Mol Gen Genet (1992) 233:252-259

MGG

Different mechanisms generating sequence variability are revealed in distinct regions of the hydroxyproline-rich glycoprotein gene from maize and related species*

Regina Raz[1], Matilde José[1], Andrés Moya[2], José Antonio Martínez-Izquierdo[1], and Pere Puigdomènech[1]

[1] Departamento de Genética Molecular, CID-CSIC, Jordi Girona, 18. E-08034 Barcelona, Spain
[2] Departamento de Genética, Universidad de Valencia, Dr. Moliner, 50. E-46100 Burjassot, Valencia, Spain

Received September 6, 1991 / December 10, 1991

Summary. The sequences of the genes coding for a hydroxyproline-rich glycoprotein from two varieties of maize (*Zea mays,* Ac1503 and W22), a teosinte *(Zea diploperennis)* and sorghum *(Sorghum vulgare)* have been obtained and compared. Distinct patterns of variability have been observed along their sequences. The 500 bp region immediately upstream of the TATA box is highly conserved in the *Zea* species and contains stretches of sequences also found in the sorghum gene. Further upstream, significant rearrangements are observed, even between the two maize varieties. These observations allow definition of a 5′ region, which is common to the four genes and is probably essential for their expression. The 3′ end shows variability, mostly due to small duplications and single nucleotide substitutions. There is an intron present in this region showing a high degree of sequence conservation among the four genes analyzed. The coding region is the most divergent, but variability arises from duplications of fragments coding for similar protein blocks and from single nucleotide substitutions. These results indicate that a number of distinct mechanisms (probably point mutation, transposon insertion and excision, homologous recombination and unequal crossing-over) are active in the production of sequence variability in maize and related species. They are revealed in different parts of the gene, probably as the result of the different types of functional constraints acting on them, and of the specific nature of the sequence in each region.

Key words: Sequence evolution – *Sorghum* – *Zea* – hydroxyproline-rich glycoproteins

Introduction

Maize is one of the plant species that has been most extensively manipulated by man and, at present, a large number of varieties are available. As a result of the analysis of this variability, the existing information on maize makes it one of the best known plants from the genetic point of view. Of particular importance was the discovery of transposons in maize as a well known source of genetic instability (McClintock B 1951; see Fedoroff 1989 for a review). For these reasons, maize is an appropriate system in which to study the mechanisms producing genetic variability in eukaryotes.

Hydroxyproline-rich glycoproteins (HRGP) are structural polypeptides present in plant cell walls. Their sequence is mostly formed by highly repeated elements. In dicotyledonous species, the most abundant HRGPs, the extensins, are encoded by multigene families (Showalter et al. 1985). Their main repetitive sequence element is SPPPP, where the proline residues are modified to hydroxyproline. Among monocotyledonous species, information is now available on HRGPs from maize and sorghum. In these two species, HRGPs are probably encoded by single genes (Stiefel et al. 1990; Raz et al. 1991). The maize HRGP sequence, as well as cDNA and genomic sequences coding for it, has recently been obtained (Stiefel et al. 1988, 1990; Kieliszewski et al. 1990). These studies have revealed that the main part of the sequence of this protein has a repetitive structure. The repeated unit is formed by a hexapeptide (PPTYTP) followed by either two or three pentapeptides (SPKPP-TPKPT or SPKPP-ATKPP-TPKPT, respectively).

In contrast to the extensins and other similar proteins, maize HRGP shows a high degree of conservation among all the elements repeated along its sequence (Stiefel et al. 1988). The presence of such a perfectly repetitive sequence in the maize *HRGP* gene made it an attractive system to study how such sequences evolve in comparison with those of the flanking regions. To this end, genomic clones for HRGP of a common maize

Offprint requests to: P. Puigdomènech

* The sequences reported in this paper have been deposited in the EMBL/GenBank Database (Bolt, Beranek, and Newman Laboratories, Cambridge, Mass., and EMBL, Heidelberg), accession nos. M36635 (maize Ac1503), X63134 (maize W22), X64173 (teosinte) and X56010 (sorghum)

variety (W22) and a teosinte *(Zea diploperennis)*, one of the supposed ancestors of maize, have been obtained, and their sequences have been compared to those already available from the maize variety Ac1503 (Stiefel et al. 1990) and sorghum (*Sorghum vulgare;* Raz et al. 1991). The variability observed in different regions of the *HRGP* gene has been analyzed. Some mechanisms which could explain how the variations have been produced are proposed.

Materials and methods

Teosinte genomic library construction. Plant material was obtained from Dr. A. Alvarez, Aula Dei, Zaragoza, Spain. Genomic DNA was prepared from adult leaves of teosinte *(Z. diploperennis)* as described (Burr and Burr 1981), and was purified through CsCl gradients (Maniatis et al. 1982). Teosinte genomic DNA was partially digested with *Sau*3A. Fragments 15–20 kb long were cloned into lambda Charon 35 (Loenen and Blattner 1983) following standard procedures (Maniatis et al. 1982).

