www.jatit.org

IDENTIFYING THE NOVEL DOMAIN INVOLVED IN HUMAN PATHOGENESIS

SWETHA R G

Department of Bioinformatics, Auxilium College (Autonomous), Affiliated to Thiruvalluvar University, Vellore, Tamil Nadu, India. E-mail: <u>swetha.rayapadi@gmail.com</u>

ABSTRACT

Functional prediction of human hypothetical proteins remains a major project for biologist to carry out their further research in various life science and healthcare segments. Human chromosome 1 is the one which is very largest and has many hypothetical proteins. So it is attempted to find the function of some of the human chromosome 1 hypothetical proteins, the proteins are clustered based on their sequence similarities and identify the consensus sequence in those proteins. Using various prediction methods their functional annotations are carried out and a domain was predicted which is involved in the pathogenic functionality.

Keywords: Function Prediction, Hypothetical Protein, Consensus Sequence, Domain.

1. INTRODUCTION

When genetics was new, time-consuming lab experiments were the only way to determine what proteins did. But, now the functional genomics can able to determine the function of genes and proteins as a means to understand life, health and illness. Currently, most approaches to protein function prediction rely on searching sequence databases for homologous sequences with prior annotation. However, the function for one protein cannot be inferred from another when similarity is <40% sequence identity. Moreover, studies on enzyme proteins have shown that the precise function diverges below identities of 60%.

. The following are the general classes of function prediction:

- Sequence based approaches
- Structure-based approaches
- · Motif-based approaches

The detection of function-related local 3D-motifs in protein structures can provide insights towards protein function in absence of sequence or fold similarity. Protein loops are known to play important roles in protein function and several loop classifications have been described, but the automated identification of putative functional 3D-motifs in such classifications has not yet been addressed. This identification can be used on sequence annotations

Most of the work on predicting protein function uses a single source of in-formation – the most common being the amino acid sequence of the protein(see Chapter 30). There are, however, a number of sources of data that are predictive of protein function.

The general categories of function are:

i) **Information:** storage, maintenance of the genetic code; DNA replication/repair; general transcription/translation.

ii) **Regulation:** regulation of gene expression and protein activity; information processing in response to environmental input; signal transduction; general regulatory or receptor activity.
iii) **Metabolism:** anabolic and catabolic processes; cell maintenance/homeostasis; secondary metabolism.

iv) Intra-cellular processes: cell motility/division; cell death; intra-cellular transport; secretion.
v) Extra-cellular processes: inter-, extra-cellular processes, e.g. cell adhesion; organismal processes, e.g. blood clotting, immune system.
vi) General: general and multiple functions; interactions with proteins/ions/lipids/small molecules.

vii) **Other/Unknown:** unknown function, viral proteins/toxin.

2. HYPOTHETICAL PROTEINS OF HUMAN CHROMOSOME 1:

www.jatit.org

Genes with unknown function are called orphan genes while their transcripts and peptides are called hypothetical proteins. There are many genes and their associated proteins that remain uncharacterized in the human genome.

In biochemistry, a hypothetical protein is a protein whose existence has been predicted, but for which there is no experimental evidence that it is expressed in vivo. The usual scenario involving a hypothetical protein is in gene identification during genome analysis. When the bioinformatics tool used for the gene identification finds a large open reading frame without an analog in the protein database, it returns "hypothetical protein" as an annotation remark.

There are many hypothetical proteins in Human chromosome1 whose function is not yet identified. Chromosome 1 is the designation for the largest human chromosome. People normally have two copies of chromosome 1, as they do with all of the autosomes, which are the non-sex chromosomes. Chromosome 1 spans about 247 million nucleotide base pairs, which are the basic units of information for DNA. It represents about 8% of the total DNA in human cells. Identifying genes on each chromosome is an active area of genetic research. Chromosome 1 is currently believed to have 4,220 genes, exceeding previous predictions based on its size. It was the last completed chromosome, sequenced two decades after the beginning of the Human Genome Project. The number of single nucleotide polymorphisms (SNPs) is about 740.000.

3. LITERATURE REVIEW:

Prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterized family for which many sequences and structures are known. However, these inferences are tenuous. Such methods provide reasonable guesses at function, but are far from foolproof.

The computational assignment of protein function may be the single most vital application of bioinformatics in the post-genome era. These assignments are made based on various protein features, where one is the presence of identifiable domains.

Protein domains, the building blocks of all globular proteins, are units of compact threedimensional structure as well as of molecular evolution. The growth of sequence databases has elevated the need for computational annotation of proteins, and the detection of conserved domains is one of the first steps towards assigning molecular function.

Domains vary in length from between about 25 amino acids up to 500 amino acids in length. The shortest domains such as zinc fingers are stabilized by metal ions or disulfide bridges. Domains often form functional units, such as the calcium-binding EF hand domain of calmoduin. Because they are self-stable, domains can be "swapped" by genetic engineering between one protein and another to make chimera proteins.

The concept of the domain was first proposed in 1973 by Wetlaufer after X-ray crystallographic studies of hen lysozyme, papain and by limited proteolysis studies of immunoglobulins. Wetlaufer defined domains as stable units of protein structure that could fold autonomously.

In the past domains have been described as units of:

- Compact structure.
- Function and evolution.
- Folding.

Each definition is valid and will often overlap, i.e. a compact structural domain that is found amongst diverse proteins is likely to fold independently within its structural environment. Nature often brings several domains together to form multidomain and multifunctional proteins with a vast number of possibilities. In a multidomain protein, each domain may fulfill its own function independently, or in a concerted manner with its neighbors. Domains can either serve as modules for building up large assemblies such as virus particles or muscle fibres, or can provide specific catalytic or binding sites as found in enzymes or regulatory proteins.

