Structure and functional annotation of hypothetical proteins having putative Rubisco activase function from *Vitis vinifera*

Suresh Kumar

Centre for Bioinformatics Research, Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia; Suresh Kumar – Email: sureshkumar@ukm.edu.my

Abstract:
Rubisco is a very large, complex and one of the most abundant proteins in the world and comprises up to 50% of all soluble protein in plants. The activity of Rubisco, the enzyme that catalyzes CO_2 assimilation in photosynthesis, is regulated by Rubisco activase (Rca). In the present study, we searched for hypothetical protein of *Vitis vinifera* which has putative Rubisco activase function. The Arabidopsis and tobacco Rubisco activase protein sequences were used as seed sequences to search against *Vitis vinifera* in UniprotKB database. The selected hypothetical proteins of *Vitis vinifera* were subjected to sequence, structural and functional annotation. Subcellular localization predictions suggested it to be cytoplasmic protein. Homology modelling was used to define the three-dimensional (3D) structure of selected hypothetical proteins of *Vitis vinifera*. Template search revealed that all the hypothetical proteins share more than 80% sequence identity with structure of green-type Rubisco activase from tobacco, indicating proteins are evolutionary conserved. The homology modelling was generated using SWISS-MODEL. Several quality assessment and validation parameters computed indicated that homology models are reliable. Further, functional annotation through PFAM, CATH, SUPERFAMILY, CDART suggested that selected hypothetical proteins of *Vitis vinifera* contain ATPase family associated with various cellular activities (AAA) and belong to the AAA+ super family of ring-shaped P-loop containing nucleoside triphosphate hydrolases. This study will lead to research in the optimization of the functionality of Rubisco which has large implication in the improvement of plant productivity and resource use efficiency.

Keywords: Rubisco activase, Rubisco, *Vitis vinifera*, hypothetical protein, homology modelling, functional annotation.

Background:
Ribulose-1,5-biphosphate carboxylase/oxygenase, most commonly known by the shorter name Rubisco is a very large and complex protein and it is one of the most abundant proteins in the world comprising up to 50% of all soluble protein in plants. It is a bi-functional enzyme (EC 4.1.1.39) that is used in the Calvin cycle to catalyze the first major step of carbon fixation, a process by which the atmospheric carbon dioxide (CO_2) is made available to organisms in the form of energy-rich molecules. Besides its carboxylase activity, Rubisco also acts as an oxygenase in a reaction involving competition between O_2 and CO_2 for reaction with RUBp (ribulose-1, 5-bisphosphate). Despite its central role for plant metabolism, the

enzyme suffers from several shortcomings that necessitate its abundance in nature. Compared to other metabolic enzymes the catalytic rate of plant Rubisco is rather slow [1].

Maturation is an important event in the process of leaf development and is closely related to photosynthesis efficiency, which is regulated by various proteins. Rubisco activase is one of the key enzyme which plays a vital role to serve as an activator and regulator of Rubisco. It helps to convert Rubisco from its inactive form to active form. In most plants, two forms of Rubisco activase (short isoform, RCAs; long isofrom RCA1) are present, and they differ only at the C-terminus [2]. Further, it consists of two types of protein subunit, called the large chain

and small chain. It is now recognized to be a member of the AAA(+) family, whose members participate in macromolecular complexes that perform diverse chaperone-like functions.

