Glycosylated Polyproline II Rods with Kinks as a Structural Motif in Plant Hydroxyproline-Rich Glycoproteins†

Patrick J. Ferris,§ Jeffrey P. Woessner,§,|| Sabine Waffenschmidt,⊥ Sven Kilz,⊥ Jutta Drees,⊥ and Ursula W. Goodenough*,§

Department of Biology, Washington University, St. Louis, Missouri 63130, and Institute für Biochemie, Universitat zu Köln, Köln, Germany

Received October 10, 2000; Revised Manuscript Received December 20, 2000

ABSTRACT: Hydroxyproline-rich glycoproteins (HRGPs) are the major proteinaceous components of higher plant walls and the predominant components of the cell wall of the green alga *Chlamydomonas reinhardtii*. The GP1 protein, an HRGP of the *C. reinhardtii* wall, is shown to adopt a polyproline II helical configuration and to carry a complex array of arabinogalactoside residues, many branched, which are necessary to stabilize the helical conformation. The deduced GP1 amino acid sequence displays two Ser-Pro-rich domains, one with a repeating $(SP)_x$ motif and the other with a repeating $(PPSPX)_x$ motif. A second cloned gene *a2* also carries the PPSPX repeat, defining a novel gene family in this lineage. The SP-repeat domains of GP1 form a 100-nm shaft with a flexible kink 28 nm from the head. The *gp1* gene encodes a PPPPPRPPFPANTPM sequence at the calculated kink position, generating the proposal that this insert interrupts the PPII helix, with the resultant kink exposing amino acids necessary for GP1 to bind to partner molecules. It is proposed that similar kinks in the higher plant HRGPs called extensins may play a comparable role in wall assembly.

Hydroxyproline-rich glycoproteins (HRGPs),[1] first identified in dicots but since found in monocots and green algae (*1*−*3*; reviewed in refs *4*−*7*), are the major proteinaceous components of the plant cell wall and often become covalently cross-linked into large meshworks (*8*, *9*). Biochemical and molecular studies of these proteins, which have been assigned to several subgroups (*4*, *5*), have documented commonalities—repeated motifs dominated by (hydroxy)-proline and serine, a polyproline II helical conformation, and arabinosyl/galactosyl side chains—that have led Kieliszewski and Lamport (*10*) to propose that they derive from an ancient gene family.

The extensin subgroup of the HRGP family includes proteins that are uniformly fibrous and apparently devoted to meshwork formation, but several members of the subgroup, called chimeric extensins (*10*), display both fibrous and globular domains, the globular portions presumably playing some additional role. These include several solonaceous lectins, where the hydroxyproline-rich fibrous domains are thought to anchor the protein to the wall and the globular domains to mediate sugar-binding activity (reviewed in refs *4* and *5*). Chimeric HRGPs are also expressed in the reproductive tissues of monocots (*11*, *12*) and dicots (*13*−*15*), where their functions have not yet been elucidated.

Algal HRGPs have been largely studied in *Chlamydomonas* and *Volvox* (reviewed in refs *16* and *17*), organisms that construct their vegetative and zygotic cell walls exclusively from HRGPs. In addition to an "inner wall" of covalently cross-linked HRGPs reminiscent of the higher plant meshworks (*18*), they possess as well an "outer wall" of HRGPs that coassemble into crystalline arrays. The arrays can be solubilized in chaotropic salts, and they reassemble when the salts are removed by dialysis (*18*−*23*). These outer-wall HRGPs represent excellent subjects for molecular and morphological studies on HRGP interactions.

Most of the volvocine HRGPs thus far characterized prove to be chimeric proteins, with globular "heads" and hydroxyproline-rich "shafts." Two classes of proline-rich motifs (when discussing HRGP-encoding genes, the residues are referred to as proline, most of which are posttranslationally converted to hydroxyproline) have been identified to date in *Chlamydomonas* and *Volvox*: in the first, the prolines are contiguous; in the second, the prolines alternate with some other amino acid, most often serine. We report here a new Pro-rich motif, identified in two cloned genes of *Chlamydomonas reinhardtii*, that is characterized by regular repeats of PPSPX, thereby generating sequences that carry both contiguous and noncontiguous prolines. One chimeric PPSPX gene, called *a2*, is expressed exclusively in gametes, where the function of its protein product has not yet been elucidated. A second chimeric PPSPX gene, called *gp1*, encodes the GP1 protein which coassembles with two other HRGPs (GP2 and GP3) to form the salt-soluble outer wall of *C. reinhardtii*

† Supported by NIH Grant GM-26150, NSF Grant MCB-9904667, and Fonds der Chemischen Industrie.

* Corresponding author. Tel: 314-935-6836. Fax: 314-935-5125. E-mail: ursula@biosgi.wustl.edu.

§ Washington University.

|| Current address: Paradigm Genetics, Research Triangle Park, NC 27709.

⊥ Universitat zu Köln.

[1] Abbreviation: HRGP, hydroxyproline-rich glycoprotein.

10.1021/bi0023605 CCC: $20.00 © 2001 American Chemical Society
Published on Web 02/10/2001

(*20*). Since much is known about the biochemistry and morphology of GP1 (*20*, *22*), we are able to offer a detailed correlation between the structure of this HRGP and its deduced amino acid sequence.

MATERIALS AND METHODS

Characterization of the a2 Gene. Region *a* is a ~20 kb segment present in the *mt*+ locus but not the *mt*− locus of *C. reinhardtii* (*24*) and is duplicated in an autosomal location in the *C. reinhardtii* genome. A 1.8 kb *Bam*H1 restriction fragment from phage HD20 which derives from the *mt*+ *a* region (*24*) was radiolabeled and used to screen a λZAPII cDNA library created using zygote polyA+ RNA. Two classes of cDNA were identified: *a1*, derived from a 0.8 kb message (to be described elsewhere), and *a2*, derived from a 2.1 kb message. The longest *a2* cDNA, which proved to be full length, was sequenced; the corresponding genomic sequence (GenBank accession number AF309495) was obtained from the BO5 phage derived from strain CC-621.

RNA was isolated from *mt*+ and *mt*− strains at various stages of the life cycle, polyA-selected using oligo(dT)−cellulose, run on formaldehyde−agarose gels, and transferred to nylon membranes. Membranes were hybridized with probes radiolabeled by random priming.

Determination of the GP1 DNA Sequence. A cDNA clone for GP1 was first isolated by Adair and Apt (*25*). This clone was radiolabeled and used to screen a λZAPII cDNA library created from RNA isolated from vegetative cells undergoing cell wall regeneration (*9*), yielding six additional cDNA clones, the longest of which was sequenced. No full-length cDNAs were obtained, and the longest terminated in a long GC-rich stretch, suggesting that further screening of the library would be unfruitful. The longest cDNA was therefore radiolabeled and used to screen a λEMBL3 genomic library (*24*). Several hybridizing phage were selected, purified, and restriction-mapped. A 3258 bp *Xho*I/*Sac*I fragment encompassing the GP1 cDNA was sequenced (GenBank accession number AF309494).

