RESEARCH ARTICLE

INSIGHT INTO SEQUENCE-STRUCTURE-FUNCTION RELATIONSHIP OF *CATHARANTHUS ROSEUS* RNA BINDING PROTEIN USING *INSILICO* APPROACH

Vibha* and H.S. Ginwal

Genetics and Tree Propagation Division, Forest Research Institute, Dehradun
Email: vibha.bioinfo@gmail.com

Received-13.03.2016, Revised-22.03.2016

Abstract: RNA binding protein regulates numerous aspects of RNA metabolism such as pre-mRNA processing, transport and translation. This study describes sequence-structure-function relationship between the *Catharanthus roseus* and their homologs plant species through computational approach. After using sequence analysis techniques, it was observed that only 11 plant species showed higher similarity with RNA binding protein of *C. roseus.* Also, multiple sequence alignment illustrate only two conserve regions between *C. roseus* and their respective homologs plant species. Hence, the structural molecular model of the RNA binding protein was developed through homology modeling using the software MODELLER (9v5). Using PROCHECK and VERIFY-3D, the energy of constructed models was minimized and qualities of each models were evaluated. The corresponding Ramachandran plot specify 93.70% amino acid residues were in the most favoured regions. Final predicted model structure was submitted to Protein Model Database having deposition number PM0080432.

Keywords: RNA binding protein; *Catharanthus roseus*, Homology modeling, RNA recognition motif

INTRODUCTION

RNA binding proteins (RBPs) are key players in aspects of post-translational gene regulation and their involvement in regulating several development processes, a large body of evidence is supporting their key function in plant adaptation to various environmental conditions (Lorkovic, 2009). RBPs are involved in a variety of hetrogenic proteins and included in diverse aspects of post-translation regulation by direct interaction with single/double strand molecule. mRNA maturation events such as splicing, capping, polyadenylation and export from the nucleus are essential for mediating their interation with proteins. RBPs also promote to post-transcriptional regulatory events in the cytoplasm, such as mRNA localization, mRNA stability, decay and translation (Burd and Dreyfuss, 1994; Kim *et al.*, 2002). Despite the diversity of their function in fascinating with RNA and regulating post-translational events, most RBPs were constructed in versatile modular structure with multiple repeats of few conserve domains and arranged in a variety of ways to complete their diverse functional requirement (Lunde *et al.*, 2007). The RBPs are unique members of hetrogenous superfamily of glycine-rich proteins (GRPs) and contains a RNA binding domain at the N-terminal, which is the form of RNA recognation motif (RRM). This is pursued by a C-terminal glycine-rich domain (Ambrosone *et al.*, 2012).

With the innovation in sequencing technologies, it is now significantly easier to obtain the uncharacterized function of protein in the plant. Still, there are the protein sequences with their functions yet to be discovered or experimentally confirmed. These uncharacterized proteins show a enormous undetermined field with various opportunities. *Insilico* analysis help in determining of the protein functions, which can be divide into three broad categories: sequence, expression and interaction based methods. Sequence based methods rely on the ability to construct alignment between protein sequences. The sequence analysis approach was used for the prediction of protein domain family and their identification was based on the amino acid sequence similarity (Oany *et al.*, 2014). However, structural analysis and comparitive study allow to confirm the function of the protein and all these predictions are based on the sequence analysis.

Hence, the *insilico* study of *C. roseus* RNA binding protein with respect to its homologs species is very interesting task to understand the molecular evolution of this protein among the species. Hence forward, the present study affirmed the sequence-structure-function relationship between *C. roseus* and their respective homologs species. Besides, the phylogenetic relationship and motif discovery between *C. roseus* RNA binding protein and their identified homologs species were also established.

MATERIAL AND METHOD

In present analysis, initially full length of 137 amino acid sequence of *C. roseus* RNA binding protein (Accession no AAF31402) was retrived from the Genbank, a protein sequence database of National Center for Biotechnology Information (NCBI).

Homology prediction

In this context, homology prediction technique was performed to identify existence of the similar regions

*Corresponding Author

among the sequences of different plant species. The smith-waterman algorithm based program BLASTp (Altschul *et al.*, 1997) was used to predict the homologs of *C. roseus* RNA binding protein. The amino acid sequences of all the identified homologs species were downloded in FASTA format and used for comparative analysis with the *C. roseus* amino acid sequence.

