Modelling family 2 cystatins and their interaction with papain

Suman Kumar Nandy, Rajabrata Bhuyan, and Alpana Seal

Department of Biochemistry and Biophysics, University of Kalyani, Kalyani 741235, India; Bioinformatics Infrastructure Facility, University of Kalyani, Kalyani 741235, India

Communicated by Ramaswamy H. Sarma

(Received 24 January 2012; final version received 28 May 2012)

Cystatins are extensively studied cysteine protease inhibitors, found in wide range of organisms with highly conserved structural folds. S-type of cystatins is well known for their abundance in saliva, high selectivity and poorer activity towards host cysteine proteases in comparison to their immediate ancestor cystatin C. Despite more than 90% sequence similarity, the members of this group show highly dissimilar binding affinity towards papain. Cystatin M/E is a potent inhibitor of legumain and papain like cysteine proteases and recognized for its involvement in skin barrier formation and potential role as a tumor suppressor gene. However, the structures of these proteins and their complexes with papain or legumain are still unknown. In the present study, we have employed computational methods to get insight into the interactions between papain and cystatins. Three-dimensional structures of the cystatins are generated by homology modeling, refined with molecular dynamics simulation, validated through numerous web servers and finally complexed with papain using ZDOCK algorithm in Discovery Studio. A high degree of shape complementarity is observed within the complexes, stabilized by numerous hydrogen bonds (HB) and hydrophobic interactions. Using interaction energy, HB and solvent accessible surface area analyses, we have identified a series of key residues that may be involved in papain–cystatin interaction. Differential approaches of cystatins towards papain are also noticed which are possibly responsible for diverse inhibitory activity within the group. These findings will improve our understanding of fundamental inhibitory mechanisms of cystatin and provide clues for further research.

Keywords: family 2 cystatin; cysteine protease inhibitor; S-type cystatin; cystatin M/E; papain; homology modelling; molecular dynamics simulation; docking.

Introduction

Cystatin superfamily is a group of homologous (Barrett, Fritz et al., 1986; Müller-Esterl et al., 1985) proteins characterized by cystatin domains, mainly known for their traditional inhibitory activity towards papain (C1) and legumain (C13) family of cysteine proteases (Abrahamson, Alvarez-Fernandez, & Nathanson, 2003; Alvarez-Fernandez et al., 1999; Barrett, Rawlings et al., 1986; Turk, Stoka, & Turk, 2008). On account of sequence similarity, number of cystatin domains, disulphide bonds and physiological localization, cystatin superfamily can be subdivided in four groups, namely stefins (family 1), cystatins (family 2), kininogens (family 3) (Barrett, Fritz et al., 1986; Müller-Esterl et al., 1985) and family 4 cystatins i.e. cystatins devoid of cysteine protease inhibitor activity (e.g. histidine-rich glycoprotein (Brown & Dziegielewska, 1997), bovine fetuin-A and its human homologue, alpha 2HS glycoprotein (AHSG) (Brown & Dziegielewska, 1997; Brown, Saunders, Møllgård, & Dziegielewska, 1992). Cystatins are broadly expressed and present in a wide range of species, organisms and tissues, from mammals, birds, fish, insects, plants to bacteria, parasites and viruses. Through their reversible inhibitory activity, cystatins have emerged as important players in a multitude of physiological and patho-physiological settings that range from cell survival and proliferation, to differentiation, cell signalling and immunomodulation (Keppler, 2006).

Despite considerable differences in amino acid sequences, first three members of cystatin superfamily share a conserved cystatin fold of five-stranded anti-parallel β-sheets wrapped around an almost perpendicular five-turn α-helix at the centre (Bode et al., 1988; Kordić & Turk, 2009; Stubbs et al., 1990). Even the plant inhibitor oryzacystatin-I, demonstrates the same cystatin fold as animal cystatins (Nagata, Kudo, Abe, Arai, & Tanokura, 2000). Within this cystatin fold, some parts of N-terminal region, highly conserved β-hairpin loop L1

*Corresponding author. Email: aseal@klyuniv.ac.in

© 2012 Taylor & Francis
(Q-X-V-X-G) and a second hairpin loop L2 (P-W with a noted exception for human stefns) altogether form a wedge-like structure, complementary to the catalytic cleft of papain (C1) family of cysteine proteases (Bode et al., 1988; Kordis & Turk, 2009; Stubbs et al., 1990).

