BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(1)

Prediction and analysis of paralogous proteins in *Trichomonas vaginalis* genome

Satendra Singh[1], Gurmit Singh[2], Atul Kumar Singh[3], Budhayash Gautam[1], Rohit Farmer[1], Sharad S Lodhi[4], Gulshan Wadhwa[4]*

[1]Department of Computational Biology & Bioinformatics, JSBB, SHIATS, Allahabad-211007, India; [2]Department of Computer Science and Information Technology, CET, SHIATS, Allahabad-211007, India; [3]Department of Biotechnology, Madhav Institute of Technology and Science, Gwalior – 474005, India; [4]Department of Biotechnology, Ministry of Science &Technology, New Delhi- 110003, India; Gulshan Wadhwa - Email: gulshan@dbt.nic.in; Phone: +91-9811301820; Fax: +91-11-24362884; *Corresponding author

Received February 07, 2010; Accepted February 21, 2010; Published March 02, 2011

Abstract:
Trichomonas vaginalis causes trichomoniasis, second most sexually transmitted disease. The genome sequence draft of *T. vaginalis* was published by The Institute of Genomic Research reveals an abnormally large genome size of 160 Mb. It was speculated that a significant portion of the proteome contains paralogous proteins. The present study was aimed at identification and analysis of the paralogous proteins. The all against all search approach is used to identify the paralogous proteins. The dataset of proteins was retrieved from TIGR and TrichDB FTP server. The BLAST-P program performed all against all database searches against the protein database of *Trichomonas vaginalis* available at NCBI genome database. In the present study about 50,000 proteins were searched where 2,700 proteins were found to be paralogous under the rigid selection criteria. The Pfam database search has identified significant number of paralogous proteins which were further categorized among different 1496 paralogous protein in pfam families, 1027 paralogous protein contains domain, 60 proteins were having different repeats and 1092 paralogous protein sequences of clans. Such identification and functional annotation of paralogous proteins will also help in removing paralogous proteins from possible drug targets in future. Presence of huge number of paralogous proteins across wide range of gene families and domains may be one of the possible mechanisms involved in the *T. vaginalis* genome expansion and evolution.

Keywords: *T. vaginalis*, pseudogenes, Paralogous proteins.

Background:
Trichomonas vaginalis is a unicellular, anaerobic, flagellated protozoan **[1]**. Infection with *T. vaginalis* cause of trichomoniasis, number one nonviral and second most sexually transmitted disease (STD) resulting in more than 250 million infections in women each year in the world **[2]**. *T. vaginalis* transmitted mostly by sexual contact. Adverse consequences to women with trichomoniasis include enhanced risk for human immunodeficiency virus transmission **[3]**; other complications resulting from infection are cervical cancer and bad pregnancy outcomes **[4]**. The recently published draft genome sequence of *T. vaginalis* by The Institute of Genomic Research (TIGR) reveals an abnormally large genome size of 160 Mb which is ten times the previously predicted size of this genome **[5]**. It is not still clear why *T. vaginalis* possesses such a large genome, and how such massive gene expansion happened. There are two possible important mechanisms which may be responsible for large scale genome expansion. It may be either through lateral gene transfer or through large scale gene duplication events. Lateral transfer is the process by which genetic information is passed from one genome to an unrelated genome, where it is stably integrated and maintained **[6]**. This genome is bigger than those of many other medically important protists but is characteristic of trichomonads. One reason for the large *Trichomonas* genome is the presence of hundreds of DNA transposons **[7]**. But in case of gene duplication a non functional copy of a gene get incorporated in the host genome. Many protein families underwent massive duplication. Pseudogenes are DNA sequences that were derived from a functional copy of a gene but which have acquired mutations that are deleterious to function. This duplicated copy of original functional gene gets incorporated into a new chromosomal location may leading to expansion of the existing gene family **[8]**. The genome also gives the platform to construct and analyze some important signal, secretary and metabolic pathway to identify and validate novel targets, which can be harvested to designed new drug molecules. Sequence similarity search methods provide some insights into putative functions for most gene products. Huge number of pseudogenes was thought to be present in *T. vaginalis* due to massive gene duplication. In case of *T. vaginalis* TIGR predicted that there are about 50,000 genes in *T. vaginalis* but did not mention about pseudogenes. It was speculated that a significant portion of the 50,000 genes might be pseudogenes. Proteins are generally comprised of one or more functional regions, commonly termed domains. Aims of the study were: (i) Identification of paralogous proteins, (ii) Prediction of families, domains and repeats of identified paralogous proteins and (iii) To investigate the role of paralogous proteins in the genome expansion of evolution of *T. vaginalis.*

