Title
The *TLA1* Protein Family Members Contain a Variant of the Plain MOV34/MPN Domain

Journal
American Journal of Biochemistry and Molecular Biology, 2(1)

ISSN
2150-4210

Authors
Mitra, Mautusi
Ng, Sam
Melis, Anastasios

Publication Date
2011-12-15

DOI
10.3923/ajbmb.2012.1.18

Peer reviewed

American Journal of Biochemistry and Molecular Biology 2 (1): 1-18, 2012
ISSN 2150-4210 / DOI: 10.3923/ajbmb.2012.1.18

The *TLA1* Protein Family Members Contain a Variant of the Plain MOV34/MPN Domain

[1,2]Mautusi Mitra, [1,3]Sam Ng and [1]Anastasios Melis

[1]Department of Plant and Microbial Biology, University of California, Berkeley CA 94720-3102, USA
[2]Department of Biology, University of West Georgia, 1601 Maple Drive, Carrollton, GA 30118, USA
[3]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA

Corresponding Author: Mautusi Mitra, Department of Biology, University of West Georgia, Carrollton, GA 30118, USA

ABSTRACT

TLA1 (Truncated Light harvesting Antenna 1) protein belongs to a novel uncharacterized protein family. It was first identified in the green microalga, *Chlamydomonas reinhardtii*. The *TLA1* like proteins are present in diverse groups of eukaryotes ranging from algae, higher plants, insects and mammals. Currently the protein domains present in the members of this unknown protein family lack a homology to any known well characterized functional domain. To date, the exact functional roles of *TLA1*-like proteins in organisms, when present, are unknown because of lack of existence of a well characterized functional domain. The aim of this study is to identify a functional domain in the *TLA1* like proteins through application of computational-based bioinformatics tools. BLAST-P helped to identify two specific domains within the conserved regions of *TLA1*-like proteins. A PHYRE hit, 2095B (the chain B, crystal structure of the metal-free dimeric human MPN domain), aligned well with the first conserved domain of the *TLA1* protein, both at the primary sequence and secondary structural level. A comparative model of the *Chlamydomonas TLA1* protein was built using the 2095B protein sequence as a template. Key putative hydrophobic and catalytic amino acid residues were identified, that are also present in certain MPN domain-containing proteins. This study has shown that *TLA1* like proteins have a plain MPN- domain. The MPN domain specific key amino acids that we have identified can be tested experimentally in the future to better define the functional role of *TLA1* like proteins.

Key words: *TLA1* protein, *Chlamydomonas*, MPN domain, 2095B protein, human NOC4 protein

INTRODUCTION

Two photosystems coordinate the light reaction of photosynthesis. Each photosystem is comprised of a reaction center and a chlorophyll antenna. Chlorophyll antenna is present in the thylakoid membranes of a chloroplast. It traps solar energy and passes it to the reaction center for photochemistry. In green and brown algae as well as in vascular plants, about 300 Chl (a and b) molecules can make up the photosystem-II (PSII) antenna, whereas the photosystem-I (PSI) antenna may contain approximately 250 Chl molecules (Melis, 2002, 2005). The number of Chl molecules in the antenna is not constant but can alter considerably in accordance with a change in the environmental, developmental, genetic, or physiological conditions (Mitra and Melis, 2008; Melis, 1991). Plants acclimatize to changing light intensity by regulating the expression of nuclear *LHCb*, *LHCa* (chlorophyll a-b light-harvesting complex) and *CAO*

(chlorophyllide *a* oxygenase) genes (Mitra and Melis, 2010; Jansson *et al*., 1992; Tanaka and Melis, 1997; Ohtsuka *et al*., 1997). Under low light condition, plants up-regulate *LHC* and *CAO* (chlorophyllide a oxygenase) gene expression and under high light condition plants turn down the LHC and CAO gene expression. This leads to an enlargement and reduction of the chlorophyll antenna sizes under low and high light conditions, respectively (Masuda *et al*., 2003; Maxwell *et al*., 1995; Huner *et al*., 1998; Wilson and Huner, 2000). Regulation of the Chl antenna size allows photosynthetic organisms to make efficient use of available solar energy in diverse ecosystems where there is fluctuation of sunlight (Polle *et al*., 2003).

Currently not much information is available about the novel genes that regulate the chlorophyll (Chl) antenna sizes of photosystems in photosynthetic higher plants and algae (Nakada *et al*., 1995; Escoubas *et al*., 1995; Yakovlev *et al*., 2002; Masura *et al*., 2002; Melis, 1991, 1996; Melis *et al*., 1999). Our laboratory identified a random DNA insertional mutant called *tla1* (truncated light-harvesting Chl antenna size1) (Tetali *et al*., 2007). *tla1* was found to be defective in a nuclear gene which was annotated later as the *TLA1* gene. *TLA1* gene codes for a protein that is 213 amino acids long and has homologs in different groups of eukaryotic organisms. All the *TLA1*-like proteins from diverse groups of organisms showed certain conserved amino acid sequences in their protein sequences but no distinct protein domain was identified (Mitra and Melis, 2010).

Although the expression of the *TLA1* gene has been shown to affect the regulation of the chlorophyll antenna size in the green alga *C. reinhardtii*, the precise functional role of the *TLA1* protein at the molecular level is not known. Likewise, the function of *TLA1*-like proteins in other organisms is not yet known. The main difficulty in the elucidation of the biochemical function of proteins belonging to a novel uncharacterized family of proteins is the apparent absence of any well conserved sequence motif in the amino acid sequence. Hence, it is necessary to establish a link between the conserved region of a *TLA1* like protein and a known or well-characterized functional domain in other protein family. This will help to narrow down the putative functional role of the uncharacterized protein and is crucial for designing experiments to test predictive hypotheses. In this study, an analysis of the structure and function of the *Chlamydomonas TLA1* protein has been presented based on the application of computational bioinformatics approaches. Results revealed a group of *TLA1* amino acids that apparently form a MOV34/MPN-like domain. This finding would permit design of experiments for testing the possible functional role of *TLA1* in protein turnover and regulation of gene expression, leading to activation of the signal transduction involved in the regulation of the Chl antenna size in *C. reinhardtii*. This approach will also help to elucidate the broader functional role of *TLA1*-like proteins in multiple eukaryotes.

MATERIALS AND METHODS

This study was started in the year April 2009 and was finally concluded in January, 2011.

Identification of *TLA1* like homologs and conserved domains in the *TLA1* protein: Webservers such as BLAST-P and PSI-BLAST (http://www.ncbi.nlm.nih.gov/), CDART [http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi], PHYRE (http://www.sbg.bio.ic.ac.uk/phyre/index.cgi) and PhyloFacts (http://phylogenomics.berkeley.edu/phylofacts/search.php) were used for searching protein databases for sequence based homology detection and conserved domain identification. Analyses of genes upstream and downstream of the *TLA1*-like gene locus on chromosomes in

Chlamydomonas (http://www.chlamy.org/chlamydb.html), *Arabidopsis* (TAIR) [http://www.arabidopsis.org/] and Human (http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/) were performed using the respective genomic databases.

Phylogenetic tree and multiple primary sequence alignment of *TLA1* homologs: Critical residues were predicted by conserved regions within the MSA built by the Unix-based MUSCLE (http://www.ebi.ac.uk/Tools/muscle/index.html) (Edgar, 2004) and viewed using GeneDoc (http://www.nrbsc.org/gfx/genedoc/) (Nicholas *et al*., 1997).

Secondary structure alignment of *TLA1* like homologs: PHYRE (Kelley and Sternberg, 2009) and Jpred (Cole *et al*., 2008) (http://www.compbio.dundee.ac.uk/www-jpred/) were used for secondary structure prediction and identification of a suitable PDB (Protein Data Bank) hit for comparative model building.

Building of comparative *TLA1* protein model and Prediction of Probable protein function: ESyPred3D (Lambert *et al*., 2002) (http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/) was used to build the comparative model. Comparative model viewing was enabled by the UCSF windows CHIMERA interface (Sali and Blundell, 1993) (http://www.cgl.ucsf.edu/chimera/download.html). ProtFun (Jensen *et al*., 2002) (http://www.cbs.dtu.dk/services/ProtFun/) was used to predict protein function from sequence of *TLA1*. Jmol (http://jmol.sourceforge.net/) was used to view structure of 2095B.

