JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTGAJC | DOI: 10.5584/jiomics.v1i1.32

Molecular Modeling of Cathepsin B protein in different Leishmania strains

Pawan Kumar Jayaswal[1,+], Mukta Rani[1], Chandra Prakash Yadav[1], Manas Ranjan Dikhit[1], Ganesh Chandra Sahoo*[1], Pradeep Das[1].

[1]Rajendra Memorial Research Institute of Medical Sciences, Agam Kuan, Patna, India- 80007; [+]Present address of the author is National Research Centre on Plant Biotechnology Genoinformatics, LBS Building, IARI, Pusa, New Delhi-110012.

Received: 13 August 2010 **Accepted:** 12 November 2010 **Available Online**: 20 November 2010

ABSTRACT

Cathepsin B like cysteine proteases representing a major component of the lysosomal proteolytic repertoire plays an important role in intracellular protein degradation. Comparative models of cathepsin B (CatB) protein of six different *Leishmania* strains were developed using MODELLER. The modeled three-dimensional (3-D) structure has the correct stereochemistry as gauged from the Ramachandran plot and good 3-D structure compatibility as assessed by PROCHECK and the DOPE score (DS2.1, Accelrys). The modeled proteins were energy minimized and validated using standard dynamic cascade protocol (DS 2.1). Seven different disulfide bonding sites are predicted in CatB protein of *Leishmania*. Two domains were identified and different motifs are present in catB protein of *Leishmania* like aspargine glycosylation sites, protein kinase phosphorylation sites, protein kinase C activation sites and N-myristoylation sites. Considering that cathepsin B is essential for survival of *Leishmania*, including for virulence to the mammalian host, it may be viewed as an attractive drug target.

Keywords: Cathepsin B; Homology Modeling; Leishmaniasis; Simulation; CatB; Cysteine protease**.**

1. Introduction

Leishmaniasis is a complex parasitic diseases caused by at least 17 different species of the protozoan parasite *Leishmania* [1]. It is transmitted by the bite of *Phlebotomine* sand flies; *Leishmania* infects approximately 12 million people and is commonly endemic in tropical and subtropical regions of America, Africa, and the Indian subcontinent, as well as in the subtropics of Southeast Asia [2]. *Leishmania* species are diploid eukaryotes and are obligate intracellular protozoa that reside in macrophages of their mammalian hosts [3]. *Leishmania donovani* and *major* are the causative agents of old world Visceral and Cutaneous Leishmaniasis (VL and CL) respectively. Among the variety of disease manifestations, VL is a systemic disease and *Leishmania donovani* complex is fatal and is a serious health problem in many tropical and subtropical countries [4] whereas CL, caused by species such as *Leishmania major, Leishmania mexicana, Leishmania braziliensis, and Leishmania panamensis*, frequently self-cures within 3–18 months, leaving disfiguring scars [5]. In vertebrate hosts Leishmaniasis are transmitted by *Phlebotomine* sand flies, which acquire the pathogen by feeding on infected hosts and transmit them by regurgitating the parasite at the site of a subsequent blood meal [5]. While obtaining blood meal, sand flies inject saliva into the host's skin containing anti clotting, anti platelet and vasodilatory compounds that increase the hemorrhagic pool [6, 7]. *Leishmania* parasites contain high levels of cysteine proteases [8] represent a major component of the lysosomal proteolytic repertoire and play an important role in intracellular protein degradation [9] The lysosomal cysteine proteases, cathepsins B, H, L, S and C, are well characterized proteins with closely related amino acid sequences, belonging to the papain super family [10]. Lysosomal enzymes are synthesized in normal cellular processes as glycosylated higher molecular weight precursors, which, during their maturation, undergo several processing steps by limited proteolysis [11]. Cathepsin B maturation includes removal of the N-terminal propeptide, the C-terminal extension and a dipeptide between residues 47 and 50 (mature enzyme numbering). The product is an en-

***Corresponding author:** Dr. Ganesh Chandra Sahoo, Scientist & Head; BioMedical Informatics Center, RMRIMS, Agam Kuan, Patna -80007, India; Contact No. +91- 612 -2631565, Fax No. +91-612-2634379; E-mail Address: ganeshiitkgp@gmail.com

zymatically active molecule with two chains covalently crosslinked by a disulfide bridge [12, 13]. It exhibits both endopeptidase and exopeptidase activities [14] and shown that the exopeptidase activity is dependent on the presence of a specialized structural element, the occluding loop, which accepts the negative charge of the P2 carboxylate. CatB protein that belongs to the papain super family and shows high homology to cathepsins L, S and O, papain and actinidin, among others [10]. This super family encompasses a large number of cysteine proteases from sources as diverse as bacteria, plants and mammals [15]. Studies involving species of *Leishmania* such as *L. major* and *L. amazonensis* have been shown to induce the production of biologically active transforming growth factor Я (TGF-β) by macrophages upon infection [16]. *L. donovani* infection is known to induce the expression of a number of cytokine genes like TNF-Ю , GM-CSF, TGF-β, and IL-6 [17]. Application of anti-TGF-β antibodies arrested the development of lesions in mice, whereas treatment with TGF-β exacerbated the infection with *L. amazonensis*, *Trypanosoma cruzi*, and *Toxoplasma gondii* [18]. *Leishmania major, L. mexicana*, and *L. braziliensis* trigger the production of TGF-β and IL-10, which inhibit killing of intracellular organisms [19]. It was also shown that TGF-β plays a role in limiting IFN- γ production during the primary infection in mice [20]. IFN-γ, on the other hand, is known to induce the expression of inducible nitric-oxide synthase, the key effectors mechanism for the killing of *Leishmania* and other parasites within the mouse macrophages, both *in vitro* and *in vivo* [21]. TGF-β isoforms are synthesized as large biologically inactive precursors, which are proteolytically processed to yield mature and active homodimer. A variety of agents and treatments are known to activate latent TGF-β, like heat, acidic pH, plasmin, subtilisin-like endopeptidases, and cathepsins [22, 23, 24]. Hence, structural and functional analysis of catB protein of different *Leishmania* sp. is vital for further structure based ligand protein interaction study.

In this work, we are concerned with the determination of the 3-D structure of cathepsin B-like cysteine protease of six different strains of *Leishmania* (*L. donovani, L. infantum, L. major*, *L. mexicana, L. chagasi,* and *L. braziliensis*). These developed models will be helpful in the screening of several antileishmanial drug molecules and alkaloids in future. Screening of antileishmanial compounds (Ligand molecules) against cathepsin protein study is going on in Biomedical Informatics Centre (BIC) of RMRIMS with the different commercial software.

2. Material and methods

In this study various three dimensional structural models of the cathepsin B protein of six different *Leishmania* strains were generated. The models were validated by Ramachandran plots of PROCHECK and DOPE scores of Discovery Studio software v 2.1. The models of cathepsin B protein were further tested for insilico docking study to know the presence of any interaction between the ligand and Cat B protein. Ligand protein interaction of KMP-11 has already been reported earlier [25]. Various methods applied in this study are given below.