Methodology:

Identification of Paralogous proteins:

The complete set of proteins predicted from the *T. vaginalis* genome was retrived from the FTP server of the TrichDB database (http://trichdb.org/trichdb/) and TIGR (FTP directory ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_vaginalis/annotation_dbs/) **[9]**. Around 50,000 proteins in the FASTA format retrieved from the database were used to carry out the all against all database searches by using the genomic BLAST-P available at NCBI server **[10]**. In case of all against all search, a comparison was made in which every predicted protein sequence was used as a query in a similarity search against a database composed of the rest of the self-proteome, and the significant matches are identified by a low E-value. The *T. vaginalis* proteome database is present at NCBI. Protein sequence was searched at E-Value 0 or less than 0. Since many proteins comprise different combinations of a common set of domains, proteins that align more than 80% of their lengths for query and subject were selected **[11]**. After this filtration only those alignment were selected which give the sequence identify more than 60%.

Prediction of families, domain and repeats in paralogous proteins:

For the purpose of functional annotation and to investigate the gene family expansion, the identified set of paralogous proteins was used to search the protein families by using the Pfam search. The Pfam database is a large collection of protein domain families. Each family is represented by multiple sequence alignments and Hidden Markov models (HMMs). The paralogous protein dataset was submitted at Pfam server which predicted the protein families, motifs, repeats and clans at the default pfam parameter (http://pfam.sanger.ac.uk/) **[12]**.

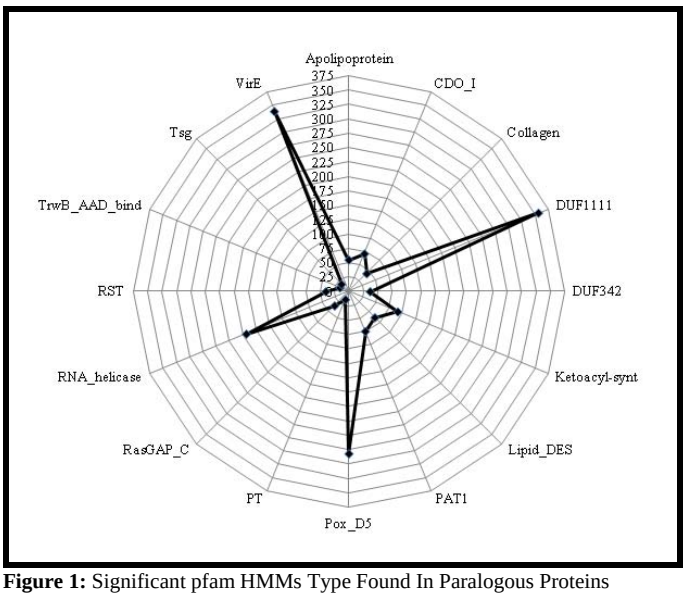

Figure 1: Significant pfam HMMs Type Found In Paralogous Proteins

Results and Discussion:

After using rigid selection criteria for BLASTP search (very low E-value,>60% sequence identity and >80% alignment length) 2700 protein sequences were found to be paralogous proteins and around 47,200 proteins were identified as non paralogous proteins as they do not match with any protein of the proteome. The various protein families, domains, repeats and clans for the paralogous protein were identified with the help of Pfam sequence search. Total 1496 paralogous protein were found in different pfam families (collection of related proteins), 1027 sequences contains different pfam domain (structural unit which can be found in multiple protein contexts), 3 sequences have pfam motif (short unit found outside globular domains) and 60 proteins contains different pfam repeats (short unit which is unstable in isolation but forms a stable structure when multiple copies are present) **Table 1 & 2 (see Supplementary material)**. Total 1092 paralogous protein sequences contain pfam clan (collection of families that have arisen from a single evolutionary origin) and 1494 proteins does not belong to any clan.