RESULTS

Domain analysis of the *TLA1* protein: CDART (Conserved Domain Architecture Retrieval Tool) analysis of the *TLA1* protein sequence [Swiss-Prot: Q8L439/trQ8L439_CHLRE], categorized the family of *TLA1* proteins as a novel uncharacterized protein family, UPF0172 (in the time period 2009-early 2011 when this work was done; see discussion section). A BLASTP search showed high homology of the *TLA1* protein with expressed protein sequences of the colorless microalga *Polytomella parva* [GenBank: ABH10987.1], higher plants (e.g., *Arabidopsis thaliana*; [TAIR: AT5G55940] and *Oryza sativa* -japonica cultivar-group [NCBI Reference Sequence: NP001052342.1], fungi (e.g., *Phytophthora infestans*, [GenBank: EEY62434.1]), insects (e.g., *Drosophila mojavensis*, [NCBI Reference Sequence: XP_002005585.1]), mammals (e.g., *Homo sapiens* [NCBI Reference Sequence: NP_057133.2], *Rattus norvegicus* -Norway rat-[NCBI Reference Sequence: NP_001012165.1], *Mus musculus* -house mouse-[NCBI Reference Sequence:NP_149158.1]), amphibians (e.g., *Xenopus laevis* [NCBI Reference Sequence: NP001089747.1), flatworms (e.g., *Schistosoma mansoni* [NCBI Reference Sequence: XP_002575710]) and fish (e.g., *Danio rerio*–zebrafish [NCBI Reference Sequence: NP956420.1).

Conserved domain and PFAM analysis indicated that *TLA1* belongs in the pfam03665, UPF0172, uncharacterized protein family (UPF0172). The distribution of BLAST-P hits was modified in order to take in 500 sequences. Some of the hits indicated the presence of two domains within the UPF0172 family. The presence of the first domain is further supported by PHYRE and PhyloFacts analysis, indicating a match with the MOV34 domain containing protein (Fig. 1). The presence of the second domain was not confirmed by PhyloFacts and PHYRE analysis and was not further considered in this work. Phylofact analysis showed a match to the MPN domain of the COP9 signalosome complex subunit-5 related protein (eukaryote) with an E-value of 0.286 and

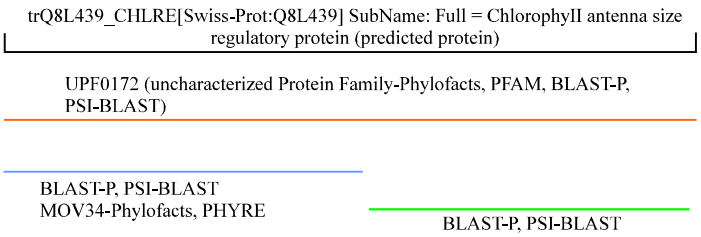

Fig. 1: Domain architecture of the *TLA1* protein using various webservers. tr:Q8L439_CHLRE/ [Swiss-Prot: Q8L439] (*TLA1* sequence from *Chlamydomonas reinhardtii*) is represented by a black line. Two domains that were identified by the BLAST-P and PSI-BLAST are shown as blue and green lines. The domain that was confirmed by PHYRE and PhyloFacts is shown as a blue line

PHYRE detected a match to chain B, crystal structure of the metal-free dimeric human MOV34/MPN domain in the PDB (Protein data bank; 2095B; Residues 1-186) with an E-value of 0.065. Though 2095B is not a highly significant PDB hit by itself, it is important, because HMM (Hidden Markov Model) could not be used to find a hit to PDB. 2095B and *TLA1* have a poor sequence alignment score but when taking into account that the secondary structure of 2095B aligns very well with the predicted secondary structure of *TLA1* (secondary structure alignment supported both by PHYRE and Jpred), the similarity becomes much more significant. Thus, the connection between *TLA1* and the PHYRE hit, 2095B, was subjected to further analysis involving Multiple Sequence Alignment (MSA) and secondary structure prediction.

Multiple sequence alignment of *TLA1* and 2095B homologs: The entire *TLA1* protein sequence was used as query sequence for both web and Unix-based analyses in order to obtain a wide and unifying view of sequence conservation of the *TLA1* protein across putative homologs. The BLAST-P interface, offered by NCBI, was first used to score the protein against NR (Non-Redundant) with default cut offs but taking into account 1000 sequences. BLAST-P analysis indicated that the *TLA1* protein is highly conserved among diverse groups of organisms ranging from green algae (*Polytomella parva*) and higher plants (*Oryza sativa, Arabidopsis thaliana*), to mammals (*Mus musculus, Homo sapiens*). PSI-BLAST was used to gather a comprehensive list of both easily detectable and remote homologs of both *TLA1* and 2095B individually, for Multiple Sequence Alignment (MSA) construction. Homologs for *TLA1* and 2095B were gathered manually using three iterations with 1000 sequences with a cut-off of 1e-4. 131 *TLA1* homolog sequences (including the seed sequence of *TLA1*) and 501 2095B (including the seed sequence of 2095B) homolog sequences with E-value more significant than 1e-4 were selected from the third iteration results for MSA analysis using MUSCLE. Twenty *TLA1* homologs were used for detailed MSA constructions (data not shown). Figure 2 shows the MSA of seven *TLA1* homologs (out of the twenty homologs) from diverse groups of organisms ranging from microalgae to mammals, with the *Chlamydomonas TLA1*. Several amino acids like histidine, aspartic acid, glutamic acid and asparagine, are highly conserved among the *TLA1* homologs. It is interesting to see a high conservation of hydrophobic and aliphatic amino acids (leucine, isoleucine, valine, proline, glycine, tyrosine, phenylalanine and tryptophan).

