FUNCTIONAL PROTEIN PREDICTION OF MYCOBACTERIUM TUBERCULOSIS USING BIOINFORMATICS TOOLS

SUGANYA, K[1], RAJESH SINGH, J*, SUBAHANISHA, A[1]
Department of Bioinformatics, Annamalai University, Tamilnadu, India
Department of Biotechnology, Rajah Serfoji Government Arts College (Autonomous), Thanjavur*

Abstract— Tuberculosis (TB) has been declared as a global health emergency by the World Health Organization (WHO). This has been mainly due to the emergence of multiple drug resistant strains and the synergy between tubercle bacilli and the Human Immunodeficiency Virus (HIV). The genomic analysis of strains for outbreak investigations is in vogue for about a decade now. However, information available from whole genome sequencing efforts and comparative genomics of laboratory and field strains is likely to revolutionize efforts towards understanding molecular pathogenesis and dissemination dynamics of this dreaded disease. Genomic information is also going to fuel discovery projects where new targets will be identified and explored towards a new drug for Tuberculosis Mycobacterium tuberculosis is a deadly infectious disease, there is rising death of humans every year because of this disease, availability of genome sequences of Mycobacterium tuberculosis has provided tremendous amount of information that can be useful in drug target and new vaccine development. Sequence similarity provides accurate annotation for genes in newly sequenced genomes. In this present work about 100 hypothetical proteins of Mycobacterium tuberculosis were taken and its functions were predicted using bioinformatics tools, BLOCKS, InterPro Scan and PFAM. From our analysis of 100 hypothetical proteins only ONE show 100% functions these proteins may serve as target for few antibiotics.

Key words: Tuberculosis, Mycobacterium tuberculosis, proteins, genome sequences

INTRODUCTION

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* is an ancient infection that has plagued humans throughput recorded and archeological history. Despite the availability of effective chemotherapy and a moderately protective vaccine, recently 8.8 million people fell ill with TB and 1.4 million died from TB. (Altschul *et al*., 1990). Mycobacterium tuberculosis is reputed to have the highest annual global mortality among all of the pathogens. The rise in tuberculosis (TB) incidence over the last two decades is partly due to TB deaths in HIV-infected patients and partly due to the emergence of multidrug resistant strains of the bacteria. (Sharma S.K and Mohan V 2004).However, rapidly evolving mycobacterial genomics with complete genome sequence known along with powerful bioinformatics approaches, one can realize better therapeutics and prophylactics in the near future.(Chakhaiyar P and Hasnain S.E 2004).Researchers are utilizing information obtained from the complete sequence of the *M.tuberculosis* genome and from new genetic and physiological methods to identify targets in *Mycobacterium tuberculosis* that will aid in the development of these sorely needed anti tubercular agents.(Issar smith., 2003).

Recent advances in DNA sequencing are leading to the ability to generate whole genome information in clinical isolates of Mycobacterium tuberculosis complex (MTBC). (Frances coll *et al*., 2014). The Mycobacterium tuberculosis complex that consists of 6 members: *Mycobacterium tuberculosis* and *Mycobacterium African* mum, which infect humans; *Mycobacterium microti*, which infects vole; *Mycobacterium bovis*, which infects other mammalian species as well as humans*; M.bovis* BCG, a variant of Mycobacterium bovis and Mycobacterium canettii, a pathogen that infects humans. M. tuberculosis and seven very closely related mycobacterial species (*M.bovis, M.africanum*, *M.microti*, *M.caprae*, *M.pinnipedii*, *M.canetti* and *M.mungi*) together comprise what is known as the M.*tuberculosis* complex. Most, but not all, of these species have been found to cause disease in humans.

Although environmental factors can increase susceptibility to disease, it is clear that resistance to tuberculosis infection is under genetic control. (Levin. M and Newport. M., 2000).*Mycobacterium tuberculosis* has circular chromosomes of about 4,200,000 nucleotides long. The G+C content is about 65% (NCBI 2007). Genes that code for lipid metabolism are a very important part of the bacterial genome, and 8% of the genome is involved in this activity (Cole S.T., 2002).The different species of the *Mycobacterium tuberculosis* complex show a 95-100% DNA relatedness based on studies of DNA homology, and the sequence of the 16SrRNA gene are exactly the same for all the species (Aranaz A.,1999).