Multiple Sequence Alignment and Phylogenetic analysis

The amino acid sequence of *C. roseus* RNA binding protein with their homologs protein sequences were subjected to multiple sequence alignment (MSA) for recognizing the conservation through CLUSTRAL-X program (Bateman, 2007).

3D Structure Prediction, Validation and Annotation

The 3D structure of *C. roseus* RNA binding protein was constructed through python based program MODELLER 9v5 (John and Sali, 2003) and a total of 50 models were generated. Modeller generated several models for the same target and the best model was selected for further analysis. The model was evaluated with the lowest value of Modeller objective function, after that used PROCHECK (Laskowski, 1993) statistics. The structure annotation was described through SAS-Sequence structure server. ProFunc server was used to identify the biochemical function of a protein from its three-dimensional structure and PDBsum was used for secondary structure analysis. ProFunc and PDBsum servers are available at European Bioinformatics Institute (Laskowski, 2009).

Physiochemical characterization

For physiochemical characterization, theoretical pI (ioelectric point), molecular weight, -R and +R (total number of positive and negative residues), EI (extinction coefficient), II (instability index), AI (aliphatic index) and GRAVY (grand average hydropathy) were computed using the Expasy's ProtParam server for set of proteins (http://us.expasy.org/tools/protparam.html).

Submission of the modeled protein in protein model database (PMDB)

The model of *C. roseus* RNA binding protein was successfully submitted in protein model database with no stereochemical errors. The submitted model can be accessed via their PMID:PM0080432.

Sequence-Structure-Function Relationship

The sequence-structure-function relationship assessment is used for understanding the molecular mechanism of protein. In this study, all the identified conserved patterns were recommended for the structure prediction through PyMol program (DeLano, 2002) to find out their structural role and functional analysis. These predicted pattern was used for identifying their domain families by Pfam analysis (Finn *et al*., 2010).

RESULT AND DISCUSSION

Sequence analysis and Homology prediction

RNA binding protein (163.08 Kda) of *C. roseus* was extracted from Genbank with 160 aa in length. The homologs protein sequences were found in other plant species and extracted through the protein BLAST program (Johnson *et al.*, 2008). The name and gene-id of all identified glycine-rich RNA-binding protein sequences have been listed in Table-1, having high sequence identity and least e-value**.** After comparative analysis among listed sequences 86% maximum identify, 153 score and 4e-46 E-value was found.

Table 1. RNA Binding protein in different plant species

S.No.	Plant Name	Sequence Id
1	*Solanum tuberosum*	gi|799015|
2	*Catharanthus roseus*	gi|6911142|
3	*Oryza sativa*	gi|169244425|
4	*Sesamum indicum*	gi|747090235|
5	*Zea mays*	sp|P10979.1|
6	*Sinapis alba*	sp|P49310.1|
7	*Arabidopsis thaliana*	sp|Q03250.1|
8	*Daucus carota*	sp|Q03878.1|
9	*Brassica napus*	sp|Q05966.1|
10	*Hordeum vulgare*	sp|Q43472.1|
11	*Sorghum bicolor*	sp|Q99070.1|
12	*Nicotiana sylvestris*	sp|P19683.1|

The primary sequence analysis of protein was calculated through Expasy's ProtParam. The programe computes the extinction coefficient of 276, 278, 279, 280 and 282 nm wavelengths and, in addition, 280 nm has been elected since proteins absorb light strongly. Extinction coefficient of protein at 280 nm was 17,085 M^{-1} cm^{-1}. The computed extinction coefficient can help in the quantitative study of protein-protein and protein-ligand interaction in solution. The instability index

provides to determine the stability of protein in a test tube. There are definite dipeptides, which is particularly divergent in the unstable protein compared with those in the stable once. This method assigned a weight value of instability, which is feasible to compute an instability index (II). A protein whose instability index is slighter than 40 is estimated as stable. The value above 40 estimates that protein may be unstable. The instability index value of the protein was found 55.80, which indicates the protein is unstable. The aliphatic index elucidated as the relative volume of a protein occupied by aliphatic side chains (A, V, I and L). It is estimated as a positive factor for the increment of the thermal stability of globular protein. Aliphatic Index (AI) of the protein sequence was 34.74, the very low aliphatic index of the protein sequence indicates that these proteins may be unstable at low temperature. The Gravy average hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids. The number of residues in the sequence were divided. A GRAVY index of protein was -0,796. This low value shows the probability of the better interaction with water.