We discovered during our work that the partial sequence of the *gp1* cDNA described in Adair and Apt (*25*), GenBank accession number M58496, was inadvertently entered as 3′ to 5′ rather than 5′ to 3′. When read in the proper direction, M58496 disagrees with our genomic and cDNA sequencing at five positions. We resequenced the Adair and Apt cDNA and determined that these five discrepancies are all sequencing errors in M58496. Adair and Apt (*25*) report that the *gp1* cDNA hybridizes to two mRNAs on Northern blots of total RNA—one of ~3.3 kb and one of ~3.5 kb. In our experiments, using either total RNA or polyA+ RNA, the *gp1* cDNA hybridizes to a single mRNA of 2.6 kb, a size more consistent with the predicted gene structure. Although we cannot settle this discrepancy with certainty, the appearance of their Northerns and the size of their "messages" suggest that Adair and Apt may have misidentified a hybridization artifact caused by overloaded ribosomal RNA as the *gp1* message.

Biophysical and Carbohydrate Analysis of GP1. GP1 was purified from vegetative cells of the *bld2 mt*− strain as described (*20*). HF/pyridine deglycosylation of GP1 was performed using a slightly modified version of the protocol of van Holst and Varner (*26*). GP1 (100 μg) was dissolved in 20 μL of anhydrous methanol and 180 μL of HF/pyridine. The deglycosylation reaction was carried out for 90 min at room temperature and terminated by the addition of 2 mL of 30 mM octyl glucoside in 1 M Tris. The protein was dialyzed against water and lyophilized.

CD spectra were recorded on a Jasco Model J-175 (Jasco, Gross-Umstadt, Germany), calibrated using ammonium d_{10}-camphorsulfonic acid. All measurements were carried out in a heat-controlled 0.1-cm path-length cylindrical cuvette at 180−260 nm at 25 °C. Typically 10 spectra were recorded at a scan speed of 50 nm/min with a step resolution of 0.1 nm and a nitrogen stream of 8 mL/min. All spectra were corrected for a protein-free spectrum obtained under identical conditions; noise reduction was applied according to the Jasco software.

For monosaccharide analysis, 150 μg of purified GP1 was dissolved in 300 μL of 2 N trifluoroacetic acid (TFA) with 10 μL of *myo*-inositol (55.5 nmol/μL) as internal standard. Hydrolysis was performed for 1 h at 125 °C. After cooling, the hydrolysate was dried at 40 °C under a nitrogen stream. The sample was then redissolved in 200 μL of methanol to remove all trace of TFA and dried again. The monosaccharides were then converted to persilyated alditols according to Kilz et al. (*27*), and these were separated using a capillary gas chromatograph (GC 9000 Series, Fisons Instruments, Mainz, Germany) equipped with a FID detector on a 30 mL DB-5 column (J&W Scientific) using a linear temperature program starting at 140 °C (3 min) and increasing to 240 °C with a heating rate of 5 °C/min.

For analysis of hydroxyproline-bound sugars, 500 μg of GP1 was dissolved in 300 μL of 0.44 N $Ba(OH)_2$ and incubated for 22 h at 130 °C. After cooling, the pH was adjusted to 7.5 using 0.1 N NH_4SO_4. Precipitated $BaSO_4$ was removed by centrifugation. The hydrolysate (1 nmol of sucrose as internal standard) was dissolved in 20 μL of quartz-distilled water and separated by HPLC on a Nucleosil 300 NH_2 column, applying a nonlinear gradient of H_2O/acetonitrile (79% acetonitrile at time 0, 65% at 200 min, 50% at 210 min, 10% at 225 min, 5% at 230 min, end 250 min) with a flow rate of 0.8 mL/min. Detection was performed online by ESI-MS (MAT 900 ST equipped with a Patric-Detector, Finigan, Bremen, Germany). The scanning rate was 4 s/d.

MALDI mass spectra were acquired using a linear time-of-flight BIFLEX III (Bruker-Franzen, Germany) instrument, equipped with delayed ion technology. Typically 50−200 laser shots (N_2 laser, 337 nm) were added per spectrum. All spectra were recorded in the positive ion mode at an accelerating voltage of 20 kV. BSA dimer was used for external calibration. For MALDI-MS analysis, GP1 samples were dissolved in 0.1% TFA to a final concentration of 1 mg/mL. Samples were prepared using the thin-layer method: 1 μL of a saturated solution of sinapinic acid in ethanol was placed on the target and dried at room temperature; 5 μL of sample and 5 μL of a saturated solution of sinapinic acid in 0.1% TFA/acetonitrile (2:1 v/v) were mixed, and 1 μL of the mixture was deposited on the dried matrix layer.

Isoelectric focusing was performed using a Pharmacia Fast System.

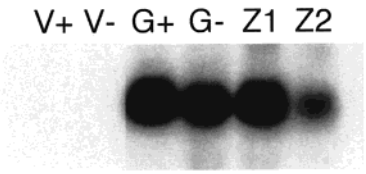


FIGURE 1: Regulation of the *a2* mRNA. A Northern blot of total RNA (20 μg) was hybridized with a full-length *a2* cDNA probe. The RNA was purified from vegetative cells of a *mt*+ (V+) or *mt*− strain (V−), gametes of a *mt*+ (G+) or *mt*− strain (G−), and zygotes either 30 min (Z1) or 3 h (Z2) after mating.

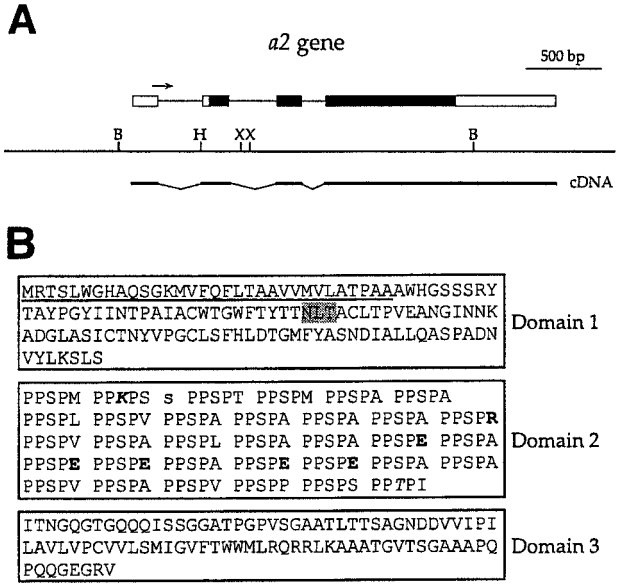


FIGURE 2: The *a2* gene. (A) Structure of the *a2* gene. The center line indicates the extent of the genomic sequencing. The locations of *Bam*HI (B), *Hin*dIII (H), and *Xho*I (X) restriction sites are indicated. The extent of the longest cDNA is depicted below the line; the intron/exon structure of *a2* is shown above the line with the 5′ end of the message on the left. Open boxes indicate untranslated sequences, closed boxes the coding regions, and thin lines the three introns. (B) Protein sequence of A2 drawn to show three domains. In domain 1, underlining marks the potential signal peptide, and a canonical N-glycosylation site is shaded. In domain 2, the PPSPX units are separated, any amino acids not within a PPSPX motif are in lower case, variants from the PPSPX consensus are in italics, and charged amino acids are in bold face.

RESULTS

The a2 Gene. The *a2* gene of *C. reinhardtii* was cloned by virtue of its homology to a sequence in the *a* region of the mating-type *plus* locus, which we have cloned and are actively investigating (*24*). The sequence in the *plus* locus is not expressed because it has been inactivated by the 5′ insertion of a second gene, called *a1* (details will be reported in a separate paper). The *a2* copy of this gene, located in an unidentified autosome, hybridizes to a 2.1 kb mRNA which is not expressed in vegetative cells, switched on in gametic (nitrogen-starved, noncycling) cells of both mating types and turned off again during early zygote development (Figure 1). The function of the A2 protein is unknown.

