Structural Basis of the Unusual Stability and Substrate Specificity of Ervatamin C, a Plant Cysteine Protease from *Ervatamia coronaria*‡

Piyali Guha Thakurta,§ Sampa Biswas,§ Chandana Chakrabarti,§ Monica Sundd,||,⊥ Medicherla V. Jagannadham,|| and Jiban K. Dattagupta*,§

Crystallography and Molecular Biology Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Kolkata-700 064, India, and Molecular Biology Unit, Institute of Medical Sciences, Banaras Hindu University, Varanasi-221 005, India

Received October 1, 2003; Revised Manuscript Received December 9, 2003

ABSTRACT: Ervatamin C is an unusually stable cysteine protease from the medicinal plant *Ervatamia coronaria* belonging to the papain family. Though it cleaves denatured natural proteins with high specific activity, its activity toward some small synthetic substrates is found to be insignificant. The three-dimensional structure and amino acid sequence of the protein have been determined from X-ray diffraction data at 1.9 Å (R = 17.7% and R_{free} = 19.0%). The overall structure of ervatamin C is similar to those of other homologous cysteine proteases of the family, folding into two distinct left and right domains separated by an active site cleft. However, substitution of a few amino acid residues, which are conserved in the other members of the family, has been observed in both the domains and also at the region of the interdomain cleft. Consequently, the number of intra- and interdomain hydrogen-bonding interactions is enhanced in the structure of ervatamin C. Moreover, a unique disulfide bond has been identified in the right domain of the structure, in addition to the three conserved disulfide bridges present in the papain family. All these factors contribute to an increase in the stability of ervatamin C. In this enzyme, the nature of the S2 subsite, which is the primary determinant of specificity of these proteases, is similar to that of papain, but at the S3 subsite, Ala67 replaces an aromatic residue, and has the effect of eliminating sufficient hydrophobic interactions required for S3−P3 stabilization. This provides the possible explanation for the lower activity of ervatamin C toward the small substrate/inhibitor. This substitution, however, does not affect the binding of denatured natural protein substrates to the enzyme significantly, as there exist a number of additional interactions at the enzyme−substrate interface outside the active site cleft.

Cysteine proteases of the papain family are widely distributed in nature. They are found in both prokaryotes and eukaryotes, e.g., bacteria, parasites, plants, invertebrates, and vertebrates (*1*). Enzymatic activity of these proteases is related to a catalytic dyad formed by a cysteine and a histidine residue which exists as an ion pair, $S^{-}\cdots H^{+}I_{m}^{-}$, in the pH interval of 3.5−8.0 (*1*, *2*). All members of the papain family share a common fold. A papain-like fold consists of two distinct domains, the left (L) comprising mainly the N-terminal half of the molecule and the right (R) comprising mainly the C-terminal half of the molecule, separated by a V-shaped active site cleft (*2*). Cys25 and His159 (papain numbering) from the L- and R-domains, respectively, form the catalytic site of the enzyme in the middle of the active site cleft. The L-domain is predominantly α-helical, and the R-domain has mainly an antiparallel β-sheet structure. Both the amino and carboxyl terminal ends of the polypeptide chain cross over to the other domain, acting as clamps holding the two domains tightly together.

The papain-like mammalian lysosomal cysteine proteases (11 cathepsins known so far) (*3*−*5*) are reported to be responsible for protein degradation (*6*), and they have also been implicated in the development and progression of many diseases that involve abnormal protein turnover (*6*−*9*). The plant cysteine proteases in this family, being more easily available, serve as useful models to study the enzyme−substrate and enzyme−inhibitor interactions.

Ervatamia coronaria, a flowering plant indigenous to India, has a wide range of medicinally important applications (*10*). Three cysteine proteases with novel properties have been isolated from the latex of this plant and designated as ervatamin A (*11*), ervatamin B (*12*), and ervatamin C (*13*). The sequence of 21 N-terminal amino acid residues of the ervatamins showed marked similarity to those of known cysteine proteases of the papain family. Studies of enzymatic and physicochemical properties of the ervatamins revealed that the activity of the enzymes toward natural substrates, their inhibition by thiol-specific inhibitors, and their autocatalytic property are similar to those of papain and other members of this family. Ervatamins, however, exhibit some striking properties that are distinctly different from those of papain and others in the family, and these enzymes also differ among themselves in many respects (*11*−*14*).

‡ The coordinates of the refined structure have been deposited at the RCSB Protein Data Bank with entry code 1O0E, and the protein sequence data reported in this paper will appear in the Swiss-Prot and TrEMBL knowledgebase under the accession number P83654.

* To whom correspondence should be addressed. E-mail: jiban@cmb2.saha.ernet.in.

§ Saha Institute of Nuclear Physics.

|| Banaras Hindu University.

⊥ Present address: Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA.

10.1021/bi0357659 CCC: $27.50
Published on Web 01/21/2004

Table 1: Summary of Diffraction Data Collection and Refinement Statistics[a]

A. Crystal Data			
space group	$P2_12_12_1$	no. of observed reflns	157915
unit cell params (Å)	$a = 43.73$, $b = 82.69$, $c = 133.05$	no. of unique reflns	38978
no. of molecules/asymmetric unit	2	completeness (%)	97.9 (99.8)
resolution (Å)	1.9 (1.95−1.90)	$\alpha\varpi$ I/σ(I)	20.15 (2.9)
V_M (Å^3 Da^{-1})	2.41	R_{merge}[b] (%)	6.0 (56.0)
solvent content (%)	50		
B. Refinement Statistics			
resolution range (Å)	15−1.9	rms dev from target values	
no. of reflections in working set	36113	bond lengths (Å)	0.006
no. of reflections in test set	1901	bond angles (deg)	1.3
no. of protein atoms	3151	dihedral angles (deg)	24.0
no. of solvent molecules	256	improper angles (deg)	0.79
no. of ligand (thiosulfate) atoms	10	Ramachandran statistics	
R factor[c] (%)	17.7	most favored regions (%)	86.4
R_{free}[c] (%)	19.0	additionally allowed regions (%)	13.3
		generously allowed regions (%)	0.3
		disallowed regions (%)	0.0
		average B factor (Å^2)	26.8

[a] Values in parentheses are for the outer resolution shell, 1.95−1.90 Å. [b] $R_{merge} = \sum|Ih - \langle Ih\rangle|/\sum(Ih)$, where $\langle Ih\rangle$ is the average intensity of reflection h and symmetry-related reflections. [c] $R = R_{free} = \sum||F_o| - |F_c||/\sum|F_o|$ calculated for reflections of the working set and test (5%) sets, respectively.