Mammalian type 2 cystatins show extensive diversity due to existence of multigene families, genetic polymorphism of coding sequences and occasional post-translational modifications (Dickinson, Thiesse, & Hicks, 2002). S-type cystatins (cystatin S, SA and SN), member of family 2 cystatins, are secreted proteins (Ryan et al., 2010) and mainly expressed in submandibular and sublingual saliva as well as in tears, urine and seminal fluid but not in parotid saliva (Dickinson et al., 2002, Ryan et al., 2010). Although, S-like cystatins exhibit more than 90% sequence similarity, yet cystatin S is a significantly poorer inhibitor than cystatin SA or SN for papain family of cysteine proteases (Baron, DeCarlo, & Featherstone, 1999; Dickinson, 2002) and all three show no activity against legumain due to absence of Asn residue as seen at 40th position of Figure 1 (Dickinson, 2002). S-type cystatins have evolved in primates to inhibit a very narrow, specialized subset of host cysteine proteases in contrast to other salivary cystatins such as cystatin C, cystatin D and mainly protects mucosa and their secretions from dietary and environmental cysteine proteases (Blankenvoorde, Henskens, van’t Hof, Veerman, & Nieuw Amerongen, 1996; Dickinson, 2002; Ruzindana-Umunyana & Weber, 2001).

Cystatin M/E is also a secreted glycoprotein distantly related to other human type 2 cystatins, and reveal a five-residue insertion in the α-helix 2/loop (Figure 1). The expression pattern of human cystatin M/E is unique among cystatins, as high expression levels are largely confined to cutaneous epithelia (epidermis, hair follicles, sebaceous glands and sweat glands), whereas other cystatins are mainly ubiquitously expressed (Cheng et al., 2008; Zeeuwen et al., 2001). Cystatin M/E controls skin barrier formation by the regulation of both cross-linking and desquamation of the stratum corneum (Elias, 2005; Zeeuwen, 2004; Zeeuwen, Cheng, & Schalkwijk, 2009). It directly regulates the activity of cathepsin V, cathepsin L and legumain, thereby controls the processing of transglutaminases (Cheng et al., 2006). Cystatin M/E is also reported as a novel candidate tumor suppressor gene for breast cancer (Zhang et al., 2004).

Several three-dimensional structures of other human cystatins are available in Protein Data Bank (PDB) in both monomeric and dimeric form and some in complex with cysteine proteases elucidating their inhibitory activity. Some of these cystatins are active mostly in monomeric state and loses their inhibitory property on dimerization by domain swapping (Abrahamson & Grubb, 1994). Here, we make an effort to model four human cystatins, namely M/E, S, SA and SN by comparative modelling and verify their stability and viability through molecular dynamics (MDs) simulations and docking studies.

**Methodology**

The amino acid sequences (121 amino acids in each case) excluding signal peptide region of cystatin M/E, S,
SA and SN (UNIPROT Accession No. Q15828, P01036, P09228 and P01037, respectively) and their basic informations like sequence length, molecular mass, signal peptide, reactive sites, amino acid modifications and disulphide bond positions were collected from the most publicly available domain UNIPROT (http://www.uniprot.org/). The obtained protein sequences were also characterized by Expasy-Protparam tool (Gasteiger et al., 2005) and XtalPred server (Slabinski et al., 2007).

**Prediction of intrinsic disordered region**

In order to specify regions of higher flexibility and complementarities, tools like DisEMBL (Linding, Jensen et al., 2003), Globplot (Linding, Russell, Neduva, & Gibson, 2003), regional order neural network (RONN) (Yang, Thomson, McNeil, & Esnouf, 2005) and protein disorder prediction system (PrDOS) (Ishida & Kinoshita, 2003) were employed. Regional order neural network (RONN) (Yang, Thomson, McNeil, & Esnouf, 2005) and protein disorder prediction system (PrDOS) (Ishida & Kinoshita, 2007) were used to specify regions of higher flexibility and complementarities. Tools like DisEMBL were employed. In order to specify regions of higher flexibility and complementarities, tools like DisEMBL (Linding, Jensen et al., 2003), Globplot (Linding, Russell, Neduva, & Gibson, 2003), regional order neural network (RONN) (Yang, Thomson, McNeil, & Esnouf, 2005) and protein disorder prediction system (PrDOS) (Ishida & Kinoshita, 2003) were employed. Regional order neural network (RONN) (Yang, Thomson, McNeil, & Esnouf, 2005) and protein disorder prediction system (PrDOS) (Ishida & Kinoshita, 2007) were used to specify regions of higher flexibility and complementarities. Tools like DisEMBL were employed.

