Plant Molecular Biology **23**: 801–812, 1993.

Characterization of a genomic sequence coding for potato multicystatin, an eight-domain cysteine proteinase inhibitor

Clive Waldron, Lynette M. Wegrich, P. Ann Owens Merlo and Terence A. Walsh
DowElanco Biotechnology Research, 9410 Zionsville Rd., Indianapolis, IN 46268-1053, USA

Received 13 May 1993; accepted in revised form 20 August 1993

Key words: Solanum tuberosum, tuber, proteinase inhibitor, cystatin, gene structure

Abstract

A gene coding for potato multicystatin (PMC), the crystalline inhibitor of cysteine proteases which is found in tubers, was isolated and characterized. The deduced polypeptide product of this genomic sequence is 757 amino acids long and has a molecular mass of 86,778 Da. It consists exclusively of eight closely related domains, with 53–89% identity of residues. Each repeated unit is homologous to the cystatin superfamily of cysteine protease inhibitors. To date, no other member of this family has been found to contain so many inhibitor domains in one polypeptide. Eight introns are proposed in the 3.5 kb of genomic DNA coding for PMC, one in each cystatin unit. There is a family of 4 to 6 such large genes in potato, while in pea and maize the homologues are much smaller, and probably code for single-domain cystatins. PMC transcripts are abundant in tubers, but scarce in undamaged leaves or stems of field-grown potatoes. The tuber messages are derived from at least four genes (including the cloned example). The pattern of gene expression, as well as the properties of the protein, suggest that PMC has a role in the plant's defense system.

Introduction

Potato multicystatin (PMC) is a proteinaceous inhibitor of cysteine proteases which is found as crystals in the subepidermal cells of potato tubers [26]. Partial amino acid sequencing of PMC revealed that it is a member of the cystatin superfamily of cysteine protease inhibitors [34]. All these inhibitors have a characteristic structural motif, QXVXG, which is essential for activity against papain and other cysteine proteinases [4]. Members of the cystatin superfamily from animal sources are classified into three groups [6]. Type I inhibitors, or stefins, are small intracellular proteins (about 11 kDa in size), without disulfide bonds or carbohydrate modifications. Type II inhibitors, or cystatins, are also small proteins (about 13 kDa), but they contain two disulfide bonds and are secreted. Type III inhibitors, or kininogens, are large (60–120 kDa) secreted proteins with multiple disulfide bonds; the amino-terminal portions of these polypeptides contain

The nucleotide sequence data reported will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under the accession number L16450.

three cystatin repeats, followed by other domains from which peptides active in inflammation can be cleaved. The small number of plant cystatins characterized to date are like animal stefins: they are small (12–18 kDa), and do not contain disulfide bonds [1, 12, 13,18]. The two rice cystatins also lack secretion signals, though a recently isolated cDNA clone appears to code for a maize cystatin with an amino-terminal extension [2]. PMC is unique among the plant cystatins with respect to its large size (a polypeptide of about 85 kDa), and its ability to form crystals at alkaline pH [26]. Recent studies indicated that the large size can be explained by the presence of multiple cystatin-like units [34]. However, this analysis did not exclude the possibility that PMC, like the animal kininogens, contains additional regions which are not related to cystatin. The genomic DNA sequence presented here demonstrates that this is not the case. The gene codes for a protein of eight stefin-like domains, without any additional elements.

Materials and methods

Construction of a tuber cDNA library

Peel from 4 cm potato tubers (*Solanum tuberosum* cv. Superior) was frozen in liquid nitrogen, and ground to a fine powder with a pestle and mortar. Disrupted tissue was thawed in 4 volumes of 50 mM Tris-HCl pH 8.0, 4% *p*-aminosalicylic acid, 1% tri-isopropylnapthalenesulfonic acid, 10 mM dithiothreitol, and 10 mM sodium metabisulfite, then extracted with an equal volume of phenol containing 0.1% 8-hydroxyquinoline. The aqueous phase was re-extracted with equal volumes of phenol/chloroform/isoamyl alcohol (25:24:1), and chloroform/octanol (24:1). RNA was first precipitated with ammonium acetate (2.5 M) at −20 °C overnight, then with 67% ethanol. Pol(A)$^+$ RNA was isolated using Hybond messenger affinity paper (Amersham, Arlington Heights, IL). cDNA was synthesized from poly(A)$^+$ RNA using the ZAP-cDNA synthesis kit (Stratagene, La Jolla, CA). Size-selected cDNA was ligated to UniZap XR vector arms, and packaged into phage particles with Gigapack Gold packaging extract (Stratagene).

Identification and characterization of PMC cDNA sequences

Amplified libraries were screened for PMC antigen production using the PicoBlue immunoscreening kit (Stratagene), and polyclonal rabbit antiserum raised against PMC (Berkeley Antibody Co., Berkeley, CA). Libraries were also screened for DNA homology to the insert in clone PC1, using gel-purified fragments labelled with digoxygenin-11-dUTP (Boehringer-Mannheim, Indianapolis, IN). Phage were purified by 3 cycles of plating at low density and rehybridization. The cDNAs were excised as phagemids, with an extra step of retransformation at low density to eliminate helper phage. Samples of phagemid DNA were prepared from XL1-Blue cells (Stratagene) by the method of Holmes and Quigley [14]. DNA sequences were determined using Sequenase (US Biochemical, Cleveland, OH). Erase-A-Base (Promega, Madison, WI) was used to generate overlapping deletions. Primers were synthesized by Oligos Etc. (Wilsonville, OR).

Identification and characterization of PMC genomic sequences

A library of potato genomic fragments (Stratagene) was screened for homology to PMC cDNA as described above. The probe was a gel-purified *Bsm* I fragment of the longest cDNA clone, pDAB1054, labelled with digoxygenin by random priming or PCR amplification. Phage were prepared on a large scale by extraction of broth cultures with polyethylene glycol (PEG 6000) and equilibrium density centrifugation in CsCl [5]. DNA was prepared from phage in buoyant density CsCl by overnight incubation at room temperature in 50% formamide, 10 mM EDTA, 0.1 M Tris-HCl pH 8.5, followed by ethanol precipitation. Fragments of this DNA

were subcloned into pBluescript II SK$^-$ (Stratagene) and sequenced, as described above.

