Plant Molecular Biology Update

## The sequence of a hydroxyproline-rich glycoprotein gene from Sorghum vulgare

Regina Raz<sup>1</sup>, Claude Crétin<sup>2</sup>, Pere Puigdomènech<sup>1</sup> and José A. Martínez-Izquierdo<sup>1,\*</sup> <sup>1</sup>Departamento de Genética Molecular, CID-CSIC, Jordi Girona 18, 08034 Barcelona, Spain (\* author for correspondence); <sup>2</sup>Laboratoire de Physiologie Végétale Moléculaire, Université de Paris-XI, Orsay, France

Received 16 October 1990; accepted 30 October 1990

Key words: hydroxyproline-rich glycoprotein, sorghum, Gramineae, 3' intron

Cell wall hydroxyproline-rich glycoproteins (HRGPs) constitute a family of proteins found as structural components in the plant primary cell walls [1]. These proteins are characterized by the high content in hydroxyproline/proline, basic amino acids, serine and tyrosine [2]. The most studied HRGPs, so far, have been those from dicot species (also called extensins), although a maize HRGP has recently been characterized at the protein and genomic levels [3, 4]. This maize HRGP has the main features of dicot extensins, although its main repeated motif is different from the dicot ones and it is very rich in threonine [3]. In common with the carrot HRGP gene [5] the maize gene contains an intron in the 3'-noncoding region [4], but the GC content of the maize gene is significantly higher than for carrot and tobacco genes [4-6]. In this study we report on the sequence of a HRGP gene from sorghum which resembles maize HRGP gene (60% overall homology), indicating the existence of a Gramineae HRGP subfamily belonging to the plant cell wall HRGP family. Moreover, genomic Southern analysis of different graminaceous DNAs with two distinct DNA genomic probes from maize and sorghum reveal two different classes of similar sequences: the first one common to both probes and the other similar to the sorghum probe, suggesting the existence of another family of HRGP-related genes in Gramineae.

A genomic library of Sorghum vulgare made in  $\lambda$  EMBL4 was screened with a maize HRGP genomic probe 842 bp long, containing part of the 5' and coding regions [4].  $\lambda$  DNA from one positive clone was restricted with several enzymes and subjected to Southern/hybridization analysis. A 4.5 kbp Eco RI/Sac I fragment which showed strong hybridization with the maize probe was cloned in M13 vectors and sequenced by the dideoxy method on both strands. Nucleotide sequence analysis showed that such a fragment contains a single long open reading frame that codes for a protein of 283 amino acids. 2 kbp of nucleotide sequence of the gene and the derived amino acid sequence are presented in Fig. 1. There is a potential TATA box 77 bp upstream of the ATG start codon and a putative canonical polyadenylation signal 448 bp downstream the stop codon (underlined). There is a 122 bp long potential intron in the 3'-untranslated region (nucleotides in lower case) in the same position (ca. 25 bp after stop codon) as the maize HRGP one, flanked by AGgt/agGT intron consensus

The nucleotide sequence data reported will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under the accession number X56010.