Covalent association of two domains represents a functional and structural advantage since there is an increase in stability when compared with the same structures non-covalently

www.jatit.org

associated. Other, advantages are the protection of intermediates within inter-domain enzymatic clefts that may otherwise be unstable in aqueous environments, and a fixed stoichiometric ratio of the enzymatic activity necessary for a sequential set of reactions.

4. MATERIALS AND METHODS:

Literature Review has been proven that the protein function prediction will be more accurate once the domain is identified.

Step 1:

The hypothetical protein sequence of human chromosome 1 is collected from the NCBI database. Though there are many protein database, NCBI is widely used, because it will develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease.

Step 2:

The collected hypothetical proteins are clustered by means of ClustalW tool. Multiple alignments of protein sequences are important tools in studying sequences. The basic information they provide is identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families.

Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment). This is true for pairwise and multiple alignments. Global alignments need to use gaps (representing insertions/deletions) while local alignments can avoid them, aligning regions between gaps.

ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences. The alignment is progressive and considers the sequence redundancy. Trees can also be calculated from multiple alignments. The program has some adjustable parameters with reasonable defaults. It calculates the best match for the selected sequence, and lines them up so that the identities, similarities and differences can b seen. Evolutionary relationships can be seen via viewing phylograms or dendrograms.

Step 3:

Using the PFP server, the function of hypothetical proteins can be predicted. PFP is designed to predict Gene Ontology annotations for a query protein sequence beyond what can be found by searching conventional databases. The PFP algorithm has been shown to increase coverage of sequence-based function annotation more than fivefold by extending a PSI-BLAST search to extract and score GO terms individually and include information from distantly related sequences. It applies a novel data mining tool, the Function Association Matrix (FAM), to score significantly associating pairs of annotations.

The current version of the prediction method uses a sequence-based method. The aim is to utilize information from relatively weak hits in PSI-BLAST, which are not conventionally used. Typically, weak hits in PSI-BLAST are not perfect orthologs to the query sequence, but rather share a common functional domain. In addition to simply transferring the function of the common domain to the query sequence, the idea is to also consider those functions which are frequently associated with the annotated functions of the domain. To this end, it has built Function Association Matrices (FAMs) that quantify the co-occurrence of Gene Ontology (GO) annotations in sequences of the UniProt database. The GO is a controlled hierarchical vocabulary describing the function of genes in three categories: function, process, and component. Approximately two thirds of associated function pairs mined from UniProt bridge functions of different categories. Thus, it can assign function using FAMs that cannot be retrieved directly from highly similar sequences or structures.

The results are listed as the top ten most probable Gene Ontology annotations in the Biological Process, Molecular Function, and Cellular Component categories. It should be note that these automated function predictions are not intended to be perfectly accurate, but only to represent the statistical probability that your sequence matches the listed function annotations according to known associations extracted from a variety of publicly available functional databases. These results were cross-referenced to the GO database version GO_200702 to create a list of primary annotations for query sequence.

Step 4:

www.jatit.org

There are some secondary structure prediction tools which are commonly used to predict the secondary structure of protein domain. They are:

SOPM:

The first step of the SOPM is to build subdatabases of protein sequences and their known secondary structures drawn from 'DATABASE.DSSP' binary by (i) making comparisons of all protein sequences and (ii) taking into account the prediction of structural classes of proteins. The second step is to submit each protein of the sub-database to a secondary structure prediction using a predictive algorithm based on sequence similarity. The third step is to iteratively determine the predictive parameters that optimize the prediction quality on the whole sub-database. The last step is to apply the final parameters to the query sequence.

SOPMA:

Recently a new method called the selfoptimized prediction method (SOPM) has been described to improve the success rate in the prediction of the secondary structure of proteins. This improved SOPM method (SOPMA) correctly predicts 69.5% of amino acids for a three-state description of the secondary structure (alpha-helix, beta-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25% identity) proteins.

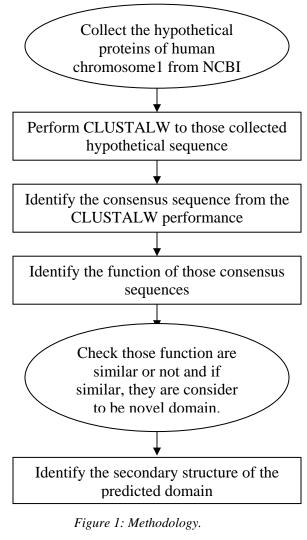
GOR IV:

The GOR method is based on information and was developed by J.Garnier, theory D.Osguthorpe and B.Robson (J.Mol.Biol.120, 97, 1978). The present version, GOR IV, uses all possible pair frequencies within a window of 17 amino acid residues. After cross validation on a data base of 267 proteins, the version IV of GOR has a mean accuracy of 64.4% for a three state prediction (Q3). The program gives two outputs, one eye-friendly giving the sequence and the predicted secondary structure in rows, H=helix, E=extended or beta strand and C=coil; the second gives the probability values for each secondary structure at each amino acid position. The predicted secondary structure is the one of highest probability compatible

SIMPA 96:

A secondary structure prediction algorithm is proposed on the hypothesis that short homologous sequences of amino acids have the same secondary structure tendencies. Comparisons are made with the secondary structure assignments of Kabsch and Sander from X-ray data [(1983) Biopolymers 22, 2577-2637] and an empirically determined similarity matrix which assigns a sequence similarity score between any two sequences of 7 residues in length. This similarity matrix differs in many respects from that of the Dayhoff substitution matrix.