Genomic DNA was prepared from etiolated shoots of potato (*Solanum tuberosum* cv. Superior) and pea (*Pisum sativum* cv. Little Marvels). Tissue frozen in liquid nitrogen was blended to a fine powder then thawed in 3 volumes of 7 M urea, 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM EDTA, 1% sarcosine. The homogenate was extracted with an equal volume of phenol/chloroform (1:1), and DNA was precipitated from the aqueous phase with an equal volume of isopropanol. Maize DNA was isolated from callus tissue (*Zea mays* cv. Black Mexican Sweet) by the procedure of Rogers [27]. DNAs cut with restriction enzymes were fractionated on agarose gels, and transferred to GeneScreen Plus membrane (DuPont, Wilmington, DE), using standard procedures [5]. To detect sequences homologous to PMC, membranes were probed with ^{32}P-labelled DNA containing the complete PMC-coding region, with introns. The membranes were washed with 2 × SSC, 0.1% SDS at 42 °C, then exposed to Kodak X-Omat film.

Characterization of PMC mRNA

The 5′ end of PMC mRNA was determined by primer extension from tuber poly(A)$^+$ RNA, prepared using the method described for construction of the cDNA library (above). Extension was from a 20-base primer complementary to the 5′ end of exon 2 (5′-TTTAAACTCCAAACTAGAATC), using AMV reverse transcriptase (US Biochemical).

For northern analysis, total RNA was prepared from tubers, leaves and stems of field-grown potatoes as described for construction of the cDNA library. The RNA was fractionated on formaldehyde-agarose gels, then transferred to GeneScreen membrane (DuPont). PMC mRNAs were detected by probing with ^{32}P-labelled insert from PC1, in the presence of 50% formamide and 10% dextran sulfate [5]. The most stringent wash was 50 mM NaCl, 20 mM sodium phosphate pH 6.5, 1 mM EDTA, 0.1% SDS at 42 °C.

Results

cDNAs coding for PMC

From a cDNA library of 1.5×10^5 phage, 12 clones producing PMC antigen were isolated. The insert from one of these plasmids (PC1) was used to isolate another 16 PMC clones by nucleic acid hybridization. None of the cDNAs was sufficiently large to code for an 85 kDa polypeptide. Three inserts were completely sequenced, and another six were partially sequenced. The 3′ untranslated sequences, which are gene specific, indicate that the PMC cDNAs were probably derived from four closely-related genes (Fig. 1). The longest clone (pDAB1054) contained an open reading frame which codes for almost four repeats of a polypeptide homologous to oryzacystatin (data not shown). As expected from its size, this insert was only part of a PMC message. At its 3′ end there was a non-translated region and a poly(A) tail, but the 5′ end of the reading frame was missing. Surprisingly, the coding region appears to include an intervening sequence which had not been processed out of the message.

Genomic clone coding for PMC

A library of genomic DNA fragments was probed with a fragment of the longest cDNA clone, pDAB1054, containing almost 4 cystatin-like repeats. From 4×10^5 phage, six clones containing hybridizing sequences were isolated. Only three of them could be grown in bulk for further characterization. Two contained part of a PMC gene, truncated at the 5′ end, but one clone (BZ2) contained all the homology in a central portion of its 13 kb insert. This hybridizing region was subcloned and sequenced (Fig. 2). An open reading frame of nine exons, coding for a polypeptide with eight cystatin-like domains, was identified. Part of the coding sequence, and the 3′ untranslated region, were identical to the cDNA clone AM9, except that one of the cystatin repeats was missing from the cDNA (Fig. 2). The deletion, between 16 bp regions of identical sequence in

```
          1                                                                    70
     PC1  TTGTTTATGC CACAAAATGA AATAAAACTA CTCTTATGCT GGTGTGAAGT AAAGCCAGTT GTTTGGTATA
pDAB1020  .......... .......... .......... .......... .......... .......... .......ATA
pDAB1023  .......... .......... .......... ........CT GGTGTGAAGT AAAGCCAGTT GTTTGGTATA
pDAB1054  ....TTATGC CACAAAATGA AATAAAACTA CTCTTATGCT GGTGTGAAGT AAAGCCAGTT GTTTGGTATA

pDAB1065  .......... .......... .......... .......... .......... .......... ..TTGGTATg
pDAB1069  .......... .......... .......... .......... .......... .......... ..........

     AM9  TTGgTgATGC tACAAAgTGA AATgAAACTA CTtTTATGtT GGTGTGAAaT AAAGCCAGTT GTTTGGTATg

pDAB1104  TTGgTgATGC tACAAAgTGA AATgAAACTA tTtTTATGtT GGTGTGAAaT AAAGCCAGTT GTTCGGTATg

          71                                                                   140
     PC1  AAGTTACTAT TATGCTTGTG TGAAATAAAG CCAGTTGTTT GGCATGCAAT TACTGTTATG TTTGTGAAAA
pDAB1020  AAGTTACTAT TATGCTTGTG TGAAATAAAG CCAGTTGTTT GGCATGCAAT TACTGTTATG TTTGTGAAAA
pDAB1023  AAGTTACTAT TATGCTTGTG TGAAATAAAG CCAGTTGTTT GGCATGCAAT TACTGTTATG TTTGTGAAAA
pDAB1054  AAGTTACTAT TATGCTTGTG TGAAATAAAG CCAGTTGTTT GGCATGCAAT TACTGTTATG TTTGTGAAAA

pDAB1065  AAGTTACTAT TATGCTTGTG TGAAATAAAG CCAGTTaTTT GGCATGCAAT TACTGTTATG TTTGTGAAAA
pDAB1069  .......... .......... .......... .......... .......... .......... ..........

     AM9  AAGTTAtTGT aATGtTTGTG AGAAATAAAG CCAaTTGTTT GGCATaaAAT TgCTaTTATG TTTGTGAgAA

pDAB1104  AAGTTAtTGT aATGtTTGTG AGAAATAAAG CCAaTTGTTT GGCATaaAAT TACTaTTATG TTTGTGAgAA

          141                                                                  210
     PC1  ATAAAGTCAG ATCTATGTAT CTGACAGTAT TTAAATGTCT TGTTGTAAGT GTATTTGAAT AATTTAAAGT
pDAB1020  ATAAAGTCAG ATCTATGTAT CTGACAGTAT TTAAATGTCT TGTTGTAAGT GTATTTGAAT AATTTAAAGT
pDAB1023  ATAAAGTCAG ATAAAGTCAG CTGACAGTAT TTAAATGTCT TGTTGTAAGT GTATTTGAAT AATTTAAAGT
pDAB1054  ATAAAGTCAG ATCTATGTAT CTGACAGTAT TTAAATGTCT TGTTGTAAGT GTATTTGAAT AATTTAAAGT

pDAB1065  ATAAAGTCAG ATCTATGTAT CTGACAGTAT TTgAATGTCT TGTTGTAAGc GTATTTGAAT AATTTAAAGT
pDAB1069  .......... .TCTATGTAT CTGACAGTAT TTgAATGTCT TGTTGTAAGc GTATTTGAAT AATTTAAAGT

     AM9  ATAAAacCAG ATCTATGTAT CTGAaAGTAT TTgAtTATCT TaTTaTAAGc GTATTTGAAT AAAAAAAAAA

pDAB1104  ATAAAacCAG ATCTATGTAT CTGAaAGTAT TT

          211                                                                  280
     PC1  TTGTAAGATT AAAAAAAAAA AAAAAAAAAA
pDAB1020  TTGTAAGATT AATAAATGTG TGTTTATGTG ATGATAAAAA AAAAAAAAAA AAAAAAAAAA AAAAA
pDAB1023  TTGTAAGATT AATAAATGTG TGTTTATGTG AAAAAAAAAA AAAAAAA
pDAB1054  TTGTAAGATT AATAAATGTG TGTTTATGTG ATGAT

pDAB1065  TTGTAAGATT AATAAATGTG TGTTTATGTG ATGATAAATA TTGATTGAAT TGTCATGCAT TGATTAACCG
pDAB1069  TTGTAAGATT AATAAATGTG TGTTTATGTG ATGATAAATA TTGATTGAAT TGTCATGCAT TGATTAACCG

          281
pDAB1065  TAAATGAATT AAAGGATAAT TTTATCGTTA AAAAAA
pDAB1069  TAAATGAATT AAAGGATAAT TTTATCGTTA AAAAAA
```

