Sequence analysis

Eicosapentapeptide repeats (EPRs): novel repeat proteins specific to flowering plants

Sunil Archak and Javaregowda Nagaraju*

Laboratory of Molecular Genetics,Centre for DNA Fingerprinting and Diagnostics, Hyderabad,India.

ABSTRACT

In this report, we describe a novel tandem peptide repeat protein, EPR, which occurs notably only in flowering plants. The EPRs are characterised by a 25-amino acid repeat unit, $X_2CX_4CX_{10}CX_2HGGG$, repeated ten times tandemly. Sequence search revealed that the repeat motif is highly conserved across its occurrence. EPRs are predicted to exist as quasi-globular stable structures owing to highly conserved amino acid positions and potential disulfide bridges. Proteins containing EPRs are predicted to be located in chloroplasts; non-enzymatic and peptide or DNA binding in molecular function; and possibly involved in transcription regulation.

Contact: jnagaraju@cdfd.org.in

Supplementary information: Architecture, identifiers and annotations of EPRs; search parameters, distribution and sequence alignment; 2D structure prediction and disulfide connectivity are provided as pdf files S1-S8.

INTRODUCTION

Tandem peptide repeats are involved in a number of essential functions both in animals (e.g. prion diseases, Alzheimer's disease, Type II diabetes) and in plants (e.g. restoration of male fertility, assembly of photosystem I). However, it has been estimated that no more than 15% of all the proteins reported so far contain repeat motifs (Marcotte *et al.,* 1999). Considering that the proteins containing tandem peptide repeats carryout multiple crucial functions (e.g. TPR; D'Andrea and Regan, 2003), discovery of such proteins is a key to understanding hitherto undetected macromolecular interactions. Efforts to search tandem peptide repeats from the protein databases have been successful (Andrade *et al.,* 2000; Katti *et al.,* 2000). However, protein database searches alone overlook DNA, cDNA and EST sequences that can give a clue about novel tandem peptide repeats. For instance, it was estimated that 11.6% of all fruit fly predicted protein sequences contain tandem peptide repeats (Ponting *et al.,* 2001).

During sequence analysis of rice (*Oryza sativa*) defensins from full-length cDNA sequences (http://cdna01.dna.affrc.go.jp/cDNA), we discovered a transcript coding for a typical peptide repeat sequence. Further, we investigated its domain architecture and its distribution in the rice genome as well as in other species. The study revealed that these tandem peptide repeats constitute a novel peptide repeat sequence family with a unique 25 aa repeat unit, and hence referred to as ***E****icosapenta* ***p****eptide* ***r****epeat*s (EPRs).

*To whom correspondence should be addressed.

RESULTS AND DISCUSSION

Occurrence and conservation of EPRs

The 25 aa EPR unit reads $X_2CX_4CX_{10}CX_2HGGG$. A thorough and non-stringent search of various sequence databases including non-redundant annotated sequences and genome sequences employing blastp, tblastn and gapped and PSI blast tools, using a single EPR unit as query, revealed that EPRs are absent in prokaryotes, fungi and animals but occur exclusively in the plant kingdom (http://www.ncbi.nlm.nih.gov/BLAST/; see supplementary file S1 for details of search parameters). Specifically, EPRs are found only in those higher plants belonging to Magnoliophyta (flowering plants). Hence, even among plants, notable exclusions include lower plant forms of green algae, mosses and liverworts as well as gymnosperms (cycads and conifers). These observations were further supported by HMM based methods for iterative construction of remote homology detection.

ESTs provide the largest source of EPR coding sequences. ESTs belonging to as many as 20 species of monocots from five families and 45 dicot species belonging to 20 families are predicted to code for EPRs (Supplementary file S2). The distribution is apparently skewed towards the species (e.g. crop plants), of which more sequences are available in the open access databases.

