BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 10(2)

# Computational analyses of protein coded by rice (*Oryza sativa japonica*) cDNA (GI: 32984786) indicate lectin like $Ca^{2+}$ binding properties for Eicosapenta Peptide Repeats (EPRs)

**Sunil Archak[1]* & Javaregowda Nagaraju[2]**

[1]Division of Genomic Resources, National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi INDIA; [2]Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad INDIA; Sunil Archak - Email: sunil.archak@gmail.com; Phone: +91-11-25846074; Fax: +91-11-25842495; *Corresponding author

Received January 19, 2014; Accepted January 26, 2014; Published February 19, 2014

**Abstract:**
Eicosapenta peptide repeats (EPRs) occur exclusively in flowering plant genomes and exhibit very high amino acid residue conservation across occurrence. DNA and amino acid sequence searches yielded no indications about the function due to absence of similarity to known sequences. Tertiary structure of an EPR protein coded by rice (*Oryza sativa japonica*) cDNA (GI: 32984786) was determined based on *ab initio* methodology in order to draw clues on functional significance of EPRs. The resultant structure comprised of seven α-helices and thirteen anti-parallel β-sheets. Surface-mapping of conserved residues onto the structure deduced that (i) regions equivalent to β12 and α4 with α4-β7 junction actually represent conserved regions as well as two functional sites, (ii) the primary function of EPR protein could be $Ca^{2+}$ binding, and (iii) the putative EPR $Ca^{2+}$ binding domain is structurally similar to calcium-binding domains of plant lectins. Additionally, the phylogenetic analysis showed an evolving taxa-specific distribution of EPR proteins observed in some GNA-like lectins.

**Keywords:** *ab initio* structure prediction, function prediction, repeat proteins, surface mapping, taxa-specific.

**Background**:
Repeat proteins form an abundant and ubiquitous class of proteins. These proteins are characterized by tandem homologous structural motifs of 20-40 amino acids and are categorized based on their structures. Owing to their modular structural properties, repeat proteins have displayed evolutionary success and a wide range of functions **[1].** Computational screening of rice full-length cDNA sequences discovered the existence of proteins containing eicosapenta peptide repeats (EPRs) a novel class of repeat proteins **[2].** EPRs genes code for ~67 kDa proteins that have 10 successive repeat units of a 25 amino acids repeat unit (X2CX4CX10CX2HGGG). The repeats are characterized by, glycine-rich motifs, and periodic occurrence of cysteine residues. EPRs are unique by specific occurrence only in flowering plants and highly conserved amino acid sequences. Although EPRs occur in multiple copies, they are far fewer compared to PPRs. Extraordinary sequence conservation at protein level as well as angiosperm specific occurrence compels the assignment of functional significance to EPRs. However, absence of even a remote homology to known DNA and protein sequences meant that conclusions on functions would be conjectural **[2].**

It is established that protein structure is much more highly conserved than protein sequence since sequence evolves faster than the corresponding structure **[3]**, and hence structural characteristics can better identify functional aspects of the proteins. Protein function can be annotated, based on different protein features such as 3D fold, sequence, structural motifs and functional sites using likelihoods **[4]**. In an effort to

 

understand the functional significance of EPRs, tertiary structure was determined using computational methodologies. Here, we describe (1) prediction and validation of 3-D structure of protein coded by a rice EPR locus (GI: 32984786) and (2) deduction of its functional role based on the structural features.