Fig. 1. Sequences at the 3′ ends of PMC cDNAs. The cDNAs are arranged into 4 groups; members of each group are identical. Nucleotides not found in the largest group are in lower case. Coding regions are underlined.

adjacent repeats, probably occurred during excision of the phagemid from the original phage vector. Deletions from PMC cDNAs were frequently observed during this manipulation in other *Escherichia coli* hosts, such as SURE (Stratagene) (data not shown).

Distribution of genes coding for PMC

DNA from potato seedlings was probed with a fragment which contains the entire PMC coding region and introns. In each digest, four to six high molecular weight fragments (> 6 kb) hybridized to the probe (Fig. 3A). Assuming that each fragment is unique in the haploid genome, and contains a single PMC gene (since we used enzymes that do not cut within the cloned sequence), potato appears to contain a relatively small family, of four to six genes, coding for PMC. This contrasts with the large family of 16–18 genes coding for the major tuber protein, patatin [32].

DNAs from pea and maize also hybridized to the PMC probe from potato (Figs. 3B and 3C). However, the hybridization signal was weaker, and can be entirely contained in a small (1 to 2 kb) fragment. These species appear to contain only a single gene for a PMC-like product, which is not as large as the potato protein. Their genes probably encode single unit cystatin inhibitors,

similar to oryzacystatin [1, 18]. A cDNA from such a gene has recently been characterized in developing maize kernels [2].

Expression of PMC genes

The 5′ end of PMC message was determined by primer extension. Reverse transcription of poly(A)$^+$ tuber RNA from the 5′ end of the second exon generated a major DNA product 170 bases long (data not shown). Assuming intron 1 was removed from the message, the PMC transcript contained 50 nucleotides beyond the translational start.

Total RNA from potato tubers, stems and leaves was probed with the insert from the cDNA clone PC1 (Fig. 4). There was strong hybridization to a 2.5 kb RNA in the preparation from tubers. A similarly sized RNA from leaf hybridized, but the signal was much weaker. There was no detectable hybridization to stem RNA.

Discussion

PMC is a unique member of the cystatin superfamily of cysteine proteinase inhibitors. Biochemical studies of the purified polypeptide have shown that PMC has eight binding sites for papain within a single polypeptide, and that it can be fragmented by proteinases into smaller polypeptides with the ability to inhibit papain. Partial amino acid sequencing of these polypeptides established that they were similar but not identical in sequence, and that they were all related to the cystatin superfamily of cysteine protease inhibitors [34]. In this work, we deduced the sequence of a full-length PMC polypeptide from the DNA sequence of a cloned gene. The result confirms the multidomain structure of PMC, and clarifies how the domains are concatenated. The interpretation of the gene structure, and some interesting features of the deduced polypeptide sequence, are described in detail below.

Coding regions of a genomic sequence are generally identified from the sequences of mature transcripts captured in cloned cDNAs. Unfortunately, none of the PMC cDNAs were full-length. The highly repetitious nature of these sequences could have created problems for reverse transcription of the mRNA *in vitro*, and for phagemid release and stability in *E. coli*. The longest inserts had other, unrelated cDNAs fused to PMC sequences, or retained an intron. Furthermore, the maximum number of intron-free units which could be cloned together in attempts to make a synthetic PMC cDNA was three (C. Waldron, unpublished data). However, the cDNAs do provide valuable information about the 3′ ends of PMC genes. At least four genes coding for PMC (including the one we cloned) are expressed in tubers, since four different classes of 3′ sequence are present in tuber-derived cDNAs (Fig. 1). Polyadenylation does not occur at a single site, even in transcripts from the same gene e.g. the Group I cDNAs (Fig. 1), and it is not determined by the AAUAAA signal which is important in some, but not all, plant genes [20]. The cDNA regions which code for cystatin-like units lack at least three intervening sequences which are present in the genomic DNA (see Fig. 2). The longest cDNA clone, pDAB1054, still contains one intron sequence which disrupts the reading frame. It probably resulted from a mutation, a transcriptional error, or an aberrant splicing event which increased stability of the insert in *E. coli* (see above).

