International Journal of Bioinformatics and Biological Science: v.1 n.1 p.87-94. March, 2013

Annotation of a hypothetical Protein (A2E4V9_TRIVA) in *Trichomonas vaginalis*

Anshika Singh[1], Neetu Singh[1], Satendra Singh[1], Budhayash Gautam[1], and Gulshan Wadhwa[*2]

[1]Department of Computational Biology & Bioinformatics, Jacob School of Biotechnology and Bioengineering, Sam Higginbottom Institute of Agriculture, Technology and Sciences-Deemed to be University, Allahabad – 211007, India
[2]Department of Molecular and cellular engineering, JSBB, SHIATS, Allahabad –211007, India
[3]Apex Bioinformatics Centre, Department of Biotechnology, Ministry of Science and Technology, CGO Complex, Lodhi Road, New Delhi – 110 003,India

[*]Corresponding Author: Gulshan Wadhwa gulshan@dbt.nic.in

Received: 24 December 2012; Accepted: 08 January 2013

ABSTRACT

Trichomonas vaginalis, the etiologic agent of trichomoniasis, is an anaerobic flagellated protozoan. Trichomoniasis marked by complications like preterm delivery, low birth weight, and increased mortality as well as predisposing to HIV infection, AIDS, and cervical cancer, is among the top most sexually transmitted diseases. The genome draft published in 2007 revealed many unusual genomic and biochemical features like, exceptionally large genome size of 160 Mb. Hypothetical proteins are predicted gene products that have no identifiable function assigned to them and comprise 10-60% of recognized open reading frames in annotated genomes. The analysis of hypothetical proteins provides an opportunity to search for novel drug targets and markers for treatment of Trichomoniasis, and better understanding of the large genome size of *T.vaginalis.* The present study shows the structural, sequential and phylogenetic analysis of a hypothetical protein of *T. vaginalis* A2E4V9_TRIVA, which on modeling revealed maximum alpha helices. The protein showed conserved domains of Rosemann and E2-binding superfamilies which contribute in ubiquitin mediated proteolysis and also in numerous dehydrogenases metabolic pathways such as glycolysis, and many other redox enzymes

Keywords: Vaginalis, *Hypothetical proteins*

BACKGROUND

Trichomonas vaginalis is an anaerobic, flagellated protozoan, a form of microorganism. The parasitic microorganism is the causative agent of trichomoniasis, and is the most common pathogenic protozoan infection of humans in industrialized countries(Soper 2004). Infection rates between men and women are the same with women showing symptoms while infections in men are usually asymptomatic. Transmission takes place directly because the trophozoite does not have a cyst. The

WHO has estimated that 160 million cases of infection are acquired annually worldwide(Harp and Chowdhury ,2011). Adverse consequences to women with trichomoniasis include enhanced risk for human immunodeficiency virus transmission (Rein 1990) other complications resulting from infection are cervical cancer and bad pregnancy outcomes(Cotch *et al* ,1997).

The *T.vaginalis* genome was found to be approximately 160 megabases in size – ten times larger than predicted from earlier gel-based chromosome sizing while the human genome is ~3.5 gigabases by comparison(Singh *et al* ,2012 and 2011) As much as two-thirds of the *T.vaginalis* sequence consists of repetitive and transposable elements(Lehker and Alderete 1999), majority of hypothetical proteins, hydrogenosomes instead of mitochondria etc.

In an attempt to define HPs (Galperin *et al* ,2004) defined conserved hypothetical proteins as a large fraction of genes in sequenced genomes encoding those that are found in organisms from several phylogenetic lineages but have not been functionally characterized and described at the protein chemical level. These structures may represent up to half of the potential protein coding regions of a genome. Prediction of protein function, structure, essentiality and sub-cellular localization for hypothetical proteins is an important component of protein description in genome annotation(Ranjana *et al* ,2011).

At present, around 50,539 total proteins are present in UNIPROTKB database, out of which 39,083 protein sequences are available as putative uncharacterized proteins which are hypothetical proteins in the *T. vaginalis* proteome. This means that gene sequence information of *T. vaginalis* is at minimum, on par with that of its host *Homo sapiens*, since the functional annotation of hypothetical proteins is yet to be performed.

The hypothetical proteins provide a opportunity to search for novel drug targets and markers for treatment of Trichomoniasis, and better understanding of the large genome size of *T.vaginalis*.