Unfortunately, the selection and spread of multidrug–resistant (MDR) *Mycobacterium tuberculosis* strains worsen the scenario, since an estimated 0.65 million cases of MDR–TB were documented for the year 2010 (WHO report, 2010Tuberculosis (TB) still remains the largest killer infectious disease despite the availability of several chemotherapeutic drugs and vaccines. In Iraq the incidence rate was estimated to be 45/100,000 (Al-Basra 2013). Antibiotic resistance is very high , in 2008 , it has been estimated that 6.6% of isolates were resistant to four drugs in use (Isoniazid, Rifampicin, Streptomycin and Ethambutol)(Al-Kareemi K 2008), besides the existence of strains which were resistant to mono,di,and tri-antibiotics at a high rate , however ,WHO estimated MDR (Multi-drug resistance) at 3.4% in 2011 in new TB cases(WHO 2012).

Such prediction is of great significance in pathogenic organisms, since function recognition in these organisms can enable identification of potential drug targets. There have been several attempts, using sophisticated homology search tools, to assign functions to gene products encoded in various genomes (Rychlewski *et al*., 1998; Pawlowski *et al.,* 1999; Hoersch *et al.,* 2000; Tatusov *et al*., 2000; Pearl *et al*., 2002; Meyer *et al*., 2003). In the present study, 7199 proteins of *Mycobacterium tuberculosis* were analyzed and out of that 2038 hypothetical proteins were presented out of the 100 hypothetical protein were selected for this study and its functions were predicted using bioinformatics tools such as BLOCKS, INTERPROSCAN and PFAM

MATERIALS AND METHODS

Hardware Configurations

Processor	Intel Corei5
Hard disk	500 GB
RAM	24GB DDR3 RAM
Key board	LOGITECH Multimedia
Monitor	794 MG SAMSUNG LCD
DVD drive	SONY 52 MA
Mouse	LOGITECH
Printer	Canon LBP 2900B

Software and tools

Protein sequence retrieval : NCBI
Function prediction : BLOCKS
Function prediction : INTERPROSCAN
Function prediction : PFAM

Tool description

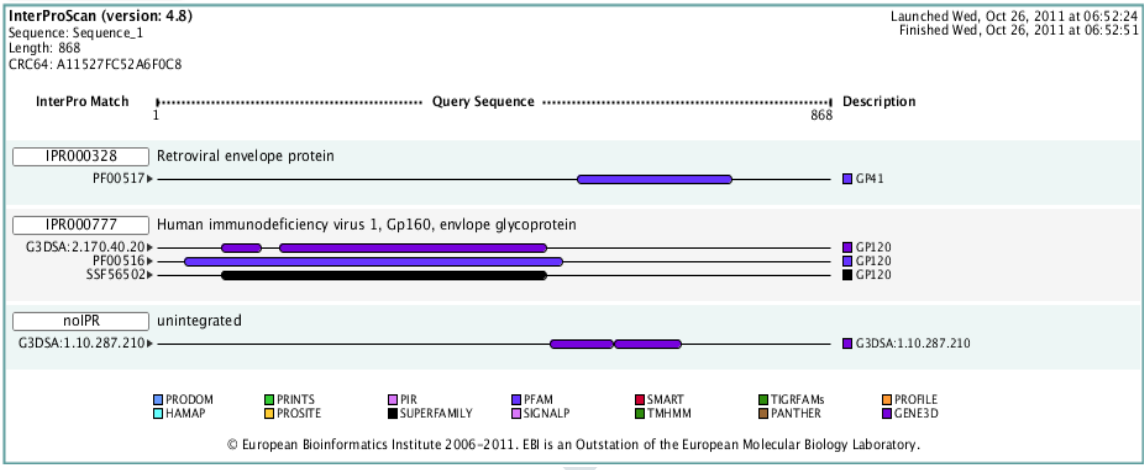
Databases Used

The following are the databases which are used to get all the necessary data such as protein sequence, protein functions and prediction, NCBI, BLOCKS, PFAM, and INTERPROSCAN

National Center for biotechnology Information

The National Center For Biotechnology Information (NCBI) houses genome sequencing data in Gen bank and an index of biomedical research articles in Pub med central and pub med as well as other information relevant to bio technology. All these databases are available online through the Entrez search engine. Gen bank coordinates with other sequence database such as those of the European Molecular Biology Laboratory (EMBL) and the DNA database of Japan (DDBJ). The NCBI assigns a unique identifier (Taxonomy ID number) to each species of organisms.