Among the ervatamins, ervatamin C shows (*13*) remarkable stability under conditions known to denature most proteins. It retains both secondary and tertiary structures along with biological activity over a wide range of pH (*2*−*12*), at high temperatures (up to 70 °C), and at high concentration of chemical denaturants. In addition, ervatamin C hydrolyzes natural protein substrates with high specific activity, whereas it exhibits very low activity against some small synthetic substrates. Moreover, the enzyme is not fully inhibited by leupeptin, a potent inhibitor of the papain family of cysteine proteases. Toward understanding the basis of such novel and significant observations, a detailed structural study of ervatamin C and a comparison with other ervatamins as well as with other members of the papain family are crucial. These studies will also be of interest from the point of view of comparative biochemistry and the evolutionary relationship of the plant cysteine proteases.

MATERIALS AND METHODS

Crystallization, Data Collection, and Structure Solution. All crystallization experiments were performed by the hanging drop vapor diffusion method. The protein, used for crystallization, was purified (*13*) in the presence of sodium tetrathionate and concentrated to 13 mg/mL in 0.01 M sodium phosphate buffer, pH 7.0. Needle-shaped crystals of ervatamin C were obtained at room temperature with 0.05 M potassium dihydrogen orthophosphate and 20% (w/v) PEG-8000. Crystals were characterized, and preliminary X-ray studies were reported (*15*). Subsequently, diffraction data were collected from a better crystal, grown under the same conditions, on a 30 cm MAR Research image plate detector mounted on a Rigaku RU-200 rotating anode generator running at 50 kV and 90 mA, using Cu Kα radiation. X-ray intensity data were indexed, integrated, and subsequently scaled using DENZO and SCALEPACK from the HKL program package (*16*) up to 1.9 Å resolution. A self-rotation function calculation using the program POLARRFN in the CCP4 package (*17*) clearly indicated the presence of two molecules in the asymmetric unit, related by a noncrystallographic 2-fold axis, distinct from the crystallographic 2-fold axis. Crystal data parameters, together with the solvent content and Matthews coefficient (*18*), are given in Table 1. The structure of ervatamin C was solved by the molecular replacement method with the program AMoRe implemented in the CCP4 suite (*17*) using the polyalanine coordinates of ervatamin B (PDB code 1IWD) as the starting model. A rotation search was carried out taking data between 10 and 4.0 Å followed by a translation search with the highest rotation peak to identify the position of the first monomer. Once the position of the first monomer was fixed, a translation search with the second highest rotation peak was performed to locate the second monomer. Rigid body refinement with AMoRe yielded an R factor of 41.2% and a correlation factor of 62.6%.

Model Building, Screening, and Refinement. The model obtained from molecular replacement was subjected to rigid body refinement in CNS (*19*) with each monomer in the asymmetric unit as an independent rigid group in the resolution range 15.0−3.0 Å, which resulted in R and R_{free} values of 46.2% and 43.3%, respectively. A random sample of 5% reflections in the data set was excluded from the refinement and used for the R_{free} calculation. Manual fitting of the polyalanine model to the $2F_o - F_c$ electron density map was performed using the interactive graphics program O (*20*). The model including two molecules in the asymmetric unit was then refined using strict noncrystallographic symmetry (NCS)[1] in CNS (*19*) during the first two rounds of positional refinement, which led to an R factor of 39.5% and R_{free} of 39.8%. The amino acid sequence of the first 21 residues was known to us, and good electron density developed for some of them. These residues, along with the conserved cysteines, were inserted into the model. Because the full amino acid sequence was unknown to us, refinement was carried out very cautiously and during each step $2F_o - F_c$ and $F_o - F_c$ maps were calculated and carefully checked to avoid incorrect identification of amino acid residues. Specific attention was paid to the residues implicated in the

[1] Abbreviations: CD, circular dichroism; BAPA, benzoylarginine-*p*-nitroanilide; MALDI-TOF, matrix-assisted laser desorption time of flight; NCS, noncrystallographic symmetry; GPII, ginger protease II.

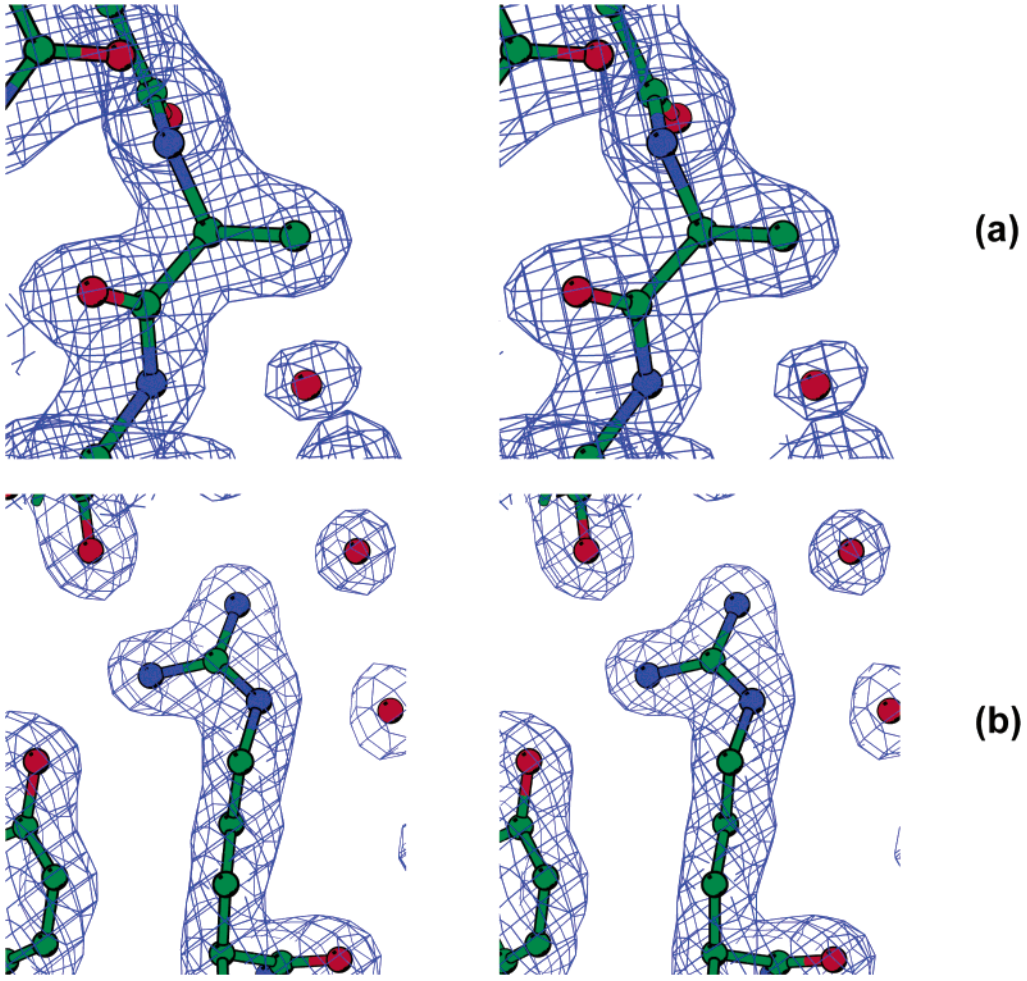

FIGURE 1: Stereoviews of the (a) composite omit map contoured at 1.0σ for substituted Ala67 in the S2 subsite and (b) $2F_o - F_c$ map contoured at 1.5σ of a representative residue, Arg172. The figure was prepared by BOBSCRIPT (*44*).

