

**IN SILICO ANNOTATION OF UN-CHARACTERIZED PROTEINS OF  
MYCOBACTERIUM TUBERCULOSIS**

A THESIS SUBMITTED IN PARTIAL FULFILLMENT  
OF THE REQUIREMENT FOR THE DEGREE OF

**Bachelor of Technology**

**in**

**Biotechnology Engineering**

**By**

**KIRAN SOY MURUM (ROLL NO. 107BT006)**

**SANTOSH KUMAR NAYAK (ROLL NO. 107BT009)**



**Department of Biotechnology & Medical Engineering**

**National Institute of Technology**

**Rourkela-769008**

**IN SILICO ANNOTATION OF UN-CHARACTERIZED PROTEINS OF  
MYCOBACTERIUM TUBERCULOSIS**

A THESIS SUBMITTED IN PARTIAL FULFILLMENT  
OF THE REQUIREMENT FOR THE DEGREE OF

**Bachelor of Technology**

**In**

**Biotechnology Engineering**

**By**

**KIRAN SOY MURUM (ROLL NO. 107BT006)**

**SANTOSH KUMAR NAYAK (ROLL NO. 107BT009)**

Under the Guidance of

**Prof. G.R.Sathpathy**



**Department of Biotechnology & Medical Engineering**

**National Institute of Technology**

**Rourkela-769008**



**National Institute of Technology  
Rourkela**

**CERTIFICATE**

This is to certify that the thesis entitled, “**Insilico Annotation of Un-characterized proteins of Mycobacterium Tuberculosis**” submitted by Santosh Kumar Nayak and Kiran Soy Murum in partial fulfillment of the requirement for the award of bachelor of technology degree in Biotechnology Engineering at National Institute of Technology, Rourkela (Deemed University) is an authentic work carried out by them under my supervision and guidance. To the best of my knowledge the matter embodied in the thesis has not been submitted to any other University/Institute for award of any Degree/Diploma.

Date:

Prof.G.R. Satpathy  
Dept. Of Biotechnology & Medical Engg.  
National Institute of Technology  
Rourkela-769008

## ACKNOWLEDGEMENT

We express our sincere gratitude to Dr. G.R.Satpathy, Professor of the Department of Biotechnology Engineering, National Institute of Technology, Rourkela, for giving us this great opportunity to work under his guidance throughout the course of this work. We are also thankful to him for his valuable suggestions and constructive criticism which have helped us in the development of this work. We are also thankful to his optimistic nature which has helped this project to come a long way through.

We are also thankful to Sri R.N.Satpathy, Assistant Professor and Department of Biotechnology Engineering of MITS Raygad for his assistance in the project work for his constructive criticism.

We are also thankful to the Prof (Dr.) Subhankar Paul, Head of the Department and our Department for providing us the necessary opportunities for the completion of the project.

Kiran Soy Murum  
Roll No. : 107bt006  
Session: 2007-2011  
Biotechnology Engineering  
National Institute of Technology  
Rourkela

Santosh Kumar Nayak  
Roll No. : 107bt009  
Session: 2007-2011  
Biotechnology Engineering  
National Institute of Technology  
Rourkela

# CONTENTS

	<b>Page No.</b>
<i>Abstract</i>	6
<i>List of Figures</i>	7
<i>List of Tables</i>	7
<b>Chapter 1</b>	<b>INTRODUCTION</b>
1.1	Mycobacterium Tuberculosis 9-11
1.2	Tuberculosis 12-13
1.3	Death Rates by Tuberculosis 13-15
<b>Chapter 2</b>	<b>REVIEW OF LITERATURE</b>
2.1	In Silico identification of potential allergens of American Cockroaches 16
2.2	Mining the Proteome of H.ducreyi for the identification of potential drug targets 16-18
2.3	Defination of the potential targets through subtractive genome 18-20
2.4	The subtractive Genomic Approach for Identification and Characterization of proteins 20-22
2.5	Functional analysis of Hypothetical Proteins 22-23
<b>Chapter 3</b>	<b>MATERIALS &amp; METHODS</b>
3.1	Bioinformatics 25-26
3.2	Steps for the extraction of Un-Characterized proteins 27-30
3.3	About ExPasy Proteomics Server 30-31
3.4	About ExPasy Proteomics Tools 31-40
3.5	Databases of PIR 41-42
3.6	Tools used for the extraction of Un-Characterized proteins 42-52

<b>Chapter 4</b>	<b>RESULTS AND DISCUSSIONS</b>	53-59
4.1	Blast output of the Hypothetical Sequences	54-55
4.2	Motif Searching output for the Sequences	56-57
4.3	Domain Searching results for smart server	57-58
4.4	Results for TMHMM server	58-59
<b>Chapter 5</b>	<b>CONCLUSION</b>	60-61
	<b>References (1pages)</b>	62

## ABSTRACT

*Mycobacterium tuberculosis* (MTB) is a pathogenic bacteria species in the genus *Mycobacterium* and the causative agent of most cases of tuberculosis. The genome of the H37Rv strain was published in 1998. Its size is 4 million base pairs, with 3959 genes; 40% of these genes have had their function characterised, with possible function postulated for another 44%. Within the genome are also 6 pseudo genes. The genome contains 250 genes involved in fatty acid metabolism, with 39 of these involved in the polyketide metabolism generating the waxy coat. Such large numbers of conserved genes show the evolutionary importance of the waxy coat to pathogen survival. The current work suggests a computational approach to annotate the putative function of the *Mycobacterium tuberculosis*. Over all 30 sequences were collected from the swiss prot data base. The insilico based annotation were performed by using BLAST, SMART, THMHM, and prediction of the motif. The result suggest that most of the uncharacterised protein resembles more to the chromosome assembly protein and also receptors. Again the motif and domains in the uncharacterise proteins has been predicted. Since the prediction of the function of this uncharacterise protein might be help ful to findout the specific drug target against this deadliest pathogen

**Key words :** *Mycobacterium*, uncharacterised protein, insilico annotation, function prediction

## List of Figures

		<b>Page No</b>
Fig 1	Blasting of protein	6
Fig 2	Protein analysis by Pfam and Motif	8
Fig 3	Conserved domain database tool	8
Fig 4	Clusters of orthologous group tool	11
Fig 5	Inter Proscan tool for proteins	14
Fig 6	Smart bioinformatic tool for analysis	15
Fig 7	Protein information resource analysis	15
Fig 8	SignalP for proteins	16
Fig 9	TMHMM tool	16
Fig 10	Protein clusters tool for protein analysis	17

## List of Tables

Table 1	List of uncharacterized protein sequences of mycobacterium tuberculosis	20
Table 4.1	BLAST OUTPUT FOR THE HYPOTHETICAL SEQUENCES	19
Table 4.2	MOTIF SEARCHING OUTPUT FOR THE SEQUENCES	24
Table 4.3	Domain searching results from SMART server	
Table 4.4	RESULT of protein sequence from TMHMM server	



# Chapter 1

## INTRODUCTION

## 1.1 *Mycobacterium tuberculosis*

*Mycobacterium tuberculosis* is a slow-growing voluntary intracellular parasite and it's the causative agent of most cases of tuberculosis. It was first discovered in 1882 by Robert Koch. The *M. tuberculosis* is highly aerobic and needs high levels of oxygen. During infection, it is exposed to many different environmental conditions depending on the stage and the severity of the disease. It is able to multiply inside the macrophage phagosome, in which the environment is generally hostile for most bacteria. The cells of *M. tuberculosis* are resistant to Gram staining as it contains a peculiar, waxy covering over the cell surface mainly mycolic acid.

*M. tuberculosis* constrains oxygen source to grow. It does not incorporate any bacteriological stain because of large lipid content in its wall, and thus is neither Gram positive nor Gram negative. They are categorized as acid-fast Gram-positive bacteria due to their lack of an outer cell membrane. It divides in every 15–20 hours, which is exceedingly slow compared to other bacteria. It is a small bacillus that can withstand weak disinfectants and can survive in an altered state for weeks. Its unique cell wall, affluent in lipids like mycolic acid, is likely liable for this tolerance and is a key virulence factor.

*M. tuberculosis* comes from the genus *Mycobacterium*, which is composed of relatively 100 recognized and recommended species. It possesses a biogeographic population configuration and different strain lineages are associated with distinct geographic regions. Their disruptions are often caused by overstimulated deadly strains of *M. tuberculosis*.

*M. tuberculosis* complex includes several species, all probably derived from a soil bacterium:

1. *Mycobacterium tuberculosis*

2. *Mycobacterium bovis*- unpasteurized milk
3. *Mycobacterium bovis*-BCG-used to treat bladder cancer
4. *Mycobacterium africanum* and *Mycobacterium Canetti*- rare causes of Tuberculosis in Africa
5. *Mycobacterium microti*- pathogen for rodents

Aerobic nature, non-motile and non-spore forming bacillus are certain characteristics of *M. tuberculosis*. They have a slow growth rate i.e. generation time of 20 hours vs. *E.coli* generation time of 20 minutes.

Basically it is a pathogen of the mammalian respiratory system, which affects the lungs. It can also manifold extracellularly in the open lung cavities that take place during the late stages of the disease. *M. tuberculosis* can be transmitted to other tissues or organs such as lymph nodes, bones, joints, skin, the central nervous system, the urinary tract and the abdomen. The general ways through which *M. tuberculosis* infection can be transmitted.