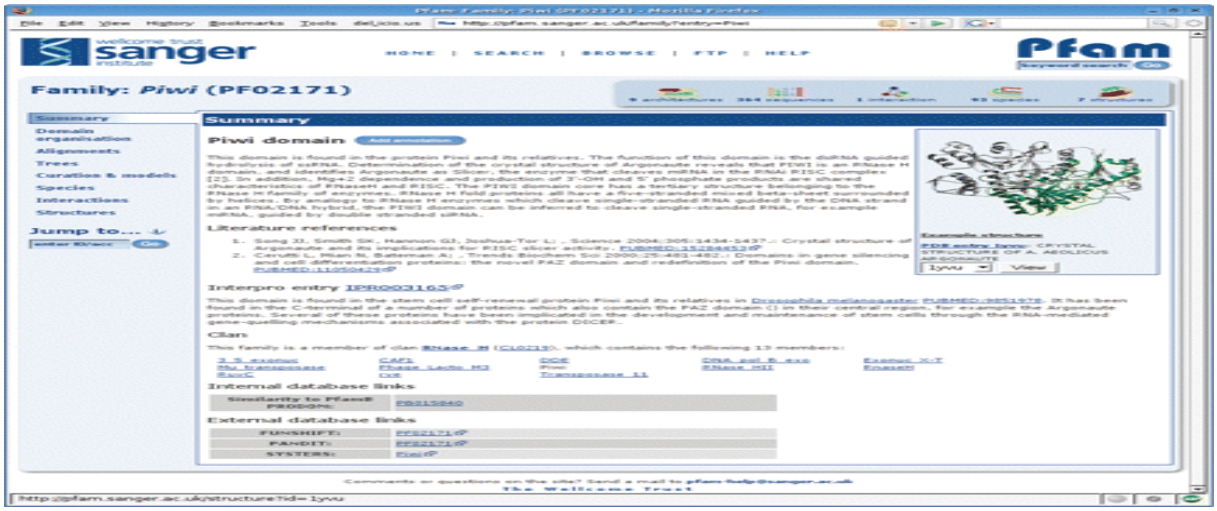
INTER PROSCAN



InterProScan is a tool that combines different protein signature recognition methods native to the Inter Pro member databases into one resource with look up of corresponding Inter Pro and GO annotation. InterProScan is a tool that combines different protein signature recognition methods into one resource. The number of signature databases and their associated scanning tools, as well as the further refinement procedures, increases the complexity of the problem.InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by several different databases (referred to as member databases), that make up the InterPro consortium.

InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to new protein sequences in order to functionally characterize them.The contents of InterPro are based around diagnostic signatures and the proteins that they significantly match. The signatures consist of models (simple types, such as regular expressions or more complex ones, such as Hidden Markov models) which describe protein families, domains or sites. Models are built from the amino acid sequences of known families or domains and they are subsequently used to search unknown sequences (such as those arising from novel genome sequencing) in order to classify them. Each of the member databases of InterPro contribute towards a different niche, from very high-level, structure-based classifications (SUPERFAMILY and CATH-Gene3D) through to quite specific sub-family classifications (PRINTS and PANTHER).

InterPro's intention is to provide a one-stop-shop for protein classification, where all the signatures produced by the different member databases are placed into entries within the InterPro database. Signatures which represent equivalent domains, sites or families are put into the same entry and entries can also be related to one another. Additional information such as a description, consistent names and Gene Ontology (GO) terms are associated with each entry, where possible.



PFAM

Proteins are generally comprised of one or more functional regions, commonly termed domains. The presence of different domains in varying combinations in different proteins gives rise to the diverse repertoire of proteins found in nature. Identifying the domains present in a protein can provide insights into the function of that protein. The Pfam database is a large collection of protein domain families. Each family is represented by multiple sequence alignments and hidden Markov models (HMMs).