functional role of the protein, and these were reconfirmed from simulated annealing omit maps. Map calculation, fitting, identifying new residues, and refinement with the improved model were performed several times, which brought the *R* factor down to 28.9% and R_{free} down to 29.8%. From a simulated annealing omit map, loop regions varying in length compared to the other members of the family were identified. Further positional refinement yielded an *R* factor of 26.3% (R_{free} = 27.8%). At this stage 145 water molecules were incorporated into the model conservatively, i.e., only if they appeared as discrete spherical peaks in both $2F_o - F_c$ (1.0σ) and $F_o - F_c$ (3.5σ) electron density maps, satisfying the hydrogen-bonding criteria as incorporated in CNS (*19*). Almost 80% of the residues could be identified at this stage. The strategy of amino acid sequence determination is discussed later in detail. After a few cycles of positional and subsequent B-group refinement alternating with model refitting into the map and successive water molecule incorporation into the model, *R* and R_{free} values decreased to 19.2% and 20.7%, respectively. In the $F_o - F_c$ map (3σ) at this stage, positive electron density beyond Cys25 SG was observed in both the molecules in the asymmetric unit. Since the enzyme, used for crystallization, was inactivated by sodium tetrathionate throughout the purification procedure (and the enzyme activity could be restored by incubation of the enzyme with β-mercaptoethanol) (*13*), it may be expected that a thiosulfate is covalently attached to Cys25 as a protecting group. Such an adduct was also reported in the case of ervatamin B (*14*) and ginger protease II (GPII) (*21*), two other cysteine proteases of the same family. Accordingly, a thiosulfate adduct was modeled taking coordinates from the HIC-Up library (*22*), which could be fitted well into the $F_o - F_c$ map (3σ) near Cys25 (thiosulfate S---Cys25 SG $\cong$ 2 Å) of each of the two molecules. In subsequent positional, B-group, and B-individual refinement the terminal S atom of this thiosulfate adduct was covalently patched to the Cys25 atom, which lowered the *R* factor to 18.1% (R_{free} = 19.6%). Last, two alternate conformations were modeled for residues Thr30, Thr129, and Arg177 of chain A and Ser29, Thr129, and Arg177 of chain B and also for the thiosulfate molecule, and their occupancies were refined. The final structure with two chains, each comprising 208 residues, 256 water molecules, and the fitted model of thiosulfate converged to an *R* factor of 17.7% and R_{free} of 19.0%. A composite omit map agreed well with the final model. Some representative sections of the final electron density maps are shown in Figure 1. The only residue that could not be identified was 117 in both chains A and B. The Ramchandran analysis showed that 86.4% of the residues lie in the most favored regions with no residue in the disallowed regions. Table 1 summarizes refinement statistics and quality of the model. The inter/intramolecular contacts and H bonds were calculated using the program CONTACT implemented in the CCP4 suite of programs (*17*). For the H bonds, the distance

criterion for the donor−acceptor atom was kept as ≤3.3 Å and the bond angle formed by the donor atom, acceptor atom, and the atom attached to the acceptor atom between 90° and 120°.

Determination of the Amino Acid Sequence. Because C, O, and N atoms have comparable atomic form factors, discrimination of Asp/Asn/Leu, Glu/Gln, and Val/Thr is not unequivocal even at this resolution and with such a good quality electron density map. The ambiguities concerning Asn/Asp/Leu, Gln/Glu, and Val/Thr were resolved by their potentiality to form strong hydrogen bonds with neighboring polar atoms, by their local environment (e.g., polar side chains should have a polar environment, whereas hydrophobic residues should be buried in a hydrophobic core/cleft), and by sequence homology. However, nonconserved amidic/acidic surface residues which formed no hydrogen bonds with neighboring polar atoms and which could not be distinguished on the basis of the above criteria were kept as amides to match the experimentally determined pI value of 9.5 for the protein.

Among the ten leucine residues in the structure, five were conserved including the one at the N-terminus. Of the remaining five, four were totally buried and one, situated near the surface, had its side chain facing toward a hydrophobic pocket.

Out of 26 residues identified as Val or Thr, 11 were conserved in the family. The discrimination of a Val/Thr was based on the criteria that Thr OG1 satisfied hydrogen-bonding parameters and valines resided in a hydrophobic environment. As is also known (*23*), individual *B* factors sometimes help in identification of Val or Thr; generally they (C or O atoms) are balanced in correctly built side chains. Except three valines with *B* factors ranging from 30 to 35 Å^2, the terminal atoms of all the other nonconserved valines and threonines possessed an average *B* factor of 21 Å^2.

Of the 38 acidic and amidic residues (Asp/Asn and Glu/Gln), 16 were conserved in the family. Out of the remaining 22, 15 were discriminated as amidic or acidic on the basis of their hydrogen-bonding environment; i.e., each polar atom of all these nonconserved residues formed at least one H bond with a neighboring protein atom. In addition to that, 3 residues were among the known 21 residues at the N-terminus. The remaining four residues, making no hydrogen bonds with protein atoms, could be either amidic or acidic, and these were kept as amides to match the experimentally determined pI value of the protein.

Circular Dichroism (CD) Spectroscopy. CD spectra of the protein were measured by a JASCO (model J720) CD spectrophotometer. The protein solution (0.1 mg/mL in 10 mM sodium phosphate buffer, pH 7.5) was scanned from 190 to 250 nm in 0.1 nm increments. The sample was heated from 20 to 90 °C in 10 °C increments, with a 10 min incubation time. The unfolding of the protein was monitored by the change in ellipticity in this range, as the sample was heat denatured.