Many environmental changes like rising global temperatures, changes in water availability and drastic variable weather events will adversely impact the inactivity of Rubsico activase that could affect plant productivity. Optimizing the

functionality of Rubisco has large implications in the improvement of plant productivity and resource use efficiency. As Rubisco activase optimizes photosynthesis in plants, its genetic engineering of in crops is of continuing interest. In this study, hypothetical proteins in *Vitis vinifera* which have putative Rubisco activase have been predicted using various structural and functional annotation tools.
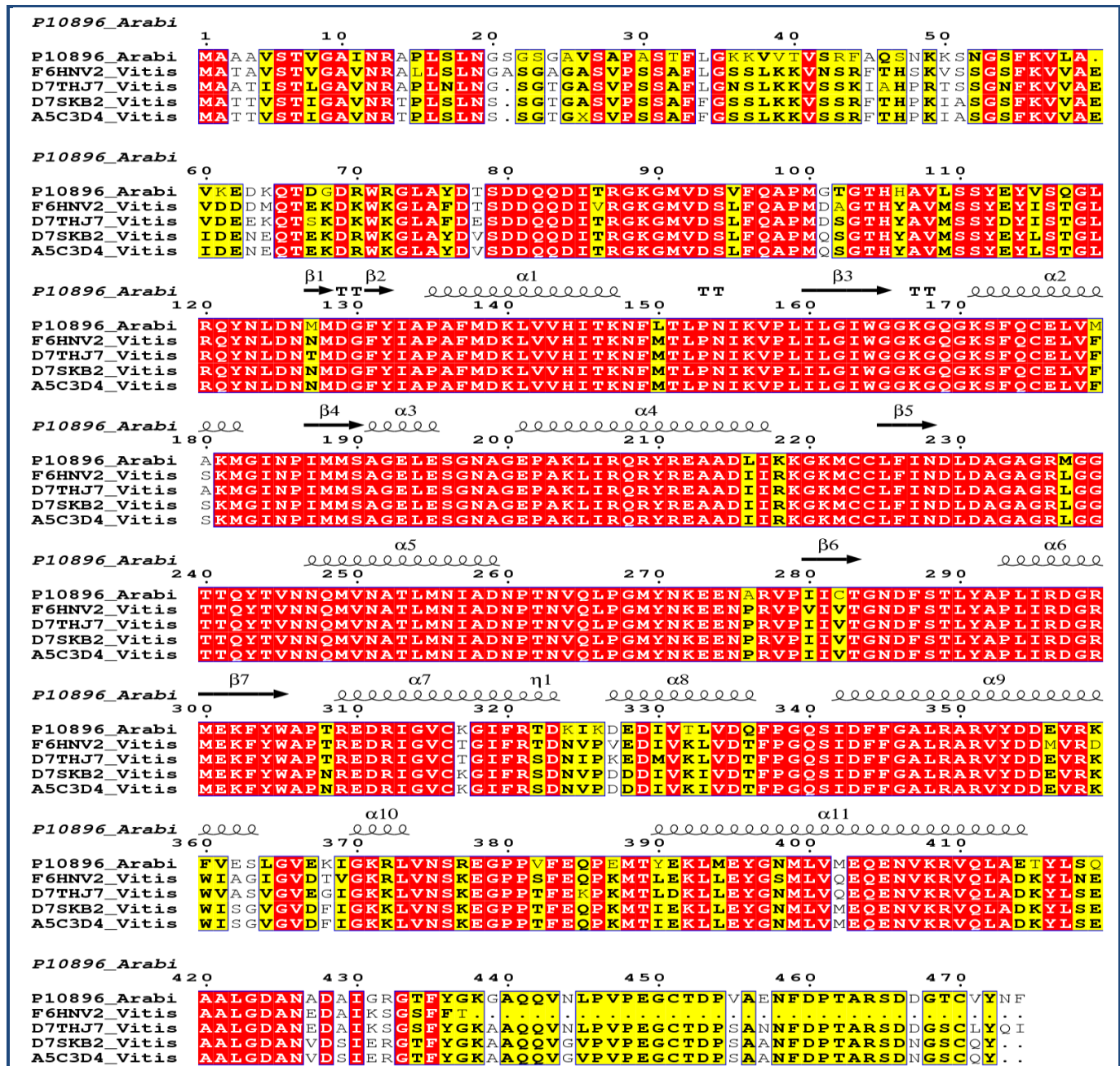
Figure 1: Multiple sequence alignment of Rubisco activase of Arabidopsis (Uniprot ID: P10896) with selected hypothetical proteins of *V. vinifera* (Uniprot ID: F6HNV2, D7THJ7, D7SKB2 and A5C3D4). Highly conserved residues are coloured in red and most conserved residues are coloured in yellow. Secondary structure predicted based on the template (PDB ID: 3T15) by using ESPript 3.0.