In the present study various Bioinformatics tools like Blast,catch,cath scop ,prodom, sopma, etc were employed for an essential hypothetical putative protein A2E4V9_TRIVA, retrieved from Unipot kb, which was analyzed to gain further insight into its secondary structure details, composition, family and its relevant domains and motifs. The protein's cellular and metabolic pathways were analyzed along with its phylogenetic relevance using multiple sequence alignment. The homology modeling of the hypothetical protein was carried out using ModWeb and evaluated using Procheck- Ramchandran plot.

MATERIALS AND METHODS

Essentiality of datasets of hypothetical proteins of *T.vaginalis* retrieved from Uniprot kb (www.uniprot.org), was checked using Blast(Altschul *et al* ,1990) results in Database of essential genes (http://tubic.tju.edu.cn/deg/). As a case study, A2E4V9_TRIVA was selected for *in-silico* sequential, structural, functional and phylogenetic analysis and characterization. The physiochemical properties were calculated using PROTPARAM (http://web.expasy.org/protparam). PS2, SOPMA and HHPred were used for secondary structure prediction respectively. The analysis of protein family, conserved domains, and motifs was carried out using PFAM (www.pfam.sanger.ac.uk/)(Coin *et al*,2003) ,PRODOM (www.prodom.prabi.fr*/),* PROSITE(Bairoch,1984),CATH(www.cathdb.info/),NCBI-CDD (www.ncbi.nlm.nih.gov/cdd), and InterPRO-SCAN (http://www.ebi.ac.uk/inter)(Mulder et al ,2005). Prediction of pathway localization was made through KEGG (www.genome.jp/kegg/) and CELLO (http://cello.life.nctu.edu.tw) was used for sub-cellular localization. ModWeb (online portal for MODELLER-www.salilab.org)(Eswar *et al,*2003) was used for homology modeling of A2E4V9_TRIVA. The results were evaluated using PROCHECK-Ramchandran plot (*www.ebi.ac.uk/thornton-srv/software/PROCHECK/*). UCSF Chimera (*www.cgl.ucsf.edu/chimera/*) was used for viewing the 3D structure of the hypothetical protein. Phylogenetic analysis was performed using multiple sequence alignment in UNIPROT.

RESULTS AND DISCUSSION

The hypothetical protein A2E4V9_TRIVA retrieved from Uniprot kb was confirmed to be 'essential' from Database of Essential Genes using NCBI-BLAST results. The physio-chemical properties were then calculated using PROTPARAM. The formula of the hypothetical protein was found to be $C_{2008}N_{529}O_{60}S_{24}$ and considered as STABLE with instability index 30.84 and GRAVY value -0.227 (Table 1).

SOPMA and PS^2 computed the secondary structure having alpha helix 39.75 % , extended strand 17.28%, Beta turn 3.21% and Random coil 39.75%.

PROSITE and PFAM described the hypothetical protein to be a part of CoA-ligase family (Succinyl-CoA-Synthetase clan). The repeated pattern was noted as G-x-[IVT]-x(2)-[LIVMF]-x-[NAK]-[GS]-[GA]-G-[LMAI]-[STAV]-x(4)-[DN]-x-[LIVM]-x(3,4)-[GD]- [GREAK].

SMART and CATH identified the protein to be a 3-layer (aba) sandwich structure. The studied protein contains CoA_binding domain. This domain has a Rossmann fold and is found in a number of proteins including succinyl CoA synthetases, malate and ATP citrate ligases. InterPROScan also confirmed presence of CoA binding domain along with E2_binding domain and Ub_act_enzyme domain. NCBI_CDD predicted the protein to be a part of NADB_Rossmann and E2_bind superfamilies. The Rossmann fold NAD(P)H / NAD(P)+ is found in numerous dehydrogenases of metabolic pathways such as glycolysis, and many other redox enzymes. In E2 binding domain, E1 and E2 enzymes play a central role in ubiquitin and ubiquitin-like protein transfer cascades (Figure 1). This is an E2 binding domain that is found on NEDD8 activating E1 enzyme. The domain resembles ubiquitin, and recruits the catalytic core of the E2 enzyme Ubc12 in a similar manner to that in which