Molecular Modeling. To analyze the binding of a specific substrate/inhibitor to ervatamin C in detail, we docked leupeptin, a known inhibitor for both ervatamin B and ervatamin C, at their respective active site clefts in a manner similar to that of its binding at the active site of papain in the crystal structure of the papain−leupeptin complex (PDB code 1POP). For molecular modeling and subsequent molecular mechanics calculations, BUILDER and DISCOVER-3 modules of the InsightII software package (MSI Corp.) with a consistent valence force field (cvff) were used. First, hydrogens were generated for both the complex structures, and their positions were optimized. During minimization and simulations, a distance-dependent dielectric constant of 1.0 was used. For each of the two complex structures, a subset was defined consisting of residues within a 5 Å radius sphere of each atom of the inhibitor and designated as the first subset. The rest of the molecule was defined as the second subset. The first subset was then minimized, keeping the second one fixed. In all subsequent calculations, the second subset was kept fixed. After minimization, the first subset was solvated by a 10 Å layer of water molecules using the SOAK option of InsightII. Positions of water molecules were minimized separately. Final minimization was carried out with the first subset along with the water molecules as an assembly using the conjugate gradient minimization method and continued until the maximum derivative dropped below 0.01 (kcal/mol)/Å. This minimized coordinate set of the whole system was then used for molecular dynamics simulation. At the beginning, each system was equilibrated for 120 ps at 300 K. Molecular dynamics simulation was continued for another 80 ps for trajectory analysis. Conformational analysis of these complexes indicated a stable conformation of leupeptin at the active site clefts of ervatamin C and ervatamin B over the last 80 ps of the trajectory. Hence, the average structure over the last 10 ps of each simulation could be used as a representative structure to find probable contacts between the enzyme and the inhibitor.

To investigate the nature of binding of ervatamin C to protein substrates/inhibitors, we docked stefin B, a cysteine protease inhibitor from the cystatin family, at the active site of ervatamin C, guided by the criteria seen in the papain−stefin B complex (PDB code 1STF). The interfaces of the initially docked complex of ervatamin C−stefin B were then optimized using the program MULTIDOCK (*24*). An analysis of the intermolecular contacts (within 4 Å) between the enzyme and inhibitor of the two complex structures (papain−stefin B and ervatamin C−stefin B) was performed using the program CONTACT in the CCP4 suite (*17*).

RESULTS AND DISCUSSION

Overall Organization. The polypeptide chain of ervatamin C has a papain-like fold (Figure 2). The L-domain consists of residues 14−106 and 206−208 and the R-domain of residues 1−9 and 111−202. Three portions of the polypeptide chain (residues 10−13, 107−110, and 203−205) act as straps to clamp the two domains together. Cys25 from the L-domain and His157 from the R-domain, situated on either side of the interdomain cleft, form the catalytic dyad of the enzyme. In addition to the three conserved disulfide bridges (between residues 22 and 63, 56 and 96, and 151 and 196), a fourth disulfide bridge is formed between residues 114 and 193, which is unique in ervatamin C. The structure of ervatamin C has been superposed on those of other plant cysteine proteases such as papain, actinidin, GPII, and ervatamin B (Figure 2) by the LSQ-EXPLICIT option in O (*20*), and the low rmsd values of Cα atoms (<1.0 Å) indicate close resemblance of the overall conformation. Deviations are mainly observed at three surface loop regions of ervatamin C, which are designated as regions 1−3 in Figure

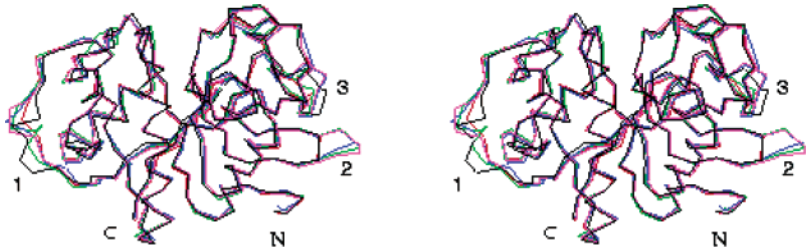

FIGURE 2: Stereoview of the superposition of C_α traces of ervatamin C (red), papain (black), actinidin (green), ervatamin B (magenta), and ginger protease II (blue). The loop regions with deviations are indicated: 1 (residues 97−98), 2 (residues 163−164 and 167−168), 3 (residues 195−196).

```
1CQD_B      ----DDLPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLISLSEQQLV 56
2ACT        ------LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELI 54
1IWD        ------LPSFVDWRSKGAVNSIKNQKQCGSCWAFSAVAAVESINKIRTGQLISLSEQELV 54
ERVC        ------LPEQIDWRKKGAVTPVKNQGSCGSCWAFSTVSTVESINQIRTGNLISLSEQELV 54
1PPO        ------LPENVDWRKKGAVTPVRHQGSCGSCWAFSAVATVEGINKIRTGKLVELSEQELV 54
1PCI-A      NEDIVNLPENVDWRKKGAVTPVRHQGSCGSCWAFSAVATVEGINKIRTGKLVELSEQELV 60
1GEC-E      ------LPESVDWRAKGAVTPVKHQGYCESCWAFSTVATVEGINKIKTGNLVELSEQELV 54
1PPN        ------IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNEYSEQELL 54
1YAL        ------YPQSIDWRAKGAVTPVKNQGACGSXWAFSTIATVEGINKIVTGNLLELSEQELV 54
                  *. :***  ***  :: *  * . ****:: ::*.* :* :*.* . ***:*:

1CQD_B      DC--TTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTVN-APVVSIDSYE 113
2ACT        DCGRTQNTRGCDGGYITDGFQFIINDGGINTEENYPYTAQDGDCDVALQDQKYVTIDTYE 114
1IWD        DC--DTASHGCDGGWMDDAFQYIIANGGIDTQSAYPYSAVQGAC--KPYRARVVSIDGFE 110
ERVC        DC--DKKNHGCLGGAFVFAYQYIINNGGIDTQANYPYKAVQGPC--QAA-SKVVSIDGYN 109
1PPO        DC--ERRSHGCKGGYPPYALEYVAKNG-IHLRSKYPYKAKQGTCRAKQVGGPIVKTSGVG 111
1PCI-A      DC--ERRSHGCKGGYPPYALEYVAKNG-IHLRSKYPYKAKQGTCRAKQVGGPIVKTSGVG 117
1GEC-E      DC--DLQSYGCNRGYQSTSLQYVAQNG-IHLRAKYPYIAKQQTCRANQVGGPKVKTNGVG 111
1PPN        DC--DRRSYGCNGGYPWSALQLVAQYG-IHYRNTYPYEGVQRYCRSREKGPYAAKTDGVR 111
1YAL        DC--DKHSYGCKGGYQTTSLQYVANNG-VHTSKVYPYQAKQYKCRATDKPGPKVKITGYK 111
            **     . **  *    . : :   * :.    *** . :  *          ..

1CQD_B      NVPSHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFTGSCNISANHALTVVGYGTEND 173
2ACT        NVPYNNEWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGG 174
1IWD        RVTRNNESALESAVASQPVSVTVEAAGAPFEHYSSGIFTGPCGTAENHGVVIVGYGTQAG 170
ERVC        GVPFCNEXALKQAVAVQPSTVAIDASSAQFQQYSSGIFSGPCGTKLNHGVTIVGY--QA- 166
1PPO        RVQPNNEGNLLNAIAKQPVSVVVESKGRPFQLYKGGIFEGPCGTKVDHAVTAVGYGKSGG 171
1PCI-A      RVQPNNEGNLLNAIAKQPVSVVVESKGRPFQLYKGGIFEGPCGTKVDGAVTAVGYGKSGG 177
1GEC-E      RVQSNNEGSLLNAIAHQPVSVVVESAGRDFQNYKGGIFEGSCGTKVDHAVTAVGYGKSGG 171
1PPN        QVQPYNEGALLYSIANQPVSVVLEAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGPN-- 169
1YAL        RVPSNXETSFLGALANQPLSVLVEAGGKPFQLYKSGVFDGPCGTKLDHAVTAVGYGTSDG 171
             *    *  :  ::: ** :* ::: .  *: * .*:* *.*.   : .:. ***  .

1CQD_B      KDFWIVKNSWGKNWGESGYIRAERNIENPDGKCGITREASYPVKKGTN 221
2ACT        VDYWIVKNSWDTTWGEEGYMRILRNVGG-AGTCGIATMPSYPVKYNN- 220
1IWD        KNYWIVRNSWGQNWGNKGYIWMERNVASSAGLCGIAQLPSYPTKA--- 215
ERVC        -NYWIVRNSWGRYWGEKGYIRMLR--VGGCGLCGIARLPYYPTKA--- 208
1PPO        KGYILIKNSWGTAWGEKGYIRIKRAPGNSPGVCGLYKSSYYPTKN--- 216
1PCI-A      KGYILIKNSWGTAWGEKGYIRIKRAPGNSPGVCGLYKSSYYPTKN--- 222
1GEC-E      KGYILIKNSWGPGWGENGYIRIRRASGNSPGVCGVYRSSYYPIKN--- 216
1PPN        --YILIKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN--- 212
1YAL        KNYIIIKNSWGPNWGEKGYMRLKRQSGNSQGTCGVYKSSYYPFKGFA- 218
              : :::***.  **:.**:   *   .  * **:   . ** *
```

