Repeats with variations: accelerated evolution of the Pin2 family of proteinase inhibitors

Endre Barta, Alessandro Pintar and Sándor Pongor

The Pin2 genes encode potato type II proteinase inhibitors that act against pathogenic attack. The first examples were found only in the Solanaceae family, but, using new EST and genomic data, we have found 11 homologous genes dispersed through almost the whole range of mono- and di-cotyledonous plants. In contrast to the repetitive precursor sequences of the Solanaceae Pin2 genes, the new homologs have only a single repeat unit. The gene family appears to have evolved from a single-domain ancestral gene through a series of gene-duplication and domain-duplication steps. A number of unequal cross-over and gene conversion events could explain the current gene and domain pattern of the Solanaceae Pin2 subfamily.

Published online: 23 September 2002

The Pin2 family of proteinase inhibitors is present in seeds, leaves and other organs of the Solanaceae. Perhaps the best known representatives are the wound-induced proteinase inhibitors [1,2], which contain up to eight sequence-repeats (the ‘IP repeats’) coded by the second exon of the gene (e.g. [3]). The 3D structure of the mature inhibitor from potato is known [4], and recently it was shown that some engineered Pin2 precursors are able to form a circularly permuted structure [5–7] that was thought to correspond to the ancestral, single-repeat protein of this family. Subsequently, a naturally occurring Pin2 protein, PSI-1.2, with this ‘ancestral’ circularly permuted structure was isolated [3]. Circular permutation of sequences had been reported in other cases (for reviews see [8,9]). But until the discovery of PSI-1.2, circular
rearrangements were observed only between species, such as favin from Vicia faba and the lectin concanavalin A from Canavalia ensiformis [10] or the plant aspartyl proteases and human lung surfactant proteins [11]. PSI 1.2 is the first example where circularly permuted members of a protein family are expressed within the same organism, moreover, within the same organ. As both proteinase inhibitors and lectins are proteins that have roles in the defense mechanisms of plants, it is tempting to speculate that the underlying sequence-rearrangements are part of a general scenario by which plants produce functional diversity against pathogenic attack.

Based on the sequences of cDNA and genomic clones, 18 members of the Pin2 family have been annotated so far in the main protein and DNA databases. While reviewing the precursor architectures of the Pin2 family (Fig. 1) and scanning the new genomic and EST databases with sensitive BLAST-based algorithms [12,13], we found 11 hitherto unannotated members of the Pin2 family in eight different monocotyledonous (Oryza indica ssp., Oryza japonica ssp., Zea mays and Sorghum halepense) and dicotyledonous (Arabidopsis thaliana, Medicago truncatula, Mesembryanthemum crystallinum, Solanum tuberosum) plants (Fig. 2). It thus appears that the Pin2 family is more widespread than previously thought. And, in contrast to the repetitive precursor sequences found in the Solanaceae, nine of the novel genes are composed of a single repeat unit, as was previously predicted for the ancestral gene [5–7].

The architecture of the Pin2 genes (Fig. 1, inset) is conserved: the first exon encoding the N-terminus of the signal peptide, and the second, major exon encoding the C-terminus of the signal peptide and a variable number of IP repeats, are always separated by a type I intron of 100–200 bp. However, the sequence of the IP repeats is quite variable, only the cysteines constituting the four disulfide bridges and a single proline residue are conserved throughout the 77 known repeat sequences (a multiple alignment is deposited as Supplementary Material at http://www.abc.hu/pin2).

Duplication of Pin2 genes seems to have occurred several times, especially within the Solanaceae. Outside the family, the ice plant is the only known example of gene duplication, the other genes are apparently present in a single copy in the haploid genome.

The comparison of repeat sequences (Fig. 3, Table 1) shows two types of cluster. First, there is a clustering according to repeat number; for example, repeat 2 of a gene can be similar to repeat 2 of another gene. This suggests that the multiplication of repeats (repeat expansion) within exon 2 preceded gene duplication. Second, repeats within the same gene are very similar to each other, which could be caused by fast repeat expansion from a single IP repeat or by a uniformization of the repeats by such mechanisms as gene conversion [14] and unequal cross-over (UECO) [15]. The latter possibility is more likely because the duplication is known to be a rare event.