5. METHODOLOGY:

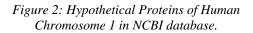


6. RESULTS AND DISCUSSIONS:

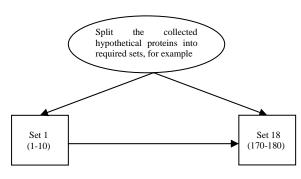
6.1) There are 180 hypothetical proteins of Human chromosome 1 in the NCBI database. They are retrieved.

www.jatit.org

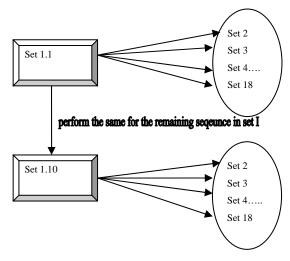
neb Disearch 🚖 Fav										
Search Star										
	otes 🙆	8.3	3							
istesjerbrez										-
al 👸 Windows 😰 Winds	ws Mariatpi	ice 👔 Windo	vd Media							
• @» Ø M • 🗗	1 800	inals• 📓	2 blocked	7 Ored	+ Anton	4 • 13 kits		ndto• 🖇		Q Sett
this from	154	-11-1	1 3G	eno	me				Market	n Resiste
far Dahart Deah			(Ge (C	lear					
V Box 500 V S	end to 👻									
osome 1, reference	assemi	bly, comple	tte sequ	ence						
		-			1.4	-				
to end	Length	ten0	10 11		Hete	sh Heset				
					7968					
	format & D	utula in CADT	1 from at de	Dentain C	unter & Darris					
ut & TAU service in EASTS										
u <mark>t ©</mark> DNA region in FASTA	Start	End		Length		GenelD	Locus	Locus_tag	C06 (4)	Links
u <mark>t ©</mark> DNA region in FASTA In F. member 5							Locus	Locus_tag		Links ++++
	Start	End	Strand	Length	G 53020740	GenelD	-	Locus_tag		
	Start 59954	End 59071	Strand	Length 305	G 53020740	GenelD 79501 100133632	-	Locus tag		****
	V Go = C M + B + C + C + C + C + C + C + C + C + C	v Go of M + B + Q too v Go of M + B + Q too v Good v Go	Image: Section of the section of t	Image: The second se	0 0	Image: State of the state o	Image: State of the state	Image: The second se	Image: Image	Image: Section of the section of t



6.2) The collected hypothetical proteins are submitted to the ClustalW program and perfoms the cluster as follows:



Like wise split all the 180 hypothetical protein and perform the ClustalW those sets, Here the example of clustering with Set I.



Similarly perform for the remaining sequence in all the set.

Address Dittp://www.abi.a		
	r an Toolon Antania Prese thabaset 🜒 Windows 100 Windows Markatplace 🜒 Windows Media	
Google G-QUETALW	- So - O M + B + O Bolevate- D Illichel O Ceck - Santre - Santre - Sentre - 3 "	Q Settinger
MATT MUSCLE T-Cutter	and a def and and a def and and and and a def	1
	OUTPUT OUTPUT TREETIPE CORRECTORT KINORE GAPS	
ClustelW Programmatic Access	ah whatters 👻 algoed 👻 noos 👻 off 👻 af 👻	
· www.clustel.org	Enter to packs a safe of sequences in any supported format. Help	
	1941 14937731 FETUER DEBUGGET. REPORTED AND ADDRESS ADDRES ADDRESS ADDRESS ADD	
	Vetting a file Browse Run Read	
	¥ you plan to one show a working a cost or planar (<u>under</u>) and. Planar cost the <u>(An)</u> buffer a working high from our support staff.	
Terrest Liter - UN Pureline	Orthol 101 - O European Bountametrics Institute 2008-2008 IBB is an Outstation of the European Mitmoder Distance Laboratory	
	© 264	

Figure 3: ClustalW Home Page.

From the number of ClustalW performance, the consensus sequence is identified and given below:

ClustalW2 Results Number of sequences 10 Alignment score 99035 Sequence format Sequence type Pearson aa JalView × Output file clustalw2-20080520-10532195.output Alignment file clustalw2-20080520-10532195.aln Guide tree file clustalw2-20080520-10532195.dnd Your input file clustalw2-20080520-10532195.input SUBMIT ANOTHER JOB To save a result file right-click the file link in the above table and choose "Save Target As". If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets Scores Table Sort by Sequence Number 🚽 🛛 View Output File

www.jatit.org

gi|14149997|ref|NP_115640.1| gi|114158664|ref|NP_001041445.

WIVHSGERELTGIREKIGEGRDASRSINGHIGALITPDEPDNSGGRDI 200 WIVHSGERELIGIREKIREGRDASHSINGHIGALITPDKHDNSGGQDF 164

interm interm interm interm interm interm
<pre> print interverter menter interverter menter interverter interverter</pre>
<pre> 1 p12974041ct p12 p7990.11 1 p139541ct p12 p7900.11 1 p139541ct p</pre>
a pitelity is off possible in the second
piperint piperint <td< td=""></td<>