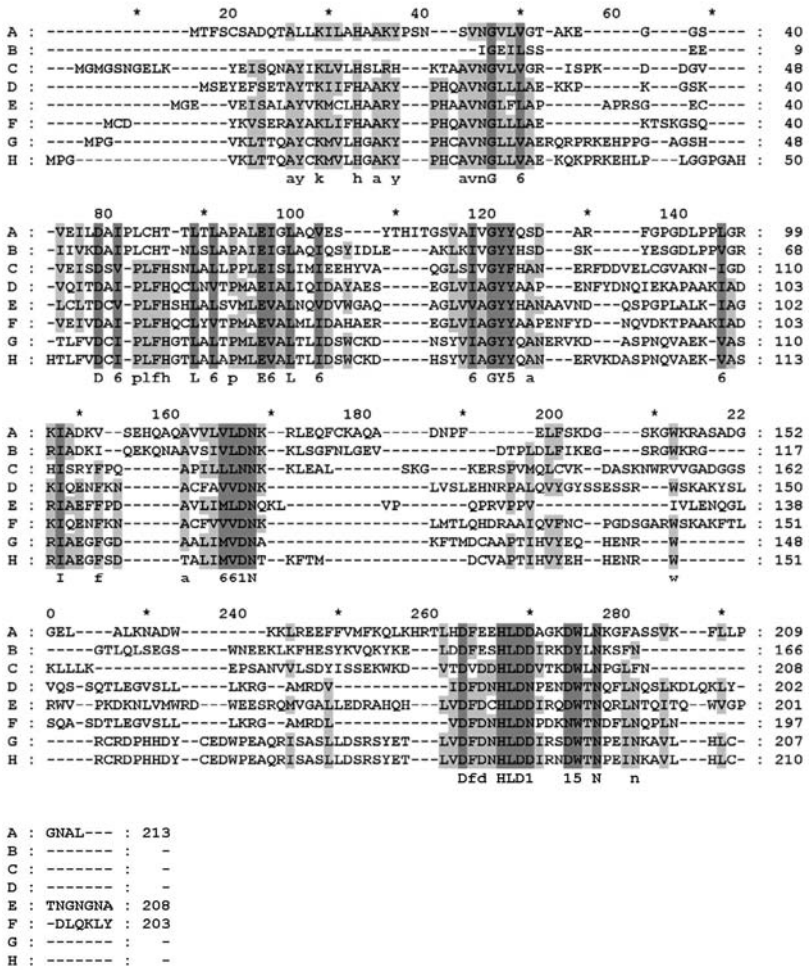

Fig. 2: Multiple sequence alignment of *TLA1* and its homologs. *TLA1* homologs were obtained by three iterations of PSI-BLAST. Seven homologs belonging to diverse group of organisms along with the seed sequence of *Chlamydomonas TLA1* were selected and aligned by MUSCLE. The asterisk is a marker that appears every 20 positions of the alignment so it goes *(10), 20, *(30), 40, *(50), 60, etc. A lowercase letter for the consensus means mostly conserved, upper case letter means completely conserved and the number is a code indicating that substitutions for similar amino acids are allowed (ie. 6 is aliphatic, 5 is aromatic and 1 is acidic, amino acids). Dark Gray is 100% conservation score from Blosum 62, Gray is 60% and uncolored is anything less than 60%. A [Swiss-Prot: Q8L439] is *TLA1* from *Chlamydomonas reinhardtii*; B [GenBank:ABH10987.1] is antenna size regulatory protein from alga *Polytomella parva*; C [NCBI reference sequence:NP_568832.1] is EMB2731 (EMBRYO DEFECTIVE 2731) protein from *Arabidopsis thaliana*; D [NCBI reference sequence:XP_002050890.1] is GJ19954 protein from *Drosophila virilis*; E [NCBI reference sequence:NP_057133.2] is hypothetical protein LOC51016 from *Homo sapiens*; F [NCBI reference sequence:611731.1] is CG3501 from *Drosophila melanogaster*; G [NCBI reference sequence:035056.1] is COX4 neighbor protein from *Mus musculus*; H [NCBI reference sequence:006058.1] is COX4 neighbor isoform 1 from *Homo sapiens*

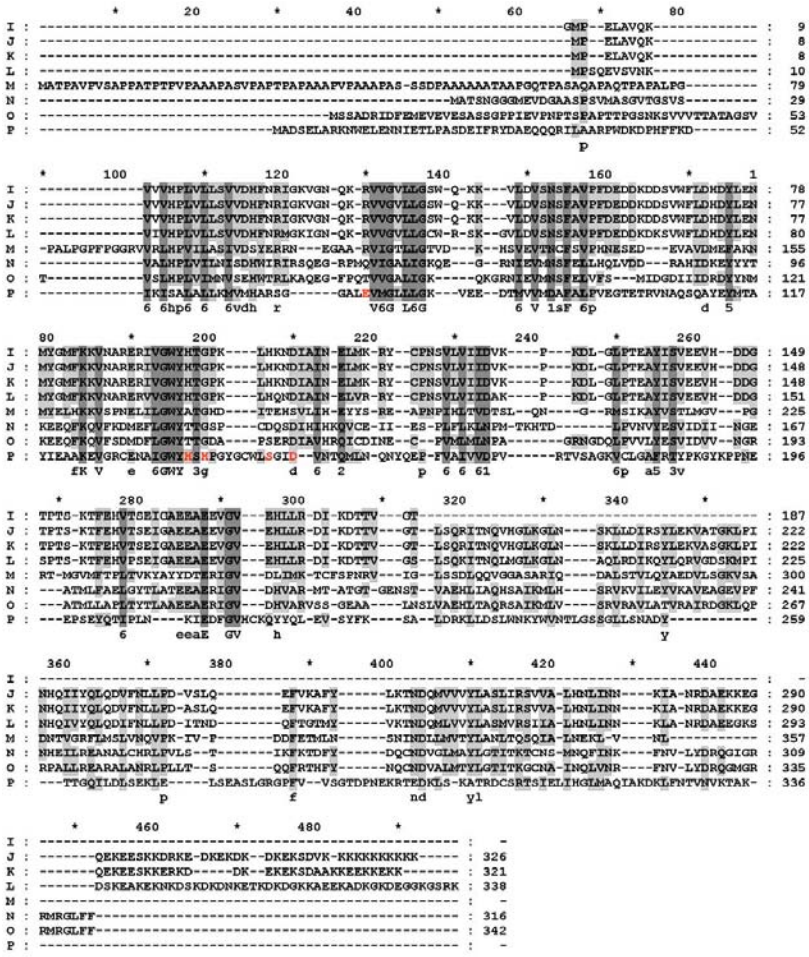

Fig. 3: Multiple sequence alignment of 2095B and its homologs. 2095B homologs were obtained by three iterations of PSI-BLAST. Seven homologs belonging to diverse group of organisms along with the 2095B seed sequence were selected and aligned by MUSCLE. The asterisk is a marker that appears every 20 positions of the alignment so it goes *(10), 20, *(30), 40, *(50), 60, etc. A lowercase letter for the consensus means mostly conserved, upper case letter means completely conserved and the number is a code indicating that substitutions for similar amino acids are allowed (ie. 6 is aliphatic, 5 is aromatic and 1 is acidic, amino acids). Dark Gray is 100% conservation score from Blosum 62, Gray is 60% and uncolored is anything less than 60%. I [PDB:2O95B] is Chain B, Crystal Structure Of The Metal-Free Dimeric Human Mov34 Mpn Domain (Residues 1-186); J [GenBank:AAH00338.1] is PSMD7 protein from *Homo sapiens*; K [NCBI reference sequence:NP_034947.1] is proteasome 26S subunit, non-ATPase, 7 from *Mus musculus*; L [NCBI reference sequence:NP_523845.2] is Mov34 protein from *Drosophila melanogaster*; M [NCBI reference sequence:NP_003745.1] is eukaryotic translation initiation factor 3 subunit 5 epsilon from *Homo sapiens*; N [NCBI reference sequence:NP_001134495.1] is COP9 signalosome complex subunit 6 from *Salmo salar*; O [NCBI reference sequence:NP_001040313.1] is transcription factor-like protein from *Bombyx mori*; P [NCBI reference sequence:XP_001649479.1] is jun activation domain binding protein from *Aedes aegypti*

Figure 3 shows the MSA of seven 2095B homologs (out of the twenty homologs) from different organisms ranging from insects, reptiles and mammals with the human 2095B. The latter contains a MOV34/MPN/JAB1/PAD1 domain which is a 120-180 hydrophobic amino acid rich region, usually present in the N-terminal end of a protein. Overall, amino acid residues in the MPN domain are not conserved but vary from homolog to homolog. This domain does not have a specific motif and represents a region in the protein, that in addition to being rich in hydrophobic amino acids, also contains conserved histidine and aspartic acid, glutamic acid, serine and glycine residues (Maytal-Kivity *et al*., 2002; Tran *et al*., 2003). The MPN domain-containing protein family can be subdivided into two groups, namely the plain MPN- and MPN+ group (Maytal-Kivity *et al*., 2002; Tran *et al*., 2003). MPN+ group has a subset motif that contains a well-defined pattern of H-x-H-x[7]-S-x[2]-D, where x[7] and x[2] indicate stretches of seven and two non-conserved residues, respectively. In addition to this conserved arrangement of four polar residues, there is an additional conserved glutamic acid residue in the N-terminal region of the domain. An aromatic residue, usually a tryptophan, is found two amino acid positions upstream of the conserved serine in many but not all MPN+ proteins. Thus, this serine residue should not be considered as part of the conserved MPN+ motif. A typical MPN+ motif is evident in the 2095B homolog sequence [NCBI Reference Sequence: gi|157106782|ref|XP_001649479.1], the c-jun activation domain binding protein from *Aedes aegypti*. The conserved amino acids in the MPN+ motif are highlighted red (Fig. 3, 4).

