BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 8(3)

Classification and comparative analysis of *Curcuma longa L*. expressed sequences tags (ESTs) encoding glycine-rich proteins (GRPs)

Basudeba Kar, Sanghamitra Nayak & Raj Kumar Joshi*

Centre of Biotechnology, School of Pharmaceutical Sciences Siksha O Anusandhan University, Bhubaneswar-751003, India; Raj Kumar Joshi – Email: rajkumar.joshi@yahoo.co.in; Phone: 09437684176; * Corresponding author

Received January 18, 2012; Accepted January 19, 2012; Published February 03, 2012

Abstract:
Glycine-rich proteins (GRPs) are a group of proteins characterized by their high content of glycine residues often occurring in repetitive blocs. The diverse expression pattern and sub cellular localization of various *GRP*s suggest their implication in different physiological processes. Several GRPs has been isolated and characterized from different monocots and dicots. However, little or no information is available about the structure and function of GRPs in asexually reproducing plants. In this study, *in-silico* analysis of expressed sequence tag database resulted in the isolation of fifty-one GRPs from *Curcuma longa L.*, an asexually reproducible plant of great medicinal and economic significance. Phylogenetic analysis grouped the GRPs into four distinct classes based on conserved motifs and nature of glycine-rich repeats. Majority of the isolated GRPs exhibited high homology with known GRPs from other plants that are expressed in response to various stresses. The presence of high structural diversity and signal peptide in some GRPs suggest their diverse physiological role and tissue specific localization. The isolated sequences can be used as a framework for cloning, characterization and expressional analysis of GRPs in response to various biotic and abiotic stresses in *Curcuma longa* as well as other asexually reproducing plants.

Keywords: *Curcuma longa*, expressed sequence tags, GRPs, TBLASTN

Background:
The glycine rich proteins (GRPs) belong to a group of super family that is characterized by the presence of semi-repetitive glycine-rich motifs. These groups of proteins have a glycine content of 20 to 70% that are arranged in (Gly)n-X repetitions. Although the first genes encoding GRPs have been isolated from plants, they have been reported in a wide variety of organisms from cynobacterias to animals **[1].** GRPs are broadly classified into four major groups based on conserved motifs and the arrangement of glycine repeats. The class I GRPs contain a signal peptide followed by a high glycine-content region with (GGX)n repeats. These proteins are attributed with structural function due to their cell wall localization **[2]**. Class II GRPs may or may not have a signal peptide. They carry a C terminal cysteine rich region following the glycine rich region and characterized by the presence of universal (GGXXXGG)n repeats. Class III GRPs may carry a signal peptide and have the lowest glycine content as compared to other classes. They are charactrized by the presence of GXGX repeats and show a high degree of structural diversity. The class IV includes the RNA binding GRPs which has the characteristic RNA recognition motif (RRM) or a cold shock domain in addition to the glycine rich domain. A few of the RNA binding GRPs are also characterized by the presence of CCHC zinc-fingers in their structure.

In the past few years, functional characterization of several plant GRPs has been investigated. It is believed that, they are developmentally regulated as well as modulated by biotic and abiotic factors. Although most of the GRPs are attributed with a structural function owing to their cell wall locations, recent development suggest that GRPs are indeed diverse in their location and function, the only similarity being the presence of glycine rich repeats **[3].** In plants, the genes encoding GRPs are induced by physical, chemical and biological factors such as temperature, wounding, pathogen infection, salinity, drought, flooding, light, salycylic acid etc **[1].** This diverse functionality suggests that, GRPs are components of different multi-molecular complexes where glycine rich domains are required for maintaining stability and flexibility of molecular interactions **[4]**. Several GRPs has been characterized in different plants such as Arabidopsis, rice, sugarcane and Eucalyptus. The differential modulation and sub cellular localization together with broad structural diversity suggest that GRPs do not represent the same family of proteins, but a group of protein that share a common structural motif **[1]**.