2.1 *Structural Modeling and Sequence Analysis of Cathepsin B Protein*

The amino acid sequence of cathepsin B of *L. donovani* (340 amino acids (aa), Genbank locus ID: AAG44365), *L. braziliensis* (340 aa, Uniprot ID: A4HH90), *L. infantum* (340 aa, Uniprot ID: A414D6), *L. chagasi* (340 aa, Uniprot ID: Q9GQN7), *L. major* (340 aa, Uniprot ID: Q4FXX7), and *L. mexicana* (340 aa, Uniprot ID: Q25319) was downloaded through NCBI & EMBL website for structural modeling. Multiple alignments of the related sequences were performed using Clustal W program accessible through the European Bioinformatics Institute [26] (http://www.ebi.ac.uk/Tools/clustalw2/index.html). No X-ray crystallographic or NMR structure of Cathepsin B protein of any *Leishmanial* strains has yet been determined. Tertiary structures of cathepsin B protein of six different *Leishmanial* strains were modeled on the basis of template pdb id: 3PBH & 1MIR using MODELER protocol of Discovery Studio 2.1. Structure of LPG2 protein of different *Leishmania* strains has already been reported earlier [27]. Structure validation was performed using Verify protein (DOPE) scores, WHATIF and, molecular modeling tools of Discovery studio. Cathepsin B protein of six different *Leishmania* strains and their two different template homologs and their PBB ID: 1MIR and 3PBH having their tertiary structures i.e. β sheets and α- helices are predicted through of Discovery Studio 2.1.

2.2 *Simulation of Cathepsin B protein*

Model of Cathepsin B protein of six different strains of *Leishmania* were further processed by applying CHARMM force field. Potential energy of a specified structure is evaluated by using calculate energy protocol of DS2.1. The calculate energy protocol can be used to compare the relative stability of different configurations of the same structure; or as a prelude to lengthy simulations to confirm the availability of appropriate force field parameters. The CHARMM molecular simulation package uses the CHARMM force field to model the energetic, forces and dynamics of biological molecules using the classical method of integrating Newton's equations of motion [28]. Energy minimization of al six different 3-D modeled protein structures are done with the help of standard dynamics cascade protocol of DS 2.1[29] which performs the following steps: minimization with steepest descent method, minimization with conjugate gradient, dynamics with heating, equilibration dynamics, production dynamics. The minimization protocol minimizes the energy of a structure through geometry optimization. The dynamics (heating or cooling) protocol allows controlling the temperature of a system when performing a molecular dynamics simulation. For the simulation cascade following parameter are used: steepest descents minimization (500steps, RMS gradient 0.1) in first minimization step & in second steepest Descents minimization (500 steps, RMS gradient 0.0001), heating (2000 steps , initial temperature 50K, final temperature 300K), equilibration (120 ps, 1fs time step, coordinates saved every 1000 steps) and Production (120 ps, 1fs time step, 300 K, NVT ensemble, non bond cutoff 14A, switching function applied between 10 and 12A, coordinates saved every 1000 steps).

2.3 *Function Assignment of Cathepsin B Protein by SVM*

To know the novel functions of Cathepsin B protein of all six different *Leishmania* strains were searched through BIDD server (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi) [30]. The web-based software, SVMProt, support vector machine (SVM) classifies a protein into functional families from its primary sequence based on physico-chemical properties of amino acids. Novel protein function assignment of different proteins of SARS virus and Japanese encephalitis virus has already been reported by using this server [31].

2.4 *Predict Protein Server*

Predict Protein provides PROSITE sequence motifs, low complexity regions (SEG), nuclear localization signals, regions lacking regular structure (NORS) and predictions of secondary structure, solvent accessibility, globular regions, transmembrane helices, coiled-coil regions, structural switch regions, disulfide-bonds, sub-cellular localization, and functional annotations [32,33,34,35].

3. Results and Discussion

Structure, function and ligand binding site analysis of cathepsin B protein will lead to identification of novel targets for design of suitable lead compounds inhibiting the specific functions of *L. donovani, L. infantum, L. major, L. mexicana, L. chagasi,* and *L. braziliensis.*

3.1 *Structural Modeling and Sequence Analysis of Cathepsin B Protein*

Structural model of all six strains (*L. braziliensis, L. infantum, L. chagasi, L. major, L. mexicana, L. donovani*) of cathepsin B protein of *Leishmania* is modeled by MODELER on the basis of different three dimensional co-ordinates of two crystal structures (templates) of proteins namely PDB ID: 3PBH & 1MIR taken for this study from RCSB PDB are same for all six strains of *Leishmania* shown in Table 1. The PDB ID: 3PBH and 1MIR are selected as a template on the basis of BLAST result. Each strain of catB protein of *Leishmania* shows that the different identity with template protein. *L. donovani, L. braziliensis and L. infantum* sequences are having 38% sequence similarity with template 3PBH and 1MIR while *L. major and L. mexicana* is having 39% sequence similarity with 3PBH and 38% with 1MIR protein. *L. chagasi* shows only 37 % sequence similarity with 1MIR and 38% with 3PBH is shown in Table 1.

Crystallographic studies have demonstrated the structural features of template proteins i.e. 3PBH & 1MIR of human liver cathepsin B protein and rat procathepsin B respectively at resolution 2.8A° and 2.15A° respectively. The three dimensional structure of human procathepsin B (PDB ID: 3PBH) revels that the propeptide folds on the catB surface [36], shielding the enzyme active site from exposure to solvent. The structure of the enzymatically active domains is virtually identical to that of the native enzyme [37]. The three dimensional coordinates (PDB ID: 1MIR) represents cysteine protease of the papain super family which is synthesized as inactive precursors with a 60-110 residues at the N-terminal pro segment. The propeptide are potent inhibitors of their parent protease [38]. On the basis of above template studies it is hypothesized that the pro region lies in between the Tyr27 – Gln95 amino acids sequence of the different *Leishmania* strains i.e. *L. donovani, L. chagasi, L. infantum, L.*

Table 1. Relative Data of six different strains of Cathepsin B proteins of *Leishmania* as well as their templates (3PBH and 1MIR) shows the highest verify score value than the expected score value. Identity of template protein is varied between 37-39%. Abbreviations: LEIDO: *L. donovani,* LEIBR: *L. braziliensis,* LEIIN: *L. infantum*, LEICH: *L. chagasi* LEIME: *L. major,* LEIME: *L. mexicana.*