1. Inhalation of droplet nuclei from infectious person with active pulmonary tuberculosis,
2. Cough: most efficient at 3000 infectious droplet nuclei per cough
3. Talking: similar quantity over 5 minutes
4. sneezing more efficient than coughing
5. Bacillus remains alive and infectious in air for long period: Ventilation key in preventing transmission and isolation of patients

The Primary infection of *M.tuberculosis* reveals different symptoms. Before immune response, Bacillus attains alveoli and then they reproduce extracellularly in alveolar space and intracellularly in Alveolar macrophage. Due to the

inadequacy of critical host immune response, alveolar macrophage consumes TB bacillus and bacillus remains in phagosome.

Phagosome usually assimilates proton-ATPase into membrane accompanying to decline pH and acidification occurred within phagosome. Acidified phagosome then normally integrates with cell lysosome, imperiling organism to lysosome's toxic enzymes. But *M.tuberculosis* anticipates insertion of proton-ATPase into phagosome. So, Phagosome never gets acidified and never merges with lysosome. It multiplies for weeks, both in initial focus in alveolar macrophages and in cells transported lymphohematogenously throughout body. Metastatic foci well established in regional nodes and then to tissues which retain bacilli and facilitate their multiplication in apical posterior areas of lungs, lymph nodes in neck, kidneys, epiphyses of long bones and vertebral bodies areas adjacent to subarachnoid space. These will be areas of reactivation disease in future as organisms implanted remain alive but dormant once immune response occurs. Reactivation can take place in any one of these areas of the body with or without reactivation in others.

## 1.2 Tuberculosis

### **History**

There are evidence for spinal TB in Egyptian mummies and pre-Columbian remains. This disease wasn't a significant problem until the 17th and 18th centuries as urbanization and crowding in unventilated living conditions increased. By the 19th century with industrialization, TB caused one quarter adult deaths in Europe.

Germ theory of diseases and discovery of TB bacillus by Koch were the few progressive works during this age.

## **Introduction**

Tuberculosis, is a deadly infectious disease which is being caused by various strains of mycobacterium, usually Mycobacterium tuberculosis in humans and is a very common disease. It mainly attacks the lungs but it can also affect the other parts of the body. It is contagious disease which is found in the air when people who have an active MTB infection cough, sneeze, or otherwise transmit their saliva through the air. Most infections in humans result in an asymptomatic, latent infection, and about one in ten latent infections eventually progresses to active disease, which, if left untreated, kills more than 50% of its victims.

## **Symptoms**

The common Symptoms of this deadly disease are:

1. Systemic symptoms non-specific includes fever, fatigue, night sweats, weight loss
2. Pulmonary symptoms: cough, productive or dry-most patients have cough but may be ignored by patient for weeks
3. Hemoptysis:
  - i.) mild-moderate, chronic blood streaking results from caseous sloughing or endobronchial erosion; seen in advanced disease
  - ii.) Sudden massive hemoptysis- erosion of pulmonary artery

## Diagnostic methods

There are certain diagnostic procedures (staining, cultures and molecular diagnostics) that can help in predicting the tuberculosis infection extent.

Acid fast stain is a method in which Acid fast implies mycobacterial species although nocardia is weakly acid fast and many other species besides *M. tuberculosis* complex will all be AFB positive. Nucleic acid amplification method can detect *M. tuberculosis* complex in fresh sputum. This diagnostic process is actually a part of developed world technology and is too costly for resource poor countries.

DNA fingerprinting is a Molecular epidemiologic tool that works on the principle of Restriction fragment length polymorphism. It's also used in developed nations in general.

## Death rates by tuberculosis

The most shocking thing is that *M. tuberculosis* infects one third world's population. It causes around 8 million new cases of active disease annually. It is one of the deadly disease that causes almost 2 million deaths just second only to HIV which is the cause of death from infectious agent worldwide among adults. HIV/TB relationship has exacerbated problem with TB increasing in areas with high AIDS incidence especially in sub-Saharan Africa.

Absolute numbers of cases of TB are highest in Asia as population density is highest there but case rates are highest in sub-Saharan Africa i.e. 300 per 100,000. Estimated incidence rates in sub-Saharan Africa vs. 100-299 per 100,000 in Asia. In most nations of developed Europe

There is a downward trend in incidence even before advent of antibiotics. 10% of infected people is responsible for the development of this active disease and mainly cavitory cases are the infectious one (only 50% cases are cavitory). Each cavitory case needs to infect 20 to maintain constant rate of cases. Data from Pre-WW2 Holland shows 1 infectious case produced 13 new infections.

Annual decrease in mortality and morbidity of 4%-6% in developed countries. In between the 1900 and World War II, various changes took place among the people of different regions. Progressively higher natural residual resistance prevailed in those who had survived infection. Better living conditions came into existences that were less conducive to airborne spread.

Advent of antibiotics in late 1940s (Streptomycin) and INH in 1952, Tuberculosis is become curable. In case of United States, it was revealed that there was steady decline in the death rates caused by the tuberculosis until 1984 when it was slowly increasing in terms of number of incidence.

The prominent causes behind was the negligence of TB control programs. Beside it, increase in urban homelessness and resultant crowding into homeless shelters were some other reasons of its spread. Currently, the restored TB control program funding and decline in number of homeless brings background rates high among immigrants from high prevalence countries. One half cases in US are now among foreign born. Dramatic change between 1993 and 2003 in New York, New England, west coast states , all have greater than 50% cases foreign born in 2003( 300 per 100,000 estimated incidence rates).

# Chapter 2

## **LITERATURE REVIEW**



## 2.1 In silico identification of potential Allergens of American cockroaches

The study in fact focused on the identification of potential allergens among the characterized proteins of *Periplaneta Americana* using web based and allergen prediction tools for the prediction of allergic proteins. With the help of UniprotKB, protein sequences of *P. Americana* were recovered. Then after these sequences acquired were examined by Allpred. Similarly another tool SDAP was used for confirmation.

Using UniprotKB , 233 cases of protein sequences of *p. Americana* were found out of which 25 were known allergens.102 are predicted as potential allergens by Allpred out of remaining 208 proteins.

However, only 9 were found to be potential allergens after screening with SDAP. This aims at the development of the bioinformatics tools to identify the potential allergens.

The challenges in our way is the identification of the various characteristics of the uncharacterized proteins of *M. tuberculosis* that may have the potential to cause the different allergies or infections that could be a part of health hazard in coming days.Our in silico identification of the uncharacterized proteins may lead to certain new deliberation of information in the field of research.

## 2.2 Mining the proteome of *H.ducreyi* for the identification of potential drug targets

*Bacterium haemophilus ducreyi* caused a severely virulent sexually transmitted disease (STD), chancroid predominant mainly in Africa, United States and in

certain parts of south Asia. It has been spotted as a cofactor for human deficiency virus transmission.

So, there is a need to develop an effective drug to encounter chancroid. The availability of proteome information of *H. ducreyi* help facilitated in silico analysis for recognition of potential vaccine models and drug targets.

Complete proteome of *H. ducreyi* was recovered from SwissProt and the complete *Homo sapiens* proteome was recovered from NCBI. The prokaryote essential proteins were conveyed from the database of Essential Genes (DEG). Metabolic pathway analysis of essential proteins of *H. ducreyi* was done by the KEGG Automatic Annotation server. Sub cellular localization analysis of the essential proteins of *H. Ducreyi* has been done by proteome Analyst Specialized Subcellular Localization server to determine the surface and membrane associated proteins which could be possible vaccine candidates. Apart of these, Functional family allocation of the putative uncharacterized essential proteins was done by using the SVMProt Web server.

1226 proteins in *H. ducreyi* are found as non-homologous with human proteome. This resulted in the recognition of 451 essential proteins by screening these proteins using the Database of Essential Genes (DEG). With the help of KEGG Automated Annotation server, 40 proteins of *H. ducreyi* acknowledged as potential drug targets by screening these proteins as they are involved in pathogen specific metabolic pathways.

Subcellular localization forecast of these 451 essential proteins revealed that 11 proteins prevailed on the outer membrane of the pathogen which could be potential vaccine models. Functional family estimation for the 50 putative uncharacterized essential proteins of *H. ducreyi* by SVM-Prot web server showed that out of 50, 3 proteins as Transmembrane proteins, which may be potential drug targets.

Through our efforts of collecting information about the uncharacterized proteins by bioinformatics tools, study of homologous or non homologous character could be possible. Besides it, light can also be put over their identification as potential drug targets by screening the uncharacterized proteins of *M. tuberculosis* by using various Web based servers. This shall really help in the field of drug and vaccine modeling.

### **2.3 DEFINATION OF POTENTIAL TARGETS IN MYCOBACTERIM TUBERCULOSIS THROUGH SUBSTRUCTIVE GENOMES:-**

We have seen that the genome sequencing technology provides some very high information for the finding of some new therapeutic targets in many pathogens over & above all human genome. The very most effective method is by Subtractive genomic approach in which we find some essential genes or proteins which are present in pathogens but are absent in host cells which are used as targets for drug delivery. But in some uncharacterized proteins we have also seen that drug targeting is also possible in pathogenic cells. So, there are around 32 uncharacterized proteins in which drug targeting is possible in their pathogenic cells. In the year December 2009 the complete genome sequence was known of about 2274

viruses ([http://www.ncbi.nlm.gov/genomes/MICROBES/microbial\\_taxtree.html](http://www.ncbi.nlm.gov/genomes/MICROBES/microbial_taxtree.html)),

1007 bacterial species & around 56 eukaryote organisms out of which half of them are fungi (<http://www.ncbi.nlm.nih.gov/genomeprj>) & for these lots of bioinformatics tools have been developed for the analyzing of those genomes. As it is the very important part of the human life i.e. HUMAN GENOME COMPLETION for the drug discovery. There are many more ways to find out the potential drug target like virulence genes, uncharacterized essential genes, some species-specific

gene & some of the unique enzyme transporter. We have also seen in some of the proposed work subtractive genomic approach is also used for the subtraction of dataset comparing of two genomes i.e. pathogen & human. There had been many minimal approaches which have been done for the target delivery for the self-replicating cell & the complete genome has been sequenced. This is basically done for the deduction of conserved genes in the analyzed genome.

