Survey of glycine-rich proteins (GRPs) in the Eucalyptus expressed sequence tag database (ForEST)

Silvia Nora Bocca¹, Claudia Magioli¹, Amanda Mangeon¹, Ricardo Magrani Junqueira¹, Vanessa Cardeal¹, Rogério Margis¹,² and Gilberto Sachetto-Martins¹

¹Universidade Federal do Rio de Janeiro, Instituto de Biologia, Departamento de Genética, Laboratório de Genética Molecular Vegetal, Rio de Janeiro, RJ, Brazil.
²Universidade Federal do Rio de Janeiro, Instituto de Química, Departamento de Bioquímica, Rio de Janeiro, RJ, Brazil.

Abstract

The occurrence of quasi-repetitive glycine-rich peptides has been reported in different organisms. Glycine-rich regions are proposed to be involved in protein-protein interactions in some mammalian protein families. In plants, a set of glycine-rich proteins (GRPs) was characterized several years ago, and since then a wealth of new GRPs have been identified. GRPs may have very diverse sub-cellular localization and functions. The only common feature among all different GRPs is the presence of glycine-rich repeat domains. The expression of genes encoding GRPs is developmentally regulated, and also induced, in several plant genera, by physical, chemical and biological factors. In addition to the highly modulated expression, several GRPs also show tissue-specific localization. GRPs specifically expressed in xylem, phloem, epidermis, anther tapetum and roots have been described. In this paper, the structural and functional features of these proteins in Eucalyptus are summarized. Since this is the first description of GRPs in this species, particular emphasis has been given to the expression pattern of these genes by analyzing their abundance and prevalence in the different cDNA-libraries of the Eucalyptus Genome Sequencing Project Consortium (ForEST). The comparison of GRPs from Eucalyptus and other species is also discussed.

Key words: glycine-rich, GRP, Eucalyptus.

Introduction

Glycine-rich proteins (GRPs) are characterized by the presence of domains that show little sequence conservation and are highly enriched in residues of the amino acid glycine. Typically, these Glycine-rich domains are arranged in (Gly)n-X repetitions. Although the first genes encoding GRPs have been isolated from plants, proteins with characteristic repetitive glycine stretches have been reported in a wide variety of organisms from cyanobacteria to animals (reviewed in Sachetto-Martins et al., 2000).

The structure and modulation of plant GRP genes have been intensively investigated showing that they are highly regulated during development as well as under the influence of several external stimuli. Also, in many cases, their expression pattern was demonstrated to be tissue-specific. These characteristics were the most intensively studied aspects of GRP genes since they point to the possible biotechnological application of their promoters.

Since the first reports describing plant GRPs as cell wall associated proteins (Showalter, 1993), many other GRPs with different domain organizations and sub-cellular localizations appeared in the literature. This diversity led to the concept that GRPs should not be considered as a family of related proteins but as a wide group of proteins that share a common structural domain (Sachetto-Martins et al., 2000).

The diverse but highly specific expression pattern of grp genes, taken together with the distinct sub-cellular localization of some GRP groups, clearly indicate that these proteins are implicated in several independent physiological processes (Condit, 1993; Keller and Heierli, 1994; Sachetto-Martins et al., 1995; Magioli et al., 2001; Franco et al., 2002). Based on what is known about their general architecture, sequence motifs, sub-cellular localization, and gene expression pattern and modulation, some inferences can be made regarding their function.
GRPs can be classified into four major groups (Figure 1) based on their primary structure (reviewed in Sachetto-Martins et al., 2000 and Fusaro et al., 2001). GRPs from class I are known as classic GRPs. They may contain a signal peptide followed by a glycine-rich region with GGGX repeats. A structural function is attributed to proteins of this class due to their cell wall localization (Cassab, 1998). The class II GRPs may or may not have a signal peptide and contain a glycine-rich region followed by a cysteine-rich region at their C-terminus. For one member of this family, AtGRP-3, this cysteine-rich domain has been shown to interact with cell wall associated receptor kinases (WAKs) (Park et al., 2001). The class III GRP contains proteins with lower glycine content that show a great diversity of structures. The best known proteins from this class are oleosin GRPs. Oleosins are alkaline proteins on the surface of oil bodies in plants. They play a structural role in stabilizing the triacylglycerols of the oil bodies together with the phospholipid layer. Previous works demonstrate that many of the major pollen coat proteins are derived from an endoproteolytic cleavage of oleosin GRPs that originally accumulate within the large cytoplasmatic lipid bodies of tapetal cells (Ferreira et al., 1997; Murphy et al., 2001). GRPs from class IV are RNA-binding GRPs. Those GRPs may contain, besides the glycine-rich region, several motifs which include RNA-recognition motif, cold-shock domain and zinc fingers (Fusaro et al., 2001).

In this article, a search for GRPs in the Eucalyptus transcriptome is reported. Several GRPs were identified and classified into the major groups previously established. The survey was extended to proteins that, despite not being considered canonical GRPs, contain domains of limited extension that are rich in glycine.

Materials and Methods

Sequence data, alignment and phylogenetic analysis

Protein sequences of reported plant GRPs were used to query the ForEST expressed sequence tag (EST) database with the TBLASTN algorithm (Altschul et al., 1997). Since glycine-rich domains are low complexity sequences, the TBLASTN default parameters were used without filtering the query for low compositional complexity. The complete list of sequences used as baits include the 86 proteins reviewed in Sachetto-Martins et al. (2000), 8 sequences recently described from a complete survey of Arabidopsis glycine-rich RNA binding proteins (Lorkovic and Barta, 2002), a wheat cold shock domain GRP (Karlson et al., 2002), a Pinus taeda cell wall GRP (Allona et al., 1998),

![Figure 1 - Schematic representation of the domain organization of plant glycine-rich proteins (GRPs).](image-url)
Arabidopsis UBA2 (Lambermon et al., 2002) and 4 Arabidopsis cold shock domain GRPs (Karlson and Imai, 2003). Additionally, several GRP sequences recently identified from a complete analysis of a sugarcane EST database were also selected to be used as baits (Fusaro et al., 2001). These sugarcane sequences belong to each of the different GRP classes and were chosen for being the less similar to other published GRPs among the complete sugarcane set. All GRP clusters found in Eucalyptus libraries were translated to obtain their putative protein sequences. When an evident frameshift was observed in the translation of the ORFs by an apparent sequencing error, a manual edition of the sequences was performed. 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 possible presence of a signal peptide in the sequences was predicted with the signalP server (http://www.cbs.dtu.dk/services/SignalP).

Multiple alignments of proteins deduced from the ForEST clusters and bait sequences were performed using the ClustalW program (Thompson et al., 1994). Unrooted trees were calculated using the Molecular Evolutionary Genetics Analysis (MEGA) software (Kumar et al., 2000). The neighbor-joining and p-distance method were used with the pairwise deletion option for the treatment of amino acid gaps during the multiple alignment GRPs. For construction of the phylogenetic tree the confidence levels for the nodes were determined with 2000 replications using the Internal Branch test (Sitnikova et al., 1995).

**Eucalyptus cDNA libraries**

All Eucalyptus sequences used during this work were obtained from the Eucalyptus Genome Sequencing Project Consortium (ForEST) and derived from cDNA libraries specific to different Eucalyptus tissues, organs or conditions of growth (for detailed information see https://forests.esalq.usp.br/Librariesinfo.html). BK1 (stem from 8 year old E. grandis trees), CL1 (E. grandis dark-growth callus), CL2 (E. grandis light-growth callus), FB1 (flower buds, flowers and fruits), LV1 (young plant leaves), LV2 (leaves from adult trees with deficiency in phosphorous, boron), LV3 (leaves colonized by Thyrinteina), RT2 (roots from young plants), RT3 (roots from green houses cultivated young plants), RT4 (roots from water stress resistant young plants), RT5 (roots from water stress susceptible young plants), RT6 (roots from frost resistant and susceptible trees), SL1 (dark growth E. grandis seedlings exposed to 3 h of light), SL4 (dark growth E. globulus seedlings), SL5 (dark growth E. saligna seedlings), SL6 (dark growth E. urophylla seedlings), SL7 (dark growth E. grandis seedlings), SL8 (dark growth E. camaldulensis seedlings), ST1 (stem from young healthy plants), ST2 (stem from young plants susceptible to water stress, mRNAs between 0.6 to 2 kb), ST5 (stem from young healthy plants), ST6 (stem from young plants susceptible to water stress, mRNAs between 0.8 to 3 kb), ST7 (stem from frost-resistant and susceptible trees), WD2 (E. grandis wood).