Isolation and sequencing of genomic clones for maize and teosinte HRGP. The genomic libraries screened were of total DNA from teosinte leaves, and from W22 6-day-old seedlings (Clontech, Palo Alto, Calif.). Libraries were plated on *Escherichia coli* strains K802*recA*$^-$ or DL538 (kindly provided by F.R. Blattner, Madison, and A. Brandt, Copenhagen) and screened for the *HRGP* gene by standard methods (Sambrook et al. 1989). A cDNA clone, MC56 (Stiefel et al. 1988), and a *Hin*dII-*Sna*BI (852 bp long) genomic fragment (Stiefel et al. 1990) of the maize *HRGP* gene, were both used as probes. A positive clone was isolated from each of the two libraries, and their regions of homology to the probes were detected by blot hybridization analysis. Fragments were subcloned into pUC18/19 vectors and sequenced on both strands over a 4.5 kb region covering the coding region of the *HRGP* gene. DNA sequences were determined by the dideoxy method (Sanger et al. 1977) after subcloning in M13mp18/19 or pUC18/19, using the T7 sequencing kit (Pharmacia, Uppsala, Sweden). Some regions of high GC content were sequenced by the chemical degradation method (Maxam and Gilbert 1980) using the DNA Sequencing System (New England Nuclear, Boston, Mass.). Teosinte and W22 maize *HRGP* gene sequences were compared to maize Ac1503 and sorghum *HRGP* genes described elsewhere (Stiefel et al. 1990; Raz et al. 1991). Sequence analysis was carried out by means of MicroGenie software from Beckman (Queen and Korn 1984).

Phylogenetic analyses. The evolutionary distances between sequences were estimated according to two procedures: the number of nucleotide substitutions per site (and for coding sequences the number of non-synonymous substitutions per non-synonymous site) between pairs of sequences (Li et al. 1985b) and the number of nucleotide insertions or deletions per site (Tajima and Nei 1984). In this second procedure, the DNA divergence due to deletions and insertions is estimated with no consideration of DNA changes due to nucleotide substitution, by the formula:

$$\gamma = -2 \log_e P$$

where P, the fraction of nucleotides shared by sequences x and y, is given by

$$P = n_{xy}/\sqrt{n_x n_y}$$

n_{xy} being the number of nucleotide shared (excluding insertions and deletions) by sequences x and y, and n_x and n_y the number of nucleotides of sequence x and y, respectively.

The unrooted phylogenetic trees were constructed following the least squares method with the FITCH program of the PHYLIP package, Version 3.3 (Felsenstein 1990).

Results

Structure of the HRGP gene

The genes coding for the maize (Stiefel et al. 1990) and sorghum (Raz et al. 1991) HRGPs have recently been cloned and sequenced. The two genes code for very similar proteins but the degree of sequence similarity varies depending on the regions compared (Raz et al.

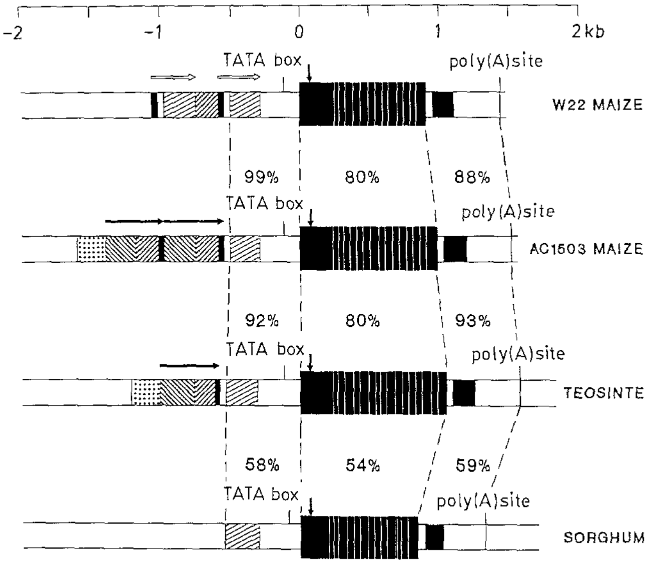

Fig. 1. Structure of the hydroxyproline-rich glycoprotein (HRGP)-encoding genes from maize varieties W22 and Ac1503, teosinte and sorghum. *Black boxes* represent the coding region and the *vertical white lines* indicate the limits of tandem repeats in the sequence. The cleavage site of the signal peptide is marked by a *vertical arrow*. Positions of the TATA box and the polyadenylation site are indicated. The intron located in the 3′ untranslated region is shown in *black*. Boxes in the 5′ region represent distinct blocks of sequence sharing more than 80% homology among different HRGP genes. *Horizontal arrows* indicate the position of sequences that are duplicated in the 5′ flanking region of the Ac1503 gene *(black)* and W22 gene *(white)*. The pairwise percentage of sequence identity in the 5′, coding, and 3′ regions is indicated between the graphics