There have been many methods for the gene target:-

➤ SEQUENCE RETRIEVAL OF HOST & PATHOGEN:-

The complete genome pathogen has been retrieved from NCBI (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION) & Swiss-Port PROTEIN Knowledgebase (<http://www.expasy.ch/sport/>). For the completion of whole genome sequence data all the genes of organisms have been coded for different proteins whose sequences were more or less greater than 100 amino acids and all are being selected out. By these methods we can find out the all those proteins whose amino acid is less than 100 in length were all unlikely to represent the essential proteins. It may be the unique organism as well.

➤ IDENTIFICATION OF DUPLICATE PROTEIN:-

In these methods we can find out the duplicate proteins within the organism. These set of duplicate proteins were used for the analysis of some other things.

➤ SIMILARITY SEARCH:-

This process is basically used to find out the similarity search of sequences. It is basically done by NCBI BlastP(<http://www.ncbi.nlm.nih.gov/blast>). It is basically done against Homo sapiens protein sequences using different threshold expectation value.

There are many more methods for finding out the gene target proteins.

From these all methods we have seen that there are lots increasing bacterial genomes which are available in all the public databases that offers all new opportunities to find the relationship between the genome type & phenotype using different in silico genome comparisons. The presence & absence of different genes can be analyzed by using the subtractive method which helps in finding the link content of different genome content & phenotypic features. This method is basically responsible for gene expression for some specific functions & is conserved during different evaluation lost in all those genes. This method also helps in finding out the genes which are available in some group of genomes but are not present in other group of genomes. In silico subtractive & differential analysis of genomes are very powerful methods which help in the identification of some genus & species-specific genes or with those groups of genes that are all responsible for a unique type of phenotype. With these methods we can search for all types of genes that are present in one group of bacteria but are absent in other group of bacteria.

#### 2.4 THE SUBTRACTIVE GENOMEICS APPROACH FOR IN.SILICO IDENTIFICATION & CHARACTERIZATION OF PROTEINS:-

There are lots of diseases which are increasing with high rate varying from 1 to 1000 per 10000 persons in different parts of the world. So for them lots of vaccines have been made for the protection of these types of deadly diseases. For the decrease of these diseases & bacterial infections lots of steps & vaccines have been made for the control of these diseases which will affect the public health problem for most of the countries. Now this Mycobacterium tuberculosis has been very rapidly controlled because before when there is no vaccine has been there it created a lot of problem in the human life. So till date the complete genome sequence of

about 863 bacteria has been determined & about 1653 bacterial genome projects are currently in progress. So for these availability of genome sequences of pathogens has been provided a tremendous amount of information can be used in drug target & vaccine target identification of the proteins. Now a day it is being seen that a lot of subtractive genomics approach database of essential gene & their pathway analysis have been studied for drug & target vaccine delivery.

For all these there are lots methods have been made for the drug delivery that includes:-

- Retrieval of Proteomes of Host & Pathogen
- Identification of Essential proteins in the species
- Functional Classification of the uncharacterized essential proteins
- Sub Cellular Localization Prediction
- Metabolic Pathways

From all these we have seen that all the proteins that are non-homologous to the human proteome could not be taken directly as targets as these also include a large number of proteins which are not essential for the visibility of the organisms. Its functional classification of the 32 uncharacterized proteins was performed by using the SVMProt web search based on the P value which is expected to be the classification as Trans membrane proteins, zinc binding & many more uses. It is also seen for the metabolic pathway analysis can be done by KEGG Automatic Server. We can find out the result of the metabolic pathways of these host & pathogen can be done by using the Kyoto Encyclopedia of Genes & Genomes Pathway databases. It is basically seen that the entire bacterial component is the very important part of the survival under some extreme conditions.

From all these we can further investigate that all the predicated proteins which are very essential are required for the reliability of the data. Therefore the complete list of the identical & identified proteins is being done by these in silico approach which is the essentially available as the supplementary method.

## 2.5 The identification & functional analysis of hypothetical genes expressed in mycobacterium tuberculosis:-

We have seen that lots of progress has been made for the uncharacterized hypothetical genes for the rapid accumulation of genebank. These genes not very much functional & also cannot be broken into simple sequence comparison alone so for these lot of significant tools have been developed for finding out the comparisons of these uncharacterized proteins which are freely available in the public databases. The hypothetical genes which are exposed to the cells are all in normal condition to the environment. So for finding out the comparison there are lots of tools which are publically available in all the databases. Now a day there have been a lots of research have been going on genome researches being going on for the sequencing of the complete genome of the cellular life form. There are lots of proteins which are being used but the rest proteins which are not in use are either homologous to genes of unknown function which are referred to as conserved hypothetical genes. Those proteins which are actually encoded but they are latter genes are called as hypothetical, uncharacterized or unknown proteins.

As we have seen that conserved hypothetical genes are the major challenge to the complete genomes. This is because they play a very important role for the function of those genes which are still obscure or it is quite unsettling as it helps in

understanding the basic idea of microbiology. These conserved hypothetical proteins are very much clearly detected to be grown in aerobically cells whose genes are found to be essential in transposon mutagenesis studies.

The methods which will help in these processes are:-

- Gene expression analysis
- Protein expression analysis
- Annotation of conserved hypothetical genes in public databases
- Structural genomics data
- Protein-protein interactions
- Uncharacterized conserved genes
- Functional characterization of genes

From these we conclude that we can identify the hypothetical genes expressed in the bacteria. We can also find out the sequence analysis of the conserved genes & also the genome context analysis



# Chapter 3

## **MATERIALS & METHODOLOGY**

### 3.1 BIO INFORMATICS:-

Bioinformatics is a scientific discipline that supports & advances biomedical research with management & (statistical) analysis of experimental data. Bioinformatics combines expertise & technologies from molecular biology, data analysis, database technology & information technology.

The course provides biomedical researchers with sufficient theoretical & practical skills to adequately apply bioinformatics in their own research. This course combines lectures with computer exercises & aims at introducing the participant in the basic principles underlying bioinformatics tools.

Bioinformatics combines the tools of Biology, Chemistry, Mathematics, Statistics & Computer Science to understand Life & its processes.

#### **Some important biological databases for analyzing biological data**

- GeneBank, EMBL, DDBJ-used for nucleotide database.
- Swissport, PIR, PRF-used for protein database.
- PDB, MMDB-used for structural database.
- SCOP, CATH, FSSP-used for classification database.
- PROSITE, PRODOM, PFAM, INTERPRO, CDD etc-used for different types of protein classification.
- KEGG-used for pathway studies.
- OMIM-used for inherit disease database.
- PUBCHEM COMPOUND, DRUG BANK, ZNIC, LIGAND-used for drug database.
- dbEST,dbSNP-used for expressional database.
- MGD,YGD,HGD,ACeDB-used for complete genome database.
- PUBMED, PUBMED CENTRAL, MEDLINE-used for literature database.

#### **ADVANTAGE OF BIOINFORMATICS**

- To solve the biological problems faced by scientist group.
- To unravel the hidden truth of life.
- To develop the value of human life by applying the knowledge in drug designing.
- It is an interdisciplinary field includes all the branches of science.

#### **APPLICATION AREAS OFBIOINFORMATICS**

- Molecular medicine
- Personalized medicine
- Preventative medicine
- Gene therapy
- Drug development
- Microbial genome applications

They contain information from research areas including:

- Genomics
- Proteomics
- Metabolomics
- Microarray
- Gene expression
- Phylogenetics

Information contained in biological databases includes gene function, structure, localization(both cellular & chromosomal),clinical effects of mutations as well as similarities of biological sequences & structures.

**BIOLOGICAL DATA:-**

- Nucleic acids:
  - DNA sequences, genes, gene products (proteins), mutation, gene coding, distribution patterns, motifs.
  - Genomics: genome, gene structure & expression, genetic map, genetic disorder.
  - RNA sequence, secondary structure, 3D structure, interactions.
- Proteins:
  - Protein sequence, corresponding gene, secondary structure, 3D structure, function, motifs, homology, interactions.
  - Proteomics: expression profile, proteins in disease processes, etc.
  - Ligands & drugs (inhibitors, activators, substrates, metabolites).

**CATEGORIZATION:-**

Based on Data Type

- Genome database
- Taxonomy database
- Sequence database
- Micro array database
- Chemical database
- Expression database
- Enzyme database
- Pathway database
- Disease database
- Literature database
- Protein database

There is around 1,936 mycobacterium tuberculosis proteins out of which there are around 32 uncharacterized mycobacterium tuberculosis proteins which we have to extract from the “EXPASY” server.

### 3.2 STEPS FOR EXTRACTION OF UNCHARACTERIZED MYCOBACTERIUM TUBERCULOSIS PROTEINS:-

- First of all we have to go to google & there we have to search “EXPASY”.
- After going to expasy we have to type “UNIPORT KB”.
- After going inside the uniprot KB we have to write mycobacterium tuberculosis + uncharacterized protein.
- After that we will get around 32 uncharacterized mycobacterium tuberculosis proteins.

