3 Yang, M. and Sack, F.D. (1995) The *too many mouths* and *four lips* mutations affect stomatal production in *Arabidopsis*. *Plant Cell* 7, 2227–2239
4 Geisler, M. *et al.* (1998) Divergent regulation of stomatal initiation and patterning in organ and suborgan regions of the *Arabidopsis* mutants *too many mouths* and *four lips*. *Planta* 205, 522–530
5 Serna, L. *et al.* (2002) Specification of stomatal fate in *Arabidopsis*: evidences for cellular interactions. *New Phytol.* 153, 399–404
6 Kobe, B. and Deisenhofer, J. (1994) The leucine-rich repeat: a versatile binding motif. *Trends Biochem. Sci.* 19, 415–421
7 Kayes, J.M. and Clark, S.E. (1998) *CLAVATA2*, a regulator of meristem and organ development in *Arabidopsis*. *Development* 125, 3843–3851
8 Jeong, S. *et al.* (1999) The *Arabidopsis CLAVATA2* gene encodes a receptor-like protein required for the stability of *CLAVATA1* receptor-like kinase. *Plant Cell* 11, 1925–1934
9 Berger, D. and Altmann, T. (2000) A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. *Genes Dev.* 14, 1119–1131
10 von Groll, U. and Altmann, T. (2001) Stomatal cell biology. *Curr. Opin. Plant Biol.* 4, 555–560
11 Clark, S.E. (2001) Cell signalling at the shoot meristem. *Nat. Rev. Mol. Cell Biol.* 2, 276–284
12 The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815

Laura Serna*
Carmen Fenoll
Universidad de Castilla-La Mancha, Facultad de Ciencias del Medio Ambiente. Avda. Carlos III, s/n, 45071 Toledo, Spain.
*e-mail: laura.serna@uclm.es