The 5′ ends of the PMC genes were identified by primer extension from poly(A)$^+$ RNA. The major transcriptional start point corresponds to nucleotide 621 in the sequenced DNA (Fig. 2), assuming that all PMC genes contain the same sized intron between the initiation point and the primer. This assumption is validated for the cloned gene by the presence of motifs associated with transcription initiation, at the expected locations relative to nucleotide 621. Potential CCAAT and TATA boxes lie upstream, at positions −81 to −67 and −35 to −23, respectively. The first translation initiator codon (ATG) downstream occurs at position + 50, and is in the reading frame which codes for a cystatin-like protein. Finally, the sequence around 621 ([!]) is identical

```
   1  AGATGATTTCGAAGATTTAAGGGTATATTGTAATTTTCTTGCAAATCTGAAAAAATATAGTTCATGTTCTTCATCTTCTT

  81  TGTATAATTGCTACAAAAATTTGATACAAAACCACCCAAAATCTTCTCCAAATTGCCTCAAATTTTAGATATAAGCTCCT

 161  CTCAACATTTCGATTCTTTTGACATATTACCCATTCAAAACGGAGCTCGTTCGTGGTTAGTCAAAATTAGAACTCAAGCT

 241  GATAAACTAACTTTAATGGTGGTTTCAAGTTGTGCTTCAAAGTTTCAACCTTTAGCTTCGATTTTCAACCTATAGTTAGT

 321  AATTTAAAAAAGAGGAAAATAATTACCGTCAAATAAACCTCAAAAAGCTAAAAGTATTATATAAATTAGGACCACTTAGT

 401  GGGGAATATTTACCATTTAATGCATTACATATGAGGGGACTCGAGGAAGGTGGATCACTTGATATGAATATTGGATAATT

 481  ATAAAAATTGAGTCTTTTTATTTAATTTTTTTTGGTTATTTTATTTTGAAGCTAAAATCCAAATCCATCAAATCAAAAT

 561  ATTTATAACGTGCTAGCCTTTTTTTTCTCTATAAAAAGCAGTTATGTTCACCTTCTTCTCATCACAAAAACATTCCTTCT

 641  TCTTAATTAAGTTATTAATTATTCGCAGTGATGGCAATCGTAGGAGGCCTTGTCGATGTTCCATTCGAAAACAAAGTCGA
                                  M  A  I  V  G  G  L  V  D  V  P  F  E  N  K  V  E

 721  GTTTGATGATCTTGCTCGTTTTGCTGTCCAAGATTACAATCAGAAAAATGTAAAGAATTATTTTTCATTTACTTCGATTA
       F  D  D  L  A  R  F  A  V  Q  D  Y  N  Q  K  N>

 801  CATCTTAGCTTTGTTATGAAAAGTTACATGTCTTAGTTAACATAATTGATAGTGTAAAATATCTACACATCATCCGTGCA

 881  CAACATTTAAAATGCATTAATGTTACAAATAAGCAGATGACTCTTCGAAGAATATATAATTAATTTTGAAATGCTTAATT

 961  ACTTTGTGAATAATTAGTTGATTCAGATGACTATTCATATTTGTTTCATTTCAACAACATATATTTTGTATTTCAGGATT
                                                                                  D

1041  CTAGTTTGGAGTTTAAAAAGGTTTTGAACGTGAAGCAACAAATAGTTGCTGGAATAATGTACTACATAACATTTGAGGCA
       S  S  L  E  F  K  K  V  L  N  V  K  Q  Q  I  V  A  G  I  M  Y  Y  I  T  F  E  A

1121  ACTGAAGGTGGAAACAAGAAAGAATATGAAGCCAAGATTTTGCTGAGGAAATGGGAGGACTTGAAGAAAGTTGTAGGATT
       T  E  G  G  N  K  K  E  Y  E  A  K  I  L  L  R  K  W  E  D  L  K  K  V  V  G  F

1201  CAAGCTTGTTGGTGATGATAGTACAATGCCTGGGGGCATTGTCAATGTTCCAAACCCAAACAACACCAAGTTTCAAGAAC
       K  L  V  G  D  D  S  T  M  P  G  G  I  V  N  V  P  N  P  N  N  T  K  F  Q  E

1281  TTGCTCGTTTTGCTATTCAGGATTATAATAAAAAACAGGTTAATTATAATTACTTACTCCTCTTTTATTTTTTCGTTAAT
      L  A  R  F  A  I  Q  D  Y  N  K  K  Q>

1361  TTCATATTTAAATCCCGATTTCACTATAGTAGTACCAACATCATACACAATCTATTTTCCAGAATGCTCATTTGGAGTTT
                                                                  N  A  H  L  E  F

1441  GTAGAAAATTTGAATGTTAAAGAGCAAGTTGTTGCTGGAATCATGTACTATATAACACTTGCGGCAACTGATGATGCTGG
       V  E  N  L  N  V  K  E  Q  V  V  A  G  I  M  Y  Y  I  T  L  A  A  T  D  D  A  G

1521  AAAGAAGAAAATATACAAAGCTAAGATTTGGGTGAAGGAATGGGAGGACTTCAAGAAAGTTGTAGAATTCAAGCTTGTTG
       K  K  K  I  Y  K  A  K  I  W  V  K  E  W  E  D  F  K  K  V  V  E  F  K  L  V

1601  GTGATGATATTGCAAAACTTGGGGGCATTACTGATGTTCCATTCCCAAATAACCCCGAGTTCCAAGATCTTGCTCGTTTT
      G  D  D  I  A  K  L  G  G  I  T  D  V  P  F  P  N  N  P  E  F  Q  D  L  A  R  F

1681  GCTATTCAAGTTTATAATAAGAAAGAGGTTAATTCAAATGGCTTACTCCTCTTTTATTTTTCGCTAGTTTCACCTTCAAA
       A  I  Q  V  Y  N  K  K  E>

1761  CTTATAATATTCAAATTCTGAATCCGCTACAGTTTCTAATAATTCTTTCATTTGAATGAATGCCATATACAATCTATTTT

1841  TCAGAATGTTCATTTGGAGTTTGTAGAAAATTTGAACGTTAAACAGCAAGTTGTTGCTGGAATGATGTACTATATAACAC
           N  V  H  L  E  F  V  E  N  L  N  V  K  Q  Q  V  V  A  G  M  M  Y  Y  I  T

1921  TTGCGGCAATTGATGCTGGAAAGAAGAAAATATATGAAACTAAGATTTGGGTGAAGGAATGGGAGGACTTCAAGAAAGTT
      L  A  A  I  D  A  G  K  K  K  I  Y  E  T  K  I  W  V  K  E  W  E  D  F  K  K  V

2001  GTAGAATTCAAGCTTGTTGGTGATGATAGTGCAAAAACTGGGGGCATTATCAATGTTCCAAACCCAAACAGCCCCGAGTT
       V  E  F  K  L  V  G  D  D  S  A  K  T  G  G  I  I  N  V  P  N  P  N  S  P  E  F

2081  CCAAGATCTTGCTCGTTTTGCTGTTCAGGATTATAATAATACACAGGTCAATTATATATAATGACTTACTTTTAGTTTCT
       Q  D  L  A  R  F  A  V  Q  D  Y  N  N  T  Q>

2161  TCTTTTTTTGTTAATTTCACATTAAAACCTATAATATTCAAATCTTTAATCTATTGCAGTATCTAATAATACTTTCATTT

2241  GAACAAATGTCATATACAATCTACTTTTCAGAATGCTCATTTGGAGTTTGTAGAAAATTTGAATGTGAAAGAACAAGTTG
                                      N  A  H  L  E  F  V  E  N  L  N  V  K  E  Q  L

2321  TTTCTGGAATGATGTACTATATAACACTTGCGGCAACTGATGCCGGGAATAAGAAAGAATATGAAGCCAAGATTTGGGTG
      V  S  G  M  M  Y  Y  I  T  L  A  A  T  D  A  G  N  K  K  E  Y  E  A  K  I  W  V

2401  AAGGAATGGGAGGACTTCAAGAAAGTTATAGACTTCAAGCTTGTTGGTAATGATAGTGCGAAAAAACTTGGGGGCTTTAC
       K  E  W  E  D  F  K  K  V  I  D  F  K  L  V  G  N  D  S  A  K  K  L  G  G  F  T

2481  CGAAGTTCCATTCCCAAACAGCCCCGAGTTTCAAGATCTTACACGTTTTGCTGTTCACCAATATAATAAGGACCAGGTTA
       E  V  P  F  P  N  S  P  E  F  Q  D  L  T  R  F  A  V  H  Q  Y  N  K  D  Q>
```