**Template selection**

NCBI Position Specific Iteration–Basic Local Alignment (PSI-BLAST) (Altschul, Gish, Miller, Myers, & Lipman, 1990) was performed against PDB proteins with default parameters to select the template protein and in turn the template was downloaded from PDB. The template–target relationship was reconﬁrmed by multiple sequence alignment between template protein and query sequences via Clustal X (Thompson, Gibson, Plewniaik, Jeanmougin, & Higgins, 1997) with the default parameters. Utopia CINEMA (Pettifer, Sinnott, & Attwood, 2004) was used to edit the alignment.

**Secondary structure prediction**

Secondary structure of the query proteins and template was predicted and compared by different methods, viz. Discrimination of protein Secondary structure Class (DSC) (King & Sternberg, 1996), GOR secondary structure prediction method version IV (GOR4) (Garnier, Gibrat, & Argos, 1996), Hierarchical neural network method (Guermeur, Geourjon, Gallinari, & Deléage, 1999), PHD (Rost & Sander, 1993), Predictor (Frishman & Argos, 1996), Self-optimized prediction method with alignment (Geourjon & Delage, 1995) and Sec.Cons (Deléage, Blanchet, & Geourjon, 1997).

**Model building**

Refined three-dimensional homology models of all the four proteins were generated using a comparative modeling program MODELLER (Eswar, Eramian, Webb, Shen, & Sali, 2008) of Discovery Studio, based on given sequence alignment and selected template. MODELLER determined probability density functions (Sali & Blundell, 1993) using statistical and classical mechanics with reference to a database of known protein structures as a guide for spatial restraints. Then distance and angle constraints were derived from template strand together with the energy terms that governed the relative spatial arrangement of atoms within the models. Later, it was optimized in Cartesian space with conjugate gradients and MD methods. Five structures were generated per protein and model with lowest value of the normalized discrete optimized protein energy (Sali & Blundell, 1993; Sali, Potterton, Yuan, van Vlijmen, & Karplus, 1995) was chosen as the best model and consequently subjected to MDs simulation.

**PDB file preparation**

The second identical chain B and crystal waters were removed from the template protein structure, conformers were deleted on basis of occupancies and further the occupancies were corrected. Then template protein and the four raw evaluated structures of DS MODELLER were repaired by Swiss PDB Viewer (Guex & Peitsch, 1997) and further by XLEAP of AMBER tool 1.5 (Case et al., 2010) mainly to clear the nomenclature issues.

**MDs simulation**

In total, five systems (one template & four models) were prepared for simulation. All MD simulations were performed using GROMACS (GROningen MAchine for Chemical Simulations) 4.5.4 (Hess, Kutzner, van der Spoel, & Lindahl, 2008) simulation packages employing Charmm27 (Chemistry at Harvard Macromolecular Mechanics) force field (MacKerell et al., 1998). The systems were solvated by explicit SPC/E (extended simple point charge model) water model (Berendsen, Straatsma, 1987) in cubic boxes maintaining a minimum 10 Å distance from cube edge. Counter ions (Cl– and Na+) were added by randomly replacing water molecules to achieve a neutral simulation cell. Each system was then minimized using a steepest descents integrator (Apol et al., 2011) in order to relieve from steric clashes or inappropriate geometry.