Genome Analysis

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

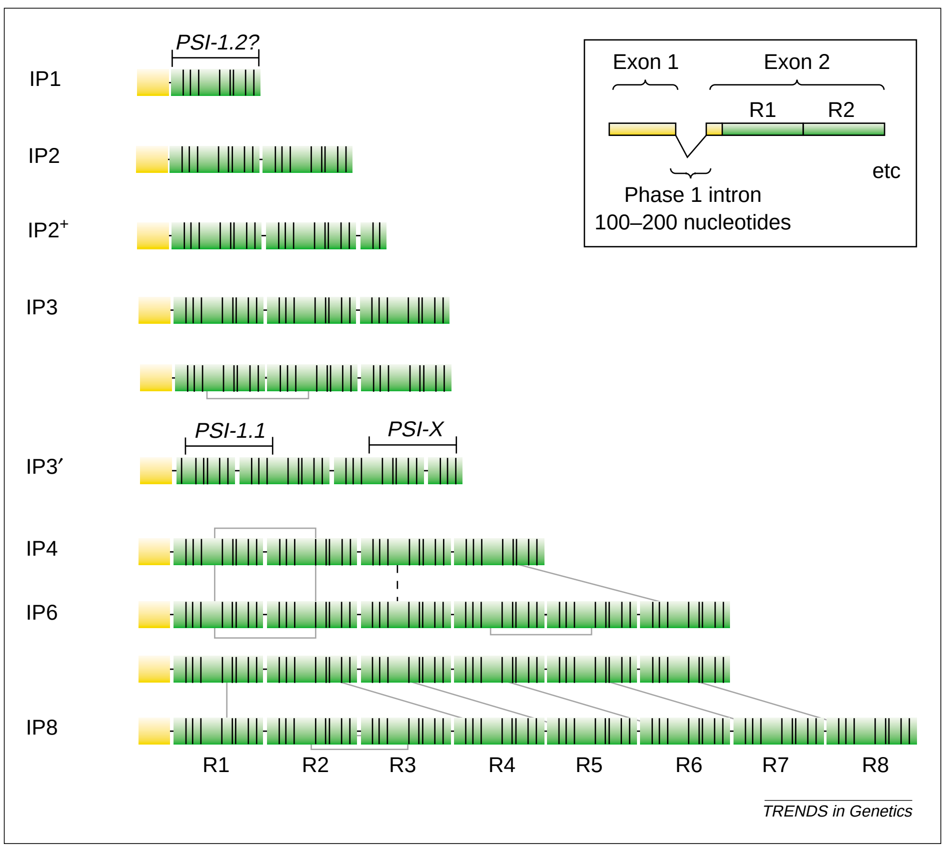

Fig. 1. The domain structure of the potato type II proteinase inhibitor (Pin2) family of precursors. The inset shows the consensus protein structure. IP1...IP8 designate the total number of IP repeats (green boxes) within each precursor. The IP1 molecules were identified while searching genomic databases for the purposes of this review (sequences shown in Fig. 2). Yellow box, signal peptide; black vertical lines, Cys residues; gray lines, sequence identity (>98%). The presence of adjacent, identical repeats is a recurrent pattern. PSI-1.1 [17] PSI-X (N. Antcheva and S. Pongor, unpublished) and PSI-1.2 [3] are paprika seed inhibitors, the gene corresponding to the major seed inhibitor PSI-1.2 is unknown. See Fig. 3 and Supplementary Material for the gene accession numbers.

  PII: S0168-9525(02)02771-3

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 proteinases 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

```
Sorghum halepense, sorghum IP1-AI724716                VACPQYCLE-VDYVTCP-SSGSEKLPA-RC-NCCLAP-KGCTLHLSDGTQQTC
Zea mays, maize IP1-AI947362                           VACPQFCLD-VDYVTCP-SSGSEKLPE-RC-NCCMTP-KGCTLHLSDGTQQTC
Arabidopsis thaliana, thale cress IP1-AI999664         IACPLYCLQ-VEYMTCP-SSGADKLPP-RC-NCCLAP-KNCTLHLSDSTTIHC
Mesembryanthemum crystallinum, ice plant IP1-BE033392  TACHEICYVDAAYMTCP-STGYQHLAP-KC-NCCEGG-PGCSLYRANGEQI-C
Mesembryanthemum crystallinum, ice plant IP1-BE033692  TICLQICYTDADYMTCP-STGYKKLNP-AC-NCCIGG-PGCSLYRANGEQI-C
Mesembryanthemum crystallinum, ice plant IP1-BE033653  QICPQICYIEAAYMLCP-STGDQQLTP-EC-NCCEAG-PGCSLYEADGKKI-C
Oryza sativa, rice IP1-AU163886                        KFCPQFCYDGLEYMTCP-STG-QHLKP-AC-NCCIAGEKGCVLYLNNGQVINC
Medicago truncatula, barrel medic IP1-BE942349         KICPQFCYDSVAYMTCP-SSGDQHLTP-KC-NCCLAS-TGCILYEADGTPI-C
Solanum tuberosum, potato IP1-BI434643                 QICPQFCEPNVDYMTCS-SSGSTILRP-TCINCCQARGRGCQLFRRDGSAI-C

Lycopersicon esculentum, tomato IP3-BI421162-R3        KACTMECDPRVAYMICP-SSGLAKLSQ-VCVNCCTAG-DGCKLYGYDGSLI-C
Solanum tuberosum, potato IP3-BI436259-R3              KACTLECDSRVSYMTCP-SSGMAKLNQ-VCVNCCTAR-EGCKLYGHDGSLI-C
Lycopersicon esculentum, tomato IP3-BI421162-R1        KPCTMECG-HLGFGICPRSQG-SPQNP-ICTNCCSGF-KGCNYYSVNGTFI-C
Solanum tuberosum, potato IP3-BI436259-R1              KACTKECG-NLGYGICPRSEG-STENP-ICTNCCAGY-KGCNYYSANGIFI-C
Lycopersicon esculentum, tomato IP3-BI421162-R2        KPCPLNCDPHIAYSKCPRSGGKTFIYPTGCTTCCTGY-KGCYYFGKDGKFV-C
Solanum tuberosum, potato IP3-BI436259-R2              KACPRNCDPHIAYSKCLRSRGKTLIYPTGCTTCCTGY-KDCYYFGKDGKFV-C
                                                         *   *     :  *  * *        * .**     .*  .  :.    *
```