The *a2* gene contains four exons (Figure 2A) and encodes a 38.8 kDa protein (p*I* 6.7) with the predicted amino acid sequence shown in Figure 2B. A signal sequence is followed by three domains. Domain 1 has a putative N-glycosylation signal, four cysteine residues, and a mixed distribution of amino acids. Using the PSIPRED method for predicting secondary structure (http://insulin.brunel.ac.uk/psipred/), this domain is predicted to carry 36% β-strand, 14% α-helix, and 49% random coil, suggesting that it adopts a globular configuration. Domain 3 also carries a mix of amino acids, including a long hydrophobic sequence, and is predicted to carry 8% β-strand, 41% α-helix, and 51% random coil, indicating that while the protein carries globular domains at both its N- and C-termini, their secondary structures are quite different.

The intervening domain 2 carries 33 nearly perfect repeats of the motif PPSPX, an extra S appearing before the third module and a K and a T replacing the S in modules 2 and 33. X is most often A (14/33), followed by E (5/33), and, in total, 10 different amino acids appear at the X position, with the missing amino acids including those with particularly bulky R groups (e.g., F, H, W, and Y). Assuming that domain 2 adopts a uniform left-handed polyproline II helix, wherein 1 nm of the extended configuration corresponds to 3.34 amino acids (Trevor Creamer, personal communication), domain 2 would translate into a 50-nm rod.

The GP1 Protein: Biophysical Properties. Previously published studies on purified GP1 have documented that the protein is 32.3% hydroxyproline, 15.8% serine, and 2.7% proline (*20*) and that, when viewed by quick-freeze deep-etch transmission electron microscopy, it displays three discrete morphological domains: a globular head, a narrow neck, and a wider shaft interrupted by a flexible juncture called a kink (Figure 3). The neck, it should be noted, was not pointed out in our earlier publication (*20*); its status as a discrete domain has been brought to our attention by the derived amino acid sequence of GP1 described in the following section.

The high content of hydroxyproline suggests that the extended shaft domain of GP1 is configured as a polyproline II (PPII) helix. This was confirmed by CD spectroscopy of GP1 (Figure 4A) and, for comparison, tomato extensin (Figure 4B). The CD spectrum of GP1 has extrema at ~197 and 224 nm, with molar ellipticities of −16 100 and 3900 deg cm^2 dmol^{-1}, respectively, in good agreement with the values for the tomato extensin (extrema at ~201 and 227 nm and molar ellipticities of −11 400 and 2300 deg cm^2 dmol^{-1}) as well as carrot extensin (*26*). When the spectrum of GP1 was further extracted for secondary structural elements using the SELCON2 program (*28*), the algorithm predicts not only PPII helix but also a significant amount of β-turn and β-sheet. The PPII helix of both GP1 and tomato extensin is heat-stable and refolds after cooling (data not shown), features also observed for carrot extensin (*26*). Following chemical deglycosylation, GP1, like carrot extensin, displays a significant decrease in the net intensities of its extrema, and the zero crossings shift to longer wavelengths (Figure 4C), reflecting a transition from a structured to a less structured molecule (*29*). These data document that the carbohydrate side chains reinforce the PPII configuration.

Since glycoproteins of this class display anomalous migration when analyzed by gel filtration or SDS/PAGE (*20*, *30*), the molecular mass of GP1 was estimated using MALDI-TOF, which has been successfully used to measure the molecular mass of soluble extensins (*31*, *32*). MALDI-TOF of native (Figure 5A) and deglycosylated (Figure 5B) GP1 gave values of 272 359 and 65 139, respectively,

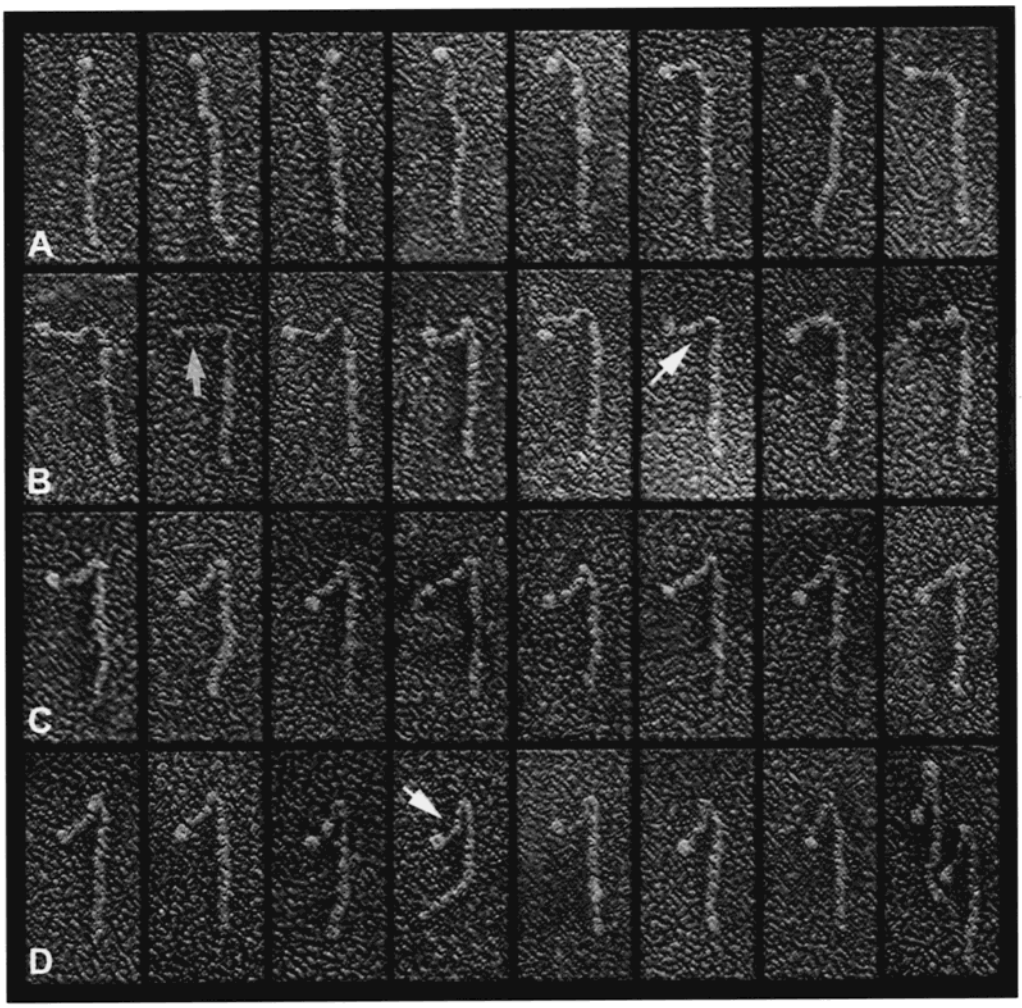


FIGURE 3: Montage of GP1 proteins adsorbed to mica flakes and subjected to quick-freeze deep-etch electron microscopy (*20*). Proteins in row A show no kink, although the kink position is often detectable as a discontinuity in the linear shaft; proteins in rows B−D show increasingly acute kink angles. Arrows point to the neck domain contiguous to the head; the neck is evident in other proteins in the montage but not in all, suggesting that it may be capable of contraction/extension. Magnification: 300000×.

indicating that ~75% of the protein's mass is contributed by carbohydrate [for comparison, this value can be up to 95% for certain HRGPs (*7*, *33*)].