	Name of Target	Template protein				Dope	Verify	Expected	Expected
S.No.	Sequence	Name	Length	Identity (%)	Model Name	Score	Score	High Score	Low Score
1	AAG44365	3PBH	317	38	3PBH.B99990001	-33420.4	150.65	139.723	62.8754
	LEIDO	1MIR	313	38					
2	A4HH90 LEIBR	3PBH	317	38	1MIR.B99990002	-32069.9	146.29	139.723	62.8757
		1MIR	313	38					
3	A414D6	3PBH	317	38	1MIR.B99990001	-32335.4	153.31	139.723	62.8757
	LEIIN	1MIR	313	38					
4	Q9GQN7	3PBH	317	38	3PBH.B99990003	-33420.4	154.79	139.723	62.8757
	LEICH	1MIR	313	37					
5	QLFXX7	3PBH	317	39	3PBH.B99990003	-33420.4	154.79	139.723	62.8757
	LEIMA	1MIR	313	38					
6	Q25319	3PBH	317	39	3PBH.B99990003	-33420.4	154.79	139.723	62.8757
	LEIME	1MIR	313	38					

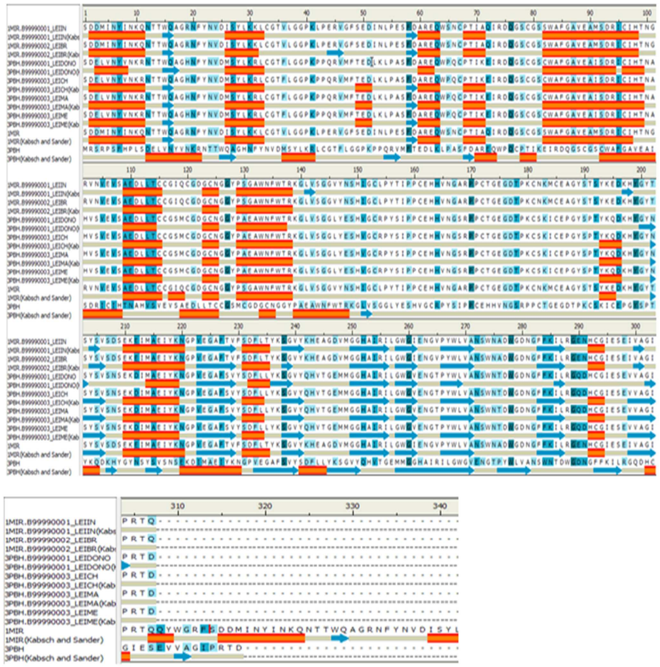


Figure 1. Comparative analysis of β-sheet and α-helix of protein PDB ID: 3PBH (3-D structure of human procathepsin B) and 1MIR (3-D coordinates of cysteine protease of papain super_ family) with cathepsin B protein of six different strains of Leishmania (DS 2.1). In figure1, Red color shows helical structure and blue arrows are the β-sheets (This Figure is available in the supplementary material).

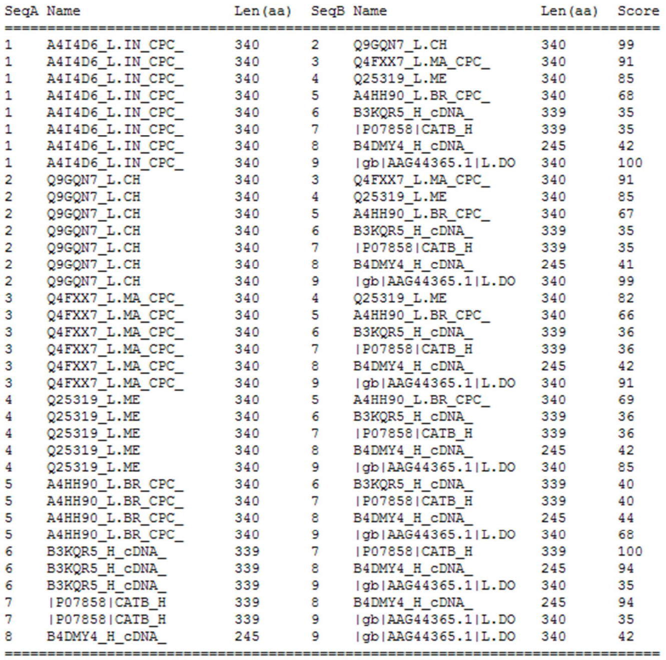
major and *L. mexicana* but in case of *L. braziliensis* Gln95 amino acid residue is replaced by Ala95 residues of the protein sequence. After the cluster analysis of amino acid sequence of catB protein of six *Leishmania* strains and two templates it is known that the mature sequence begins at the N-terminal Met1 and ends at C-terminal Glu-340 amino acids. Secondary structure of catB protein (α -helices and β -sheets) are represented in Figure 1, 340 amino acid sequences of catB protein of six *Leishmania* strains and two template PDB ID: 3PBH and 1MIR in the modeled catB protein of *Leishmania,* 7-9 α helices and 6-8 β sheets have been observed.

Multiple alignment of amino acid sequences of cathepsin B protein of different *Leishmania* strains show that there is much identity (66-100%) among each other, catB protein of different *Leishmania* strains show 35- 45% identity with catB protein of human. Cathepsin B of *L. infantum* and *L. donovani* are identical (100%) to each other, hence demographic separation do not have any impact on protein structure at these two strains. Cathepsin B Protein of six different strains of *Leishmania* and three different strains of *Homo sapiens* i.e. B3KQRS, P07858 and B4DMY4 have been aligned through multiple sequence alignment by using Clustal W. It is learnt from multiple alignments that six strains of *Leishmania* are similar to each other and dissimilar with three different strains in human which is shown in Table 2. From cladogram, catB protein of *L. braziliensis* and other *Leishmania* strains are forming a cluster different from other cysteine proteases of human. Three strains of human cysteine proteins are found to be far from other *Leishmania* strains shown in Figure 2.

Developed 3- D models of all six *Leishmania* strains is verified with the help of Verify protein (MODELER) score protocol of DS2.1 and their score is higher than the expected high score. The DOPE (Discrete Optimized Protein Energy) score (-33420.4) is same for 3-D structure models of catB protein of *L. mexicana, L. major, L. chagasi* and *L. donovani* and is different i.e. (-32335.4) from other two modeled structures that of *L. braziliensis* and *L. infantum* is shown in Table1.