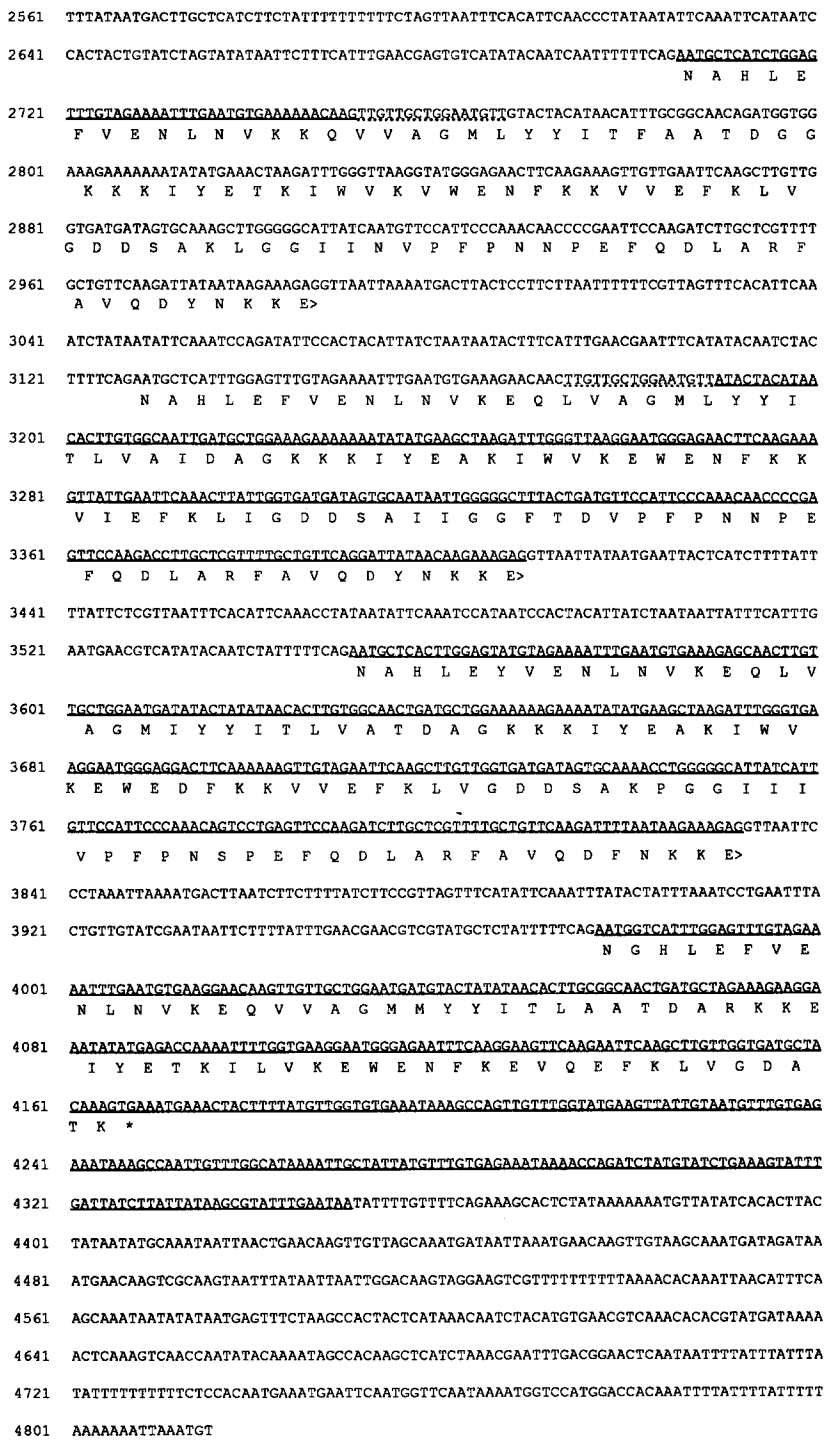

Fig. 2. The nucleotide sequence, and deduced translation product, of the potato genomic DNA encoding PMC. The potential CCAAT and TATA boxes, and the transcription start site are indicated by double underlining. Singly-underlined sequences were also found in the cDNA clone AM9. The dotted underline sequences are the two repeats between which DNA was deleted in AM9 (the cDNA contained only one copy of this repeat).