The GP1 Protein: Carbohydrate Composition. Previous studies of crude cell-wall preparations from *Chlamydomonas* have shown that the serines are O-glycosylated with single galactose residues and that various heteroarabinosides are O-linked to hydroxyprolines (*3*, *34*). A later study of cell-wall HRGP mixtures also detected the presence of glucose, mannose, rhamnose, xylose, and methylated sugars (*35*).

To characterize the carbohydrate composition of purified GP1, two approaches were taken. To identify the major monosaccharides associated with GP1, the protein was hydrolyzed with 2 N trifluoroacetic acid, and the released sugars were identified as persilyated alditols by gas chromatography. As reported elsewhere (*27*), this analysis documents that arabinose (41.1%) and galactose (45.5%) are the prominent species, with glucose (7.9%), xylose (4.5%), and mannose (1.1%) as minor species. The galactose is presumably linked predominantly to serine residues as in other HRGPs (*10*, *34*).

To identify the heteroarabinosides associated with the Hyp residues, the protein was treated with saturated $Ba(OH)_2$, and the released residues were analyzed by ESI mass spectroscopy. Methylation analysis, described in ref *27*, permitted a distinction between linear and branched isomeric structures. HPLC-separated heteroarabinosides were further analyzed and quantified on-line by ESI-MS.

Table 1 displays the characteristic fragments encountered, and their structures, masses, and relative abundances (see ref *27* for nomenclature). Notable is the abundance of different species; indeed, the structures shown in Table 1 represent only 84% of the total, the remaining 16% being present at <0.5% frequency, too rare for fragmentation analysis. Also of interest is the large proportion (21%) that are branched, a proportion that is in fact closer to 37% since most of the rare species were long chain and presumably branched as well. By contrast, the GP2 and GP3 glycoproteins that coassemble with GP1 to form the salt-soluble outer wall carry a far lower proportion of branching residues; for purified GP3, for example, where 95% of the total heteroarabinosides were abundant enough for fragmentation analysis, only 4% were found to be branched (Kilz and Waffenschmidt, unpublished).

The GP1 Gene. In an earlier study (*25*), the GP1 protein was purified (*20*) and HF-deglycosylated, the deglycosylated protein was used to raise polyclonal antibody, and the antibody was used to screen a cDNA expression library. The

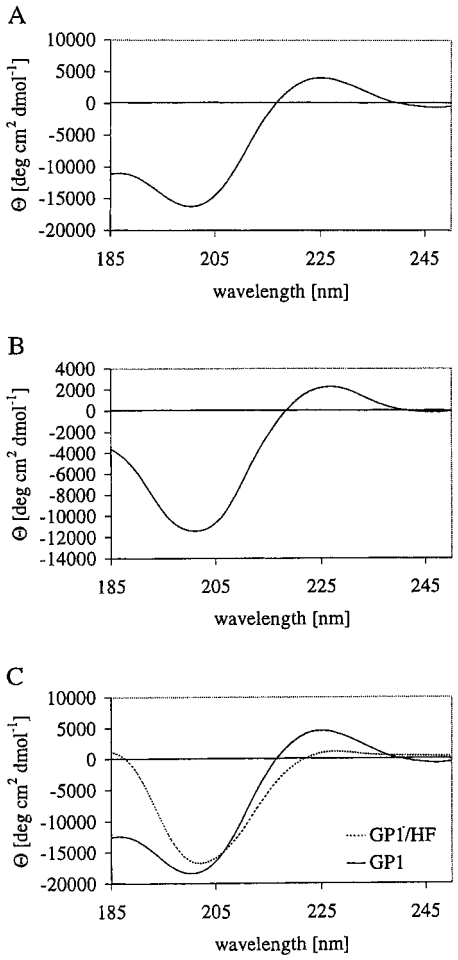


FIGURE 4: Far-UV CD spectra of (A) purified GP1, (B) tomato extensin, and (C) deglycosylated (dashed lines) versus native (solid lines) GP1 at pH 7. The protein concentration was 0.2 mg/mL.

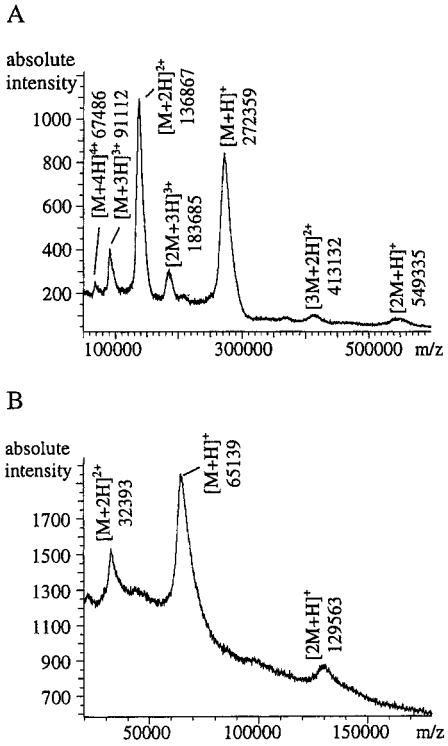


FIGURE 5: MALDI-TOF mass spectra of (A) native and (B) deglycosylated GP1. The spectrum of native GP1 shows various protonated GP1 adducts expressed as *m*/*z*, where *m* stands for mass and *z* is the number of positive charges on the respective ion. The major peaks correspond to doubly $(M + 2H)^{2+}$ and singly charged $(M + H)^{+}$ molecular ions centered at *m*/*z* 136 867 and 272 359, respectively. The spectrum of deglycosylated GP1 shows three peaks at *m*/*z* 32 393, 65 139, and 129 563 corresponding to the doubly and singly charged species and a singly charged dimer.

resultant cDNA clone was used in the present study to identify additional cDNA and genomic clones and thus characterize the full *gp1* sequence.

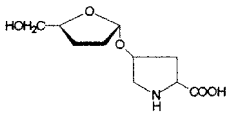
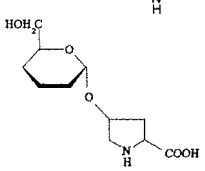
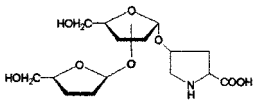
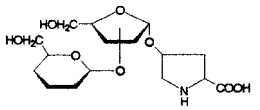
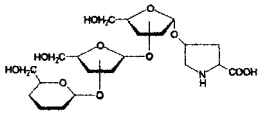
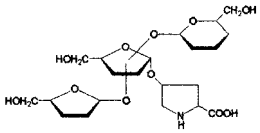
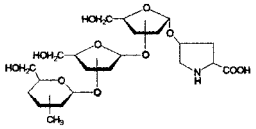
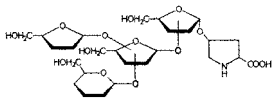
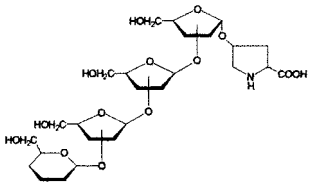
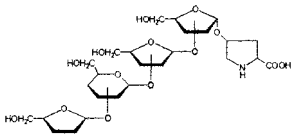
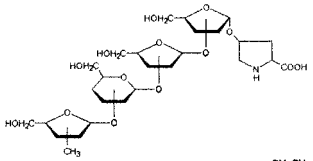
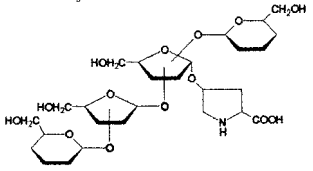
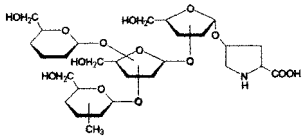
The *gp1* gene, which contains two introns (Figure 6A), hybridizes to a 2.6 kb mRNA (data not shown) and encodes a protein with the amino acid sequence shown in Figure 6B, where 42% of the residues are proline. The calculated molecular mass of this sequence—minus the signal peptide and assuming [on the basis of amino acid analysis (*20*)] that 90% of the prolines are hydroxylated—is 54.6 kDa, where the discrepancy between this and the MALDI-TOF value of 64.9 kDa (Figure 5b) is most likely the result of incomplete deglycosylation. The calculated p*I* of the protein (minus the signal peptide) is 10.9, in keeping with the marked retention of GP1 on cation-exchange columns (*20*); by isoelectric focusing, the purified protein has a p*I* of 11.0 (data not shown).