A 100 ps NVT equilibration was performed at 300 K with position restraints applied to all of the backbone atoms in order to relieve any bad contacts at the side chain solvent interface. The velocity rescale thermostat (Bussi, Donadio, & Parrinello, 2007) was used with a temperature coupling time constant (τT) of 0.1 ps. All bond lengths were constrained using the linear constraint solver (LINCS) (Hess, Bekker, Berendsen, & Fraaije, 1997) algorithm, which allowed for a 2 fs time step. Long-range electrostatic interactions were approximated using the particle mesh Ewald (PME) method (Darden, York, & Pedersen, 1993) with a fourth-order spline interpolation and a 0.15 nm Fourier grid spacing. The short range non-bonded interactions were defined as van der Waals (VDW) and electrostatic interactions between particles within 10 Å. Then position restraints were withdrawn and a 100 ps NPT simulation was conducted.
relaxing the system into an isotropic Parrinello-Rahman barostat (Parrinello & Rahman, 1981) set to 1.0 bar of pressure in all directions and a pressure coupling time constant (\( \tau_P \)) of 2.0 ps. The production MD that followed was performed with the velocity rescale thermostat and Parrinello–Rahman barostat, as well as the LINCS and PME treatments as described. Snapshots of the trajectory were taken in every 2 ps.

The GROMACS suite of tools along with a secondary structure recognition algorithm (DSSP) (Kabsch & Sander, 1983) was used for all types of MD simulation studies. Electrostatic potential, molecular surfaces were calculated using VMD (Sass, 1998) and structural alignments were done in DS suite. Microsoft Excel program was used for preparation of the graph and for structure visualization, DS Visualizer and Chimera (Pettersen et al., 2004) was employed.

**Structure validation**

Stereochemistry and conformations of the crude models (MODELLER output) and refined structures (post simulation) were evaluated using PROCHECK (Laskowski, MacArthur, Moss, & Thornton, 1993), WHAT_CHECK (Hooft, Vriend, Sander, & Abola, 1996), Verify3D (Lüthy, Bowie, & Eisenberg, 1992), ERRAT (Colovos & Yeates, 1993) and PROtein Volume Evaluation (PROVE) (Pontius, Richelle, & Wodak, 1996) through Structural Analysis and Verification Server (SAVES) at http://nih-server.mbi.ucla.edu/SAVES/ and also with Atomic Non-Local Environment Assessment (ANOLEA) (Melo & Feytmans, 1998), ProSA (Sippl, 1993, 1995) and MolProbity (Davis et al., 2007) servers.

**Binding site study: molecular docking**

Computed Atlas of Surface Topology of Proteins (CASTp) (Dundas et al., 2006), cons-PPISP (Chen & Zhou, 2005) and the InterProSurf server (Negi, Schein, Oezquen, Power, & Braun, 2007) were utilized in order to locate the binding-site amino acids. To validate the binding sites and investigate the mode of binding, each of the four optimized models was docked with papain (PDB ID 1KHQ) (Janowski, Kozak, Jankowska, Grzonka, & Jaskólski, 2004) using ZDOCK program (Chen & Weng, 2002) in Discovery Studio 2.5. ZDOCK is a rigid-body docking algorithm that uses Fast Fourier Transform (FFT) to perform an exhaustive six-dimensional grid-based search in the translational and rotational space between the two molecules. ZDOCK searches orientational space by rotating the ligand around its geometric centre keeping the receptor protein fixed in space. For each sampled angle, the algorithm rapidly scans translational space using FFT based on a 3D grid with step size of 1.2 Å and only the ligand translation corresponding to the best geometric match between the two proteins is retained. The binding sites of papain (CYS 25, HIS 159 and ASN 175) (Drenth, Jansonius, Koekoek, Swen, & Woltthers, 1968; Kamphuis, Kalk, Swarte, & Drenth, 1984) and the possible inhibitory sites of cystatins (Bode et al., 1988; Kordiš & Turk, 2009; Stubbs et al., 1990) were defined to screen the potential docking configurations. A total of 2000 docked poses were generated for each of the cystatin-papain complex models. All the predicted complexes were ranked based on a scoring function (Chen, Li, & Weng, 2003) combining shape complementarity, desolvation energy and electrostatics, and also re-ranked by ZRANK program (Pierce & Weng, 2007) which uses a more detailed weighted energy function with VDW, electrostatics and desolvation terms to obtain the best docked pose. For further refinement through RDOCK (Li, Chen, & Weng, 2007) program, five high scoring complexes resulting from ZRANK ranking were selected on basis of known papain-cystatin conformations by binding site root-mean-square deviation (RMSD) calculation. RDOCK employs a three-stage energy minimization scheme, followed by the evaluation of electrostatic and desolvation energies using CHARMM polar H force field. Ionic side chains were kept neutral in the first two stages of minimization, and reverted to their full charge states in the last stage of brief minimization. The best docking pose with the lowest RDOCK energy was considered.