**Table 1** - Distribution of glycine-rich protein genes on ForEST database.

| General characteristics | Representative member | Number of genes published | Number of SUCEST clusters | Number of ForEST clusters |
|-------------------------|-----------------------|---------------------------|---------------------------|---------------------------|
| GGGX repeats, signal peptide, cell wall or membrane located | *Pv* GRP1.8 | 20 | 37 | 30 |
| GGGXXXGG repeats and/or signal peptide and C-terminal cysteine-rich domains homologues to nodulins | *Ar* GRP-3 | 11 | 8 | 9 |
| GXGX repeats, lower glycine content | - | 20 | 20 | 46 |
| GXGX repeats, lower glycine content and mixed pattern of repeats | - | | | 18 |
| GXGX repeats Proteins with lower glycine content with similarities to dehydrin | - | | | 15 |
| Oleosin-GRP, Oleosin conserved sequenceTapetal-specific expression | *Ar* OlnB-2 | 12 | 0 | 0 |
| RNA-binding GRP with RRM and GGYGG repeats | MA16 | 29 | 62 | 16 |
| RNA-binding GRP with RRM, CCHC zinc finger and GGYGG repeats | RZ-1 | 2 | 11 | 2 |
| RNA-binding GRP with cold-shock domain, CCHC zinc fingers and GGYGG repeats | *Ar* GRP-2 | 7 | 10 | 4 |
| RNA-binding GRP with multiple RRM motifs | SCCCLR1C01G05.g | 0 | 2 | 5 |
| GRP with nucleic acid binding domains | - | | | 8 |
| Total | | 101 | 150 | 153 |
Results and Discussion

Distribution of glycine-rich proteins genes on ForEST database

GRPs were previously subdivided into four major groups according to the presence of conserved domains and the pattern of sequence repeats. The four different classes of GRPs are shown in Table 1 and Figure 1. Three groups are based on the pattern of the glycine-rich repeats (class I, GGGX; class II, GXXXGG; class III, GXGX) and the two other groups are based on the type of functional conserved motif (one sub-group from class III, the oleosin glycine-rich proteins and class IV, the RNA-binding GRPs).

The distribution of each EST sequence between the different ForEST libraries was also analyzed (see Tables 2 to 10). The ForEST database comprises 123,889 EST sequences, arranged in 33,080 clusters. These EST sequences (reads) came from 19 different cDNA libraries constructed from different plant tissues under different culture conditions. Since several GRP genes present tissue-specific expression in other plants, the distribution of the reads from each cluster per library was analyzed. All clusters that were found in only one or two libraries were considered as pre-

Table 2 - Eucalyptus ESTs encoding GRPs with GGGX repeats.

| Eucalyptus cluster | Obs | Signal peptide | Homologous sequence | Accession number | e value | Library expression pattern |
|--------------------|-----|----------------|---------------------|------------------|---------|----------------------------|
| EGBGSL1020F05.g²   | -   | Petunia hybrid grp-1 | X04335              | 2e-09            | ST6, SL1 |
| EGBMFB1134F04.g²   | No  | Arabidopsis thaliana unknown protein | BT000113 (At2g36120) | 5e-22            | FB1     |
| EGCEFB1016H06.g²   | Gly-His rich | Brassica napus GRP22 | Z15045              | 1e-44            | FB1 (2X), LV2 |
| EGCEST2224F07.g²   | Gly-His rich | Phaseolus vulgaris GRP 1.8 | X13596              | 2e-69            | FB1 (3X), RT6 (2X), ST2 (12X), ST6 (11X) |
| EGEPSF6172E09.g²   | No  | Medicago truncatula clone mth2-10a12 | AC146308            | 7e-69            | SL5, ST6 |
| EGEQKB1002G03.g²   | No  | Hordeum vulgare grp | X52580              | 6e-20            | BK1, LV2, LV3, SL1 (2X), ST6 |
| EGEQLV2202B05.g²   | No  | Nephrilis madagascariensis flagelliform silk protein | AF218623            | 8e-35            | BK1, CL1, LV2 (22X), LV3 (4X), RT6 (2X), SL1 (2X), SL4 (2X), ST2 (3X) |
| EGEQRT5001F09.g²   | Gly-His rich | Arabidopsis thaliana unknown protein | AY136328 (At2g36120) | 1e-43            | RT5 (4X), ST (2X) |
| EGEQRT5002G04.g²   | Gly-His rich | Hordeum vulgare grp | X52580              | 1e-27            | FB1 (2X), RT5 (2X), SL1 , ST2 (2X), ST6 |
| EGEQRT5200A04.g²   | Gly-His rich | Hordeum vulgare grp | X52580              | 2e-33            | CL1, FB1 (4X), RT3 (5X), RT4 (2X), RT5 (2X), RT6, SL0, SL1 (29X), SL4 (3X), SL7, SL8, ST2 (5X), ST6 (10X), WD2 |
| EGEQST2205H11.g⁴   | Gly-His rich | Arabidopsis thaliana protease inhibitor/lipid transfer protein | NM_179606 (At2g05440) | 4e-27            | BK1 (2X), ST2 (3X) |
| EGEQWD2247G05.g⁴   | Yes  | Arabidopsis thaliana protease inhibitor/lipid transfer protein | NM_104929 (At1g62500) | 5e-55            | WD2     |
| EGEZRT5003A03.g²   | Gly-His rich | Arabidopsis thaliana unknown protein | BT000113 (At2g36120) | 7e-45            | RT5 (5X), SL7, ST2 (2X), ST6 (4X) |
| EGEZSL1043A10.g²   | No  | Arabidopsis thaliana ATGRP-5 | S47414 (At3g20470) | 3e-30            | SL1, ST7 |
| EGEZST6039B04.g⁶   | Gly-His rich | Brassica napus GRP22 | Z15045              | 6e-58            | ST6     |
| EGEZWD2203C11.g⁶   | Gly-His rich | Brassica napus GRP22 | Z15045              | 8e-46            | WD2     |
| EGEJEL2108E07.g⁷   | Yes  | Lycopersicon esculentum glycine-rich protein (clone wM) | X55688              | 5e-19            | SL5, CL1 (4X) |
| EGJFFB1008B03.g²   | No  | Oryza sativa putative glycine-rich protein (OJ1174_D05.13) | NM_189553            | 3e-15            | FB1     |
| EGJMLV2235G09.g²   | No  | Nicotiana tabacum grp | X74106              | 3e-26            | LV2     |
### Table 2 (cont.)

| Eucalyptus cluster       | Obs          | Signal peptide | Homologous sequence                  | Accession number | e value | Library expression pattern |
|--------------------------|--------------|----------------|--------------------------------------|-------------------|---------|---------------------------|
| EGJST6274H06.g           | No           |                | Hordeum vulgare grp                  | X52580            | 1e-32   | ST6                       |
| EGMCST7237H08.g          | Gly-His rich |                | Arabidopsis thaliana unknown protein  | BT000113          | 3e-39   | ST7                       |
| EGRFT6265E06.g           | No           |                | Nephila clavipes flagelliform silk protein | AF027973         | 2e-41   | ST6                       |
| EGCBST2089E03.g          | Yes          |                | Phaseolus vulgaris GRP 1.0            | X13595            | 2e-39   | RT6, SL7, ST2, ST6        |
| EGUTBK1007D04.g          | Gly-His rich |                | Arabidopsis thaliana unknown protein  | BT000113          | 2e-51   | BK1 (2X), RT3             |
| EGAGST6080E06.g          | No           |                | Nephila clavipes flagelliform silk protein | AF027973         | 2e-51   | ST6                       |
| EGBMFB1134E09.g          | No           |                | Medicago truncatula clone mth2-15i12  | AC130809          | 6e-20   | CL1, FB1, LV3             |
| EGACST7207C06.g          | No           |                | Nephila clavipes flagelliform silk protein | AF218622         | 6e-47   | ST6, ST7 (2X)             |
| EGABLVL2284A04.g         | -            |                | Anopheles gambiae clone FK0AAA48AC02  | BX032251          | 2e-31   | LV2                       |
| EGBGFB1050C05.g          | No           |                | Nephila clavipes flagelliform silk protein | AAC38847         | 9e-33   | FB1                       |
| EGCBSL4285H01.g          | -            |                | Vigna unguiculata gripe              | X87948            | 3e-15   | SL4                       |

*Incomplete sequence, a-Edited.*

### Table 3 - Eucalyptus ESTs encoding GRPs with C-terminal domains rich in cysteine.

| Eucalyptus cluster       | Signal peptide | Homologous sequence                  | Accession number | e value | Library expression pattern |
|--------------------------|----------------|--------------------------------------|-------------------|---------|---------------------------|
| EGEQSL1007D03.g          | Yes            | Citrus unshiu gene for glycine-rich protein | AB007818          | 2e-31   | CL1 (6X), SL0 (3X), SL1 (49X), SL4 (8X), SL5 (45X), SL7 (38X), SL8 (9X), WD2 |
| EGCESL5205D06.g          | Yes            | Citrus unshiu gene for glycine-rich protein | AB007818          | 9e-31   | LV2, SL5, SL6, SL7 (2X)   |
| EGUTSL6225D10.g          | Yes            | Citrus unshiu gene for glycine-rich protein | AB007818          | 2e-30   | SL4, SL6 (3X), SL8 (3X)   |
| EGBMSL6209D07.g*         | Yes            | Citrus unshiu gene for glycine-rich protein | AB007818          | 6e-29   | SL6                       |
| EGUTSL7221B03.g          | Yes            | Citrus unshiu gene for glycine-rich protein | AB007818          | 2e-29   | SL7 (3X)                  |
| EGUTSL6223E11.g          | Yes            | Citrus unshiu gene for glycine-rich protein | AB007818          | 4e-29   | SL6                       |
| EGJFSL4205F03.g          | Yes            | Citrus unshiu gene for glycine-rich protein | AB007818          | 1e-29   | SL4 (7X)                  |
| EGEPLVL2297H06.g         | No             | Nephila clavipes flagelliform silk protein (Flag) gene | AF218621          | 1e-22   | LV2                       |
| EGEQST7201F02.g          | Yes            | Sus scrofa clone TP23 basic proline-rich protein | AY035847          | 3e-20   | SL4 (3X), ST7             |