Methodology:
Sequence retrieval
The Arabidopsis and tobacco Rubsico activase (Uniprot ID: P10896 and Q40460) long isoform protein sequence was used as

seed sequence to do blast search against *Vitis vinfiera* in UniprotKB database. Among the hits for hypothetical proteins only four hypothetical were present in *V.vinifera* in UniprotKB (accessed December, 2014). The selected hypothetical protein

12

sequences (Uniprot ID: D7THJ7, D7SKB2, A5C3D4 and F6HNV2) were retrieved in FASTA format for further characterization.
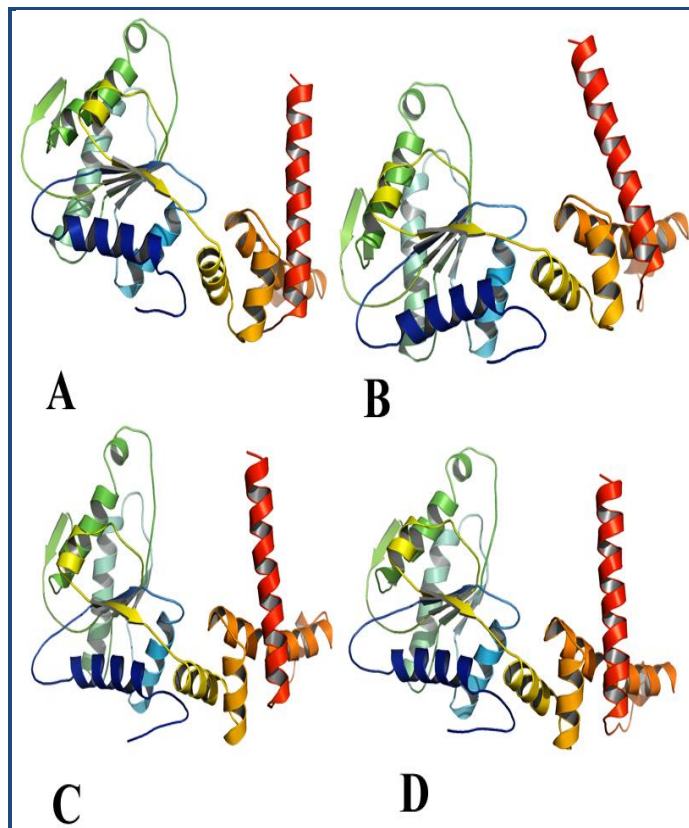
Figure 2: Homology model of selected hypothetical proteins of *V.vinifera* based on the template of green-type Rubisco activase from tobacco (PDB ID: 3T15) by using SWISS-MODEL server.

Physicochemical characterization
The selected hypothetical proteins were subjected to physicochemical characterization using Expasy protparam tool **[3]** for theoretical measurements such as molecular weight, isoelectric point, extinction coefficient, instability index, aliphatic index and grand average of hydropathicity (GRAVY).

Sub-cellular localization
The selected hypothetical proteins of *V. vinifera* were subjected to sub-cellular prediction tools like WoLF PSORT and CELLO **[4]**. For predicting signal peptide, SignalP 4.1 was used and SecretomeP was used for identifying protein involvement in non-classical pathway. HMMTOP and TMHMM were used to predict the propensity of a protein to be membrane protein.