Curcuma longa L. (turmeric) of the family Zingiberaceae is one of the most important crop with great medicinal and economic significance. Its medicinal uses are indeed diverse, ranging from cosmetic face cream to the prevention of Alzheimer's disease. Turmeric is also qualified as the queen of natural Cox-2 inhibitors **[5]**. India is the world's largest producer, and exporter of turmeric followed by China, Indonesia, Bangladesh and Thailand **[6]**. However, turmeric is completely sterile and propagated exclusively by vegetative means using rhizome. This seems to have eroded their genetic base making them suceptible to major biotic and abiotic stresses. Characterization and comparative analysis of GRPs in turmeric can provide a wide array of informations on the regulation of different stress responses in vegetatively propagated plants. Recent advances in Curcuma genomic technologies have generated a large number of expressed sequence tags (ESTs) that has been made available in public database, thereby offering an opportunity to classify and compare glycine rich protein sequences in turmeric. As of July 2010, GenBank had released 12593 EST sequences from Curcuma longa. In the present study, we describe the isolation, classification and characterization of glycine rich proteins in *Curcuma longa* EST database using known GRP sequences.

Methodology:

A basic local alignment search tool (BLAST) TBlastN search **[7]** was performed using protein sequences of reported plant GRPs as baits against the *Curcuma longa* expressed sequence tag (EST) database. 12593 *Curcuma longa* EST sequences were mined consisting of two tissue libraries of rhizomes 6870 (DY395309-DY388440) and leaves 5723 (DY388439-DY382717). The EST sequences were screened against the UniVec database from NCBI (ftp://ftp.ncbi.nih.gov/pub/ UniVec/) for detecting vector and adapter sequences by using the program Cross_Match. CAP3 program was used to assemble the EST sequence into contigs for creating a non-redundant dataset. The GRP sequences used as baits includes those reviewed by Sachetto-Martins **[1]**, rice GRPs **[8, 9]**, wheat GRPs **[10, 11]**, *Arabidopsis* GRP rich RNA binding proteins **[12]**,a nodule specific GRP from *Medicago* **[13]**, a root specific GRP from *Zea mays* **[14]**, sugarcane GRP sequences **[15]**, *Eucalyptus* GRPs **[16]** and a *Petunia* cold shock GRP sequence **[17]**.

All the turmeric GRPs isolated was subsequently translated to obtain their putative protein sequences. The Open Reading Frames (ORFs) for each searched contig was predicted using the Expasy Translate Tool (bo.expasy.org/tools/dna.html). Protein sequences obtained were used in a second round of TBLASTN search against the non-redundant protein database at the National Center for Biotechnology Information (NCBI) to identify their closest homologues. Additional domains were detected using the Prosite (http://bo.expasy.org/prosite) and Pfam (http://www.sanger.ac.uk/Software/Pfam/search.shtml) prediction programs. The signal peptides were predicted using signalP server (http://www.cbs.dtu.dk/services/SignalP). ClustalX program **[18]** was used to align GRPs deduced from the turmeric EST database. The phylogenetic tree was constructed using the Molecular Evolutionary Genetics Analysis (MEGA) software package version 2.1 **[19]**. The neighbor joining distance method was used with pair wise deletion to treat the amino acid gaps during multiple alignment of turmeric GRPs. For construction of the phylogenetic tree, the confidence levels for the nodes were determined with 1000 replications using the internal branch test **[20]**.

Results and Discussion:

Typical GRP protein sequences were used to search the *Curcuma longa* EST database for genes encoding glycine–rich proteins. Fifty-one potential turmeric GRP gene sequences were isolated and distributed into four distinct classes- class I (GGGX); class II (GGXXXGG); class III (GXGX and class IV, RNA binding GRPs) **Table 1 (see supplementary material)**. Similar *in-silico* approach has also been utilized earlier to identify GRPs in other important plants such as sugarcane **[15]** and eucalyptus **[16]**.