Conservation analysis and Phylogenetic analysis

Multiple sequence alignment of *C. roseus* RNA binding protein with its homologs 11 plant species showed highly conserved residues and patterns at different positions. Beside, the conserve patterns with their respective position in *C. roseus* showed in **Fig-1**. These results concluded that *C. roseus* RNA binding protein showed high similarity with their respective homologs. Hence, it was also suggested that *C. roseus* RNA binding protein must undergo evolution in between plant species and their respective function in them.

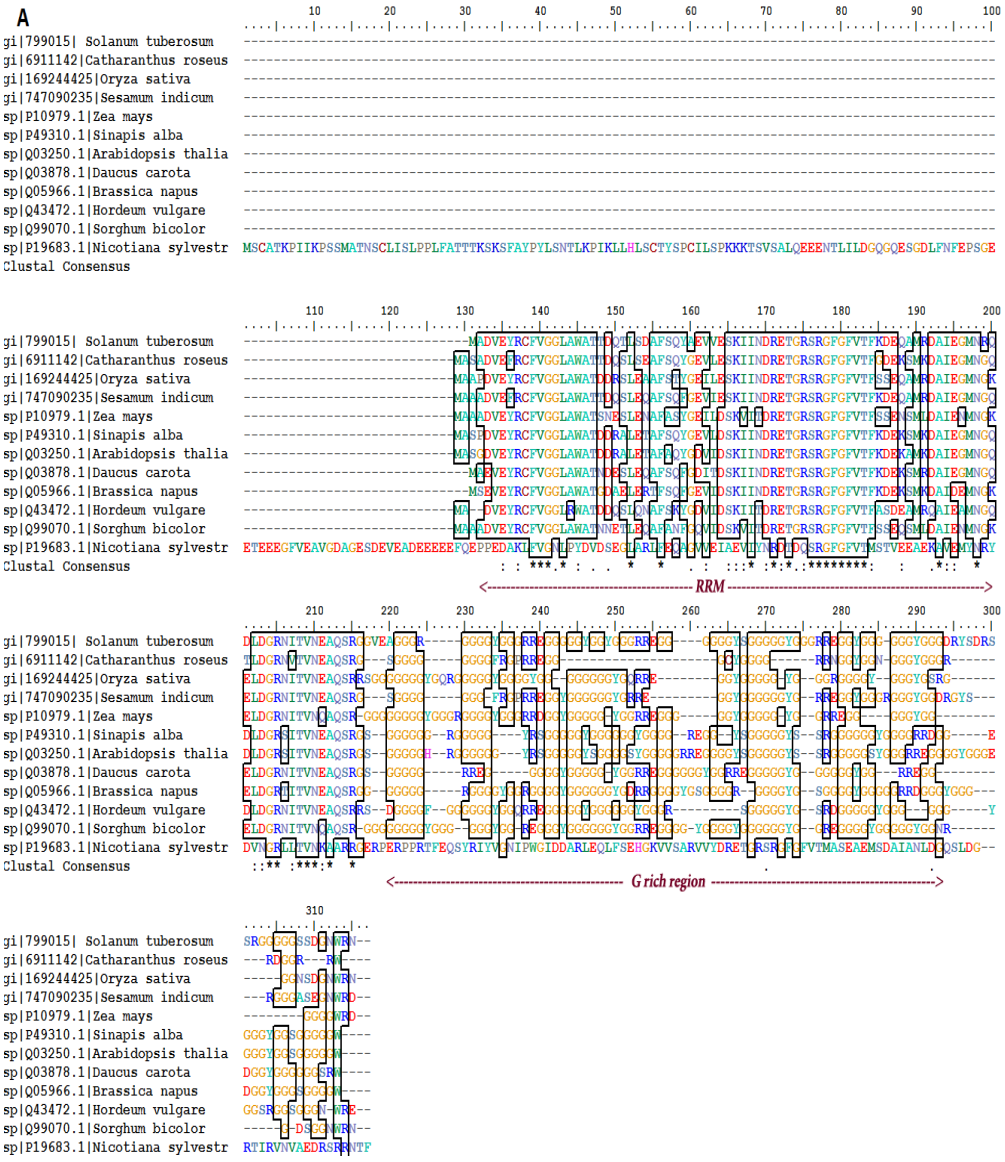


Fig 1. Identification of conserved residues in RRM and glycine-rich regions in the selected orthologus sets of RNA binding protein