TRENDS in Genetics

Fig. 2. Sequences of potato II type proteinase inhibitor (Pin2) homologs found in the EST databases of ice plant (EMBL accession number: BE033392), rice (EMBL accession number: AU163886), barrel medic (*Medicago truncatula*) (EMBL accession number: BE942349), sorghum (EMBL accession number: AI724716), maize (EMBL accession number: AI947362) and *Arabidopsis thaliana* (EMBL accession number: AI999664). The *A. thaliana* EST-sequence was also found in the genomic sequence of chromosome 1 of *A. thaliana* [18], the rice EST sequence was found in both recently published rice genomes [19,20]; both genes are apparently present as a single copy within the genomes. IP indicates the total number of repeats within the gene, R indicates the serial order of the repeat starting from the N-terminus. The multiple alignment was produced by the CLUSTAL program [21] (see Supplementary Material at http://www.abc.hu/pin2). The proteinase-contact residues P_1 and P_1' are highlighted in green, 100% conserved amino acid positions (yellow) are marked with asterisks, double or single dots mark conservative and semi-conservative amino acid substitutions, respectively [21].

Table 1. The distribution of the proteinase contact residues P1 and P1′ within the clusters[a]

Cluster		P1				P1′		
No. (Fig. 3)	Type	Basic	Acidic	Hydrophobic	Polar	Acidic	Hydrophobic	Polar
1	R3	40%		40%	20%	100% E		
2	IP1	10%	10%	10%	70%		90%	10%
3	R2	23%		46%	15%		15%	85%
4	R1	69%		31%		100% E		
5	IP3	78%		22%		33%	11%	55%
6	*N. alata, N. glutinosa*	75% R		25% L				100% N
All		53%	1%	34%	3%	28%	16%	56%

[a]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.