Position optimizations of all the four complexes were carried out by using all-atom CHARMM27 (MacKerell et al., 1998) force field with Adopted Basis Newton–Raphson minimization algorithm (Chu, Trout, & Brooks, 2003) in DS 2.5. In order to identify the amino acids at the protein–protein interface, hydrogen bond (HB) analysis was performed and interaction energies were computed in vacuum using implicit distance-dependent dielectrics for each of the minimized complexes in DS 2.5. Further, relative percentage of solvent accessibility was considered by NACCESS program (Pollastri, Baldi, Fariselli, & Casadio, 2002). Binding free energies between the papain and cystatins were determined at physiological ionic strength with the Molecular Mechanics Poisson–Boltzmann non-polar Surface Area (MM-PBSA) method (Grandis, Bizzarri, & Cannistraro, 2007; Srinivasan, Cheatham, Cieplak, Kollman, & Case, 1998) in DS 2.5. The procedure is based on a combination of molecular mechanics and continuum solvent approaches to give estimation for the binding free energy of a protein complex.

**System specification**

All the work was done using windows operating system except simulation, which was performed by Fedora 10 OS in IBM D10 workstations with Xenon quad core processor.

**Results and discussion**

All the proteins are hydrophilic in nature and except cystatin M/E, all including the template cystatin C are
declared as unstable in both monomeric and dimeric forms by Expy-ProtParam tool and XtalPred server.

**Detection of disordered regions in protein**

All the intrinsic disorder predicting servers detect consistently, the initial N-terminal region and the final two or three C-terminal residue as disordered region. Some other portions such as regions 76–83 for Cystatin M/E and 80–90 (approximately) for the rest of the cystatins, come common in RONN and DisEMBL which can be elucidated as AS loop from later studies (Figure 4). DisEMBL also marks the post-helix region of all proteins.

**Template selection & homology modelling**

In PSI-BLAST, all the queries are recognized as member of cystatin superfamily and PDB entries from cystatin C and cystatin D show maximum similarity with the target proteins. Among these the higher resolution (1.70 Å) structure, PDB ID 3GAX (Kolodziejczyk et al., 2010) is chosen as template despite the presence of some missing residues and mutations at L47C and G69C, as the rest of the structures are either in domain swapped dimeric form or an inhibitory site mutant. The mutations of 3GAX are reversed; the missing residues are fabricated in Discovery Studio Biopolymer module and subjected to 2000 step energy minimization with smart minimizer of Discovery Studio before using for modelling.

The identities of PSI-BLAST search and an alignment between template and target proteins are presented in Table 1 and Figure 1, respectively. It is evident that the identities are in safe zone and the queries can be homology modelled with this template (Rost, 1999). Secondary structure prediction of all five template-target proteins also shows a good correlation with template, which also recommends 3GAX as a template for these cystatins (Table SI, Supplementary Material).

Further, conserved regions and cysteine residues responsible for forming disulphide bonds in the target proteins are perfectly aligned with template (Figure 1). It is also clear from alignment as the N-terminal region in the template is truncated; the models will be abnegated from the proper structure for the mentioned region. Hence, the target sequences fed to MODELLER starts from the 13th column of the alignment (Figure 1) for all target proteins.

**Stability of the models**

The stability of the modelled structures is first assessed with DSSP profile (Figure 2), then with RMSD, root-mean-square fluctuation (RMSF) analysis and radius of gyration ($R_g$) studies (Figure 3). It is evident from DSSP studies that the main cystatin fold of five anti-parallel β-sheets wrapped around a α-helix at the centre as well as L1 (45–50) and L2 loops (97–101) remains conserved throughout the simulation period in all the modelled structures (Figure 2). At the same time, regular fluctuations are observed in pre- (5–10) and post- (25–30) α-helix regions near N-terminal and AS loops (65–75) (Figures 2 and 3(B)). It might also be noted that the AS loop and post-helix regions are designated as disordered regions by the intrinsic disorder prediction servers.