There are two levels of quality to Pfam families: Pfam-A and Pfam-B. Pfam-An entries are derived from the underlying sequence database, known as Pfamseq, which is built from the most recent release of UniProtKB at a given time-point. Each Pfam-A family consists of a curated seed alignment containing a small set of representative members of the family, profile hidden Markov models (profile HMMs) built from the seed alignment, and an automatically generated full alignment, which contains all detectable protein sequences belonging to the family, as defined by profile HMM searches of primary sequence databases. Pfam-B families are un-annotated and of lower quality as they are generated automatically from the non-redundant clusters of the latest ADDA release. Although of lower quality, Pfam-B families can be useful for identifying functionally conserved regions when no Pfam-A entries are found.

BLOCKS

Number of found motifs: 75

BLOCKS

BLOCKS	Position(Score)		Description	Related Sequence
BL00347L	861..914(2226)	Detail	Poly(ADP-ribose) polymerase zinc finger domai	7
BL00347K	799..847(1994)	Detail	Poly(ADP-ribose) polymerase zinc finger domai	7
BL00347I	655..694(1925)	Detail	Poly(ADP-ribose) polymerase zinc finger domai	7
BL00347J	705..759(1879)	Detail	Poly(ADP-ribose) polymerase zinc finger domai	7
BL00347G	536..586(1875)	Detail	Poly(ADP-ribose) polymerase zinc finger domai	7

As an aid to detection and verification of protein sequence homology, the BLOCKS Searcher compares a protein or DNA sequence to the current database of protein blocks. Blocks are short multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

The rationale behind searching a database of blocks is that information from multiply aligned sequences is present in a concentrated form, reducing background and increasing sensitivity to distant relationships. This information is represented in a position-specific scoring table or "profile", in which each column of the alignment is converted to a column of a table representing the frequency of occurrence of each of the 20 amino acids. For searching a database of blocks, the first position of the sequence is aligned with the first position of the first block, and a score for that amino acid is obtained from the profile column corresponding to that position. Scores are summed over the width of the alignment, and then the block is aligned with the next position. This procedure is carried out exhaustively for all positions of the sequence for all blocks in the database, and the best alignments between a sequence and entries in the BLOCKS database are noted. If a particular block scores highly, it is possible that the sequence is related to the group of sequences the block represents. Typically, a group of proteins has more than one region in common and their relationship is represented as a series of blocks separated by unaligned regions. If a second block for a group also scores highly in the search, the evidence that the sequence is related to the group is strengthened, and is further strengthened if a third block also scores it highly, and so on. BLOCKS are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

