Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure of the mexicain–E-64 complex and comparison with other cysteine proteases of the papain family

J. A. Gavira,[a] L. A. González-Ramírez,[a] M. C. Oliver-Salvador,[b] M. Soriano-García[c] and J. M. García-Ruiz[a]*

[a]Laboratorio de Estudios Cristalográficos, IACT, CSIC–Universidad de Granada, Granada, Spain, [b]Unidad Profesional Interdisciplinaria de Biotecnología, Instituto Politécnico Nacional, Av. Acueducto s/n, Barrio la Laguna, Ticoman, México D.F., Mexico, and [c]Instituto de Química, Universidad Nacional Autónoma de México, Circuito Exterior, Ciudad Universitaria, México, Mexico

Correspondence e-mail: jmgruiz@ugr.es

Received 22 December 2006
Accepted 2 February 2007

PDB Reference: mexicain–E-64 complex, 2bdz, r2bdzsf.

Mexicain is a 23.8 kDa cysteine protease from the tropical plant *Jacaratia mexicana*. It is isolated as the most abundant product after cation-exchange chromatography of the mix of proteases extracted from the latex of the fruit. The purified enzyme inhibited with E-64 [*N*-(3-carboxyoxirane-2-carbonyl)-leucyl-amino(4-guanido)butane] was crystallized by sitting-drop vapour diffusion and the structure was solved by molecular replacement at 2.1 Å resolution and refined to an *R* factor of 17.7% ($R_{\rm free}$ = 23.8%). The enzyme belongs to the $\alpha+\beta$ class of proteins and the structure shows the typical papain-like fold composed of two domains, the α-helix-rich (L) domain and the β-barrel-like (R) domain, separated by a groove containing the active site formed by residues Cys25 and His159, one from each domain. The four monomers in the asymmetric unit show one E-64 molecule covalently bound to Cys25 in the active site and differences have been found in the placement of E-64 in each monomer.

1. Introduction

Proteolysis in plants is a complex highly regulated process involving many enzymes and proteolytic pathways in various cellular compartments, with cysteine proteases playing an essential role. Cysteine proteases are a class of proteases in which a cysteine residue is located in the active centre. They are involved in protein maturation, degradation and protein rebuilding in response to various external stimuli and also play a housekeeping function to remove abnormal misfolded proteins (Wisniewski & Zagdanska, 2001). They also participate in developmental stages such as germination, morphogenesis and cell biogenesis and senescence, as well as in programmed cell death (Solomon *et al.*, 1999; Palma *et al.*, 2002). In addition, they are involved in perception, signalling and response to biotic and abiotic stress, leading to plant defence (Grudkowska & Zagdanska, 2004; Konno *et al.*, 2004; van der Hoorn & Jones, 2004).

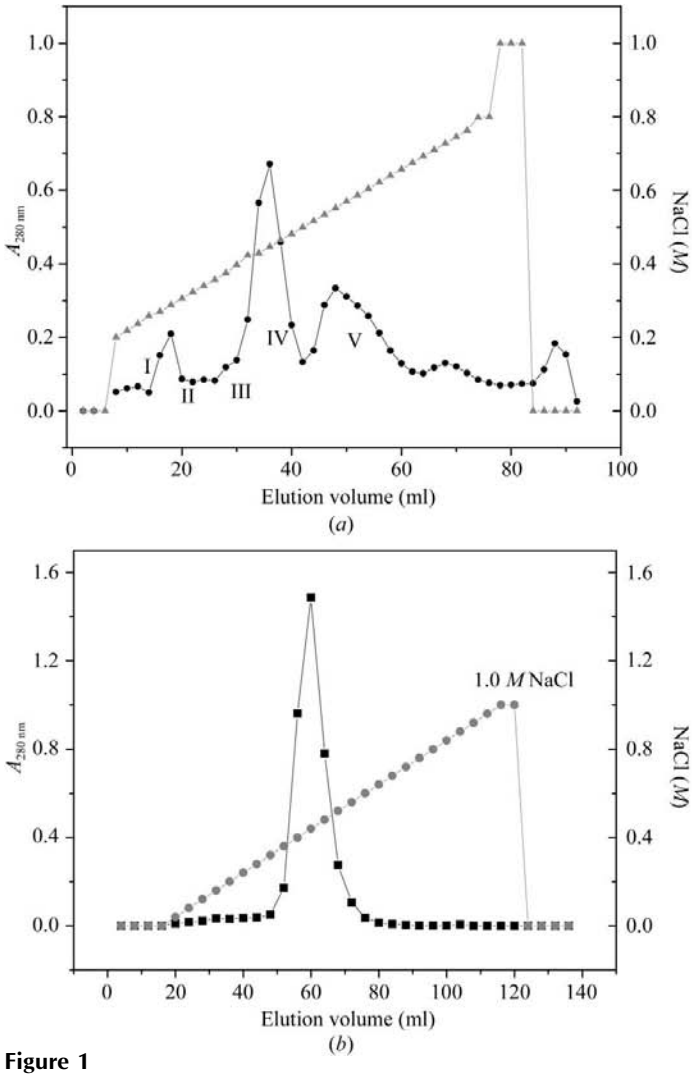
In addition to the important physiological roles played by plant cysteine proteases, the enzymes have also received special attention in the food and biotechnology industries (particularly in the treatment of leather and bioremediation processes) owing to their property of being active over a wide range of temperature and pH; they also have applications in the pharmaceutical industry for the preparation of medicines, for example for the debridement of wounds (Ford *et al.*, 2002) and the prevention of infection of burns (Starley *et al.*, 1999). In addition, cysteine proteases from Caricaceae have been shown to be important in stimulating IgE responses (Furmonaviciene *et al.*, 2000). Their mitogenic properties have also been demonstrated in mammalian cells (Silva *et al.*, 2003; Gomes *et al.*, 2005).

Cysteine proteases are primarily classified on the basis of their sequences and/or tertiary structures and are grouped into superfamilies or clans. The papain family (C1), calpain family (C2) and streptopain family (C10) are closely related evolutionarily and are usually described as 'papain-like' or clan CA proteases. The papain family (C1) is the best studied and enzymes of this family are found in a wide variety of life forms: baculoviruses, eubacteria, yeasts and probably all protozoa, plants and mammals, including humans (Barrett & Rawlings, 1996; Barrett *et al.*, 1998). Although most of the papain-family proteases are cysteine endopeptidases (EC 3.4.22), some members show a variety of other activities. These include endopeptidases with broad specificity (such as papain), endopeptidases with very narrow specificity (such as glycyl endopeptidase), aminopeptidases, a dipeptidyl peptidase and peptidases with both endopeptidase and exopeptidase activities (such as cathepsins B and H; Rawlings & Barrett, 1994).

Several structures of cysteine proteases have been reported and together with kinetic studies they have allowed a mechanism of action to be proposed. In cysteine proteases the attacking nucleophile is the S atom of a cysteine side chain and a histidine is involved in a hydrogen-acceptor/shuttle role (Drenth *et al.*, 1968). Using the Schechter and Berger definition of substrate-binding sites for proteases (Schechter & Berger, 1967), the residue at the P2 position of the substrate has been shown to be the most significant in terms of determining the specificity toward peptide substrates. This amino acid is aromatic (Tyr, Phe or Trp) in most papain-like cysteine proteases.