Cathepsin B protein of structural models of six *Leishmania* strains were validated by verify protein (MODELER) score of DS 2.1 (Accelrys), WHATIF and PROCHECK. In the Ramachandran plots (Procheck) show that 65-88.1% amino acid residues belong to core region, 9-30% residue in allowed region, 0.4 - 4.4% are in generously allowed regions and 0.4–1.2% in disallowed region is reported in Table 3. These amino acids which occur in invalid region of Ramachandran plot were further refined by side chain and loop refinement tools of DS2.1 (Accelrys) to get validate of the 3-D structure of cathepsin B protein. The best model of different *Leishmania* strains were screened by Verify protein (MODELER) score and the best was selected for further analysis. About 7 to 9 α-helices have been observed for catB protein of different models of various *Leishmania* strains, seven, eight, and nine helices have been observed in catB protein of modeled structure in *L. donovani, L. braziliensis, L. infantum* respectively. The modeled structures of catB protein of different *Leishmania* strains have shown close identity with each other, one modeled structure of *L. braziliensis* is given in Figure 3. The verify protein (MODELER) score of best predicted models of catB

Table 2. ClustalW results of multiple sequence alignment scores of Cathepsin B Protein of six different strains of Leishmania and three strains of human. Leishmania strains are closely associated with each other and far from the human catB protein sequence.



```
SeqA Name                Len(aa)  SeqB Name                 Len(aa)  Score
==========================================================================
1    A4I4D6_L.IN_CPC_    340      2    Q9GQN7_L.CH          340      99
1    A4I4D6_L.IN_CPC_    340      3    Q4FXX7_L.MA_CPC_     340      91
1    A4I4D6_L.IN_CPC_    340      4    Q25319_L.ME          340      85
1    A4I4D6_L.IN_CPC_    340      5    A4HH90_L.BR_CPC_     340      68
1    A4I4D6_L.IN_CPC_    340      6    B3KQR5_H_cDNA_       339      35
1    A4I4D6_L.IN_CPC_    340      7    |P07858|CATB_H       339      35
1    A4I4D6_L.IN_CPC_    340      8    B4DMY4_H_cDNA_       245      42
1    A4I4D6_L.IN_CPC_    340      9    |gb|AAG44365.1|L.DO  340      100
2    Q9GQN7_L.CH         340      3    Q4FXX7_L.MA_CPC_     340      91
2    Q9GQN7_L.CH         340      4    Q25319_L.ME          340      85
2    Q9GQN7_L.CH         340      5    A4HH90_L.BR_CPC_     340      67
2    Q9GQN7_L.CH         340      6    B3KQR5_H_cDNA_       339      35
2    Q9GQN7_L.CH         340      7    |P07858|CATB_H       339      35
2    Q9GQN7_L.CH         340      8    B4DMY4_H_cDNA_       245      41
2    Q9GQN7_L.CH         340      9    |gb|AAG44365.1|L.DO  340      99
3    Q4FXX7_L.MA_CPC_    340      4    Q25319_L.ME          340      82
3    Q4FXX7_L.MA_CPC_    340      5    A4HH90_L.BR_CPC_     340      66
3    Q4FXX7_L.MA_CPC_    340      6    B3KQR5_H_cDNA_       339      36
3    Q4FXX7_L.MA_CPC_    340      7    |P07858|CATB_H       339      36
3    Q4FXX7_L.MA_CPC_    340      8    B4DMY4_H_cDNA_       245      42
3    Q4FXX7_L.MA_CPC_    340      9    |gb|AAG44365.1|L.DO  340      91
4    Q25319_L.ME         340      5    A4HH90_L.BR_CPC_     340      69
4    Q25319_L.ME         340      6    B3KQR5_H_cDNA_       339      36
4    Q25319_L.ME         340      7    |P07858|CATB_H       339      36
4    Q25319_L.ME         340      8    B4DMY4_H_cDNA_       245      42
4    Q25319_L.ME         340      9    |gb|AAG44365.1|L.DO  340      85
5    A4HH90_L.BR_CPC_    340      6    B3KQR5_H_cDNA_       339      40
5    A4HH90_L.BR_CPC_    340      7    |P07858|CATB_H       339      40
5    A4HH90_L.BR_CPC_    340      8    B4DMY4_H_cDNA_       245      44
5    A4HH90_L.BR_CPC_    340      9    |gb|AAG44365.1|L.DO  340      68
6    B3KQR5_H_cDNA_      339      7    |P07858|CATB_H       339      100
6    B3KQR5_H_cDNA_      339      8    B4DMY4_H_cDNA_       245      94
6    B3KQR5_H_cDNA_      339      9    |gb|AAG44365.1|L.DO  340      35
7    |P07858|CATB_H      339      8    B4DMY4_H_cDNA_       245      94
7    |P07858|CATB_H      339      9    |gb|AAG44365.1|L.DO  340      35
8    B4DMY4_H_cDNA_      245      9    |gb|AAG44365.1|L.DO  340      42
==========================================================================
```

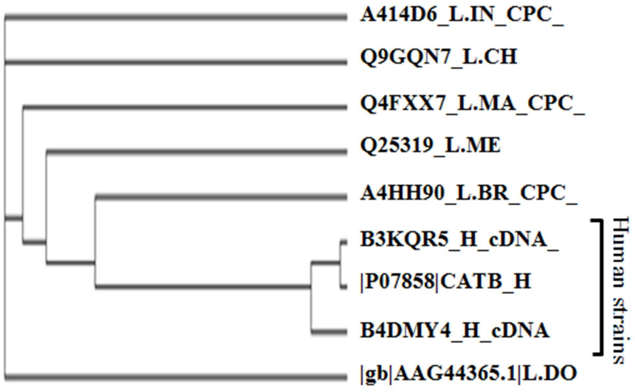


Figure 2. Phylogram showing phylogenetic relationship of Cathepsin B protein of six strains of Leishmania (L. infantum, L. chagasi, L. mexicana, L. braziliensis, L. major and L. donovani) with three different strains i.e. B3KQR5_H(cDNA), P07858,CATB_H and B4DMY4_H(cDNA) of cathepsin B protein in Homo sapiens.

Table 3. Referring to Ramachandran Plots of cathepsin B protein of six different strains of *Leishmania*. Abbreviations: Ldv: *L. donovani*, Lbrzl: *L. braziliensis*, Linf: *L. infantum*, Lch: *L. chagasi*, Lma: *L. major*, Lme: *L. mexicana*.

Residues	Number of Amino acids involved						Percentage of amino acids involved					
	Ldv	Lbrzl	Linf	Lch	Lma	Lme	Ldv	Lbrzl	Linf	Lch	Lma	Lme
Residues in most favoured regions [A,B,L]	164	223	220	220	220	220	65.1	88.1	87.0	87.3	87.3	87.3
Residues in additional allowed regions [a,b,l,p]	75	25	28	30	30	30	29.8	9.9	11.1	11.9	11.9	11.9
Residues in generously allowed regions [~a,~b,~l,~p]	11	3	2	1	1	1	4.4	1.2	0.8	0.4	0.4	0.4
Residues in disallowed regions	2	2	3	1	1	1	0.8	0.8	1.2	0.4	0.4	0.4
Number of non-glycine and non-proline residues	252	253	253	252	252	252	100% for all strains					
Number of end-residues (excl. Gly and Pro)	2	2	2	2	2	2						
Number of glycine residues (shown as triangles)	35	38	38	35	35	35						
Number of proline residues	18	14	14	18	18	18						
Total number of residues	307	307	307	307	307	307						

protein of various *Leishmania* strains Table 4, shows that highest scores (154.79) has been found in case of cathepsin protein model of *L. chagasi, L. major, L. mexicana* and lowest scores (146.29) has been found in case of *L. braziliensis*.