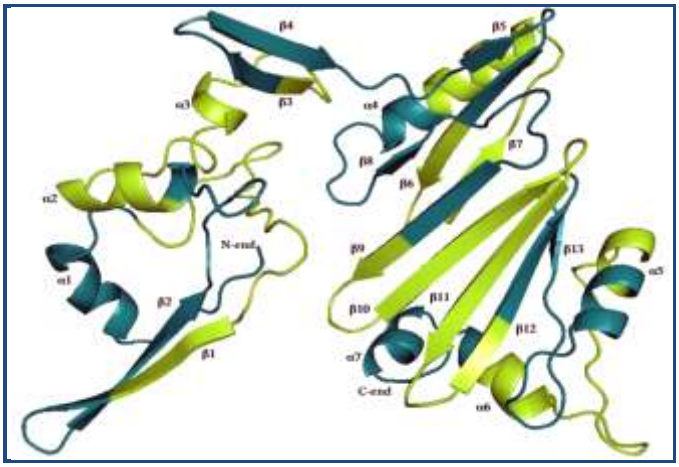


**Figure 1**: Predicted structure of EPR protein. EPR protein is made up of seven alpha helices and 13 beta sheets.

**Methodology:**

***Fold prediction***

Protein sequence was subjected to LOMETS (Local Meta-Threading-Server) analysis (http: / /zhanglab.ccmb.med.umich.edu/LOMETS/). LOMETS generates consensus protein structure predictions from nine locally-installed threading programs of FUGUE, HHsearch, PAINT, PPA-I, PPA-II, PROSPECT2, SAM-T02, SPARKS, and SP3. The output of LOMETS include: Consensus threading models based on TM-score; Spatial C-alpha and side-chain contact and distance restraints; and Full-length models built by MODELLER with consensus restraints.

***3-D structure prediction***

*Ab initio* sequence-based tertiary structure prediction was carried out using a Monte Carlo fragment insertion protein-folding program, ROSETTA **[5]** at http: //www.bioinfo.rpi.edu/bystrc/hmmstr/server.php. The constructed model was evaluated for its backbone conformation (structural quality) using Ramachandran plot **[6]**; Procheck Validation Suite for Protein Structures (http: //biotech.ebi.ac.uk:8400) as well as Verify3D at http://nihserver.mbi.ucla.edu/Verify_3D **[7].**

***Prediction of functional sites***

To identify the conserved residues that are often functionally significant, the multiple sequence alignment of rice EPR homologs (based on BAYES Model of substitution for proteins) was mapped onto the structure using ConSurf at http: //consurf.tau.ac.il **[8].** Putative pockets of interactions with ligands in a protein structure were identified by Putative Active Sites with Spheres (PASS) at http: // bioserv.rpbs.univ-paris-diderot.fr/RPBS/cgi-bin/Ressource.cgi?chzn_lg=an&chzn_rsrc=PASS **[9].** Structural pockets and cavities, expected to be associated with binding sites, were identified by Computed Atlas of Surface Topography of Proteins (CASTp) server at http: //sts.bioengr.uic.edu/castp **[10].**

***Annotation of function***

Bayesian weights were computed for putative GO terms indicating the likelihood of predicted biochemical properties using the ProKnow server (http://www.doe-mbi.ucla.edu/Services/ProKnow/) **[4].** The interpretation of the *Bayesian Score* was done in conjunction with the *Evidence Rank* and the *Number of clues* (Clue Count). An altogether different function prediction methodology that uses local 3D templates by comparing fold matching, residue conservation and surface cleft analysis was employed using ProFunc server https://www.ebi.ac.uk/thornton-srv/databases /profunc/index.html **[11].**

***Construction of phylogeny***

EPR sequences of rice and Arabidopsis EPR loci, as well as ESTs belonging to 20 species of monocots from five families and 45 dicot species belonging to 20 families were downloaded from the NCBI databases. To ensure correct alignment of recursive units, sequences were anchored at the carboxyl-end of the repeat stretch, X2CWX for phylogenetic analysis. The sequences were aligned using ClustalX (version 1.81). The phylogenetic tree was constructed by the neighbor-joining method using MEGA version 4.0, with the setting of complete gap deletion and Poisson correction. Bootstrapping (1,000 replicates) was performed to evaluate the degree of support for a particular grouping in the neighbor-joining analysis.