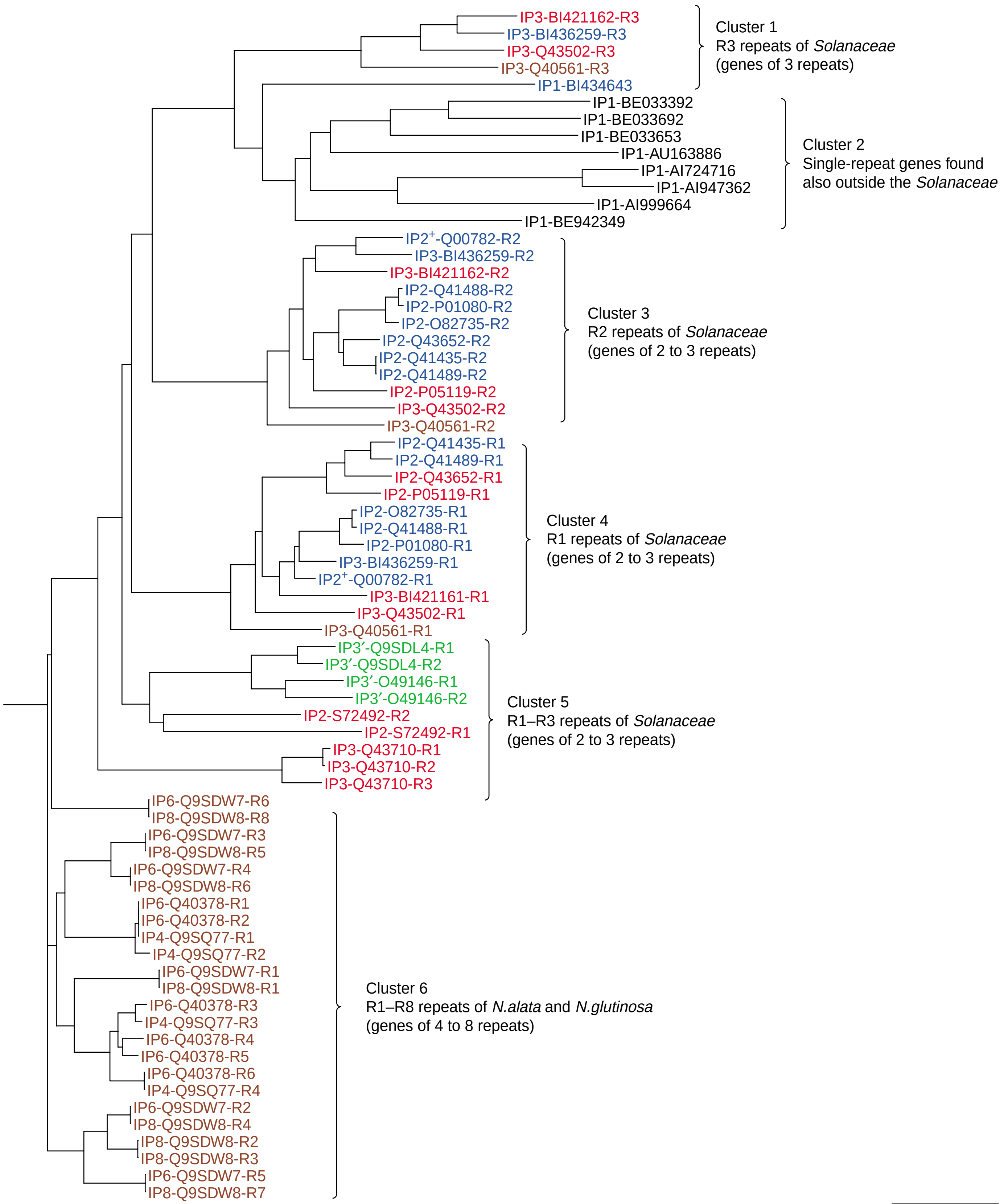

Fig. 3. Clustering of the DNA sequences coding for IP repeats. The sequences were clustered using the CLUSTAL program [21]. IP indicates the total number of repeats within the gene, R indicates the serial order of the repeat starting from the N-terminus. Brown, tobacco (N. tabacum; IP3; N. alata: one 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.