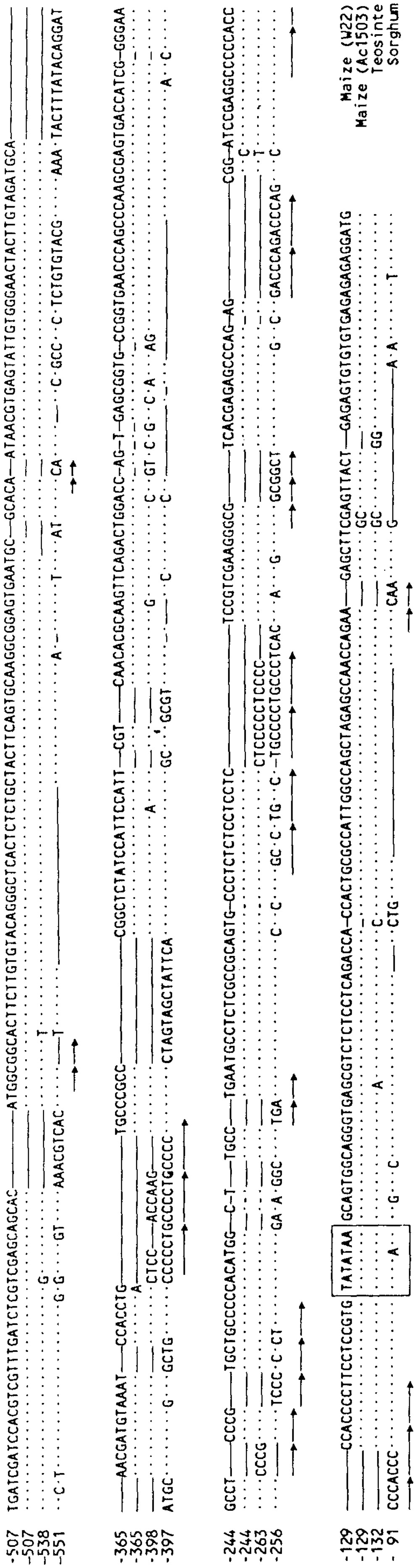
-507
-507
-538
-551
-365
-365
-398
-397
-244
-244
-263
-256
-129
-129
-132
- 91
TATATAA
Maize (W22)
Maize (Ac1503)
Teosinte
Sorghum

Fig. 2.

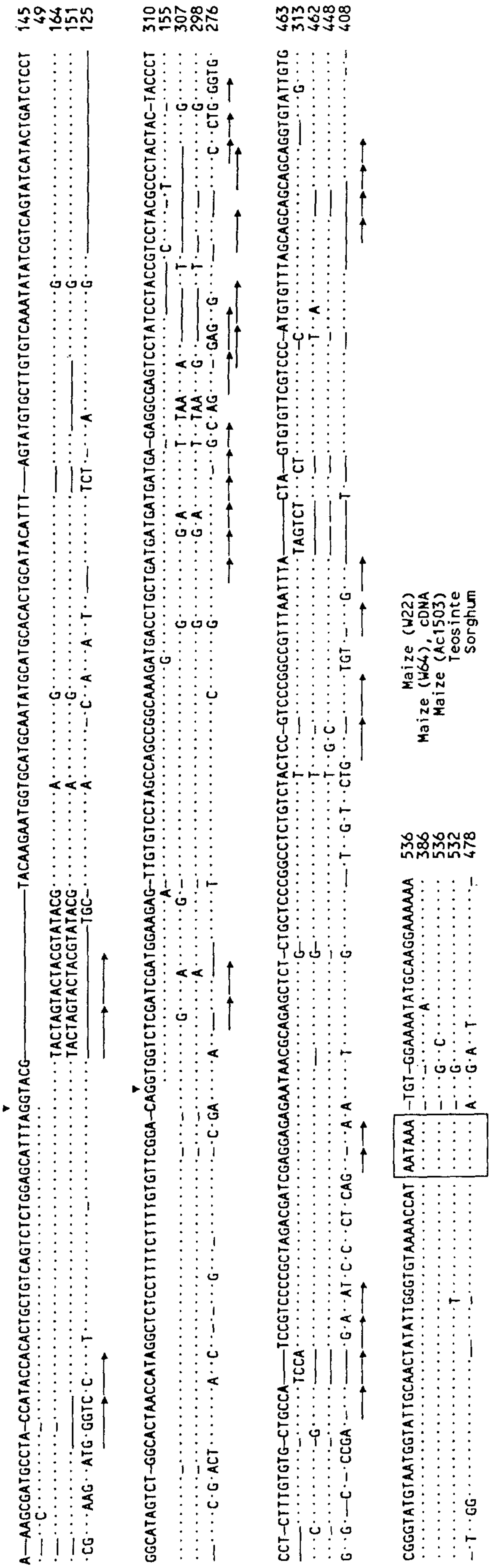
145
49
164
151
125
310
155
307
298
276
463
313
462
448
408
AATAAA
536
386
536
532
478
Maize (W22)
Maize (W64), cDNA
Maize (Ac1503)
Teosinte
Sorghum

Fig. 3.

1991). To investigate this observation further, the sequences of the *HRGP* genes from the teosinte *Z. diploperennis* and the maize W22 pure inbred line have been determined and compared with the two other sequences. The overall similarity among the four genes in the coding region and its vicinity is such as to allow alignment of the sequences and comparison of the different regions separately.