The turmeric GRP sequences were almost equivalent to the other monocotyledonous GRP sequences already published. Fourteen sequences encoding GRPs with GGGX repeats were identified in the *Curcuma longa* EST database. The sequences were quite different and related to previously known GRPs from monocots and dicots. Four sequences showed high similarity with *AtGRP6*, the cold shock glycine rich protein from Arabidopsis. Likewise, four and three sequences showed high similarity with GRPs from *Oryza sativa* and *Zea mays* respectively. Eleven of the 14 class I GRPs showed the presence of a signal peptide at their N-terminal end suggesting their location in the cell wall or cell membrane. Ten sequences encoded GRPs that were highly enriched in histidine having a GGGH repeats. Similar results were also retrieved in Eucalyptus GRPs **[16]**.Searching the turmeric EST database using the previously reported GRPs with cysteine rich domans and C terminal homology to nodulins resulted in the identification of five GRPs with GGXXXGG repeats. The tripeptide between the glycine residues were composed of Y, N and R amino acid. One among the five-class II turmeric GRPs- CL.CON.1727 showed high homology with *Triticum aestivun* predicted protein grp having a distinct signal peptide. The remaining four appears closely related to *HvGRP1 of Hordeum vulgare*. The class II GRPs has been found to interact with cell wall associated kinase molecule that initiate the recognition of various environmental signals in response to external stresses and to transduce them into the cell **[21]**.

The class III GRPs with GXGX repeats consists of the lowest glycine content of only 20%. In turmeric, twenty-one different sequences were identified encoding this type of GRP. These GRPs were also rich in alanine and arginine amino acids besides having the glycine rich domains. The GRP sequences were highly diverse representing heterogeneous groups of proteins with no significant sequence similarity except within the glycine rich motif. Five GRPs with GHGH repeat showed highest homology to a *Zea mays* aluminium induced GRP. Likewise, two turmeric GRPs-CL.CON.1712 and CL.CON.1872 exhibited highest similarity with a cold-drought regulated grp from *Medicago sativa*. This suggests that classes III group of GRPs from turmeric must be getting expressed in response to abiotic stresses. Five out of the 21 class III GRPs possessed signal sequence in the N-terminal end reflecting their extra cellular localization. Many GRPs isolated from other plants had signal peptide and found to be located outside the cell **[2, 22].** No Oleosin GRPs were isolated from turmeric as like other monocotyledonous plants. As the Oleosin GRPs are meant for tapetal development in dicots, the absence of Oleosin GRPs in turmeric further supports the existing difference in the anthers and pollen grain development between monocots and dicots **[23]**.

The class IV GRPs consists of a RNA binding motif in the N-terminal end followed by a C-terminal rich in glycine repeats. It is broadly classified into four sub-classes- sub class I with RNA recognition motif (RRM) in the N-terminal end, sub-class II with RRM conserved motif and a CCHC zinc finger motif within the glycine rich region, sub-class III with a cold shock domain in the N terminus and CCHC zinc fingers within glycine rich motif and sub-class IV with two RRM motifs in the N terminus **[1, 15, 24]**. Eleven *Curcuma longa* sequences encoding RNA binding GRPs were identified and classified according to their domain organizations. Five sequences from turmeric EST database, CL.CON.447, CL.CON.794, CL.CON.1078, CL.CON.1582 and CL.CON.3068 showed highest homology with *Oryza sativa* Japonica group, putative RNA binding glycine rich protein. All the five GRPs encoded a RRM conserved motif in the N-terminus and a CCHC zinc finger motif within the glycine motif classifying them as the sub-class II RNA binding GRPs. Three sequences exhibited homology with an absisic acid inducible RNA binding GRP from Zea mays. They were classified as the subclass I type RNA binding GRPs due to the presence of a single RNA recognition motif in the N-terminus of the protein sequence. The rest of the three sequences were classified as subclass IV RNA binding GRPs. They showed homology with different sequences from *Medicago sativa*, Sorghum bicolor and Triticum aestivatum respectively each having atleast two RRM motif followed by C-terminal glycine rich region. No subclass III GRPs were identified in *Curcuma longa* EST data bank. Absence of subclass III GRPs has also been reported in other monocotyledonous plants **[25]**. However, Fusaro *et al* **[15]** has reportedly isolated ten RNA binding GRPs of subclass III in the sugarcane EST data bank. The greater diversity among the GRPs of turmeric suggests their origin through DNA recombination. The high GC contenet in the grp genes make them as ctive site for recombination events resulting in high variability in different classes of turmeric GRPs. The existance of high recombination in glycine rich regions has been already proved in mammals **[3].** Alignment conducted with turmeric EST encoded GRPs using MEGA ver 2.1 resulted in a distinct unrooted tree **(Figure 1)**. Four well-separated cluster groups were obtained in the phylogenetic tree each representing the member of the particular classes of GRPs. The GRPs with GGXXXGG repeats and the RNA binding GRPs were relatively closure as compared to other classes. Likewise, the turmeric GRPs with GGGX repeats formed a completely separated cluster as compared to other three classes.