#### LIST OF UNCHARACTERIZED MYCOBACTERIUM TUBERCULOSIS:-

Accession	Entry name	Protein names	Gene name	Organism	Length
O53766	Y0569_MYCTU	UNCHARACTERIZED PROTEIN Rv0569/MT0595	Rv0569 MT0595	Mycobacterium tuberculosis	88
Q79F93	PE35_MYCTU	UNCHARACTERIZED PE FAMILY PROTEIN PE35	<b>PE35</b> Rv3872 MT3986	Mycobacterium tuberculosis	99
P96243	Y3835_MYCTU	UNCHARACTERIZED MEMBRANE PROTEIN Rv3835/MT394	Rv3835 MT3943	Mycobacterium tuberculosis	449
O53618	Y073_MYCTU	UNCHARACTERIZED ABC TRANSPORTER ATP-BINDING PROTEIN	Rv0073 MT0079	Mycobacterium tuberculosis	330
O53617	Y072_MYCTU	UNCHARACTERIZED ABC TRANSPORTER PERMEASE Rv00 PROTEIN	Rv0072 MT0078	Mycobacterium tuberculosis	349
O33209	O33209_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	<b>scpB</b> MT1751 Rv1710	Mycobacterium tuberculosis	231
O33208	O33208_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	<b>scpA</b> MT1750 Rv1709	Mycobacterium tuberculosis	278

P96874	P96874_MYCTU	10 Kda IRON REGULATED PROTEIN Irp 10	<b>IrpA</b> Rv3269	Mycobacterium tuberculosis	93
P96875	Q7TZRI_MYCBO	PUTATIVE UNCHARACTERIZED PROTEIN Mb 1737	Mb1737	Mycobacterium tuberculosis	231
P96876	A5WN29_MYCTF	PUTATIVE UNCHARACTERIZED PROTEIN	TBFG_11725	Mycobacterium bovis	231
P96877	C6DRW2_MYCTK	PUTATIVE UNCHARACTERIZED PROTEIN	TBMG_02285	Mycobacterium tuberculosis (strain F11)	231
P96878	A1KJC7_MYCBP	PUTATIVE UNCHARACTERIZED PROTEIN BCG_1749	BCG_1749	Mycobacterium tuberculosis (strain KZN 1435 / MDR)	231
P96879	C1ANY3_MYCBT	PUTATIVE UNCHARACTERIZED PROTEIN	JTY_1724	Mycobacterium bovis (strain BCG / Pasteur 1173P2)	231
P96880	Q7TZR2_MYCBO	PUTATIVE UNCHARACTERIZED PROTEIN Mb1736	Mb1736	Mycobacterium bovis (strain BCG / Tokyo 172 / ATCC 35737 / TMC 1019)	278
P96881	A1KJC6_MYCBP	PUTATIVE UNCHARACTERIZED PROTEIN BCG_1748	BCG_1748	Mycobacterium bovis	278
P96882	C1ANY2_MYCBT	PUTATIVE UNCHARACTERIZED PROTEIN	JTY_1723	Mycobacterium bovis (strain BCG / Pasteur 1173P2)	278
P96883	A2VIJ4_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBCG_01664	Mycobacterium bovis (strain BCG / Tokyo 172 / ATCC 35737 / TMC	231

				1019)	
P96884	A4KMH3_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBHG_01668	Mycobacterium tuberculosis C	231
P96885	A5WN28_MYCTF	PUTATIVE UNCHARACTERIZED PROTEIN	TBFG_11724	Mycobacterium tuberculosis str. Haarlem	278
P96886	D5YFM2_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBGG_00917	Mycobacterium tuberculosis (strain F11)	231
P96887	D5XTZ3_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBDG_03303	Mycobacterium tuberculosis EAS054	231
P96887 P96889	D5ZHK6_MYCTU D5Z4J7_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBJG_00128	Mycobacterium tuberculosis T92	231
P96890	B2HRU5_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBIG_01103	Mycobacterium tuberculosis T17	231
P96891	A2VIJ3_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	MMAR_2524	Mycobacterium tuberculosis GM 1503	275
P96892	A4KHM2_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBCG_01663	Mycobacterium marinum (strain ATCC BAA-535 / M)	278
P96893	D5YFM1_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBHG_01667	Mycobacterium tuberculosis C	278
P96894	D5XTZ2_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBGG_00916	Mycobacterium tuberculosis str. Haarlem	278
P96895	D7ERE3_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBDG_03302	Mycobacterium tuberculosis EAS054	278

P96896	D5ZHK5_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBAG_00626	Mycobacterium tuberculosis T92	278
P96897	D5Y410_MYCTU	PUTATIVE UNCHARACTERIZED PROTEIN	TBJG_00127	Mycobacterium tuberculosis 94_M4241A	278

**TABLE-1 UNCHARACTERIZED PROTEIN LISTS OF MYCOBACTERIUM TUBERCULOSIS**

Before proceeding forward we must have to know what is “EXPASY” & “UNIPORT”

**EXPASY SERVER:-**The EXPASY (Expert Protein Analysis System) is a proteomics server of the Swiss Institute of Bioinformatics which analyzes protein sequences & structures & two-dimensional gel electrophoresis. The server functions in collaboration with the “EUROPEAN BIOINFORMATICS INSTITUTE”. Expasy also produces the protein sequence knowledgebase, UniportB/Swissport, & its computer annotated supplement, UniportKB/Trembl.

### 3.5 DATABASES OF PIR

The protein database of PIR is categorized into three groups:-

- UNIVERSAL PROTEIN RESOURCE
- iProClass
- PIRSF protein family

- **UNIVERSAL PROTEIN RESOURCE:-**

It is basically a central repository of protein sequences & functions. It is being enriched by information shared from those contained in Swiss-Port, TrEMBL, PIR & many more sources. These databases consist of mainly three database layers:-

- **UniPORT Knowledgebase(UniPortKB):-**

These database mainly provides the central database protein sequences with the annotation & functional information of the sequences. PIR-PSD are the sequences which are mainly missing from Swiss-Port & TrEMBL are being found to be in the UniProt database. It has basically two parts:-

- ❖ First part contains manually annotated records & is referred to as “UniProt/Swiss-Prot”.
- ❖ The second part contains the computationally analysed records which have to be manually annotated & is referred to as”UniPort/TrEMBL”.

The knowledgebase aims to be in a single record of all protein products which are derived from a certain gene from a certain species & gives not only the whole record of an accession number, but also assigns some

alternative splicing, proteolytic cleavage & post-translational modification isoform identifiers to each form of the derived proteins.

▪ UniPORT Reference Clusters(UniRef):-

These databases provide some non-redundant data collections based on the UniPort Knowledgebase & UniParc to obtain complete coverage of the sequence space at several resolutions. There are 3 separate datasets which compress sequence space at different resolutions. The sequences that are 100% which are named as UniRef100 database. The sequences which are  $\geq 90\%$  are named as UniPortRef90. & the sequences which are  $\geq 50\%$  are named as UniRef50 are identical regardless of source organism & are merged with each other. UniParc records that represents sequences are over-presented in the Knowledgebase, DDBJ/EMBL/GenBank Whole Genome Shotgun data. Ensembl protein translations which form various organisms & are also the International Protein Index data. UniRef90 & UniRef50 databases provide a more even sampling of sequences that can be reduce the number of closely related sequences. This sequences speeds up the similarity searches & these searches are made more informative.

▪ UniPORT Archive(UniParc):-

This database provides a stable, comprehensive, non-redundant sequence collection by storing the complete body which are publically available protein sequence data. In these database if we add some new or revised protein sequences a UniParc sequence version is provided or increased & thus makes it possible to track the history of sequence changes in all the sources which are available in the databases. In order to avoid redundancy with each unique sequence is assigned to a unique identifier & it is stored only once in a lifetime. The basic information which are stored with each UniParc entry are the indentifires, the sequences, the cyclic redundancy check number, the source databases with their accession & version numbers & a time stamp. The other informations can be retrieved out from the other source databases. Each source databases accession number is being given with some code in that database.

• iProClass(Integrated Protein Knowledgebase):-

It provides some compreshive description of a protein family, function & structure for the UniPort protein networking environment. These iProClass database contains value-added description of proteins which includes family relationships at global & local levels. And also the structural & functional classifications & the features. These databases were first released on October 2000 & it contains data from the PIR protein sequence database & Swiss-Port. It basically presents two types of protein sequence reports. The first type is the information on family, structure, function, gene, genetics, disease, ontology, taxonomy & literature with the crossreferences to the relevant molecular databases & executive summary lines & it also has a graphical display of domain & motif sequence regions & a link to the related sequences in the pre-computed FASTA clusters. The second type is a super-family report which presents PIR superfamily membership information with the length, taxonomy & the keyword statistics, complete member listing