The RMSD of the models gets stabilized with the time, shows a maximum standard deviation of 0.2 Å for cystatin M/E (Figure 3(A)) in comparison to other three modelled structures. A RMSF analysis reveals β-sheets are the most stable secondary structures in the models and loops; terminal regions fluctuate more than others (Figure 3(B)). Radius of gyration ($R_g$) almost simply replicates the RMSD pattern (Figure 3(C)) and records a highest standard deviation of 0.24 Å for Cystatin S among all four models. All these observations altogether prove the reliability of the structures.

In human cystatin C, under mild denaturing condition (lowering of pH or increase in temperature or both) or due to disruption of the hydrophobic core in presence of mutation (L68Q for human cystatin C, 69th column in Figure 1) or in presence of both, the L1 loop transforms into an extended β-strand and participates in domain swapping during dimerization (Janowski et al., 2001). In this study, no notable irregularity is observed in the specific site (61st position for cystatin M/E and 57th for the rest in DSSP Profile Figure 2) although cystatin M/E has a more hydrophilic methionine (M) residue at the same position (69th column in Figure 1) compared to valine (V) of others. But, L1 loop (45–50 in Figure 2(B)) of cystatin S shows somewhat a different nature from others (Figures 2(B) and 3(B)) which may be an indication towards dimerization.

Apart from these generalized notions, some specific features about these modelled structures can also be noted. Cystatin M/E, although showing the most near to perfect DSSP as well as RMSD profile (Figures 2 and 3(A)) among the four models, the AS loop region differs markedly from others in the length and by the absence of second α-helix immediately after the AS loop perhaps due to the presence of R-V-T-G-D portion as seen in the template-target alignment (Figure 1). This

| Sequence alignment | Identity (%) |
|--------------------|--------------|
| 3GAX – Cystatin M/E | 36           |
| 3GAX – Cystatin S  | 55           |
| 3GAX – Cystatin SA | 56           |
| 3GAX – Cystatin SN | 54           |
portion is also noted as disordered region from earlier study.

In Cystatin S, the L1 loop shows some flip-flop between bend and turn after 2 ns run and regains its bend status towards the final stages of simulation (Figure 2).

The residues 60–65 (the end portion of 3rd β-sheet) demonstrate a fluctuation of secondary structure after 14 ns (Figure 2) and it contribute significantly in the RMSD plot (Figure 3(A)). Cystatin S also exhibits highest ebb and flow around AS loop (Figure 3(B)). More-
over, cystatin S is the only model which shows a consistent increase in the radius of gyration over the time of simulation. All these features perhaps point towards its propensity towards dimerization by domain swapping.

C-terminal residues (105–109) of cystatin SA (114–121) are transformed into coil conformation after 12 ns and remain so (Figure 2(C)). Residues in between 75–80 shift towards turn from helix conformation after 17 ns (Figure 2(C)) and both of these regions (residue 72–109) are also noted as disordered region. These abnormalities as well have their part in RMSD (Figure 3(A)).

Finally for cystatin SN, DSSP detects a consistent β-bridge instead of β-sheet present in the N-terminal region of the rest of the models (Figure 2(D)).

As an additional elucidation of stability, two disulphide bonds characteristic to family 2 cystatins can also be taken into account. Both, report a bond length of 2.03 ± 0.004 Å in all four modelled structures during 20 ns of simulation. The 3GAX monomer also shows good threading with all four models. The main-chain atom-based RMSD values of cystatin M/E, cystatin S, cystatin SA and cystatin SN with reference to template structure are calculated to be 2.286, 1.717, 1.359, and 1.302 Å, respectively (Figure 4). The functional domains of all models show good alignment with the template. All these results approve the stability of the modelled structure in contrast to Expasy-ProtParam tool and XtalPred server prediction.

The electrostatic molecular surfaces of the final models differ from each other at physiological pH (Figure 5) mainly due to their difference in isoelectric pH (pI) and variation in sequence (Figure 1). This gives an indication that binding strategies of these models towards receptor may possibly differ from each other.

Structure validation

The stereochemical qualities of the models are evaluated with Ramachandran’s plot (Ramachandran & Sasisekharan, 1968) calculated using PROCHECK (Figure S1, Supplementary Material). Backbone Psi and Phi dihedral angles of the refined structures reveal that at least 95% residues of each protein are in most favored regions and
none is in the disallowed regions. Moreover, the overall G-factor of each model is greater than the recommended value (−0.50) which approves the correctness of the models (Table 2).