Table 1: Physiochemical properties of A2E4V9_TRIVA

S.No.	Properties	
1	NUMBER OF AMINO ACIDS	405
2	MOLECULAR WEIGHT	45100.8
3	THEORETICAL PI	5.61
4	NEGATIVELY CHARGED RESIDUES (ASP+GLU)	52
5	POSITIVELY CHARGED RESIDUES (ARG+LYS)	44
6	ALANINE	6.9%
7	ARGININE	2.7%
8	ASPARAGINE	5.2%
9	ASPARTIC ACID	6.2%
10	CYSTEINE	3.7%
11	GLUTAMINE	4.4%
12	GLUTAMIC ACID	6.2%
13	GLYCINE	7.7%
14	HISTIDINE	2.0%
15	ISOLEUCINE	8.1%
16	LEUCINE	7.7%
17	LYSINE	8.1%
18	METHIONINE2.	2%
19	PHENYLALANINE	3.5%
20	PROLINE	4.7%
21	SERINE	3.7%
22	THREONINE	4.9%
23	TRYPTOPHAN	0.7%
24	TYROSINE	4.2%
25	VALINE	6.7%
26	FORMULA	$C_{2008}H_{3163}N_{529}O_{601}S_{24}$
27	EXT. COEFFICIENTS	42705 $M^{-1}cm^{-1}$
28	ABS. COEFFICIENTS	0.947
29	EXT. HALF LIFE	30 hours
30	INSTABILITY INDEX	30.84 (STABLE)
31	GRAVY	-0.227

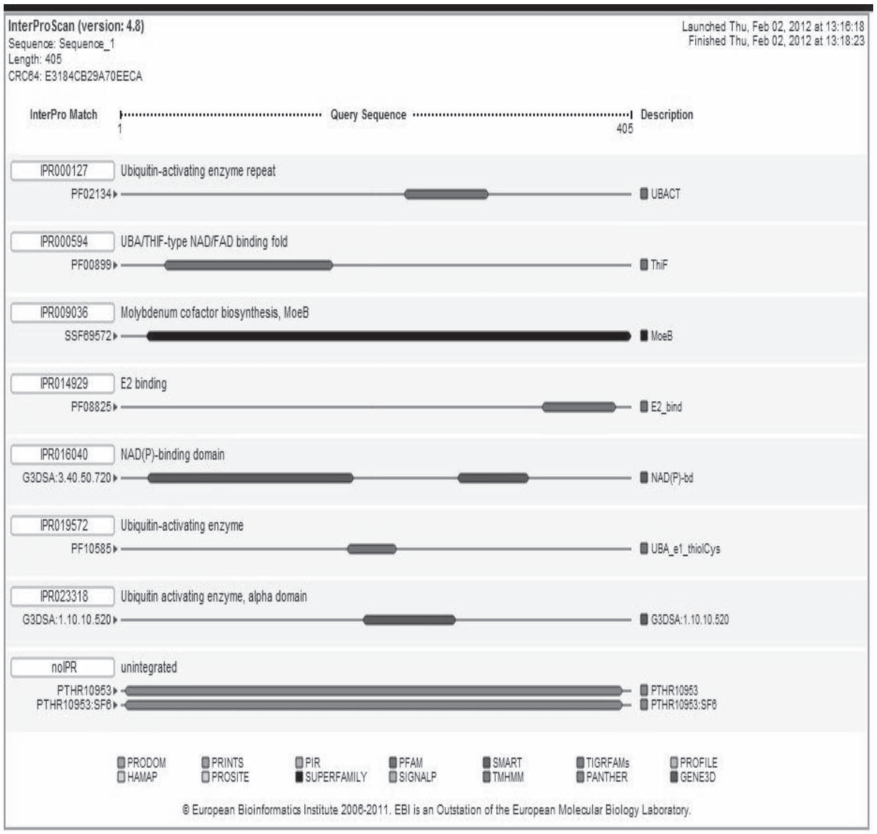
ubiquitin interacts with ubiquitin binding domains.

Cytoplasmic localization of the hypothetical protein A2E4V9_TRIVA was confirmed by PSORTII and CELLO program. According to KEGG, the protein was found to be a Ubiquitin mediated enzyme contributing in Ubiquitin mediated proteolysis (Figure 2).

The protein model was generated (Figure 3). For validation, SAVS-PROCHECK was used, where as Ramachandran plot validated 83.4% residues in most favored regions, 16.3% residues in additional allowed regions and 0.3% in generously allowed regions (Figure 4). The 3D structure was viewed through UCSF CHIMERA.