Four domains are evident in the GP1 amino acid sequence. The N-terminal domain 1, after cleavage of the predicted signal peptide, would comprise only 12 amino acids, 2 of which are cysteine residues. The C-terminal domain 4 contains 174 amino acids, including 2 cysteines and 3 putative N-glycosylation sites (possibly the location of some of the minor sugar residues). All of the 5 tryptophan residues in the protein are located in domain 4, and since the fluorescence spectrum of GP1 has an emission maximum of 330 nm which shifts to 345 nm during thermal denaturation (data not shown), these residues are presumed to be sequestered in a hydrophobic interior in the native protein. As noted above, CD analysis predicts the presence of β-sheets and β-turns in GP1, and the PSIPRED method predicts that domain 4 will be 32% β-sheet, 3% α-helix, and 65% random coil, a secondary structure more similar to the N-terminal than the C-terminal domain of A2. Since deep-etch images of GP1 show a single large globular domain which we call the head (Figure 3), we propose that the C-terminal domain 4 corresponds to the GP1 head and carries the protein's β-sheets and turns.

Domain 2 carries 49 repeats of the PPSPX motif. In four of the repeats, the S is replaced by V, A, or T—about the same level of substitution at this position as is observed in the A2 protein. The use of amino acids at the X position is more conservative in GP1 than in A2, however, with 27/49 being A, 9/49 being P, 8/49 being S, and the rest being G, K, T, and V. The most conspicuous difference between A2 and GP1 is that GP1 carries numerous intercalated sequences: a T precedes module 43; an SP and a PV precede modules 41 and 46, respectively; and 13 three amino acid inserts are dispersed throughout the rest of the domain, 5 of which are SPA, 4 PPV, 2 PPA, 1 PPL, and 1 PSP.

The final difference between the PPSPX domains of A2 and GP1 is that the GP1 domain carries a major discontinuity, marked by braces in Figure 6B, displaying the sequence PPPPPRPPFPANTPM. The centrally located R is preceded,

Table 1: Hydroxyproline-Bound Sugar Side Chains of GP1

structure	mass $(M + Na)^+$	relative abundance (%)	characteristic fragments (m/z)
	286	22.8	154 (Y_0)
	316	0.9	154 (Y_0)
	418	13.4	154 (Y_0) 286 (Y_1)
	448	10.2	154 (Y_0) 286 (Y_1) 317 (B_2)
	580	1.8	154 (Y_0) 286 (Y_1) 418 (Y_2)
	580	14.3	154 (Y_0) 286 ($Y_{1\alpha/1\beta}$) 418 ($Y_{1\alpha}$) 448 ($Y_{1\beta}$)
	594	3.1	154 (Y_0) 286 (Y_1) 458 (Y_2)
	712	0.6	154 (Y_0) 286 (Y_1) 418 ($Y_{2\alpha/2\beta}$) 550 ($Y_{2\alpha}$) 580 ($Y_{2\beta}$)
	712	1.4	154 (Y_0) 286 (Y_1) 418 (Y_2) 550 (Y_3)
	712	1.8	154 (Y_0) 286 (Y_1) 418 (Y_2) 550 (Y_3)
	726	7.9	154 (Y_0) 286 (Y_1) 418 (Y_2) 580 (Y_3)
	742	5.2	154 (Y_0) 286 (Y_1) 418 ($Y_{2/1\beta}$) 580 ($Y_{1\beta}$)
	756	0.9	154 (Y_0) 286 (Y_1) 418 ($Y_{2\alpha/2\beta}$) 580 ($Y_{2\alpha}$) 594 ($Y_{2\beta}$)

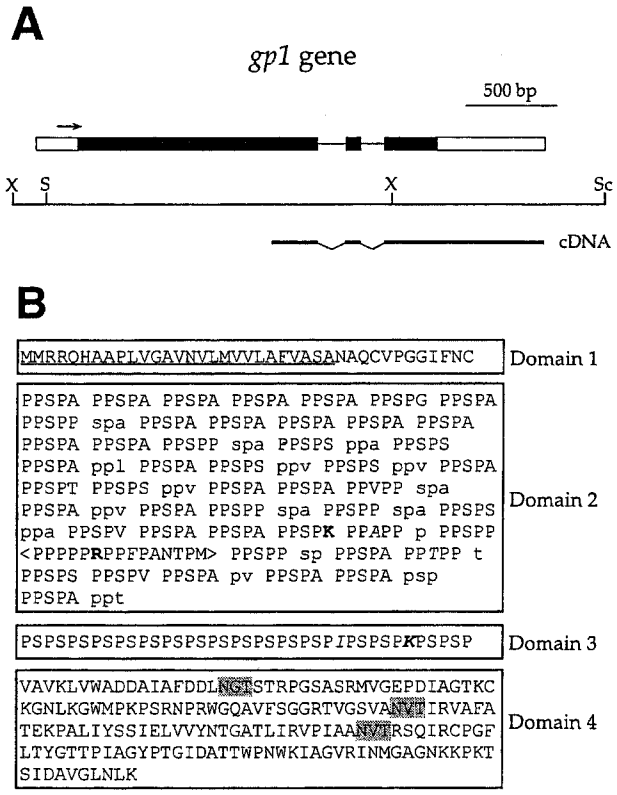


FIGURE 6: The *gp1* gene. (A) Structure of the *gp1* gene. The center line indicates the extent of the genomic sequencing. The locations of *Sac*I (Sc), *Sal*I (S), and *Xho*I (X) restriction sites are indicated. The extent of the longest cDNA is depicted below the line; above the line the intron/exon structure of *gp1* is shown with the 5′ end of the message on the left. Open boxes indicate untranslated sequences, closed boxes the coding regions, and thin lines the two introns. (B) Protein sequence of GP1 drawn to show four domains. In domain 1, underlining marks the potential signal peptide. In domain 2, the PPSPX units are separated, any amino acids not within a PPSPX are in lower case, variants from the PPSPX consensus are in italics, and charged amino acids are in bold face. The kink site is contained within braces. In domain 3, amino acids not fitting the $(SP)_x$ consensus are in italics, and charged amino acids are in bold face. In domain 4, the canonical N-glycosylation sites are shaded.

three modules earlier, by a PPSPK, where the R and the K are the only two charged amino acids in domain 2. The F represents the only aromatic amino acid in either the A2 or the GP1 PPSPX domains, and the N represents the only asparagine.

The PPSPX domain of GP1 is followed by domain 3, which carries the canonical $(SP)_x$ repeat found in all *C. reinhardtii* wall proteins to date (*9*, *28*, *36*−*38*). In this case the repeat is $(SP)_{19}$, where the S is replaced by I and K at two positions.