The structure of the four genes, studied over 2.5 kb, is shown diagramatically in Fig. 1. Two zones can be distinguished in the 5′ region of the gene, upstream of the site of translational initiation. The proximal region of ca. 500 bp (including the TATA box) is defined by the existence of a significant degree of similarity among the four genes compared. The distal region shows less than 40% similarity between the sorghum and the maize or teosinte sequences. This region is characterized by the presence of long repeated motifs in the sequences from the genus *Zea* (Fig. 1), which correspond to different overlapping motifs in each of the genes analyzed. The sequence of the coding region is highly repetitive and variable in length, and this is reflected in the structure of the protein, which shows varying copy numbers of peptide motifs, depending on the species or variety. This region is not interrupted by introns. However, an intron is present in the 3′ non-translated region, a situation which is quite unusual in plants but which is also found in other genes for HRGPs, such as carrot extensin (Chen and Varner 1985). In the gene coding for a proline-rich protein expressed in the lateral roots of tobacco (Keller and Lamb 1989), a sequence that follows the rules for plant introns is also present in the same position. Downstream of the polyadenylation site no significant homology can be observed between *Zea* and sorghum sequences.

5′ Region

As stated above, two zones can be distinguished in the 5′ region of the HRGP gene. The proximal region (indicated by vertical dotted lines in Fig. 1) shows a high degree of homology between the different genes. The alignment for maximum homology of the four genomic sequences studied is shown in Fig. 2. To align the sequences of maize varieties with that of teosinte, a number of insertions/deletions have to be introduced that correspond to duplications of short sequences (shown underlined with arrows in Fig. 2) present both in teosinte and in maize. By contrast, when the *Zea* and sorghum sequences are compared, zones of conserved sequences alternate with variable zones, generally coincident with the areas where most of the changes among *Zea* sequences are also observed. Therefore, besides single nucleotide substitutions, two types of variations can be observed in this region: either insertions/deletions corresponding to short sequences that may be repeated several times, or long insertions/deletions that appear on comparison of *Zea* sequences with that of sorghum. In some cases, these long insertions/deletions are flanked by short direct repeats, indicating that they may have been produced by a mechanism of homologous recombination, as has been proposed in other cases (Efstratiadis et al. 1980; Dixon and Hohn 1984).

Upstream of this region, long repeats and large sequence rearrangements are observed. They are shown schematically in Fig. 1. In Ac1503 maize, the repeats (marked by black arrows in Fig. 1) are 423 bp long, and a homologous sequence is present once in teosinte. In the W22 variety, another region of 308 bp (marked by white arrows in Fig. 1), including part of the previous one, is duplicated. No repetition or similarity with these repeated sequences is observed in sorghum. Upstream of this region, no significant similarity could be observed for more than 1700 bp, not even among maize varieties. In the W22 variety, a 39 bp sequence repeated 17 times, and not present in any of the other sequences, is found 100 bp upstream of the two large repeats (not shown).

3′ Untranslated region

Single nucleotide changes and insertions/deletions in zones of sequence duplications (Fig. 3) are a common feature shared between 3′ and 5′ regions of the *HRGP* genes. However, the number of single nucleotide changes in the 3′ region is, in general, higher than at the 5′ end. In the *Zea* species, they are restricted in the majority of cases to zones of sequence duplications, while in sorghum they are spread all over the sequence, except for some highly conserved stretches. Most of the insertions/deletions are generated by duplication of short sequences that may be repeated several times in any of the sequences studied. In contrast with the 5′ region, long insertions/deletions (longer than 15 nucleotides) are not observed in the 3′ untranslated end, except for the single intron, both in *Zea* species and in sorghum.

The sequence of a 200 bp region downstream of the polyadenylation site is practically identical between maize W22 and teosinte with only two small duplications and nine single nucleotide changes (results not shown). No significant similarity has been found in this region between any *Zea* species and sorghum.

Coding region

The coding region of the *HRGP* gene of the different varieties and species studied retains the features observed

←

Fig. 2. Alignment of nucleotide sequences of the 5′ region upstream of the initiation codon of the *HRGP* genes from maize (W22 and Ac1503), teosinte and sorghum. Gaps (indicated by *dashes*) have been introduced to obtain optimal alignment. *Points* represent identity with maize W22 nucleotide sequence. The putative TATA sequence is *boxed*. Motifs duplicated in any of the four sequences compared are indicated by *arrows*