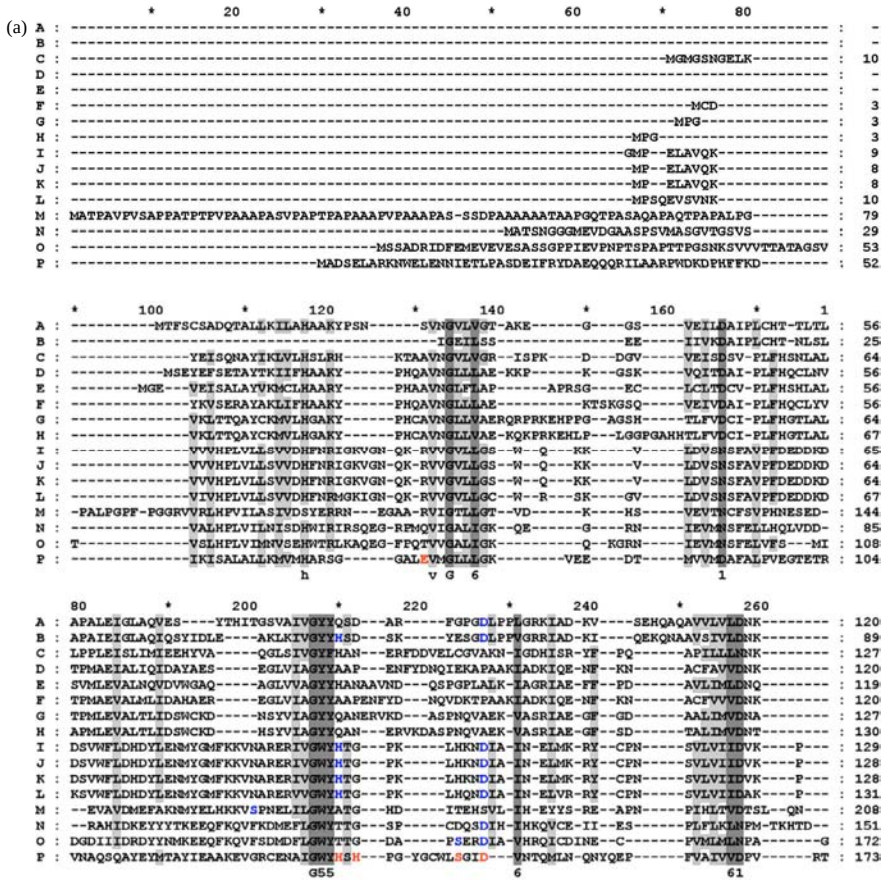

Fig. 4(a-b): Continued

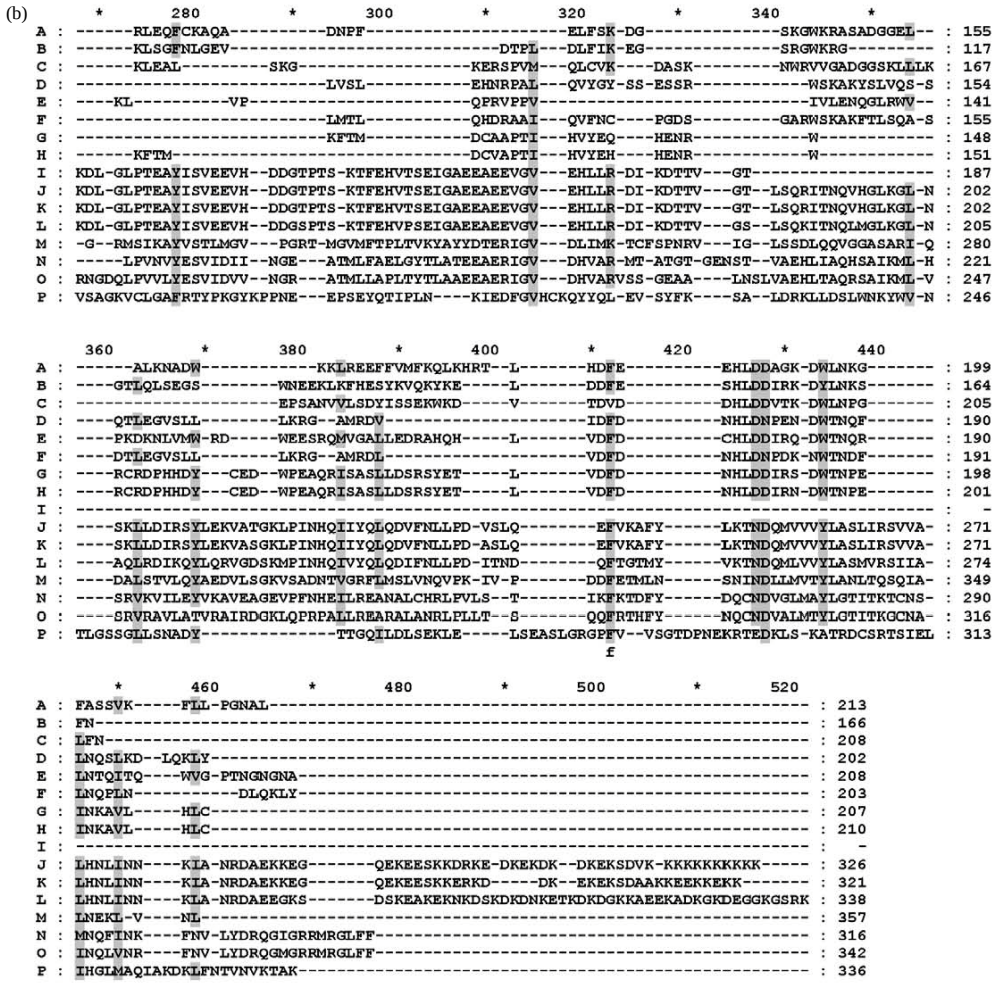

Fig. 4(a-b): Multiple sequence alignment of *TLA1* and 2O95B homologs. Seven *TLA1* (sequences B-H; see Fig 2 legend for accession number and protein information) and seven 2O95B homologs (sequences J-P; see Fig 3 legend for accession number and protein information) were aligned to the *TLA1* (sequence A) and 2O95B (sequence I) seed sequences by MUSCLE. The asterisk is a marker that appears every 20 positions of the alignment so it goes *(10), 20, *(30), 40, *(50), 60, etc. A lowercase letter for the consensus means mostly conserved, upper case letter means completely conserved and the number is a code indicating that substitutions for similar amino acids are allowed (ie. 6 is aliphatic, 5 is aromatic and 1 is acidic, amino acids). Dark Gray is 100% conservation score from Blosum 62, Gray is 60% and uncolored is anything less than 60%. The MPN+ motif is shown in red; the conserved amino acids reminiscent of the MPN+ motif, are highlighted in blue color

MSA of all the seven homologs of *TLA1* and 2095B, along with the individual seed sequences, revealed highly conserved histidine, valine, phenylalanine and aspartic acid residues in the protein domain (Fig. 4). These amino acids are known to play catalytic role in many enzymes. For example,