Multiple Sequence Alignment of the hypothetical protein A2E4V9_TRIVA was performed using UNIPROT. *T.vaginalis* was aligned with various sequences for phylogenetic analysis, where it was found to be closest to Chromosome undetermined scaffold_56 (*Paramecium tetraurelia*).

Therefore, with increase in proteome and genome information of organisms, deciphering hypothetical proteins is crucial. Annotation of hypothetical proteins provides information regarding its sequence, structure, motif, pattern, subcellular and pathway localization, signature and phylogenetic profiling. Having more



InterProScan (version: 4.8)
Sequence: Sequence_1
Length: 405
CRC64: E3184CB29A70EECA

Launched Thu, Feb 02, 2012 at 13:16:18
Finished Thu, Feb 02, 2012 at 13:18:23

© European Bioinformatics Institute 2006-2011. EBI is an Outstation of the European Molecular Biology Laboratory.

Figure 1: InterProScan results showing various domains

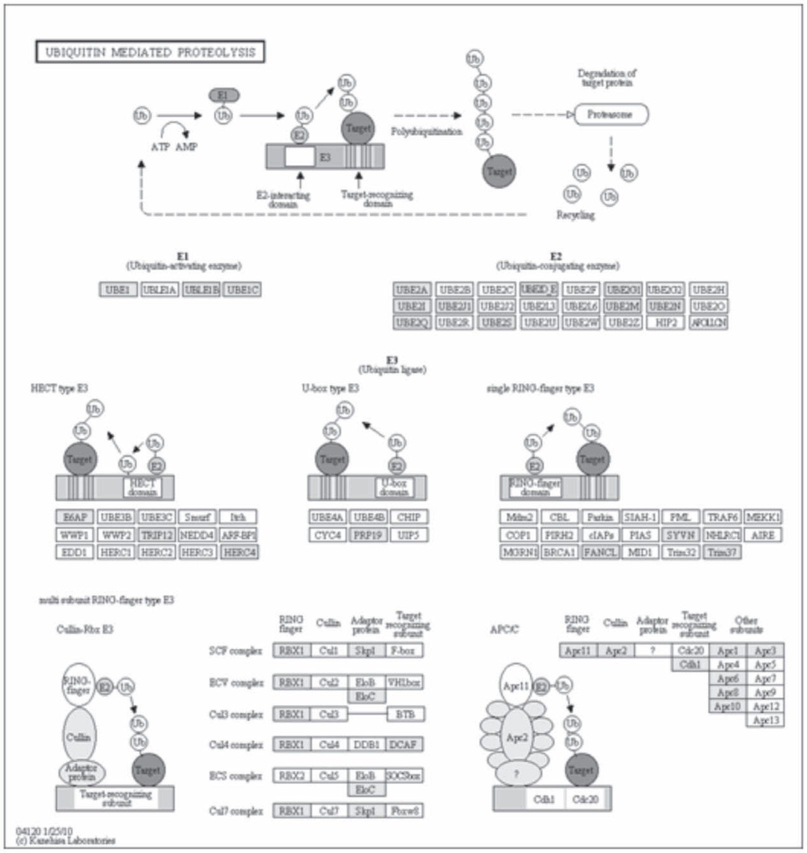

UBIQUITIN MEDIATED PROTEOLYSIS
Ub
ATP AMP
E1
E2
E3
Target
E2-interacting domain
Target-recognizing domain
Polyubiquitination
Degradation of target protein
Proteasome
Recycling
E1
(Ubiquitin-activating enzyme)
UBE1 UBLE1A UBLE1B UBE1C
E2
(Ubiquitin-conjugating enzyme)
UBE2A UBE2B UBE2C UBE2D_E UBE2F UBE2G1 UBE2G2 UBE2H
UBE2I UBE2J1 UBE2J2 UBE2L3 UBE2L6 UBE2M UBE2N UBE2O
UBE2Q UBE2R UBE2S UBE2U UBE2W UBE2Z HIP2 APOLLON
E3
(Ubiquitin ligase)
HECT type E3
HECT domain
E6AP UBE3B UBE3C Smurf Itch
WWP1 WWP2 TRIP12 NEDD4 ARF-BP1
EDD1 HERC1 HERC2 HERC3 HERC4
U-box type E3
U-box domain
UBE4A UBE4B CHIP
CYC4 PRP19 UIP5
single RING-finger type E3
RING-finger domain
Mdm2 CBL Parkin SIAH-1 PML TRAF6 MEKK1
COP1 PIRH2 cIAPs PIAS SYVN NHLRC1 AIRE
MGRN1 BRCA1 FANCL MID1 Trim32 Trim37
multi subunit RING-finger type E3
Cullin-Rbx E3
RING-finger
Cullin
Adaptor protein
Target-recognizing subunit
RING finger
Cullin
Adaptor protein
Target recognizing subunit
SCF complex RBX1 Cul1 Skp1 F-box
BCV complex RBX1 Cul2 EloB VHLbox
EloC
Cul3 complex RBX1 Cul3 BTB
Cul4 complex RBX1 Cul4 DDB1 DCAF
ECS complex RBX2 Cul5 EloB SOCSbox
EloC
Cul7 complex RBX1 Cul7 Skp1 Fbxw8
APC/C
RING finger
Cullin
Adaptor protein
Target recognizing subunit
Other subunits
Apc11 Apc2 ? Cdc20 Apc1 Apc3
Cdh1 Apc4 Apc5
Apc6 Apc7
Apc8 Apc9
Apc10 Apc12
Apc13
Apc11
Apc2
?
Cdh1 Cdc20
04120 1/25/10
(c) Kanehisa Laboratories