Comprehensive information about genomic locations of EPR containing proteins can be obtained only in fully sequenced plant genomes – *Oryza sativa* and *Arabidopsis thaliana* (annotated in the present study; Supplementary file S3). Rice has as many as seven EPR loci across 4 chromosomes (Table 1). Each rice locus codes for a protein with all ten repeat units except OsEPR-6, which codes for a 379 aa long protein containing only four repeats. *Arabidopsis* has four EPR loci distributed on three chromosomes (Table 1). Three AthEPR loci code for proteins containing full-length EPRs whereas, AthEPR-4 carries only seven repeat units. Such variation in the number of repeats even between paralogues of proteins containing tandem repeats, is known to be common as exemplified by WD40 alleles (Saupe *et al.,* 1995).

Presence of repeat units as the major coding part is a typical feature of almost all EPRs of rice and *Arabidopsis*. However, AthEPR-4 is actually a known transcription factor that carries domains for histone deacetylase (SIN3), WRKY DNA binding, Toll-interleukin 1, ATPase, NB-ARC, LRR, Ser-Thr protein kinase, Tyr kinase and protein kinase in that order from residue 300 downstream, with EPR units occurring between residues 91-284 of the transcription factor. This suggests that EPR may function as a domain in addition to being a full assembly. Further, there is a poten-

Table 1. EPR genes of *Oryza* and *Arabidopsis*

Locus	Chr.	cDNA length (bp)	Amino acid length	Number of EPR units
Rice				
OsEPR-1	1	2693	646	10
OsEPR-2	2	2365	519	10
OsEPR-3	2	2635	655	10
OsEPR-4	4	2644	631	10
OsEPR-5	4	2274	637	10
OsEPR-6	6	1629	379	4
OsEPR-7	6	1861	463	10
Arabidopsis				
AthEPR-1	1	2743	1941	10
AthEPR-2	5	2533	1905	10
AthEPR-3	5	3005	1641	10
AthEPR-4	4	5567	5397	7

An expanded version is given as Supplementary file 3

tial case of EPR coding region fused within an otherwise housekeeping gene. There are two malate dehydrogenase genes (MDH-1 and MDH-2) in *Brassica napus*, coding for mitochondrial and glyoxysomal forms of the enzyme. Coding sequences for six EPR units are found in the 5' region of both the genes (e.g. 739-1200 bp in MDH-1 which is 4773 bp long). MDH-1 is coded by an ORF spanning from 2533 bp to 4492 bp through a 1288 base mRNA. Possibly, EPRs are coded by an upstream ORF (331-1260 bp).

Occurrence of EPRs as a complete assembly of repeat units is unique to flowering plants. However, EPR as a domain of one to two units are found in non-plant sources of four environmental sequences from Sargasso sea sequencing programme (GenBank IDs 43092420, 44156513, 43677013, 44070348), one *Ectocarpus siliculosus* virus EsV-1-115 sequence (gi 13177389) and one *Thalassiosira pseudonana* (Diatom) whole genome shotgun sequence (gi 53853431). Absence of even such singletons in higher organisms other than plants is baffling.

High sequence conservation is a characteristic of EPRs (Fig 1 and Supplementary files S4 and S5). Invariable amino acid positions in the EPR units, cysteine at 3, 8 and 19, and histidine followed by three glycine residues at the end of each repeat unit, constitute the signature of the EPR. Among the variant amino acids, positions 1, 2, 10, 13 and 16 are almost always occupied by polar amino acids. Similarly, positions 7, 11, 12 and 14 are occupied mostly by glycine or alanine (Supplementary file S6).

The conservation does not improve within a particular protein let alone in a single plant species. Level of variability at non-consensus residues of a repeat unit within a peptide repeat or among paralogues or orthologues does not show any trend. Under selection pressure throughout evolution, tandem peptide repeats might have conserved all the functionally or structurally important amino acids allowing only a few substitutions to occur. Apart from substitutions, addition of three amino acids is also seen in the 8th repeat unit (GGV, GGL, GGI or DDP) in full length EPRs. There are instances of single residue addition also (in two AthEPRs, addition of proline or leucine in the fourth repeat unit). Shorter OsEPRs, which present a degenerated appearance, display a few other additions (Fig 1). Among the full length EPRs analysed, there has been no case of deletions. Whether deletions are not tolerated and if the presence of additional three residues in some EPRs imparts functional specificity can be answered only by *in vivo* analysis. What is clear by our analysis, though, is that level of conservation among repeat units is a reflection of functional constraints and such constraints for conservation pattern of the EPRs are indicative of the fact that individual unit structures are important in addition to whole assembly (Andrade *et al.,* 2001).