Fig. 3. Alignment of nucleotide sequences of the 3′ region from the stop codon to the polyadenylation site of the *HRGP* genes from maize (W22 and Ac1503), teosinte and sorghum. Gaps (indicated by *dashes*) have been introduced to obtain optimal alignment. *Points* represent identity with maize W22 nucleotide sequence. Sites of intron splicing are marked *(vertical arrowtips)*. The putative polyadenylation signal is *boxed*. Motifs duplicated in any of the four sequences compared are indicated by *arrows*

```
                                            ↓
MAIZE (W22)     M-GGSGRAALLLALV--AVSLA-VEIQA DAGYGYGGGY TPTPTPATPTPKPEKPPTK GPKPDKPPKEHKPP-KEH GPKPEKPPKEHKPT    84
MAIZE (AC1503)  ·-·············VV·····-····· ·········· ··················· ·····------···-··· ··············    80
TEOSINTE        ·-····T········--·····-····· ····----·· ··················· ····E·········-··· ··············    80
SORGHUM         ·M·--·K········--··T··V····· ·········· -·····--------···A· ····E------···T·G· ·H············    70

MAIZE (W22)     PPTYTP SPKPT | PPTYTP TP-PP TPKPT | PPTYTP APTP- H-K--P TPKPT PT | PPTYTP SPKPP ----- TPKPT   146
MAIZE (AC1503)  ······ ····· | ······ ··T·- -···· | ······ ····- ·-·--· ····· ·· | ······ T···· ----- ·····   141
TEOSINTE        ······ ····· | ······ ··T·· ····· | ······ ····- ·-·PT· ····· ·· | ······ ····· ----- ·····   145
SORGHUM         ······ ····· | ··---· A-T-- -···· | ······ S·K·K SPVYP· P··AS -· | ······ ····· ATKPP -----   129

MAIZE (W22)     PPTYTP SPKPP ATKPP TPKPT | PPTYTP SPKPP TPKPT | ------ ----- | PPTYTP SPKPP ----- TPKPT   199
MAIZE (AC1503)  ······ ····· ----- ····· | ······ ····· ····· | ------ ----- | ······ ····· ATKPP ·····   194
TEOSINTE        ····A· ····· ····· ····· | ······ ····· ····· | PPTYTP SPKPT | ······ ····· ----- ·····   209
SORGHUM         --··-· T···· ····· ·---- | ------ ----- ----- | ------ ----- | ··V··· ····· VTKPP ·····   164

MAIZE (W22)     PPTYTP SPKPP ----- TPKPT | ------ ----- ----- | PPTYTP SPKPP ATKPP TPKPT | PPTYTP SPKPP ----- TPKPS   252
MAIZE (AC1503)  ······ ····· ----- ····· | PPTYTP SPKPP TPKPT | ······ ····· -·H·- ····· | ······ ····· ----- ····T   261
TEOSINTE        ······ ····· ATKPP ····· | PPTYTP SPKPP TPKPT | ······ ····· ··--- ---·· | ······ ····· ----- ····T   278
SORGHUM         ··V··· N···· VTK-- ----- | PPTHTP SPKPP TSKPT | ··V··· ····· ----- ---·S | ······ T···· ATKPP ·ST·T   229

MAIZE (W22)     ------ ----- ----- | PPTYTP SPKPP ----- TPKPT | PPTYTP TPKP- P-ATK | PPTYTP TPPVSHTPSPPPP-----YY    303
MAIZE (AC1503)  PPTYTP SPKPP TPKPT | ······ ····· ----- ····· | ······ ····- ·-··· | ······ ·············-----··    328
TEOSINTE        PPTYTP SPKPP TPKPT | ······ ····· ATKPP ····· | ······ ····- ·-··· | ······ ·············-----··    350
SORGHUM         ------ ----- HPKPT | ·--H·· ----- ----- I···· | ····K· A···S ·P·PT | ····-- ·········S···PPPPP··    283
```

Fig. 4. Comparison of the HRGP amino acid sequences among maize (W22 and Ac1503), teosinte and sorghum. *Points* represent identity with maize W22 amino acid sequence. Deletions are indicated by *dashes*. Cleavage site for the signal peptidase is shown by an *arrow*. *Vertical lines* indicate the repeated peptides characteristic of these HRGPs

in the maize cDNA and genomic sequences (Stiefel et al. 1988, 1990). At the N-terminus of the protein, including the signal peptide, the different proteins are very similar. The first part of the mature protein includes a small stretch of glycines and tyrosines, which is shorter in teosinte than in the other species studied. In the following region, which has a highly hydrophilic character, one of the most interesting features is that the motif KPPKEH may be present either two or three times depending on the species.

In the main part of the protein, characterized by a highly repeated sequence, a number of variations can be observed (Fig. 4). The sequences of the different *Zea* proteins, though varying by a small number of amino acid changes, differ mainly in the number of repeats present in each protein. However, the general structure of the protein is perfectly conserved. A larger number of amino acid changes is observed when the sorghum HRGP sequence is compared with those of *Zea* species.

The structure of the repetitive region of the protein consists of several repeated units formed by the hexapeptide PPTYTP, followed by either two or three pentapeptides having similarity among themselves, and probably sharing the same origin. In the W64 cDNA sequence (Stiefel et al. 1988), it was possible to observe that the presence of three pentapeptides correlates with the use of the AGC codon for the serine of the first pentapeptide. The second pentapeptide is absent when this serine is coded by a TCN codon (Stiefel et al. 1988). This phenomenon indicates that both classes of repeated units have probably been amplified independently. The third pentapeptide would have appeared, at a preliminary stage of evolution, in a unit having the serine of the first pentapeptide encoded by AGY. Both kinds of units probably have a common origin. The existence of two classes of codons for serine (AGY or TCN) may be explained if both are derived independently from a more ancestral one, probably a codon for threonine (ACN). In fact, all the pentapeptides in the *Zea* and sorghum HRGPs could have derived from a TPTPT motif by processes of duplication and mutation.