FIGURE 3: A multiple-sequence alignment of ervatamin C (ERVC) and other members of the papain family of plant cysteine proteases. Residues forming S2 and S3 subsites are shaded. The sequences listed (identified by PDB code) are of ginger protease II (1CQD_B), actinidin (2ACT), ervatamin B (1IWD), papaya protease ω (1PPO), procaricain (1PCI_A), glycyl endopeptidase (1GEC_E), papain (1PPN), and chymopapain (1YAL).

2. The deletion of amino acid residues in these three surface loops reduces the flexibility of the enzyme, which in turn makes it more compact compared to the others.

The Primary Structure. From the good quality electron density map (Figure 1), a total of 208 amino acid residues could be traced for both chains A and B, only the side chain of residue 117 of which could not be identified unambiguously. The molecular weight of ervatamin C calculated from 208 deduced amino acid residues is 22500 Da, which is in agreement with the molecular weight (23000 Da) determined by the matrix-assisted laser desorption time-of-flight (MALDI-TOF) mass spectrometry method. The theoretically calculated pI value from the amino acid sequence is 8.8, which is close to the experimentally determined pI value of 9.5. The primary structure of ervatamin C determined in this way was aligned by CLUSTALW (http://www.ebi.ac.uk/clustalw/) with the known amino acid sequences of eight other plant cysteine proteases of the papain family (Figure 3). The sequence identity of ervatamin C is 66% with ervatamin B, 50% with papain, 57% with actinidin, and 57% with GPII.

Structural Basis of the Stability. Unlike the other plant cysteine proteases of the papain family, ervatamin C is able to retain its activity over a wide pH range (2−12), at high temperatures (up to 70 °C), and at high concentration of chemical denaturants without any major changes in secondary and tertiary structures (*13*, *25*). All these features indicate a high rigidity of the protease. The CD spectrum of ervatamin C (190−250 nm) was monitored over a range of temperature in our laboratory, and the secondary structure of the protein was found to remain stable up to 80 °C.

The improved stability of a protein is due to a concerted effort of several local adjustments on the sequence and structure of the enzyme, and the origins of stability differences are subtle and come from many factors. A comparative study of thermophilic proteins with their mesophilic counterparts suggested that increased hydrogen bonding and salt bridge formation might provide the most general explanation for thermal stability of a protein (*26*−*31*). Though in ervatamin C the number of salt bridges are comparable to those in other plant cysteine proteases of the family, there are some natural substitutions of amino acid residues, conserved in the others, at both the left and the right domains and at the region of the interdomain cleft, which has resulted in an increase in the number of intra- and interdomain hydrogen-bonding interactions.

(i) Intradomain Interactions. Residue 39 in the left domain is a conserved lysine in the family, which is replaced by a glutamine in ervatamin C. The OE1 atom of Gln39 makes two hydrogen bonds with the backbone nitrogen atom of Ile40 and water molecule 167. This water molecule in turn forms a hydrogen bond with the main chain oxygen atom of Lys10. The NE2 atom of Gln39 is hydrogen bonded to the main chain oxygen of Gly43 and water molecule 30 via the water molecule 165. Other members of the family, having lysine at this position, are able to make only one hydrogen bond through their only polar NZ atom with the main chain oxygen of conserved Gly43. There are three substitutions of residues in the right domain of ervatamin C, 114, 128, and 136. Residue 114 is replaced by a cysteine in ervatamin C and forms a disulfide bond with the Cys193 of the same domain. The presence of this unique fourth disulfide bond in ervatamin C is one of the important contributing factors responsible for the higher stability of the protein. Residues 128 and 136 are conserved valine and glycine, respectively, in the family, and both of them are replaced by serine in ervatamin C. Ser128 OG forms a hydrogen bond with the main chain oxygen of Gln126, whereas Ser136 OG hydrogen bonds with Phe139 N, water molecule 140, Gln138 OE1, and the main chain oxygen of Pro150. Water molecule 140 also forms three main chain hydrogen bonds with residues Asp133, Ser135, and Ser136. Substitution of the valine and glycine residues by serine in ervatamin C, however, does not destabilize the hydrophobic core of the protein, as these residues are present near the surface.