Sequence Comparison
The hypothetical proteins were predicted using similarity search in BLASTp **[5]** against Non-redundant (NR) database and HHPred based on hidden Markow model against protein databases such as PDB. The Arabidopsis sequence of Rubisco activase and selected four hypothetical proteins of *V.vinifera* were subjected to multiple alignments using ClustalO and analyzed for sequence conservation. The secondary structures were predicted using EsPript 3.0 **[6]**. The hypothetical proteins were also subjected to protein disorder prediction using consensus prediction method **[7]**.

Function Prediction
The domain analysis was done for the precise function prediction. The function domain of four hypothetical proteins was predicted by using various publically available protein family databases and tools like PFAM **[8]**, SUPERFAMILY, CATH, NCBI Conserved Domains Database (NCBI-CDD) **[9]**, CDART and INTERPROSCAN **[10]**.
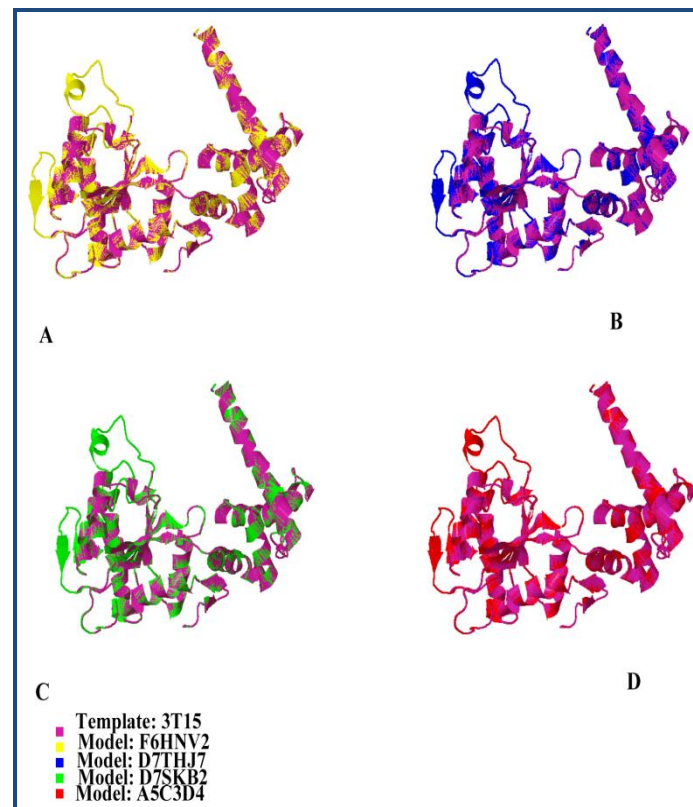
Figure 3: Homology modelled proteins **(A)** F6HNV2 **(B)** D7THJ7 **(C)** D7SKB2 and **(D)** A5C3D4 are superimposed with the template of green-type Rubisco activase from tobacco (PDB ID: 3T15) by using UCSF chimera-1.10.

Structure prediction
The hypothetical proteins were subjected to blast search against PDB database. As all the proteins had more than 80% identity against the templates, the protein sequences were subjected to automated modelling in SWISS-MODEL **[11]**. The homology modelled proteins were subjected to energy minimization using Modrefiner. The energy minimized structures were assessed with PROCHECK Ramachandran plots **[12]**, ERRAT and PROSA-web **[13]**. All the homology modeled proteins were superimposed with the template using UCSF Chimera 1.10 **[14]**. As the detection of structural similarities in proteins can give elucidation of the biochemical functions, the homology modeled hypothetical proteins were subjected to identification of structurally similar binding sites in protein using ProBis server.