Architecture of the EPRs

A typical EPR containing protein (~67 kD) is coded by a gene of ~4.5kb (Supplementary file S6). Canonically, ten EPR units (~250 aa, 40% of the protein) are arranged tandemly without any gap. In spite of the fact that the amino acid residues of repeat units are not exact duplications, identifying the recursive unit was straightforward due to the degree of sequence conservation. A full length EPR peptide begins with X_2C (first X is usually any of K, R and Q; second mainly a polar residue) and possesses a characteristic X_2CWX motif, of which the first two residues are polar amino acids and the last residue is mostly a glycine or alanine, to mark the end of the repeat unit (Fig 1).

Repeat units are flanked by approximately 260 and 125 amino acids to N-end and C-end respectively. There are conserved Leu-rich motifs DTXLXLX_2L, LXL and PXL among AthEPRs, and LXLGLG among OsEPRs in the N-end. These and a C-end motif ARGX_2GLCX_2H (conserved in both AthEPRs and OsEPRs) did not show any sequence similarity with known non-EPR domains. In contrast, another C-end motif EGRVHGGGLLXLL was found to be present in many non-EPR proteins performing as varied functions as transporters, enzymes, DNA-processing proteins, corneodesmosins etc. in different organisms—bacteria, protozoa, fungus and mammals (including human).

Secondary structure analysis was carried out using the multiple sequence alignment of amino acid sequences derived from rice and *Arabidopsis* EPR loci listed in Table 1. On the whole amino acid distribution in the EPR is as follows: 57.2% polar, 11.5% non-polar and 30.1% glycine and alanine (PHD; Rost, 1996). Solvent accessibility analysis shows that about 62% of the residues are exposed with more than 16% of their surface (PHDacc; Rost, 1994). PHDsec predicted a secondary structure composition predominantly of a random coil (67.56%) followed by strand (31.30%) and helix (1.15%) structures (Rost and Sander, 1993) (Supplementary file S7). Multiple sequence alignment viewing and printing was carried out using JalView 2.07 (Clamp *et al*., 2004). Secondary structure analyses were carried out by using PredictProtein (Rost *et al.*, 2004). Many servers such as SWISS-MODEL, PROSITE, PRODOM, ASP, HMMpFAM failed to return any predictions suggesting lack of any known homologue corroborating the novelty of the EPR.