*Edited.*
| Eucalyptus cluster                  | Signal peptide | Homologous sequence                     | Accession number | e value | Library expression pattern |
|------------------------------------|----------------|-----------------------------------------|------------------|---------|----------------------------|
| Gly-His rich repeats (GHGH)        |                |                                         |                  |         |                            |
| EGEZRT4203B02.g                    | yes            | Panax ginseng GBR5                      | AF485332         | 2e-14   | CL2, RT3 (3X), SL7, RT4    |
| EGBGRTE6259A06.g                   | yes            | Panax ginseng GBR5                      | AF485332         | 7e-13   | SL4, CL2 (2X), RT6         |
| EGABSL6211E06.g                    | yes            | Panax ginseng GBR5                      | AF485332         | 2e-12   | SL6                        |
| EGEQSL1006B03.g                    | yes            | Capsella bursa-pastoris                 | AF180444         | 8e-15   | CL1 (2X), SL1 (12X), SL6 (4X), SL7 (2X), SL8 (2X) |
| EGCBSL4283H08.g                    | No             | Capsella bursa-pastoris                 | AF180444         | 6e-15   | SL4 (2X)                   |
| EGJMSL7045D12.g                    | yes            | Capsella bursa-pastoris                 | AF180444         | 6e-15   | SL7                        |
| EGEQFB1201C05.g                    | yes            | Capsella bursa-pastoris                 | AF180444         | 1e-14   | SL5, SL1 (10X), SL0, RT5, SL8, CL1 (2X), SL7(8X), ST6(3X), RT6(3X), RT3(2X) |
| EGEQRT5200B04.g                    | yes            | Citrus unshiu gene for glycine-rich protein | AB007818       | 7e-16   | SL5, SL1 (10X), SL0, RT5, SL8, CL1 (2X), SL7(8X), ST6(3X), RT6(3X), RT3(2X) |
| EGEQSL1054D09.g                    | yes            | Quercus robur phase-change related protein |               | 3e-14   | CL2, ST2, SL1, ST6 (2X), LV2 (5X) |
| EGACSL5245C06.g                    | yes            | Medicago sativa cold and drought-regulated protein (corA) | L03708          | 1e-14   | SL5                        |
| EGSBLS7015D12.g                    | yes            | Medicago sativa cold and drought-regulated protein (corA) | L03708          | 1e-22   | SL7 (2X)                   |
| EGBFR76224A05.g                    | No             | Caenorhabditis elegans putative nuclear protein (4B256) | NM_067535       | 1e-15   | RT6                        |
| EGJFFB11118H02.g                   | No             | Caenorhabditis elegans putative nuclear protein (4B256) | NM_067535       | 2e-14   | FB1                        |
| EGBMRT3131G07.g                    | No             | Caenorhabditis elegans putative nuclear protein (4B256) | NM_067535       | 6e-14   | ST6(2X), RT3 (5X)          |
| EGUTSL1042G04.g                    | No             | Eucalyptus globulus bicostata symbiont (F00078), ESTun052 | LA1713          | 5e-21   | SL5, SL4, RT3 (4X), RT6 (2X), SL7 (2X), SL1 |
| EGEQRT3200A07.g                    | No             | Eucalyptus globulus bicostata symbiont (F00078), ESTun052 | LA1713          | 5e-21   | RT3 (7X), SL1 (5X), SL4 (3X) |
| EGCBRT6017B04.g                    | No             | Eucalyptus globulus bicostata symbiont (F00078), ESTun052 | LA1713          | 2e-21   | RT6 (3X)                   |
| EGJMST2266A01.g                    | No             | Bacillus anthracis virB gene             | AF238885         | 2e-11   | ST2                        |
| EQQHSL1102H06.g*                   | -              | Caenorhabditis elegans putative protein  | NM_171608       | 7e-23   | SL1 (2X)                   |
| EGRFSL4277F01.g                    | No             | Caenorhabditis elegans putative protein  | NM_171608       | 5e-24   | SL4(2X)                    |
| EGCCVL2224A10.g                    | No             | Danio rerio clone MGC:66347              | BC055611         | 3e-14   | CLV2                       |
| EGEZSL7230B03.g*                   | No             | Phaseolus vulgaris PVGRP1.8              | X13596           | 3e-52   | SL6, SL7, WD2              |
| EGQSTI1001D01.g                    | No             | Medicago sativa cold-inducible protein   | AF411552         | 2e-10   | SL4, ST1                   |
| EGEZRT6273H01.g*                   | -              | Caenorhabditis elegans putative protein  | NM_171608       | 9e-21   | CL2, LV3, RT6 (2X)         |
| EGEZRT4203E09.g                    | No             | Rattus norvegicus similar to AT hook motif, putative (LOC289506) | XM_223239       | 3e-26   | RT6 (2X), ST6 (2X), FB1 (3X), RT3 (4X), RT4, LV2 (3X) |

| Gly-Pro rich repeats (GPGP)        |                |                                         |                  |         |                            |
|------------------------------------|----------------|-----------------------------------------|                  |         |                            |
| EGBFFB1043B08.g                    | No             | Drosophila melanogaster                  | NM_141224        | 2e-28   | FB1                        |
| EGCCRT6012A07.g*                   | No             | Arabidopsis thaliana putative protein    | AY136343         | 1e-40   | RT6, ST6                   |
| EGEZRT5202F02.g*                   | -              | Volvox carteri f. nagariensis pheorophin-dzL1 protein | AUSJ29250        | 2e-20   | RT5 (2X)                   |

*Incomplete sequence.
### Table 5 - Eucalyptus ESTs encoding GRPs with lower glycine content (GXGX).

| Eucalyptus cluster | Signal peptide | Domain | Homologous sequence | Accession number | e value | Library expression pattern |
|--------------------|----------------|--------|---------------------|------------------|---------|---------------------------|
| EGRFRT302F11.g      | -              | -      | Botrytis cinerea strain T4 | AL116868          | 2e-09   | RT3                       |
| EGEFBL2124H05.g     | No             | -      | Oryza sativa (japonica cultivar-group) cDNA clone:002-104-F11 | AK064219          | 1e-47   | SL1 (2X), FB1              |
| EGEZLS8267A07.g     | No             | -      | Oryza sativa (japonica cultivar-group) cDNA clone:002-104-F11 | AK064219          | 3e-47   | SL8                       |
| EGJMST2266A01.g     | No             | His-rich | Bacillus cereus strain ATCC 43881 putative VrrB gene | AF238888          | 2e-12   | ST2                       |
| EGJEST2212B07.g     | Yes            | -      | Arabidopsis thaliana unknown protein | BT000091 (Art4g30460) | 8e-36   | ST6 (3X), ST2              |
| EGCCRT6008E03.g     | No             | Asp-rich, Glu-rich | Arabidopsis thaliana clone RAFL08-11-G02 (R11308) unknown protein | BT000731 (Art1g47970) | 2e-37   | RT6 (2X)                  |
| EGEPRTE2260F11.g    | -              | -      | Mus musculus per-hexamer repeat gene 5 (Phx5) | NM_008836          | 2e-22   | RT6 (3X)                  |
| EGCCFB1223G04.g     | No             | Pro-rich | Glycine max SRC2 | AB000130          | 6e-62   | RT6 (2X), SL5, FB1 (2X), BK1, RT3, SL5 (2X), LV2, SL1 |
| EGEQRT3200C10.g     | No             | Met-rich, Pro-rich | Oryza sativa (japonica cultivar-group) cDNA clone:001-039-H07l | AK04798           | 2e-65   | ST2 (3X), ST6 (5X), RT6 (4X), CL1 (7X), RT3 (3X), SL5 (2), ST7, LV2 (3X), LV1, ST1 |
| EGCESL5057B05.g     | No             | Pro-rich | Arabidopsis thaliana nuclear protein ZAP-related | NM_180919 (Art5g62760) | 2e-14   | SL5 (2X)                  |
| EGJMLV2236E04.g     | No             | -      | Prunus persica abscisic stress ripening-like protein | AF317062          | 5e-57   | RT6, LV2 (2X), CL1, ST7    |
| EGMCRRT3145H05.g    | -              | -      | Arabidopsis thaliana unknown protein | AY123003 (Art3g13224) | 6e-26   | RT3, LV3, SL6 (2X)        |
| EGSBRT3313G03.g     | No             | Arg-rich | Oryza sativa (japonica cultivar-group) cDNA clone:J033084H19 | AK102114          | 1e-88   | ST6 (3X), SL1 (4X), FB1 (2X), LV2, LV3, RT3, ST2 |
| EGQQL4009C08.g      | No             | Ala-rich | Human herpesvirus 6 mRNA for ORF99, immediate early 2 | AB075776          | 4e-06   | SL4 (4X)                  |
| EGBMRT3131F10.g     | No             | Pro-rich, His-rich | Oryza sativa (japonica cultivar-group) cDNA clone:J023110G03 | AK071715          | 8e-77   | WD2 (2X), RT3 (3X), SL1 (2X), ST2 (2X), SL7 |
| EGJMLV2226G07.g     | No             | -      | Mus musculus cDNA clone MGC:12025 | BC005782          | 6e-65   | LV2                       |
| EGUTSL1042C12.g     | No             | -      | Petunia hybrida grp-1 | X04335           | 4e-14   | LV2 (2X), SL1 (2X)        |
| EGUTFB1293G12.g     | Yes            | -      | Lycopersicon esculentum Tm5 gene | X95262           | 9e-15   | FB1                       |

a-Incomplete sequence, b-Edited.