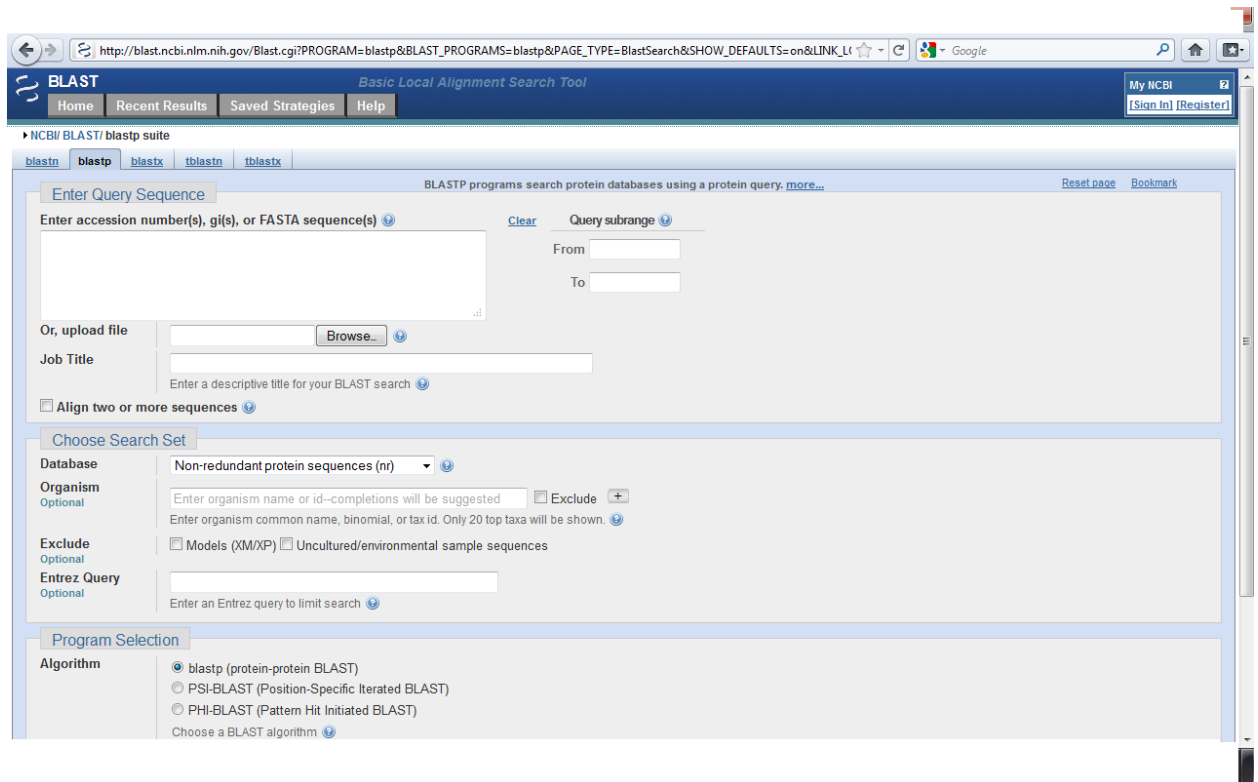
separated into a major kingdoms, family relationships at the whole protein & domain & motif levels with the direct mapping to the other classifications, structures & function cross-references, graphical display of domain & motif architecture of the members. It also provides a link to dynamically generated multiple sequence alignments & phylogenetic trees for super-families with the curated seed members.

### 3.6 TOOLS USED FOR THE EXTRACTION OF UNCHARACTERIZED PROTEIN SEQUENCES OF MYCOBACTERIUM TUBERCULOSIS:-

- BLAST
- PFAM
- CDD
- COG
- INTERPROSCAN
- SMART
- PIR
- SIGNAL P
- TMHMM
- PROTEIN CLUSTURE

❖ BLAST(Basic Local Alignment Search Tool) {<http://blast.ncbi.nlm.nih.gov/Blast.cgi> } :-

It is an algorithm developed for comparing the primary biological sequence information in the amino acid sequence of protein or nucleotide of DNA sequences. A blast search enables a researcher to compare a query sequence with a library or database of sequences & identify library sequences that resemble the query sequence above a certain threshold sequences. It was developed by Myers E, Altschul S.F, Gish W, Miller E.W, Lipman D.J.NCBI. Its stable release is 2.2.24/23 August 2010. It works on UNIX, LINUX, Mac, MS-windows operating system. It is a public domain tool where it can be used by everybody at all places.



**FIG-1 BLAST SERVER IN WHICH THE SEQUENCES ARE BEING PUT FOR BLASTING**

❖ **PFAM:-**

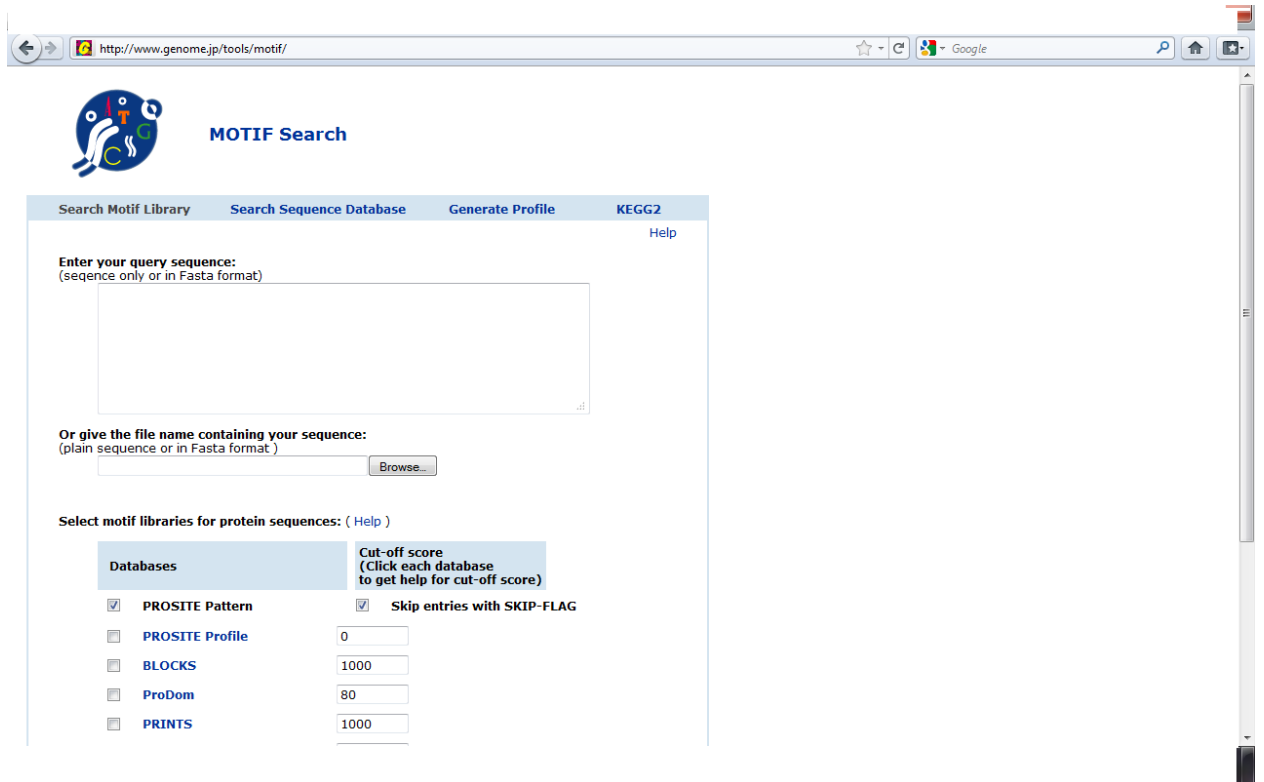
It is a database of protein families that includes their annotations & multiple sequence alignment generated using hidden Markov models. 74% of the protein sequences have at least one match of Pfam. This number is called the sequence coverage. It is the mutually curated portion of the database sequence alignment & a hidden Markov model is stored.

**Hidden Markov Model:-**

It is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states. An HMM can be considered as the simplest dynamic Bayesian network.

These PFAM can also be done by using MOFIF tool

**MOTIF:-** It is a sequence pattern of nucleotides in a DNA sequence or amino acids protein. Its structural part is formed by the spatial arrangement of amino acids. It recur within a network much more often than the expected at random part. It is basically the user interface toolkit used in the software development. It is also used in the element to move in the consideration of why the piece moves & how it supports the fulfillment of the problem stipulation.



**FIG-2 MOTIF SERVER WHERE THE SEQUENCE ARE BEING PUT AND LOTS OF MOFIF ARE FOUND**

❖ CDD(Conserved Domain Database):-

It is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains & full-length proteins. These are available as position-specific score matrices for the fast identification of conserved domains in protein sequences via RPS-Blast. CDD content includes NCBI-curated domains which use 3d-structure information to define domain boundaries & provide insights into sequence/structure/function relationship as well as domain models imported from a number of external source database(Pfam, SMART, COG, PRK, TIGRFAM).

The screenshot shows the NCBI Conserved Domains and Protein Classification website. At the top, there is a navigation bar with 'Structure Home', '3D Macromolecular Structures', 'Conserved Domains', 'PubChem', and 'BioSystems'. A search bar is present with 'Conserved Domains' selected. The main content area is divided into 'Resources' and 'Highlights'. The 'Resources' section includes:
 

- Conserved Domain Database (CDD):** A protein annotation resource consisting of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. It is available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM).
- CD-Search & Batch CD-Search:** NCBI's interface to searching the Conserved Domain Database with protein query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as specific hits.
- CDART: Domain Architectures:** Conserved Domain Architecture Retrieval Tool (CDART) performs similarity searches of the Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in protein queries. CDART finds protein similarities across significant evolutionary distances using sensitive domain profiles rather than direct sequence similarity. Proteins similar to the query are grouped and scored by architecture. You can search CDART directly with a query protein sequence, or, if a sequence of interest is already in the Entrez Protein database, simply retrieve the record, open its "Links" menu, and select "Domain Relatives" to see the precalculated CDART results (illustrated example). Relying on domain profiles allows CDART to be fast and, because it relies on annotated functional domains, informative.