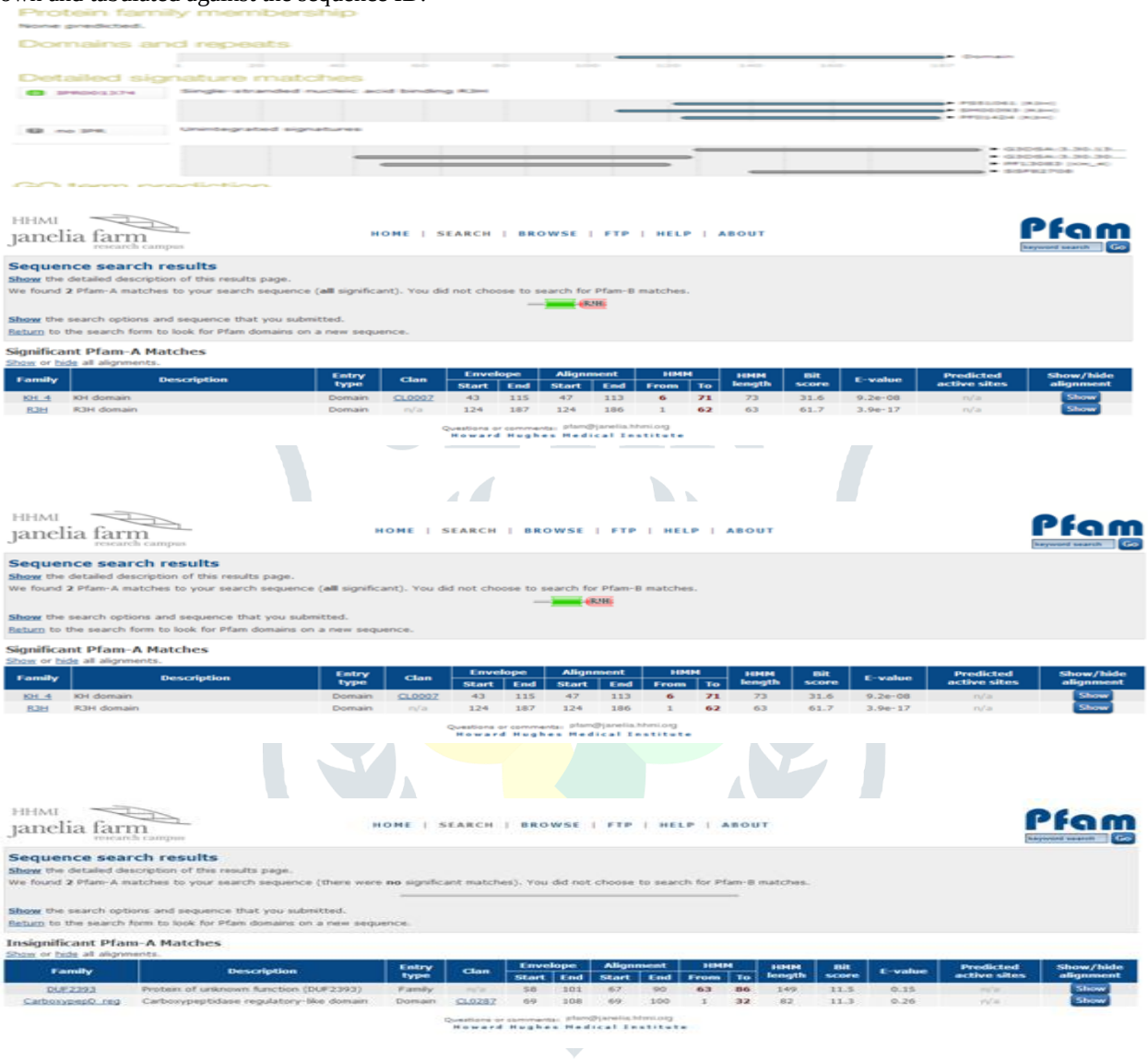
The blocks for the BLOCKS Database are made automatically by looking for the most highly conserved regions in groups of proteins documented in the Prosite Database. The Prosite pattern for a protein group is not used in any way to make the BLOCKS Database and the pattern may or may not be contained in one of the blocks representing a group. These blocks are then calibrated against the SWISS-PROT database to obtain a measure of the chance distribution of matches. It is these calibrated blocks that make up the BLOCKS Database."

The complete genome for mycobacterium tuberculosis was downloaded from the NCBI database. In the genome sequence of mycobacterium tuberculosis 2038 hypothetical proteins were present. Out those only 100 hypothetical proteins of genome sequence was

downloaded from the site NCBI. The functions for these hypothetical proteins were predicted using bioinformatics tools such as BLOCKS, interProScan and PFAM. The confidence level can be measured on the basis of above tools,100% confidence level was given to those proteins that indicate same functions in all the given three tools, 60% confidence level was given to those proteins that indicate same functions in any of the given two tools other are different.,20% confidence level was given to those proteins that indicates same functions in any of the given tools. Moreover 0% confidence level was given to those proteins that indicate unknown functions in all the given tools.

RESULTS

Mycobacterium tuberculosis has totally 7199 proteins. In this 2038 proteins are hypothetical proteins from which 100 hypothetical of mycobacterium tuberculosis were retrieved for the present study from NCBI Database hypothetical proteins were submitted to bioinformatics tools BLOCKS, Interproscan, and PFAM. The results of these tools were analysed and the functions observed in each output were noted down and tabulated against the sequence ID.