Figure 2: KEGG results showing ubiquitin mediated proteolysis

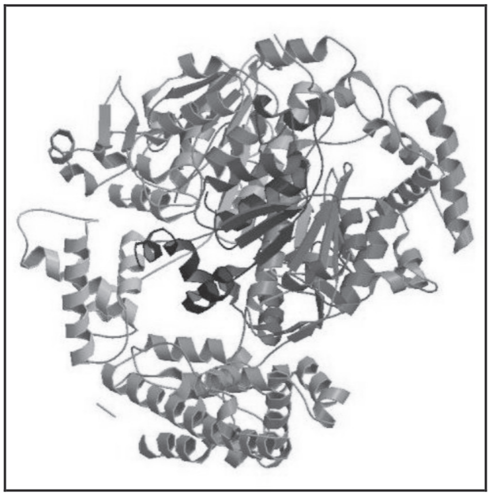
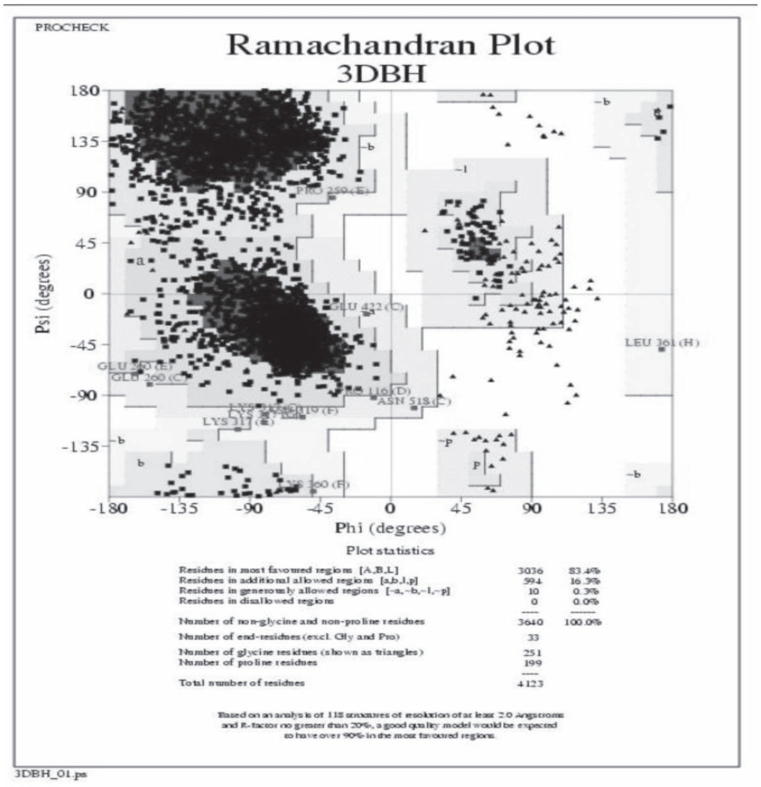