Results & Discussion:
Sequence analysis
Arabidopsis Rubisco activase long isoform protein sequence and the selected hypothetical protein sequence of *V. vinifera* were used to create multiple protein sequence alignment using clustalO with default settings were applied for the alignment. A

conserved tyrosine in the C-terminal extension (residue number 416) was identified, which is essential for the ATPase activity of Rca as per previous findings **(Figure 1).** The secondary structure prediction and multiple alignment was prepared using EsPript3.0. Protparam was used to analyze different phyisochemical properties from the amino acid sequence. Predicted phyisochemical properties for four hypothetical proteins are shown in **Table 1 (See supplementary material)**. All the four hypothetical proteins are stable protein as inferred from instability index. All the protein sequences have negative GRAVY index which is an indication of hydrophilic and soluble protein. Then selected hypothetical protein sequence searched for consensus protein disorder prediction revealed that most disorder residues occur in both N (between 16-22) and C terminus (between 450-472) loop regions. The predictions from various subcellular localization tools indicated all the hypothetical proteins are a cytoplasmic protein as shown in **Table 2 (See supplementary material)**.
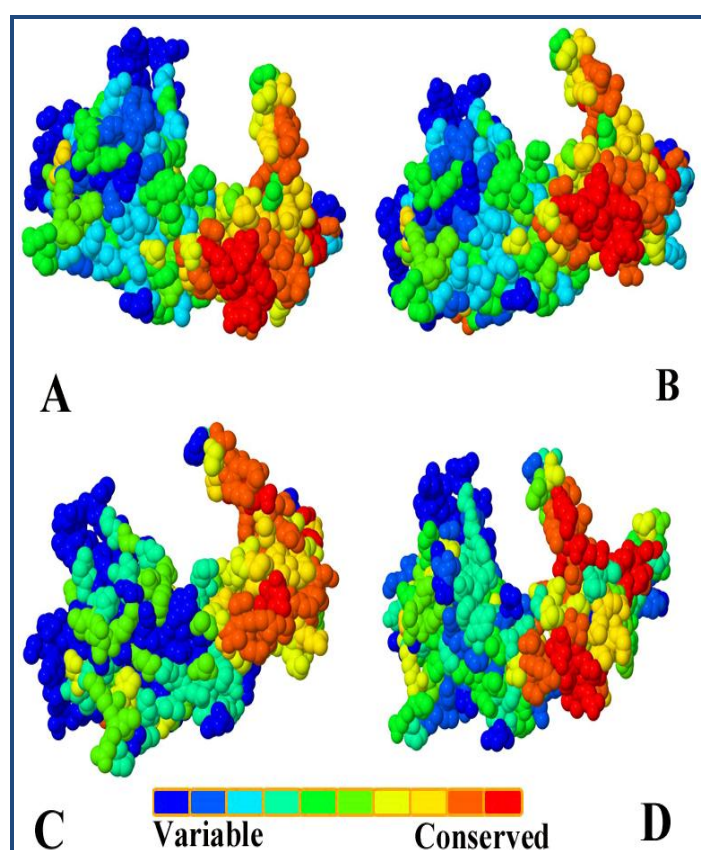
Figure 4: Predicted functional regions of selected hypothetical protein of *V. vinifera* **(A)** D7THJ7 **(B)** D7SKB2 **(C)** A5C3D4 and **(D)** F6HNV2 by using ProBis server. The highly conserved residues are shown in red and least conserved regions are show in blue.

Family classification predicted in PFAM suggests that all the hypothetical proteins belong to ATPase family associated with various cellular activities (AAA). Further, the search in CATH database suggests transcript termination protein A18- like domain and superfamily search suggests P-loop containing nucleotide triphosphate hydrolases **(Table 2)**. The functional prediction using various tools suggests that all the hypothetical proteins of *V. vinifera* have putative function to Rubisco activase.

Structure analysis

Protein three-dimensional structure provides precise information on how proteins interact and localize in their stable conformation. Homology modelling is one the most common structure prediction methods in structural genomics and proteomics. There are many homology modelling server and modelling through these servers are reliable if the sequence identity is greater than 40%. Initial search of the hypothetical proteins against Protein Data Bank (PDB) using BLASTp showed that all our query sequence were above 80% identity with the template structure of green-type Rubisco activase from tobacco **Table 3 (See supplementary material)**. SWISS-MODEL is an automated homology modelling server that builds 3D models using the template identified from the PDB. The SWISS-MODEL was used for homology model construction for selected four hypothetical proteins of *V. vinifera* **(Figure 2)**.