From Ramachandran plot, it is known that maximum residues in cathepsin B protein are responsible for construction of helices. It is found that the best model of cathepsin B protein of all these strains consisted of only one chain. In all the models 19 – 23 % is helical. The best models of catB protein of *L. chagasi, L. major, L. mexicana* are having highest number (eight) of helices where minimum three and maximum seventeen residues take part in formation of a helix. In all six strains of *Leishmania* 3_{10} helices have also been found where as in *L. donovani* maximum eighteen residues has been involved in forming a helix and helices in this strain accounts 19.2 % of all is shown in Table 4. CatB protein of six *Leishmania* strains that varies six to eight β sheets. Six β sheets are observed in *L. infantum* where as in L. *chagasi, L. major, L. mexicana* are having eight β sheets similarly seven β sheets are observed in *L. donovani and L. braziliensis.*

3.2 *Simulation of Cathepsin B protein*

Cathepsin B protein of six different strains of *Leishmania* are simulated by standard dynamic cascade protocol, in this process each simulation consists of 500 steps which is extended up to 5000 after that in each step 1000 increment has been given which was continue up to 10,000 steps of energy minimization. Each step can calculate the Van der Waals energy, CHARMM energy, potential energy and kinetic energy of the protein. Net partial charge and Net formal charge of catB protein of *L. donovani*, *L. chagasi, L.major, L. mexicana* are having -9 and that of other two strains are having -11 in *L. braziliensis and L. infantum.* Initial CHARMM energy of catB *L. chagasi, L. major, L. mexicana* was 14341 Kcal/mol and that of *L. donovani* was -19493.7 Kcal/mol and in *L. braziliensis* was -9705.34 Kcal/mol, after the minimization of energy up to 10,000 steps. CHARMM force field of cathepsin-B protein of each strain was changed and it varies between -16343.9 kcal/mol to -17110.1 kcal/mol. Van der Waals energy of cathepsin B of *L. donovani* changed from 2007.61 kcal/mol to -1732.12 kcal/mol, similar type of Van der Waals energy variation has been observed in other cathepsin B of five strains is shown in Table 5.

3.3 *Functional Assignment of Cathepsin B protein by SVM*

From the comparative analysis of cathepsin B protein of different *Leishmania* strains functional assignment shows that it belongs to transmembrane region protein. Cathepsin B protein of *L. donovani, L. chagasi, L. infantum* strains belongs to metal binding (65.4 %), manganese binding (62.2 %), copper binding (58.6 %) and magnesium binding (58.6 %) protein function families. Other protein functional families like hydrolases - acting on peptide bonds (peptidase) has been detected in *L. donovani*, *L. chagasi* and *L. infantum* (76.2%) and 85.4 % in *L. major* and 80.4 % in *L. mexicana*. Cathepsin B protein of *L. braziliensis, L. major and L. mexicana* belongs to metal binding functional motifs are 62.2 %, 73.8% and 71.3% respectively. Cathepsin B protein of *Leishmania major* has manganese binding (65.4 %), and 58.6% copper binding in *L. mexicana*. Calcium binding property has

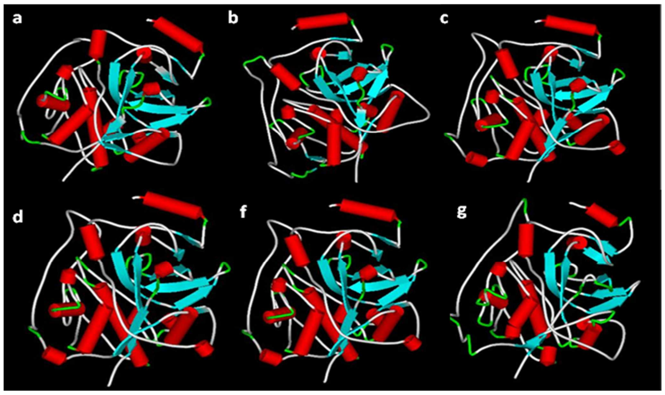


Figure 3. Ribbon representations of the homology model of Cathepsin B protein images of all six different leishmania strains using Discovery Studio 2.1 (Accelyrs) software (a) *L. braziliensis* (b) *L. infantum* (c) *L. chagasi* (d) *L. major* (e) *L. mexicana* and (f) *L. donovani.*

Table 4. Promotif search result summary and Profiles- 3D scores of modeled structure of cathepsin B proteins of all six strains of *Leishmania*.

Model Features Strain Names	No. and % of alpha helices	No. and % of 3,10(310) helices	No. of chain	Profile 3D Scores
L. donovani	7 / 19.2% 3(min)-18(max) Residues take part in formation of helices	2 /1.6% 3 residues	1	150.65
L. braziliensis	7/ 21.2 % 3(min)-17(max) Residues take part in formation of helices	4 / 3.9 % 3,5 residues	1	146.29
L. infantum	7/ 21.2 % 3(min)-17(max) Residues take part in formation of helices	4 / 3.9 % 3,5 residues	1	153.31
L. chagasi	8 /22.8 % 3 (min)-17(max) Residues take part in formation of helices	5/5.2 % 3-5 residues	1	154.79
L.major	8 /22.8 % 3 (min)-17(max) Residues take part in formation of helices	5/5.2 % 3-5 residues	1	154.79
L. mexicana	8 /22.8 % 3 (min)-17(max) Residues take part in formation of helices	5/5.2 % 3-5 residues	1	154.79

been detected for amino acid sequence of *L. braziliensis* and *L. mexicana*. In *L. braziliensis* few amino acids of catB protein participates in the formation of outer membrane. DNA repair as well as transportation activity are the novel function reported by us in different strains of catB protein of *Leishmania* with the help of support vector machine tool (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi). Lipid binding property (88-92%) of cathepsin B has been predicted in *L. donovani* (91.3%), *L. chagasi* (88.1 %), *L. major* (83.9%) and L. *infantum* (91.3%) is shown in Figure 4. From NCBI and EMBL, it is also known that Cathepsin B protein is a specific synthetic inhibitor protein which indicates that the inhibitor itself does not affect the growth of the parasites during the promastigote stages of the parasite.

Table 5. It shows the simulation of cathepsin B protein of six different *Leishmania* strains.