Figure 3: PDB image showing 3DBH.PDB structure



Plot statistics

```
Residues in most favoured regions  [A,B,L]              3036    83.4%
Residues in additional allowed regions  [a,b,l,p]        594    16.3%
Residues in generously allowed regions  [~a,~b,~l,~p]     10     0.3%
Residues in disallowed regions                             0     0.0%
                                                        ----   ------
Number of non-glycine and non-proline residues          3640   100.0%

Number of end-residues (excl. Gly and Pro)                33

Number of glycine residues (shown as triangles)          251
Number of proline residues                               199
                                                        ----
Total number of residues                                4123
```

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

Figure 4: Ramchandran plot for 3DBH.PDB

annotated information and characterization of the proteome will help in gaining further insight into the various metabolic, gene regulatory and functional aspects of the organisms. Phylogenetic analysis reveals the evolution and development of a particular protein and also helps in finding its related organism protein derivatives. The hypothetical proteins can also prove to be novel drug targets and markers for further medical research which is the need of the hour.

CONCLUSION

The 'essential' hypothetical protein A2E4V9_TRIVA was found to be the part of NADB_Rossmann and E2_binding domain superfamilies, with CoA_binding, E2_binding and Ub_act_binding domains, which contribute in ubiquitin mediated proteolysis and dehydrogenases metabolic pathways, indicates toward its crucial importance in the functioning of *T.vaginalis*. The *in-silico* structural, sequential, functional and phylogenetic annotation of hypothetical proteins hence prove to be very beneficial in gaining further insight in the genome structure of *T.vaginalis.*

ACKNOWLEDGEMENT

The authors are deeply grateful to Sam Higginbottom Institute of Agriculture, Technology and Sciences for providing the necessary guidance, facilities and support required for completing the research work.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:**403-410.

Bairoch, A and Bucher, P., 1994. PROSITE: recent developments. *Nucleic Acids Res*. **22:** 3583-3589.

Coin, L., Bateman, A., and Durbin, R. 2003. Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl. Acad. Sci. USA*. 100 (2003) 4516-4520.

Cotch, M.F., Pastorek, J.G. R.P. Nugent, S.L. Hillier, R.S. Gibbs, D.H. Martin, D.A. Eschenbach, R. Edelman, J.C. Carey, J.A. Regan, M.A. Krohn, M.A. Klebanoff, A.V.and Rao, G.G. 1997. Rhoads, and the Vaginal Infections and rematurity Study Group, *T. vaginalis* associated with low birth weight and preterm delivery. *Sexually Transmitted Disease.* **24:**353–360.

Eswar, N., Bino, J., Nebojsa, M., Andras, F., Valentin, A.I., Ursula, P., Ashley, C.S., Marc, A.M.R., Madhusudhan, M.S., Bozidar, Y and Andrej, S. 2003. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res*., **31:**3375-3380.

Galperin, M.Y and Koonin, E.V. 2004. Conserved hypothetical proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* **32:**5452–5463.

Harp, D.F and Chowdhury, I. 2011. Trichomoniasis: Evaluation to execution. *European Journal of Obstetrics & Gynecology and Reproductive Biology*. 157 (2011) 3–9.

Lehker, M.W. and Alderete, J.F. 1999. Resolution of Six Chromosomes of *T. vaginalis* and Conservation of Size and Number among Isolates. *The Journal of Parasitology,* **85** 976–979.

M.F. Rein and M. Muller, 1990. *T. vaginalis* and Sexually transmitted diseases. McGraw-Hill, New York, 481–492.

Mulder, N.J. , R. Apweiler, T.K. Attwood, A. Bairoch and A. Bateman, *et al.*, 2005. Inter-Pro, progress and status in 2005. *Nucleic Acids Res*. **33:**201-205.

N. Ranjana, J. Singh, A. Shefali, S. Maneet, Annotation of hypothetical proteins orthologous in *Pongo abelii* and *Sus scrofa*. *Bioinformation*. 6 (2011) 297-299.

Singh, S., Singh, G., Sagar, N., Yadav, P. K., Jain P.A., Gautam, B. and Wadhwa, G. 2012. Insight into *Trichomonas vaginalis* genome evolution through metabolic pathways comparison. *Bioinformation* **8**(4):189-195

Singh, S., Singh, G., Singh, A. K., Gautam, B., Farmer, R., Lodhi, S. S. and Wadhwa, G. 2011. Prediction and Analysis of Paralogous Proteins in *Trichomonas vaginalis* Genome. *Bioinformation,* **6**(1):31-34

Soper, D. 2004. Trichomoniasis: under control or under-controlled? *American Journal of Obstetrics and Gynecology*. **190:**281-290.