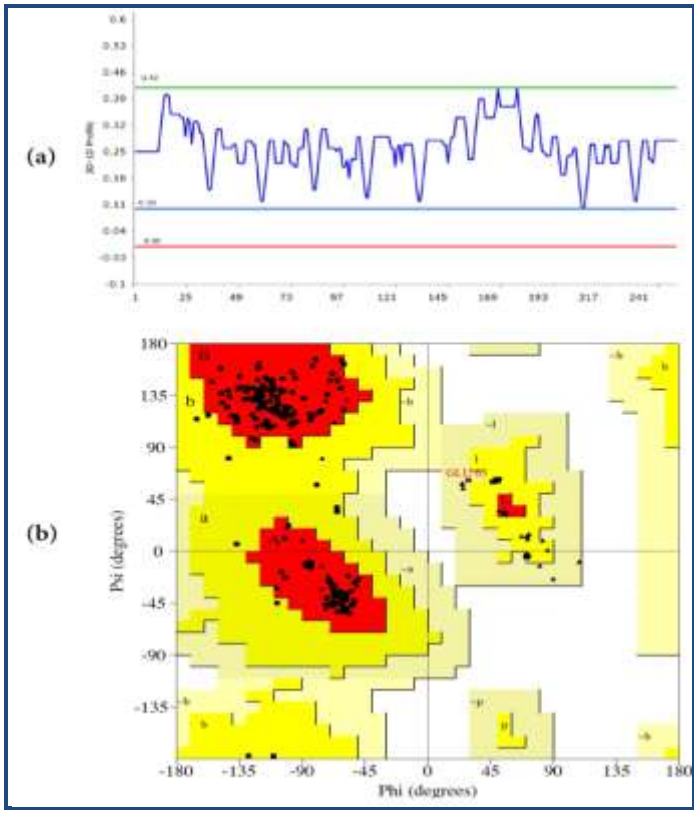


**Figure 2:** Quality evaluation of EPR protein structure. Assessment by **a)** Verify3D and **b)** Ramachandran statistics confirmed a good quality model. Out of 194 non-glycine and non-proline residues, 175 (90.2%) residues were in most favored regions [A,B,L]; 18 9.3% in additional allowed regions [a,b,l,p]; one residue in generously allowed regions [~a,~b,~l,~p]. There were 60 glycine and four proline residues along with a lone end-residues (excluding Gly and Pro).

**Results & Discussion:**

***Fold prediction for EPR structure determination***

Fold recognition returned eight models based on fold homology to five templates, all of which were receptors. Top hits were mouse herceptin2 epidermal growth factor receptor (1N8Y), human Toll like receptor (1ZIW), and Arabidopsis TIR1 plant hormone receptor (2P1M). However, the length of sequence overlap and quality of the fold match (RMSD of the model template alignment were always in double digits) in all the models developed by MODELLER were inadequate.

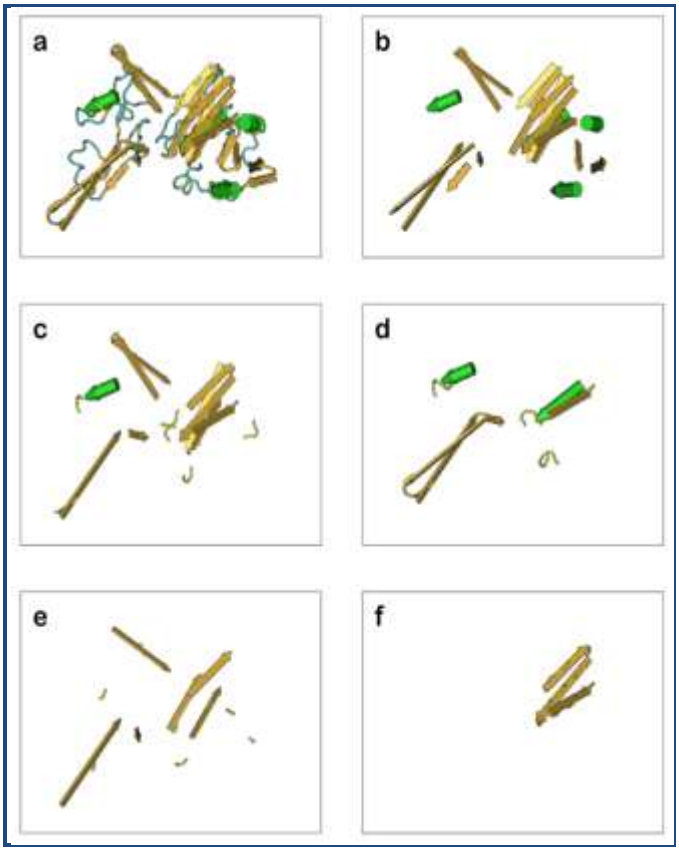


**Figure 3:** Contribution of recursive residues and patterns in achieving structural stability of EPR protein: **a)** Predicted structure of EPR is represented as cartoon showing helices (green), sheets (ochre) and loops (blue); **b)** Cysteine was found be a part of all the helices and beta strands. Other patterns such as **c)** triple glycine; **d)** motif end, AHGGG; **e)** motif start, XRC; and **f)** additional three amino acids in the 8th repeat unit (GGV, GGL, GGI or DDP) were all found to occur in beta sheets 10, 11 and 12.