Since GP1 forms a PPII helix (Figure 4A) with 3.34 amino acids per nanometer, we can point out three correlations between its amino acid sequence and its morphology. First, the 336 amino acids in domains 2 and 3 would generate a 100-nm fiber, concordant with the 100-nm length of GP1 measured in deep-etch replicas (*20*). Second, we suggest that the 19 SP repeats of domain 3, predicted to be 11 nm in length, generate the short narrow neck that emerges from the head (Figure 3, arrows), while the PPSPX repeats of domain 2, predicted to be 89 nm in length, generate the wider shaft. And third, the major sequence discontinuity in domain

2 lies 95 amino acids away from the start of the head domain, which would correspond to 28 nm of the PPII helix. The marked kink in the fibrous portion of the GP1 protein (Figure 3) has been measured to lie 28 nm from the head (*20*). We propose, therefore, that the kink arises at the site of the discontinuity in the PPSPX sequence.

DISCUSSION

HRGP Gene Families. The analysis of gene families can provide important evolutionary insights: the persistence and spread of a family through a lineage indicate that its conserved motifs participate in key cellular activities, and the variations on the theme can be analyzed for their contributions to particular phenotypes that have presumably been sculpted by natural selection.

The HRGP gene family is particularly suited to this kind of analysis because one of its conserved ideas—the formation of a linear glycosylated PPII helix—allows the investigator to glean considerable information about the protein product from its gene and amino acid sequence. For example, inspection of the sequence of a chimeric HRGP gene readily allows identification of "head" and "shaft" domains and allows the length of the shaft to be calculated, and analysis of glycosylation patterns in various HRGPs has generated proposals for "glycosylation rules" based on the deduced amino acid sequence (*10*, *33*, *39*).

A sufficient number of HRGP gene sequences from the Volvocales have now been obtained, including the two reported in this paper, that it becomes possible to recognize the existence of three HRGP subfamilies in this lineage. Their topology is described below, with the A2 and GP1 proteins considered in the appropriate contexts.

The Contiguous (Hydroxy)Proline Motif. One subfamily of volvocine HRGP sequences displays strings of from 2 to 40 contiguous prolines (*9*, *30*, *37*, *40*−*46*). The proline clusters may be interrupted by a few or by many other amino acids, often serines, and in three studies (*40*, *41*, *46*), the prolines were found to be converted into hydroxyproline in the context of at least 10 flanking amino acids [including lysine, an amino acid that appears to block hydroxylation in higher plants (*10*)]. Plant prolylhydroxylases require the PPII helix as a substrate (*47*−*49*), indicating that the clustered regions adopt this conformation. Sugar analysis of one of these proteins documents the presence of Ara and Gal in a 2:1 ratio (*46*).

The Alternating (Hydroxy)proline Motif. In a more regular subfamily, prolines alternate in dyads with other amino acids to generate strings of $(XP)_x$. The X is most often serine in *C. reinhardtii*, although the current sampling is biased in that an antibody against $(SP)_x$ was used to select many of the clones that have been analyzed (*36*). However, several genes with $(SP)_x$ repeats have been identified in other kinds of screens (*37*, *45*), indicating that this is likely to be a common dyad.

Certainly the most striking sequences in the $(XP)_x$ subfamily are found in the *vsp3* gene from *C. reinhardtii* and in its apparent homologue, *w6*, in *Chlamydomonas eugametos*: they encode long (171 and 134 amino acid) domains of $(SP)_x$ sequences, with frequent iterations of the variant SPSPSPKA (*50*). More commonly, $(SP)_x$ sequences are short and encountered in several locations in a gene sequence (*9*, *16*, *36*, *38*, *45*).

The sequence of the GP1 protein includes a highly invariant $(SP)_{19}$ motif, and we propose that this corresponds to the narrow "neck" domain of the mature protein (Figure 3, arrows). The neck domains of the GP2 and GP3 proteins are similarly narrow (*20*), suggesting that they may also represent $(SP)_x$ polymers. That GP2 contains SP_x motifs has been indirectly demonstrated by the ability of an anti-SP_{10} antibody and an anti-GP2 antibody to immunoprecipitate the same product, with the size expected for deglycosylated GP2, from an in vitro translation of vegetative mRNA (*36*).

The PPSPX Family. We report here a new HRGP subfamily, represented by the A2 and GP1 proteins, which displays long and regular repeats of the motif PPSPX. It is, of course, the case that most domains rich in S and P will include occasional PPSPX 5-mers (e.g., *37*, *40*, *41*), but the highly repetitive nature of these repeats in A2 and GP1 is clearly a distinctive idea. The high Hyp content of GP1 (*20*) indicates that most of the prolines in this domain are hydroxylated. That they form a PPII helix is documented directly by CD spectroscopy (Figure 4a) and inferred by the concordance between the predicted (100 nm) and measured (100 nm) length of the GP1 fibrous domain.

We show here that the PPII helix of GP1, like the PPII helix of extensins from carrot (*26*) and tomato (present report), is stabilized by its carbohydrate side chains. We also show that most of these are arabinogalactosides associated with the hydroxyprolines of GP1 and that many are unexpectedly complex, with up to 37% of the heteroarabinosides being branched, in contrast to the paucity of branched sugars in GP3. These observations generate two suggestions. First, the branched sugars may be responsible for the thicker caliber of the GP1 shaft, compared with the necks of GP1, GP2, and GP3, in rotary-shadowed replicas (*20*), possibly because the longer branched configurations trap more platinum. Second, the PPSPX motif may serve as a "glycosylation code" (*39*) for adding both straight and branched sugars to a maturing HRGP in *Chlamydomonas* and perhaps in the lineage in general. The HRGPs of higher plants display two kinds of glycosylation patterns: the side chains associated with the extensins are invariably straight, whereas the side chains associated with the arabinogalactan proteins are often branched (*7*, *33*, *34*, *51*, *52*). Shpak et al. (*39*) have recently shown in a tobacco system that contiguous Hyp residues in a PPII helix are glycosylated with short-chain arabinosides, whereas long-chain arabinosides are added to noncontiguous Hyps. Since the PPSPX motif carries both contiguous and noncontiguous Hyp residues, it may turn out that this sequence signals a mixed glycosylation pattern in the green algae as well.

GP1 Design. Our previous analysis of GP1 indicated that the protein carries a globular head and a fibrous shaft that is interrupted by a kink roughly 28 nm from the head (*20*). The present study confirms this organization at the molecular level and documents that the shaft adopts a PPII helix, the first such demonstration for an algal protein.

The molecular design of GP1 is consonant with its configuration in the *C. reinhardtii* cell wall. As noted earlier, this wall possesses a crystalline layer, called W6A, wherein the proteins GP2 and GP3 self-assemble to form a 24 × 28 nm lattice of regular parallelograms (*18*, *22*, *53*). The GP1 protein is itself incapable of self-assembly; instead, its incorporation into the wall is dependent on its ability to

interact with binding sites displayed by the GP2/GP3 lattice, the result being an overlying double-stranded hexagonal weave called W6B (*18*, *22*). Images of this weave adsorbed to mica surfaces (*22*) indicate that each GP1 head binds to one position on the W6A lattice and each kink binds to a second position, 28 nm from the first, to set up the basic parameters of the W6B layer. As GP1 proteins fill up these available binding sites, their shaft domains associate to form the remaining double-stranded sides of the hexagonal weave.