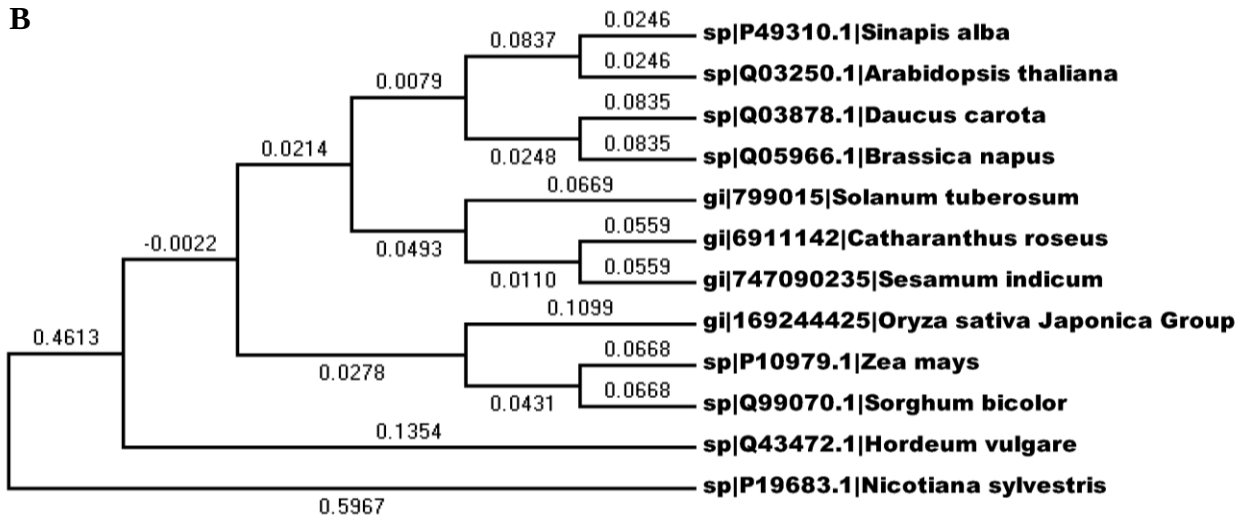


Fig 2. Phylogenetic analysis of RNA binding protein in different plant species

For phylogenetic analysis, special sequences have surpassed through morphological and other plant characters as the most popular form of data. In this section, a single profile of different plants was created and developed a Neighbor-joining tree through MEGA 5.0 software and showed evolutionary relationship among the plant species (Fig-2).

3D Structure Prediction, Validation and Annotation

Comparative modeling of the protein provides a significant hypothesis of homology between the target and the template. This approach provides reasonable results based on the hypothesis and the tertiary structure of the two proteins will be similar if their sequences are related. Absence of the experimentally determined three dimensional protein structure of *C. roseus* in PDB (Protein Data Bank), comparative modeling methods were utilized to construct its theoretical three dimensional structure. BLAST scanning results had shown higher similarity with the crystallographic structure of *Nicotiana tabacum* (PDBid: 4C7Q), while the template was selected on the basis of higher sequence identity. It has been 85.7% sequence identity with 81 conserved residues and 96.4% sequence similarity. Three dimensional structure of RNA binding protein (target) was constructed using comparative modeling and mainly based on the alignment of template. The resulting 50 models were sorted according to the Modeller Objective Function. and root mean square deviation (RMSD). The final model that have the lowest root mean square deviation, related to the trace of the crystal structure was selected for the further study. In constructed model phi and psi torsion angles were checked through Ramachandran plot. The corresponding Ramachandran plot is shown in Fig-3 with following parameters, the phi and psi angles of 93.70% residues in most favoured regions, 4.20% residues in core regions and additional allowed regions, 2.10% residues in disallowed regions. These values showed good quality of the model. The model structure was validated through structure verification servers such as Verify_3d and ProSA webserver (Fig-4A, B). Both programs showed the structure quality of the model was good. Visualization and analysis of model has been done through PyMol software (Fig-5).