***Ab initio tertiary structure prediction of EPRs***

*Ab initio* methods have been consulted only under difficult conditions such as very low sequence homology and low scores of meta-prediction. Nevertheless, *ab initio* structures, despite their low quality, provide biological hints. The objective here was to obtain clues on the functional aspects of the EPR protein rather than study the structure *per se*. The 3D structure of only the 259 residue repeat region of protein coded by a rice EPR locus (GI: 32984786) was determined. The resultant structure comprised of seven alpha helices and thirteen anti-parallel beta sheets **(Figure 1).** The recursive units observed in the sequence showed no specific pattern in the structure. Further, unlike typical alpha helix groove of a TPR protein, EPR protein did not show particular structural pattern. Tandem repeat proteins have been classified based on their tertiary structures **[1]** EPR did not seem to fit into any of these categories. With α and β protein type of folds connected by long loops, EPR was found to be altogether different from typical tandem peptide repeats such as TPR, HEAT, LRR, ANK, WD40, Fib, etc., whose 3-D structures have been solved **[1].** The observations on EPR putative structure as well as the absence of homologs for any of its domains indicate a completely novel structure, forming an interesting subject for further investigation on solving the actual structure.

It was crucial that the quality of the predicted structure was validated before drawing any functional cues from the structural characteristics. However there is no single method able to consistently and accurately predict the errors in a protein structure. Hence, two distinct approaches were used, which could complement each other to impart greater confidence to the predicted error of specific regions in the protein. From Verify3D analysis, it was found that 223 residues scored more than 0.2, denoting that 86.1% of the residues complemented with the 1D-3D profile **(Figure 2a).** Since satisfactory models are expected to have a minimum of 80% score, quality of EPR structure was considered good. Ramachandran's statistics showed that more than 90% of the residues were found to be located in the most favored regions, and none in disallowed regions **(Figure 2b).** These observations confirmed predicted EPR protein structure to be a good quality model. From the predicted model, it was evident that recursive residues, conserved pattern and disulphide bonds played major role in determining the 3D structure of EPR protein **(Figure 3).**

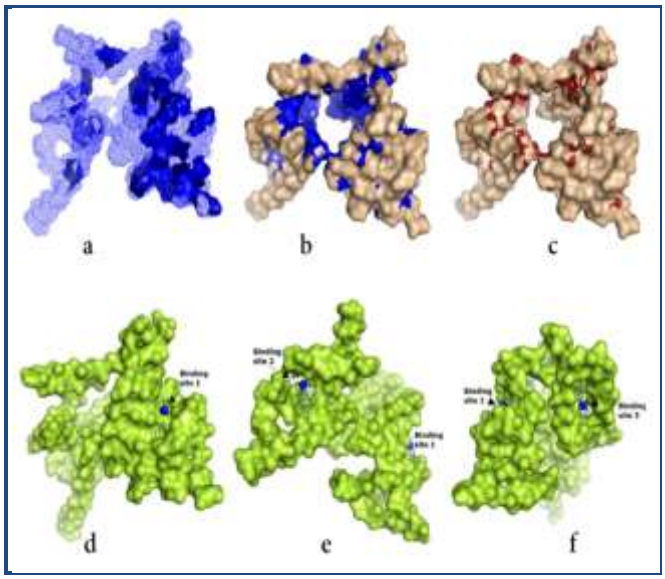


**Figure 4:** Prediction of functional sites of EPR protein: **a)** Conserved residues (blue surface representation) that are often functionally important, are mapped onto the EPR structure (mesh representation); **b)** Structural pockets of EPR protein that could be functionally significant (blue surface areas); **c)** Mouths of the structural pockets showing the accessibility of the putative functional sites for interactions (red surface areas); **d)** to **f)** Molecular probes (blue grains) and ligand binding sites (blue balls) show three binding sites. Predicted structural pockets, conserved regions and ligand binding sites overlap at many sites.