The complex evolution of HRGP *gene*

The evolutionary distance between HRGP genes from the different species was estimated according to two procedures: the number of nucleotide substitutions per site (and for coding sequences the number of non-synonymous substitutions per non-synonymous site) between pairs of sequences (Li et al. 1985b); and the number, γ, of nucleotide insertions or deletions per site (Tajima and Nei 1984).

Table 1 shows the abovementioned estimates for pairs of sequences and different regions of the *HRGP* gene. Contrary to the common observation that insertion/deletion events are less frequent than nucleotide substitution events, it is evident from our data that, irrespective of the region considered (except, in the case of sorghum, the 3′ sequence excluding the intron), insertions/deletions are always more frequent than substitutions, giving rise to higher distance estimates. This is particularly true for the repetitive coding region, where the γ values are the highest in the entire gene, comparable only with those of the 5′ region; the converse is observed for most other genes (Tajima and Nei 1984 and references therein). This indicates that mechanisms (i.e. unequal crossing-over) that produce insertions/deletions and reduce the accumulation of mutations are very active within the coding region.

The effects of such mechanisms on the *HRGP* gene should, obviously, distort the phylogenetic tree based exclusively on the accumulation of single nucleotide substitutions, or under a low rate of accumulation of insertions/deletions. To verify this, we have constructed un-

Table 1. Comparison of sequence divergence in different regions of the hydroxyproline-rich glycoprotein gene from maize (W22 and Ac1503), teosinte and sorghum

Region	Subregion	Species	Species			
			W22	Ac1503	Teosinte	Sorghum
5′	All	W22	–	0.0060	0.0214	0.1273
		Ac1503	0.1173	–	0.0182	0.1267
		Teosinte	0.0708	0.1763	–	0.1537
		Sorghum	0.4711	0.5805	0.4325	–
Coding	Less repetitive	W22	–	0.0000	0.0111	0.1149
		Ac1503	0.0970	–	0.0121	0.1059
		Teosinte	0.0476	0.1520	–	0.1136
		Sorghum	0.2628	0.1864	0.3346	–
	Repetitive	W22	–	0.0135	0.0177	0.1148
		Ac1503	0.2092	–	0.0209	0.1162
		Teosinte	0.2566	0.1596	–	0.1332
		Sorghum	0.5783	0.4470	0.4784	–
	All	W22	–	0.0097	0.0158	0.1147
		Ac1503	0.1802	–	0.0187	0.1133
		Teosinte	0.2039	0.1581	–	0.1280
		Sorghum	0.4898	0.3783	0.4430	–
3′	Intron	W22	–	0.0275	0.0297	0.1802
		Ac1503	0.1275	–	0.0000	0.1747
		Teosinte	0.1896	0.0491	–	0.1796
		Sorghum	0.3264	0.3888	0.4801	–
	Rest	W22	–	0.0527	0.0443	0.1887
		Ac1503	0.0804	–	0.0157	0.1934
		Teosinte	0.0584	0.0546	–	0.1685
		Sorghum	0.1493	0.1726	0.1654	–
	All	W22	–	0.0322	0.0241	0.1928
		Ac1503	0.0955	–	0.0161	0.2088
		Teosinte	0.0959	0.0531	–	0.1922
		Sorghum	0.1956	0.2352	0.2501	–

Values above the diagonal are numbers of nucleotide substitutions per site or numbers of non-synonymous substitutions per non-synonymous site for non-coding and coding sequences, respectively. Values below the diagonal are numbers of insertion/deletion events per site, γ

rooted phylogenetic trees following the least squares method (Felsenstein 1990). From these results (not shown) it appears that different trees are obtained when different regions of the *HRGP* gene are selected for the analysis. The maize W22 and Ac1503 sequences do not always cluster first (as expected from an evolutionary model based on single base substitutions or reduced rates of insertion/deletion). On the contrary, sometimes W22 clusters first with teosinte, and sometimes Ac1503 clusters first with teosinte. This conclusion can also be reached using distances estimates based on both substitutions per site and γ estimates.

Discussion

The sequences of genomic clones corresponding to an *HRGP* gene from different varieties of maize, teosinte and sorghum have been obtained and compared. Advantage has been taken of the simplicity of the system, both in terms of copy number and in terms of gene structure. Indeed, results from restriction fragment length polymorphism mapping, Southern blot (Stiefel et al. 1990) and cDNA and genomic cloning failed to detect more than one homologous gene in any of these species.