Cathepsin B Protein in six different strains of leishmania		CHARMM Energy Kcal/mol	Electrostatic Energy Kcal/mol	Initial energy Kcal/mol	Vander Wall Energy Kcal/mol
L. donovani	**Initial**	-19493.7	-22057		2007.61
	Final	-16436.7	-20840.7	-16472.9	-1732.12
L. braziliensis	**Initial**	-9705.3	-11186		-1288.92
	Final	-17110.1	-21271.7	-17137.8	-1768.44
L. infantum	**Initial**	-1580.56	-10111.5		5452.26
	Final	-16886.8	-21141.1	-16881.9	-1701.91
L. chagasi	**Initial**	14341	-10295.5		21547.8
	Final	-16816.9	-21211.9	-16825.1	-1796.56
L. major	**Initial**	14341	-10295.5		21547.8
	Final	-16944.8	-21190	-16960	-1795.06
L. mexicana	**Initial**	14341	-10295.5		21547.8
	Final	-16953.4	-21152.5	-16994.2	-1815.47

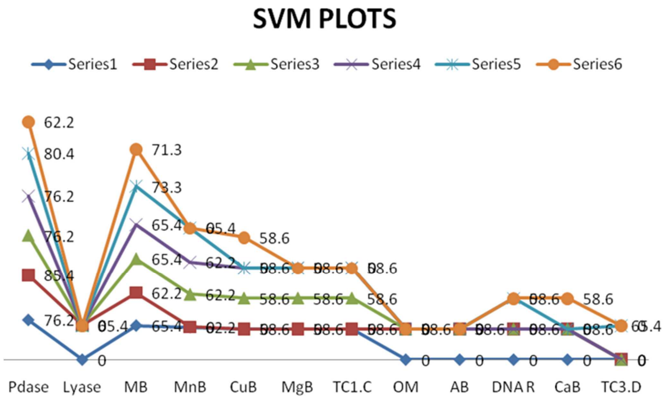


Figure 4. Comparative analysis of functional assignment of Cathepsin B protein in various strains by SVMProt. [Pdase→EC3.4, Hydrolases acting on peptide bonds, Lyase→ EC4.1 Lyase carbon-carbon lyases, MB→ Metal Binding, MnB→ Manganese binding, CuB→Copper Binding, MgB→ Magnesium Binding, Tc 1.c→ channels/pores-pore forming toxins (proteins & peptides), OM→ Outer membrane, AB→ Actin Binding, DNAR→ DNA Repair, CaB→ Calcium binding, Tc3-D→ Primary Active transporter oxidoreduction driven transporter.] (Series1→ L. donovani, Series2→ L.braziliensis, Series3→ L. chagasi, Series4→ L.major, Series5→ L. mexicana, Series6 → L. infantum).

3.4 *Predict Protein Server*

The amino acid sequence of cathepsin B protein of different *Leishmania* strains were submitted at protein predict server, to know whether there is presence of any post translation modification. In all six strains of *Leishmania* total eights post translational modification sites are observed. There are five aspargine glycosylation sites (N [^P] [ST] [^P]) from 18 amino acid to 159 amino acid in catB protein of six different strains of *Leishmania*. NFSV and NTTC glycosylation pattern is predicted at 28th and 159th residues in *L. chagasi, L. major, L. mexicana* where as in *L. donovani, L. braziliensis* (NFSV and NMST) glycosylation patterns were observed at 18th and 28th positions respectively. Glycosylation pattern (NSSK and NTTC) were observed at the 141th and 149th positions in catB protein of *L. donovani* and *L. mexicana*. Two protein kinase phosphorylation sites, one cAMP- and one cGMP-dependent with RRIS motif have been found at 88th position in *L. infantum*, *L. chagasi* and *L. mexicana* where as in *L. major* RRIS motif has been replaced by RRMS motif. In *L. donovani* and *L.braziliensis* at 78th position (motif RRIS), 260th position (RRGT motifs) are also present which imply that cAMP and cGMP dependent protein kinase phosphorylation occur at these sites. Five Protein kinase C activation sites have been found in three *Leishmania* strains i.e. *L. infantum*, *L. chagasi*

Table 6. Comparative analysis of different motifs of cathepsin B protein of six *Leishmania* strains. The motifs were predicted by Predicted Protein Server.

Motifs name	***L.donovani***	***L.braziliansis***	***L.infantum***	***L.chagasi***	***L.major***	***L.mexicana***
	PATTERN					
ASN_GLYCOSYLATION	18→NFSV 149→NTTC	18→NMST 159→NSTC	28→NMST 159→NSTC	28→NFSV 159→NTTC	28→NFSV 159→NTTC	28→NFSV 141→NSSK 159→NTTC
(N-glycosylation site) N[^P][ST][^P]						
CAMP_PHOSPHO_SITE	78→RRIS	260→RRGT	88→RRIS	88→RRIS	88→RRMS	88→RRIS
(cAMP- and cGMP-dependent protein kinase phosphorylation site) [RK]{2}.[ST]						
PKC_PHOSPHO_SITE	66→SDR 132→SDK 45→TPK 169→ SVK	3→SGK 25→SPR 45→SDK 76→SDR 179→SLR 208→SYK 278→SLK	3→SGK 76→SDR 142→ SDK 155→TPK 179→SVK	3→SGK 76→SDR 142→SDK 155→TPK 179→SVK	3→TGK 76→SDR 142→SEK 155→TPK 179→SVK	3→TGK 45→SEK 76→SDR 142→SSK 155→TPK 179→SIK
(Protein kinase C phosphorylation site) [ST].[RK]						
CK2_PHOSPHO_SITE	20→SVDE 43→TISE 141→TIYD 150→TTCE 155→SEMD	6 →SDEE 30→SAEE 40→TSFD 53→TISE 15 →TIYD 161→TCAD 217→TTGE	6 →SLEE 30→VDE 53→TISE 151→TIYD 160→TTC 165→SEMD	6 →SLEE 30→VDE 53 →TISE 151→TIYD 160→TTCE 165→SEMD	6 →SLGE 30→SVEE 53→TISE 151→TIYD 160→TTCE 165→SEMD	6→SLEE 30→SVEE 53→TIGE 160→TTCD
(Casein kinase II phosphorylation site) [ST].{2}[DE]						
TYR_PHOSPHO_SITE	154→KSEMDLVKY	171→KHKGEKSY	164→KSEMDLVKY	164→KSEMDLVKY	164→RSEMDLVKY	-
(Tyrosine kinase phosphorylation site) [RK].{2,3}[DE].{2,3}Y						
MYRISTYL	54→GSCWAI 95→GCYGGI 165→TSYSV 223→TQGGV 252 →SNECG 261→GVAGT	64→GSCWAI 105→CQGGI 233→GVQNGT 262 →GTDECG	64→SCWAI 105→CYGGI 175→GTSYSV 233→GTQGGV 262→SNECG 271→GVAGT	64→SCWAI 105 →CYGGI 175 →TSYSV 233→TQGGV 262 →SNECG 271→GGVAGT	64→GSCWAI 105→GCHGGGI 223→GTQDGV 262→GNNECK 271→GGVAGI	64→GSCWAI 105→GCYGGI 233→GVKDGI 262→GNDECG
(N-myristoylation site) G[^EDRKHPFYW].{2}[STAGCN][^P]						
THIOL_PROTEASE_CYS	50→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA
(Eukaryotic thiol (cysteine) proteases cysteine active site) Q.{3}[GE].C[YW].{2}[STAGC][STAGCV]						
THIOL_PROTEASE_HIS	213→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG
(Eukaryotic thiol (cysteine) proteases histidine active site) [LIVMGSTAN].H[GSACE][LIVM].[LIVMAT]{2}G.[GSADNH]						
PREDICTED SECONDARY STRUCTURE	H→13.33 E→23.33 L→63.33	H→15.71 E→18.21 L→66.07	H→15.36 E→21.79 L→62.86	H→15.36 E→22.50 L→62.14	H→16.07 E→19.29 L→64.64	H→17.14 E→20.00 L→62.86
GLOBULARITY	nexp=62 nfit =115 diff =47.00	nexp=172 nfit=119 diff=53.00	nexp=164 nfit=119 diff =45.00	nexp=164 nfit=119 diff=45.00	nexp=165 nfit=119 diff=46.00	nexp=170 nfit=119 diff=51.00
	`nexp` -number of predicted exposed residues					
	`nfit` -number of expected exposed residues					
	diff -difference nexp-nfit					