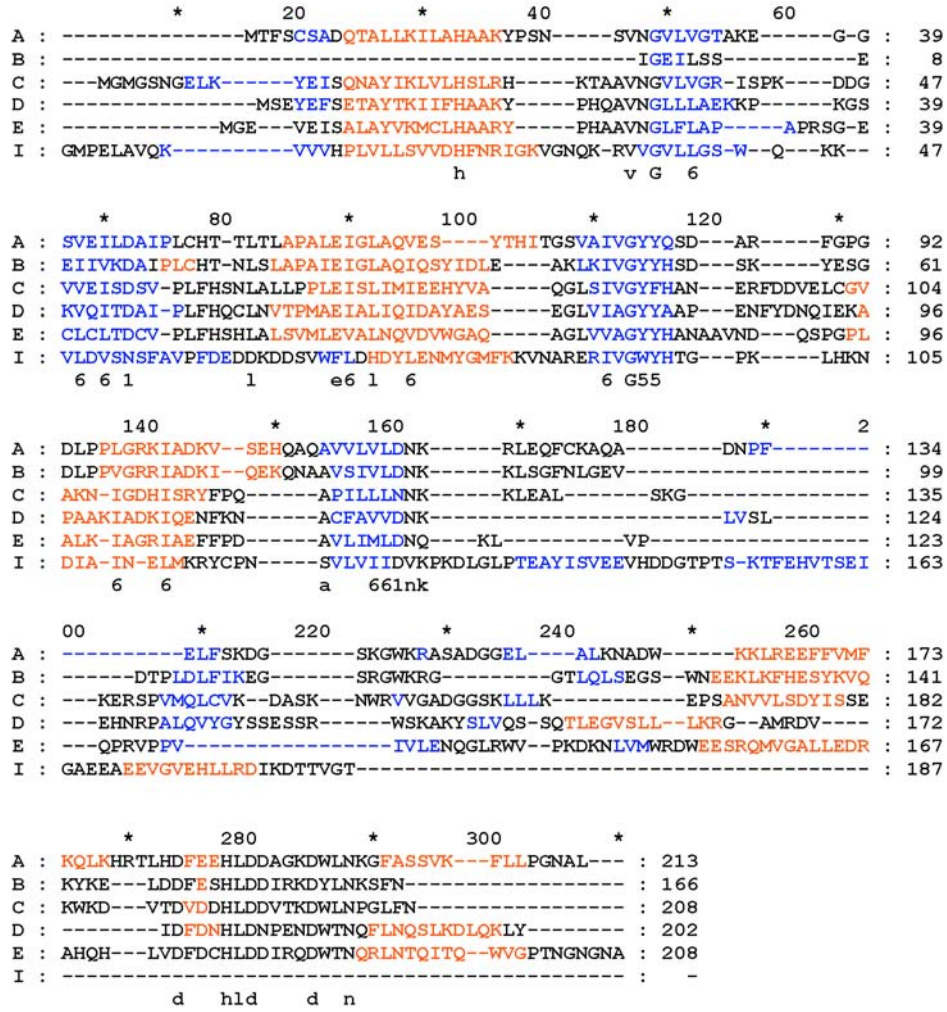

Fig. 5: Secondary structure alignment of *TLA1* and its homologs. Sequence I (2O95B) secondary structure was used to compare the secondary structure of *TLA1* (sequence A) and four of its homologs (B, C, D and E) belonging to diverse groups of organisms ranging from algae to mammals. B [Genbank:ABH10987.1] is antenna size regulatory protein from alga *Polytomella parva*; C [NCBI Reference sequence: NP_568832.1;TAIR: AT5G55940] is EMB2731 (EMBRYO DEFECTIVE 2731) protein from *Arabidopsis thaliana*; D [NCBI Reference sequence: XP_002050890.1] is GJ19954 protein from *Drosophila virilis*; E [NCBI Reference sequence:NP_057133.2] is hypothetical protein LOC51016 from *Homo sapiens*; Red amino acids denote amino acids involved in alpha helix formation and blue amino acids denote amino acids involved in beta strand formation

acidic amino acids like aspartic acid and basic amino acids like histidine are known to play a role in the activation of substrates of hydrolases (Varfolomeev *et al.*, 2005). Hydrophobic residues often play a role in keeping water out of the catalytic sites of enzymes. Amino acids like glycine and proline which are known to play a role in structural organization of catalytic sites in hydrolases, are conserved at few places in the *TLA1* and 2095B protein sequences (Fig. 4). Conserved glycine

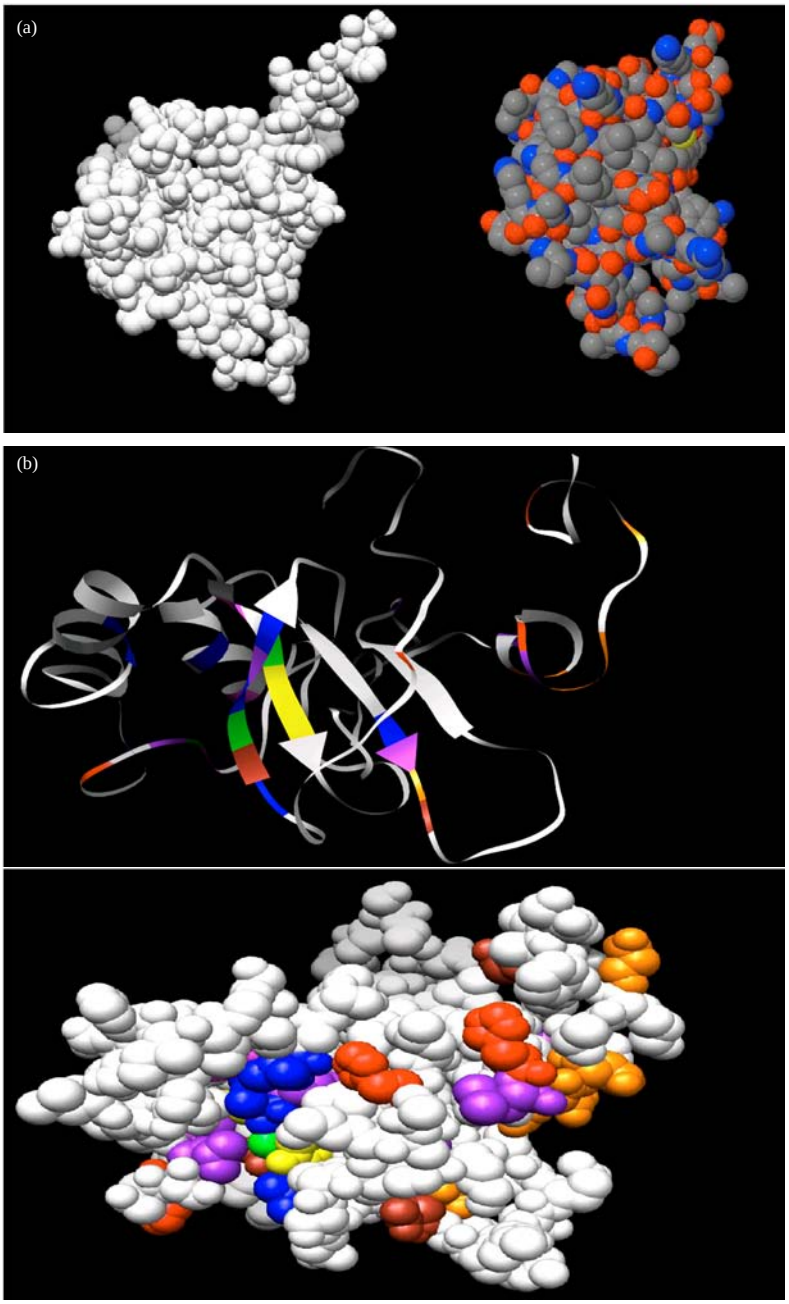

Fig. 6(a-b): Comparative protein structural model of *TLA1*. (a) Left: Comparative model built with EsyPred3D of *TLA1* using 2O95B template Right: JMOL structure of 2095B. (B) Top panel shows the ribbon model and the bottom panel shows the space-filling model of the *TLA1* protein superposed on 2095B model. The predicted critical residues from MSA in Figure 4 are shown in colors. Red: Histidine, Blue: Valine, Green: Glycine, Dark Green: Proline, Purple: Leucine, Magenta: Isoleucine, Orange: Aspartic Acid, Brown: Asparagine, Gray: Tryptophan and Yellow: Tyrosine

and the tryptophan/tyrosine amino acid residues in the MPN+ domain exists in a groove in the secondary structure in close proximity to the histidine residue and plays a role in metal coordination and substrate recognition (Varfolomeev *et al*., 2005; Sanches *et al*., 2007). These amino acids are conserved in the protein domain of *TLA1* homologs and the plain MPN domains of 2095B homologs (Bellare *et al*., 2006; Fig. 4). Some of the MPN+ reminiscent conserved amino acids like aspartic acid, serine and histidine, are denoted in blue font Fig. 4.

Secondary structure alignments: The 2095B secondary structure was compared against the predicted structure from Jpred for *TLA1* and four other *TLA1* homologs from diverse group of organism ranging from algae, to higher plants, insects and mammals (Fig. 5). As evident from Fig. 5, although there is no high primary sequence identity between the MPN domain region in 2095B homologs and the corresponding region in *TLA1* homologs, there exists a high degree of identity in secondary structure (alpha helices and the beta strands). Crystal structure analysis of 2095B has shown that it is composed of nine beta strands and four alpha helices (Sanches *et al*., 2007). Seven beta strands and four alpha helices are well aligned in the secondary structure alignment of 2095B and *TLA1* homologs shown in Fig. 5. This may suggest a similar tertiary conformation for the two sets of proteins.

Comparative model study: As primary sequence alignment and secondary structure analyses of 2095B and *TLA1* have shown promising indications of remote homology. A three dimensional model of *TLA1* was built individually with EsyPred3D using the template sequence of 2095B (Fig. 6a, left). Figure 6a (right) shows the Jmol structure of 2095B hit from PHYRE which is the chain B crystal structure of the metal-free dimeric human MOV34 MPN domain. Both of these models were superimposed on each other (Fig. 6b). Figure 6b (upper panel) shows the secondary structures as a ribbon model, whereas Fig. 6b (lower panel) shows the space-filling model. With lack of experimental information on catalytic residues, the MSA was used to predict residues that might play a role in catalysis (shown in color in Fig. 6b) using highly conserved positions. The locations of these amino acids within the 3D structure were visualized using the windows interface UCSF CHIMERA (Fig. 6b). Highlighting these residues on the comparative model also shows that these residues occur within a cleft which could be the active site of the protein.