0 0
pilletitistering pilletititistering pilletitistering <
111272001211E110000011E110000011E11000001110000000
upper definition upper definition <td< td=""></td<>
u115706311ctfff_0000013, 00 0 0 0 0 011096011ctfff_000002000 02000000
u115026011cr010_0000013 00 0 01170260120 00000120000000000000000000000000000000
111252621111111 0000 000 000
appl://production.com/production.co
pilspresticter (m versus, no
pillspredstrict product pillspredstrict pillspre
pillspredstrict product pillspredstrict pillspre
pi:1979001:ret:#_0021.it pi:001 pi:1979001:ret:#_0021.it pi:001001:ret#_0021.it pi:001001:ret
ap1134991114-000 ap11444000 ap1114440000 ap11444000 ap114
pi14502101ctfly 000513. pi1411438641ctfly 000513. pi14114438641ctfly 000513. pi1414438641ctfly 000513. pi1414438641ctfly 000513. pi1414438641ctfly 000513. pi1414438641ctfly 000513. pi1414438641ctfly 000513. pi1414458641ctfly 000513. pi141458641ctfly 000513. pi1414458641ctfly 000513. pi141458641ctfly 000
pi11572620361rcf IP, 0000133, 76 7 pi11572620351rcf IP, 0000123, 76 7 pi11572620351rcf IP, 0000124, 74 pi11572620351rcf IP, 0000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11
pi11572620361rcf IP, 0000133, 76 7 pi11572620351rcf IP, 0000123, 76 7 pi11572620351rcf IP, 0000124, 74 pi11572620351rcf IP, 0000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 000124, 74 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11 pi11572620351rcf IP, 0001276, 11
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
april 10001401 (eff [0] 0001401. april 10001401. bp bp april 10001401. bp <
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
a) 114355641rcf1m_001445. a) 2) 7 a) 1144 b) 2) 7
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
a) 11434564/1et fbp_00014446, 41 10 a) 1105141001et fcb_00117450, 67 67 69 a) 115707171et fbp_0540,11 91 6 a) 115707171et fbp_001023, 161 CEDD LISKHERER - OFEELLa
g119701711rcf1W_00543.1.1 91 7 g112070021rcf1W_00543.1.1 91 7 g112070021rcf1W_00543.1.1 92 g112070021rcf1W_001006133.1.1 92 g112070017rcf1W_001006133.1.1 92 92 g112070017rcf1W_001007rcf1W_001007rcf1W_001007rcf1W_001007rcf1W_001007rcf1W_001007rcf1W_001007rcf1W_001007rcf1W_001007rcf1W_00
iii 197001711cct HP 05030.11 91 9112002971cct HP 05030.11 9112010971cct HP 001095133. 911201201971cct HP 05030.11 911201201971cct HP 05030.11 911201201971cct HP 05030.11 911201201971cct HP 001095133. 911201201971cct HP 001097154.1 911201201971cct HP 001097154.1 911201201971cct HP 001097154.1 911201201971cct HP 001097154.1 9112101971cct HP 00109713.1 9
g1:37307:10:ft HP_050307:10:ft HP_051030.1: 9:1144:09971:10:ft HP_051037:10:ft HP_0110:ft HP_010:ft HP
p1:01:01:01:ret IP 0603:1 p1:02:02:02:01:ret IP 72:00:1 p1:02:02:02:02:02:02:02:02:02:02:02:02:02:
p1:statustisering_control p1:sta
pi statusti ret im gedulation in the down the resonance of the source in the source the source
g1139780451ret (107_75500.1) 670 9 g1141499971ret (107_11064.1) 633 73 g1139780451ret (107_75500.1) 670 10 g110561110101et (157_00171246.0, 647) 21 g1129789451ret (107_11064.1) EXEMPT (107_11066.1) EXEMPT (107_11066.1) EXEMPT (107_11066.1) EXEMPT (107_11066.1) EXEMPT (107_11066.1) EXEMPT (107_11066.1
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
g1 (834 (51/4) ref IVP_00103764.1 EEE EFE TUDERKIEEGRASSIMENLOALITPEPERSOF 412 G1 (834 (51/4) ref IVP_00103764.1 EEE EEE TUDERKIEEGRASSIMENLOALITPEPERSOF 412 G1 (834 (51/4) ref IVP_00103764.1 EEE TUDERKIEEGRASSIMENLOALITPEPERSOF 412 G1 (834 (51/4) ref IVP_0010313.1 EEE EEE TUDERKIEEGRASSIMENLOALITPEPERSOF 337 G1 (13507) 400 ref IVP_0010513.3 EEE EEE TUDERKIEEGRASSIMENLOALITPEPERSOF 337 G1 (13507) 400 ref IVP_0010513.3 EEE EEE TUDERKIEEGRASSIMENLOALITPEPERSOF 337 G1 (13507) 400 ref IVP_0010513.3 EEE EEE TUDERKIEEGRASSIMENLOALITPEPERSOF 337 G1 (13507) 400 ref IVP_0010513.3 EEE EIF 400 1157266285 (ref IVP_00104145. DV3 334 :: : 1161 G1 (135701) ref IVP_00105613.3 CYSTPSDYLEIPDLGQPYSAVYSLEOYLGLALAVDR 400 1157266285 (ref IVP_001096132. 1161/27794/051 (ref IVP_001096132. 1161/27794/051 (ref IVP_0010971/450. 401 G1 (13570405) ref IVP_00109613.3 CYSTPSDYLEIPDLGQPYSAVYSLEOYLGLALAVDR 400 1157266285 (ref IVP_001096132. 1161/27794/051 (ref IVP_0010971/450. 1161/27794/051 (ref IVP_0010971/450.
Lack NUTE Some Korts may be maning from the above table if the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple CPU more. Respected in the alignment was dow samp multiple CPU more. Respected in the output left of the alignment was dow samp multiple. Respected in the output left of the alignment was dow samp multiple. Respected in the output left of the alignment was dow samp multiple. Respected in the output left of the alignment was dow samp multiple. Respected in the output left of the alis align alis aligned dow same was dow samp multiple CP
LAGE OVEC 2004 scores may be mising from the above table of the adjust method. g1 921100 9 ref IIF _ 060410. 2 EEE TULERKLEEGRASSIMENUCALITPEDERSIG 337 Sontby Sequence Number V Wew Output File g1 150014001 ref IIF _ 001095133. EEE
g11197098251 cef NP_901095132, EEE TULERKIEEGRDASSIENTELIALITPDEPRS0C=37 iijument g11141396641 cef NP_001095133, g1141439971 cef NP_001095133, g1141439971 cef NP_001095133, g111572662851 cef NP_001095132, g11572662851 cef NP_001095133, g11572662851 cef NP_001095133, g11572662851 cef NP_001095132, g11572662851 cef NP_001095133, g11572662851 cef NP_001095133, g11572662851 cef NP_001095132, g11572662851 cef NP_001095133, g11572662851 cef NP_001095133, g11572670171 cef NP_055198,11 g11572662851 cef NP_001095133, g11572674,1 g11572670171 cef NP_055198,11 g11572674,1 g11572674,1 g1157267411 cef NP_001095133, g11572674,1 g11572674,1 g1157267411 cef NP_001095133, g11572674,1 g11572674,1 g115726741,1 g11572674,1 g11572674,1 g11572674,1 g11572674,1 g11572674,1 g115671401 cef NP_001095133, g114499971 cef NP_001095133, g11449971 cef NP_001095133, g111459641 cef NP_001095131, g11449971 cef NP_001095133, g11449971 cef NP_001095133, g111459641 cef NP_001095131, g1114596641 cef NP_001095133, g1114572662851 cef NP_001095133, <td< td=""></td<>