The 5′ region of the HRGP gene can be divided into an upstream zone, where large rearrangements are observed in the *Zea* sequences, and a 500 bp region having a significant degree of homology among *Zea* and sorghum genes. Recent results indicate that the latter region is able to direct the expression of a *GUS* reporter gene after bombardment of maize calli and shoot tissue (D. Tagu, this laboratory, personal communication) and it thus probably contains the *cis* elements controlling expression of the gene. The coding region consists of a main repetitive sequence with non-repeated segments at both ends, with the 5′ part coding for a signal peptide. The 3′ untranslated region containing an intron is also well conserved among *Zea* and sorghum genes. Outside these zones, the sequence divergence is high. In the 5′ region upstream of the observed rearrangements and in the 3′ region downstream of the polyadenylation site, no significant sequence homology can be found between *Zea* species and sorghum. This observation allows us to define, in a relatively precise way, the limits of the gene (as shown in Fig. 1) probably including its regulatory sequences. The *HRGP* gene in these species appears as a unit of conserved sequences positioned between a region showing significant rearrangements and non-conserved sequences.

The types of sequence variations observed between maize and teosinte *HRGP* genes have to be distinguished from those observed between *Zea* and sorghum. For instance, in the 5′ region of the *Zea* genes, only small insertions/deletions are observed in zones rich in short duplications, an effect already observed on comparison of sequences from different maize varieties (Schwarz-Sommer et al. 1985; Prat et al. 1987). However, comparison of *Zea* and sorghum HRGP genes reveals, besides a larger number of single nucleotide changes, long insertions/deletions, probably reflecting a greater evolutionary distance between the genera. In contrast, at the 3′ end, the number of large insertions/deletions is very reduced even between *Zea* and sorghum, and mostly short duplications are present (Fig. 3). In the coding region, the changes in the sequence arise mainly from large duplications of the repetitive elements that constitute its main part. In maize and teosinte these duplications retain the basic repetitive structure almost perfectly. Between *Zea* and sorghum, single amino acid replacements and small duplications locally vary the overall structure.

The variability observed in different parts of the *HRGP* gene probably reflects distinct kinds of selective pressure acting on them and allows us to distinguish between several mechanisms producing sequence variability. At least three types of sequence variation have to be considered. First, single nucleotide changes that occur at diverse rates in different parts of the gene (Table 1). Second, the appearance of small duplications that occur mostly at both ends of the gene. Some of them may be produced by the action of transposons, which leave behind short perfect or imperfect direct repeats after excision (Schwarz-Sommer et al. 1985). Third, large insertions/deletions that occur mostly in the 5′ region (when the sorghum sequence is compared with the others) and in the coding region, and which are probably generated by different mechanisms, as discussed below. Some of the 5′ deletions may have been generated by homologous recombination between short direct repeats flanking them, as has been described elsewhere (Efstratiadis et al. 1980; Dixon and Honh 1984). Nevertheless, several invariable boxes may correspond to *cis* elements important for the control of gene expression. For instance, the sequence GGGAAGCCTCC found at −249 bp in W22 maize and conserved in the four genes analyzed (Fig. 2), is almost identical to a motif found in the ethylene regulatory region of a bean chitinase (Broglie et al. 1989).

In the coding region, the main mechanism of sequence variability appears to be the generation of different numbers of repetitive motifs. These vary between teosinte and maize W22 from 17 to 12, and between two maize varieties (Ac1503 and W22) from 15 to 12. Taking into account the short evolutionary time that probably separates these two varieties (or maize from teosinte), the mechanism generating these variations appears to be very active in the genus *Zea*, as is reflected by the γ estimates observed in the coding region for the maize-teosinte pair (see Table 1). We propose that these variations could have been generated by a non-reciprocal interchange of DNA repetitive motifs between different alleles, as has been described for other repetitive sequences (Dover 1986). This process can result in sequence homogenization, as has been observed in the sequence of the *HRGP* gene repetitive motifs. This is illustrated by the codon used in a given position of the repetitive units, which is constant depending on the species. For instance, proline number 6 in the hexapeptide is encoded mostly by CCT/A in *Zea* species (79% of cases on average), and by CCG in sorghum (92% of cases). Among the different proposed mechanisms for non-reciprocal interchange of DNA in repetitive sequences, unequal crossing-over could have played a role in the evolution of the *HRGP* coding sequence because it allows the variation in the number of repeats observed in the coding region of *HRGP* genes (Li et al. 1985a).

The observations discussed above indicate that mechanisms that generate short sequence duplications at the 5′ and 3′ ends and large repetitions in the coding region are very active in the production of genomic variability in *Zea HRGP* genes. Some sequences may favour the action of a particular mechanism. For instance, the appearance of small duplications may raise the probability of increasing their number due to polymerase slippage (Schmid and Shen 1985). Furthermore, the boundaries between the repeated nucleotide sequences in the coding region consist of stretches rich in GC which seem to be a target for recombination enzymes (Ruiz-Carrillo and Renaud 1987). However, it has to be taken into account that a distinct selective pressure may exist in different regions of the gene. This phenomenon is illustrated by: (a) the conservation of small boxes at the 5′ end, regardless of the distance between them; (b) the conservation of an overall structure of the 3′ end, where large insertions/deletions are not tolerated except for the intron, suggesting that conservation of RNA secondary structure is important for gene function; and (c) the preservation of a well-defined repetitive structure in the protein region by maintenance of a certain number of repeated units and a defined spacing between them, independent of its length. Altogether, the distinct nature of the sequences present in particular regions of the *HRGP* gene, and the different selective pressures acting on them, could have determined the pattern of sequence evolution observed in these regions.