### Table 6 - Eucalyptus ESTs encoding GRPs with a mixed pattern of repeats.

| Eucalyptus cluster | Sequence repeat | Signal peptide | Homologous sequence | Accession number | e value | Library expression pattern |
|--------------------|-----------------|----------------|---------------------|------------------|---------|---------------------------|
| EGBFSL1078H02.g    | GYPPX/GGX/GXGX/ GHS | No             | Oryza sativa (japonica cultivar-group) cDNA clone:023104E19 | AK071554          | 3e-63   | SL1                       |
| EGCEST2256B04.g    | GYPPX/GGX/GXGX/ GSH/GKX | No             | Oryza sativa (japonica cultivar-group) cDNA clone:J023104E19 | AK071554          | 1e-63   | CL2 (2X), SL1 (2X), FB1 (4X), LV2, ST2 (3X), ST6 (2X), RT3, WD2, ST7, LV1 |
dominantly expressed in a tissue-specific pattern. Several clusters identified in this search presented this characteristic.

The search for genes encoding GRPs in Eucalyptus resulted in 153 potential genes (clusters) that were distributed in the classes mentioned above (Table 1). While no sequences were found to present the characteristic pattern of repeats GGXXXGG, our search retrieved a number of other Eucalyptus sequences having a mixed pattern of repeats (Table 6). Among these sequences, clusters with conserved motifs that characterize dehydrins were found (Table 7). As expected for an angiosperm with wet-type stigmas, no Eucalyptus ESTs with similarity to oleosin-GRPs were found.

The analysis was also extended to twelve other proteins that contain domains of limited extension that are rich in glycine even though these domains represented a small proportion of the complete protein (Table 10).

Table 6 (cont.)

| Eucalyptus cluster | Sequence repeat | Signal peptide | Homologous sequence | Accession number | e value | Library expression pattern |
|--------------------|----------------|----------------|---------------------|-----------------|---------|---------------------------|
| EGEQLV2202G11.g    | GYPPX/GXGX/GHS/GKH | No            | Oryza sativa (indica cultivar-group) GPRP | AY348312         | 3e-57   | LV2                       |
| EGJEB1029A04.g     | GYPPX/GXGX     | No            | Daucus carota GPR A3 | X72383          | 4e-35   | FB1 (2X)                  |
| EGCEST2257B12.g    | GYPPX/GXGX/GKF | No            | Mus musculus cDNA MGC:12025 | BC005782       | 2e-47   | ST2 (2X), ST6 (3X)        |
| EGBMSL1091B03.g    | GYPPX/GXGX     | No            | Mus musculus cDNA MGC:12025 | BC005782       | 1e-40   | SL1                       |
| EGEZSL8268H10.g    | GYPPX/GXGX/GKX | No            | Mus musculus cDNA MGC:12025 | BC005782       | 6e-47   | SL8 (2X)                  |
| EGUTRT3113E07.g    | GYPPX/GXGX/GKF | No            | Arabidopsis thaliana GPRP | NM_121771 (At5g17650) | 2e-46 | RT3                       |
| EGEZST7214D02.g    | GYPPX/GXGX     | -             | Arabidopsis thaliana GPRP | NM_121771 (At5g17650) | 2e-33 | ST7                       |
| EGCCSL1005C02.g    | GYPPX/GXGX     | -             | Arabidopsis thaliana GPRP | NM_121771 (At5g17650) | 1e-35 | LV2 (2X), SL1             |
| EGUTFB1292D04.g    | GGP            | No            | Arabidopsis thaliana expressed protein | NM_123319 (At5g39570) | 1e-31 | FB1 (2X)                  |
| EGCBCL1215C03.g    | GXGX/GGX Glu-rich, Asp-rich | No | Pinctada fucata Shell matrix protein | AB094512 | 5e-31 | CL1                       |
| EGZFJ21240H06.g    | GGGX/GXGX     | Yes           | Arabidopsis thaliana cDNA GSLTSIL53ZA11 | BX841600 (At4g1620) | 3e-41 | RT6, SL1, ST6, FB1        |
| EGEQST7200H05.g    | GXGX/GGX     | -             | Arabidopsis thaliana cDNA GSLTFB20ZF11 | BX822222 (At5g07560) | 8e-54 | ST7                       |
| EGECL1208E07.g     | GGX/GXGX      | Yes           | Lycopersicon esculentum GPR wM | X55688       | 5e-19 | SL5, CL1 (4X)             |
| EGRFLV3242F08.g    | GGX            | Yes           | Chlamydomonas reinhardtii agglutinin (SAG-1) | AY450930 | 4e-14 | LV3                       |
| EGEZST7211E06.g    | GGX/GXGX      | No            | Arabidopsis thaliana cDNA GSLTL521ZE06 | BX827206 (At4g13530) | 3e-14 | CL1, ST7                  |
| EGACRT3321F02.g    | GGGX/GGX HMA_2 domain, Asp-rich, Asp-rich, Gln-rich, Met-rich, Pro-rich | No | Arabidopsis thaliana heavy-metal-associate d domain-containing protein | NM_121914 (At5g19090) | e-130 | SL5, RT6 (3X), FB1, ST6, RT3 (4X), SL4 (2X), ST2 |

a-Incomplete sequence.
of GRP usually has a predicted signal peptide at their N-terminal end. The best characterized protein of this class is PvGRP1.8, a structural protein from bean specifically associated with the primary cell walls of elongating protoxylem elements (Keller et al., 1989). Recent studies using antibodies against PvGRP1.8 indicated that PvGRP1.8 form a three-dimensional protein network that stabilizes the protoxylem elements (Ryser and Keller, 1992; Ryser et al., 1997 and Ringli et al., 2001).

Thirty Eucalyptus clusters with GGGX repeats were found. Several clusters (11) encode GRPs that are highly enriched in histidine, resulting in a repetition pattern GGGH (Table 2). Fourteen clusters presented an apparent tissue specific expression, with 9 being expressed exclusively in one library. Interestingly, two clusters (EGEQWD2247G05.g and EGEZWD2203C11.g) were observed only in libraries prepared from wood tissues making them interesting genes for study in relation to wood biogenesis.

As previously noted (Sachetto-Martins et al., 2000; Fusaro et al., 2001), this class of GRPs represents a rather heterogeneous set of proteins with sequence similarity limited to the repetitive glycine amino acids. The alignments obtained presented many gaps and regions with no sequence overlapping, which made the construction of a dendrogram impossible. The functional characterization of members of this class could help to establish a clear classification of these proteins.

### Eucalyptus clusters encoding GRPs with C-terminal domains rich in cysteine

Some GRP proteins are grouped together based on the similarity of their N- and C-terminal domains with soybean nodulin 24 (Sandal et al., 1992). Usually, the C-terminal end of GRPs that are similar to nodulins are cysteine-rich and the glycine-rich repeats found in these sequences are GGXXXGG with Y, H, R, N or Q as the most frequent amino acids in the tripeptide between the glycine residues (Sachetto-Martins et al., 2000).

The direct interaction of AtGRP3, a protein belonging to this class of GRPs, with the cell wall-associated kinase WAK1 was recently demonstrated. The interaction occurs between the cysteine-rich C-terminal end of AtGRP3 and the extracellular domain of WAKs (Park et al., 2001). WAK1 is a member of the WAK receptor kinase family that links the plasma membrane to the extracellular matrix (Verica and He, 2002). WAK kinases are proposed to recognize different environmental signals through the interaction of their diverse extracellular domains with cell wall molecules and transduce those signals to the cell. 