The isolation of two proteases, named mexicain and chymomexicain, from the latex of the fruit of *Jacaratia mexicana* (formerly *Pileus mexicanus*) was first reported by Castañeda-Agulló *et al.* (1945). The complete amino-acid sequences were determined by solid-phase peptide-sequence analysis and the two proteins show 86% sequence identity; based on sequence homology, both proteins are cysteine proteases. Mexicain has a high proteolytic activity towards casein, haemoglobin and gelatin (Ortega & del Castillo, 1966; Soriano *et al.*, 1975) and when compared with other cysteine proteases shows high pH and temperature stability while maintaining a high proteolytic activity.

As in the case of other cysteine proteases, the proteolytic activity of mexicain can be inhibited by E-64 [*N*-(3-carboxyoxirane-2-carbonyl)-leucyl-amino(4-guanido)butane] with an affinity constant in the region of $\sim 10^{-9}\ M^{-1}$. This inhibitor was first isolated from its natural source as described by Hanada and coworkers (Hanada, Tamai & Ohmura, 1978; Hanada, Tamai & Yamagishi, 1978), who also determined its structure. The mechanism of inhibition by E-64 remains unclear, but it is assumed that the *trans*-epoxide group alkylates the cysteine at the proteinase active site. To date, several structures of cysteine proteases inhibited by E-64 have been solved, such as those of papain (Varughese *et al.*, 1989), actinidin (Varughese *et al.*, 1992), caricain D158E mutant (Katerelos *et al.*, 1996) and cathepsin K (Zhao *et al.*, 1997). E-64 has been found to bind to the active site in two different orientations. It usually binds to the active site in the opposite direction to the substrate, but can also bind in the same direction, as observed with cathepsin B (Turk *et al.*, 1997).



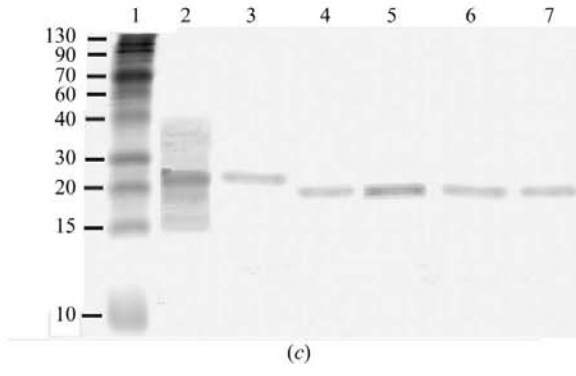


Figure 1
(*a*) Cation-exchange profile of *J. mexicana* latex extracted by cation-exchange chromatography. Mexicain is peak IV. (*b*) Chromatogram of mexicain–E-64 obtained by HPLC cation-exchange chromatography. (*c*) SDS–PAGE of the purified *J. mexicana* proteinases. Lane 2, mexicain (peak IV); lane 3, proteinase V (peak V); lane 1, standard molecular-weight markers (kDa): phosphorylase b (94 kDa), bovine serum albumin (67 kDa), ovoalbumin (43 kDa), carbonic anhydrase (30 kDa), soybean trypsin inhibitor (20.1 kDa) and α-lactoalbumin (14.4 kDa).

Table 1
Data-collection and refinement statistics of the mexicain–E-64 complex.

Values in parentheses are for the highest resolution shell.

Data collection	
Wavelength (Å)	1.54
Temperature (K)	100
Space group	$P2_1$
Unit-cell parameters (Å, °)	a = 57.36, b = 90.45, c = 80.39, α = 90.00, β = 92.64, γ = 90.00
Monomers per ASU	4
Resolution (Å)	2.1 (2.18–2.10)
No. of observed reflections	279184
Redundancy	5.9 (3.1)
Completeness (%)	99.3 (97.1)
$R_{\rm merge}$† (%)	9.5 (29.6)
Average $I/\sigma(I)$	23.9 (1.5)
Refinement	
R (%)	17.7
$R_{\rm free}$ (%)	23.8
No. of reflections in working set	40977
No. of reflections in test set	4576
No. of solvent molecules	379
Average B factor (Å^2)	14.05
R.m.s.d. bond lengths (Å)	0.012
R.m.s.d. bond angles (°)	1.092
Ramachandran plot, residues in (%)	
Most favoured regions	98.0
Allowed regions	2.0
Generously allowed regions	0.0
Disallowed regions	0.0

† $R_{\rm merge} = \sum_{hkl}\sum_i |\langle I\rangle - I_i|/\sum_{hkl}\sum_i I_i$.

The aim of the present study is to describe the crystal structure of the cysteine protease mexicain, which will enhance our understanding of the structure–activity relationship of plant proteases.

2. Materials and methods

2.1. Purification, crystallization and data collection

Purification, crystallization and data collection have previously been reported for mexicain (Oliver-Salvador *et al.*, 2004). Briefly, a crude mixture of proteases was extracted from the latex of *J. mexicana* fruits (cuaguayote) after inhibition with $HgCl_2$ following a previously described protocol (Soriano *et al.*, 1975; Oliver-Salvador, 1999). Material from the above preparation (about 14 mg protein) was applied onto a BioRad Econo-Pac High S column equilibrated with 50 m*M* sodium phosphate buffer pH 6.3 installed on a Pharmacia LKB GradiFrac system and eluted with a linear gradient of 0.05–1.0 *M* NaCl. Five peaks were obtained, of which the largest, identified as proteinase IV (mexicain), was pooled and again applied onto the same chromatographic column. The protein was eluted by isocratic elution with 0.47 *M* NaCl. This purified enzyme preparation was dialyzed against distilled water and treated with dithioerythritol in order to recover fully active mexicain. The enzyme was concentrated to 10 mg ml^{-1}, inhibited with six equivalents of E-64 [N-(3-carboxyoxirane-2-carbonyl)-leucyl-amino(4-guanido)butane] and dialyzed against 20%(*v*/*v*) methanol–water followed by dialysis against water and 50 m*M* sodium phosphate buffer pH 6.3. Mexicain–E-64 was then purified by cation-exchange HPLC on a Protein-Pak SP 8R column (10.0 × 100 mm, 8 µm particle size, 1000 Å pore size) equilibrated with 50 m*M* sodium phosphate buffer pH 6.3 and eluted with mobile phases of 50 m*M* sodium phosphate buffer pH 6.3 (*A*) and 50 m*M* sodium phosphate buffer, 1.0 *M* NaCl pH 6.3 (*B*) using a linear gradient of 0–100% *B* at a flow rate of 1.0 ml min^{-1} on a Varian HPLC system. Protein elution was monitored by measurement of the absorbance at 280 nm. Peaks containing mexicain were pooled, dialyzed against distilled water and concentrated to 8–10 mg ml^{-1}. Fig. 1 summarizes the relevant purification steps.

The proteolytic activity or absence of activity in the case of the mexicain–E-64 complex was evaluated using a modification of the Kunitz method (Ortega & del Castillo, 1966; Oliver-Salvador, 1999). Protein concentrations were determined either by using ε_{280} = 1.92 g^{-1} l cm^{-1} according to the Pace formula (Pace *et al.*, 1995) or using the bicinchoninic acid (BCA) protein-assay reagent (Smith *et al.*, 1985).