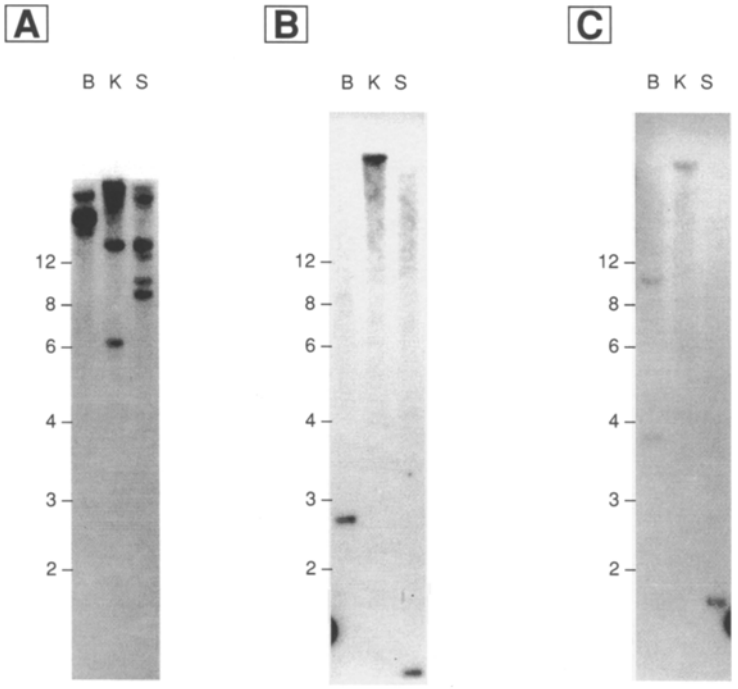

Fig. 3. Genomic DNA fragments with homology to the cloned PMC gene. The autoradiographs of Southern blots show hybridization to fragments obtained by digestion with *Bam* HI (B), *Kpn* I (K) or *Sst* I (S). The migration of size markers (in kb) is shown on the left. A. Potato DNA, exposed for 24 h. B. Pea DNA, exposed for 3 days. C. Maize DNA, exposed for 3 days.

to the consensus for transcription initiation sites in other plant genes [15].

Between the mapped 5′ end of the PMC gene, and the 3′ end cloned in AM9, the genomic DNA contains a sequence encoding eight cystatin-like domains (Fig. 2). However, to maintain this reading frame through translational stop codons and frameshifts, some introns must be postulated in the 5′ portion of the gene. Potential introns were identified by: (1) having 5′-GT and 3′-AG ends, (2) having splice sites similar to the consensus sequences AG|GTAAG and TGCAG|G [10], and (3) maintaining the reading frame between cystatin-like repeats. Eight intervening sequences in the transcribed region satisfy these criteria. They are similar to the introns in other dicot plant genes, being 104 to 267 bases long and AT-rich. Three of them are identical to the introns identified by comparison with a PMC cDNA (Fig. 2). One intron occurs in each of the eight cystatin-coding sequences, at the same position in each repeat. An intron is found at exactly the same

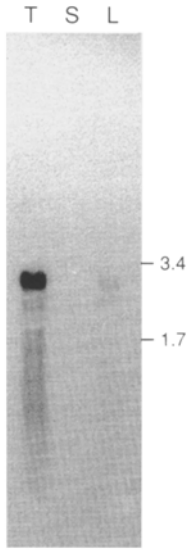

Fig. 4. Analysis of PMC transcripts. An autoradiograph showing hybridization of a PMC-specific probe to total RNAs from tuber (T), stem (S) and leaf (L), fractionated on a denaturing agarose gel. The sizes of ribosomal RNA markers (in kb) are shown on the right.

location in the oryzacystatin genes [17, 19]. In contrast to the chicken ovoinhibitor gene, where the 7 Kazal-type serine proteinase inhibitor domains are flanked by introns [29], the PMC gene does not contain introns separating the regions coding for the cystatin-like domains. The introns in the PMC and oryzacystatin genes occur at different sites from the introns in animal cystatin genes [28], suggesting that they arose after the plant and animal progenitor genes had diverged. However, not all intron locations are conserved in plant cystatin genes – the PMC genomic DNA does not contain an intron in the 3′ untranslated region, where one is found in the oryzacystatin genes [17, 19].

The proposed open reading frame in the PMC gene has a pattern of codon usage similar to that in other dicot genes [23]. It encodes a polypeptide of 757 amino acids, with a calculated M_r of 86778 (Fig. 5). The polypeptide consists of eight homologous cystatin domains of ca. 10.8 kDa, with calculated pI values generally

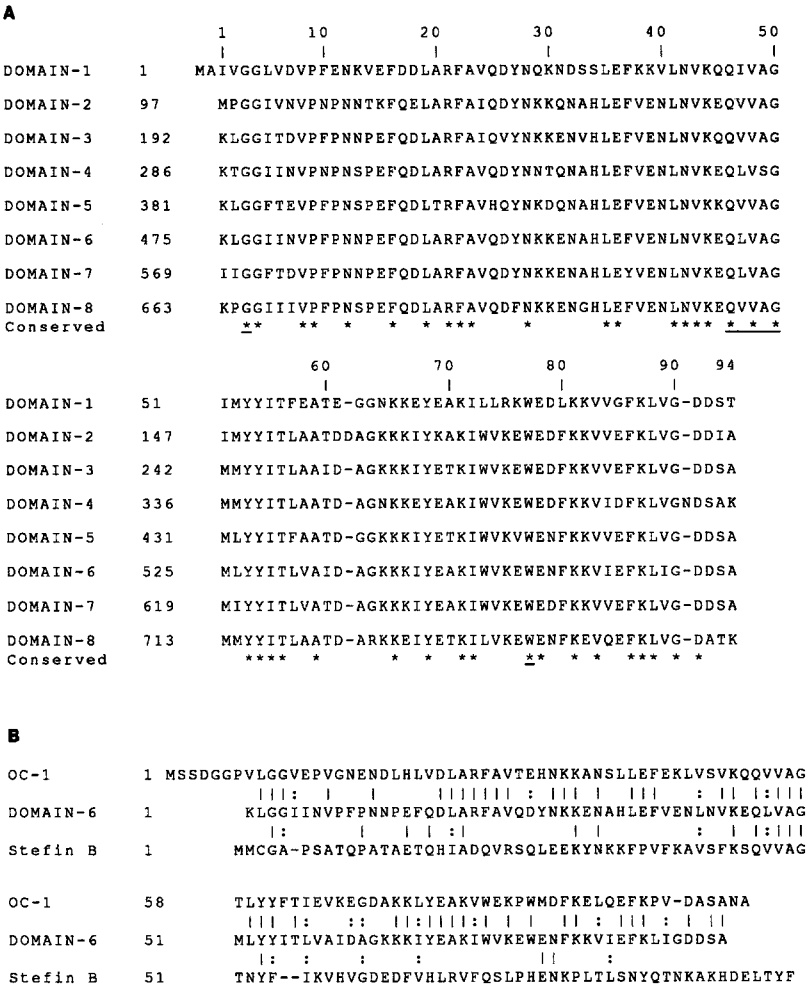

Fig. 5. Comparison of the deduced amino acid sequences of the cystatin-like domains in PMC. A. The eight PMC domains, numbered from the amino terminus of the polypeptide, are aligned to illustrate their similarity. Residues which are identical in all domains are indicated by an asterisk (*). Those residues which contribute to the papain-binding region in cystatins [7, 31] are underlined. B. The amino acid sequence of a representative PMC domain (6) is aligned with the sequences of a plant cystatin (oryzacystatin-1 [1]) and a human cystatin (stefin B [31]). Identical residues are indicated by vertical lines (|) and conservative substitutions are indicated by a colon (:).