BLOCKS

Query=Unknown Unknown

Size=187 Amino Acids

Blocks Searched=27288

Alignments Done= 5813855

Cutoff combined expected value for hits=1

Cutoff block expected value for repeats/other=1

==========

Family		Strand	Blocks	Combined E-value
IPB001374	Single-stranded nucleic acid binding	1	2 of 2	5.6e-07
IPB002177	DNA-binding protein Dps	1	1 of 3	0.28
IPB004931	Prothymosin/parathymosin	1	1 of 1	0.29

3 possible hits reported

Query=Unknown Unknown

Size=802 Amino Acids

Blocks Searched=27288

Alignments Done=22595975

Cutoff combined expected value for hits=1

Cutoff block expected value for repeats/other=1

Family	Strand	Blocks	Combined E-value

0 possible hits reported

Query=Unknown Unknown

Size=802 Amino Acids

Blocks Searched=27288

Alignments Done=22595975

Cutoff combined expected value for hits=1

Cutoff block expected value for repeats/other=1

Family	Strand	Blocks	Combined E-value

0 possible hits reported

Query=Unknown Unknown

Size=169 Amino Acids

Blocks Searched=27288

Alignments Done= 5322671

Cutoff combined expected value for hits= 1

Cutoff block expected value for repeats/other= 1

Family		Strand	Blocks	Combined E-value
IPB008859	Thrombospondin, C-terminal	1	1 of 14	1e-05
IPB010531	NOA36	1	1 of 6	0.11
IPB001910	Inosine/uridine-preferring nucleosi	1	1 of 3	0.36

3 possible hits reported

Query=Unknown Unknown

Size=176 Amino Acids

Blocks Searched=27288

Alignments Done= 5513687

Cutoff combined expected value for hits= 1

Cutoff block expected value for repeats/other= 1

Family	Strand	Blocks	Combined E-value

0 possible hits reported

Functional protein prediction of Mycobacterium tuberculosis

S.NO	ACCESSION NO	GENE ID	BLOCKS	INTERPROSCAN	PFAM	CONFIDENCE LEVEL
1	YP_007353728.1	479057922	Single-stranded nucleic acid binding	Single-stranded nuclic acid binding K3H.	KH domain.	60%
2	YP_007353724.1	479057918	Dienelactone hydrolase.	Unintegrated .	No hits.	20%
	YP_007353717.1	479057911	No hits.	Unintegrated .	Protein of unknown function(DUF2393).	20%
4	YP_007353714.1	479057908	Thrombospondin, C-terminal.	Unintegrated .	No hits.	20%
5	YP_007353710.1	479057904	No hits.	Unintegrated .	Immunity protein 37.	20%
6	YP_007353707.1	479057901	Nsp1-like, C-terminal	No hits.	Aminoacyl-tRNA editing domain.	20%

96	YP_007353443.1	479057637	No hits.	Unintegrated	No hits.	0%
97	YP_007353442.1	479057636	Histidine acid phosphatase.	Unintegrated	No hits.	20%
98	YP_007353439.1	479057633	No hits.	Unintegrated.	No hits.	0%
99	YP_007353433.1	479057627	Phytanoyl-CoA dioxygenase	Phytaonyl-coA dioxygenase.	Phytanoyl-CoA dioxygenase(phyH).	100%
100	YP_007353427.1	479057621	D-Ala-D-Ala carboxypeptidase 3 (S13)	Peptidase S13,D-Ala-D-Ala carboxy peptidase C.	D-Ala-D-Ala carboxypeptidase3(S13)family, and D-Ala-D-Ala carboxypeptidase3(S13)family.	100%

Table-2.Percentage of similarity

S.NO	NUMBER OF PROTEINS IN 100%	NUMBER OF PROTEINS IN 60%	NUMBER OF PROTEINS IN 20%	NUMBER OF PROTEINS IN 0%
1	4	41	48	7

DISCUSSION

The existence of hypothetical proteins in genomes constitutes a major issue for comparative and functional genomics analyses. In particular for pathogenic organisms, these hypothetical proteins hamper the search for new and effective drug targets and weaken progress towards the advancement of research on these organisms and enhancement of our understanding of their virulence and pathogencity. In this present study, the complete genome for mycobacterium tuberculosis was downloaded from the NCBI database. In the genome sequence of mycobacterium tuberculosis 7199 proteins were present. Out of those only 100 hypothetical proteins of genome sequence was downloaded from the site NCBI. The function for these hypothetical proteins was predicted using bioinformatics tools such as BLOCKS, interProScan and PFAM. The confidence level can be measured on the basis of above tools.This was achieved by performing percentage scores of these proteins and the confidence level of predicted annotations for these proteins. The present study showed that, some of these proteins may contribute to the survival of the bacterial pathogen within the host system and they may have a particular role in helping the organism evade the host immune response and in persistence and latency. They are thus likely to be important for the specific lifestyle of the organism and adaptability of this pathogen in the host, so functional characterization of these proteins is essential.