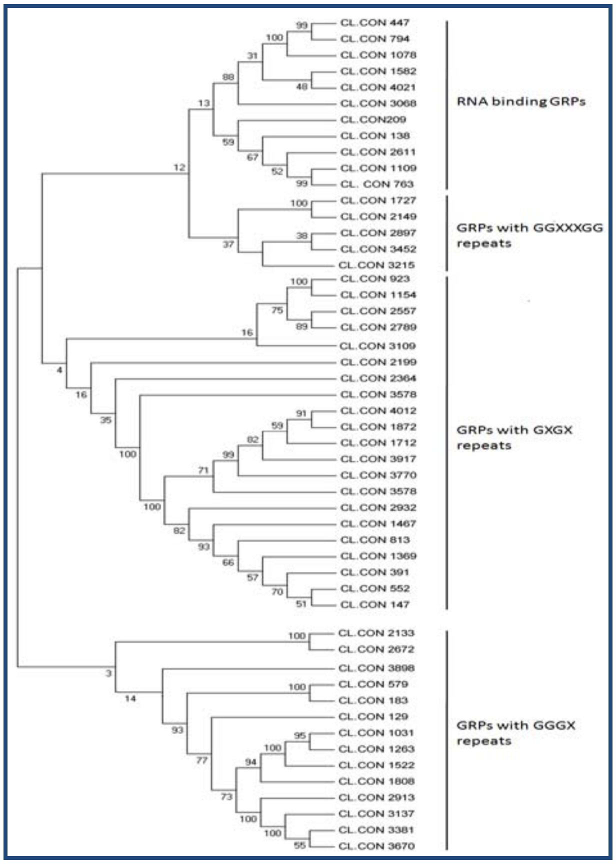

Figure 1: Unrooted dendrogram of glycine rich protein sequences encoded by Curcuma longa ESTs. The relationships were calculated using MEGA (p distance, neighbour joining method and bootstrap test with 1000 replications, pairwise deletions). The analysis was based on the ClustalX alignment of sequences

Conclusion

Turmeric is a sterile monocot, exhibit high stigmatic incompatibility and undergoes vegetative means of reproduction. It has a smaller triploid genome with n=21 that exhibit secondary polyploidy. Thus, it can be effectively used as a model plant for characterizing and analyzing various genes that are expressed in response to different stresses in asexually reproducing monocots. In the present study, we identified fifty-one *Curcuma longa* glycine-rich protein sequences from the EST database of the plant. Although several genes encoding GRPs has been isolated from different species, only a few has been cloned and characterized and their functions determined. With the availability of large number of *in-silico* derived GRPs from different plants species, greater information on the role of GRPs in diverse processes such as stress responses, signal transduction and developmental regulation can be determined. The greater diversity in structure, modulation and localization among the grp genes expressed from turmeric suggest that they are directly or indirectly involved in several physiological processes. Thus, the *in-silico* derived GRPs isolated in the present study will act as a starting point towards isolation, cloning, characterization and functional validation of different glycine rich proteins that are expressed in response to various stresses in turmeric as well as other asexually reproducing plants.

Acknowledgement
The authors are grateful to Dr. Manoj Ranjan Nayak, President, Siksha O Anusandhan University for his encouragement and support

References:

[1] Sachetto-Martins G, *et al. Biochim Biophys Acta* (2000) **1492**:1 [PMID: 10858526]
[2] Cassab GI, *Annu Rev Plant Physiol Plant Mol Biol* (1998) **49**:281 [PMID: 15012236]
[3] Steinert PM *et al. Int J Biol Macromol* (1991) 13:130 [PMID: 1716976]
[4] Alberts B, *Cell* (1998) **92**:291 [PMID: 9476889]
[5] JA Duke. *CRC handbook of medicinal plants* (2003).
[6] Selvan *et al. Indian Spices- production and utilization* (2002).
[7] Altschul SF *et al. Nucl Acids Res* (1997) **25**: 3389. [PMID: 9254694]
[8] Liu Z *et al. Science in China Series: Life Sciences* (2010) **46**: 584. [PMID:18758715]
[9] Fang RX *et al. Plant Mol Biol* (1991) **17**: 1255 [PMID: *1840687*]
[10] Guiltinan MJ & Niu X, *Plant Mol Biol* (1996) **30**:1301. [PMID: 8704137]
[11] Karlson D *et al. J Biol Chem* (2002) **277**: 35248 [PMID: 12122010]
[12] Lorkovic ZJ & Barta A, *Nuc Acids Res.* (2002) **30**: 623. [PMID: 11809873]
[13] Kevei Z *et al. Mol Plant Microbe Int.* (2002) **15**: 922. [PMID: 12236598]
[14] Goddemeir ML *et al. Plant Mol Biol.* (1998) **36**:799. [PMID:9526513]
[15] Fusaro A *et al. Genet Mol Biol.* (2001) **24**: 1.
[16] Bocca SN *et al. Genet Mol Biol.* (2005) **28**: 608.
[17] Condit CM & Meagher RB, *Plant Physiol.* (1990) **93**: 178 [PMID:16667509]
[18] Thomson JD, *et al. Nuc Acids Res.* (1997) **25**: 4876. [PMID:9396791]
[19] Kumar S *et al. Bioinformatics* (2001) **17**: 1244. [PMID:11751241]
[20] Sitnikova T *et al. Mol Biol Evol* (1995) **12**: 319. [PMID:7700156]
[21] Park AR *et al. J Biol Chem.* (2001) **276**: 26688. [PMID: 11335717]
[22] Showalter AM, *Plant Cell.* (1993) **5**: 9. [PMID: 8439747]
[23] Ting JTL *et al. Plant J* (1998) **16**: 541. [PMID: 10036772]
[24] Karlson D & Imai R, *Plant Physiol* (2003) **131**:12 [PMID: 12529510]
[25] Ni Z *et al. Mol Gen Genet* (2000) **263**: 934 [PMID:1095407]

Edited by P Kangueane
Citation: Kar *et al.* Bioinformation 8(3): 142-146 (2012)

Supplementary material:

Table 1: *Curcuma longa* ESTs encoding different classes of glycine rice protein's including the data about the homologous sequence, accession numbers and e-value.