between 4.7 and 5.7 (though domains 2 and 5 have a rather basic pI of 8.2). Unlike the kininogens [6], there are no extraneous, non-cystatin sequences. We found no evidence for a secretory pre-peptide at the N-terminus, agreeing with histological observations showing crystals within the cytoplasm of potato cells [8]. Nor does there appear to be any specialized region which might act as a center for crystallization; this unusual ability of PMC must be an intrinsic property of the multiple-cystatin polypeptide. Crystal formation may permit PMC accumulation to high levels (for defensive purposes, see below), without any adverse affects on normal cellular functions.

Of the about 94 amino acid residues within each cystatin-like domain, 38 (40%) are completely conserved in all eight domains, and are quite uniformly distributed throughout the entire polypeptide sequence (Fig. 5). Domains 3 and 6 have the highest homology (89% amino acid identity), while domains 1 and 8 are least homologous (53% identity). The PMC domains are 43–50% identical to oryzacystatins I and II, and maize cystatin (within the 94 residue region of comparison; all three monocot cystatins have somewhat extended N-termini relative to the PMC domains). Such a high degree of conservation suggests that all plant cystatin genes are derived from a common ancestor. However, in some species, such as potato, the cystatin-coding regions have been amplified to form genes for a multidomain inhibitor (Fig. 3A). The PMC domains also share some homology (10–50% identity) with animal members of the cystatin superfamily. Crystallographic studies with some of these cystatins show that they contain specific regions which form a contiguous wedge that effectively blocks the active site of the target cysteine proteinase [7, 31]. The residues involved in this interaction (Gly-3, Gln-Xaa-Val-Xaa-Gly (residues 46–50), and Trp-77 in PMC domain numbering) are conserved in all eight PMC domains (Fig. 5). Quite small variations in the sequence of these regions, either through natural variation or by mutagenesis, can result in significantly different affinities for target proteases [4, 18]. In the seven-residue sequence (residues 45–51) which includes the cystatin signature motif QXVXG, there are six different sequence variants within the eight PMC domains. This variability may be of functional significance by producing a broad spectrum of inhibitory activity within one multidomain polypeptide.

The interdomain connections in the PMC polypeptide, as in other multidomain proteinase inhibitors [29, 33], are susceptible to cleavage by relatively non-specific proteinases such as trypsin, chymotrypsin and subtilisin Carlsberg. Therefore, even though they do not interfere with papain binding, they must be in exposed positions such as a surface loop [24]. Within the seven interdomain regions (considered as residues 1–2 and 80–94 in Fig. 5), there are many conserved residues that account for PMCs susceptibility to cleavage, such as lysine for trypsin, and tryptophan or phenylalanine for chymotrypsin. The N-terminal amino acid sequences obtained from PMC after treatment with trypsin [34] are consistent with the presence of a lysine at residue 1 in five of the domains. Despite these exposed sites, PMC accumulates in the tuber and leaf as a full-length 85 kDa polypeptide [26, 34]. It does not act as a precursor for the release of individual domains in intact potato cells, in contrast to the recently characterized multidomain serine protease inhibitor from *Nicotiana alata* stigmas [3]. The 42 kDa *N. alata* inhibitor is a secreted precursor protein, from which 6 kDa inhibitor domains homologous to the potato protease inhibitor II are processed, leading to the accumulation of mature single-domain inhibitors.

The cloned PMC gene is expressed in tubers, since its 3′ untranslated region was found in a tuber cDNA library (Fig. 2). We therefore looked for sequence motifs which might be involved in the regulation of PMC gene expression, by comparing the DNA upstream of the transcriptional start site with the promoters of tuber-expressed class I patatin genes [21, 32]. No significant homology was detected, probably because PMC and patatin genes are expressed in different types of tuber cells (PMC is restricted to subphellogen cells, whereas patatin is synthesized throughout the parenchyma). PMC also accumulates in in-

jured leaves (T.A. Walsh and J.A. Strickland, Plant Physiol., in press). One wound-induced PMC cDNA has been sequenced [11], and has the same 3′-untranslated region as the Group II tuber cDNAs (Fig. 1). This gene, and potentially other members of the PMC family, can therefore be expressed in response to both developmental and environmental signals. However, we were unable to identify any potential regulatory elements in the region upstream of the cloned PMC gene by sequence homology with the promoters of wound-inducible potato genes [9, 25, 30], or a proteinase inhibitor II gene which is expressed both in tubers and in response to wounding [16]. The wound inducibility of at least one PMC gene, and the accumulation of PMC protein in the layer of cells just below the tuber surface, suggest that PMC is part of the plant's defense against attack by insects and pathogens. PMC could be particularly effective against Coleopteran insects because the guts of these insects are mildly acidic (allowing ingested crystals to solubilize), and their major digestive proteases are generally of the cysteine type [22].

Acknowledgements

We are extremely grateful to W. Michael Ainley for his collaboration on the genomic analyses, and to Don Merlo for his critical reading of this manuscript.