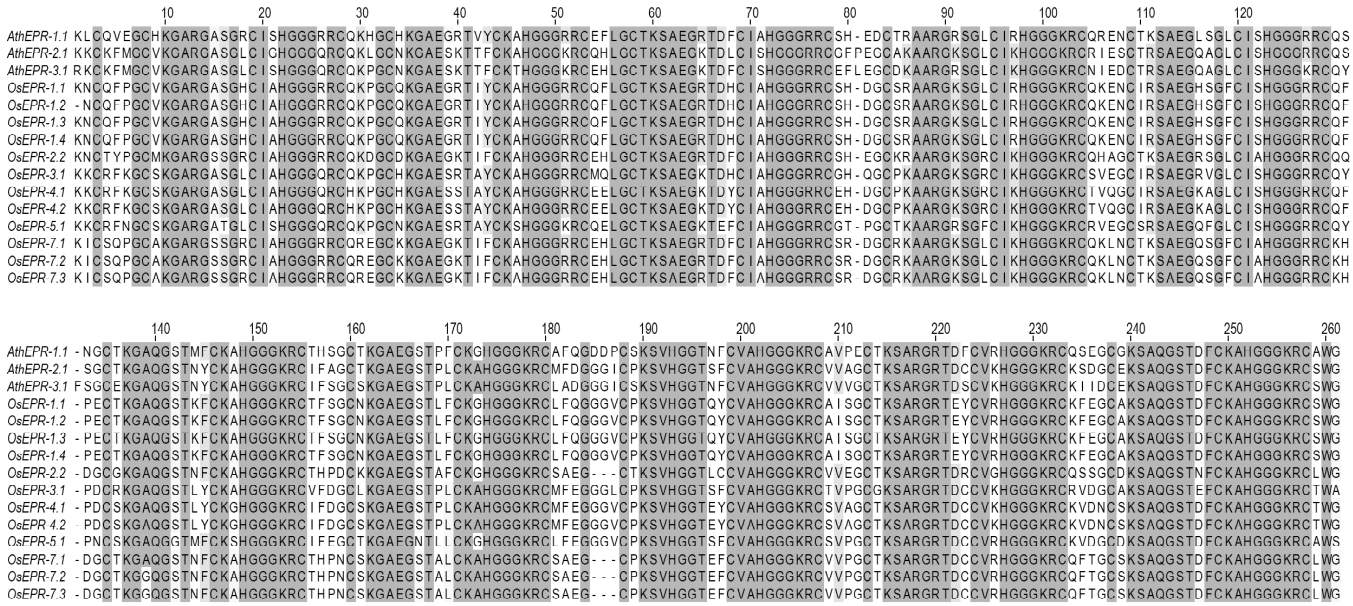

Fig 1. Multiple sequence alignment of full-length EPRs from *Oryza sativa* (OsEPR) and *Arabidopsis thaliana* (AthEPR). Shaded areas highlight the conserved residues. Detailed annotations are given in supplementary file S3.

Since all the prediction methods are trained on globular proteins, predictions of protein structures having long and repetitive domains tend to be inconsistent and therefore results obtained for EPR are treated with caution. Among repeat proteins, features of periodicity and signature residues are major determinants of the final packing of folds. However, these are not adequately incorporated in the existing structure prediction methods.

There are certain primary requirements for an amino acid sequence to achieve a stable structure such as <20% proline, ≥ 30% non-polar amino acids and absence of high iso-charged residues (Kajava, 2001). It is not surprising therefore that quite a few repeat proteins do not attain 3-D structure at all (e.g. elastin, glutenin) or have a definite structure only while they are bound to the substrates (e.g. histone H1 to DNA, or ice nucleation protein to bacterial membrane). However, presence of conserved polar residues, Ser, Thr, Asn, Asp, His and Cys, can result in intra- and inter-repeat molecular ionic and covalent interactions may stabilise the structure as in Zn-finger domains and insect anti-freeze proteins (Kajava, 2001).

Although EPR consists of very few proline residues, presence of ≤ 30% of non-polar amino acids and preponderance of positive charged amino acids (Estimated charge at pH 7.00 = +39.6), considered in isolation, might deny stability to the structure supposed to be full of coils (>65%). However, the below mentioned factors infer that EPRs achieve a stable 3-D fold that could well be different from the PHDsec predictions: (i) EPRs enjoy highly conserved amino acid positions that are more likely to be key positions for fold conservation; (ii) Even if amino acid residue is a variant, presence of an equivalent type of amino acid ensures that the side chain requirement for the structure is met. Regularised protein structures (TPR and ANK) have shown the importance of signature residues in imparting a stable structure (Main *et al.,* 2003); (iii) EPRs are also rich in polar residues contributing to stabilising ionic and covalent interactions, and high solvent accessibility indicates local interactions stabilising the structure (Gilis and Rooman, 1997); (iv) The structure is held together by disulfide bonds predicted for virtually every cysteine residue (DISULFIND; Vullo and Frasconi, 2004) (Supplementary file S8); (v) EPR is predicted to exist as a globular structure though not as compact as a domain (GLOBE, http://cubic.bioc.columbia.edu/papers/1999_globe/paper.html); and (vi) In tandem repeat peptides, a minimum number of repeats have to be reached before correct folds as a protein is achieved (e.g. 1.5 repeats for TPRs, 3 repeats for ANK). Because, unlike globular proteins, repeat proteins can be trimmed off one or more repeats without affecting the scaffold to a great extent (Main *et al.,* 2003). However, in nature repeat proteins are always obtained in higher multiples of the minimum repeat number in order to achieve stability. Majority of the rice and *Arabidopsis* EPRs possess ten repeats plausibly to attain better intra- and inter-repeat packing and to minimise unfavourable inter-molecular interactions.