Sorty Sequence Number Wew Output File G11550714201 Lef INP_001095133. EEE TULERKLEEGRASESLIFELIQALITPOPERSION
gill41499971cefINF_0104445. gill572662851cefINF_001096132. CYSTF3DYLELFDLGQPYSSAYYSLEEQYLGLALDVDR
Nigmment gill 114158664/ref NP_001041445. D75 334 Show Colors View Alignment File 334 LINTAL 2.0.5 walkiple sequence alignment gill 522662851ref NP_001096133. CYSTPSDYLELPDL60PYSSATYSLEEOYLGLALDVDR
Show Colors View Alignment File : Show Colors View Alignment File ::
Bits Construction
clustal 2.0.5 multiple sequence alignment gi 1/55/01/17ter HP_US195.11 CVTSTSULE/DL0/DUSTSVSTSEQUIOLALD/DWF 400 clustal 2.0.5 multiple sequence alignment gi 1/25/01/17ter HP_US195.11 CVTSTSULE/DL0/DUSTSVSTEQUIOLALD/DWF 400 gi 1/25/01/17ter HP_US195.11 CVTSTSULE/DL0/DUSTSVSTEQUIOLALD/DWF 430 gi 1/25/01/17ter HP_US195.11 CVTSTSULE/DL0/DUSTSVSTEQUIOLALD/DWF 430 gi 1/25/01/01/17ter HP_US195.11 CVTSTSULE/DL0/DUSTSVEEDDE/DEDE/DUSTSUEDDE/DEDE/DUSTSUEDDE/DEDE/DUSTSUEDDE/DEDE/DUSTSUEDDE/DEDE/DUSTSUEDDE/DEDE/DUSTSUEDDE/DEDE/DUSTSUEDDE/DEDE/DUSTSUEDDE/DUST
LINTEL 2.5.5 multiple expense alignments g1 [324]51/41/ef IIF _00103764.1
g11572662051cef1NP_001096133, g11572662051cef1NP_001096133, g11671201cef1NP_001096133, g11671201cef1NP_001096133, g11572662051cef1NP_00102744.1 g115210019cef1NP_00102744.1 g115210019cef1NP_0010204ECCLAQUIVQUSPENDEDDUVQUEEAE 451 g115210019cef1NP_00102744.1 g115201011cef1NP_00010204ECCLAQUIVQUSPENDEDDUVQUEEAE 533 g115007140001100104ECCLAQUIVQUSPENDEDDUVQUEEAE 533 g115210019cef1NP_000102744.1 g111500714001cef1NP_000103133, 00100004ECCCLAQUIVQUSPENDEDUVQUEEAE 533 g11150074301ref1NP_0001021445. g1114001401ref1NP_0001041445. 05070FCEGTFLALVEQXYCSAQ0VAS
g1192260311cef1NP_004913, g1192100311cef1NP_06404.0.21 01020L4ECCLAQULVQKISPRIDEDEDUVQVEALF 451 g119200311cef1NP_06410.21 01020L4ECCLAQULVQKISPRIDEDEDUVQVEALF 533 01020L4ECCLAQULVQKISPRIDEDEDUVQVEALF 533 g119200351cef1NP_0040374.4 01020L4ECCLAQULVQKISPRIDEDEDUVQVEALF 533
g1115070112et ND_054194.11 g111507040511ct ND_055251ct N
1194341741crflWP_00103744.1 011741699701crflWP_0103744.1 011741699701crflWP_01041445. 1190010101crflWP_001041445. 03170410147 031704101445. 111107002514crflWP_115640.1 03170410445. 031704101445. 111107002514crflWP_001041445. 03170410445. 03170410445. 111107002514crflWP_115640.1 03170410445. 03170410445. 111107002514crflWP_001041445. 03170410445. 03170410445. 11140399971crflWP_115640.1 03170410445. 03170410445. 11140399971crflWP_115640.1 03170410445. 03170410445. 11140399971crflWP_01041445. 03170410445. 03170410445. 11140399971crflWP_01041445. 031710710704704070470470470470470470470470470470
1194341741crflWP_00103744.1 011741699701crflWP_0103744.1 011741699701crflWP_01041445. 1190010101crflWP_001041445. 03170410147 031704101445. 111107002514crflWP_115640.1 03170410445. 031704101445. 111107002514crflWP_001041445. 03170410445. 03170410445. 111107002514crflWP_115640.1 03170410445. 03170410445. 111107002514crflWP_001041445. 03170410445. 03170410445. 11140399971crflWP_115640.1 03170410445. 03170410445. 11140399971crflWP_115640.1 03170410445. 03170410445. 11140399971crflWP_01041445. 03170410445. 03170410445. 11140399971crflWP_01041445. 031710710704704070470470470470470470470470470470
illusses/itelm/lises.il militalises/itelm/lises.il militalises/itelm/lises.il militalises/itelm/lises.il gills72662851ref18P_0010945. militalises/itelm/lises.il militalises/itelm/lises.il militalises/itelm/lises.il
ril43069/refm/100013 arrivforeilarvalaber/form3 3 gill572662851reflNP_00109613. RFFTLTVTSHLVF0MGVIFP0_670
Interpreter Second
Interpreter Second
cil76570171refINP_056198.11 BFFTLTVTSLHLVFOMGVTFP0_921
gi/297894051ref/NP 775909.1 RFFTLTVTSLHLVFOMOVIFP0 670
111572662851ref1NP 001096133WVVSAGPLSSEKAEMNILEINEKLRPOLAEKKOPFOLAEKKO
[11657017]red.HW_054294.1 FTITTYTSLHAFVDWLVTPU-942 [11657037]red.HW_054294.1 FTITYTSLHAFVDWLVTPU-942 [11579945]red.HW_054294.1 FTITYTSLHAFVDWLVTPU-942
183415174LCECINF 001032764.1WV336FW33EKAEMULLEINEKLEFOLAENKOOFGNLKERCL 44 011921100191refINF 060410.21 BELTLMGTSLHLVFOMGVTFP0 1214
1169161100/ref/XP 001717450WVY3AGPV3SEKAEMNILEINEKLRPOLAENKOOPGNLKERCFL 44 grill 10700825 Lyof UND 800228 21
19210091ce1WF_00040.21
gi 14149997 ref NP 115640.1 RFFTLTVIRHHLAFOMGVIFPH 633
III41505641rcfINF_001041445. SATIVISHVAADPUSCHAEMDILEIMUSCHAEMDOLPHAMMORFUX 64
1114156664]ccc1NF_001041445. SATNYTHYVSAOFNSSEXAEMUILEIMERLEFQLAAMNQOPSMDXQFLV 64
1572662851crc18P_001096133. TULAGFLANDGKTKYZECKDLIKFHLMENDOFKZEKLAEQLKQAZELDQ 94 OK EASE ANOTE: SASUKAT S
7657017 ref NP_056198.1
227894051ree1Hp_775909.1] TQLAGFLANRQKKYKYEECKDLIKFHLRNERQFKEEKLAEQLKQAEELRQ 94
183415174[ref]NP_001032764.1 TQLAGFLANROKKYKYEECKDLIKFMLRNEROFKEEKLAEQLKQAEELRQ 94 Show Colors View Alignment File
169161100/ref/XP_001717450. TQLAGFLANRQKXYKYEECKDLIKFHLRNERQFKEEKLAEQLKQAEELRQ 94
1921100191ref HP_06940.21 T0LAGFLARROKYKYECKDLIKPHLINREOFFEEKLAELIKOARELIKO 94
11197092521rct187_99223.31
1360/7420/TECHW/D0109513. TVXYFLARQUMYDYDCKDLIKSINA-SIGKDG/07-10L1Q 19 CHARGE 16
1410657/1CE/IF_ID001041445 T0MAFLANDURYTEECOLAULIALAULAULAULAULAULAULAULAULAULAULAULAULAU
t t = t =*t =*
Show as Phylogram Tree Show Distances View DND File
1572662851zef1NP_001096133. VKVLVHAGERELTOLREKLREGRDASRSLNEHLOALLTPDEPDKSGGDL_144
17657017 ref.HP_U56198.1 129789405 ref.HP_175509.1 YYVIUYHAGEFELTQEFEKLEEGEDASKSLMEHLGALITUCEVOKOGCOL 144
2278794961 ief 117 775590.1 YVV WHALEPEITLIKERKARSSLIKEULALITVEKKSJOL 144 AVALSIVIET VII VII VII VII VII VII VII VII VII VI
192110019/ref/IIP 060410.2/ VRVIVHSQERELTQLREKLREGROASRSLINGHLQALLTEDEPOKSQCQL 144 FIGURE 4: Clustal W Results, Yellow Shaaing
1113/99/2012/11/2019/2010/2013. 10 11/2014/2010/2010/2010/11/2014/2010/2010