Some of significant protein families are Adeno_E4 (362), CDO_I (71), DUF1111 (357), PAT1 (75), VirE (339) followed by domains Alpha-2-MRAP_C (213), Pox_D5 (282), RNA_helicase (193), Lipid_DES (64), Ketoacyl-synt (92), Apolipoprotein (55) and significant repeats are PT (15), Collagen (44) **Figure 1**. Similarly some of significant predicted clan are CL0318 (356), CL0123 (280), CL0023 (209), CL0046 (92), CL0029 (79), CL0194 (19) and CL0044 (17) **Figure 2**. Some other clan also present but not in significant value are CL0028 (5), CL0219 (5), CL0125 (4), CL0236 (3), CL0281 (3), CL0020 (3), CL0063 (1), CL0119 (1), CL0072 (1), CL0183 (1) and CL0295 (1). Here we can clearly see the evidences of evolutionary relationship among paralogous protein in the form of sequence motifs, protein families, domain and repeats **[12]**.

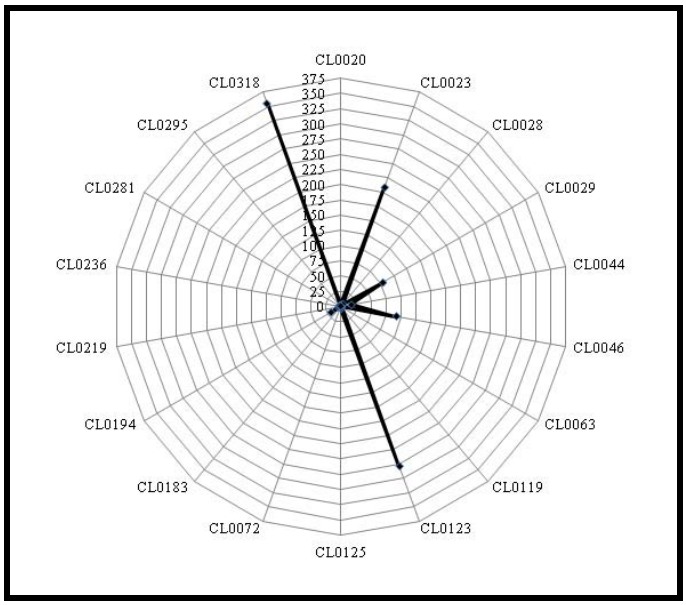

Figure 2: Pfam Clan predicted for paralogous proteins

There are other protein families were only one member of paralogous protein is present. Some of such protein families found in *T. vaginalis* genome are Adeno_E3_CR2, 3H, AnfG_VnfG, Dak2, Dor1, DUF1151, DUF1524, DUF2078, DUF3508, DUF357, DUF562, DUF752, DUF912, DUF947, FliD_C, FliS, Glycophorin_A, KiIA-N, Mid2, MtrF, MyTH4, PfUIS3, Phage_30_3, Podoplanin, Rabaptin, Roughex, T4SS, Tobravirus_2B, Tom37, Transposase_1, and Transposase_7. Similarly CRM1_C, CTP_transf_2, Hat1_N, KorB, MetRS-N, NurA, PAS, PBP5_C and Ribosomal_L30_N are the domains where only single member of paralogous protein is identified. The CCT is the only identified pfam motif having three members of paralogous proteins.