References

1. Abe K, Emori Y, Kondo H, Susuki K, Arai S: Molecular cloning of a cysteine proteinase inhibitor of rice (oryzacystatin). J Biol Chem 262: 16793–16797 (1987).
2. Abe M, Abe K, Kuroda M, Arai S: Corn kernel cysteine proteinase inhibitor as a novel cystatin superfamily member of plant origin. Eur J Biochem 209: 933–937 (1992).
3. Atkinson AH, Heath RL, Simpson RJ, Clarke AE, Anderson MA: Proteinase inhibitors in *Nicotiana alata* stigmas are derived from a precursor protein which is processed into five homologous inhibitors. Plant Cell 5: 203–213 (1993).
4. Auerswald A, Genenger G, Assfalg-Machleidt I, Machleidt W, Engh RA, Fritz H: Recombinant chicken egg white cystatin variants of the QLVSG region. Eur J Biochem 209: 837–845 (1992).
5. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (eds) Current Protocols in Molecular Biology. John Wiley, New York (1987).
6. Barrett AJ, Fritz H, Muller-Esterl W, Grobb A, Isemura S, Jarvinen M, Katunuma N, Machleidt W, Sasaki M, Turk V: Nomenclature and classification of the proteins homologous with the cysteine proteinase inhibitor chicken cystatin. In: Turk V (ed) Cysteine Proteinases and their Inhibitors, pp. 1–2. Walter de Gruyter, Berlin (1986).
7. Bode W, Engh R, Musil D, Thiele U, Huber R, Karshikov A, Brzin J, Kos J, Turk V: The 2.0 A X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. EMBO J 7: 2593–2599 (1988).
8. Boulter D, Harvey PJ: Accumulation, structure and utilization of tuber storage proteins, with particular reference to *Dioscora rotunda*. Physiol Vég 23: 61–74 (1985).
9. Cleveland TE, Thornberg RW, Ryan CA: Molecular characterization of a wound-inducible inhibitor I gene from potato and the processing of its mRNA and protein. Plant Mol Biol 8: 199–207 (1987).
10. Goodall GJ, Filipowicz W: Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. EMBO J 10: 2635–2644 (1991).
11. Hildmann T, Ebneth M, Pena-Cortes H, Sanchez-Serrano J, Willmitzer L, Prat S: General roles of abscisic and jasmonic acids in gene activation as a result of wounding. Plant Cell 4: 1157–1170 (1992).
12. Hines ME, Osuala CI, Nielsen SS: Isolation and partial characterization of a soybean cystatin cysteine proteinase inhibitor of coleopteran digestive proteolytic activity. J Agric Food Chem 39: 1515–1520 (1991).
13. Hirashiki I, Ogata F, Yoshida N, Makisumi S, Ito A: Purification and complex formation analysis of a cysteine proteinase inhibitor (cystatin) from seeds of *Wisteria floribunda*. J Biochem 108: 604–608 (1990).
14. Holmes DS, Quigley M: A rapid boiling method for the preparation of bacterial plasmids. Anal Biochem 114: 386–389 (1981).
15. Joshi CP: An inspection of the domain between putative TATA box and translation start site in 79 plant genes. Nucl Acids Res 15: 6643–6653 (1987).
16. Keil M, Sanchez-Serrano JJ, Willmitzer L: Both wound-inducible and tuber-specific expression are mediated by the promoter of a single member of the potato proteinase inhibitor II gene family. EMBO J 8: 1323–1330 (1989).
17. Kondo H, Abe K, Emori Y, Arai S: Gene organization of oryzacystatin-II, a new cystatin superfamily member of plant origin, is closely related to that of oryzacystatin-I but different from those of animal cystatins. FEBS Lett 278: 87–90 (1991).
18. Kondo H, Abe K, Nishimura I, Watanabe H, Emori Y,

Arai S: Two distinct cystatin species in rice seeds with different specificities against cysteine proteinases. J Biol Chem 265: 15832–15837 (1990).
19. Kondo H, Emori Y, Abe K, Susuki K, Arai S: Cloning and sequence analysis of the genomic fragment encoding oryzacystatin. Gene 81: 259–265 (1989).
20. MacDonald MH, Mogen BD, Hunt AG: Characterization of the polyadenylation signal from the T-DNA-encoded octopine synthase gene. Nucl Acids Res 19: 5575–5581 (1991).
21. Mignery GA, Pikaard CS, Park WD: Molecular characterization of the patatin multigene family of potato. Gene 62: 27–44 (1988).
22. Murdock LL, Brookhart G, Dunn PE, Foard DE, Kelley S, Kitch L, Shade RE, Shukle RH, Wolfson JL: Cysteine digestive proteases in Coleoptera. Comp Biochem Physiol 87B: 783–987 (1987).
23. Murray EE, Lotzer J, Eberle M: Codon usage in plant genes. Nucl Acids Res 17: 477–499 (1989).
24. Neurath H: In: Jaenicke R (ed) Protein Folding, pp. 501–524. Elsevier/North Holland, Amsterdam/New York (1980).
25. Palm CJ, Costa MA, An G, Ryan CA: Wound-inducible nuclear protein binds DNA fragments that regulate a proteinase inhibitor II gene from potato. Proc Natl Acad Sci USA 87: 603–607 (1990).
26. Rodis P, Hoff JE: Naturally occurring protein crystals in the potato. Plant Physiol 74: 907–911 (1984).
27. Rogers SO, Bendich AJ: Extraction of DNA from plant tissues. In: Gelvin SB, Schilperoort RA (eds) Plant Molecular Biology Manual, pp. A6: 1–10. Kluwer Academic Publishers, Dordrecht (1988).
28. Saitoh E, Sabatini LM, Eddy RL, Shows TB, Azen EA, Isemura S, Sanada K: The human cystatin C gene (CST3) is a member of the cystatin gene family which is localized on chromosome 20. Biochem Biophys Res Comm 162: 1324–1331 (1989).
29. Scott MJ, Huckaby CS, Kato I, Kohr WJ, Laskowski M, Tsai M-J, O'Malley BW: Ovoinhibitor introns specify functional domains as in the related and linked ovomucoid gene. J Biol Chem 262: 5899–5907 (1987).
30. Stanford A, Bevan MW, Northcote DH: Differential expression within a family of novel wound-induced genes in potato. Mol Gen Genet 215: 200–208 (1989).
31. Stubbs MT, Laber B, Bode W, Huber R, Jerala R, Lenarcic B, Turk V: The refined 2.4 A X-ray crystal structure of recombinant human stefin B in complex with the cysteine proteinase inhibitor papain: a novel type of proteinase inhibitor interaction. EMBO J 9: 1939–1947 (1990).
32. Twell D, Ooms G: Structural diversity of the patatin gene family in potato cv. Désirée. Mol Gen Genet 212: 325–336 (1988).
33. Vogel R, Assfalg-Machleidt I, Esterl A, Machleidt W, Muller-Esterl W: Proteinase-susceptible regions in the heavy chain of low molecular weight kininogen map to the inter-domain junctions. J Biol Chem 263: 12661–12668 (1988).
34. Walsh TA, Strickland JA: The 85-kDa crystalline cysteine protease inhibitor from potato contains eight cystatin domains. In Brew K *et al.* (eds) Miami Short Reports, Vol 3, p. 53. IRL Press, Oxford (1993).