Functional prediction: ProtFun analysis was also performed to obtain information about the functional role of the *TLA1* protein (Table 1). ProtFun server analysis of a protein sequence yielded a score consisting of two numbers. The first number is the estimated probability that the entry belongs to the class in question and is influenced by the prior probability of that class. The second number represents the odds that the sequence belongs to that class/category and is independent of the prior probability. ProtFun analysis showed that *TLA1* has the highest probability of being a "lyase" type (probability value of 0.069 and the odds value of 1.474) of an enzyme (probability value of 0.448 and the odds value of 1.563) that is involved in translation (probability value of 0.296 and the odds value of 6.729) which regulates cell growth (probability value of 0.027 and the odds value of 1.915). It also showed that the *TLA1* protein has the least probability of being a "hydrolase" (probability value of 0.077 and the odds value of 0.243) that can play a role in transport and binding (probability value of 0.017 and the odds value of 0.042) of metal ions (probability value of 0.009 and the odds value of 0.020).

Table 1: Prot Fun predictions of probable functional role of *TLA1*.

Function role	Probabilities	Odds
Functional category		
Amino acid biosynthesis	0.022	1.019
Biosynthesis of cofactors	0.110	1.530
Cell envelope	0.032	0.524
Cellular processes	0.077	1.048
Central intermediary metabolism	0.053	0.839
Energy metabolism	0.329	3.658
Fatty acid metabolism	0.056	4.322
Purines and pyrimidines	0.271	1.114
Regulatory functions	0.018	0.112
Replication and transcription	0.124	0.463
Translation	0.296	6.729
Transport and binding	0.017	0.042
Enzyme/nonenzyme		
Enzyme	0.448	1.563
Non enzyme	0.552	0.774
Enzyme class		
Oxidoreductase	0.150	0.722
Transferase	0.208	0.603
Hydrolase	0.077	0.243
Lyase	0.069	1.474
Isomerase	0.010	0.321
Ligase	0.026	0.512
Gene ontology category		
Signal transducer	0.062	0.288
Receptor	0.004	0.025
Hormone	0.001	0.206
Structural protein	0.001	0.044
Transporter	0.025	0.231
Ion channel	0.016	0.277
Voltage gated ion channel	0.004	0.178
Cation channel	0.011	0.239
Transcription	0.066	0.512
Transcriptional regulation	0.061	0.486
Stress response	0.027	0.310
Immune response	0.022	0.258
Growth factor	0.027	1.915
Metal ion transport	0.009	0.020

DISCUSSION

Protein database searches revealed that the *C. reinhardtii TLA1* protein has homologs in many eukaryotic organisms, including algae (Chlorophytes, diatoms and Prasinophytes), mosses, vascular plants, fungi, slime molds, fruit flies, aphids, flatworms, nematodes, apicomplexans, kinetoplastids, hydrozoans, amphibians, fishes and mammals. Todate, *TLA1* like proteins have not been detected in prokaryotes. Currently, the precise functional role of all members of the *TLA1*-like protein family is unknown. There are two homologs of the *TLA1*-like protein in the plant *Arabidopsis thaliana* [TAIR:AT5G55940 and AT5G51620] and in *Homo sapiens* (CGI-112 [FAM158A] and COX4NB [FAM158B]), and three homologs in *Danio rerio* (zebrafish). In *Arabidopsis*, one of the *TLA1*

homologs (TAIR: AT5G55940) has been localized in the Endoplasmic Reticulum (ER) (Dunkley *et al*., 2006). Localization of the other *Arabidopsis TLA1* homolog (TAIR:AT5G51620) is not known at present. One of the human *TLA1* homolog proteins (COX4NB) has been localized in the cytoplasm, mitochondria and nucleus (Bachman *et al*., 1999). Localization of the other human *TLA1* homolog, CGI-112 is not known. Thus, *TLA1* represents a highly conserved protein across virtually all eukaryotic but not prokaryotic systems, performing an as yet unknown function. There are no past results related to the exact functional role of *TLA1* like proteins because of an apparent lack of a functional domain. The results presented in this manuscript, is the first original report that provides some clue about the functional domain present in the *TLA1* like proteins.

Phylofact, PHYRE and JPRED analysis of *TLA1* protein sequence showed a match to the chain B crystal structure of the metal-free dimeric human MOV34/ plain MPN domain (Residues 1-186) (Fig. 1) (Alves *et al*., 2006). Our bioinformatics analysis was initiated in 2009 and got completed in early, 2011. During that time period, BLAST analysis classified *TLA1* as a member of an uncharacterized protein family. As protein databases get updated periodically, we repeated the protein BLAST analysis before submission of this manuscript few weeks back and we found that NCBI categorizes *TLA1* belonging to MPN_UPF0172 family (formerly called UPF0172). Hence our results have been confirmed by NCBI. We also detected that the *TLA1* protein from *C. reinhardtii* is currently showing a 74% identity with a hypothetical protein from *Volvox carteri f. nagariensis* (Gen Bank: EFJ45716.1). Our analysis was done at a time when the hypothetical *Volvox* protein sequence was not in the protein database and the highest match we obtained at that time was with the colorless alga *Polytomella parva* (50% identity) [GenBank: ABH10987.1](see Result section).

MPN (Mpr1, Pad1 N-terminal)/JAB1(jun activation binding domain protein 1)/PAD1/MOV34 domain is a widespread 120 amino acid protein module found in archaea, bacteria, phage tail assembly protein K from the bacteriophage lambda and its closely related homologs from other phages, prophages and eukaryotes. In eubacteria and archaea, the JAB1/MPN domain is usually found in proteins that do not form a macromolecular complex (Tran *et al*., 2003; Maytal-Kivity *et al*., 2002). In eukaryotes the JAB1/MPN domain is found in subunits of several multiprotein complexes of the proteasome, the COP9 signalosome (CSN), the subunits f and h (also known as subunits p47 and p40, respectively) of the eukaryotic translation initiation factor 3 (eIF3) and in regulators of transcription and translation (Hershey *et al*., 1996; Ponting *et al*., 1999). Within the MPN domain super family there are two main subclasses: the JAMM (JAB1[jun activation domain binding protein 1]/MOV34/MPN+ metalloenzyme domain and the plain MPN domain containing proteins. Eukaryotic proteins containing a MPN+ motif, exist as subunits of large complexes. They play a role in the selective hydrolysis of the isopeptide bond between the C-terminus diglycine attached to ubiquitin/ubiquitin like proteins and a lysine residue in the target protein that facilitate release of the ubiquitin/ubiquitin like proteins (Lam *et al*., 1997a, b; Eytan *et al*., 1993). The iso-peptidase activity of MPN+ proteins plays a major role in regulation of ubiquitin mediated physiological processes like light-mediated development in *Arabidopsis* and cell differentiation, proliferation and apoptosis in other eukaryotes (Naumann *et al*., 1999; Kouvaraki *et al*., 2003).

In contrast, the plain MPN domain lacks the well-conserved JAMM/MPN+ motif (Fig. 3 and 4), that coordinates metal binding, and lacks the isopeptidase activity (Sanches *et al*., 2007). Examples of some plain MPN domain-containing proteins existing as a subunit of macromolecular complexes are CSN6 (COP 9 complex homolog subunit 6; a subunit of COP9 signalosome), eIF3f, eIF3h and Rpn8/MOV34/PSMD7 (a non-ATPase subunit of the lid of the 26S proteasome regulatory particle

[19S]). *TLA1* protein shares some charged and hydrophobic amino acids with the plain MPN domain (see result section). Hence *TLA1* like proteins could well play a similar functional role like the proteins that contain a "plain MPN domain".