868 995	GGT Caa	CTC( tcg	CAC	:TC1 :cg1	TGCT tttt	GTC/	AGTC1 gtgt1	TCGG: tgca:	AGCA gaag	TTTA GTGA	Ggta TCGA	cgtg TCGA/	ctaca AGAG	aaga TTTG	atgg TGTC	agca CTAG	tgca CCAG	cata CCGG	cata CCAA	cata GATG	cata AGCT	catt GCTG	ttct ATGA	atat TGAT	atgc GAGA	ttgt GGCG	gtca AGGA	aatg GCGA	tatc GTCG	atac TACC	tgta .CCTC	ctca CCTC	ctaa CTCT	
	Ala	Pro	o Th	nr f	Pro	Pro	Thr	Tyr	Thr	Pro	Pro	Val	Ser	His	Thr	Рго	Ser	Ser	Pro	Pro	Tyr	Tyr						283						
769	GCG	сса	: AC	:c (	CCG	CCG	ACA	TAC	ACG	CCG	CCG	GTG	TCG	CAC	ACC	ccc	AGC	TCG	CCG	сст	TAC	TAC	TAG	ACG	AAGA	AGTG	ATGA							
_	Thr	Ser	• Th	nr I	Pro	Thr	His	Рго	Lys	Ρгο	Thr	Рго	His	Thr	Pro	Ile	Рго	Lys	Рго	Thr	Pro	Pro	Thr	Туг	Lys	Рго	Ala	Pro	Lys	Рго	Ser	Pro	Pro	256
673	ACC	TC/	A AC	:T (	CCG	ACG	CAC	сст	AAG	CCG	ACG	CCG	CAC	ACG	сст	ATC	ccc	AAG	ссс	ACC	200	CCA	ACC	TAC	AAG	CCG	GCA	CCG	AAG	CCG	AGC	CCG	CCG	
	Pro	Thr	Pr	·0	Pro	Val	Туг	Thr	Pro	Ser	Pro	Lys	Pro	Pro	Lys	Pro	Ser	Pro	Рго	Thr	Tyr	Thr	Pro	Thr	Pro	Lys	Pro	Рго	Ala	Thr	Lys	Pro	Pro	224
577	CCG	ACT	r cc	:G (	CCG	GTG	TAC	ACA	CCG	AGC	ссс	AAG	CCA	CCG	AAG	CCG	TCC	CCG	CCG	ACG	TAC	ACG	CCG	ACC	CCG	AAG	CCA	CCG	GCT	ACC	AAG	CCG	ccc	
	Рго	Ly	s Pr	·0 '	Thr	Pro	Рго	Val	Tyr	Thr	Рго	Asn	Pro	Lys	Pro	Pro	Val	Thr	Lys	Pro	Pro	Thr	His	Thr	Pro	Ser	Рго	Lys	Pro	Pro	Thr	Ser	Lys	192
481	ссс	AA	5 CC	:G /	ACC	CCG	CCG	GTG	TAC	ACG	CCG	AAC	000	AAG	CCA	CCG	GTT	ACC	AAG	сст	CCG	ACA	CAC	ACG	CCG	AGC	ссс	AAG	сст	CCG	ACA	тсс	AAG	
	Pro	Th	r T)	r I	Pro	Thr	Pro	Lys	Pro	Pro	Ala	Thr	Lys	Pro	Pro	Thr	Pro	Pro	Val	Tyr	Thr	Pro	Ser	Pro	Lys	Pro	Pro	Val	Thr	Lys	Pro	Рго	Thr	160
385	сст	AC	C TA	IC I	CCG	ACC	ссс	AAG	CCA	CCA	GCT	ACC	AAG	CCT	CCG	ACA	CCT	100	GTG	TAC	ACT	200	AGC	CCC	-	CCA	200	GTT	ACA	AAG	CCT	CCT	ACA	