Although little is yet known about the molecular interactions involved in W6B assembly, the structure of the GP1 gene generates some predictions. First, it seems likely that the head domain will prove to carry binding motif(s) that allow specific recognition of the W6A lattice. Once this site is "tacked down", the arginine in the kink would represent the sole charged moiety in an otherwise sugar-insulated shaft, meaning that if a negatively charged node were present in the W6A lattice some 28 nm away from the initial binding site, the kink would be likely to bind to it and the basic pattern of the W6B layer would be established.

Of the volvocine cell walls analyzed to date, GP1 is peculiar to *C. reinhardtii*: even though *Gonium pectorale* (*54*) and *Volvox carteri* (*22*) form crystalline W6A layers with very similar parameters, these carry no overlying W6B layers and contain no polypeptides with the electrophoretic properties of GP1. Thus not only is GP1 an "epiphyte" of the W6A layer, it is possibly an epiphyte that evolved in just one evolutionary lineage. Since *C. reinhardtii* is primarily found in soils whereas *Gonium* and *Volvox* are found in ponds, the acquisition of the GP1-containing W6B may represent an adaptation wherein water becomes trapped in the hexagonal weave, with its rich endowment of branched carbohydrates, and staves off the hazard of soil dehydration.

Shafts with Kinks. The HRGPs are optimized as components of the extracellular matrix in that they are fibrous (thereby space-filling) and glycosylated [thereby resistant to proteases (*40*)]. The sugar residues also stabilize the PPII conformation and hence help to maintain the fibrous configuration (*26*, *55*, and present study). Although the carbohydrates may well mediate facets of wall assembly, this sugar-coated design introduces a cost in that such a fiber is unable to display its component amino acids to other matrix molecules and hence is compromised in its capacity to participate in molecular interactions. The globular domains of chimeric HRGPs are not so constrained, but most of the HRGPs in higher plants are not chimeric, yet they form covalent bonds with one another and presumably interact with other cell-wall components as well. How is this accomplished?

The structure of the fibrous domain of the GP1 protein provides a model. Embedded in the regular PPSPX domain is the PPPPPRPPFPANTPM sequence which, as noted above, occurs at the location of the pronounced kink in the fibrous shaft of the molecule. As illustrated in Figure 3, this kink is not an invariant feature of the isolated protein—indeed, the shaft can be straight (Figure 3A) or kinked to various extremes (Figure 3B−D), as if the region had the properties of a hinge. By contrast, the kink is always present, and at an invariant angle, when the protein is associated with the W6A lattice (*22*).

Darsey and Mattice (*56*) predict that PPII helices will make sharp turns in regions where the helix is not stable, and three considerations suggest that the helix will not be stable in the PPPPPRPPFPANTPM domain.

(1) Kieliszewski and Lamport (*10*) note that prolines are never hydroxylated next to phenylalanines or tyrosines. If this "rule" extends to algae—possibly because the bulky aromatic rings themselves destabilize the helix so that the prolylhydroxylases are not active—then the F in this sequence might generate several nonhydroxylated, and hence nonglycosylated, prolines that would form a deformable "weak spot" in the helix. Darsey and Mattice (*56*) also note that the pyrrolidine ring of proline is more flexible than that of hydroxyproline, further contributing to the tendency to kink. It is perhaps relevant that F and Y are disfavored in the PPII-helical domains of bacterial and animal proteins (*57*).

(2) The shift from the PPSPX to the PPPPPRPPFP motif may also signal a different pattern of glycosylation (*39*), one that provides a less robust stabilization of the helical configuration.

(3) Creamer (*58*) predicts, from computer simulations, that PPII helices will be stable when prolines are either contiguous or else alternate with single "guest" amino acids (other than glycine). However, if two or more guests are inserted into the chain, the PPII configuration is predicted to propagate only through the first residue, and in general, a proline can influence only the preceding residue in an oligopeptide to adopt a PPII conformation. The GP1 neck and shaft domains carry either contiguous prolines or single guests flanked by prolines *except* for the ANT sequence in the PPPPPRPPFP*ANT*PM domain (Figure 6B). The ANT residues may therefore also destabilize the helix and thereby contribute to forming the kink in the GP1 shaft.

Our interest in the kink domain of GP1 is whetted by the observation that extensins from several dicots display kinks along their lengths that are strikingly similar to the kink in the GP1 shaft (*59*, *60*), kinks that persist after the proteins are deglycosylated (*55*, *60*). The amino acid sequences of extensins consist of modular repeats that begin with SPPPP and are followed by anywhere from 2 (e.g., SPPPP*EH*) to 15 (e.g., SPPPP*KHSPAPEHHYKYKYK*) guest amino acids (*61*). Creamer's study predicts that these guests would destabilize the helix, generating a tendency to kink (*56*), a tendency that would expose the Y, K, E, and H residues that are believed to participate in interchain cross-linking (*8*, *61*). Because different modules might be expected to destabilize and hence kink at different times and with different probabilities, this would explain why the positions of the kinks display no regular pattern (*59*). Such a design would also give extensins the option of forming covalent interactions at variable rather than at fixed positions, a flexibility that might undergird the variable pore size observed in higher plant cell walls (*62*, *63*).

Analysis of cell-wall gene and protein sequences with the Creamer predictions in mind yields some interesting patterns. Some proteins—e.g., A2 (present study) and the CELP (*15*) and class I (*14*) proteins from tobacco flowers—display undeviant polyproline and/or $(PX)_x$ patterns, suggesting that they will generate straight shafts only, whereas others—e.g., the maize pollen-specific Pex1 protein (*11*) and the histidine-rich and threonine-rich HRGPs (*64*)—are quite sloppy in this regard, and indeed the latter show little evidence of PPII helix by CD spectroscopy despite a high overall content of Hyp (*64*). Of particular note is the *RK* dyad embedded in an

otherwise invariant polyP/PX domain (77 residues) of the SSG185 ECM protein of *V. carteri* (*40*). The homologous HRGP domain in a second *V. carteri* ECM protein, pherophorin-S, lacks this dyad (*43*). SSG185 binds a highly sulfated polysaccharide at a position corresponding to the *RK* dyad (*40*) whereas pherophorin-S does not associate with such a polysaccharide, suggesting that the *RK* guest may impose sufficient instability on the PPII helix to expose its positively charged groups to the negatively charged sugar.

Extending this concept to the many animal and bacterial proteins that carry PPII domains (reviewed in refs *57* and *65*)—including a variety of transcriptional regulators and proteins involved in signal transduction cascades—it is possible that kink-like transitions in these regions, mediated by guest amino acids, might help to expose ligands to their targets.

ACKNOWLEDGMENT

We dedicate this paper to the treasured memory of Steven Adair (1946−1990). We thank M. D. Brownleader for the gift of tomato extensin, C. Luo for performing the Northern analysis, and J. E. Heuser for generating Figure 3.