***Prediction of Functional sites on EPR protein***

Identifying mutual interactions with proteins/peptides, nucleic acids or ligands play a vital role in determining the function of a protein. Therefore, clues on the function can be obtained by identifying functionally important amino acids on the protein surface that are responsible for these interactions. It is also known that functional sites are the most conserved residues among sequence homologs. Residue conservation data obtained from all rice EPR protein sequences were surface-mapped onto the predicted structure by ConSurf. The analysis revealed predominantly conserved residues in the carboxyl-terminal half of the protein particularly in β11, β12, α5, α4 with α4-β7 junction, β8 and β9. A few scattered conserved sites in the amino-terminal half were also observed **(Figure 4a).**

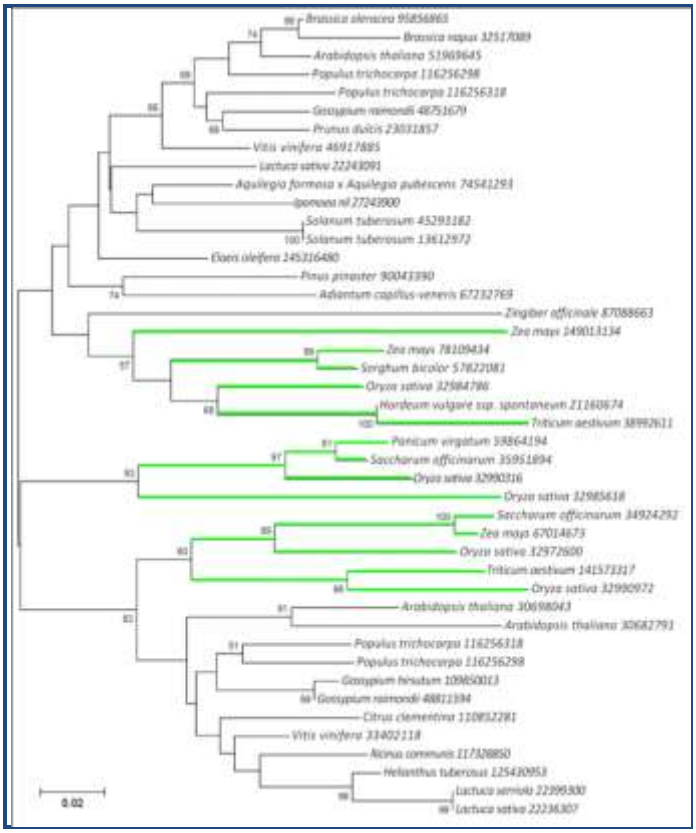


**Figure 5:** Phylogenetic analysis of EPR proteins. Only those EPR sequences that possess carboxyl-end terminal motif X2CWX are used to align and construct the neighbor-joining tree. Each OTU is represented by the binomial name and the genbank accession number of the sequence. Bootstrap values (1000 replications) are shown on the nodes. Grouping of monocots is indicated by the green lines.

Binding sites and active sites of proteins are often associated with structural pockets and cavities, and it has been shown that by locating, defining and measuring concave surface regions, functional sites can be recognized. Surface topography analysis revealed that EPR protein possesses putative functional regions in the interior of the molecule **(Figure 4b, blue colored areas)** as if to resemble a groove or a grotto, in which ligands could interact. Mouth openings of pockets were predicted to be typically on the surface, particularly around the groove opening **(Figure 4c, red colored areas)**, indicating that the binding sites, although in the interior, are readily accessible to ligands and substrates.

The binding sites were also identified using a completely different algorithm that probes the protein surface for identifying putative pockets of interactions. PASS analysis of EPR structure showed the possibility of three putative active site points **(Figure 4d-f, blue spheres)** among the final set of probes (sky blue dots). Location of these active sites completely agreed with the CASTp output and thus, three common putative functional sites were delineated upon the EPR protein at (i) β12, (ii) α4 and α4-β7 junction, and (iii) a groove defined by a β-turn between α2- α3 and a β-hairpin between α1-β1.

