEGRAL MEMBRANE PROTEIN	Cellular process
ZP_02553206.1	CYTOPLASMIC PROTEIN	Cellular process
ZP_02553205.1	INTEGRAL MEMBRANE PROTEIN	Cellular process
ZP_02553204.1	SECRETED PROTEIN	Metabolism Molecule
ZP_02550120.1	CYTOPLASMIC PROTEIN	Information and storage
ZP_02549605.1	CYTOPLASMIC PROTEIN	Metabolism molecule
ZP_02549604.1	CYTOPLASMIC PROTEIN	Cellular process
ZP_02549603.1	PROTEIN ATTACHED TO MEMBRANE BY LIPID ANCHOR	Cellular process
ZP_02549602.1	INTEGRAL MEMBRANE	Cellular process
ABR14062.1	CYTOPLASMIC PROTEIN	Metabolism Molecule
ABR14061.1	CYTOPLASMIC PROTEIN	Cellular process
AAB07556.1	CYTOPLASMIC PROTEIN	Cellular process
CAB05953.1	INTEGRAL MEMBRANE PROTEIN	Metabolism Molecule
ZP_02553232.1	INTEGRAL MEMBRANE PROTEIN	Metabolism Molecule

ZP_02553231.1	CYTOPLASMIC PROTEIN	Metabolism Molecule
ZP_02553230.1	INTEGRAL MEMBRANE	Metabolism Molecule
ZP_02553229.1	INTEGRAL MEMBRANE	Cellular process
ZP_02553228.1	CYTOPLASMIC PROTEIN	Metabolism Molecule
ZP_02553227.1	CYTOPLASMIC PROTEIN	Cellular process
ZP_02553226.1	CYTOPLASMIC PROTEIN	Metabolism Molecule
ZP_02553225.1	CYTOPLASMIC PROTEIN	Metabolism Molecule
ZP_02553224.1	CYTOPLASMIC PROTEIN	Cellular process
ZP_02553223.1	INTEGRAL MEMBRANE	Cellular process
ZP_02553222.1	CYTOPLASMIC PROTEIN	Metabolism Molecule
ZP_02553221.1	CYTOPLASMIC PROTEIN	Metabolism Molecule
ZP_02553220.1	CYTOPLASMIC PROTEIN	Cellular process

Table 2
Subcellular localization Predicted

Protein Locations	Secretory Proteins	Cytoplasmic proteins	integral membrane protein	protein attached to membrane by Lipid
Number of protein in particular location	**1**	**1417**	**13**	**1**

Table 3
Functional classes of proteins Predicted

Functional classes	cellular process	information and storage	metabolism	virulence factors
Number of Proteins predicted	**498**	**130**	**728**	**76**

CONCLUSION

Previously, all the study has been done on the basis of in vitro assay for identification of subcellular localization and function of proteins. Since, there was no computational tool available to predict the location of protein for targeting the vaccine or targeting the drugs. Very few reports were available on localization of proteins *in vitro* experiments. In conclusion, we include the specified prediction of subcellular localization and functional class prediction results in the most putative proteins of *M. tuberculosis* H37RA. This initiative might be useful in annotating newly sequenced or hypothetical mycobacterial proteins. Also it is important in drug and vaccine point of view to select virulence proteins from the pool of proteins or the proteome of an organism. Thus the search for a potential vaccine/ drug target for an important bacterial pathogen by *in vitro* researchers will greatly be appended by this prediction.

ACKNOWLEDGMENT

The authors are highly thankful to the Department of Biotechnology, Delhi, India for providing support in the form of Bioinformatics infrastructure facility to carry out this work.

REFERENCES

1. Bhasin M, Garg A, Raghava GP: PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, 21:2522-2524,18-26 (2005).
2. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci. Dec; 60(12):2637-50. Review (2003).
3. Saha, S. and Raghava, G.P.S. VICMpred: SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition, Genomics Proteomics & Bioinformatics. 4(1):42-7 (2006).
4. Bhasin M, Raghava GP: ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Research*, 32:W414-W419 (2004).
5. Bloom, B. R., and J. D. Mckinney. The death and resurrection of tuberculosis. *Nat. Med*. 5:872-874(1999).
6. Bloom, R. B., and C. J. L. Murray. Tuberculosis: commentary on a reemergent killer. *Science* 257:1055-1064 (1992).
7. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS: PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, 21:617-623(2005).
8. Gomez M, Johnson S, Gennaro ML Identification of secreted proteins of Mycobacterium tuberculosis by a bioinformatics approach. Infect Immun. 68(4):2323-7(2000).
9. Guo J, Lin Y: TSSub: eukaryotic protein subcellular localization by extracting features from profiles. *Bioinformatics*, 22:1784-5 (2006).
10. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J: *Molecular Cell Biology* 3rd edition. Scientific American Books,New York; 739-777(1995).
11. Mamoon Rashid, Sudipto Saha and Gajendra PS Raghava. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics. 8: 337(2007).
12. Shen HB, Yang J, Chou KC. Euk-PLoc: an ensemble classifier for large-scale

eukaryotic protein subcellular location prediction. *Amino Acids*, 33:57-67(2007).

13. Smith, I. *Mycobacterium tuberculosis* Pathogenesis and Molecular Determinants of Virulence. *Clinical Microbiology Reviews*, 16(3): 463-496 (2003).
14. Xie D, Li A, Wang M, Fan Z, Feng H, LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Research*, 33:W105-W110 (2005).
15. Rashid M, Saha S, Raghava GPS, Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs, *BMC Bioinformatics*, 8:337, (2007).