REFERENCES

1. Lamport, D. T. A., and Northcote, D. H. (1960) *Nature 188*, 665−666.
2. Lamport, D. T. A. (1967) *Nature 216*, 1322−1324.
3. Miller, D. H., Lamport, D. T. A., and Miller, M. (1972) *Science 176*, 918−920.
4. Showalter, A. M. (1993) *Plant Cell 5*, 9−23.
5. Sommer-Knudsen, J., Bacic, A., and Clarke, A. E. (1997) *Phytochemistry 47*, 483−497.
6. Cassab, G. I. (1998) *Annu. Rev. Plant Physiol. Mol. Biol. 49*, 281−309.
7. Serpe, M. D., and Nothnagel, E. A. (1999) *Adv. Bot. Res. 30*, 207−289.
8. Schnabelrauch, L. S., Kieliszewski, M. J., Upham, B. L., Alizedeh, H., and Lamport, D. T. A. (1996) *Plant J. 9*, 477−489.
9. Waffenschmidt, S., Woessner, J. P., Beer, K., and Goodenough, U. W. (1996) *Plant Cell 5*, 809−820.
10. Kieliszewski, M. J., and Lamport, D. T. A. (1994) *Plant J. 5*, 157−172.
11. Rubenstein, A. L., Broadwater, A. H., Lowrey, K. B., and Bedinger, P. A. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 3086−3090.
12. Rubenstein, A. L., Marquez, J., Suarez-Cervera, M., and Bedinger, P. A. (1995) *Plant Cell 7*, 2211−2215.
13. Baldwin, T. C., Coen, E. S., and Dickinson, H. G. (1992) *Plant J. 2*, 733−739.
14. Goldman, M. H. de S., Pezzotti, M., Seurinck, J., and Mariani, C. (1992) *Plant Cell 4*, 1041−1051.
15. Wu, H.-M., Zou, J., May, B., Gu, Q., and Cheung, A. Y. (1993) *Proc. Natl. Acad. Sci. U. S. A. 90*, 6829−6833.
16. Woessner, J. P., and Goodenough, U. W. (1994) *Protoplasma 181*, 245−258.
17. Sumper, M., and Hallmann, A. (1998) *Int. Rev. Cytol. 180*, 51−85.
18. Goodenough, U. W., and Heuser, J. E. (1985) *J. Cell Biol. 101*, 1550−1568.
19. Hills, G. J., Phillips, J. M., Gay, M. R., and Roberts, K. (1975) *J. Mol. Biol. 96*, 431−434.
20. Goodenough, U. W., Gebhart, B., Mecham, R. P., and Heuser, J. E. (1986) *J. Cell Biol. 101*, 924−942.
21. Adair, W. S., Steinmetz, S. A., Mattson, D. M., Goodenough, U. W., and Heuser, J. E. (1987) *J. Cell Biol. 105*, 2373−2382.
22. Goodenough, U. W., and Heuser, J. E. (1988) *J. Cell Sci. 90*, 717−733.
23. Goodenough, U. W., and Heuser, J. E. (1988) *J. Cell Sci. 90*, 735−750.
24. Ferris, P. J., and Goodenough, U. W. (1994) *Cell 76*, 1135−1145.
25. Adair, W. S., and Apt, K. E. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 7355−7359.
26. van Holst, G.-J., and Varner, J. E. (1984) *Plant Physiol. 74*, 247−251.
27. Kilz, S., Waffenschmidt, S., and Budzikiewcz, H. (2000) *J. Mass Spectrom. 35*, 689−697.
28. Sreerama, N., and Woody, R. W. (1994) *Biochemistry 33*, 10022−10025.
29. Johnson, W. R. (1988) *Annu. Rev. Biophys. Chem. 17*, 145−166.
30. Woessner, J. P., and Goodenough, U. W. (1989) *Plant Cell 1*, 901−911.
31. Brownleader, M. D., Byron, O., Rowe, A., Trevan, M., Welham, K., and Dey, P. M. (1996) *Biochem. J. 320*, 577−583.
32. Kieliszewski, M., and Orlando, R. (1997) *Phytochemistry 45*, 9−14.
33. Kieliszewski, M., O'Neill, M., Laykam, J., and Orlando, R. (1995) *J. Biol. Chem. 270*, 2541−2549.
34. Lamport, D. T. A., and Miller, D. H. (1971) *Plant Physiol. 48*, 454−456.
35. O'Neill, M. A., and Roberts, K. (1981) *Phytochemistry 20*, 25−28.
36. Woessner, J. P., and Goodenough, U. W. (1992) *Plant Sci. 83*, 65−76.
37. Kurvari, V. (1997) *Mol. Gen. Genet. 256*, 572−580.
38. Suzuki, L., Woessner, J. P., Uchida, H., Kuroiwa, H., Yuasa, Y., Waffenshcmidt, S., Goodenough, U. W., and Kuroiwa, T. (2000) *J. Phycol. 36*, 571−583.
39. Shpak, E., Leykam, J. F., and Kieliszewski, M. J. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 14736−14741.
40. Ertl, H., Mengele, R., Wenzl, S., Engel, J., and Sumper, M. (1989) *J. Cell Biol. 109*, 3493−3501.
41. Ertl, H., Hallmann, A., Wenzl, S., and Sumper, M. (1992) *EMBO J. 11*, 2055−2062.
42. Huber, O., and Sumper, M. (1994) *EMBO J. 13*, 4212−4222.
43. Godl, K., Hallmann, A., Wenzl, S., and Sumper, M. (1997) *EMBO J. 16*, 25−34.
44. Amon, P., Haas, E., and Sumper, M. (1998) *Plant Cell 10*, 781−789.
45. Rodriguez, H., Haring, M. A., and Beck, C. F. (1999) *Mol. Gen. Genet. 261*, 267−274.
46. Ender, F., Hallmann, A., Amon, P., and Sumper, M. (1999) *J. Biol. Chem. 274*, 35023−35028.
47. Tanaka, M., Sata, K., and Uchida, T. (1981) *J. Biol. Chem. 256*, 11397−11400.
48. Sauer, A., and Robinson, D. G. (1985) *Planta 164*, 287−294.
49. Blankenstein, P., Lang, W. C., and Robinson, D. G. (1986) *Planta 169*, 238−244.
50. Woessner, J. P., Molendijk, A. J., van Egmond, P., Klis, F. M., and Goodenough, U. W. (1994) *Plant Mol. Biol. 26*, 947−960.
51. Fincher, G. B., Stone, B. A., and Clarke, A. E. (1983) *Annu. Rev. Plant Physiol. 34*, 47−70.
52. Knox, J. P. (1995) *FASEB J. 9*, 1004−1012.
53. Roberts, K. (1974) *Philos. Trans. R. Soc. London, B: Biol. Sci. 268*, 129−146.
54. Adair, W. S., and Appel, H. (1990) *Planta 179*, 381−386.
55. Stafstrom, J. P., and Staehelin, L. A. (1986) *Plant Physiol. 81*, 242−246.
56. Darsey, J. A., and Mattice, W. L. (1982) *Macromolecules 15*, 1626−1631.
57. Stapley, B. J., and Creamer, T. P. (1999) *Protein Sci. 8*, 587−595.
58. Creamer, T. P. (1998) *Proteins 33*, 218−226.
59. Stafstrom, J. P., and Staehelin, L. A. (1986) *Plant Physiol. 81*, 234−241.
60. Heckman, J. W., Jr., Terhume, B. T., and Lamport, D. T. A. (1988) *Plant Physiol. 86*, 848−856.

61. Chen, J., and Varner, J. E. (1985) *EMBO J. 4*, 2145−2151.
62. McCann, M. C., Wells, B., and Roberts, K. (1990) *J. Cell Sci. 96*, 323−334.
63. Cooper, J. B., Heuser, J. E., and Varner, J. E. (1994) *Plant Physiol. 104*, 747−752.
64. Kieliszewski, M. J., Kamyab, A., Leykam, J. F., and Lamport, D. T. A. (1992) *Plant Physiol. 99*, 538−547.
65. Williamson, M. P. (1994) *Biochem. J. 297*, 249−260.

BI0023605