Detailed information on protein crystallization and data collection can be found in the previous report (Oliver-Salvador *et al.*, 2004); a summary is given in Table 1.

2.2. Molecular replacement

The initial model for molecular replacement was constructed by the *SWISS-MODEL* server (Schwede *et al.*, 2000) using as input five PDB coordinate files selected by the program with a sequence identity higher than 64% (Oliver-Salvador *et al.*, 2004). The PDB coordinates were those of the homologous chymopapain (PDB code 1yal), protease ω (PDB code 1ppo), caricain mutant D158E (PDB code 1meg), procaricain (PDB code 1pci) and glycyl endopeptidase (PDB code 1gec). The fast direct-rotation algorithm from *CNS* (Brünger *et al.*, 1998) was run using 25–4.0 Å data with $F/\sigma(F) < 0$ rejection criteria and PC refinement with e2e2 target. The top ten picks were used to search for translation solutions with the same criteria and with a fastf2f2 target for the general translation function. Using the molecular weight of 23.7 kDa and assuming the presence of four molecules in the asymmetric unit, the Matthews coefficient ($V_{\rm M}$) was calculated to be 2.24 Å^3 Da^{-1} (Matthews, 1968), corresponding to a solvent content of 45% (Westbrook, 1985); the solution with the highest correlation coefficient and packing values was thus fixed and a new translation search was run with the first molecule fixed. The procedure was repeated until four molecules were fixed in the asymmetric unit and the final location of the four monomers in the asymmetric unit was performed using the nearest positioning of the four monomers in the cell as criteria. After rigid-body refinement of the four monomers and two cycles of minimization of the maximum-likelihood target function, an R value of 0.419 and an $R_{\rm free}$ of 0.416 were obtained.

2.3. Crystallographic refinement and structure validation

The model was refined to a final resolution of 2.1 Å using *CNS* (Brünger *et al.*, 1998) and *REFMAC* (Murshudov *et al.*,

1997). Refinement was conducted using *CNS* (Brünger *et al.*, 1998) with iterative model building using *XTALVIEW* (McRee, 1999) based on $|2F_o - F_c|$ and $|F_o - F_c|$ electron-density maps. After several cycles of simulated-annealing and temperature-factor refinement, the final model had an R value of 0.28 (R_{free} = 0.30). The $|F_o - F_c|$ difference Fourier map contoured at 1.5σ revealed the location of the inhibitor E-64 in the active site of the four monomers (Fig. 2). Strict noncrystallographic symmetry (NCS) averaging was initially used and was gradually shifted to loose restraints before complete elimination to account for different conformations of loop 99–104. Further refinement including the inhibitor E-64 and water molecules using the water-pick protocol of *CNS* lowered the R value to 0.198 (R_{free} = 0.25). Water molecules with B factors higher than 45 Å^2 were removed. Libraries and *CNS* input files for the inclusion and energy minimization of E-64 were generated using *PRODRG* (Schüttelkopf & van Aalten, 2004) and the *HIC-Up* server (Kleywegt & Jones, 1998). The final cycle of refinement was performed using *REFMAC*5, applying restrained refinement from the *CCP*4 software suite (Collaborative Computational Project, Number 4, 1994). After refinement, the quality of the model was checked using *MolProbity* (Lovell *et al.*, 2003) and the model was checked before deposition using *PROCHECK* (Laskowski *et al.*, 1993) and *WHAT_CHECK* (Hooft *et al.*, 1996).

2.4. Comparative studies

Sequence-homology analysis was performed with *ClustalW* (Thompson *et al.*, 1994). Superposition of the mexicain structure with homologous papain-like peptidases was performed using the program *LSQMAN* (Kleywegt & Jones, 1997). The sequence and three-dimensional structure comparison was performed with the Protein Structure Comparison service *SSM* (Krissinel & Henrick, 2004) at the European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/ssm). The active-site cavity volume was determined using the program *QUANTA*-2002 (Accelrys Inc.).

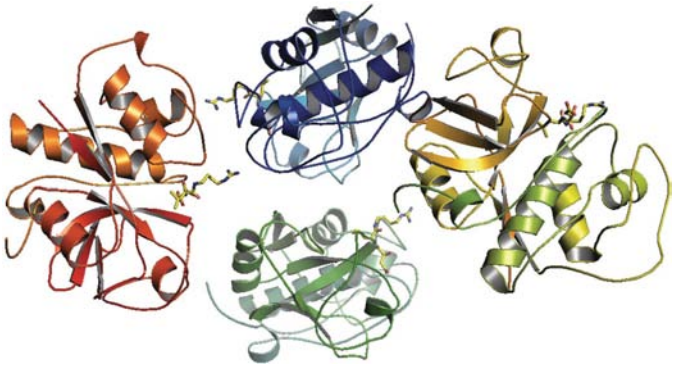


Figure 2
Spatial distribution of the four monomers in the asymmetric unit. The maximum r.m.s.d. of the C$^\alpha$ atoms between monomers was 0.37 Å; it was 1.03 Å when considering the contribution of the side chains.

3. Results and discussion

3.1. Structure of mexicain

The crystallization of mexicain was originally reported by Castañeda-Agulló *et al.* (1945); however, the mexicain–E-64 complex did not crystallize under these previously reported conditions. After improving the purification procedure, new crystallization trials succeeded in producing crystals of the mexicain–E-64 complex that were suitable for X-ray data collection. Inhibited mexicain–E-64 crystallizes at a pH close to 11 in the monoclinic space group $P2_1$, with unit-cell parameters a = 57.36, b = 90.45, c = 80.39 Å, β = 92.64° (Oliver-Salvador *et al.*, 2004). The structure was determined by molecular replacement and refined at 2.1 Å resolution to a final R factor of 17.7% (R_{free} = 23.8%). The model contains