Proteins with tandem peptide repeats are under-represented in the structural databases (~0.5% of all structures) owing to their non-standard shape and larger size (Kajava, 2001). Hence, structure prediction of novel repeat proteins relies on theoretical and computational approaches. However, in the absence of a homologue whose 3-D structure is known, we could not predict a reliable tertiary structure for EPR.

Prediction of function of EPRs

Origin of tandem peptide repeats is attributed to intra-genic duplication and recombination (Andrade *et al.,* 2001) and that selection for repeats is a relatively recent evolutionary occurrence (Kajava, 2001). 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.

Although absence of analogy to EPRs highlighted their novelty, it made the prediction of their function challenging at every step. We sought to collect as many clues as possible from: (a) sequence homology, (b) multi-server querying to generate integrated information from various post-translational and localisational aspects of the protein, (c) cellular location, and (d) predicted structure.

Going by the periodicity of specific residues in the sequence, EPRs might mimic the functions of zinc-finger motifs participating in nucleic acid binding (http://prodata.swmed.edu/zndb /zndb_view.php). Repeat families commonly represent either enzymes or non-enzymes but rarely both (Andrade *et al.,* 2001). Prediction of function on ProtFun 2.2 server, irrespective of whether the input sequence is an entire EPR containing protein or a full EPR array or a single EPR unit, classified EPRs as non-enzymatic and ascribed them to be particularly involved in transcription regulation. (Jensen *et al.,* 2002; http://www.cbs.dtu.dk/services/ ProtFun/). This was further supported by the observation that EPRs with several disordered regions (46.7% of the residues) could effectively be DNA binding regions (http://genesilico.pl/meta). Amino acid analysis for protein localisation based on PSORT server predicts plastid/chloroplast as the possible destination for rice proteins having EPRs (e.g. OsEPR-1 and OsEPR-2; http://cdna01.dna.affrc.go.jp/cDNA). Likewise, although biological and molecular functions are unknown, *Arabidopsis* EPR harbouring protein AthEPR-1 is predicted to be located in the chloroplast. Similar to PPRs, another illustrious peptide repeat protein from plants that are known to be located in the organelles and functioning in RNA editing, EPRs could well be participating in the processes involving nucleic acid binding in the plastids.

Prediction of a single function merely based on sequence similarities could be speculative. E.g. in well-characterised repeat family of TPRs, the bihelical structure has proliferated to form various sequence sub-families with a wide range of function (Andrade *et al.,* 2001) suggesting that proteins like EPRs could be multifunctional. It is understood that periodic sequence pattern is a mechanism to provide definite arrangement of spatial and functional groups, useful not only for structural packing but also for molecular interactions with targets. EPRs, like TPRs, possess tandem repeat motifs that potentially form scaffolds upon which, components of metabolic processes they are involved in, may assemble. Therefore, EPRs might participate both in nucleic acid and peptide binding as shown in Pumilio, a helical repeat protein, which binds to an extended peptide as well as RNA (Edwards *et al.,* 2001). Leucine-rich conserved positions in 5' and 3' flanking sequences of EPRs may add to a common function of forming complexes with other proteins (Andrade *et al.,* 2001). EPRs also exhibit flexibility in the form of full-length EPRs or 5' EPRs (as seen in *Arabidopsis* WRKY and brassica MDH) possibly to reflect the fact that functional success of repeat families is fundamentally due to their ability to perform different roles (Andrade *et al.,* 2001).