The predicted homology models were further refined using energy minimization server ModRefiner. The energy minimized structures were assessed for both geometric and energy aspects. The PROSA-web accesses the protein structure with Z-score which is indicative of overall model quality. It was used to check whether the input structure is within the range of scores typically found for native proteins of similar size. Z-scores of template and query model were obtained from PROSA web server. The template Z-score was -6.57 and for the homology modelled proteins it was around -7.93 to -8.16 **(Table 3),** suggesting similarity between template and query structure. Finally, the Ramachandaran plots were obtained for both the template and homology model protein as a quality assessment using PROCHECK server. From the Table 4, it is inferred that all the homology models are reliable. Similarly, all the homology modelled hypothetical proteins have passed in ERRAT and PROVE structure assessment. The homology modelled proteins were superimposed with template structure of green-type Rubisco activase from tobacco using UCSF Chimera 1.10 and all the superimposed structure had Root Mean Squared Deviation (RMSD) value of 0.65 **(Figure 3)**.

Functional annotation

Several bioinformatics tools and databases were used with the goal of performing functional annotation of hypothetical proteins. These proteins were searched for the conserved domains and protein functions. Consensus predictions suggested that selected hypothetical proteins of *V. vinifera* contain ATPase family associated with various cellular activities (AAA) belong to the AAA+ super family of ring-shaped P-loop. They contain nucleoside triphosphate hydrolases, which exert their activity through the energy-dependent remodeling and translocation of macromolecules **Table 4 (See supplementary material)**.

They are essential for many cellular functions and are involved in processes such as DNA replication, protein degradation, membrane fusion, microtubule severing, peroxisome biogenesis, signal transduction and the regulation of gene expression. Moreover, the functional regions of protein predicted using ProBis server, indicated structure conservation by pairwise local structural alignment of structural homologs in Protein Data Bank. The structure conservation is depicted in the structure **(Figure 4)**.

BIOINFORMATION

Conclusion:

In this study, we have used sequence analysis, functional prediction and structure prediction to assign putative Rubisco activase function to selected hypothetical proteins of *Vitis vinifera*. Rubisco activase plays an important role in regulating Rubisco. This study will lead for further research in optimizing the functionality of Rubisco which has large implications for the improvement of plant productivity and resource use efficiency. Experimental validation will provide more insight into the actual function of these proteins.

References:

[1] Carmo-silva E *et al. Plant Cell Environ*. 2014 doi: 10.1111/pce.12425. [Epub ahead of print] [PMID: 25123951]

[2] Portis AR Jr, *Photosynth Res.* 2003 **75:** 11 PMID: 16245090]

[3] Wilkins MR *et al. Methods Mol Biol.* 1999 **112:** 531 [PMID: 10027275]

[4] Yu CS *et al. Proteins.* 2006 **64**: 643 [PMID: 16752418]

[5] Altschul SF *et al. Nucleic Acids Res*. 1997 **25**: 3389 [PMID: 9254694]

[6] Gouet P *et al. Bioinformatics* 1999 **15**: 305 [PMID: 10320398]

[7] Kumar S & Carugo O, *Open Biochem J.* 2008 **2**: 1 [PMID: 18949069]

[8] Finn RD *et al. Nucleic Acids Res.* 2010 **38**: D211 [PMID: 19920124]

[9] Marchler-Bauer A *et al. Nucleic Acids Res.* 2011 **39**: D225 [PMID: 21109532]

[10] Zdobnov EM & Apweiler R, *Bioinformatics* 2001 **17**: 847 [PMID: 11590104]