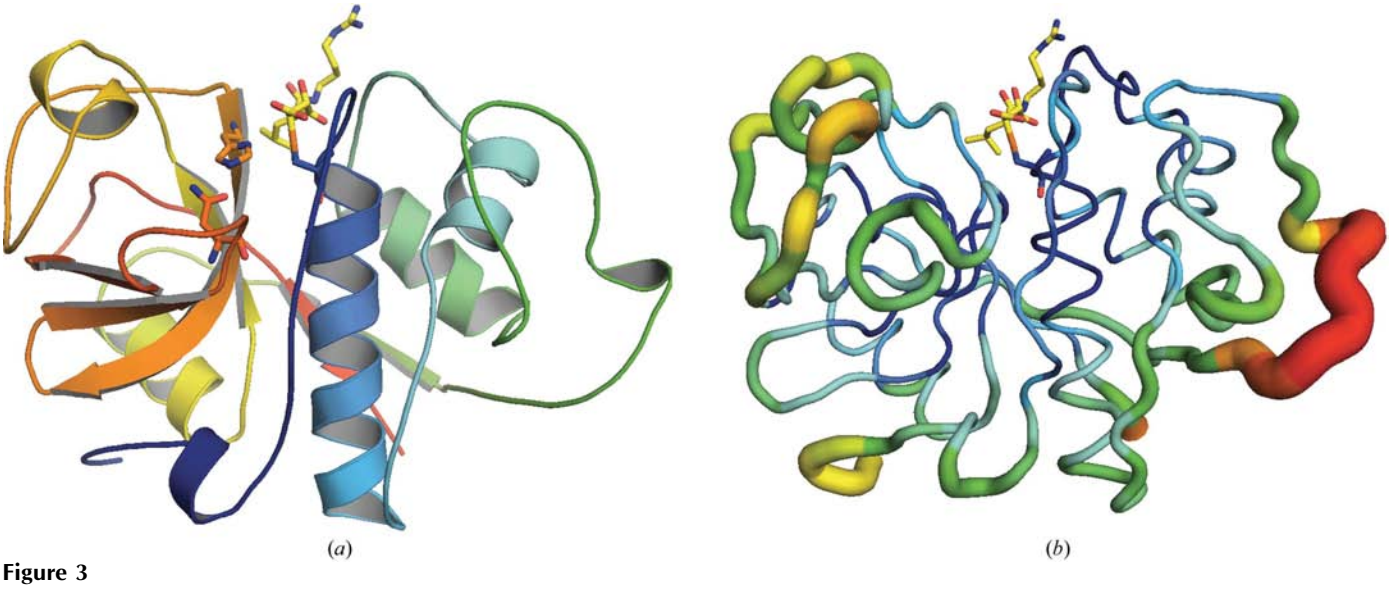


Figure 3
(*a*) Cartoon representation of the monomers of mexicain. The active-site Cys25 is covalently bonded to the inhibitor E-64. Residues His159 and Asn175 are also represented as sticks. (*b*) The amino-acid B-factor distribution is shown as a colour gradient from red to blue and as a tube diameter that decreases when the B factor diminishes.

Table 2
Sequence and structural homology comparison of mexicain with other proteases of the papain family.

PDB code	*Q* score	R.m.s.d.	Residues aligned	Length	Identity (%)	%seq† (%)	Ligand at Cys25	Protein	Reference
1meg	0.95	0.557	212 (139)	216	72.22	0.6557	E-64	Caricain D158E	Katerelos *et al.* (1996)
1yal	0.94	0.575	212 (147)	218	74.03	0.6887	SCH	Chymopapain	Maes *et al.* (1996)
1ppo	0.94	0.647	212 (140)	216	72.13	0.6604	Hg	Protease ω	Pickersgill *et al.* (1991)
1khq	0.91	0.644	207 (128)	212	64.98	0.5913	GLM	Papain	Janowski *et al.* (2004)
1gec	0.91	0.603	208 (142)	216	71.07	0.6683	CBZ	Glycyl endopeptidase	O'Hara *et al.* (1995)
1aec	0.85	0.881	206 (109)	218	56.11	0.468	E-64	Actinidin	Varughese *et al.* (1992)
1atk	0.81	1.002	203 (97)	215	45.81	0.468	E-64	Human cathepsin K	Zhao *et al.* (1997)

† Sequence identity %seq is a quality characteristic of C^{α} alignment. It is calculated as the fraction of pairs of identical residues among all aligned: %seq = $N_{\rm ident}/N_{\rm align}$ (Krissinel & Henrick, 2004).

212 amino-acid residues and 379 water molecules and one E-64 molecule is covalently bound to Cys25 of each monomer in the active site. The refinement parameters and statistics of the final model of mexicain are summarized in Table 1.

The structure of mexicain shows the typical papain-like fold described by Drenth *et al.* (1968), composed of two domains, an α-helix-rich (L) domain and a β-barrel-like (R) domain, separated by a groove containing the active site formed by residues Cys25 and His159, one from each domain (Fig. 3*a*). The L domain (residues 11–108) is composed of four helices and the R domain (residues 1–10 and 115–214) is formed of a β-barrel (seven β-sheets) and three small helices at the surface, which are typical features of the C1 papain-like fold. The C- and N-termini of the R and L domains, respectively, bind to the L and R domains to stabilize the binding region.

Mexicain contains the seven cysteine residues common to the papain family, with six forming disulfide bonds (Cys22–Cys63, Cys56–Cys95 and Cys153–Cys200); Cys25 is the active catalytic residue.

The four molecules in the asymmetric unit, *A*, *B*, *C* and *D*, are crystallographically independent (Fig. 2). Size-exclusion chromatography and dynamic light-scattering studies indicate that mexicain is a monomer in solution with a mean radius ($R_{\rm H}$) of 2.5 nm and a molecular weight of approximately 28 kDa (Moreno *et al.*, 2000) in all the conditions assayed and, therefore, that the asymmetric unit formed by four molecules is just a crystallographic association. The protomers are very similar, as shown by the superposition of C^{α} atoms with a maximum r.m.s. deviation of 0.37 Å and of 1.03 Å when considering the contribution of the side chains.

Initially, the model was built into the electron density using the sequence of mexicain deposited in the SWISS-PROT database with accession code P84346. The quality of the electron density for the main chains and the side chains allows an accurate interpretation of the structure, with the exception of the C-termini and the loop region formed by residues 99–104. At an early stage of refinement, the $|F_{\rm o} - F_{\rm c}|$ Fourier map contoured at 1.5σ showed a discrepancy between the input sequence and the side chains in all the monomers at position 177. A comparison with the sequences of homologous proteases and of chymomexicain suggests the presence of tryptophan instead of isoleucine. As the model was completed, the electron density was improved and four more amino acids (Glu9, Tyr58, Pro70 and Pro103) did not show electron density corresponding to their side chains. An OMIT map from which the suspicious amino acids were removed shows the electron density to be more similar to other side chains in all these positions. The first attempt to check for alternate amino acids in the electron density was performed using those that appear in the same position in homologous sequences (Fig. 4). When Tyr58 and Pro70 were changed to Arg58 and Thr70, a better match in the electron density was obtained in all the monomers, resulting in a decrease in the *R* factors. In the case of Glu9 and Pro103, identification is more difficult. Glu9 is located at the surface without the possibility of establishing crystal contacts, while Pro103 belongs to the disordered loop 99–104 (Fig. 3*b*).

The average *B* factor of the structure is 14.0 Å^2 (the average *B* factor for chain *A* is 12.3 Å^2, for chain *B* 14.2 Å^2, for chain *C* 14.8 Å^2 and for chain *D* 14.9 Å^2), the lowest *B* factors being for the residues in the active-site cleft (Fig. 3*a*). The overall geometry was good, with 100% of the residues in the allowed regions and 98.1% in the favoured regions of the conformational space in Ramachandran plots generated using *MOLPROBITY* (Lovell *et al.*, 2003).

3.2. Comparative analysis with the highest structural homologous

Mexicain shows high temperature and pH stability. The enzyme is active in the pH range 3–10 at 293 K, with maximum proteolytic activity towards casein in the pH range 8.5–9.0 (Oliver-Salvador, 1999). The optimal activity temperature with the same substrate is 338 K. Other cysteine proteases show a lower optimal activity temperature. For example, the proteases freesia protease B from *Freesia reflacta* and araujiain h I from *Araujia hortorum* show optimal activity temperatures of 323–328 and 333 K, respectively, using the same substrate (Kaneda *et al.*, 1997; Priolo *et al.*, 2000).