	Ser	Pre	5 L)	/s I	Pro	Lys	Ser	Pro	val	١yr	Pro	PLO	Pro	PLO	LYS	ALA	ser	INC	Pro	PLO	inr	iyr	Inc	Pro	ser	PLO	LYS	800	Pro	жіа	101	LYS	PT'0	120
289	AGC	CC.	T AF	G I	ссс	AAG	TCG	CCG	GTC	TAC	CCT	ccc	CCG	CCG	AAA	GCC	AGT	ACT	CCT	CCG	ACC	TAC	ACC	CCG	AGC	000	AAG	CCT	CCG	GCG	ACC	AAG	CCG	179
	Lys	61		15	Lys	Pro	186	Pro	Pro	101	T <b>y</b> r	101.	PTU	361	FIU	LYS	PIQ		FIU	FIU	FIU	ALG		FIU	L ¥ 5	10		10	FIU	••••			r i o	
193	AAG	GA	G CA	AC A	AAG	CCG	ACT	CCG	000	ACC	TAC	ACC	000	AGC	CCC	AAG	CCC	ACG	CCG	CCG	CCG	GCT	ACT	CCC	AAG	CCG	ACG	CCG	CCC	ACC	TAC	ACG	CCG	96
	uty			U	1 1 1 1	FIQ	1111	- 10	.,.				-,3	.,	1.0	.,.			- / 0				2,0	•••,		•••		-,-			-/-			
97			C C(	CG I	ACG Thr	CCG	ACG	CCG	AAG	CCG Pro	CCG	GCC	AAG	GGG	Pro	AAG LVS	Pro	GAG	AAG Lvs	Pro	Pro	Thr	AAG LVS	GLV	UAC His	GLV	His	LYS	Pro	Glu	LVS	Pro	Pro	64
																														~ • ~		~~~		
1	Met	Me	t Gl	si ly∣	Gly	Lys	Ala	Ala	Leu	Leu	Leu	Ala	Leu	Val	Ala	Val	Thr	Leu	Ala	Val	Val	Glu	Ile	Gln	Ala	Asp	Ala	Gly	Tyr	Gly	Tyr	Gly	Gly	32
				- 7					<b>6</b> 16	<b>6</b> 16	676		<b>CT</b> C	616	600	GTC	ACC	CTC	666	oto	010	GAG	ATC	CAG	1	GAC	200	222	тас	222	TAC	222	200	
- 127	ACC	CAG	CGGG	CAT	CCG	AGGC	CCCC	ACCC	CCAC	CCCC	ACCC	CTTC	CTCCI	STGT	ATAA	AAGCO	GTG	CAG	GTG	GCG	rctci	тссто	CACAC	CTGAC	CTGC/	CCAG	AACA	AGAC	SCTTO	AGAC	ITGAC	GAGTO	GAGG	
- 254	100	CUT		CAC	ATG	GGAC.	ATGG	CTGC	CTGA	TGAA	TGCC	TETE	GCCGI	CACTI	GCCC		CCCT	10010	TGC		SCCCI	TCAC	TCCAT	CGG/	GGGG	100A0	GCTT	CACO	GAGAG	CCG	GCAC	GAC	CAG	
-508	CAC	ATG	GCTO		CITO	GCTA	CTTC	AGTG		GGGA	GTGT	ATGC	ATGCI	ACAC	AATA	ATGC	GGCC	I GCG1	CTGI	GTAC	CAC	AGAA/		CITI	'ATAC "CAAC	CAGGA	TATO		GACG	GAA	RECT	GCAU		
-635	AAA	AAA	AAT	TTG	GTT	TTTG	CGGT	GAAC	TAAA	CAAG	GCCA	TATA	ETAG	IGIG	CIGI	GIII	CCCC	TAC	SAAG	ATT	166	46100	31661	IGAI	LCAL	.6166		AILI	6666	GAGU	TOCA	CARP		