Predicted GRPs	*Curcuma longa* contig	Signal peptide	Homologous sequence	Accession number	E.value
GRPs with GGGX repeats	CL.CON.129	Yes	Glycine rich cell wall structural protein, *Zea mays*	NM001158268	4e-22
	CL.CON.183	Yes	Glycine rich cell wall structural protein, *Zea mays*	NM001158268	3e-13
	CL.CON.579	Yes	Glycine rich cell wall structural protein, *Zea mays*	NM001158268	1e-19
	CL.CON.1031	Yes	*AtGRP6*, *Arabidopsis thaliana* glycine rich protein	P27483	1e-29
	CL.CON.1263	Yes	*AtGRP6*, *Arabidopsis thaliana* glycine rich protein	P27483	3e-32
	CL.CON.1522	Yes	*AtGRP6*, *Arabidopsis thaliana* glycine rich protein	P27483	7e-09
	CL.CON.1808	Yes	*AtGRP6*, *Arabidopsis thaliana* glycine rich protein	P27483	5e-63
	CL.CON.2133	No	*Phaseolus vulgaris* Glycine rich structural protein	P10495	1e-98
	CL.CON.2672	No	*Oryza sativa* chromosome 10 BAC OsJNBb0014l11, translated sequence	AC037426	7e-10
	CL.CON.2913	Yes	*OsGRP2*, Glycine rich protein (*Oryza sativa* Indica group)	CAA38315	9e-21
	CL.CON.3137	Yes	*OsGRP2*, Glycine rich protein (*Oryza sativa* Indica group)	CAA38315	8e-16
	CL.CON.3381	Yes	*OsGRP2*, Glycine rich protein (*Oryza sativa* Indica group)	CAA38315	9e-33
	CL.CON.3670	Yes	*OsGRP2*, Glycine rich protein (*Oryza sativa* Indica group)	CAA38315	7e-03
	CL.CON.3898	No	*Ricinus communis* grp1 precursor	XM002524779	6e-72
GRPs with GGXXXGG repeats	CL.CON.1727	Yes	*Triticum aestivun* predicted protein grp	AK333576	1e-78
	CL.CON.2149	No	*Hordeum vulgare* Glycine rich protein *HvGRP1*	Z48625	5e-26
	CL.CON.2897	No	*Hordeum vulgare* Glycine rich protein *HvGRP1*	Z48625	2e-12
	CL.CON.3452	No	Barley *grp1*	X52580	4e-39
	CL.CON.3215	No	Barley *grp1*	X52580	4e-16
GRPs with GXGX repeats	CL.CON.147	Yes	*OsGRP2*, Glycine rich protein (*Oryza sativa* Indica group)	CAA38315	2e-39
	CL.CON.391	Yes	*OsGRP2*, Glycine rich protein (*Oryza sativa* Indica group)	CAA38315	2e-41
	CL.CON.552	Yes	*OsGRP2*, Glycine rich protein (*Oryza sativa* Indica group)	CAA38315	2e-32
	CL.CON.813	No	*Zea mays*, aluminium induced grp	AAB86493	5e-19
	CL.CON.923	No	*Petunia hybrida grp1* protein	X04335	8e-13
	CL.CON.1154	No	*Petunia hybrida grp1* protein	X04335	8e-24
	CL.CON.1369	No	*Zea mays*, aluminium induced grp	AAB86493	6e-12
	CL.CON.1467	No	*Zea mays*, aluminium induced grp	AAB86493	5e-43
	CL.CON.1712	Yes	*Medicago sativa* cold-drought regulated grp	L03708	7e-63
	CL.CON.1872	Yes	*Medicago sativa* cold-drought regulated grp	L03708	7e-26
	CL.CON.2199	No	*AtGRP1*, glycine rich protein	S47405	3e-14
	CL.CON.2364	No	*AtGRP1*, glycine rich protein	S47405	3e-37
	CL.CON.2557	No	*Ricinus communis* GRP35	XM002529873	8e-22
	CL.CON.2789	No	*Ricinus communis* GRP35	XM002529873	8e-41
	CL.CON.2932	No	*Zea mays*, aluminium induced grp	AAB86493	6e-33
	CL.CON.3109	No	*Medicago truncatula* hypothetical protein	XM003605901	5e-34
	CL.CON.3578b	No	*AtGRP1*, glycine rich protein	S47405	2e-11
	CL.CON.3578	-	*Brassica oleracea* glycine rich protein	Z74892	6e-23
	CL.CON.3770	-	*Brassica napus* GRP22	Z15045	5e-89
	CL.CON.3917	-	*Brassica napus* GRP22	Z15045	5e-63
	CL.CON.4012	No	*Zea mays*, aluminium induced grp	AAB86493	5e-36
RNA binding GRPs	CL.CON.138	No	*Medicago sativa GRBP*	AAF06329	3e-74
	CL.CON.209	No	*Sorghum bicolor grbp1*	AAG23220	2e-17
	CL.CON.447	No	*Oryza sativa* Japonica group, putative RNA binding glycine rich protein	AAM19060	1e-32
	CL.CON.763	No	ABA inducible grbp, *Zea mays*	ACG28088	2e-51
	CL.CON.794	No	*Oryza sativa* Japonica group, putative RNA binding glycine rich protein	AAM19060	2e-17
	CL.CON.1078	No	*Oryza sativa* Japonica group, putative RNA binding glycine rich protein	AAM19060	1e-23
	CL.CON.1109	No	ABA inducible grbp, *Zea mays*	ACG28088	4e-34
	CL.CON.1582	No	*Oryza sativa* Japonica group, putative RNA binding glycine rich protein	AAM19060	1e-32
	CL.CON.2611	No	ABA inducible grbp, *Zea mays*	ACG28088	2e-26
	CL.CON.3068	No	*Oryza sativa* Japonica group, putative RNA binding glycine rich protein	AAM19060	1e-33
	CL.CON.4021	No	*Triticum aestivum* glycine rich RNA binding protein	BAF30986	3e-21