***Prediction of Function of EPR protein***

Bayesian weights were computed for putative GO terms indicating the likelihood of predicted biochemical properties. Since the GO assignment was very strong (number of clues was always six) and reliable (the rank scores were never greater than 2.9), the Bayesian weight seemed to have confident inference of the ontology in case of EPR. Top prediction for molecular function was calcium ion binding with a *Bayesian Score* of 0.7. EPR protein was further predicted to have functional role in development (*Bayesian Score* of 0.73). Analyses using local 3D templates by comparing fold matching, residue conservation and surface cleft analysis, confirmed binding as the most likely function of EPR protein, although it could not be precise about the ligand. However, in the process, the ProFunc server identified PDB structure matches that could provide clues on the nature of EPR proteins. All the top hits turned out to be plant agglutinins (lectin) indicating that at least some part of EPR structure has similarity to lectin structures. The top matching proteins were WGA isolectin 3 (1K7T and 1WGT) and Pokeweed lectin-C (1ULK).

***Phylogenetic distribution of plants based on EPR sequences***

Detailed database searches confirmed that EPR proteins and EPR motifs are extraordinarily taxa-specific to only Magnoliophyta, compared to other tandem repeat protein domains such as PPRs and TPRs that are distributed in wider taxa. Such observations proposed that EPRs must be of recent origin. This was further supported by the existence of remarkable sequence similarity across species and occurrence of perfect recursive units in almost all the available sequences. Sequence alignment could not show any sequence pattern specific to monocots and dicots, the two major taxonomical sub-units of flowering plants. The phylogenetic analysis to construct a neighbor-joining tree showed two major clades, both of which contained monocot and dicot species **(Figure 5).** However, it was observed that, in the overall presentation, monocots appeared to be partially differentiated from dicots, which occur in two groups. This kind of phylogenetic distribution is indicative of an on-going evolution process towards taxa-specificity in sequence. Although, it is apparent that EPR proteins must have undergone unit-duplication, lack of sequence data from basal eudicots and other primitive flowering plants restricts the analysis from drawing conclusions on common ancestor where EPRs are likely to have had their origin.

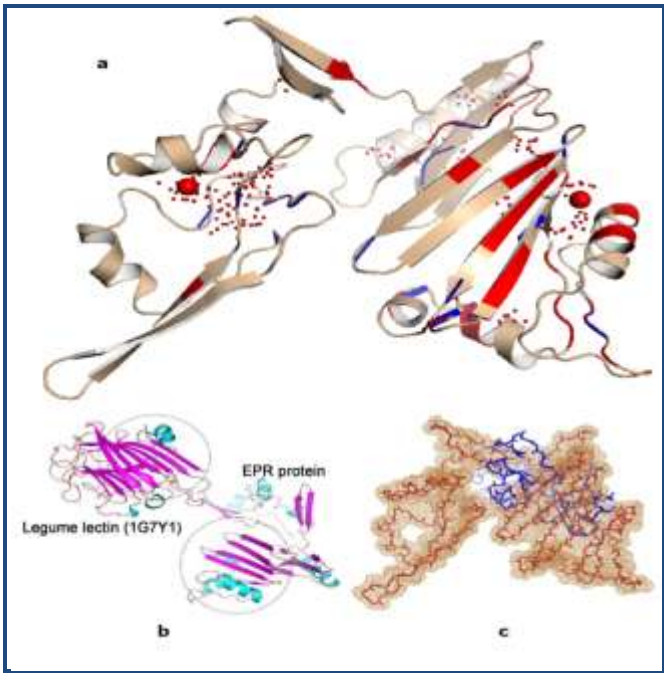


**Figure 6:** Is EPR protein a lectin-like protein? (a) The cartoon representation of EPR protein marked with conserved domains (in red color) and structural pockets (in blue color). Functional pockets (red balls surrounded by red colored probes) that are putatively calcium-binding overlap with conserved domains as well as structural pockets. (b) Calcium binding domains of EPR protein and a legume lectin. (c) It is the same region (functional pockets 1 and 2, see Figure 4) that aligns with a plant lectin (blue frame).