WAK1 and Atgrp-3 are both induced by salicylic acid treatment. Moreover, exogenously added AtGRP-3 up-regulates the expression of WAK1, Atgrp-3 and PR-1 in Arabidopsis protoplasts. Taken together, this data suggest that AtGRP-3 regulates WAK1 function through binding to the cell wall

### Table 7 - Eucalyptus ESTs encoding GRPs with similarity to dehydrin.

| Eucalyptus cluster | Dehydrin motif | Homologous sequence | Accession number | e value | Library expression pattern |
|--------------------|---------------|---------------------|------------------|---------|---------------------------|
| EGACST2103C05.g    | -             | Solanum commersonii DHN1 protein | Y15813          | 5e-12 | ST2                       |
| EGSBST6076H12.g    | -             | Solanum commersonii DHN1 protein | Y15813          | 2e-11 | ST6 (2X)                  |
| EGSBST2112G04.g    | -             | Citrus sinensis dehydrin (DHN) | AY297793        | 2e-17 | ST2                       |
| EGGBST2107B05.g    | -             | Herpesvirus papio EBNA1 | U23857          | 7e-15 | ST2                       |
| EGBMST6205F07.g    | -             | Arabidopsis thaliana dehydrin (RAB18) | NM_126038 (At5g66400) | 9e-21 | ST6                       |
| EGEQRT5201H10.g    | 2             | Arabidopsis thaliana dehydrin (RAB18) | AY093779 (At5g66400) | 9e-43 | SL1, ST2 (77X), ST6 (100X), SL7 (3X), SL4 (3X), RT5 (2X), FB1 (12X), ST7 (2X), LV2, SL5 |
| EGBGFB1253G12.g    | 2             | Arabidopsis thaliana dehydrin (RAB18) | AY093779 (At5g66400) | 4e-38 | ST6 (10X), FB1, SL0, ST2 |
| EGABST2226G10.g    | 2             | Arabidopsis thaliana dehydrin (RAB18) | AY093779 (At5g66400) | 2e-32 | ST2                       |
| EGMLV2223A07.g     | 1             | Arabidopsis thaliana dehydrin (RAB18) | AY093779 (At5g66400) | 5e-30 | LV2 (2X)                  |
| EGMST16020B07.g    | 2             | Arabidopsis thaliana dehydrin (RAB18) | AY093779 (At5g66400) | 1e-30 | ST6                       |
| EGEQRT3201H04.g    | -             | Arabidopsis thaliana dehydrin (RAB18) | AY093779 (At5g66400) | 9e-21 | RT3                       |
| EGEZRT5003B09.g    | 2             | Arabidopsis thaliana dehydrin (RAB18) | AF428458 (At5g66400) | 1e-32 | ST6 (20X), ST2 (22X), SL6 (2X), RT5 |
| EGEZRT5003F10.g    | -             | Arabidopsis thaliana dehydrin (RAB18) | AF428458 (At5g66400) | 1e-27 | RT5, ST6 (4X), ST2 (5X) |
| EGUTFB1136A01.g    | 1             | Arabidopsis thaliana dehydrin (RAB18) | AF428458 (At5g66400) | 1e-27 | FB1                       |
| EGJMST2269C12.g    | 1             | Arabidopsis thaliana dehydrin (RAB18) | BT002226 (At5g66400) | 8e-28 | ST2 (5X), ST6 (5X)        |

*Edited.*
Table 8 - Eucalyptus ESTs encoding RNA-binding GRPs.

| Eucalyptus cluster | Homologous sequence | Accession number | e value | Library expression pattern |
|--------------------|---------------------|------------------|---------|---------------------------|
| **Subclass I RNA-binding GRPs** | | | | |
| EGQST2201B10.g | *Ricinus communis* glycine-rich RNA-binding protein (grp1 gene) | AJ245939 | 9e-76 | SL1, SL5 (2X), WD2, ST2 |
| EGQSL1006E12.g | *Ricinus communis* glycine-rich RNA-binding protein (grp1 gene) | AJ245939 | 9e-73 | ST2 (6X), SL1 (5X), FB1 (4X), CL1 (3X), ST6, RT3 (2X), LV2, WD2 |
| EGQSL1007A08.g | *Ricinus communis* glycine-rich RNA-binding protein (grp1 gene) | AJ245939 | 7e-69 | SL1, FB1 |
| EGQFB1203H12.g | *Ricinus communis* glycine-rich RNA-binding protein (grp1 gene) | AJ245939 | 7e-69 | FB1 |
| EGMFB1226F11.g.h | *Ricinus communis* glycine-rich RNA-binding protein (grp1 gene) | AJ245939 | 4e-60 | FB1 (4X), SL1, SL6 |
| EGQSL1050E12.g | *Ricinus communis* glycine-rich RNA-binding protein (grp1 gene) | AJ245939 | 6e-46 | LV2, SL1 (3X), SL6, SL8, SL7 |
| EGEZF1204G11.g | *Arabidopsis thaliana* putative RNA-binding protein | NM_125496 (At5g61030) | 1e-75 | ST6 (4X), LV3, FB1 |
| EGRFST6266B12.g | *Solanum tuberosum* RNA-binding protein | AY048973 | 2e-53 | ST6 |
| EGBFFB1042A06.g | *Arabidopsis thaliana* maf19_30 | AY060565 (AT5g61030) | 1e-53 | WD2, CL1, FB1 |
| EGRFST2079A03.g | *Pisum sativum* glycine rich RNA-binding protein (PsrGBP) | PSU81287 | 2e-49 | ST2 (2X), SL4 |
| EGJCL2028G10.g | *Nicotiana tabacum* RNA-binding protein | AY048972 | 1e-59 | CL2 |
| EGECCL1282G03.g.h | *Medicago truncatula* clone mth2-10p20 | AC134242 | 6e-46 | CL1 (2X) |
| EGEST2229C10.g | *Medicago truncatula* clone mth2-10p20 | AC134242 | 9e-43 | ST2 (2X), SL1, WD2, FB1, SL7 (2X) |
| EGEZF1006G11.g.h | *Glycine max* glycine-rich RNA-binding protein | AF169205 | 2e-41 | FB1 |
| EGCRT3342D03.g.h | *Neurospora crassa* strain OR74A | XM_331179 | 3e-52 | RT3 |
| EGBGL1020B05.g | *Rumex obtusifolius* putative glycine rich protein (grp gene) | AJ441311 | 9e-45 | SL1, SL4 |

**Subclass II RNA-binding GRPs**

| EGEFB1029H07.g | *Nicotiana sylvestris* RZ-1 | D28861 | 2e-58 | FB1, ST6 |
| EGBSL1048F09.g | *Arabidopsis thaliana* RNA-binding protein-like | AY114645 (At5g04280) | 2e-83 | RT3, LV3, ST2, CL1, SL5, SL1 |

**Subclass III RNA-binding GRPs**

| EGUTBK1006H11.g | *Triticum aestivum* WCSP3 | AB161683 | 1e-77 | SL4, LV3, BK1, ST2, SL1 |
| EGEPRT3325H02.g | *Triticum aestivum* WCSP3 | AB161683 | 9e-77 | CL1, SL5, ST6 (2X), RT3 |
| EGEQFB1001F04.g | *Nicotiana sylvestris* GRP2 | X60007 | 3e-53 | FB1, WD2 |
| EGJEL2215H02.g | *Nicotiana sylvestris* GRP2 | X60007 | 3e-38 | CL2 |

**Subclass IV RNA-binding GRPs**

| EGEST2228E05.g | *Arabidopsis thaliana* Ribonucleoprotein-like | AY136466 (At5g40490) | e-124 | ST6 (3X), CL2, ST7, ST2 (2X), RT6, FB1, RT3, SL6 |
| EGEQRT3201H05.g | *Arabidopsis thaliana* Ribonucleoprotein-like | AY136466 (At5g40490) | 3e-35 | SL5, RT6, CL1, LV2(2X), FB1, RT3 |
| EGACST2105B03.g | *Arabidopsis thaliana* putative RNA-binding protein | AY063846 (At3g15010) | e-131 | WD2 (3X), CL1, ST2 |
| EGCEFB1016C10.g | *Arabidopsis thaliana* putative RNA-binding protein | AY063846 (At3g15010) | e-124 | SL5 (2X), ST7 (2X), WD2 (3X), CL1 (3X), SL4, FB1 (2X), ST2 (2X), RT3 (3X), ST6, RT6 |
| EGUTLV1248B11.g | *Arabidopsis thaliana* ribonucleoprotein 1 (rnp1) | AJ303457 (At4g14300) | e-153 | CL1 (4X), LV1 |

* - Not included in phylogenetic analyse, a-Incomplete sequence, b-Edited.
domain of Wak1 and that the interaction of Wak1 with AGRP-3 occurs in a pathogenesis-related process in planta (Park et al., 2001).

Ten GRPs containing C-terminal Cys-rich end were found in the ForEST database (Table 3). None of them presents the typical pattern of repetition GGXXXGG usually found in this group of GRPs. In order to analyze the similarities of these 10 sequences with the reported GRPs that are similar to nodulins, all the sequences were aligned and an unrooted tree was constructed (Figure 2). Seven clusters were found to be more related to petunia PtGRP-2 and tobacco gGRP-8; two other are closer to a group of GRPs sequences from Medicago sativa; and one seems to be more divergent from all the previously reported sequences of this group.

**Eucalyptus clusters encoding GRPs with GXGX repeats**

This last pattern of glycine repeats, GXGX, is generally observed in GRPs with an average glycine content of 20%. Similar to the GGGX group (Table 2) this GRP group shows a high degree of structural diversity and probably contains several different types of GRPs. In Eucalyptus, forty-six different clusters were identified encoding this type of GRP (Table 4 and Table 5).

As noticed for the Eucalyptus sequences with GGGX repeats, several sequences of this group are also rich in histidine, resulting in the repetition pattern GHGH. Three other clusters show Pro/Gly-rich sequences. Sequences that in addition to the glycine-rich domains are also enriched in different aminoacids (arginine, alanine or methionine) were also found (Table 4).

A predicted N-terminal signal sequence which may reflect their possible extracellular localization was observed in twelve clusters from the GXGX Eucalyptus GRPs (Table 4 and Table 5).

As occurs with all GRPs grouped only on the basis of their pattern of repeats, most of the GXGX GRP sequences comprise a heterogeneous group of proteins with no significant sequence similarity outside the Glycine-rich repetitive domains.