 The 'Highlights' section includes:
 

- What is a conserved domain?** (with a 3D protein structure image)
- 3-D structures and conserved core motifs:** (with a 3D protein structure image)
- Conserved features (binding and catalytic sites):** (with a 3D protein structure image)

**FIG-3 CDD SERVER**

❖ COG(Clusters of Orthologous Group):-

These groups of proteins were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages & thus corresponds to an ancient conserved domain. There are around:-

- 66 genomes(microbial)
- 38 orders(microbial)
- 28 classes(microbial)
- 14 phyla(microbial)

Upcoming microbial genomes are:-

- Genomes-261
- Orders-63
- Classes-33
- Phyla-17
- Genera-126(new)

www.ncbi.nlm.nih.gov/COG/

Phylogenetic classification of proteins encoded in complete genomes

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

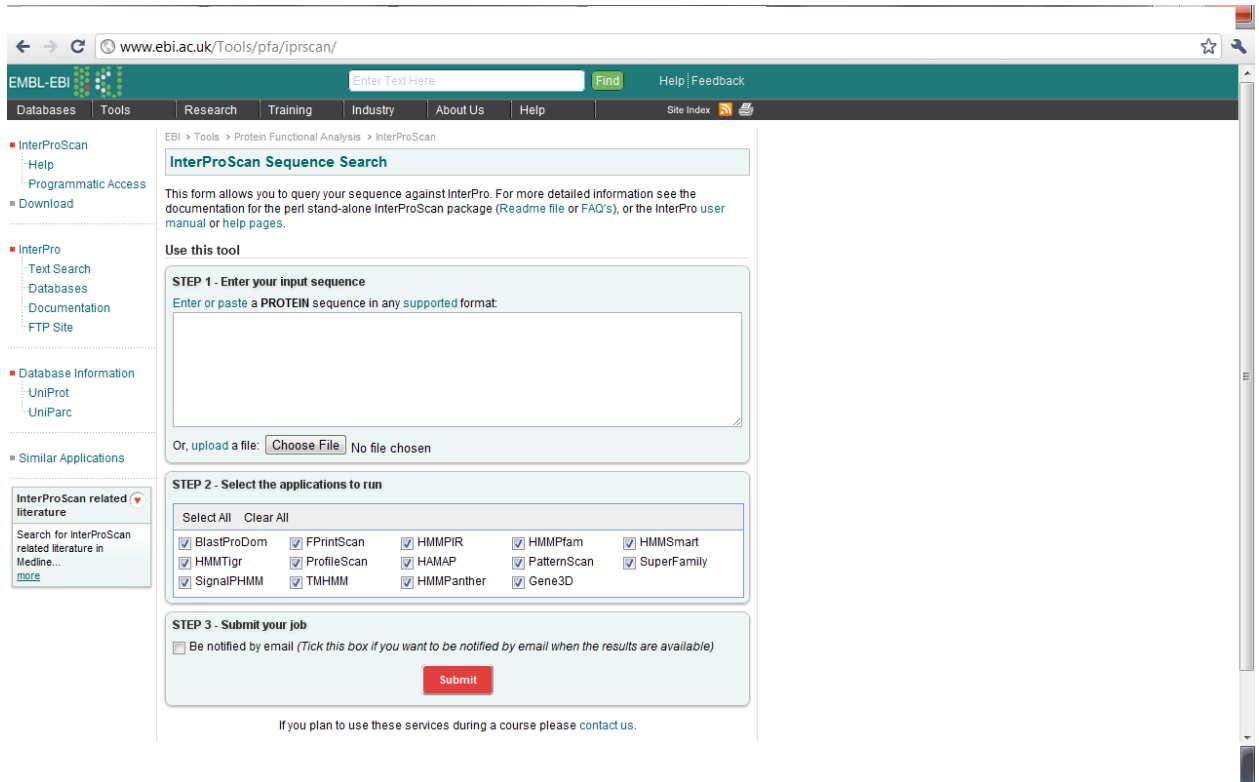
Unicellular clusters		FTP	
66 genomes		Initial	
38 orders		version	
28 classes			
14 phyla			
<a href="#">Science 1997 Oct 24;278(5338):631-7.</a> <a href="#">BMC Bioinformatics 2003 Sep 11;4(1):41.</a>			

Eukaryotic Clusters		FTP	
Code	Name	Abbreviation	
A	<i>Arabidopsis thaliana</i> (thale cress)	ath	
C	<i>Caenorhabditis elegans</i> (worm)	cel	
D	<i>Drosophila melanogaster</i> (fruit fly)	dme	
H	<i>Homo sapiens</i> (human)	hsa	
Y	<i>Saccharomyces cerevisiae</i> (baker yeast)	sce	
P	<i>Schizosaccharomyces pombe</i> (fission yeast)	spo	
E	<i>Encephalitozoon cuniculi</i> (Microsporidia)	ecu	
Upcoming eukaryotic genomes			
O	<i>Oryza sativa</i> (rice)	osa	
Q	<i>Anopheles gambiae</i> (mosquito)	aga	
Z	<i>Pan troglodytes</i> (chimpanzee)	ptr	

FIG-4 COG SERVER

❖ INTERPROSCAN:-

It is a tool that combines different protein signature recognition methods native to the InterPro member databases into one resource with look up of corresponding Inter Pro & Go annotation. It is a bioinformatics tool that provides a one stop-stop for the automated sequence analysis of protein & nucleic acid, the latter via a full six-frame translation. It offers the ability to identify both structural & functional regions of interest based upon the methods & models that have been generated by a large number of member groups. These members' databases use a variety of different bioinformatics techniques & algorithms which are optimized for specific feature types. It is therefore able to offer the researcher the ability to quickly characterize a new novel sequence with considerable confidence. Inter Proscan is being developed as an open source project at "EMBL EUROPEAN BIOINFORMATICS INSTITUTE".



**FIG-5 INTERPROSCAN SERVER**

❖ **SMART:-**

It can be divided into two different modes:-

- Normal SMART:-this database contains Swiss-Port, SP-TrEMBL & stable Ensembl proteomes. The protein database in Normal SMART has significant redundancy even though identical proteins are removed.
- Genomic SMART:-The only proteomes of completely sequenced genomes are used Ensembl for metazoans & Swiss-Port for the rest.

---

## FIG-6 SMART SERVER

---

❖ PIR(Protein Information Resource):-

It is a non-redundant annotated protein sequence database & is an analytical tool which is maintained by the collaboration of MIPS in Munich & the Japanese. The UniPort provides the scientific community with a single, centralized & authoritative resource for protein sequences & functional information.

pir.georgetown.edu/pro/pro.shtml

**PIR** A UniProt Consortium Member  
Protein Information Resource

TEALPN---PRAVADHLLM  
LIGGLRNCASVTAARQDAE  
VTQFSN---ARTTAQRVKK

Protein Search Site Search

About PIR Databases Search/Analysis Download Support

NOTICE: The PRO ID format has changed from PRO: to PR: (e.g. PRO:000000563 is now PR:000000563).

HOME / Protein Ontology (NIH grant #R01 GM080646-01)

PRO provides the ontological representation of proteins (including specific modified forms and orthologous isoforms) and protein complexes, meaning that it explicitly defines these protein-related entities and shows the relationships between them. Each PRO term/entry represents a defined entity (e.g. the constitutive active form of rho-associated protein kinase 1 isoform 1 resulting from proteolytic cleavage is described in PR:000000563, whereas the unmodified form is defined by PR:000002529) (current release: 18.0).

- Consortium
- Dissemination
- PRO Wiki
- Documentation
- Downloads
- PRO tutorial
- PRO Paper
- PRO Statistics

Browse PRO  
-- Quick Browse  
Example: methylated (sample output)

Retrieve a PRO entry (enter a PRO ID):  
Example: PR:000000563 (sample output)

Search PRO (enter text or ID):  
Example: smad (sample output)

Annotation: RACE-PRO PRO tracker

PRO encompasses three sub-ontologies: proteins based on evolutionary relatedness (ProEvo); protein forms produced from a given gene locus (ProForm); and protein-containing complexes (ProComp).

(Explanation of the figure)

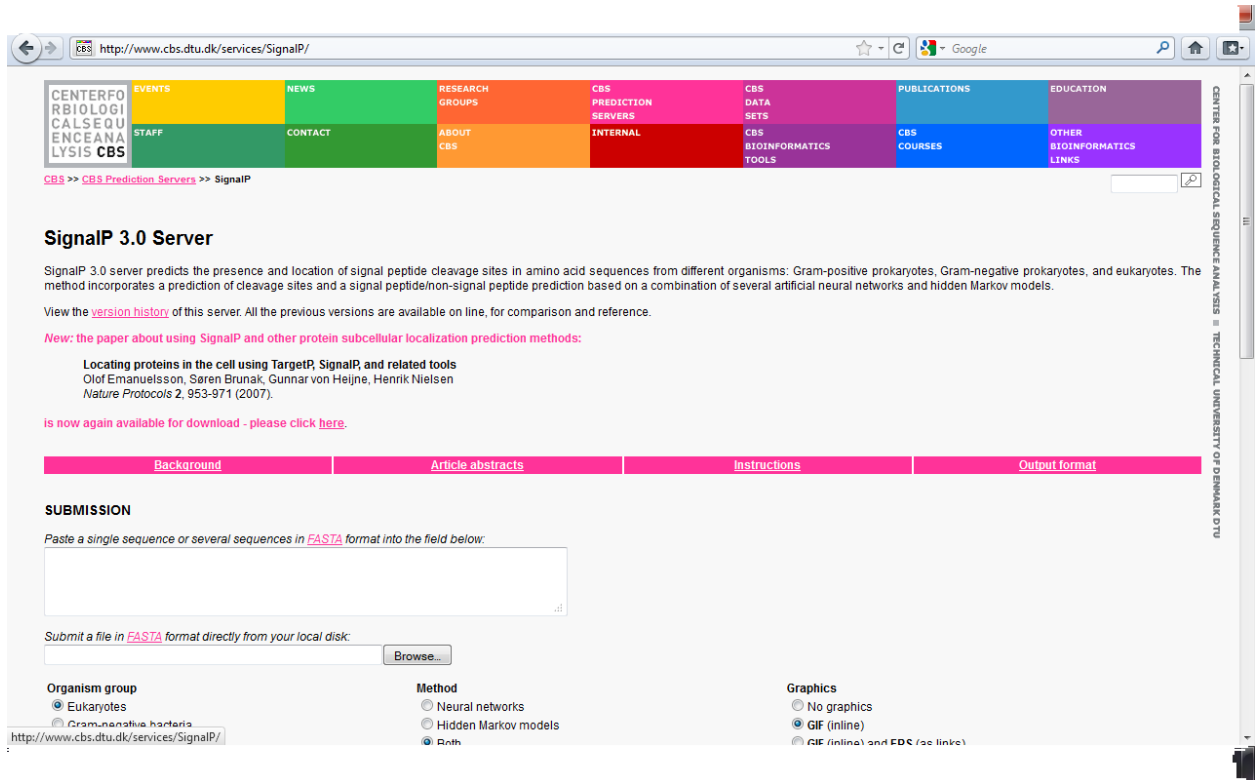
The diagram illustrates the PRO ontology structure. It shows three main levels: Root Level - protein, Root Level - protein complex, and Complex-Level Distinction. The Root Level - protein includes 'protein' and 'translation product of an evolutionarily-related gene'. The Root Level - protein complex includes 'protein domain' and 'general protein complex'. The Complex-Level Distinction includes 'molecular function'. Relationships are shown with 'is\_a' and 'has\_part' arrows.