To get some idea about the broad roles of these plain MPN domain containing proteins, we did a comparative study of the functional roles of some *TLA1* like proteins with those of other plain MPN domain containing proteins. The *TLA1* mutant of *C. reinhardtii* is not a knockout mutant for the *TLA1* gene, but has reduced expression of the *TLA1* protein because of reduced translation of the *tla1* gene. The *tla1* mutant possesses a smaller than wild type Chl antenna size for both photosystems, with the PSII antenna size reduced by 50% and PSI antenna size reduced by 35%. It also showed lower levels of light-harvesting proteins and of Chl *b* relative to the wild type. In the *tla1* mutant of the green alga *Chlamydomonas*, the compact stacking of the thylakoid membrane is lost and chloroplast development has been affected due to reduced expression of the *TLA1* gene (unpublished data). Homozygous mutant (EMBRYO DEFECTIVE 2731) of *Arabidopsis thaliana TLA1* gene (AT5G55940) leads to defective embryo development in the seeds (Meinke, 1996). In *Arabidopsis thaliana,* it has been shown that complete depletion of CSN6B, the two CSN6 paralogs with plain MPN domain, result in loss of stability of CSN and degradation of various CSN components, along with the complete loss of CUL1, CUL3 and CUL4 derubylation (Gusmaroli *et al*., 2007). The *MOV34* gene in mammals encodes a non-ATPase subunit of the 19S regulator (Rpn8/PSMD7) with a plain MPN domain. The *mov34* mutation in mouse is a recessive embryonic lethal mutation that is due to a retroviral integration in the murine germline (Gridley *et al*., 1990). Mutation in *PAD1* gene, a MOV34 homolog in yeast results in temperature sensitive growth and inability to degrade ubiquitin conjugates (Penney *et al*., 1998). eIF3f, eIF3h are translation factors with plain MPN domain. All these data indicate that the MPN+ residues have catalytic functions and the plain MPN subunits play a redundant structural role in complexes where they exist. Our ProtFun analysis (see result section) also seems to indicate that *TLA1* like proteins play a role in cell growth via translation regulation. Hence these plain MPN proteins play an important role in cell/organ/organelle development in eukaryotes by an yet unidentified mechanism.

It is interesting to note that some of *TLA1* like proteins with plain MPN domains retain some of the conserved amino acids of the MPN+ domain (e.g., hisitidine/ serine/ aspartic acids; these amino acids are colored in blue font in Fig. 4) and can be termed as a variant of MPN+ domain. An example of such an MPN+ variant protein is found in the PRP8 U5snRNP protein, present in the major splicesome, which processes pre-mRNA by splicing out introns from them (Bellare *et al*., 2006). Although Prp8 does not possess the typical MPN+ motif and the isopeptidase activity associated with MPN+ motif, a Prp8p fragment containing the Jab1/MPN domain binds directly to ubiquitin with an affinity that is similar to that of other known UBDs (Bellare *et al*., 2006). Several mutations within this domain that hamper efficient splicing also reduce interaction of the Prp8p fragment with ubiquitin-Sepharose column (Bellare *et al*., 2006). These results along with fact that *TLA1* like genes are playing a role in embryo development in *A. thaliana* (Meinke, 1996) or chloroplast development in *C. reinhardtii*, raises the question if *TLA1*-like proteins regulate protein turnover in eukaryotic cells via ubiquitin or SUMO binding. Ubiquitin/SUMO binding in turn might control the organ/organelle development.

We also performed a detailed analysis of the neighbor protein coding genes present on the chromosome that contains the *TLA1* locus in *C. reinhardtii*, *A. thaliana* and *H. sapiens*. This analysis was done to get an idea about the functional roles of these genes as sometimes genes that

play a role in a particular metabolic pathway or have same functions are located close to each other on the same chromosome. In such situations, neighboring gene functions can give a clue about the probable function of gene whose function is unknown. The *TLA1* gene is located on chromosome 5 of *C. reinhardtii*. Interestingly, proteins coded by genes located in the immediate vicinity of *TLA1* locus in *C. reinhardtii* include superoxide dismutase, (copper/zinc binding proteins), four Peptidase M (neutral zinc metallopeptidases) with zinc finger-RING type; Zinc finger C2H2-type, Ubiquitin-conjugating enzyme E2, Peptidase C2, (cysteine proteases; calpain family) and RDP1, a RING-like domain protein (Mitra and Melis, 2010). The *Arabidopsis thaliana TLA1* homolog (AT5G55940) also has near-neighbor genes encoding for C3HC4-type RING finger proteins, two WD-40 repeat family proteins (they bind nucleic acid and are located in the CUL4 RING ubiquitin ligase complex), ubiqutin conjugating enzyme 30, and SUMO1 (small ubiquitin-like modifier 1). There are two homologs of *TLA1* like protein in *Homo sapiens* (CGI-112 [Fam158A] and COX4NB [FAM158B]). COX4NB is located on the 5' upstream region of COX4, the gene encoding for the cytochrome *c* oxidase subunit IV, on Chr16q24. Human COX4NB protein shares 44% amino acid identity with CGI-112 (Comparative Gene Identification isolate 112) which is located on (14q11.2). Genes located near the CGI-112 locus encode for PSMB11 [proteasome (prosome, macropain) subunit, beta type, 11], a WD repeat-containing protein that interacts with the COP9 signalosome referred to as DDB1 and CUL4 associated factor 11 and a proteasome (prosome, macropain) activator subunit 2 (PA28 beta). Genes near COX4NB locus on chromosome 16 include the human MOV34 homolog (16q23-q24), the zinc finger DHHC-type containing protein 7, the zinc finger CCHC domain containing protein 14, and IRF8 (Interferon regulatory factor 8). In summary all of these genes near the *TLA1* locus in *C. reinhardtii, A. thaliana and H. sapiens* code for proteins that function as some kind of proteases or metal binding proteins or translational/transcriptional regulators. These observations give support to our ProtFun analysis (Table 1; see result section) which indicates that *TLA1* like proteins might play a role in translation regulation.

CONCLUSIONS

Synthesis of information collected from a variety of sources and related bioinformatic analysis presented in this study show that *TLA1* has a variant of plain MOV34 domain. Hence the *TLA1* protein family is no longer an uncharacterized protein family and can be classified as part of MPN super family. The MPN domain in *TLA1* is unique in the fact that although it shares conserved amino acids mainly with the plain MPN domain, it still shares few amino acids with the MPN+ domain. It would be interesting to see if other non-canonical Jab1/MPN domain containing proteins including *TLA1*, can serve in ubiquitin/SUMO-binding functions. Recombinant *TLA1* protein can be over-expressed in *E. coli*/yeast and purified. Biochemical experiments can then be performed to test if the recombinant *TLA1* protein can bind ubiquitin /SUMO. If *TLA1* protein does bind to ubiquitin/SUMO, site directed mutagenesis of some of the conserved amino acids in the MPN domain can be performed to determine the function of the various amino acid residues, identified in this study. Our study has opened up a path for performing "wet lab" experiments to study the biochemical function of the *TLA1* like proteins. The three dimensional model generated in our study can be now used for functional analysis until crystal structures of the *TLA1*/*TLA1* like protein becomes available.

ACKNOWLEDGMENTS

The work was supported by the US Department of Energy, Hydrogen Program, grant # DE-FG36-05GO15041.

REFERENCES

Alves, B., S. Oyama-Jr, F.C. Gozzo, M. Sanches, B.G. Guimaraes and N. Zanchin, 2006. Characterization of the human ortholog of Mov34 reveals eight N-terminal residues important for MPN domain stability. Biochem. Biophys. Res. Commun., 347: 608-615.

Bachman, N.J., W. Wu, T.R. Schmidt, L.I. Grossman and M.I. Lomax, 1999. The 5' region of the COX4 gene contains a novel overlapping gene NOC4. Mammalian Genome, 10: 506-512.

Bellare, P., K. Alan, A.K. Kutach, A.K. Rines, C. Guthrie and E.J. Sontheimer, 2006. Ubiquitin binding by a variant Jab1/MPN domain in the essential pre-mRNA splicing factor Prp8p. RNA, 12: 292-302.

Cole, C., J.D. Barber and G.J. Barton, 2008. The Jpred 3 secondary structure prediction server. Nucleic Acids Res., 36: W197-W201.

Dunkley T.P.J., S. Hester, I.P. Shadforth, J. Runions and T. Weimar *et al.*, 2006. Mapping the *Arabidopsis* organelle proteome. PNAS, 103: 6518-6523.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res., 32: 1792-1797.

Escoubas, J.M., M. Lomas, J. LaRoche and P.G. Falkowski, 1995. Light intensity regulation of cab gene transcription is signaled by the redox state of the plastoquinone pool. Proc. Nat. Acad. Sci., 92: 10237-10241.

Eytan, E., T. Armon, H. Heller, S. Beck and A. Hershko, 1993. Ubiquitin C-terminal hydrolase activity associated with the 26 S protease complex. J. Biol. Chem., 268: 4668-4674.

Gridley, T., A. Douglas, D.A. Gray, T. Orr-Weaver and P. Soriano *et al.*, 1990. Molecular analysis of the *Mov* 34 mutation: transcript disrupted by proviral integration in mice is conserved in *Drosophila*. Development, 109: 235-242.