Table 7. It shows the prediction of disulphide bond and different motifs of cathepsin B protein of six different *Leishmania* strains.

Leishmania strains	Disulfide bond	Length	Sequence	Domains
L. donovani	10-188	178	AKSALCLVAVF – ITTEVCQPYPF	1: 1-206 2: 207-340
	111-218	107	EHWPMCVTISE –YDTPKCNTTCE	
	123-166	43	RDQSNCGSCWA –ICGFGCYGGIP	
	126-162	36	SNCGSCWAIAA – SCCFICGFGCY	
	158-208	50	SNLLSCCFICG – DKYPPCPNTIY	
	159-326	167	NLLSCCFICGF – RGSNECGIESG	
	196-222	26	YPFGPCSHHGN KCNTTCEKSEM	
L. braziliensis	3-111	108	XXXMRCYTKF-DKWPKCRTISE	1: 1-206 2: 207-340
	123-166	43	RDQSNCGSCWA-VCGMGCQGGIP	
	126-326	200	SNCGSCWAIAA – RGTDECGIEST	
	140-218	78	MSDRYCTVAGI – YDTPTCNSTCA	
	158-222	64	GHLLSCCFVCG - TCNSTCADSHT	
	159-188	29	HLLSCCFVCGM - LTSEVCQPYPF	
	196-208	12	YPFPPCGHHTD - GKYPACPSTIY	
L. infantum	10-188	178	AKSALCLVAVF – ITTEVCQPYPF	1: 1-206 2: 207-340
	111-218	107	EHWPMCVTISE - YDTPKCNTTCE	
	123-166	43	RDQSNCGSCWA – ICGFGCYGGIP	
	126-162	36	SNCGSCWAIAA – SCCFICGFGCY	
	158-208	50	SNLLSCCFICG – DKYPPCPNTIY	
	159-326	167	NLLSCCFICGF – RGSNECGIESG	
	196-222	26	YPFGPCSHHGN - KCNTTCEKSEM	
L. chagasi	10-188	178	AKSALCLVAVF – ITTEVCQPYPF	1: 1-206 2: 207-340
	111-326	215	EHWPMCVTISE – RGSNECGIESG	
	123-166	43	RDQSNCGSCWA – ICGFGCYGGIP	
	126-162	36	SNCGSCWAIAA – SCCFICGFGCY	
	158-208	50	SNLLSCCFICG – DKYPPCPNTIY	
	159-208	49	NLLSCCFICGF – YDTPKCNTTCE	
	196-222	26	YPFGPCSHHGN - KCNTTCEKSEM	
L. major	10-159	149	AKSALCLVAVF – NLLSCCFICGL	1: 1-206 2: 207-340
	111-326	215	EHWPMCLTISE – RGNNECKIESG	
	123-188	65	RDQSNCGSCWA – IATEDCQPYPF	
	126-140	14	SNCGSCWAIAA – ISDRYCTFGGV	
	158-166	8	SNLLSCCFICG – ICGLGCHGGIP	
	196-208	12	YPFDPCSHHGN - EKYPPCPSTIY	
	218-222	4	YDTPKCNTTCE - KCNTTCERSEM	
L. mexicana	10-162	152	TKSALCLVAVF – SCCFICGFGCY	1: 1-206 2: 207-340
	111-218	107	EKWPMCVTIGE – YNTPKCNTTCD	
	123-166	43	RDQSNCGSCWA – ICGFGCYGGIP	
	126-326	200	SNCGSCWAIAA – RGNDECGIESS	
	158-208	50	TNLLSCCFICG - SKYPPCPNTIY	
	159-188	29	NLLSCCFICGF – VTTELCQPYPF	
	196-222	26	YPFGPCSHHGN - KCNTTCDNVEM	

and *L. major* and their patterns are shown in (Table 6), but in *L. major* at first protein kinase C phosphorylation site serine has been replaced threonine. In *L. infantum*, *L. chagasi* and *L. major* six identical motifs of casein kinase II phosphorylation site were observed likewise at 6th and 30th positions codes SLGE and SVEE in L.major. In *L. donovani* and *L. braziliensis* and *L. mexicana* predicted 5, 7 and 4 different patterns are present respectively (Table 6). Three different tyrosine kinase phosphorylation sites (154→ KSEMDLVKY, 171 → KHKGEKSY and 164→RSEMDLVKY) have been observed in cathepsin B of *L. donovani, L. braziliensis* and *L. major* respectively. No tyrosine kinase phosphorylation site was detected in catB protein of *L. major.* Six N-myristoylation sites with same pattern have been observed in catB protein of *L. infantum, L. chagasi* and *L. major* but in *L. major* 175th pattern are absent. In *L. braziliensis* and *L. mexicana* having same four sites were observed. Amino acid composition of N-myristoylation site in *L. donovani* is completely different from all five strains of *Leishmania.* One motif of eukaryotic thiol (cysteine) proteases active site and eukaryotic thiol (histidine) proteases active site is present in different strains of *Leishmania* (Table 6).

Seven disulfide bonds formed between different amino acids. Two domains were identified in six *Leishmania* strains 1st domain is formed between 1-206 amino acids and 2nd domain is formed between 207-340 amino acids shown in Table 7.

4. Future Perspectives

Homology modeling of six different strains of *Leishmania* cathepsin B protein provided for the first time its 3-D structure model which could be tested for screening different molecules for the *Leishmania* specific cathepsin B inhibitory activity (docking analysis). The developed model showed good overall structural quality, and is validated using PROCHECK, WHATIF program. Prediction of different functional sites like binding motifs, hydrolases sites, metal binding, glycosylation sites, protein kinase phosphorylation sites, N-myristoylation sites and different disulphide bridges are likely to be validated by experimental work. This knowledge could be used in biochemical studies to test the hypotheses of possible ligand binding sites. On the other hand, these experimental findings can then in turn be used to refine our models for virtual screening of chemical databases and rational drug design purposes. Advances in the field of insilico study will contribute to understanding between 3-D structure and ligand specificity of antileishmanial compound and it facilitate the development of various analogous of the presently available drug molecule on the basis of different binding sites of catB protein of different *Leishmania* strains.