WHAT CHECK provides an overall summary of the quality of the structure in comparison to a set of reliable structures. Structural Z-scores offer a number of constraint-independent quality indicators whereas RMS Z-scores give an impression of how well the model conforms to common refinement constraint values. The detailed results were given in Table 2.

Verify3D examines the compatibility of an atomic model (3D) with its own amino acid sequence by assigning a structural class (i.e. a 3D profile) to each residue based on its location and environment (alpha, beta, loop, polar, non-polar etc.). The server reports more than 80% residues have a score >0.2 for each refined model (Table 2 and Figure S2, Supplementary Material) which indicate towards the good quality of models.

ERRAT by analyzing the statistics of non-bonded atom–atom interactions in reference to a database of reliable high-resolution structures identifies a marked improvement in each structure after simulation (Table 2 and Figure S3, Supplementary Material).

Z-score RMS (volume Z-score RMSD) values, computed by PORVE program, measures the average magnitude of the volume irregularities in the structure, also gives near acceptable values (≈1) for all of the final structure (Table 2).

Another well-known validation server ProSA determines Z-scores as a measure of overall model quality by calculating the deviation in total energy of the structure with respect to an energy distribution derived from random conformations. The negative Z-scores for all models stand for the correctness of the models (Table 2).

Further, the packing quality of the model is assessed by ANOLEA, which uses a very sensitive and accurate atomic mean force potential to calculate the non-local energy profile (NL-profile) of a proteins structure. All negative non-local normalized energy Z-score establishes the accuracy of the models (Table 2 and Figure S4, Supplementary Material).

Finally, Molprobity server also does not report the presence of any bad angles or bonds in the modelled structures. All these results altogether prove the reliability of the structures. Table 2 summarizes all these results.

**Binding site identification**

The conserved regions around L1, L2 loop of each of the model are recognized by CASTp, cons-PPISP and the InterProSurf server as binding pockets. Additionally, CASTp detects some binding cavities around AS loop and cons-PPISP and the InterProSurf credits some N-terminal residues as protein–protein interaction sites. These results are in agreement with previous crystallographic and mutagenic studies (Bode et al., 1988; Kordiš & Turk, 2009; Stubbs et al., 1990).

Some physical properties of protein–protein interface of these four docked complexes (Figure 6) are analyzed and listed in Table 3. The interface accessible surface area (ASA) represents the difference in the water ASA between the complex and the single proteins; provides information on the protein–protein geometric fit. A higher ASA is, in fact, an indication of a higher shape complementary between the molecules. ASA values between 1000 and 2000 Å² are a feature of stable enzyme–inhibitor complexes (Nooren & Thornton, 2003) and all the docked complexes are in harmony with the observation. We can also note that the complexes mainly characterized by polar interfaces, which approve the non-obligatory nature of enzyme–inhibitor complexes and suggesting the presence of number of HBs.
To identify the residues involved in interaction in the protein-protein interface, interaction energy (IE) and relative solvent accessible surface areas (rSASA) are calculated and HB analysis is performed (Table 4). The IE is the sum of the VDW and electrostatic (E) energy and the value is broken down per-residue for each residue located at the interface (Table SII, Supplementary Material). rSASA areas (%) of each residue in the complex (rSASc) and individual proteins (rSASI) are calculated in NACCESS and higher ΔSAS values (rSASc/rSASI > 30) are listed in Table 4 and SIII, Supplementary Material. NACCESS uses tripeptide model to normalize the surface area to eliminate the effects of different residue types.

Residues exhibiting high IE mostly involved either in hydrogen bonding or in hydrophobic interactions or in both (Table 4 and SII, SIII, Supplementary Material). Short strong hydrogen bonds (SSHB)<2.7 Å are observed within some residues, (Table SIV, Supplementary Material) which also scores highly in IE analysis (Table SII, Supplementary Material). In addition, triadic HBs are also present (Table 4 and SV, Supplementary Material), which may strengthen the H-bond strength and sometimes show the catalytic effect in certain biosystems (Kim, Kim, Lee, Tarakeshwar, & Oh, 2002; Kim, Oh, & Lee, 2000). Moreover, all SSHBs forms triadic HBs (marked in Table 4 and SV, Supplementary Material) and we can safely predict that these residues play a very important role in the interaction.