The mechanism of repeat expansion has not been studied in detail. However, it is conspicuous that the Pin2 precursor architectures – and especially the pattern of repeat identities shown in Fig. 1 – can be reconstructed by a series of putative UECO within exon 2 (Fig. 4). UECO is well known for genes and has been invoked to explain recombination of introns [16]. The peculiarity of this case of UECO is that the repeats that proliferate within the same exon constitute individual folding units, and that virtually all the precursor architectures can be explained by the mechanism. We note that repeat expansion seems only to have occurred within the Solanaceae, as the genes of other families all code single-repeat proteins. This leads us to suppose that the accelerated evolution of this gene might have been triggered by...

### Table 1. The distribution of the proteinase contact residues P1 and P1' within the clusters*

<table>
<thead>
<tr>
<th>Cluster</th>
<th>P1</th>
<th>P1'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Basic</td>
<td>Acidic</td>
</tr>
<tr>
<td>R3</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>IP1</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>R2</td>
<td>23%</td>
<td>46%</td>
</tr>
<tr>
<td>R1</td>
<td>69%</td>
<td>31%</td>
</tr>
<tr>
<td>IP3</td>
<td>78%</td>
<td>22%</td>
</tr>
<tr>
<td>N. alata, N. glutinosa</td>
<td>75% R</td>
<td>25% L</td>
</tr>
</tbody>
</table>

*The amino acid types were as follows: basic, R, K; acidic, E, D; hydrophobic, L, I, F, V, M; polar, N, S, T. The amino acids not listed did not occur in the two positions, the conserved amino acids are explicitly indicated.
Fig. 3. Clustering of the DNA sequences coding for IP repeats. The sequences were clustered using the CLUSTAL program [21]. IP indicates the total number of repeats within the gene, R indicates the serial order of the repeat starting from the N-terminus. Brown, tobacco (N. tabacum; IP3; N. alata: one IP4, one IP6; N. glutinosa: one IP6, one IP8); blue, potato (one IP1, six IP2, one IP2’); red, tomato (two IP2 three IP3); green, paprika (two IP3, two IP3); black, non-solanaceous plants (three IP1 in ice-plant, one IP1 in the others). Table 1 shows the distribution of the proteinase contact residues P1 and P1’ within the clusters.
Fig. 4. Some of the potential unequal cross-over (UECO) events that explain the emergence of sequence identity patterns of the potato II family precursors (Fig. 1 and Supplementary Material at http://www.abc.hu/pin2). The two partners are colored green and brown. Two types of UECO event involving either adjacent (Ri+1 × Ri+1) or nonadjacent (Ri + Ri) repeats are shown by dashed lines. Gray lines indicate sequence identity (>98%). The IP1 type structure that corresponds to the homologs of the putative ancestral gene (Fig. 2), can arise as a result of various UECO events. For example, as the PSI-1.2 protein is similar to the R3 repeats [3], it could have occurred in ancestral sequences (Fig. 1, bottom) could have emerged from a putative RI protein as a result of an UECO event (not shown).

an initial repeat multiplication that occurred in ancestral Solanaceae.

The impressive variability of the Pin2 genes in Solanaceae can be understood by comparing the proteinase contact residues that are the major determinant of the specificity of proteinase inhibitors (see Supplementary Material at http://www.abc.hu/pin2 for details). First, the ancestral single-repeat proteins have a different specificity to the multirepeat Solanaceae proteins. Second, the repeat units within the multirepeat proteins have different specificities, so a mature protein represents a mix of inhibitors. It thus appears that, in response to pathogenic attack, plants resort to protease inhibitor cocktails similar to those used to fight retroviral infections in humans. The Pin2 genes are part of the plant’s innate immune response, which is in general characterized by broad specificity. UECOs and gene conversions seem to be plausible mechanisms both for generating and for fine-tuning this broad specificity against various pathogens.

References
3 Antcheva, N. et al. (2001) Proteins of circularly permuted sequence present within the same organism: the major serine proteinase inhibitor from Capsicum annuum seeds. Protein Sci. 10, 2280–2290
14 Hurles, M.E. (2001) Gene conversion homogenizes the CMT1A paralogous repeats. BMC Genomics 2, 11

Endre Barta
Agricultural Biotechnology Center, 2100 Gödöllő, Hungary.

Alessandro Pintar
Sándor Pongor*
Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, 34012 Trieste, Italy.
*e-mail: pongor@icgeb.trieste.it

TiG Early Edition
Did you know that Trends in Genetics articles are now published online ahead of print?
To access the Early Edition, log on to http://reviews.bmn.com/journals, select Trends in Genetics then click on the ‘Early Edition’ button.