Acknowledgements

This study was supported by a grant for setting up biomedical informatics centre from Indian Council of Medical Research (ICMR), Govt. of India. We acknowledge Dr. Sindhu Prava Rana for helping us in preparation of the manuscript. We thank Dr. Meera Singh, Scientist F of ICMR, New Delhi for helping us in setting up our biomedical informatics dept in RMRIMS, Patna, India.

References

1. P.J. Guerina, P. Olliarob, S. Sundard, M. Boelaerte, S.L. Croftf, P. Desjeuxg, M.K. Wasunnah, A. DM. Bryceson, *Lancet Infect.* Dis. 2 (2002) 494-501.
2. World Health Organization 1993 UNDP? World Bank/WHO 8, Leishmaniasis, Special Programme for Research and Training in Tropical Disease. Tropical Disease Research: Progress 1991-1992. Eleventh prgramme Report, 77-87.
3. B.L. Herwaldt, Lancet. 354 (1999) 1191-1199.
4. Y. Tselentis, A. Gilkas, B. Chaniotis, Lancet. 343 (1994), 1635.
5. R. Killick-Kendrick, 1979. Biology of *Leishmania* in phlebotomine sand flies. In Biology of the Kinetoplastida, W. Lumsden and D. Evans, editors. Academic Press, New York, USA.
6. J.M. Ribeiro, P.A. Rossignol, A. Spielman, Comp. *Biochem. Physiol.* 4 (1986) 683–686.
7. R. Charlab, J.G. Valenzuela, E.D Rowton, J.M. Ribeiro, *Proc. Natl. Acad. Sci.USA* 26 (1999) 15155–15160.
8. J. C. Mottram, D. R. Brooks, G. H. Coombs, Curr. *Opin. Microbiol.* 1 (1998) 455-460.
9. M. Knop, H.H. Schiffer, S. Rupp, D.H. Wolf, *Curr. Opin. Cell Biol.* 5 (1993) 990–996.
10. P.J. Berti, A.C. Storer, J. Mol. Biol. 246 (1995) 273-283.
11. J.S. Bond, P.E. Butler, Annu. Rev. Biochem. 56 (1987) 333-364.
12. K. Takio, T. Towatari, N. Katunuma, D.C. Teller, K. Titani , *Proc. Natl. Acad. Sci. USA* 80 (1983) 3666-3670.
13. H. Kirschke, A.J. Barrett, 1987. Chemistry of lysosomal proteases in Lysosomes: Their role in protein breakdown (Glaumann, H. & Ballard, F.J., eds), Academic *Press,* London. 193–238.
14. K.M. Karrer, S.L. Peiffer, M.E. DiTomas, *Proc. Natl.Acad. Sci. USA 90* (1983) 3063–3067.
15. M. Barral-Netto, A. Barrel, C. E. Brownell, Y. A. W. Skeiky, L. R. Ellingsworth, D. R. Twardzik, S. G. Reed, *Science.* 257 (1992) 545-548.
16. M. E. Wilson, B. M. Young, B. L. Davidson, K. A. Mente, S. E. McGowan, *J. Immunol.* 161 (1998) 6148-6155.
17. C. A. Hunter, H. Bermudez, H. Beernink, W. Waegell, J. S. Remington, Eur. J. Immunol. 25 (1995) 994-1000.
18. C. Bogdan, M. Rollinhoff, Parasitol. Today 15 (1999) 22-28.
19. C. Bogdan, J. Paik, C. Vodovotz, C. Nathan, *J. Biol. Chem.* 267 (1992) 23301-23308.
20. C. Bogdan, *Behring Inst. Res. Commun.* 99 (1997) 58-72.
21. J. Massague, Annu. Rev. Biochem. 67 (1998) 753-791.
22. J. S. Munger, J. G. Harpel, P. E. Gleizes, R. Mazzieri, I. Nunes, D. B. Rifkin, *Kidney Int.* 51 (1997) 1376-1382.
23. T. M. Chu, E. Kawinski, *Biochem. Biophys. Res. Commun.* 253 (1998) 118-134.
24. G.C. Sahoo, M. Rani, M.R. Dikhit, W.A Ansari, P.Das, Structural Modeling, Evolution and Ligand Interaction of KMP11 Protein of Different *Leishmania* Strains. J Comput Sci Syst Biol 2 (2009) 147-158.
25. J.D. Thompson, D.G. Higgins, T.J. Gibson, *Nucleic Acids Res.* 22 (1994) 4673-4680.
26. C.S.Ganesh, R.D.Manas, R.Mukta, D.Pradeep, Homology Modeling and Functional Analysis of LPG2 Protein of Leishmania Strains. J Proteomics Bioinform 0: (2009) 032-050.
27. A. D. MacKerell Jr., B. R. Brooks, C. L. III Brooks, L. Nilsson, B. Roux, Y. Won, M. Karplus, CHARMM: The Energy Function and Its Parameterization with an Overview of the Program, *The Encycl. Of Comp.Chem,* 1998, pp. 1271-1277.
28. Discovery Studio, Accelrys, San Diego, CA, USA.
29. C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen, *Nucleic Acids Res.* 31 (2003) 3692-3697.
30. G.C.Sahoo, M.R. Dikhit, P. Das, Functional assignment to JEV proteins using SVM. Bioinformation 3 (2008) 1-7.
31. P. Puntervoll, R. Linding, C. Gemünd, D.S. Chabanis, M. Mattingsdal *Nucleic Acids Res.* 31 (2003) 3625-3630.
32. B. Rost, G. Yachdav, J. Liu, *Nucleic Acids Research* 32 (2004) W321-W326.
33. A. Bairoch, P. Bucher, K. Hofmann, *Nucleic Acids Research* 25 (1997) 217-221.
34. A. Ceroni, P. Frasconi, A. Passerini, A. Vullo, *Bioinformatics* 20 (2004) 653-659.
35. D. Turk, M. Podobnik, R. Kuhelj, M. Dolinar, V. Turk, *FEBS Letters* 384 (1996) 211-214.
36. D. Musil, D. Zucic, D. Turk, R.A. Engh, I. Mayr, R. Huber, T. Popovic, V. Turk, T. Towatari, N. Katunuma, W. Bode, *EMBO J.* 10 (1991) 2321-2330.
37. M. Cygler, J. Sivaraman, P. Grochulski, R. Coulombe, A.C. Storer, J.S. Mort, *Structure* 4 (1996) 405-416.