**Conclusion:**

Origin of tandem peptide repeats is attributed to intra-genic duplication and recombination **[1]** and that selection for repeats is a relatively recent evolutionary occurrence **[12].** High conservation combined with narrow phylogenetic specificity of EPRs observed in the study brings forth two facts: first, EPRs have resulted from recent evolutionary events and second, they are functionally significant. When all types of predictions were put together, interesting results emerged. It was evident that regions of EPR coded by rice cDNA (GI: 32984786) exhibiting a great degree of conservation were those that are putatively involved in function such as functional sites 1 and 2 (equivalent to β12 and α4 with α4-β7 junction). However, putative functional site 3 (a groove defined by a β-turn between α2- α3 and a β-hairpin between α1-β1) did not seem to be conserved as much although CASTp analysis marked the site for likely ligand interaction. The alignment between EPR protein and legume lectin as well as WGA isolectin 3 overlapped with functional sites 1 and 2. On the other hand, EPR protein aligned with pokeweed lectin-C in the region of functional sites 2 and 3. Lectins, have a common requirement for divalent metal ions, usually $Ca^{2+}$ and $Ca^{2+}$ binding residues are highly conserved than sugar binding residues **[13].** The observation that EPR protein has domains for $Ca^{2+}$ binding leads us to consider if the conserved regions that are functionally active actually participate in $Ca^{2+}$-binding and that EPR protein could in fact be a lectin-type protein **(Figure 6).** Even the threading meta-server identified the top template (1N8Y) that contained a lectin fold. Additionally, the phylogenetic analysis **(Figure 5)** showed a taxa-specific distribution of EPR proteins that is typical of lectins observed earlier, for instance, monocot-specific (now GNA-like) mannose binding lectins. It looks obvious that EPRs as repetitive units must have been a product of duplication events as found in for instance mannose binding lectins. Plants are capable of acquiring novel domains from existing structural scaffolds with a different activity **[14].** E.g. Curculin from *Curculigo latifolia* is a homolog of GNA related lectin that has no sugar binding activity but has sweet tasting properties **[15];** Arcelins are lectins by structure but do not bind to carbohydrates and have insecticidal activity **[16].** EPR proteins could therefore have lectin-like binding properties **(Figure 6).**

Prediction of a single function merely based on computational analyses could be speculative for repeat proteins. For instance in well-characterized repeat family of TPRs, the bi-helical structure has proliferated to form various sequence sub-families with a wide range of function **[1].** Proteins like EPRs could also be multifunctional; our analysis revealed calcium ion binding function as a definite possibility.

**Acknowledgement:**

Javaregowda Nagaraju passed away on 31 December 2012 and this article is dedicated to his memory.

**References:**

**[1]** Andrade MA *et al. J Struct Biol*. 2001 **134:** 117 [PMID: 11551174]
**[2]** Archak S & Nagaraju J, *Bioinformatics* 2006 **22**: 2455 [PMID: 16809390]
**[3]** Chothia C & Lesk AM, *EMBO J.* 1986 **5:** 823 [PMID: 3709526]
**[4]** Pal D & Eisenberg D, *Structure* 2005 **13**: 121 [PMID: 15642267]
**[5]** Simons KT *et al. J Mol Biol.* 1997 **268**: 209 [PMID: 9149153]
**[6]** Ramachandran GN *et al. J Mol Biol.* 1963 **7**: 95 [PMID: 13990617]
**[7]** Luthy R *et al. Nature* 1992 **356**: 83 [PMID: 1538787]
**[8]** Glaser F *et al. Bioinformatics* 2003 **19**: 163 [PMID: 12499312]
**[9]** Brady GP Jr & Stouten PF, *J Comput Aided Mol Des*. 2000 **14**: 383 [PMID: 10815774]
**[10]** Binkowski TA *et al. Nucleic Acids Res*. 2003 **31**: 3352 [PMID: 12824325]
**[11]** Laskowski RA *et al. J Mol Biol*. 2005 **351**: 614 [PMID: 16019027]
**[12]** Kajava AV, *J Struct Biol*. 2001 **134**: 132 [PMID: 11551175]
**[13]** Maliarik MJ *et al. Plant Physiol*. 1991 **95**: 286 [PMID: 16667966]
**[14]** Van Damme EJ *et al. Plant Physiol*. 2007 **144**: 662 [PMID: 17098856]
**[15]** Harada S *et al. J Mol Biol*. 1994 **238**: 286 [PMID: 8158656]
**[16]** Mirkov TE *et al. Plant Mol Biol*. 1994 **26**: 1103 [PMID: 7811969]

**Edited by P Kangueane**

**Citation**: **Archak & Nagaraju,** Bioinformation 10(2): 063-067 (2014)