An alignment of the sequence of mexicain with those of structural homologues (Fig. 4) shows high homology with chymopapain, with 74.03% of the residues conserved. Mexicain shares the two amino-acid elongations at the C-terminus, Tyr and Arg, with chymopapain, but this has not been modelled owing to lack of electron density.

The next most homologous protease is papain. Mexicain shares with papain the deletion of four residues (169–172), which are only absent from these two proteins of the papain family (Fig. 4*b*). The major differences between mexicain and papain are observed in the flexible loops 99–106 and 191–196.

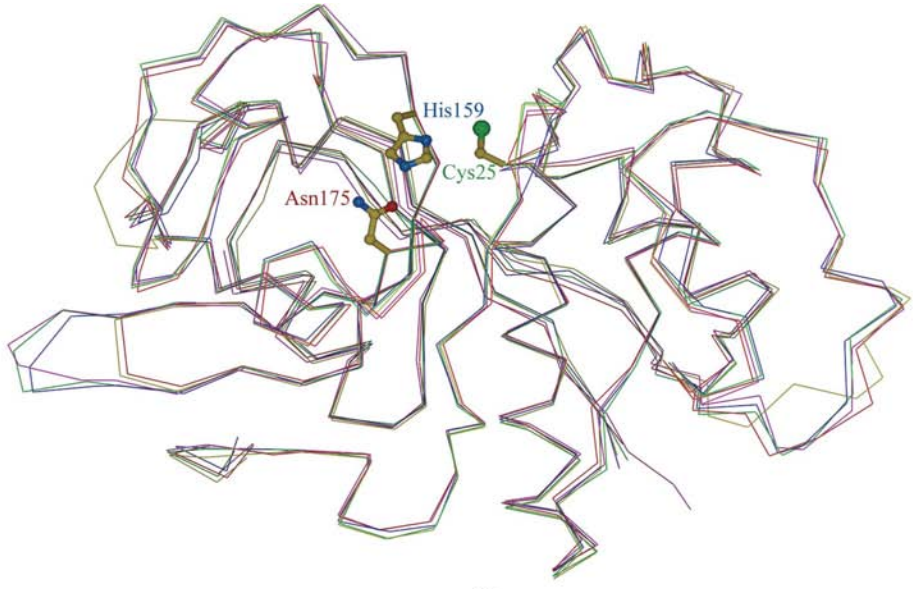
The first loop consists of eight amino acids with a very low homology. The second loop shows lower *B*-factor values when compared with papain and homologues and is therefore less flexible.

Using the *Q* score as a criterion that takes into account not only the sequence homology but also the three-dimensional structure homology, the best (highest) score is obtained for caricain (mutant D158E, PDB code 1meg), followed by chymopapain and protease ω (Table 2 and Fig. 4). The average r.m.s.d. of the mexicain and caricain–D158E coordinates is below 0.7 Å, showing high structural homology.

3.3. Comparative analysis of the active site–E-64 interaction

Electron density for the inhibitor E-64 in the four monomers was found at an early stage of the refinement in both $|2F_o - F_c|$ and $|F_o - F_c|$ maps and the E-64 molecule was modelled attached to Cys25. The average distance between C2 of E-64 and the SG atom of Cys25 is 2.2 Å, with the lowest distance found for monomer *B* (2.07 Å). These values are larger than those observed in papain bound to E-64c (1.8 Å), which is very close to the theoretical value.

The E-64 amino-4-guanidinobutane moiety is highly surface-exposed, well defined in the electron density and appears between both domains in each molecule. The electron density for the leucyl group of the inhibitor E-64 is different in chain *C* and in this chain the E-64 molecule has been modelled in a different conformation, similar to that in other E-64 structures (Fig. 5). The observed leucyl side chain of E-64 is also well defined and is located at the entrance of the hydrophobic pocket named the S_2 subsite. Mexicain is a low-specificity peptidase and, similar to other cysteine proteases of



(*a*)

```
                    10        20        30        40        50        60        70        80        90       100       110
                    |         |         |         |         |         |         |         |         |         |         |
Mexicain            YPESIDWREKGAVTPVKNQNPCGSCWAFSTVATIEGINKIITGQLISLSEQELLDCERRSHGCDGGYQTTSLQYVVDNG-VHTEREYPYEKKQGRCRAKDKKGPKVYITG
Mex1_JM             ..........................................................Y..........P..........................................
Chymomexicain       ........D.....................V.....R..K............D.....K.....G.I...A...G....K.........K....E...T..Q...
Papain              I..YV...Q..........GS.........A.V.....I..R..N.NEY........D...Y..N...PWSA..L.AQY..I.YRNT....GV.RY..SRE.GPYAAKTD.
Chymopapain         ..Q.....A..........GA...X.....I..V......V..N.LE......V..DKH.Y..K...........AN......SKV...QA..YK...T..P....K...
Protease Omega      L..NV...K........RH.GS.........A...V......R..K.VE......V..........K...PPYA.E..AK...I.LRSK...KA...T....QVG..I.KTS.
Glycyl endo.        L...V...A........H.GY.E..........V......K..N.VE......V..DLQ.Y..NR...S......AQ...I.LRAK...IA..QT...NQVG....KTN.
Caricain-D185D      L..NV...K........RH.GS.........A...V......R..K.VE......V..........K...PPYA.E..AK...I.LRSK...KA...T....QVG..I.KTS.
                     *: :*** ********::*. * * ****:: .*:*** ** **:* . *****:**: :*:**. **   ::: *.. * :*    ***   *  **: :      .  *
Prim.cons.          YPES2DWR3KGAVTPVKNQG2CGSCWAFSTVATVEGINKIRTG2L2ELSEQEL2DC2RRSHGCKGGYQTTSLQYVADNGG2HTR32YPYEAKQGTCRAK2KGGPKVK2TG

                    120       130       140       150       160       170       180       190       200       210
                    |         |         |         |         |         |         |         |         |         |
Mexicain            YKYVPANDEISLIQAIANQPVSVVTDSRGRGFQFYKGGIYEGPCGTNTDHAVTAVGYG----KTYLLLKNSWGPNWGEKGYIRIKRASGRSKGTCGVYTSSFFPIKGYR
Mex1_JM             ..........................................................................................................
Chymomexicain       ..R..........G.G.....LHE.K..A..L.....FN....YKN.....I........AQ..D.............K......K.E......K..Y.......
Papain              VRQ.QPYN.GA.LYS.........LEAA.KD..L.R...FV....NKV....A.....PN..--.I.I.....TG...N.......GT.N.Y.V..L.....Y.V.N--
Chymopapain         ..R..S.X.T.FLG.L....L..LVEAG.KP..L..S.VFD.....KL..........TSDG.N.III.............M.L..Q..N.Q......K..YY.F..FA
Protease Omega      VGR.QP.N.GN.LN...K......VE.K..P..L.....F......KV..........KSGG.G.I.I.....TA.............P.N.P.V..L.K..YY.T.N--
Glycyl endo.        VGR.QS.N.G..LN...H......VE.A..D..N.....F..S...KV..........KSGG.G.I.I......G...N.....R....N.P.V....R..YY...N--
Caricain-D185D      VGR.QP.N.GN.LN...K......VE.K..P..L.....F......KVE.........KSGG.G.I.I.....TA.............P.N.P.V..L.K..YY.T.N--
                       *  .   *  :: .:.:**:**: :: *: ** *:.*:: *.** : :***:*:*** .      :: *****. ***:**::::* .*.* *.**:* **::* *.
Prim.cons.          2KRV22NNEGSLL2AIANQPVSVVVESKGRPFQLYKGGIFEGPCGTKVDHAVTAVGYGKSGGKGYILIKNSWGPNWGEKGYIRIKRASGNSPG2CGVYKSSYYPIK2YR
```

(*b*)

Figure 4
(*a*) Superposition of the C^{α} traces of mexicain (yellow), papain (red), chymopapain (pink), glycyl endopeptidase (green) and caricain (blue). The positions of residues Cys25, Asn175 and His159 are also shown. The coordinates correspond to the PDB codes in Table 2. (*b*) Sequence alignment of the homologues, including chymomexicain, and the consensus sequence generated using *ClustalW*.