It is noteworthy that 3 GRP clusters with GHGH repeats share high sequence identity with a Gly/His-rich protein of an endosymbiotic fungus of Eucalyptus (Table 4). One could speculate that those sequences may represent fungal contamination in the plant mRNA population and should be considered as possible non-plant GRPs.

In several species, cell wall associated proteins with preferential expression in vascular tissues have been reported (Showalter, 1993). GRPs localized in vascular tissues are thought to provide elasticity and tensile strength during vascular development (Cassab, 1998) and most of the wood quality-related traits are linked to the properties of the cell wall during this process. Despite the economic importance of wood biogenesis, few reports exist to date on the role of cell wall associated proteins in the development of vasculature.

A GRP with GXGX repeats from *Pinus taeda* (Allona et al., 1998; Zhang et al., 2000), as well as its proposed orthologous in *Pinus pinaster* (Le Provost et al., 2003), were found to be differentially expressed in the xylem of different wood types. It has been proposed that both *Pinus* proteins, reported as GRPs, might be involved in the determination of wood properties (Le Provost et al., 2003). However, only the *Pinus taeda* protein (AAB66348) presents high glycine content with a pattern of GXGX repetitions. The protein from *Pinus pinaster* (AAF75823) was apparently misclassified as GRP on the basis of its partial similarity with the *Pinus taeda* sequence. Searching the ForEST database with the *Pinus taeda* GRP protein sequence allowed us to identify a closely related *Eucalyptus* cluster (EGJEST2212B07.g) with 61% similarity throughout 119 aminoacids, showing a high degree of similarity to Eucalyptus glycine-rich proteins.
the Arabidopsis gene At4g30460 (Table 5). Both Pinus and Eucalyptus proteins are rich in glycine and serine and present a predicted N-terminal signal peptide as expected for a putative cell-wall protein. The high degree of conservation between the Pinus and Eucalyptus sequences indicates that the Eucalyptus cluster identified may be the Pinus taeda ortholog and that this gene is an interesting candidate to be studied due to its possible involvement in wood biogenesis in conifers and angiosperm trees.

Eucalyptus clusters encoding GRPs with a mixed pattern of repeats

In addition to the classic repeats observed in the previous described plant GRPs, the ForEST database also contains a set of GRPs with a mixed pattern of repetition (Table 6).

Ten of them encode GRPs with GXGX repeats combined with domains that contain 8 to 15 tandem repeats of the pentapeptide GYPPX (where X is usually Q). Strictly, these proteins should be considered as glycine/proline-rich proteins (GPRPs). The motif XYPPX is found in a wide variety of proteins including annexin and the carboxy tail of certain rhodopsins. The motif was proposed to form polyproline beta-turn helices but its molecular function is unknown (Matsushima et al., 1990). Eucalyptus sequences with GYPPQ repeats may be functionally related to PtaADH1 (AF101786), a proline-rich sequence from Pinus taeda recently characterized as a cell wall structural protein with GYPQ repetitions. The observation that PtaADH1 mRNA is mainly expressed in vascular tissue and that its expression is modified in different types of wood led to a proposal that it may be involved in the process of wood biogenesis (Zhang et al., 2000).

Fifteen other sequences present a mixed pattern of GGGX and GXGX repeats, sharing identity with dehydrins (Table 7). Dehydrins are classified as the late embryogenesis abundant proteins group 2 (Wise, 2003). They are also termed responsive to abscisic acid (RAB). These proteins form a subset of evolutionarily conserved glycine-rich, hydrophilic proteins induced in maturing seeds or vegetative tissues following abscisic acid treatment as well as in response to salinity, dehydration or cold stress (reviewed in Allagulova et al., 2003). Dehydrins are characterized by the presence of a highly conserved Lys-rich 15 amino acids motif that appears repeated from 1 to 12 times in the C-terminus of the protein. This dehydrin motif, referred to as the K-segment (EKKGIMDKIKEKLPG), was found in 8 out of the 15 Eucalyptus GRP clusters that present sequence similarity with dehydrins (Table 7). The same clusters also present a conserved Ser stretch that is commonly found in many dehydrins and is thought to be involved in nuclear localization. The N-terminal sequence of many proteins of this group present a third conservative sequence termed the Y-segment (V/T DEYGNP).

It is known that some dehydrins are preferentially induced under specific stresses while others have a constitutive expression. Among the Eucalyptus GRPs identified as possible dehydrins, one cluster is strikingly over-expressed in libraries of stems of plants susceptible to dehydration (EGEQRT5201H10.g). Its closest similar sequence is RAB18, an A. thaliana dehydrin strongly induced both in water-stressed and ABA-treated plants but only slightly responsive to cold (Welin et al., 1994).
Several different types of plant RNA-binding GRPs have been identified. They contain an RNA-binding motif in their N-terminal half followed by a C-terminal region rich in glycine residues. Most of these proteins have the conserved RNA-binding motif termed RRM (RNA Recognition Motif) encompassing 80-100 amino acid residues in which two short sequences, RNP-1 and RNP-2, are highly conserved regions (Alba and Pages, 1998). A different type of RNA-binding motif observed in the N-terminus of plant GRPs is the CSD (Cold-Shock Domain), with only the RNP-1 sequence conserved (Sachetto-Martins et al., 2000). In addition to their RNA-binding motifs, some GRPs contain a variable number of CCHC (CX2CX4HX4C) retroviral-like zinc-fingers inside the C-terminal glycine-rich region.

### Table 10 - Eucalyptus ESTs encoding short glycine-rich domains.

| Eucalyptus cluster | Domain | Protein length (aa) | Gly-rich domain* | Homologous sequence | Accession number | e value | Library expression pattern |
|--------------------|--------|---------------------|------------------|---------------------|------------------|--------|--------------------------|
| EGEZL1202A01.g     | 3 TPR repeats | 415 | 65 aa (51%) | *Abdomopsis thaliana* HSP associated protein like | AY059803                         | 1e-179 | ST6(4X), FB1(4X), WD2(4X), RT3(6X), LV2, CL1(2X), SL4(4X), LV1, ST2(2X), SL7, SL0 |
| EGEZSL5201D09.g    | 244 | 63 aa (36%) | *Abdomopsis thaliana* unknown protein | AF412102                         | 4e-87   | SL5(2X), CL1               |
| EGMCSL1062E05.g    | 209 | 55 aa (49%) | *Chlamydomonas reinhardtii* putative amn protein | AF509496                         | 4e-13   | SL1                       |
| EGEQFB1003D01.g    | 308 | 59 aa (64%) | *Abdomopsis thaliana* fibrillarin 2 | NM_116695                         | 1e-148  | BK1, FB1, RT3, RT4, SL1, ST2(7X), ST6(2X), WD2 |
| EGUITS6226B01.g    | ABA/WDS | 194 | 55 aa (32%) | *Prunus persica* abscisic stress ripening-like protein | AF317062                         | 1e-57   | SL6                       |
| EGCCR3330E08.g*   | - | 42 aa (40%) | *Abdomopsis thaliana* putative DEAD box RNA helicase | AL137082                         | 4e-46   | RT3(3X), SL5(2X)           |
| EGUITS2502D05.g*   | - | 65 aa (58%) | *Abdomopsis thaliana* putative ethylene-responsive DEAD box RNA helicase | NM_125706                         | 1e-32   | ST2                       |
| EGABST6008C06.g    | 185 | 42 aa (40%) | *Abdomopsis thaliana* putative DEAD box RNA helicase | NM_129813                         | 1e-32   | ST6, ST7                  |
| EGEPSL4003H05.g    | 191 | 59 aa (71%) | *Oryza sativa* cDNA clone | AK111212                         | 4e-65   | SL4                       |
| EGJMBF1115H10.g    | HLH | 324 | 92 aa (38%) | *Abdomopsis thaliana* basic helix-loop-helix (bHLH) family protein | AF367328                         | 1e-114  | WD2(2X), FB1, RT6          |
| EGCCSL4029G02.g    | 247 | 40 aa (38%) | *Eucalyptus grandis* zinc transporter | AF197329                         | 1e-90   | SL5, SL4(4X)              |
| EGEQST7200D08.g    | 249 | 32 aa (81%) | *Oryza sativa* cDNA | AK067094                         | 9e-55   | ST6, LV3, CL2, WD2(2X), CL1, SL1, ST7(2X), SL7(2X) |
| EGEZRT6214E11.g    | Cation efflux | 556 | 130 aa (35%) | *Abdomopsis thaliana* epsin N-terminal homology (ENTH) domain-containing protein | NM_180055                         | 1e-104  | RT6                       |
| EGQHST2015H03.g    | Ubiquitin | 540 | 50 aa (40%) | *Abdomopsis thaliana* putative ubiquitin protein | AY142486                         | 1e-99   | FB1, CL1, ST2(3X), SL5(3X), RT6 |
| EGEQLV2222B06.g*  | - | 48 aa (58%) | *Lilium longiflorum* mRNA for nucleotide excision repair protein | AJ002990                         | 1e-86   | LV2, ST6                  |
| EGSBRT3121C08.g    | 2 RRM | 379 aa | 63 aa (55%) | *Pisum sativum* L (clone na-481-5) | L43510                         | 1e-100  | ST6(3X), WD2, SL8(2X), RT6(2X), SL7, CL1, RT3 |

* - Incomplete sequence; b - Edited; * % of glycine in Gly-rich domain; TPR: tetratricopeptide repeat; ABA / WDS : domain present in a family of plant proteins induced by water deficit stress (WDS) or abscisic acid (ABA) stress and ripening; HLH: Helix loop helix domain; RRM: RNA Recognition Motif.
RNA binding GRPs can be classified in four different sub-classes based on the combination of the structural domains they present (Figure 1, Table 1). Proteins from the first sub-class show an RRM conserved motif at the N-terminal end, followed by a glycine-rich region with GGYGG repeats (Sachetto-Martins et al., 2000). GRPs from the second sub-class show a similar organization, but present a CCHC zinc finger inside their glycine-rich region. Proteins from the third sub-class are organized with a cold-shock domain at the N-terminus and a number of CCHC zinc fingers in their glycine-rich region that varies from 1 to 7 (Sachetto-Martins et al., 2000; Karlson and Imai, 2003). Finally, sub-class IV RNA-binding GRPs present two copies of the RRM motif followed by a C-terminal glycine-rich region, unlike the previously described proteins (Fusaro et al., 2001).