(ii) Interdomain Interactions. Ser32 and Ser36 from the left domain and Arg172 from the right domain of ervatamin C are situated on opposite sides of the interdomain cleft, and they correspond to conserved glycine, glycine, and lysine, respectively, in the other members of the papain family of plant cysteine proteases (Figure 3). On the other hand, in ervatamin B, an enzyme from the same source, there are substitutions of residues at positions 36 and 177 (172 in ervatamin C) by serine and arginine and an increase in the stability of the protein has been proposed (*14*). The substitution of lysine by arginine needs special mention. A detailed study by Argos et al. (*32*) and Menendez and Argos (*33*) established that Lys → Arg is the most preferred substitution in thermophilic proteins. The Lys → Arg mutation is also consistent with the decrease in solvation energy of thermophiles because Arg has three polar side chain nitrogen atoms to contribute to the protein surface compared with only one for Lys. Mrabet et al. (*34*) studied engineered Lys → Arg variants in three different proteins and showed by actual examination of tertiary structures that enhanced hydrogen-bonding and electrostatic interactions are likely to be responsible for the observed increased stability in the Arg-containing mutants. In ervatamin C, Ser32, Ser36, and Arg172 together constitute an intricate network of hydrogen-bonding interactions between the two domains (Figure 4). They are involved in both protein−protein and water-mediated interdomain contacts. The NH2, NH1, and NE atoms of Arg172 interact directly with residues Thr14, Pro15, and Ser32 of the left domain. Apart from that, these three side chain nitrogen atoms of Arg172 also interact, through water molecules 1, 9, 14, 22, and 148, with residues of the left domain such as Lys17, Phe28, Ser32, Glu35, Ser36, and Glu50. The NH1 atom of Arg172, in addition, forms a hydrogen bond with Tyr184 of the same domain. Ser32 interacts with both Ser36 (through W148) of the left domain and Arg172 of the right domain and plays the role of a bridging residue in the interdomain interactions. This is illustrated in Figure 4. Other cysteine proteases having glycine, glycine, and lysine at respective positions on opposite sides of the interdomain cleft cannot have such interactions.

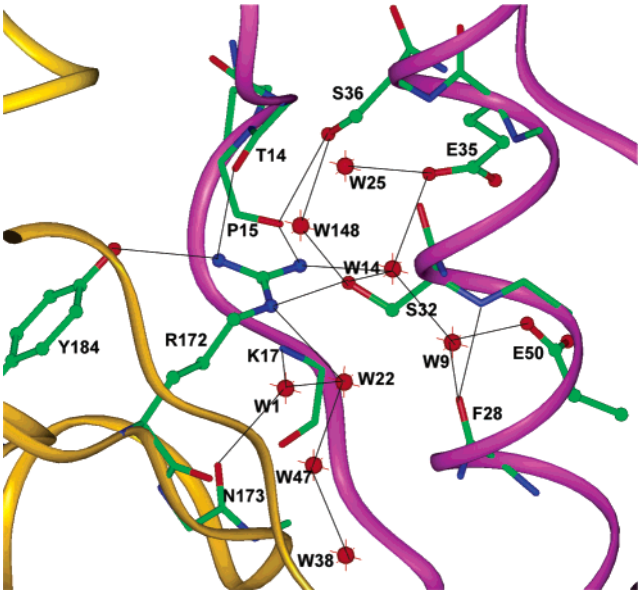

FIGURE 4: Interdomain interactions in ervatamin C. Left and right domains are indicated by purple and yellow ribbons, respectively. Sticks represent the main chain atoms, and side chain atoms are shown as ball-and-stick models. The figure was generated by InsightII (MSI, Inc.).

(iii) Additional Factors. Increased compactness would lead to an increase in van der Waals interactions and higher protein stability (*35*). A possible measure of compactness is the reduction in the number and volume of cavities within the protein (*36*). Calculation of the total gap volume (*37*) of some plant cysteine proteases of the papain family using the program SURFNET (*37*) shows (Table 2) that ervatamin C possesses the lowest gap volume and is much more compact with respect to the others. We also compared the amino acid composition of these proteases by using the Protparameter tool of the Expasy server (http://us.expasy.org/tools/protparam.html) to analyze primary sequence parameters such as the aliphatic index (*38*), the instability index (*39*), etc. which are related to the stability of a protein. Table 2

Table 2: Comparison of Some Factors Affecting the Stability of Plant Cysteine Proteases of the Papain Family

protease	PDB code	no. of residues	total vol (Å^3)	gap vol (Å^3)	instability index	aliphatic index	no. of S−S bridges
actinidin	2ACT	218	19781.32	3822.09	27.42	74.91	3
papain	1PPN	212	19932.28	4264.96	26.02	73.54	3
ginger protease II	1CQD	218 (one chain)	19565.38	4317.68	32.2	68.78	3
chymopapain	1YAL	218	20113.94	4943.36	18.0	63.49	3
ervatamin B	1IWD	215	19477.87	3988.47	28.33	71.21	3
ervatamin C	1O0E	208 (one chain)	18971.27	3462.13	16.67	76.87	4

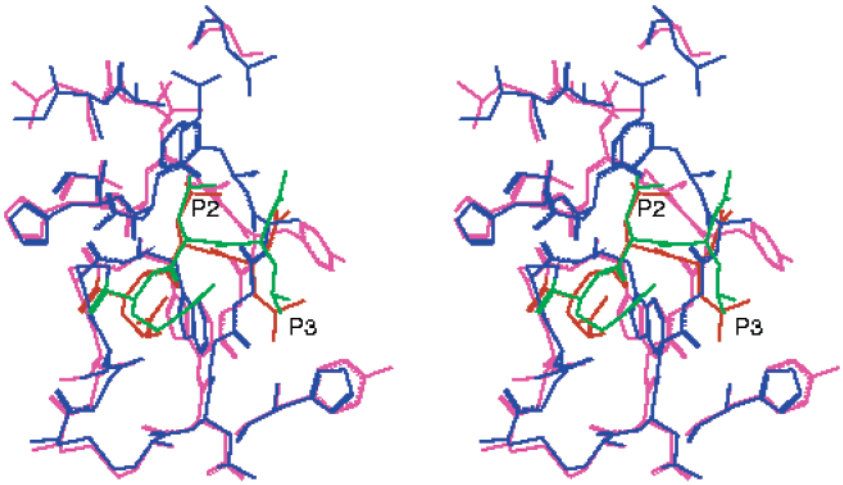

FIGURE 5: Superposition of the last 10 ps average structure of ervatamin C−leupeptin (blue−green) complex and crystal structure of papain−leupeptin (magenta−orange) complex. P2 and P3 positions of the leupeptin molecule are indicated.

summarizes the parameters and shows that the aliphatic index of ervatamin C is the highest and it also possesses the lowest instability index, both of which indicate higher stability for the protein.

Thus, it can be concluded that ervatamin C owes its stability to a repertoire of factors such as an increase in intra- and interdomain hydrogen-bonding interactions, increased compactness, and the presence of a unique fourth disulfide bond.

Substrate Specificity. Biochemical studies on ervatamin C have shown (*13*) that it has insignificant activity toward some small synthetic substrates and the enzyme is not fully inhibited by leupeptin, retaining 25% of its enzymatic activity. However, the enzyme hydrolyses denatured natural protein substrates such as casein, hemoglobin, azoalbumin, and azocasein with high specific activity.