Also, ProMotif documentation of the protein via Profunc server for the secondary structure analysis and showed that the 160-residues span of the structure was made of 39 residues (24%) that are involved in the formation of the strands. Likewise, 24 residues (15%) participated in the formation of alpha helices. Besides, 4 beets sheeta and 2 alpha helices were also recorded. After analysis of the predicted structure, it was confirmed that *C. roseus* RNA binding protein can be distinguished into the two domains for the structural framework. It was observed that the first domain found between the arg9 to gly71 residues region belong to the RNA recognation motif (RRM) and the second domain noticed from thr73 to gln85 residues region belong to glycine-rich domain family. The topology of the enzyme structure is illustrated in Fig-3.

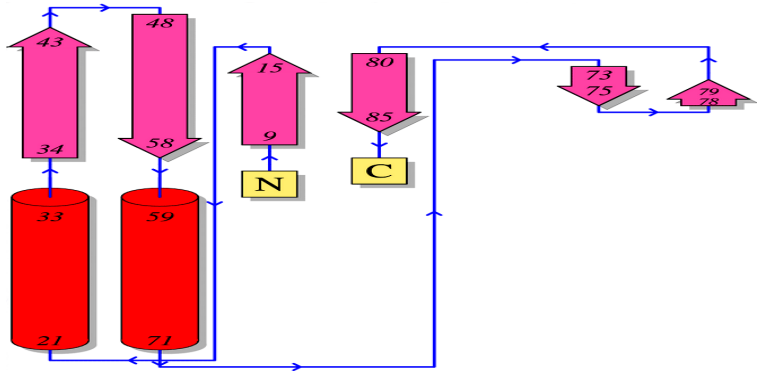
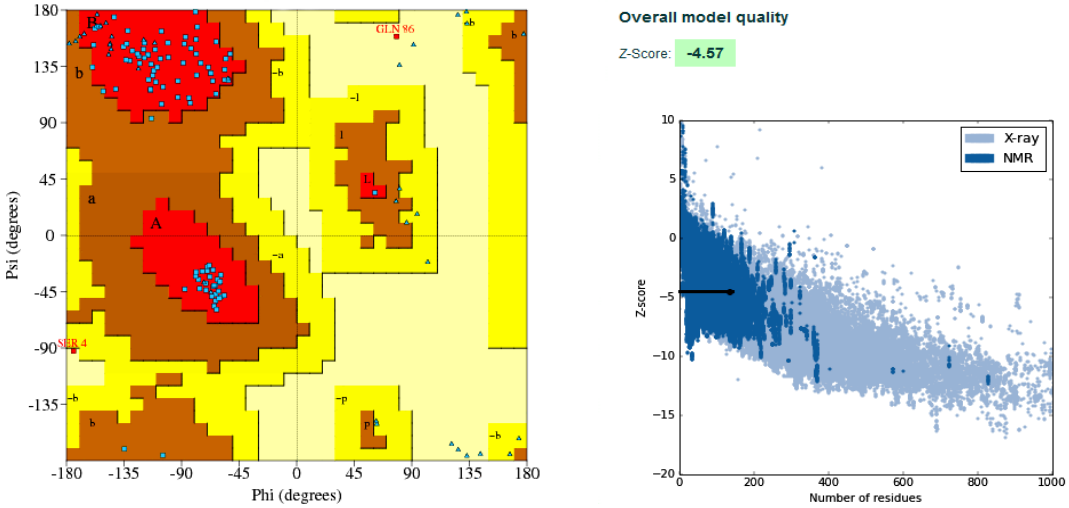


Fig 3. Topology Structure of *C. roseus* RNA binding protein generated through Profunc server



(A) (B)

Fig 4. (A) Showing Ramachandran plot of the predicted model (B) Overall model quality of the structure

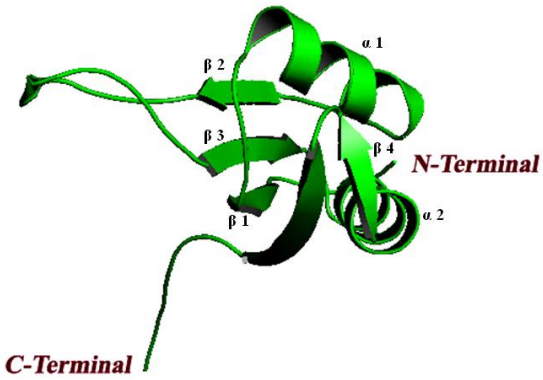


Fig 5. Homology model of *C.roseus* RNA binding protein with two alpha helices and four beta sheets