Figure 4: ClustalW Results, Yellow shading indicates the Conserved region.

From the result of CLUSTALW, these ten sequences show the conserved region at **QYKVLVHAQERELTQLREKLREGRDASRS** LNEHLQALLTPDEPDKSQGQDLQEQLAEG CRLTQHLVQKLSPENDNDDDEDVQVEVAE KVQKSSAPREMQKAEEKEVPEDSLEECAIT CSNSHGPYDSNQPHKKTKITFEEDKVDSTLI GS.

This may be act as a common domain if they share the same functional property.

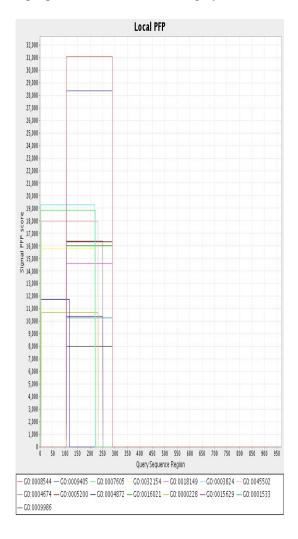


www.jatit.org

6.3) The sequences that are found have the conserved regions are submitted to the **protein function prediction server** and the functions are predicted for those sequences. The result of one such sequence is given below.



Local function assignment to regions by PFP for top 5 predictions in each GO category.



Prediction scores for top 5 predictions in each GO category GO term	Short Def	Raw Score	Term Type
GO0008544	epidermis development	31093.65	BP
GO0009405	pathogenesis	28348.88	BP
GO0007605	perception of sound	18814.05	BP
GO0032154	0032154	15815.04	BP
GO0018149	peptide cross- linking	14607.87	BP
GO0003824	catalytic activity	19268.02	MF
GO0045502	dynein binding	17966.18	MF
GO0004674	protein serine/threonine kinase activity	16382.07	MF
GO0005200	structural constituent of cytoskeleton	16321.62	MF
GO0004872	receptor activity	11752.9	MF
GO0016021	integral to membrane	16010.64	CC
GO0000228	nuclear chromosome	10695.14	СС
GO0015629	actin cytoskeleton	10389.45	CC
GO0001533	cornified envelope	10252.62	CC

Figure 5: PFP Results.