Gusmaroli, G., P. Figueroa, G. Serino and X.W. Deng, 2007. Role of the MPN subunits in COP9 signalosome assembly and activity and their regulatory interaction with arabidopsis Cullin3-Based E3 ligases. Plant Cell, 19: 564-581.

Hershey, J.W., K. Asan, T. Naranda, H.P. Vornlocher, P. Hanachi and W.C. Merrick, 1996. Conservation and diversity in the structure of translation initiation factor EIF3 from humans and yeast. Biochimie, 78: 903-907.

Huner, N.P.A., G. Oquist and F. Sarhan, 1998. Energy balance and acclimation to light and cold. Trends Plant Sci., 3: 224-230.

Jansson, S., E. Pichersky, R. Bassi, B.R. Green and M. Ikeuchi *et al.*, 1992. A nomenclature for the genes encoding the chlorophyll a/b-binding proteins of h igher plants. Plant Mol. Biol. Rep., 10: 242-253.

Jensen, L.J., R. Gupta, N. Blom, D. Devos and J. Tamames *et al.*, 2002. Prediction of human orphan protein function from post-translational modifications and localization features. J. Mol. Biol., 319: 1257-1265.

Kelley, L.A. and M.J.E. Sternberg, 2009. Protein structure prediction on the web: A case study using the Phyre server. Nat. Protoc., 4: 363-371.

Kouvaraki, M.A., G.Z. Rassidakis, L.L. Tian, R. Kumar, C. Kittas and F.X. Claret, 2003. Jun activation domain-binding protein 1 expression in breast cancer inversely correlates with the cell cycle inhibitor p27Kip1. Cancer Res., 63: 2977-2981.

Lam, Y.A., W. Xu, G.N. DeMartino and R.E. Cohen, 1997a. Editing of ubiquitin conjugates by an isopeptidase in the 26S proteasome. Nature, 385: 737-740.

Lam, Y.A., G.N. DeMartino, C.M. Pickart and R.E. Cohen, 1997b. Specificity of the ubiquitin isopeptidase in the PA700 regulatory complex of the 26S proteasome. J. Biol. Chem., 272: 28438-28446.

Lambert, C., N. Leonard, X. De Bolle and E. Depiereux, 2002. ESyPred3D: Prediction of proteins 3D structures. Bioinformatics, 18: 1250-1256.

Masuda T., A. Tanaka and A. Melis, 2003. Chlorophyll antenna size adjustments by irradiance in *Dunaliella salina* involve coordinate regulation of chlorophyll a oxygenase (*CAO*) and *Lhcb* gene expression. Plant Mol. Biol., 51: 757-771.

Masura, T., J.E.W. Polle and A. Melis, 2002. Biosynthesis and distribution of chlorophyll among the photosystem during recovery of the green algae (*Dunaliella salina*) from irradiance stress. Plant Physiol., 128: 603-614.

Maxwell D.P., S. Falk and N.P.A. Huner, 1995. Photosystem II excitation pressure and development of resistance to photoinhibition. 1. Light harvesting complex II abundance and zeaxanthin content in *Chlorella vulgaris*. Plant Physiol., 107: 687-694.

Maytal-Kivity, V., N. Reis, K. Hofmann and M.H. Glickman, 2002. MPN † a putative catalytic motif found in a subset of MPN domain proteins from eukaryotes and prokaryotes, is critical for Rpn11 function. BMC Biochem., 3: 28-28.

Meinke, D.W., 1996. Molecular genetics of plant embryogenesis. Annu. Rev. Plant Physiol. Plant Mol. Biol., 46: 369-394.

Melis, A., 1991. Dynamics of photosynthetic membrane composition and function. Biochim. Biophys. Acta, 1058: 87-106.

Melis A., 1996. Excitation Energy Transfer: Functional and Dynamic Aspects of *Lhc* (cab) Proteins. In: Oxygenic Photosynthesis: The Light Reactions, Ort, D.R. and C.F. Yocum Dordrecht, (Eds.). Kluwer Academic Publishers, Netherlands, pp: 523-538.

Melis, A., J. Neidhardt and J.R. Benemann, 1999. *Dunaliella salina* (Chlorophyta) with small chlorophyll antenna sizes exhibit higher photosynthetic productivities and photon use efficiencies than normally pigmented cells. J. Applied Phycol., 10: 515-552.

Melis, A., 2002. Green alga hydrogen production: Progress, challenges and prospects. Int. J. Hydrogen Energy, 27: 1217-1228.

Melis, A., 2005. Bioengineering of Green Algae to Enhance Photosynthesis and Hydrogen Production. In: Artificial Photosynthesis: From Basic Biology to Industrial Application, Collins, A.F. and C. Critchley (Eds.). Wiley-Verlag and Co., Berlin, Germany, pp: 229-240.

Mitra, M. and A. Melis, 2008. Optical properties of microalgae for enhanced biofuels production. Optics Exp., 16: 21807-21820.

Mitra, M. and A. Melis, 2010. Genetic and biochemical analysis of the *TLA1* gene in *Chlamydomonas reinhardtii*. Planta, 231: 729-740.

Nakada, E., Y. Asada, T. Arai and J. Miyake, 1995. Light penetration into cell suspensions of photosynthetic bacteria and relation to hydrogen production. J. Ferment. Bioeng., 80: 53-57.

Naumann, M., D. Bech-Otschir, X. Huang, K. Ferrell and W. Dubiel, 1999. COP9 signalosome-directed c-Jun activation/stabilization is independent of JNK. J. Biol. Chem., 274: 35297-35300.

Nicholas, K.B., H.B. Nicholas Jr. and D.W. Deerfield, 1997. GeneDoc: Analysis and visualization of genetic variation. EMBNEW News, 4: 14-14.

Ohtsuka, T., H. Ito and A. Tanaka, 1997. Conversion of chlorophyll b to chlorophyll a and the assembly of chlorophyll with apoproteins by isolated chloroplasts. Plant Physiol., 113: 137-147.

Penney, M., C. Wilkinson, M. Wallace, J.P. Javerzat and K. Ferrell *et al.*, 1998. The pad11 gene encodes a subunit of the 26S proteasome in fission yeast. J. Biol. Chem., 273: 23938-23945.

Polle, J.E.W., S.D. Kanakagiri and A. Melis, 2003. *TLA1* a DNA insertional transformant of the green alga *Chlamydomonas reinhardtii* with a truncated light-harvesting chlorophyll antenna size. Planta, 217: 49-59.

Ponting, C.P., L. Aravind, J. Schultz J, P. Bork and E.V. Koonin, 1999. Eukaryotic signaling domain homologs in archaea and bacteria. Ancient ancestry and horizontal gene transfer. J. Mol. Biol., 289: 729-745.

Sali, A. and T.L. Blundell, 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol., 234: 779-815.

Sanches, M., B.S. Alves, N.I. Zanchin and B.G. Guimaraes, 2007. The crystal structure of the human Mov34 MPN domain reveals a metal-free dimer. J. Mol. Biol., 370: 846-855.

Tanaka, A. and A. Melis, 1997. Irradiance-dependent changes in the size and composition of the chlorophyll a-b light-harvesting complex in the green alga *Dunaliella salina*. Plant Cell Physiol., 38: 17-24.

Tetali, S., M. Mitra and A. Melis, 2007. Development of the light-harvesting chlorophyll antenna in the green alga *Chlamydomonas reinhardtii* is regulated by the novel *TLA1* gene. Planta, 225: 813-829.

Tran, H.J., M.D. Allen, J. Lowe and M. Bycroft, 2003. Structure of the Jab1/MPN domain and its implications for proteasome function. Biochemistry, 42: 11460-11465.

Varfolomeev, S.D., I.A. Gariev and I.V. Uporov, 2005. Catalytic sites of hydrolases: Structures and catalytic cycles. Russian Chem. Rev., 74: 61-76.

Wilson, K.E. and N.P.A. Huner, 2000. The role of growth rate, redox-state of the plastoquinone pool and the trans-thylakoid ÄpH in photoacclimation of *Chlorella vulgaris* to growth irradiance and temperature. Planta, 212: 93-102.

Yakovlev, A.G., A.S. Taisova and Z.G. Fetisova, 2002. Light control over the size of an antenna unit building block as an efficient strategy for light harvesting in photosynthesis. FEBS Lett., 512: 129-132.