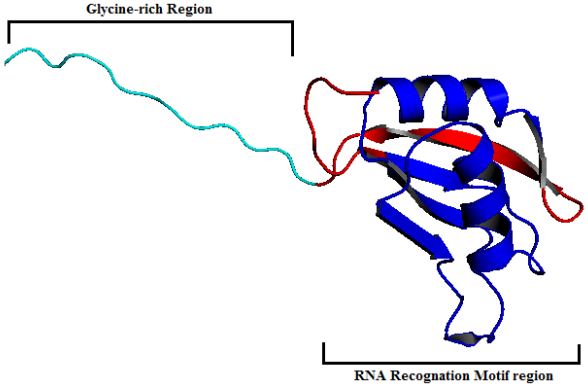


Fig 6. Highlighted structural regions represent the role of patterns in the structural confirmation of *C. roseus* RNA binding protein

Sequence-Structure-Function Relationship

The sequence-structure-function relationship was established by identifying the structural and functional role of conserved patterns, which were found in the multiple sequence alignment profile of all the analyzed sequences. The highlighted structural roles of all the identified conserve patterns have shown in Fig-6. The function of RBPs responces in plant stress and their putative role in enhancing plant tolerance to environmental stress. Additionally, auxiliary domains, such as glycine-rich, arginine-rich or serine-arginine repeats are frequently found in RBPs (Alba and Pages, 1998). Glycine-rich RBPs, harbouring a RRM domain and concomitantly a glycine-rich regions at the C-terminus are widely distributed in cynobacteria, plants and metazoa (Burd and Dreyfuss, 1994; Kim *et al.*, 2010) and have been found to be transcriptionally regulated by environmental stress in plants.

ACKNOWLEDGEMENT

We are thankful to Director, Forest Research Institute, Dehradun for providing the necessary facilities to pursue the above research work.

REFERENCES

Lorkovic, Z.J. (2009). Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends Plant Sci.* **14**: 229–236.

Burd, C.G. and Dreyfuss G. (1994). Conserved structures and diversity of functions of RNA-binding proteins, *Science*. **265**: 615–621.

Dreyfuss, G.; Kim, V.N.; Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry, *Nat. Rev. Mol. Cell Biol.* **3**: 195–205.

Lunde, B.M.; Moore, C. and Varani, G. (2007). RNA-binding proteins: modular design for efficient function, *Nat. Rev. Mol. Cell Biol*. **8**: 479–490.

Oany, AR.; Ahmad, SAI.; Siddikey, MAA.; Hossain, MU. and Ferdoushi, A. (2014). Computational Structure Analysis and Function Prediction of an Uncharacterized Protein (I6U7D0) of *Pyrococcus furiosus* COM1. *Austin J Comput Biol Bioinform*. **1(2):** 5.

Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. **25(17)**: 3389-3402.

Bateman, A. (2007). ClustalW and ClustalX version 2.0. *Bioinformatics*. **21:** 2947–2948.

John, B., Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res*. **31:** 3982-3992.

Laskowski, R.A. (2009). PDBsum new things. *Nucleic Acids Res*. **37:** D355-D359.

DeLano, W. (2002). Pymol Molecular Graphics System: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallogr*.

Finn, R.D.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J.E.; Gavin, O.L.; Gunasekaran, P.; Ceric, G. and Forslund, K. (2010). The Pfam protein families database. *Nucleic Acids Res*. **38:** D211-D222.

Laskowski, R.A.; MacArthur, M.W.; Moss, D.S. and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr*. **26:** 283-291.

Alba, M. and Pages, M. (1998). Plant proteins containing the RNA-recognition motif. *Trends Plant Sci.* **3:** 15-21.

Kim, J.Y.; Kim, W.Y.; Kwak, K.J.; Oh, S.H.; Han, Y.S. and Kang, H. (2010). Glycine-rich RNA-binding proteins are functionally conserved in *Arabidopsis thaliana* and *Oryza sativa* during cold adaptation process. *J. Exp. Bot.* **61:** 2317-2325.