Table 3
Hydrogen-bonding interactions of E-64 with mexicain, papain, actinidin and human cathepsin K.

For clarity, common residues have only been named in the the first column. In the case of papain, values marked with asterisks are not known because the structure has not been deposited in the PDB.

	Mexicain								Papain		Actinidin		Human cathepsin K	
E-64	Chain *A*	*d* (Å)	Chain *B*	*d* (Å)	Chain *C*	*d* (Å)	Chain *D*	*d* (Å)		*d* (Å)		*d* (Å)		*d* (Å)
O1	His159 ND1†	3.29		3.49		3.29		3.59		2.90		2.75		2.95
O1	H_2O 1132	2.60	H_2O 1372	2.49					H_2O 217	2.67	H_2O 582	2.73		
O1			H_2O 1310	2.51	H_2O 1258	2.88	H_2O 1160	2.81	*		H_2O 215	3.13	H_2O2 32	2.66
O2	Cys25 N	3.13		2.97		3.00		2.98		2.85		2.89		3.06
O2	Gln19 NE2	2.92		2.68		2.88		2.58		2.86		2.87		2.95
O2	Ser24 N	3.44		3.23		3.47		3.32		3.11	Gly24 N	3.33		3.31
O3			Asp158 O	3.45										
O3	H_2O 1101	2.43	H_2O 1018	2.86	H_2O 1068	2.49	H_2O 1040	2.49	*		H_2O 586	2.65	H_2O 225	2.91
O3	H_2O 1371	2.82			H_2O 1347	2.94			*					
O3			H_2O 1310	3.27										
O4	Gly66 N‡	2.85		2.88		2.92		2.92		2.88		2.90		3.07
O5	H_2O 1371	2.77			H_2O 1347	2.62			*					
N1	Asp158 O§	3.17		2.92		2.92		3.01		3.40		3.10		3.29
N1	H_2O 1371	3.28			H_2O 1371	3.35			*					
N2	Gly66 O‡	2.65		2.88		2.88		2.75		3.04		2.87		2.83
N5	Asp64 O	2.71		2.70		2.70		2.77	Tyr67 OH	2.93	Tyr66 O	3.35	Glu59 O	3.46

† Numbered 162 in actinidin and human cathepsin K. ‡ Numbered 68 in actinidin. § Numbered 161 in actinidin and human cathepsin K. This interaction is not described in the original work of Varughese *et al.* (1989), but was described during the comparison of papain with actinidin by the same author (Varughese *et al.*, 1992).

the papain family, is considered to possess a relatively broad substrate specificity with no amino-acid restriction at the substrate P_1 site. However in mexicain Asp, Glu and Ile seem to be disfavoured at the P_2 site (work to be published), unlike papain and ginger protease II, which show a preference at this site for aromatic amino acids (Berger & Schechter, 1970) and proline (Choi *et al.*, 1999), respectively. Small differences are also detected at the hydrophobic site S_2. In mexicain this site is composed of Tyr67, Gln68, Thr69, Val133 and Thr157, with Thr157 being a particular feature of mexicain. In chymopapain the first three amino acids belonging to the L domain are identical, but Val133 and Thr157 are replaced by Leu in both positions. In papain the subsite S_2 is formed by Tyr, Pro, Trp, Val and Val, residues that are more hydrophobic than those found in mexicain, which could explain its preference for aromatic residues at subsite P_2.

In mexicain and chymomexicain, position 20 is occupied by Asn, whereas in most cysteine proteases (papain, chymopapain, protease ω, glycyl endopeptidase, caricain and another 31 plant proteases, including ginger proteases GP-1 and GP-2) this residue is glycine (Choi *et al.*, 1999; Lopez *et al.*, 2000). The presence of Asn20 and its side-chain orientation, which is fixed by a hydrogen bond to the carbonyl group of Asn18, may contribute to the stability of the active site and catalytic activity and will be investigated by direct mutagenesis.

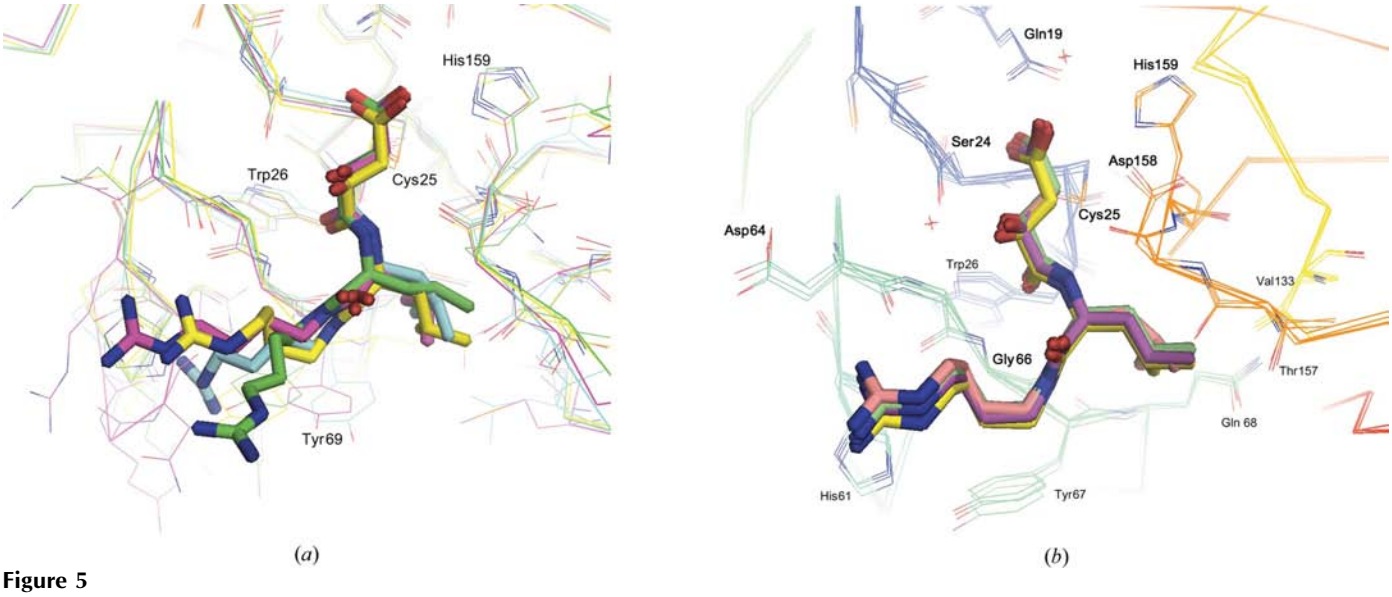


Figure 5
(*a*) Superposition of E-64 at the active site of the four monomers of mexicain that constitute the asymmetric unit. Residues involved in hydrogen bonding with E-64 are labelled in bold. Other residues have been labelled for reference. (*b*) Superposition of E-64 at active site of mexicain (C atoms in yellow), caricain (green), human cathepsin K (cyan) and actinidin (pink). Some residues are labelled for reference.