Even though the lysosomal cysteine proteases of the papain family utilize both unprimed and primed subsites to bind a substrate/inhibitor (*40*), the specificity and activity of the plant cysteine proteases of this family are mainly determined by Sn–Pn interactions with $n \leq 3$ (*41*). The structure of ervatamin C shows that its S1 pocket resembles that of papain and its S2 pocket, comprising residues Ala67, Phe68, Ala131, Leu155, and Leu201, is a hydrophobic one also similar to that of papain. The S3 subsite in papain too is a hydrophobic region at the middle of the left wall of the interdomain cleft formed by Tyr61 and Tyr67. The residue at the equivalent position 67, contributing to both the S2 and S3 subsites, is in general an aromatic residue in the plant cysteine proteases (Figure 3), and together with residue 61, it provides necessary hydrophobic interactions to bind a substrate/inhibitor having a hydrophobic residue at the P3 position. By analogy with papain the S3 subsite of ervatamin C, formed by His 61 and Ala67, can also be assigned to a similar region of the cleft. But it is to be noted that Tyr67Ala is an important substitution at the active site cleft of ervatamin C, since it alters the nature of its S3 subsite and hence the S3 specificity.

The binding mode of a substrate or a substrate analogue inhibitor to ervatamin C has been analyzed from the modeled structure of ervatamin C−leupeptin complex and compared with our model of ervatamin B−leupeptin complex and the crystal structure of the papain−leupeptin complex (PDB code 1POP). It is seen that residues such as Phe68, Ala131, Leu155, and Leu201 at the S2 pocket of ervatamin C provide sufficient hydrophobic contacts for the leucine at the P2 position of the leupeptin molecule. The gap volume (191 Å^3), calculated by SURFNET (*37*), beyond the Leu side chain of leupeptin at the S2 groove of ervatamin C is comparable to that of papain (218 Å^3). Practically no such gap beyond leucine is observed in the ervatamin B−leupeptin complex. So the S2 subsite pockets of ervatamin C and papain not only are similar in nature but also have enough space to bind bulkier hydrophobic side chains such as phenylalanine at the P2 position of the inhibitor. This is in accordance with the biochemical observation that ervatamin C (large S2 pocket) has no significant activity toward the small synthetic substrates having a residue with a small side chain such as Ala at the P2 position (*13*). This also explains another biochemical observation that ervatamin C can hydrolyze benzoylarginine-*p*-nitroanilide (BAPA) (*13*), a potent substrate of papain, with a Phe at the P2 position, whereas ervatamin B (small S2 pocket) is inert against BAPA (*12*). It has further been found that the leucine residue at the P3 position of the inhibitor rests against the two phenyl rings

of Tyr61 and Tyr67 in papain and the imidazole ring of His61 and the indole ring of Trp67 in ervatamin B and is stabilized by a number of hydrophobic interactions involving both residues of the corresponding enzymes. But, as mentioned before, in ervatamin C the hydrophobic nature of the S3 subsite is altered due to substitution at residue 67 by alanine. Leucine at the P3 position of the inhibitor therefore interacts with His61, which is the only available residue in the vicinity to make hydrophobic contacts (Figure 5), resulting in weaker S3−P3 interactions compared to those of ervatamin B and papain. This corroborates the observation that ervatamin C is partially inhibited by leupeptin (*13*) compared to ervatamin B (*12*) and papain (*42*).

As is evident from the papain−stefin B complex structure (*43*), the binding of a protein substrate/inhibitor to cysteine proteases is mediated by a number of interactions, in addition to the S*n*−P*n* interactions at the active site cleft. Our docking studies of stefin B at ervatamin C also suggest that the two hairpin loops of the inhibitor make extensive interactions with the surface residues near the active site cleft of the enzyme while the N-terminal trunk residues (P*n*) block the S*n* subsites in a manner similar to that of papain. Though the number of S*n*−P*n* contacts in ervatamin C is found to be less compared to that of papain due to the Tyr67Ala substitution at the S3 subsite, it is compensated by the interactions between the enzyme and the two hairpin loops of the inhibitor. So the Ala67 substitution in ervatamin C does not affect the hydrolysis of denatured protein substrates significantly.

ACKNOWLEDGMENT

We acknowledge Dr. D. Mukhopadhyay and Dr. K. I. Varughese of Scripps Research Institute, San Diego, for the MALDI-TOF MS measurements. We also thank A. Bhattacharya of the Crystallography and Molecular Biology Division, Saha Institute of Nuclear Physics, for help in preparing the manuscript.