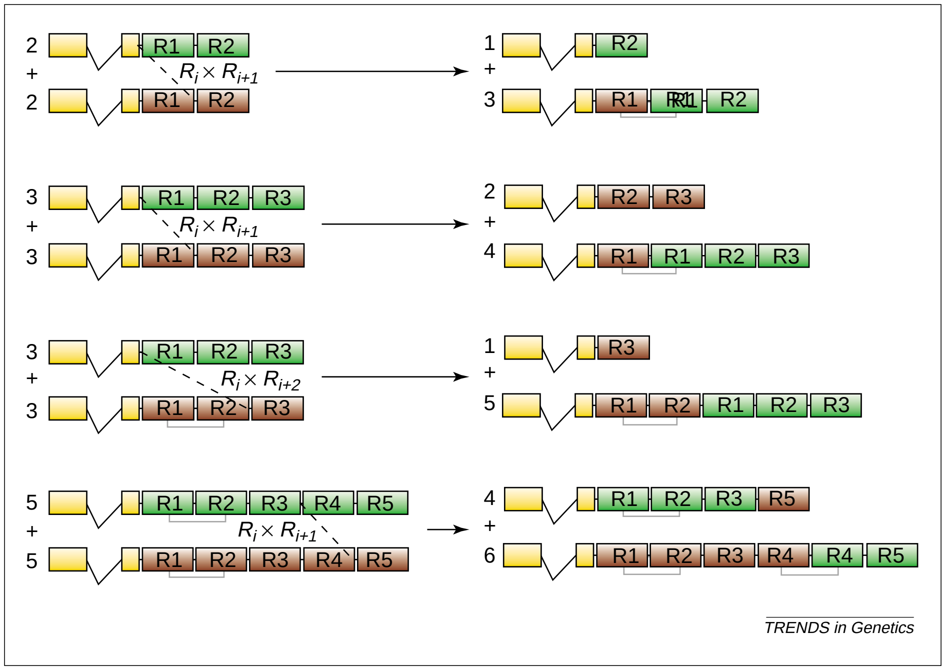

Fig. 4. Some of the potential unequal cross-over (UECO) events that explain the emergence of sequence identity patterns of the potato II family precursors (Fig. 1 and Supplementary Material at http://www.abc.hu/pin2). The two partners are colored green and brown. Two types of UECO event involving either adjacent ($R_i \times R_{i+1}$) or nonadjacent ($R_i \times R_{i+2}$) 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 emerged from the $R_i \times R_{i+2}$ type recombination of two IP3 genes. IP5 type products have not been observed, but their putative $R_i \times R_{i+1}$ type recombination products are found in *N. alata* (bottom). In similar manner, the IP6 and IP8 type products found in *N. glutinosa* (*Q9SDW7* and *Q9SDW8*, respectively; Fig.1, bottom) could have emerged from a putative IP7 protein as a result of an UECO type 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 multi-repeat 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

1 Moura, D.S. *et al.* (2001) Characterization and localization of a wound-inducible type I serine-carboxypeptidase from leaves of tomato plants (*Lycopersicon esculentum* Mill.) *Planta* 212, 222–230
2 Moura, D.S. and Ryan, C.A. (2001) Wound-inducible proteinase inhibitors in pepper. Differential regulation upon wounding, systemin, and methyl jasmonate. *Plant Physiol.* 126, 289–298
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
4 Greenblatt, H.M. *et al.* (1989) Structure of the complex of *Streptomyces griseus* proteinase B and polypeptide chymotrypsin inhibitor-1 from Russet Burbank potato tubers at 2.1 Angstrom resolution. *J. Mol. Biol.* 205, 201–228
5 Scanlon, M.J. *et al.* (1999) Structure of a putative ancestral protein encoded by a single sequence repeat from a multidomain proteinase inhibitor gene from *Nicotiana alata*. *Structure Fold Des.* 7, 793–802
6 Schirra, H.J. *et al.* (2001) The solution structure of C1-T1, a two-domain proteinase inhibitor derived from a circular precursor protein from *Nicotiana alata*. *J. Mol. Biol.* 306, 69–79
7 Lee, M.C. *et al.* (1999) A novel two-chain proteinase inhibitor generated by circularization of a multidomain precursor protein. *Nat. Struct. Biol.* 6, 526–530
8 Heringa, J. and Taylor, W.R. (1997) Three-dimensional domain duplication, swapping and stealing. *Curr. Opin. Struct. Biol.* 7, 416–421
9 Lindqvist, Y. and Schneider, G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.* 7, 422–427
10 Edelman, G.M. *et al.* (1972) The covalent and three-dimensional structure of concanavalin A. *Proc. Natl. Acad. Sci. U. S. A.* 69, 2580–2584
11 Ponting, C.P. and Russell, R.B. (1995) Swaposins: circular permutations within genes encoding saposin homologues. *Trends Biochem. Sci.* 20, 179–180
12 Murvai, J. *et al.* (2000) A simple probabilistic scoring method for protein domain identification. *Bioinformatics* 16, 1155–1156
13 Murvai, J. *et al.* (2001) Prediction of protein functional domains from sequences using artificial neural networks. *Genome Res.* 11, 1410–1417
14 Hurles, M.E. (2001) Gene conversion homogenizes the CMT1A paralogous repeats. *BMC Genomics* 2, 11
15 Smith, G.P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* 191, 528–535
16 Patthy, L. (1999) *Protein Evolution*, Blackwell Science
17 Antcheva, N. *et al.* (1996) Primary structure and specificity of a serine proteinase inhibitor from paprika (*Capsicum annuum*) seeds. *Biochim. Biophys. Acta* 1298, 95–101
18 Theologis, A. *et al.* (2000) Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* 408, 816–820
19 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* 296, 79–92
20 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* 296, 92–100
21 Thompson, J.D. et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680

Endre Barta
Agricultural Biotechnology Center, 2100 Gödöll, 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