FIG-6 PIR SERVER

❖ **SIGNAL P:-**

These server predicts the presence & location of signal peptide cleavage sites in amino acid sequences from different organism i.e. Gram Positive prokaryotes, Gram Negative prokaryotes & eukaryotes. This method incorporates a prediction of cleavage sites & a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks & hidden Markov models.





**FIG-7 SIGNAL P SERVER**

❖ **TMHMM:-**

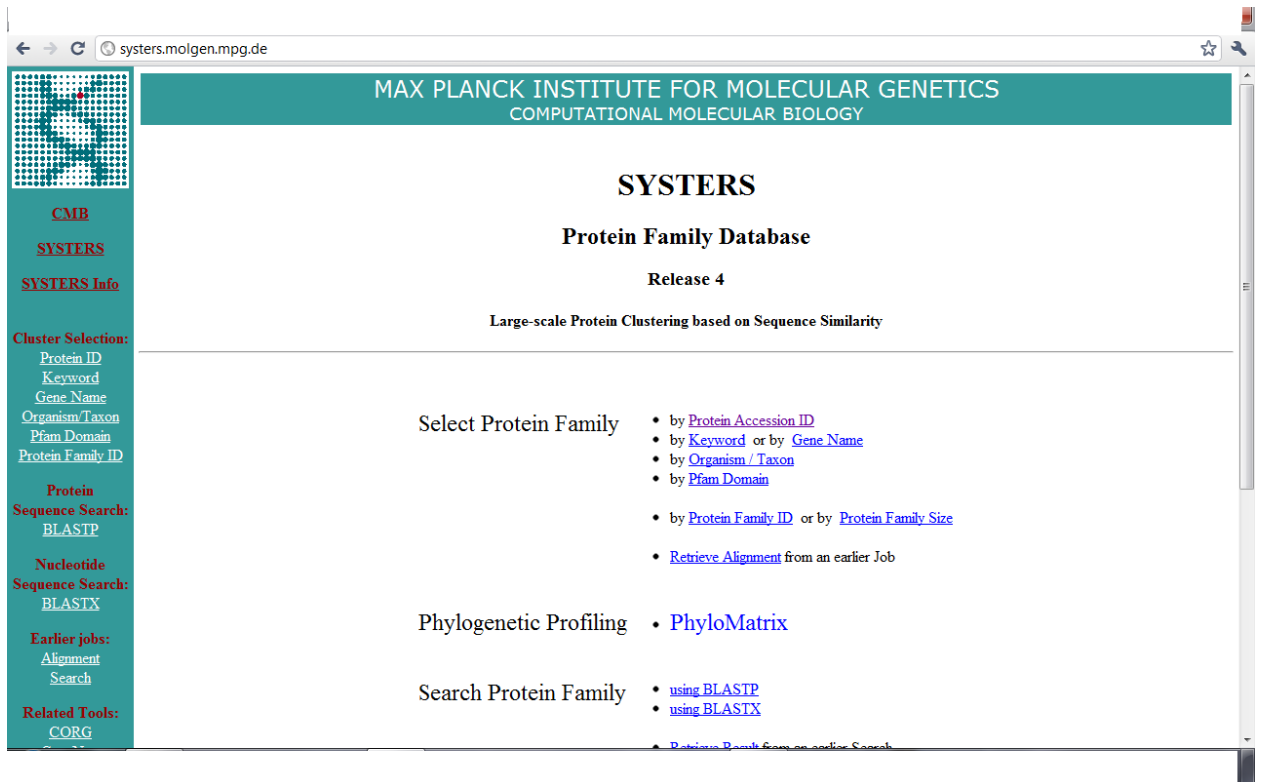
This server helps in the prediction of trans membrane helices in proteins. In July 2001 these TMHMM server has been rated very good in an independent comparison of the programs for prediction of TM helices. In these the program are taken into the proteins in the FASTA format. It recognizes around 20 amino acids out of which B,Z & X are equally unknown proteins. After leaving these unknown proteins the rest of the character is changed with X .So, that we can make sure that the sequences are of the sensible protein type or something else.

The screenshot shows the web interface for the TMHMM Server v. 2.0. The browser address bar displays the URL <http://www.cbs.dtu.dk/services/TMHMM/>. The page has a multi-colored navigation bar with links for EVENTS, NEWS, RESEARCH GROUPS, CBS PREDICTION SERVERS, PUBLICATIONS, EDUCATION, STAFF, CONTACT, ABOUT CBS, INTERNAL, CBS DATA SETS, CBS BIOINFORMATICS TOOLS, CBS COURSES, and OTHER BIOINFORMATICS LINKS. The main heading is "TMHMM Server v. 2.0" with the subtitle "Prediction of transmembrane helices in proteins". A note states: "NOTE: You can submit many proteins at once in one fasta file. Please limit each submission to at most 4000 proteins. Please tick the 'One line per protein' option. Please leave time between each large submission." Below this is a pink "Instructions" bar. The "SUBMISSION" section offers two methods: "Submission of a local file in FASTA format (HTML 3.0 or higher)" with a "Browse..." button, and "OR by pasting sequence(s) in FASTA format:" with a large text input area. Under "Output format:", there are three radio buttons: "Extensive, with graphics" (selected), "Extensive, no graphics", and "One line per protein". Under "Other options:", there is a checkbox for "Use old model (version 1)". On the right side, there is a 3D ribbon diagram of a protein structure. The footer of the page reads "CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS - TECHNICAL UNIVERSITY OF DENMARK DTU".

**FIG-8 TMHMM SERVER**

❖ PROTEIN CLUSTER:-

This is a collection of related protein sequences which consists of Reference Sequence proteins which are encoded by the complete genomes. This database contains both curated & non-curated clusters. The protein clusters database provides easy access to annotation information, publications, domains, structures & external links & analysis tools which include multiple alignments, phylogenetic trees & genomic neighbourhoods. These protein clusters can be searched like any other Entrez Database.



← → ↻ systers.molgen.mpg.de ☆

**MAX PLANCK INSTITUTE FOR MOLECULAR GENETICS**  
COMPUTATIONAL MOLECULAR BIOLOGY

**SYSTERS**  
**Protein Family Database**  
**Release 4**  
Large-scale Protein Clustering based on Sequence Similarity

**CMB**  
**SYSTERS**  
**SYSTERS Info**

**Cluster Selection:**  
Protein ID  
Keyword  
Gene Name  
Organism/Taxon  
Pfam Domain  
Protein Family ID

**Protein**  
**Sequence Search:**  
BLASTP

**Nucleotide**  
**Sequence Search:**  
BLASTX

**Earlier jobs:**  
Alignment  
Search

**Related Tools:**  
CORG

Select Protein Family

- by [Protein Accession ID](#)
- by [Keyword](#) or by [Gene Name](#)
- by [Organism / Taxon](#)
- by [Pfam Domain](#)
- by [Protein Family ID](#) or by [Protein Family Size](#)
- [Retrieve Alignment](#) from an earlier Job

Phylogenetic Profiling

- [PhyloMatrix](#)

Search Protein Family

- [using BLASTP](#)
- [using BLASTX](#)
- [Retrieve Result from an earlier Search](#)