Fig. 1. Nucleotide and derived amino acid sequence of the sorghum HRGP gene. Nucleotides are numbered from the A of the start codon (+1) on the left and amino acids on the right. Putative TATA box and polyadenylation signal are underlined. The intron in the 3'-untranslated region is in lower case. The putative site of processing of the mature protein between Ala 24 and Asp 25 is indicated by a vertical arrow.

sequences. When the different functional regions of the gene, namely 5', coding, intron and 3' (intron) regions are compared separately with the corresponding regions of the maize HRGP gene, no significant variation was observed with the overall homology value (60%) except for the 3' (intron) region with 66% homology. The higher homology observed in that region can be explained by the absence of long insertion/deletions shown in the other three regions.

The amino acid composition of the sorghum HRGP is similar to the maize one: rich in proline/hydroxyproline, lysine, tyrosine and threonine [3, 4], although the threonine content is reduced (16 versus 22%) and the valine content is doubled (3.9 versus 1.8%). In this respect the amino acid composition of sorghum HRGP is closer to carrot HRGP than the maize one [3]. However, the sorghum HRGP should be considered as belonging to the same HRGP family as

the maize one, because the main repetitive motifs of the maize HRGP, one undecapeptide and two pentapeptides, are essentially the same in the sorghum HRGP. The loss of some of those motifs results in length reduction of the sorghum HRGP which is 45 amino acid shorter than the maize one [4]. The N-terminal part of the protein, rich in hydrophobic amino acids, has the characteristics of a signal peptide with a potential cleavage site (by comparing with the maize HRGP) between the residues Ala and Asp (marked by a vertical arrow in Fig. 1). Previous results have shown that maize HRGP gene is also present in species of the Panicoideae subfamily of Gramineae [4], as it is the case of the sorghum gene reported here. Figure 2A shows a genomic Southern hybridized with the maize probe described above. It can be seen that related sequences to maize HRGP gene are also present in species of Festucoideae subfamily, such as wheat, barley and rice, although

366



Fig. 2. Genomic Southern blots of several graminaceous DNAs hybridized with HRGP probes. Species analyzed are W64A maize inbred line (W), Palomero Toluqueño maize variety (PT), Zea diploperennis teosinte (T), sorghum (S), wheat (Wh), barley (Ba), rice (Ri) and tobacco (To), as external dicot control. DNAs ( $10 \mu g$  per lane) were digested with *Hind* III. The same filter was hybridized with two different probes: the *Hind* III-Sna BI 842 bp long fragment from maize inbred line AC1503 (panel A) and the Alu I-HinfI 168 bp long fragment from sorghum. The position of size markers (kbp) is indicated.

they are not present in dicot species such as tobacco. There is a region within the sorghum gene which shows low homology to the maize one, compared to the overall homology between the two genes. When a 168 bp long probe (nucleotides 411-579 in Fig. 1) encompassing that region was used, the same bands which appeared with the maize genomic probe, lit up (Fig. 2B). In addition, some other new bands can be clearly seen for the cases of teosinte, sorghum and barley DNAs (Fig. 2B). These results suggest the existence of another HRGP-related gene family in Gramineae. Whether those genes code for proteins belonging to the same family as the dicot and monocot cell wall HRGPs, or to other hydroxyproline-rich glycoproteins [1] must be elucidated.

## Acknowledgements

The authors are indebted to Mr. Patricio Gomez for technical assistance. This work has been supported by grants from PLANICYT (BIO 88/0242) and EECC (BAP 374). R.R. is recipient of a fellowship from PFPI.

## References

- Showalter AM, Varner JE: Plant hydroxyproline-rich glycoproteins. In: A. Marcus (ed), The Biochemistry of Plants, Vol. 15A, pp. 485-520. Academic Press, San Diego (1989).
- Varner JE, Lin L-S: Plant cell wall architecture. Cell 56: 231–239 (1989).
- Kieliszewski MJ, Leykam JF, Lamport DTA: Structure of the threonine-rich extensin from Zea mays. Plant Physiol 92: 231–239 (1990).
- Stiefel V, Ruiz-Avila L, Raz R, Valles MP, Gomez J, Pages M, Martinez-Izquierdo JA, Ludevid MD, Langdale JA, Nelson T, Puigdomenech P: Expression of a maize cell wall hydroxyproline-rich glycoprotein gene in early leaf and root vascular differentiation. Plant Cell 2: 785–793 (1990).
- Chen J, Varner JE: An extracellular matrix protein in plants: characterization of a genomic clone for carrot extensin. EMBO J 4: 2145–2151 (1985).
- Keller B, Lamb CJ: Specific expression of a novel cell wall hydroxyproline-rich glycoprotein gene in lateral root initiation. Genes Dev 3: 1639–1646 (1989).