The result of all the sequences are tabulated and find weather the conserved regions of all the sequences possess same function or not.



www.jatit.org

	NCBI gi			Term
S.NO	number	Function predicted	Raw score	type
1	7657017	epidermis development	31093.65	BP
		Pathogenesis	28348.88	BP
		perception of sound	18814.05	BP
		32154	15815.04	BP
		peptide cross-linking	14607.87	BP
		catalytic activity	19268.02	MF
		dynein binding	17966.18	MF
		protein serine/threonine kinase activity	16382.07	MF
		structural constituent of cytoskeleton	16321.62	MF
		receptor activity	11752.9	MF
		integral to membrane	16010.64	CC
		nuclear chromosome	10695.14	CC
		actin cytoskeleton	10389.45	CC
		cornified envelope	10252.62	CC
		cell surface	7997.1	CC
2	92110019	epidermis development	31093.65	BP
		Pathogenesis	28348.88	BP
		perception of sound	18814.05	BP
		32154	15815.04	BP
		peptide cross-linking	14607.87	BP
		catalytic activity	19268.02	MF
		dynein binding	17966.18	MF
		protein serine/threonine kinase activity	16382.07	MF
		structural constituent of cytoskeleton	16321.62	MF
		receptor activity	11752.9	MF
		integral to membrane	16010.64	CC
		nuclear chromosome	10695.14	CC
		actin cytoskeleton	10389.45	CC
		cornified envelope	10252.62	CC
		cell surface	7997.1	CC
3	29789405	Synaptogenesis	22741.89	BP
		protein-nucleus import	17711.21	BP
		positive regulation of I-kappaB kinase/NF-kappaB cascade	16273.19	BP
		carbohydrate metabolism	15979.02	BP
		SOS response	14951.04	BP
		Catalytic activity	43390.77	MF
		receptor activity	18639.08	MF
I			10057.00	1711

Vol 7. No 1. (pp 018 - 030)



www.jatit.org

S.NO	NCBI gi number	Function predicted	Raw score	Term type
4	14149997	perception of sound	341455.4	BP
		Synaptogenesis	170897.63	BP
		protein transport	163190.59	BP
		plasma membrane repair	157105.91	BP
		protein-nucleus export	147918.06	BP
		protein-tyrosine kinase activity	203176.02	MF
		GTPase activity	171778.6	MF
		signal transducer activity	165762.94	MF
		protein serine/threonine kinase activity	134087.24	MF
		exonuclease activity	119295.99	MF
		actin cytoskeleton	324442.32	CC
		cornified envelope	186395.29	CC
		septin ring	150602.05	CC
		ciliary rootlet	124790.73	CC
		cell surface	102968.93	CC
5	169161100	perception of sound	341455.4	BP
		Synaptogenesis	170897.63	BP
		protein transport	163190.59	BP
		plasma membrane repair	157105.91	BP
		protein-nucleus export	147918.06	BP
		protein-tyrosine kinase activity	203176.02	MF
		GTPase activity	171778.6	MF
		signal transducer activity	165762.94	MF
		protein serine/threonine kinase activity	134087.24	MF
		exonuclease activity	119295.99	MF
		actin cytoskeleton	324442.32	CC
		cornified envelope	186395.29	CC
		septin ring	150602.05	CC
		ciliary rootlet	124790.73	CC
		cell surface	102968.93	CC
			102/001/0	
6	157266285	epidermis development	51134.59	BP
Ū	157200205	Synaptogenesis	20540.84	BP
		protein-nucleus import	16706.75	BP
		positive regulation of I-kappaB kinase/NF-kappaB cascade	15838.39	BP
		carbohydrate metabolism	15620.98	BP
		catalytic activity	42484.21	MF
		receptor activity	42484.21 17603.75	MF
I			1/005./5	IVIT

Vol 7. No 1. (pp 018 - 030)



www.jatit.org

S.NO	NCBI gi number	Function predicted	Raw score	Term type
7	83415174	ciliary or flagellar motility	691958.57	BP
		Neurogenesis	683191.29	BP
		regulation of embryonic development	355724.17	BP
		membrane protein ectodomain proteolysis	326061.39	BP
		perception of sound	310562.69	BP
		GTPase activity	312354.27	MF
		metallopeptidase activity	286230.15	MF
		protein-tyrosine kinase activity	285182.71	MF
		receptor activity	233443.7	MF
		protein transporter activity	213372.14	MF
		Midbody	341319.99	CC
		septin ring	330489.87	CC
		growth cone	275600.06	CC
		Synaptosome	200571.35	CC
		ciliary rootlet	187282.01	CC
8	119709825	Pathogenesis	54825.47	BP
		RNA splicing	42467.13	BP
		Neurogenesis	27276.7	BP
		ion transport	21731.84	BP
		Spermatogenesis	20643.32	BP
		GTPase activity	49460.02	MF
		protein heterodimerization activity	25501.28	MF
		protein-tyrosine kinase activity	24800.97	MF
		structural molecule activity	20357.76	MF
		hydrolase activity	18802.54	MF
		Myosin	29385.68	CC
		septin ring	26512.47	CC
		alpha-glucosidase complex	24887.99	CC
		growth cone	18243.4	CC
		ciliary rootlet	18141.28	CC
9	156071420	Pathogenesis	52781.82	BP
		RNA splicing	39914.09	BP
		Neurogenesis	26114.97	BP
		positive regulation of I-kappaB kinase/NF-kappaB cascade	23577.9	BP
		ion transport	20794.78	BP
		GTPase activity	47120.8	MF
		protein heterodimerization activity	24498.88	MF

Vol 7. No 1. (pp 018 - 030)

LATT

www.jatit.org

	NCBI gi			Term
S.NO	number	Function predicted	Raw score	type
10	114158664	Pathogenesis	2306360.89	BP
		protein transport	416503.48	BP
		polysaccharide catabolism	203038.47	BP
		glycerol-3-phosphate catabolism	130474.59	BP
		brain development	106689.33	BP
		metal ion binding	607029.88	MF
		ATPase activity	463755.4	MF
		metallopeptidase activity	297063.2	MF
		GTP binding	252123.28	MF
		zinc ion binding	225511.21	MF
		cell surface	658803.27	CC
		Chromosome	356615.29	CC
		actin cytoskeleton	143564	CC
		Myosin	74258.4	CC
		cornified envelope	66980.2	CC

Table 1: PFP Results for the 10 Consensus Sequence.