Twenty-seven Eucalyptus clusters encoding RNA-binding GRPs were identified and were classified according to the structural organization of their domains. In order to analyze the relationships between them and other related RNA-binding GRPs already characterized, a phylogenetic tree was constructed (Figure 3).

Sixteen clusters belong to the sub-class I (Table 8). Among these, 7 presented a pattern of expression limited to only one or two libraries indicating that they can probably represent tissue-specific genes. It was observed that sequences from Eucalyptus sub-class I of RNA-binding GRPs split into two separated groups (Figure 3). One group is closely related to the Arabidopsis glycine rich RNA-binding proteins (AtGR-RBPs) 2, 3, 4, 5 and 6. The other group is more related to genes coding for RNA binding proteins from Nicotiana sylvestris (RGP-1a, -1b and -1c), Nicotiana glutinosa (NgRBP) and Euphorbia (EcGRRBP-1 and -2). Interestingly, the N. sylvestris genes were reported to present tissue-specific alternative splicing and were suggested to produce truncated polypeptides as well as functional RNA-binding polypeptides (Hirose et al., 1993). The high number of clusters belonging to this sub-class of RNA-binding proteins and the close relationship they present may reflect that at least some of these sequences correspond to alternative spliced forms of the same gene. Both Eucalyptus groups of sub-class I RNA binding GRPs are more related to other previously reported sequences from dicot plants, while several sugarcane sequences included in the phylogenetic tree are preferentially related to sequences from monocot plants like Zea mays (MA16 and CHEM2) and Shorgum vulgare (S1 and S2).

RNA binding GRPs from sub-class II are the least abundant among all the RNA-binding GRPs and are apparently plant-specific (Lorkovic and Barta, 2002). The domain organization of these proteins presents a CCHC-type zinc finger inside the glycine-rich C-terminal domain in combination with the N-terminal RRM motif. Only two clusters were found in the ForEST database with these characteristics (Table 8). One of the clusters (EGJEFB...
1029H07.g) is very similar to the tobacco nuclear protein RZ-1 (Hamano et al., 1996) while the other (EGSBSL1048F09.g) has a close similarity with a still non-characterized *Arabidopsis* protein (Table 8 and Figure 3).

Sub-class III RNA-binding GRPs were represented by 4 clusters in the ForEST database. Two of them were isolated from only one or two libraries corresponding to putative tissue-specific expressed genes (Table 8). Three clusters (EJECL2215H02.g, EGEPRT3325H02.g and EGUTBK1006H11.g) grouped close to the *Arabidopsis* cold-induced proteins AtGRP-2 and AtGRP-2b, proteins that have two zinc fingers in their glycine-rich domains. The remaining cluster (EGEQFB1001F04.g) appears more related to two other sequences from *Arabidopsis* (At2g17870 and At4g36020) that were also shown to be cold-regulated (Karlson and Imai, 2003) but have a longer C-terminal end with 7 zinc fingers interspersed in the glycine-rich region. Two zinc fingers were observed in all the Eucalyptus sequences with the exception of one cluster that is incomplete in its C-terminal end which made the analysis of the zinc finger number of this cluster impossible.

Five *Eucalyptus* clusters encoding GRPs with multiple RRM domains were classified as belonging to sub-class IV (Table 8). Among them, two clusters (EGACST22105B03.g and EGEFB1016C10.g) share high similarity with *Arabidopsis* UBA1 proteins. Comparison analysis indicates that they group together with *Arabidopsis* UBA2c (Figure 3). UBA1 and UBA2 proteins bind RNA with specificity for oligouridylates *in vitro* and interact with UBP1, an hnRNP-like protein associated with poly(A)(+) RNA in the cell nucleus. It has been suggested that UBA proteins may act as components of a complex that recognizes U-rich sequences in plant 3'-UTRs, contributing to the stabilization of mRNAs in the nucleus (Lambermon et al., 2002). The three remaining clusters from the RNA-binding GRPs sub-class IV (EGECST222E05.g, EGEQRT3201H05.g and EAGTBL124BF11.g) appear to be *Arabidopsis* heterogeneous nuclear ribonucleoproteins (hnRNPs), RNA-binding proteins that form complexes with RNA polymerase II transscripts and are proposed to regulate pre-mRNA processing (Krecic and Swanson, 1999). While metazoan hnRNPs have a Glycine-rich C-terminal domain in addition to the two N-terminal RRMs, only two out of the six *Arabidopsis* predicted hnRNPs have a C-terminal domain rich in glycine (Lorkovic and Barta, 2002). The only two sugarcane sequences identified as sub-class IV RNA-binding GRPs (Fusaro et al., 2001) grouped together with the hnRNP similar proteins.

In addition to sequences classified in the four previous described sub-classes of RNA-binding GRPs, 8 clusters encoding GRPs that present other conserved domains usually found in RNA-binding proteins were found in *Eucalyptus* (Table 9). One cluster (EQLHVL2253F10.g) has a conserved domain characteristic of Gar1, a small nucleolar RNP that possesses a typical glycine/arginine-rich domain and is required for pre-rRNA processing and pseudouridylation (Bagni and Lapeyre, 1998). Two clusters (EGEQQFB1001B12.g and EGEERT3101F06.g) have a CCCH (CX3CX2CX3HX3) type zinc finger. It has been shown that different CCCH zinc finger-containing proteins interact with the 3' untranslated region of various mRNA. Three clusters (EGELVL2220C06.g, EGERRT3101F05.g and EGEQRT3300C03.g) were identified with a domain found in proteins that includes the HABP4 family proteins, and the PAI-1 mRNA-binding protein. HABP4 has been observed to bind hyaluronan as well as RNA, but the latter with a lower affinity. PAI-1 mRNA-binding protein specifically binds the mRNA of type-1 plasminogen activator inhibitor (PAI-1), and is thought to be involved in regulation of mRNA stability. Finally, one cluster (EGEQLV2243F09.g) was found with the conserved LSM domain present in proteins that bind and stabilize snRNPs involved in pre-mRNA splicing.

Since proteins containing such domains as the unique RNA-binding motifs could not be predicted unequivocally as having an RNA-binding function, they were classified as putative RNA-binding GRPs. Particularly interesting is the cluster EGTSL1044E03.g. It could be a conserved RNA-binding GRP since it presents an RM motif, but unlike RNA-binding GRPs of classes I, II or IV this domain is located at the C-terminal end of the protein. The sequence with higher similarity to this cluster corresponds to a rice mRNA that encodes a glycine-rich protein with a C-terminal located RRM motif in combination with RanBP2 type zinc fingers at the N-terminal end. This kind of domain organization was never reported before for a GRP and could represent a new class of still uncharacterized RNA-binding GRPs. Since the *Eucalyptus* cluster is incomplete at the N-terminal the presence of zinc fingers could not be determined.

**Eucalyptus** clusters encoding proteins with glycine-rich domains

In addition to the GRPs showing glycine-rich domains with semi-repetitive structure described here, several proteins that present short domains with high glycine content and usually without a characteristic pattern of repetition were also found (Table 10). These proteins were classified as proteins with glycine-rich domains. Those clusters presented glycine-rich domains ranging from 32 to 130 aminoacids with 35-81% of glycine. Glycine-rich stretches shorter than 30 aminoacids were not included in this classification.

Out of the 16 *Eucalyptus* sequences that have glycine-rich domains in their structure, 7 are similar to known RNA binding proteins including the ribosomal RNA processing fibrillarin, several DEAD box RNA helicases, a nucleotide excision repair protein, a BHLL transcriptional regulator and a nucleolin-like protein. The presence of a short glycine-rich domain in a number of pro-
teins involved in RNA metabolism suggests that this domain may play a role in the RNA binding function of these proteins.

Concluding Remarks

Although the number of genes encoding GRPs in plants is large up to date, only a few GRPs have been characterized so far and their functions remain speculative. However, it is becoming clear that GRPs exert important roles in very diverse processes such as signal transduction, stress response, transcriptional regulation and development.