Large number of pseudogenes were already reported in many families of protein for example, ankyrin repeat proteins, hypothetical protein, conserved hypothetical protein, adenylate cyclase, vsaA, surface antigen BspA, ANK-repeat protein, CG1651-PDrelated, ABC transporter protein, kinases, major facilitator superfamily protein, leucine rich repeat family protein, and Transmembrane amino acid transporter protein **[7, 13]**. These pseudogenes may be playing active role in the formation of paralogous protein. The New gene functions are thought to be gained by duplication of an existing gene creating different tandem copies. Functional differentiation then occurs between the copies by mutation and selection.

We found 2700 paralogous protein which is present across wide range of different protein families, domain, clan and repeats. This clearly reflects that many protein families underwent massive duplication in the *T. vaginalis* genome. The expansion of genetic material and amplification of specific gene may be the example of adaptations of the *T. vaginalis* during its transition to a urogenital environment from enteric environment (the habitat of most trichomonads) **[5, 14]**. We hope that after a larger survey on individual duplicated protein families and having more experimental data on the paralogous protein, we could shed light on biological issues like, how genes were duplicated and their evolution histories.

The presence of different domains in varying combinations in different proteins gives rise to the diverse repertoire of proteins found in genome. Identifying the domains present in a protein can provide insights into the function of that protein. Such identification of paralogous proteins and their functional

annotation will not only give insight into the biological mechanism of genome but also help in identification of the novel drug targets. The identified paralogous proteins can be excluded from the possible list of drug targets, as paralogous proteins represents non functional product of duplicated genes known as pseudogenes **[15]**.

The identified paralogous proteins and their sequence in the FASTA format can be retrieved using the *T. vaginalis* protein accession number from http://trichdb.org/trichdb/ for future analysis**.** The amino acid sequence of the predicted hypothetical proteins encoded by the predicted genes can be used as a query of the protein sequence databases in a database similarity search. A match of a predicted protein sequence to one or more database sequences not only serves to identify the gene function, but also validates the gene prediction. The genome sequence can further be annotated with the information on gene content and predicted structure, gene location, and functional predictions **[16]**.

Conclusion:

Collectively, these data suggest the presence of a very large number of paralogous proteins in unicellular eukaryote *T. vaginalis*. Presence of paralogous proteins across wide range of protein families, domain, repeats, clans and motifs reflects large scale gene duplication events leading to gene family expansion. The identification of paralogous proteins indicates the possible role of gene duplication in the evolutionarily expansion of the *T. vaginalis* genome because organisms considered to be deep-branching have both paralogs. For further investigation the paralogous proteins can be subjected to cluster analysis in order to identify the most closely related groups of proteins.

Acknowledgement:

The authors are grateful to the Sam Higginbotom Institute of Agriculture, Technology & Sciences, Deemed University, Allahabad for providing the facilities and support to complete the present research work.

References:

[1] A Donne. *C R Hebd Seances Acad Sci.* 1836 **3**: 385
[2] Wisdom AR & Dunlop EM. *Br J Vener Dis.*1965 **41**: 90 [PMID: 14332084]
[3] Sorvillo F & Kerndt P. *Lancet.* 1998 **351**: 213 [PMID: 9449891]
[4] Cotch MF *et al. Sex Transm Dis.*1997 **24**: 353 [PMID: 9243743]
[5] Carlton JM *et al. Science* 2007 **315**: 207 [PMID: 17218520]
[6] de Koning AP *et al. Mol Biol Evol.* 2000 **17**: 1769 [PMID: 11070064]
[7] Silva JC *et al. Mol Biol Evol.* 2005 **22**: 126 [PMID: 15371525]
[8] Cui J *et al. Genome Inform.* 2007 **18**: 35 [PMID: 18546472]
[9] Aurrecoechea C *et al. Nucleic Acids Res.* 2009 **37**: D526 [PMID: 18824479]
[10] Altschul SF *et al. Nucleic Acids Res.* 1997 **25**: 3389 [PMID: 9254694]
[11] Rubin GM *et al. Science* 2000 **287** (5461): 2204 [PMID: 10731134]
[12] Finn RD *et al. Nucleic Acids Res.* 2010 **38**: 211 [PMID: 19920124]
[13] Lawrence JG. *Curr Opin Microbiol.* 1999 **2**: 519 [PMID: 10508729]
[14] Singh S *et al. Int J Pharmaceutical Sci Rev and Res.* 2010 **3** (1): 38
[15] Li W *et al. Bioinformatics* 2001 **17**: 282 [PMID: 11294794]
[16] Ashurst JL *et al. Nucleic Acids Res.* 2005 **33**: 459 [PMID 15608237]