From the results of Protein Prediction server, it is found that those consensus sequences possess the similar function **pathogenesis**, so it might be considered as a domain.

6.4) The identified domain is submitted to the secondary structure prediction tool and the results are given below:

TOOL	Alpha I helix % s	Extended trand %	Beta urn %	Random coil %
SOPM	52.67	10.67	4.67	32.00
SOPMA	56.00	8.67	6.00	29.33
GOR IV	46.67	11.33	0.00	42.00
SIMPA96	56.95	7.95	0.00	34.44

Table 2: Secondary Structure Prediction Results.

Thus the secondary structure of the domain is predicted and it is found that it will contain majority of helices and random coils.

7. CONCLUSION:

The hypothetical proteins from the Human Chromosome 1 are retrieved and the sequences are clustered based on sequence similarity. A set of ten sequences are found to be more similar and they contain consensus sequence. These sequences are fortunately having the similar function, all because of the presence of that consensus sequence.

Thus the consensus sequence might be the novel domain for those ten sequences of Human Chromosome 1 which is not yet identified. The novel domain involved in the pathogenesis functionality. The secondary structure of the identified domain contains majority of the helix and random coil.

Thus this project may provide the pave to identify the novel domains which are not yet identified. This project is just an assumption and it is very difficult to cluster the large sets of hypothetical sequence at a time and even though the identified common domain possesses identical residues more, it shows few similarity residues.

www.jatit.org

REFERENCES:

- Anfinsen CB, Haber E, Sela M, White FH (1961). "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain". Proc Natl Acad Sci U S A 47 (9): 1309-14.
- [2]. Banner et al. (1975). "Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data". Nature 255: 609-614.
- [3]. Coleman W. 1965. Cell, nucleus, and inheritance: an historical discovery. Proc. Am. Philos. Soc. 109, 124-158.
- [4]. Copley, R. R. and Bork, P (2000).
 ""Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways"". J Mol Biol 303: 627-641.
- [5]. Cordes, M. H., Davidson, A. R., and Sauer, R. T (1996). ""Sequence space, folding and protein design"". Curr Opin Struct Biol 6: 3-10.
- [6]. Dubcovsky J, Lou MC, Zhong GY, et al (1996). "Genetic map of diploid wheat, Triticum monococcum L., and its comparison with maps of Hordeum vulgare L"
- [7]. Ellison, W. (1935). "A study of the chromosome numbers and morphology in certain British varieties of the common cultivated potato (solanum buberosum L.)" (PDF)
- [8]. Haeckel E. 1866. Generelle Morphologie der Organismen: Allgemeine Gründzuge der organischen Formen-Wissenschaft. 2 vols, Reimer, Berlin.
- [9]. Hegyi H, and Gerstein M. (1999). "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome". J Mol Biol 288 (1): 147-64.
- [10]. Kato A, Lamb JC, Birchler JA (September 2004). "Chromosome painting

using repetitive DNA sequences as probes for somatic chromosome identification in maize"

- [11]. Kenton A, Parokonny AS, Gleba YY, Bennett MD (August 1993).
 "Characterization of the Nicotiana tabacum L. genome by molecular cytogenetics". Mol. Gen. Genet. 240 (2): 159–69.
- [12]. Lesk AM, Brändén CI, Chothia C. (1989). "Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet". Proteins 5 (2): 139-48.
- [13]. Nägeli C. 1842. Zur Entwickelungsgeschichte des Pollens bei den Phanerogamen. Ovell & Füssli, Zürich.
- [14]. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar H, Moran N, Hattori M (2006). "The 160-kilobase genome of the bacterial endosymbiont Carsonella". Science 314 (5797): 267.
- [15]. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. (1997). "CATH--a hierarchic classification of protein domain structures". Structure 5 (8): 1093-108.
- [16]. Ostermeier M, Benkovic SJ. (2000). "Evolution of protein function by domain swapping". Adv Protein Chem 55: 29-77.
- [17]. Phillips DC. (1966). "The threedimensional structure of an enzyme molecule". Scientific American 215 (5): 78-90.
- [18]. Pradella S, Hans A, Spröer C, Reichenbach H, Gerth K, Beyer S (2002). "Characterisation, genome size and genetic manipulation of the myxobacterium Sorangium cellulosum So ce56". Arch Microbiol 178 (6): 484–92.
- [19]. Sandman K, Pereira SL, Reeve JN (1998). "Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome". Cell. Mol. Life Sci. 54 (12): 1350–64.

www.jatit.org

- [21]. Sandman K, Reeve JN (2000). "Structure and functional relationships of archaeal and eukaryal histones and nucleosomes". Arch. Microbiol. 173 (3): 165–9.
- [22]. Smith J, Burt DW (1998). "Parameters of the chicken genome (Gallus gallus)". *Anim. Genet.* 29 (4): 290–4.
- [23]. Wetlaufer DB. (1973). "Nucleation, rapid folding, and globular intrachain regions in proteins". *Proc Natl Acad Sci U S A* 70 (3): 697-701.