The highly specific but diverse expression pattern of grp genes, taken together with the distinct sub-cellular localization of some GRP groups, clearly indicate that these proteins are implicated in several independent physiological processes. Notwithstanding the absence of a clear definition of the role of GRPs in plant cells, studies conducted with these proteins have provided new and interesting insights on the molecular and cell biology of plants. Complexly regulated promoters and distinct mechanisms of gene expression regulation have been demonstrated (Keller and Heierli, 1994; Franco et al., 2002). New protein targeting pathways, as well as the exportation of GRPs from different cell types have been discovered (Ryser et al., 1997; Murphy and Ross, 1998). These data show that GRPs can be useful markers for many physiological processes and/or models to improve the understanding of distinct aspects of plant biology (Sachetto-Martins et al., 2000). The results obtained here point to interesting roles for GRPs in plant physiology. The characterization of the grp genes in Eucalyptus could lead to new strategies for the manipulation of growth and stress signaling in this culture.

Acknowledgments

S.N.B is a recipient of a CNPq post-doctoral fellowship. C.M. is a recipient of the CAPES Prodoc fellowship. V.C.J. is supported by the PIBIC-CNpq fellowship. A.M. and R.M.J. were supported by Master degree fellowships from CAPES. G.S.M. are indebted to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Função Carlos Chagas de Amparo a Pesquisa do Rio de Janeiro (FAPERJ) for financial support.

References

Albà MM, Culiáñez-Macià FA, Goday A, Freire MA, Nadal, B and Pagès M (1994) The maize RNA-binding protein, MA16, is a nucleolar protein located in the dense fibrillar component. Plant J 6:825-834.

Albà MM and Pagès M (1998) Plant proteins containing the RNA-recognition motif. Trends Plant Sci 3:15-21.

Allagulova ChR, Gimalov FR, Shakirova FM, and Vakhitov VA (2003) The Plant Dehydrogenases: Structure and Putative Functions. Biochemistry (Mosc) 68:945-951.

Allona I, Quinn M, Shoop I, Swope K, St Cyr S, Carlis J, Riedl J, Retzel E, Campbell M, Sedero R and Whetten RW (1998) Analysis of xylem formation in pine by cDNA sequencing. Proc Natl Acad Sci USA 95:9693-9698.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Bagioni C and Lapeyre B (1998) Galrp binds to the small nucleolar RNAs snR10 and snR30 in vitro through a nontypical RNA binding element. J Biol Chem 273:10868-10873.

Cassab GI (1998) Plant cell wall proteins. Annu Rev Plant Physiol Plant Mol Biol 49:281-309.

Condit CM and Meagher RB (1986) A gene encoding a novel glycine-rich structural protein of petunia. Nature 323:178-181.

Condit CM (1993) Developmental expression and localization of petunia glycine-rich protein 1. Plant Cell 5:277-288.

de Oliveira DE, Franco LO, Simeons C, Seurink J, Coppitiets J, Bottermann J and van Montagu M (1993) Influence-specific genes from Arabidopsis thaliana encoding glycine-rich proteins. Plant J 3:495-507.

Freire MA, Almeira-Engler J, Miguens FC, van Montagu M, Engler G and de Oliveira DE (1997) Oleosin gene expression in Arabidopsis thaliana coincides with accumulation of lipids in plastids and cytoplasmic bodies. Plant Physiol Biochem 35:729-739.

Franco LO, de O Manes CL, Hamidi S, Sachetto-Martins G and de Oliveira E (2002) Distal regulatory regions restrict the expression of cis-linked genes to the tapetal cells. FEBS Lett 517:13-18.

Freire MA and Pages M (1995) Functional characterization of the maize RNA binding protein MA16. Plant Mol Biol 29:797-807.

Fusaro A, Mangeon A, Magrani Junqueira R, Benício Rocha CA, Cardoso Coutinho T, Margis R and Sachetto-Martins G (2001) Classification, expression pattern and comparative analysis of sugarcane expressed sequences tags (ESTs) encoding glycine-rich proteins (GRPs). Genet Mol Biol 24:263-273.

Hanano S, Sugita M and Sugiuma M (1996) Isolation of a novel RNA-binding protein and its association with a large ribonucleoprotein particle present in the nucleoplasm of tobacco cells. Plant Mol Biol 31:57-68.

Hirose T, Sugita M and Sugiuma M (1993) cDNA structure, expression and nucleic acid-binding properties of three RNA-binding proteins in tobacco: Occurrence of tissue-specific alternative splicing. Nucleic Acids Res 21:3981-3987.

Karlson D, Nakaminami K, Toyomasu T and Imai R (2002) A cold-regulated nucleic acid-binding protein of winter wheat shares a domain with bacterial cold-shock proteins. J Biol Chem 277:35248-35256.

Karlson D and Imai R (2003) Conservation of the cold shock domain protein family in plants. Plant Physiol 131:12-15.

Keller B, Schmid J and Lamb CJ (1989) Vascular expression of a bean cell wall glycine-rich protein - β-glucuronidase gene fusion in transgenic tobacco. Embo J 8:1309-1314.

Keller B and Heierli D (1994) Vascular expression of the grp1.8 promoter is controlled by three specific regulatory elements and one unspecific activating sequence. Plant Mol Biol 26:747-756.
Krecic AM and Swanson MS (1999) hnRNP complexes: Composition, structure, and function. Curr Opin Cell Biol 11:363-371.

Kumar S, Tamura K, Jacobsen I and Nei M (2000) MEGA2: Molecular Evolutionary Genetics Analysis, version 2.0. Pennsylvania and Arizona State Universities, University Park, Pennsylvania and Tempe, Arizona.

Lambermon MH, Fu Y, Wieczorek Kirk DA, Dupasquier M, Filipowicz W and Lorkovic ZJ (2002) UBA1 and UBA2, two proteins that interact with UBPI, a multifunctional effector of pre-mRNA maturation in plants. Mol Cell Biol 22:4346-4357.

Le Provost G, Paiva J, Pot D, Brach J and Plomion C (2003) Seasonal variation in transcript accumulation in wood-forming tissues of maritime pine (Pinus pinaster Ait.) with emphasis on a cell wall glycerin-rich protein. Planta 217:820-830.

Lorkovic ZJ and Barta A (2002) Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant Arabidopsis thaliana. Nucleic Acids Res 30:623-635.

Magioli C, Barrôco RM, Benício Rocha CA, de Santiago-Fernandes LD, Mansur E, Engler G, Margis-Pinheiro M and Sachetto-Martins G (2001) Somatic embryo formation in Arabidopsis and eggplant is associated with expression of a glycerin-rich protein gene (Atgrp-5). Plant Sci 161:559-567.

Matsushima N, Creutz CE and Kretzinger RH (1990) Polyproline, beta-turn helices. Novel secondary structures proposed for the tandem repeats within rhodopsin, synaptophysin, synexin, gliadin, RNA polymerase II, hordein, and gluten. Proteins 7:125-155.

Murphy DJ and Ross JHE (1998) Biosynthesis, targeting and processing of oleosin-like proteins, which are major pollen coat components in Brassica napus. Plant J 13:1-16.

Murphy DJ, Hernández-Pinzón I and Patel K (2001) Role of lipid bodies and lipid-body proteins in seeds and other tissues. J Plant Physiol 158:471-478.

Ni Z, Sun Q, Liu Z, Wu L and Wang X (2000) Identification of a hybrid-specific expressed gene encoding novel RNA-binding protein in wheat seedling leaves using differential display of mRNA. Mol Gen Genet 263:934-938.

Obokata J, Ohme M and Hayashida N (1991) Nucleotide sequence of a cDNA clone encoding a putative glycerin-rich protein of 19.7 kDa in Nicotiana sylvestris. Plant Mol Biol 17:953-955.

Park AR, Cho SK, Yun UJ, Jin MY, Lee SH, Sachetto-Martins G and Park OK (2001) Interaction of the Arabidopsis Receptor Protein Kinase Wak1 with a Glycerin-rich Protein, AtGRP-3. J Biol Chem 276:26688-2669.

Ringli C, Keller B and Ryser U (2001) Glycerin-rich proteins as structural components of plant cell walls. Cell Mol Life Sci 58:1430-1441.

Ryser U and Keller B (1992) Ultrastructural localization of bean glycerin-rich protein in un lignified primary walls of protophylem cells. Plant Cell 4:773-783.

Sachetto-Martins G, Fernandez LD, Felix DB and de Oliveira DE (1995) Preferential transcriptional activity of a glycerin-rich protein gene from Arabidopsis thaliana in protoderm-derived cells. Int J Plant Sci 156:460-470.

Showalter AM (1993) Structure and function of plant cell wall proteins. Plant Cell 5:9-23.

Smitikova T, Rzhetsky A and Nei M (1995) Interior-branch and bootstrap tests of phylogenetic trees. Mol Biol Evol 12:319-333.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Verica JA and He ZH (2002) The cell wall-associated kinase (WAK) and WAK-like kinase gene family. Plant Physiol 129:455-459.

Welin BV, Olson A, Nylander M and Palva ET (1994) Characterization and differential expression of dhn/lea/rab-like genes during cold acclimation and drought stress in Arabidopsis thaliana. Plant Mol Biol 26:131-144.

Zhang YI, Sederoff R and Allona I (2000) Differential expression of gene encoding cell wall proteins in vascular tissues from vertical and bent loblolly pine trees. Tree Physiol 20:450-457.