Edited by AU Khan

Citation: Singh *et al.* Bioinformation 6(1): 31-34 (2011)

BIOINFORMATION

open access

Supplementary material:

Table1: Predicted pfam Families in Paralogous Protein

S.No	HMM Name	HMM Accession	Number of proteins
1.	ACTH_domain	PF00976.11	2
2.	Adeno_E4	PF05385.4	362
3.	CDO_I	PF05995.5	71
4.	Cytochrom_B562	PF07361.4	2
5.	DegS	PF05384.4	4
6.	DNA_pol_B	PF00136.14	12
7.	DNA_pol_B_2	PF03175.6	7
8.	DUF1111	PF06537.4	357
9.	DUF1381	PF07129.4	4
10.	DUF1421	PF07223.4	5
11.	DUF2203	PF09969.2	3
12.	DUF2223	PF09985.2	2
13.	DUF2785	PF10978.1	2
14.	DUF3015	PF11220.1	3
15.	DUF342	PF03961.6	37
16.	DUF605	PF04652.9	7
17.	DUF654	PF04910.7	2
18.	EpuA	PF11772.1	10
19.	ESAG1	PF03238.6	7
20.	FadA	PF09403.3	4
21.	Fic	PF02661.11	4
22.	FTCD_C	PF04961.5	3
23.	FtsK_SpoiIIIE	PF01580.11	2
24.	MCPsignal	PF00015.14	4
25.	PAT1	PF09770.2	75
26.	Peptidase_C78	PF07910.6	4
27.	Peptidase_S28	PF05577.5	5
28.	Phage_pRha	PF09669.3	4
29.	Pirin	PF02678.9	9
30.	Pox_A32	PF04665.5	3
31.	PRKCSH	PF07915.6	3
32.	RasGAP_C	PF03836.8	35
33.	RE_LlaJI	PF09563.3	3
34.	ResIII	PF04851.8	2
35.	TF_AP-2	PF03299.7	3
36.	Transposase_24	PF03004.7	4
37.	Transposase_5	PF01498.11	2
38.	Tsg	PF04668.5	18
39.	VirE	PF05272.4	339

Table 2: Predicted pfam Domains in Paralogous Protein

S.No	HMM Name	HMM Accession	Number of proteins
40.	Alpha-2-MRAP_C	PF06401.4	213
41.	Apolipoprotein	PF01442.11	55
42.	BRO1	PF03097.11	5
43.	Cdc6_C	PF09079.4	2
44.	DEAD	PF00270.22	2
45.	DUF1910	PF08928.3	3
46.	EMP24_GP25L	PF01105.17	8
47.	Flu_M1_C	PF08289.4	3
48.	GAGA_bind	PF06217.5	2
49.	Ketoacyl-synt	PF00109.19	92
50.	LBR_tudor	PF09465.3	4
51.	Lipid_DES	PF08557.3	64
52.	Pox_D5	PF03288.9	282
53.	RNA_helicase	PF00910.15	193
54.	RNA_pol_Rpc4	PF05132.7	6
55.	RST	PF12174.1	40
56.	Rve	PF00665.19	5
57.	Spc7	PF08317.4	2
58.	Talin_middle	PF09141.3	3
59.	TFIIH_BTF_p62_N	PF08567.4	8
60.	TrwB_AAD_bind	PF10412.2	17