The intricate relationship between stability, repeat motif number and function even in well investigated repeat peptides notwithstanding (Main *et al.,* 2003), if aforementioned possibilities are to be deemed as cues, functional analysis of EPRs is going to be rewarding.

ACKNOWLEDGEMENTS

JN acknowledges the financial support from the Department of Biotechnology, Government of India. SA is supported by ICAR study leave.

REFERENCES

Andrade, M.A., Iratxeta, C.P. and Ponting, C.P. (2001) Protein Repeats: Structures, Functions, and Evolution, *J Struct Biol*, **134**, 117–131.

Andrade, M.A., Ponting, C.P., Gibson, T.J. and Bork, P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates, *J Mol Biol*, **298**, 521-537.

Broekaert, W.F., Terras, F.R., Cammue, B.P. and Osborn, R.W. (1995) Plant defensins: novel antimicrobial peptides as components of the host defense system, *Plant Physiol*, **108**, 1353-1358.

Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The Jalview Java Alignment Editor, *Bioinformatics*, **12**, 426-427.

D'Andrea, L.D. and Regan, L. (2003) TPR proteins: the versatile helix, *Trends Biochem Sci*, **28**, 655-662.

DeLano, W.L. (2005) MacPyMOL: A PyMOL-based Molecular Graphics Application for MacOS X. DeLano Scientific LLC, South San Francisco, CA, USA.

Edwards, T.A., Pyle, S.E., Wharton, R.P. and Aggarwal, A.K. (2001) Structure of Pumilio reveals similarity between RNA and peptide binding motifs, *Cell*, **105**, 281-289.

Gilis, D. and Rooman, M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence, *J Mol Biol*, **272**, 276-290.

Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H.H., Rapacki, K., Workman, C., Andersen, C.A., Knudsen, S., Krogh, A., Valencia, A. and Brunak, S. (2002) Prediction of human protein function from post-translational modifications and localization features, *J Mol Biol*, **319**, 1257-1265.

Kajava, A.V. (2001) Review: proteins with repeated sequence--structural prediction and modeling, *J Struct Biol*, **134**, 132-144.

Katti, M.V., Sami-Subbu, R., Ranjekar, P.K. and Gupta, V.S. (2000) Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications, *Protein Sci*, **9**, 1203-1209.

Main, E.R., Jackson, S.E. and Regan, L. (2003) The folding and design of repeat proteins: reaching a consensus, *Curr Opin Struct Biol*, **13**, 482-489.

Marcotte, E.M., Pellegrini, M., Yeates, T.O. and Eisenberg, D. (1999) A census of protein repeats, *J Mol Biol*, **293**, 151-160.

Ponting C. P., Mott R., Bork P., and Copley R. R. (2001). Novel protein domains and repeats in Drosophila melanogaster: insights into structure, function, and evolution, *Genome Res* **11:** 1996-2008.

Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks, *Methods Enzymol*, **266**, 525-539.

Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy, *J Mol Biol*, **232**, 584-599.

Rost, B. and Sander, C. (1994) Conservation and prediction of solvent accessibility in protein families, *Proteins*, **20**, 216-226.

Rost, B., Yachdav, G. and Liu, J. (2004) The PredictProtein server, *Nucleic Acids Res*, **32**, W321-326.

Saupe, S., Turcq, B. and Begueret, J. (1995) Sequence diversity and unusual variability at the het-c locus involved in vegetative incompatibility in the fungus Podospora anserina, *Curr Genet*, **27**, 466-471.

Simons, K.T., Kooperberg, C., Huang, E. and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, *J Mol Biol*, **268**, 209-225.

Vullo, A. and Frasconi, P. (2004) Disulfide connectivity prediction using recursive neural networks and evolutionary information, *Bioinformatics*, **20**, 653-659.