Several structures of cysteine proteases bound to the E-64 inhibitor have been described previously (Varughese *et al.*, 1989, 1992; Katerelos *et al.*, 1996; Zhao *et al.*, 1997). Table 3 lists the hydrogen-bonding distances involving E-64 and the active centre of mexicain and comparisons with the interactions of E-64 with papain, actinidin and human cathepsin K have also been included. As a general rule, most of the interactions involving the residues of the active site are similar to those observed in the homologues. However, taking into account the distances measured in the four mexicain monomers, it is not clear that those interactions are necessarily conserved. Moreover, in mexicain the binding of E-64 is in the opposite direction to that found in cathepsin B (Turk *et al.*, 1997). In all four mexicain–E-64 monomers, the correct orientation of the His159 side chain with regard to the nucleophilic cysteine Cys25 is maintained by hydrogen bonding to the side-chain O atom of a conserved asparagine residue, Asn175 (hydrogen-bond distances range from 2.77 to 2.89 Å). Asn175 is at a distance of ∼7 Å from the inhibitor E-64, but forms a hydrogen bond to His159 (∼2.8 Å) at the active site. Asn175 is well conserved among cysteine proteases and has been found to form a hydrogen bond with the histidine in the active site in all protease–E-64 complex structures. Beers *et al.* (2004) have proposed that Asn175 plays an important role in the proper orientation of the histidine side chain. Furthermore, the tryptophan residue Trp177 and the 11-residue loops (Asn175–Gly185) are also well conserved. These residues wrap the hydrogen bond formed between the Asn-His pair protecting the active site of the enzyme. Together with the main-chain N atom of the catalytic cysteine Cys25, the side-chain amide of the conserved glutamine Gln19 stabilizes the negative charge developing on the scissile carbonyl O atom during nucleophilic attack. This 'oxyanion hole' is present in all four monomers in the asymmetric unit (Fig. 5*a*).

The longer His159 ND1–E-64 carboxyl O1 distance compared with other E-64 complex structures makes this interaction in mexicain weaker than in papain, actinidin and cathepsin K (Table 3). The second carboxylic oxygen (O2) shows three hydrogen-bonding interactions with Cys25 N, Gln19 NE2 and Ser24 N which are also present with similar distance values in the other three homologues.

Several water molecules have been found to interact with key residues in the active site of mexicain (Table 3). The O1 atom of the carboxylic group of the inhibitor interacts with the ND1 atom of the imidazole group of His159 and with a water molecule in monomers *A* and *B* but not in *C* and *D*. This water molecule also interacts with Gln19 NE2, His159 ND1 and Trp177 NE1 and is also present in papain and actinidin. In monomers *B*, *C* and *D* a second water molecule participates in this interaction and forms a hydrogen bond with Ser136 OG in monomers *C* and *D*, while in monomer *B* it is also bonded to Asp158 O. The equivalent water molecule is found in actinidin and cathepsin K, bonded to residues Asp161 O and Asn161 O, respectively. The O3 atom of E-64 shows a conserved hydrogen bond to a water molecule that interacts with residues Gly23 O and Cys63 O in the four monomers and with other water molecules: either one in chains *C* and *D* or two in chain *B*. In chains *A* and *C*, E-64 O3 is also bonded to a water molecule acting as a bridge with O5, N1 and Asp158 O that is not present in chain *B*, where the hydrogen bond is formed between O3 and Asp158 O and water 1310, the same water that interacts with the atom O1 of the E-64 molecule. The same water molecule is also present in caricain and cathepsin, but in caricain it only interacts with other water molecules, while in cathepsin it only interacts with Gly64 O.

Table 4
Estimation of the cavity volume at the active site using *QUANTA*.

Molecule	Cavity volume (Å^3)	No. of water molecules
1	873.9	47
2	905.2	49
3	1401.5	76
4	1228	66
1khq	751.5	40

The small discrepancy between the interactions in the active site for the four monomers in the asymmetric unit shows that the active sites of the four monomers are similar but not identical. This feature reveals that the interaction at the active site is very flexible and certain interactions considered to be essential can be replaced by others in which water molecules play an essential role. Moreover, the flexibility at the active site can also be considered when taking into account the volume of the cavity that accommodates the inhibitor (Table 4), which ranges from 873.9 Å^2 for monomer *A* to 1401.5 Å^2 for monomer *C* and is larger when compared with the same cavity in papain (751.5 Å^2). The cavity volumes were estimated using *QUANTA* (Accelrys Inc., San Diego, CA, USA).

The four monomers of mexicain show an open active-site cleft and this agrees with the general observation that an open cleft will accept a wider variety of conformations of bound substrates and inhibitors and may explain the low specificity of mexicain, which allows attack by a variety of proteins as part of their protective role in the plant (Azarkan *et al.*, 2004; Konno *et al.*, 2004; van der Hoorn & Jones, 2004).

4. Conclusions

Mexicain is a cysteine protease belonging to the papain family. This enzyme shows a high stability in the pH range 3–10 and high thermal stability as demonstrated by its enzymatic activity. Mexicain is a low-specificity peptidase with no amino-acid restrictions at the substrate P_1 site, although Asp, Glu and Ile seem to be disfavoured at the P_2 site. This broad protease specificity is structurally supported on the basis of a flexible active site and an open cavity that can allow easy access to the substrate. The presence of asparagine at position 20 is a particular feature of mexicain in the papain family and may play a role in the stability of the active-site structure or the catalytic activity. These studies will form the basis of understanding the biological function and specificity of mexicain towards natural and synthetic substrates. These results will help to compare mexicain with other cysteine proteases in order to obtain a complete view of their biological function in plants.

This work was supported by the Spanish Ministry of Science and the Technology Project ESP2003-04759, the Andalucia Regional Government Project RNM1344 and by the Mexican National Polytechnic Institute (IPN) Projects CGPI 20021341 and 20030510. This paper is a result of the Factoría de Cristalización funded by the program Consolider-Ingenio 2010. JAG was supported by the Andalusian Regional Government, Spain. We also would like to thank Dr Ana Cámara-Artiga for her helpful comments and revision of the manuscript.

References

Azarkan, M., Wintjens, R., Looze, Y. & Baeyens-Volant, D. (2004). *Phytochemistry*, **65**, 525–534.

Barrett, A. J. & Rawlings, N. D. (1996). *Perspectives in Drug Discovery and Design*, Vol. 6, edited by J. H. McKerrow & M. N. G. James, pp. 1–9. Leiden: ESCOM.