Currently, there is a need for novel and effective drugs with new biological mechanisms of action against drug susceptible and drug-resistant strains. These need to be reliably administered with a shorter regimen to overcome the disease caused by this particular organism, which constitutes a public health challenge, claiming millions of lives and new cases every year. Such quantitative analysis may help us better understand the biology of the organism as a whole system and identify potential drug targets at the molecular level for the disease.

This was achieved by performing percentage scores of these proteins and the confidence level of predicted annotations for these proteins. The present study showed that,among 100 protein four(Table No-1 in page no:36&39)hypothetical proteins showed 100% confidence level, those four proteins can be used for the development of vaccines and antibiotics development. Some of these proteins may contribute to the survival of the bacterial pathogen within the host system and they may have a particular role in helping the organism evade the host immune response and in persistence and latency. They are thus likely to be important for the specific lifestyle of the organism and adaptability of this pathogen in the host, so functional characterization of these proteins is essential. The present study opened the gateway for the production develops new vaccines and drugs based on the four functional protein predictions.

ACKNOWLEDGEMENT

The Author is appreciative to Prof and Head Dr. R. Karuppasamy Department of Zoology, Annamalai University, Tamilnadu, India in support of providing confidence and indispensable amenities to conduct the present research work.

REFERENCE

[1] Altschul, S;Gish,W;Miller,W;Myres,E;Lipman, D,1990."Basic local alignment seach tool". *Journal of Molecular Biology* 215(3): 403-410.

[2] Aranaz A, Liébana E, Gomez-Mampaso E, 1999. "*M. Tuberculosis* subsp. caprae subsp. nov.: a taxonomic study of a new member of the *Mycobacterium tuberculosis* complex isolated from goats in Spain. *Int J Syst Bacteriol*; 49:1263 73.

[3] Al-Kareemi K. 2008 .Effect of Diacetyl on Mycobacterium tuberculosis In Vitro. MSc. thesis. Dept. of Microbiology, College of Medicine, University of Baghdad , Iraq

[4] Chakhaiyar P, Hasnain SE. 2004. defining the mandate of tuberculosis research in a post genomic era. *Med Princ Prac*; 13:177-84.

[5] Cole,S.T., 2002. "Comparative and functional genomics of the *Mycobacterium tuberculosis* complex." *Microbiology* 148: 2919-2928.

[6] Sharma S.K, Mohan V. 2004. Multidrug-resistant tuberculosis. *Indian J Med Res*; 120 : 354-76.

[7] Levin M, 2000. Newport M. Inherited predisposition to mycobac- terial infection: historical considerations. *Microbes Infect*;

[8] NCBI. 24 May 2007. Welcome Trust Sanger Insitute. 2 June 2007

[9] WHOreport 2010, http://www.who.int/tb/publications/global_report/2010 *webcite*

[10] WHO. Global Tuberculosis Control .Geneva, WHO Report, 2012.

[11] Rychlewski L, Zhang B and Godzik A 1998. Fold and function predictions for Mycoplasma genitalium proteins; *Fold Des*. 3 229–238

[12] Pawlowski K, Zhang B, Rychlewski L and Godzik A 1999 The Helicobacter pylori genome from sequence analysis to structural and functional predictions; *Proteins* 36 20–30

[13] Hoersch S, Leroy C, Brown N P, Andrade M A and Sander C 2000. The GeneQuiz web server protein functional analysis through the Web; *Trends Biochem.*

[14] Tatusov R L, Galperin M Y, Natale D A and Koonin E V 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution; *Nucleic Acids Res.* 28 33–36

[15] Pearl F M, Lee D, Bray J E, Buchan D W, Shepherd A J and Orengo C A 2002. The CATH extended protein-family data- base providing structural annotations for genome sequences; *Protein Sci*. 11 233–244

[16] Meyer F, Goesmann A, McHardy A C, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, 2003. GenDB- an open source genome annotation system for prokaryote genomes; *Nucleic Acids Res.* 31 2187–2195