[11] Arnold K *et al. Bioinformatics* 2006 **22:** 195 [PMID: 16301204]

[12] Laskowski RA *et al. J Biomol NMR*. 1996 **8**: 477 [PMID: 9008363]

[13] Wiederstein M & Sippl MJ, *Nucleic Acids Res.* 2007 **35**: W407 [PMID: 17517781]

[14] Pettersen EF *et al. J Comput Chem*. 2004 **25**: 1605 [PMID: 15264254]

BIOINFORMATION

Supplementary material:

Table 1: List of predicted physiochemical parameters by Expasy's ProtParam tool for four selected hypothetical proteins from *Vitis Vinifera*.

S.NO	UNIPROT ID	Molecular weight, M_w (Da)	Theoretical PI	Extinction coefficient (M^{-1} cm^{-1})	Instability Index		Aliphatic index	Grand average of hydropathicity (GRAVY)
					computed	classification		
1	**D7THJ7**	51880.9	5.69	44725	33.31	Stable	76.96	-0.346
2	D7SKB2	51884.8	5.54	46215	35.40	Stable	75.81	-0.330
3	A5C3D4	51925.1	5.54	46215	35.54	Stable	75.59	-0.334
4	F6HNV2	48014.7	5.50	41620	28.38	Stable	80.80	-0.219

Table 2: List of predicted sub-cellular localization of selected hypothetical protein of *Vitis vinifera.*

S.No	UNIPROT ID	Sub-cellular localization		Signal Peptide	Secretory Protein (SecretomeP)	Trans membrane helices prediction	
		WoLF PSORT	CELLO			HMMTOP	TMHMM
1	**D7THJ7**	Chloroplast	chloroplast	No	No	Not present	Not present
2	**D7SKB2**	Chloroplast	Chloroplast	No	No	Not present	Not present
3	**A5C3D4**	Chloroplasta	Chloroplast	No	No	Not present	Not present
4	**F6HNV2**	Chloroplast	Chloroplast	No	No	Not present	Not present

Table 3: Structural prediction of selected hypothetical proteins of *Vitis vinifera*. Table shows query sequence in Uniprot id and template used in PDB and sequence identity between query and template in percentage used for homology modelling. PROCHECK, PROVE, ERRAT and PROSA was used for structural evaluation.

Uniprot ID	Template (PDB ID)	Sequence identity %	PROCHECK	PROVE	ERRAT	PROSA
D7THJ7	3T15	90.44	Favoured region: 96.6%, Allowed region: 2.7%, Outlier region: 0.7%	PASS	PASS At least 80% of the amino acids have scored >= 0.2 in the 3D/1D profile	-8.16
D7SKB2	3T15	90.78	Favoured region: 93.5%, Allowed region: 4.5%, Outlier region: 2.1%	PASS	PASS At least 80% of the amino acids have scored >= 0.2 in the 3D/1D profile	-8.05
A5C3D4	3T15	90.78	Favoured region: 95.5%, Allowed region: 3.5%, Outlier region: 0.7%	PASS	PASS At least 80% of the amino acids have scored >= 0.2 in the 3D/1D profile	-8.05
F6HNV2	3T15	89.08	Favoured region: 94.2%, Allowed region: 5.2%, Outlier region: 0.7%	PASS	PASS At least 80% of the amino acids have scored >= 0.2 in the 3D/1D profile	-7.93

Table 4: Function annotation of selected hypothetical proteins of *Vitis vinifera* predicted by using CATH, Superfamily, PFAM and CDART

Uniprot ID	CATH	Superfamily	Pfam	CDART
D7THJ7, D7SKB2, A5C3D4, F6HNV2	Transcript termination protein A18 -like domain	Superfamily:P-loop containing nucleoside triphosphate hydrolases Family:Extended AAA-ATPase domain	ATPase family associated with various cellular activities (AAA)	ATP-binding cassette transporter nucleotide-binding domain