Barrett, A. J., Rawlings, N. D. & Woessner, J. F. (1998). *Handbook of Proteolytic Enzymes.* London: Academic Press.

Beers, E. P., Jones, A. M. & Dickerman, A. W. (2004). *Phytochemistry*, **65**, 43–58.

Berger, A. & Schechter, I. (1970). *Philos. Trans. R. Soc. London Ser. B*, **257**, 249–264.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Castañeda-Agulló, M., Hernandez, A., Loaeza, F. & Salazar, W. (1945). *J. Biol. Chem.* **159**, 751–751.

Choi, K. H., Laursen, R. A. & Allen, K. N. (1999). *Biochemistry*, **38**, 11624–11633.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M. & Wolthers, B. G. (1968). *Nature (London)*, **218**, 929–932.

Ford, C. N., Reinhard, E. R., Yeh, D., Syrek, D., De Las Morenas, A., Bergman, S. B., Williams, S. & Hamori, C. A. (2002). *Ann. Plast. Surg.* **49**, 55–61.

Furmonaviciene, R., Sewell, H. F. & Shakib, F. (2000). *Clin. Exp. Allergy*, **30**, 1307–1313.

Gomes, M. T., Mello, V. J., Rodrigues, K. C., Bemquerer, M. P., Lopes, M. T., Faca, V. M. & Salas, C. E. (2005). *Planta Med.* **71**, 244–248.

Grudkowska, M. & Zagdanska, B. (2004). *Acta Biochim. Pol.* **51**, 609–624.

Hanada, K., Tamai, M. & Ohmura, S. (1978). *Agric. Biol. Chem.* **42**, 529–536.

Hanada, K., Tamai, M. & Yamagishi, M. (1978). *Agric. Biol. Chem.* **42**, 523–528.

Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

Hoorn, R. A. L. van der & Jones, J. D. (2004). *Curr. Opin. Plant Biol.* **7**, 400–407.

Janowski, R., Kozak, M., Jankoswska, E., Grzonka, Z. & Jaskolski, M. (2004). *J. Pept. Res.* **64**, 141–150.

Kaneda, M., Yonezawa, H. & Uchikoba, T. (1997). *Biosci. Biotechnol. Biochem.* **61**, 1554–1559.

Katerelos, N. A., Taylor, M. A., Scott, M., Goodenough, P. W. & Pickersgill, R. W. (1996). *FEBS Lett.* **392**, 35–39.

Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 525–545.

Kleywegt, G. J. & Jones, T. A. (1998). *Acta Cryst.* D**54**, 1119–1131.

Konno, K., Hirayama, C., Nakamura, M., Tateishi, K., Tamura, Y., Hattori, M. & Kohno, K. (2004). *Plant J.* **37**, 370–378.

Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Lopez, L. M., Sequeiros, C., Natalucci, C. L., Brullo, A., Maras, B., Barra, D. & Caffini, N. O. (2000). *Protein Expr. Purif.* **18**, 133–140.

Lovell, S. C., Davis, I. W., Adrendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*, **50**, 437–450.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

Maes, D., Bouckaert, J., Poortmans, F., Wyns, L. & Looze, Y. (1996). *Biochemistry*, **35**, 16292–16298.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Moreno, A., Mas-Oliva, J., Soriano-Garcia, M., Salvador, C. O. & Bolanos-Garcia, V. M. (2000). *J. Mol. Struct.* **519**, 243–256.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

O'Hara, B. P., Hemmings, A. M., Buttle, D. J. & Pearl, L. H. (1995). *Biochemistry*, **34**, 13190–13195.

Oliver-Salvador, M. C. (1999). PhD thesis. Escuela Nacional de Ciencias Biológicas, IPN, Mexico.

Oliver-Salvador, M. C., González-Rámirez, L. A., Gavira, J. A., Soriano-García, M. & García-Ruiz, J. M. (2004). *Acta Cryst.* D**60**, 2058–2060.

Ortega, M. L. & del Castillo, L. M. (1966). *Ciencia (Mexico City)*, **24**, 247–251.

Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. (1995). *Protein Sci.* **4**, 2411–2423.

Palma, J. M., Sandalio, L. M., Corpas, F. J., Romero-Puertas, M. C., McCarthy, I. & del Rio, L. A. (2002). *Plant Physiol. Biochem.* **40**, 521–530.

Pickersgill, R. W., Rizkallah, P., Harris, G. W. & Goodenough, P. W. (1991). *Acta Cryst.* B**47**, 766–771.

Priolo, N., del Valle, S. M., Arribere, M. C., Lopez, L. & Caffini, N. (2000). *J. Protein Chem.* **19**, 39–49.

Rawlings, N. D. & Barrett, A. J. (1994). *Methods Enzymol.* **244**, 461–486.

Schechter, I. & Berger, A. (1967). *Biochem. Biophys. Res. Commun.* **27**, 157–162.

Schüttelkopf, A. W. & van Aalten, D. M. F. (2004). *Acta Cryst.* D**60**, 1355–1363.

Schwede, T., Diemand, A., Guex, N. & Peitsch, M. C. (2000). *Res. Microbiol.* **151**, 107–112.

Silva, C. A., Gomes, M. T., Ferreira, R. S., Rodrigues, K. C., Val, C. G., Lopes, M. T., Mello, V. J. & Salas, C. E. (2003). *Planta Med.* **69**, 926–932.

Smith, P. K., Krohn, R. I., Hermanson, G. T., Mallia, A. K., Gartner, F. H., Provenzano, M. D., Fujimoto, E. K., Goeke, N. M., Olson, B. J. & Klenk, D. C. (1985). *Anal. Biochem.* **150**, 76–85.

Solomon, M., Belenghi, B., Delledonne, M., Menachem, E. & Levine, A. (1999). *Plant Cell*, **11**, 431–443.

Soriano, M., Cruz, M. T., Bustamante, Y., del Castillo, L. M. & Castañeda-Agulló, M. (1975). *Rev. Latinoamer. Quím.* **6**, 143–151.

Starley, I. F., Mohammed, P., Schneider, G. & Bickler, S. W. (1999). *Burns*, **25**, 636–639.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). *Nucleic Acids Res.* **22**, 4673–4680.

Turk, B., Turk, V. & Turk, D. (1997). *Biol. Chem.* **378**, 141–150.

Varughese, K. I., Ahmed, F. R., Carey, P. R., Hasnain, S., Huber, C. P. & Storer, A. C. (1989). *Biochemistry*, **28**, 1330–1332.

Varughese, K. I., Su, Y., Cromwell, D., Hasnain, S. & Xuong, N.-H. (1992). *Biochemistry*, **31**, 5172–5176.

Westbrook, E. M. (1985). *Methods Enzymol.* **114**, 187–196.

Wisniewski, K. & Zagdanska, B. (2001). *J. Exp. Bot.* **52**, 1455–1463.

Zhao, B., Janson, C. A., Amegadzie, B. Y., D'Alessio, K., Griffin, C., Hanning, C. R., Jones, C., Kurdyla, J., McQueney, M., Qiu, X., Smith, W. W. & Abdel-Meguid, S. S. (1997). *Nature Struct. Biol.* **4**, 109–111.