---

**FIG-10 PROTEIN CLUSTER SERVER**

---

# Chapter 4

## RESULT AND DISCUSSION

#### 4.1BLAST OUTPUT FOR THE HYPOTHETICAL SEQUENCES

Serial no.	Seq id(hypothetical protein )	No. of hits	Type of protein
1.			
2.	gi 75766092	30	PE like protein & domain from <i>M.tuberculosis</i>
3.	<a href="#">ZP_03418071</a>	20	Putative secretory protein from <i>Cornybacterium sp.</i>
4.	gi 15607215	50	glutamine-transport ATP-binding protein ABC transporter from Mycobacterium
5.	<a href="#">NP_214586</a>	45	ABC transporters from Bacillus sp. And Mycobacterium
6.	<a href="#">NP_216226</a>	08	putative transcriptional regulator from Mycobacterium sp
7.	<a href="#">NP_216225</a>	25	segregation and condensation protein Corynebacterium sp.
8.	<a href="#">NP_216226</a>	20	chromosome segregation and condensation protein from multiple species
9.	A5WN29	18	transcriptional regulator from Mycobacterium sp
10.	C6DRW2	20	chromosome segregation and condensation protein from multiple species
11.	EGE50258	15	segregation and condensation protein B from Mycobacterium sp.
12.	AAA50918	14	segregation and condensation protein B [Mycobacterium
13.	ZP_05772470	10	chromosome segregation and condensation protein from multiple species
14.	Zp_05772470	08	chromosome segregation and condensation protein

			from multiple species
15.	ZP_05772470	20	chromosome segregation and condensation protein from multiple species
16.	A2VIJ4	25	segregation and condensation protein B from multiple species
17.	EGE50258	30	segregation and condensation protein
18.	ZP_05772470	-----	-----
19.	EGE50258	24	chromosome segregation and condensation protein from multiple species
20.	ZP_05223897	23	chromosome segregation and condensation protein from multiple species
21.	YP_003647317	21	chromosome segregation and condensation protein from multiple species
22.	YP_003273925	09	Putative secretory protein from
23.	YP_002766704	34	chromosome segregation and condensation protein from multiple species
24.	YP_003134356	29	Putative secretory protein from
25.	YP_700916	10	Putative secretory protein from
26.	YP_002834855	21	chromosome segregation and condensation protein from multiple species
27.	ZP_04388575	20	Putative secretory protein from
28.	YP_700916	17	chromosome segregation and condensation protein from multiple species
29.	YP_002834858	28	Putative secretory protein from
30.	YP_0028333456	35	chromosome segregation and condensation protein from multiple species

**TABLE-4.1 BLAST OUTPUT FOR THE HYPOTHETICAL PROTEINS**

## 4.2 MOTIF SEARCHING OUTPUT FOR THE SEQUENCES

Sequences	Name of the motif	Function
1.	DUF 1918	Unknown
2.	PE	PE family
3.	PRO_RICH	Proline rich family
4.	4.1 ABC tran 4.2 CNMP binding 4.3 SMC N 4.4 Mg chelatase 4.5 AAA 4.6 PduV-EutP 4.7 DUF258 4.8 NACHT 4.9 Urease beta 4.10 Rad17 4.11 Arch ATPase 4.12 AAA 5 4.13 UPF0079 4.14 DL IC	ABC transporter Cyclic nucleotide-binding domain RecF/RecN/SMC N terminal domain Magnesium chelatase ATPase family associated with cellular activities Ethanamine utilisation - propanediol utilization Unknown NACHT domain Urease beta subunit Rad17 cell cycle checkpoint protein Archaeal ATPase AAA domain Uncharacterised P-loop hydrolase Dynein light intermediate chain
5.	5.1 Fts X 5.2 MacB PCD 5.3 Histidinol dh	FtsX-like permease family MacB-like periplasmic core domain Histidinol dehydrogenase
6.	6.1 DUF387 6.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
7.	ScpA _ ScpB	ScpA/B protein
8.	8.1 DUF387 8.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
9.	9.1 DUF387 9.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
10.	10.1 DUF387 10.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
11.	11.1 DUF387 11.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
12.	12.1 DUF387 12.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
13.	ScpA ScpB	ScpA/B protein
14.	ScpA ScpB	ScpA/B protein
15.	ScpA ScpB	ScpA/B protein
16.	16.1 DUF387 16.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
17.	17.1 DUF387 17.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
18.	ScpA ScpB	ScpA/B protein

19.	19.1 DUF387 19.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
20.	ScpA ScpB	ScpA/B protein
21.	21.1 DUF387 21.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
22.	22.1 DUF387 22.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
23.	23.1 DUF387 23.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
24.	24.1 DUF387 24.2 HTH 15	Putative transcriptional regulators (Ypuh-like) Helix-turn-helix domain of alkylmercury lyase
25.	ScpA ScpB	ScpA/B protein
26.	ScpA ScpB	ScpA/B protein
27.	ScpA ScpB	ScpA/B protein
28.	ScpA ScpB	ScpA/B protein
29.	ScpA ScpB	ScpA/B protein

---

**TABLE-4.2 MOTIF SEARCHING OUTPUT FOR THE SEQUENCE**

---

### 4.3 Domain searching results from SMART server

SEQUENCE	DOMAIN NAME	FUNCTION
1.	Q7U1R5	UNKNOWN FUNCTION The protein is found to be in Hypothetical bacterial protein
2.	Q79F93	NIL
3.	A5WU55	NIL
4.	A4KNA2	NIL
5.	A4KNA1	UNKNOWN FUCTION Uncharacterized domain in proteins
6.	O33209	NIL
7.	A5U638	NIL
8.	O33209	NIL
9.	O33209	NIL
10.	O33209	NIL
11.	O33209	NIL
12.	O33209	NIL
13.	A5U368	NIL
14.	A5U368	NIL
15.	A5U368	NIL
16.	O33209	NIL
17.	O33209	NIL
18.	A5U368	NIL



19.	O33209	NIL
20.	A05SX6	NIL
21.	O33209	NIL
22.	O33209	NIL
23.	O33209	NIL
24.	O33209	NIL
25.	A5U368	NIL
26.	A5U368	NIL
27.	A5U368	NIL
28.	A5U368	NIL
29.	A5U368	NIL
30.	A5U368	NIL
31.	Q93GL0	NIL

---

**TABLE-4.3 DOMAIN SEARCHING RESULTS FOR SMART SERVER**

---

#### **4.4 RESULT FROM TMHMM SERVER**

SEQUENCES	NO. OF PREDICTED THMHM	REMARK
1.	NIL	
2.	NIL	
3.	1	TM HELIX 44-61
4.	NIL	
5.	4	TM HELIX 16-38,231-253,274-296,306-328
6.	NIL	
7.	NIL	
8.	NIL	
9.	NIL	
10.	NIL	
11.	NIL	
12.	NIL	
13.	NIL	
14.	NIL	
15.	NIL	
16.	NIL	
17.	NIL	
18.	NIL	
19.	NIL	
20.	NIL	
21.	NIL	
22.	NIL	
23.	NIL	

24.	NIL	
25.	NIL	
26.	NIL	
27.	NIL	
28.	NIL	
29.	NIL	
30.	NIL	
31.	1	TM HELIX 23-45

---

**TABLE-4.4 RESULTS OF TMHMM SERVER**

---

### **DISCUSSION:-**

From the results that

- We got 30 good hits from the BLAST server.
- We got very high no of hits from Motif server.
- We got two good hits from the Smart server.
- We got Three good hits from TMHMM server.

This clearly shows that Motif server is a very good and desirable server with which we can find a lots of Hypothetical sequences from the Un-Characterized proteins.

# Chapter 5

**Conclusion**

## **5.1 Conclusion**

The above results shows that the uncharacterized protein sequences show good sequence relationship with other proteins. The current BLAST study from the retrieved sequences, most sequences shows good similarity to the chromosome segregation and condensation protein and transcription regulatory proteins. Similarly the motif shows mostly Putative transcriptional regulators (Ypuh-like), Helix-turn-helix domain of alkyl mercurylyase and ScpA/B protein type of domain and motif is dominant among the uncharacterized proteins of Mycobacterium species .Also the transmembrane helix prediction shows some of the protein contains the transmembrane helix. Since the uncharacterized proteins have not been studied properly we have annotated by using some computational tool. Further structural analysis and modeling method could be used for more analysis. The current work may be used as a preliminary study to find the novel drug target against Mycobacterium tuberculosis.

## REFERENCES

- Sarangi, A. , Gupta, S., Mining the Proteome of Haemophilus ducreyi for Identification of Potential Drug Targets, The Open Bioinformatics Journal, 2010, 4, 1-4
- Ahmed, A., Minhas, K., In Silico Identification of Potential American Cockroach (Periplaneta americana) Allergens, Iranian J Publ Health, Vol. 39, No.3, 2010, pp. 109-115
- Weig,M., Ja,L. Systematic identification in silico of covalentlybound cell wall proteins and analysis of protein–polysaccharide linkages of the human pathogen Candida glabrata, Microbiology (2004), 150, 3129–3144.
- Mario A. Rodríguez-Pérez, In silico analysis of protein neoplastic biomarkers for cervix and uterine cancer, Clin Transl Oncol (2008) 10:604-617, DOI 10.1007/s12094-008-0261-2.
- Jane Mulder, N., Pruess, M., In Silico Characterization of Proteins, J. M. Walker Totowa.
- Sunil Kumar,G., Sarita1,S. Definition of Potential Targets in Mycoplasma pneumoniae Through Subtractive Genome Analysis, www.omicsonline.org, J Antivir Antiretrovir ISSN: 1948-5964 JAA, an open access journal.
- Narayan Sarangi, A., Aggarwal , R., Subtractive Genomics Approach for in Silico Identification and Characterization of Novel Drug Targets in Neisseria Meningitides Serogroup B, JCSB/Vol.2 September-October 2009.
- [www.ncbi.org.gov.in](http://www.ncbi.org.gov.in)
- [www.wikipedia.com](http://www.wikipedia.com)
- [http://www.ncbi.nlm.gov/genomes/MICROBES/microbial\\_taxtree.html](http://www.ncbi.nlm.gov/genomes/MICROBES/microbial_taxtree.html)
- <http://www.ncbi.nlm.nih.gov/genomeprj>
- <http://www.expasy.ch/sport/org>.
- <http://www.ncbi.nlm.nih.gov/blast>