Acknowledgements. The authors are indebted to Dr. A. Brandt (Carlsberg Research Centre, Copenhagen, Denmark) for his help in genomic library preparation, to Mr. Patricio Gómez for sequencing assistance, and to Dr. A. Alvarez (Aula Dei, Zaragoza, Spain) for providing teosinte and sorghum plant material. The work has been financed by Plan Nacional (grant BI088/242) and by EECC (grant BAP 347). R.R. and M.J. are recipients of fellowships from PFPI (MEC) and from AFPI (CICYT), respectively.

References

Broglie KE, Biddle P, Cressman R, Broglie R (1989) Functional analysis of DNA sequences responsible for ethylene regulation of a bean chitinase gene in transgenic tobacco. Plant Cell 1:599–607

Burr B, Burr FA (1981) Controlling-element events at the shrunken locus in maize. Genetics 98:143–156

Chen JA, Varner JE (1985) An extracellular matrix protein in plants: characterization of a genomic clone for carrot extensin. EMBO J 4:2145–2151

Dixon LK, Hohn T (1984) Initiation of translation of the cauliflower mosaic virus genome from a polycistronic mRNA: evidence from deletion mutagenesis. EMBO J 3:2731–2736

Dover GA (1986) Molecular drive in multigene families: how biological novelties arise spread and are assimilated. Trends Genet 2:159–165

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ (1980) The structure and evolution of the human β-globin gene family. Cell 21:653–658

Fedoroff NV (1989) About maize transposable elements and development. Cell 56:181–191

Felsenstein J (1990) PHYLIP Manual. University Herbarium of the University of California, Berkeley, Calif

Keller B, Lamb CJ (1989) Specific expression of a novel cell wall hydroxyproline-rich glycoprotein gene in lateral root initiation. Genes Dev 3:1639–1646

Kieliszewski MJ, Leykam JF, Lamport DTA (1990) Structure of the threonine-rich extensin from Zea mays. Plant Physiol 92:316–326

Li W-H, Luo C-C, Wu C-I (1985a) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular Evolutionary Genetics. Plenum Press, New York pp 65–76

Li, W-H, Wu C-I, Luo C-C (1985b) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Loenen WAM, Blattner FR (1983) Lambda Charon vectors (Ch 32, 33, 34 and 35) adapted for DNA cloning in recombination-deficient hosts. Gene 26:171–179

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY

Maxam AM, Gilbert W (1980) Sequencing end-labelled DNA with base-specific chemical cleavages. Methods Enzymol 65:499–560

McClintock B (1951) Chromosome organization and genic expression. Cold Spring Harbor Symp Quant Biol 16:13–47

Prat S, Perez-Grau L, Puigdomènech P (1987) Multiple variability in the sequence of a family of maize endosperm proteins. Gene 52:41–49

Queen C, Korn LJ (1984) A comprehensive sequence analysis program for the IBM personal computer. Nucleic Acids Res 12:581–599

Raz R, Crétin C, Puigdomènech P, Martínez-Izquierdo JA (1991) The sequence of a hydroxyproline-rich glycoprotein gene from Sorghum vulgare. Plant Mol Biol 16:365–367

Ruiz-Carrillo A, Renaud J (1987) Endonuclease G: a $(dG)_n \cdot (dC)_n$ specific DNase from higher eukaryotes. EMBO J 6:401–407

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-termination inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schmid CW, Shen C-KJ (1985) The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates. In: MacIntyre RJ (ed) Molecular Evolutionary Genetics. Plenum Press, New York pp 323–358

Schwarz-Sommer Z, Gierl A, Cuypers H, Peterson PA, Saedler H (1985) Plant transposable elements generate the DNA sequence diversity needed in evolution. EMBO J 4:591–597

Showalter AM, Bell JN, Cramer CL, Bailey JA, Varner JE, Lamb CJ (1985) Accumulation of hydroxyproline-rich glycoprotein mRNAs in response to fungal elicitor and infection. Proc Natl Acad Sci USA 82:6551–6555

Stiefel V, Pérez-Grau L, Albericio F, Giralt E, Ruiz-Avila L, Ludevid D, Puigdomènech P (1988) Molecular cloning of cDNAs encoding a putative cell wall protein from Zea mays and immunological identification of related polypeptides. Plant Mol Biol 11:483–493

Stiefel V, Ruiz-Avila L, Raz R, Vallés MP, Gómez J, Pagés M, Martínez-Izquierdo JA, Ludevid D, Langdale JA, Nelson T, Puigdomènech P (1990) Expression of a maize cell wall hydroxyproline-rich glycoprotein gene in early leaf and root vascular differentiation. Plant Cell 2:785–793

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269–285

Communicated by H. Saedler