REFERENCES

1. Berti, P. J., and Storer, A. C. (1995) Alignment/phylogeny of the papain superfamily of cysteine proteases, *J. Mol. Biol. 246*, 273−283.
2. Turk, B., Turk, V., and Turk, D. (1997) Structural and functional aspects of papain-like cysteine proteases and their protein inhibitors, *Biol. Chem. 378*, 141−150.
3. Turk, B., Turk, D., and Turk, V. (2000) Lysosomal cysteine proteases: more than scavengers, *Biochim. Biophys. Acta. 1477*, 98−111.
4. Turk, V., Turk, B., and Turk, D. (2001) Lysosomal cysteine proteases: facts and opportunities, *EMBO J. 20*, 4629−4633.
5. Turk, D., Janjic, V., Stern, I., Podobnik, M., Lamba, D., Dahl, S. W., Lauritzen, C., Pedersen, J., Turk, V., and Turk, B. (2001) Structure of human dipeptidyl peptidase I (Cathepsin C): exclusion domain added to an endopeptidase framework creates the machine for activation of granular serine proteases, *EMBO J. 20*, 6570−6582.
6. Kirschke, H., Barrett, A. J., and Rawlings, N. D. (1995) in *Protein Profiles* (Sheterline, P., Ed.) Vol. 2, pp 1587−1643, Academic Press, London.
7. Afonso, S., Romagnano, L., and Babiarz, B. (1997) The expression and function of cystatin C and cathepsin B and cathepsin L during mouse embryo implantation and placentation, *Development 124*, 3415−3425.
8. Henskens, M. C., Veerman, E. C. I., and Amerongen, A. V. N. (1996) Cystatins in health and disease, *Biol. Chem. Hoppe-Seyler 377*, 71−86.
9. Grubb, A. (2000) Cystatin C-properties and use as diagnostic marker, *Adv. Clin. Chem. 35*, 63−99.
10. Anonymous (1952) in *The Wealth of India*, Vol. III, pp 192−193, CSIR, New Delhi.
11. Nallamsetty, S., Kundu, S., and Jagannadham, M. V. (2003) Purification and characterization of a highly active cysteine protease from the latex of *Ervatamia coronaria*, *J. Protein Chem. 22*, 1−13.
12. Kundu, S., Sundd, M., and Jagannadham, M. V. (2000) Purification and characterization of a stable cysteine protease Ervatamin B, with two disulfide bridges from the latex of *Ervatamia coronaria*, *J. Agric. Food Chem. 48*, 171−179.
13. Sundd, M., Kundu, S., Pal, G., and Jagannadham, M. V. (1998) Purification and characterization of a highly stable cysteine protease from the latex of *Ervatamia coronaria*, *Biosci. Biotechnol. Biochem. 62*, 1947−1955.
14. Biswas, S., Chakrabarti, C., Kundu, S., Jagannadham, M. V., and Dattagupta, J. K. (2003) Proposed amino acid sequence and the 1.63Å X-ray crystal structure of a plant cysteine protease, Ervatamin B: some insights into the structural basis of its stability and substrate specificity, *Proteins: Struct., Funct., Genet. 51*, 489−497.
15. Chakrabarti, C., Biswas, S., Kundu, S., Sundd, M., Jagannadham, M. V., and Dattagupta, J. K. (1999) Crystallization and preliminary X-ray analysis of ervatamin B and C, two thiol proteases from *Ervatamia coronaria*, *Acta Crystallogr. D55*, 1074−1075.
16. Otwinoski, Z., and Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode, *Methods Enzymol. 276*, 307−326.
17. Collaborative Computational Project, Number 4 (1994) *Acta Crystallogr. D50*, 760−763.
18. Matthews, B. W. (1968) Solvent content of protein crystal, *J. Mol. Biol. 33*, 491−497.
19. Brünger, A. T., Adams, P. F., 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., Simenson, T., and Warren, G. L (1998) Crystallography and NMR system: A new software suite for macromolecular structure determination, *Acta Crystallogr. D54*, 905−921.
20. Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard, M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models, *Acta Crystallogr. A47*, 110−119.
21. Choi, K. H., Laursen, R. A., and Allen, K. N. (1999) The 2.1Å structure of a cysteine protease with proline specificity from ginger rhizome, *Zingiber officinale*, *Biochemistry 38*, 11624−11633.
22. Kleywegt, G. J., and Jones, T. A. (1998) Databases in protein crystallography, *Acta Crystallogr. D54*, 1119−1131.
23. Hilge, M., Perrakis, A., Abrahams, J. P., Winterhalter, K., Piontek, K., Gloor, S. M. (2001) Structure elucidation of β-mannanase; From the electron density map to the DNA sequence, *Acta Crystallogr. D57*, 37−43.
24. Jackson, R. M., Gabb, H. A., and Sternberg, M. J. E. (1998) Rapid refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem, *J. Mol. Biol. 276*, 265−285.
25. Kundu, S., Sundd, M., and Jagannadham V. Medicherla. (1999) Structural Characterization of a Highly Stable Cysteine Protease Ervatamin C, *Biochem. Biophys. Res. Commun. 264*, 635−642.
26. Yip, K. S. P. (1995) The structure of Pyrococcus furious glutamate dehydrogenase reveals a key-role for ion-pair networks in maintaining enzyme stability at extreme temperatures, *Structure 3*, 1147−1158.
27. Querol, E., Perez-Pons, J. A., and Mozo-Villarias, A. (1996) Analysis of protein conformational characteristics related to thermostability, *Protein Eng. 9*, 265−271.
28. Vogt, G., and Argos, P. (1997) Protein Thermal Stability: hydrogen bonds or internal packing, *Folding Des. 2*, 40−46.
29. Vogt, G., Woell, S., and Argos, P. (1997) Protein Thermal Stability, Hydrogen Bonds, and Ion Pairs, *J. Mol. Biol. 269*, 631−643.
30. Russell, R. J., Ferguson, J. M., Haugh, D. W., Danson, M. J., and Taylor, G. L. (1997) The crystal structure of citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at 1.9Å resolution, *Biochemistry 36*, 9983−9994.
31. Russell, R. J., Gerike, U., Danson, M. J., Hough, D. W., and Taylor, G. L. (1998) Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium, *Structure 6*, 351−361.

32. Argos, P., Rossmann, M., Grau, U., Zuber, H., Frank, G., and Tratschin, J. (1979) Thermal stability and protein structure, *Biochemistry 25*, 5698−5703.
33. Menendez-Arias, L., and Argos, P. (1989). Engineering protein thermal stability: sequence statistics point to residue substitutions in alpha-helices, *J. Mol. Biol. 206*, 397−406.
34. Mrabet, N. T., Van den Broeck, A., Vanden brande, I., Stanssens, P., Laroche, Y., Lambeir, A. M., Matthijssens, G., Jenkins, J., Chiadmi, M., and van Tilbeurgh, H. (1992) Arginine residues as stabilizing elements in proteins, *Biochemistry 31*, 2239−2253.
35. Spassov, V. Z., Karshikoff, A. D., and Landenstein, R. (1995) The optimization of protein−solvent interactions: thermostability and the role of hydrophobic and electrostatic interactions, *Protein Sci. 4*, 1516−1527.
36. Russel, R. J. M., Hough, D. W., Danson, M. J., and Garry, L. T. (1994) The crystal structure of citrate synthase from the thermophilic archaeon *Thermoplasma acidophilum*, *Structure 2*, 1157−1167.
37. Laskowski, R. A. (1995) SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions, *J. Mol. Graphics 13*, 323−330.
38. Ikai, A. (1980) Thermostability and aliphatic index of globular proteins, *J. Biochem. (Tokyo) 88*, 1895−1898.
39. Guruprasad, K., Reddy, B. V., and Pandit, M. W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence, *Protein Eng. 4*, 155−161.
40. Turk, D.; Gunčar, G. (2003) Lysosomal cysteine proteases (cathepsins): promising drug targets. *Acta Crystallogr. D59*, 203-213.
41. Matsumoto, K., Murata, M., Sumiya, S., Kitamura, K., and Ishida, T. (1994) Clarification of substrate specificity of papain by crystal analyses of complexes with covalent type inhibitors, *Biochim. Biophys. Acta 1208*, 268−276.
42. Schroder, E., Philips, C., Garman, E., Harlos, K., and Crawford, C. (1993) X-ray crystallographic structure of a papain-leupeptin complex, *FEBS Lett. 315*, 38−42.
43. Stubbs, M. T., Laber, B., Bode, W., Huber, R., Jerala, R., Lenarcic, B., and Turk, V. (1990) The refined 2.4Å crystal structure of recombinant human stefin B in complex with the cysteine proteinase papain: a novel type of proteinase inhibitor interaction, *EMBO J. 9*, 1939−1947.
44. Esnouf, R. M. (1997) An extensively modified version of Molscript that includes greatly enhanced coloring capabilities, *J. Mol. Graphics 15*, 132−134.

BI0357659