Here, we can notice that the catalytic residues (CYS 25, HIS 159 and ASN 175) of papain (Drenth et al., 1968; Kamphuis et al., 1984) do not directly interact with cystatins, rather the residues in their neighbourhood interact with cystatins more frequently. Four definite zones can be located around the active site of papain that take part in interaction with cystatins, namely (i) residues 20–25 (G-S-C-G-S-C), (ii) residues 64, 67 (N, Y), (iii) residues 139, 142 (K, Q) and (iv) residues 156–159 (K-V-D-H). Among these residues SER21, GLY23, ASN64, LYS139, LYS156, VAL157 and ASP158 come common in most of the complexes and ranked highly in IE, HB and SASA analysis (Table 4 and SII-SIV, Supplementary Material).

Figure 5. Electrostatic potentials with solvent accessible area of modeled proteins calculated by DS Studio Biopolymer at physiological pH; positive potential charges are in blue, negative values are in red and intermediates are in white.
Table 2. Comparative results of structure validation.

| Server          | Template | Cystatin M/E | | Cystatin S | | Cystatin SA | | Cystatin SN | |
|-----------------|----------|--------------|-----------------|--------------|-----------------|-----------------|--------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| PROCHECK        |          |              |                 |              |                 |                 |              |                 |                 |                 |                 |                 |                 |                 |
| % of Residues in most favoured regions | 88.00 | 93.14 | 96.08 | 93.00 | 96.00 | 95.05 | 97.03 | 92.00 | 95.00 |
| % of Residues in additional allowed regions | 10.00 | 4.90 | 2.94 | 6.00 | 3.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| % of Residues in generously allowed regions | 2.00 | 1.96 | 0.98 | 0.00 | 1.00 | 0.99 | 0.00 | 1.00 | 0.00 |
| % of Residues in disallowed regions | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| Overall G-factor |          |              |                 |              |                 |                 |              |                 |                 |
| WHAT_CHECK      |          |              |                 |              |                 |                 |              |                 |                 |
| Structure Z-scores: |          |              |                 |              |                 |                 |              |                 |                 |
| Second generation packing quality | -0.015 | -1.695 | -0.563 | -0.822 | -0.199 | -0.592 | 0.447 | -0.799 | 0.059 |
| Ramachandran plot appearance | -2.035 | 0.021 | 0.986 | 0.550 | -0.279 | 0.856 | 0.426 | 0.380 | 1.130 |
| Chi-1/chi-2 rotamer normality | -0.206 | -0.184 | 0.182 | -0.714 | -1.524 | -0.251 | -1.423 | 0.114 | -1.286 |
| Backbone conformation | -11.868 (bad) | -11.383 (bad) | -8.679 (bad) | -6.496 (bad) | -5.163 (bad) | -5.923 (bad) | -5.298 (bad) | -7.568 (bad) | -5.891 (bad) |
| RMS Z-scores: |          |              |                 |              |                 |                 |              |                 |                 |
| Bond lengths | 1.062 | 0.964 | 0.699 | 0.969 | 0.730 | 0.978 | 0.714 | 0.975 | 0.715 |
| Bond angles | 1.118 | 1.248 | 1.304 | 1.258 | 1.153 | 1.243 | 1.106 | 1.238 | 1.153 |
| Omega angle restraints | 1.283 (loose) | 0.739 | 1.483 (loose) | 0.665 (tight) | 1.397 (loose) | 0.684 (tight) | 1.561 (loose) | 0.693 (tight) | 1.577 (loose) |
| Side-chain planarity | 1.499 | 0.288 (tight) | 1.119 | 0.405 (tight) | 1.486 | 0.411 (tight) | 1.335 | 0.369 (tight) | 1.412 |
| Improper dihedral distribution | 0.958 | 0.737 | 0.716 | 0.759 | 0.932 | 0.688 | 0.883 | 0.662 | 0.921 |
| Inside/outside distribution | 1.115 | 1.069 | 1.052 | 1.107 | 1.110 | 1.117 | 1.077 | 1.085 | 1.087 |
| Verify3D |          |              |                 |              |                 |                 |              |                 |                 |
| Residues having a score > 0.2 (%) | 79.13 | 70.18 | 80.70 | 90.91 | 80.91 | 71.82 | 88.18 | 85.45 | 98.18 |
| ERRAT |          |              |                 |              |                 |                 |              |                 |                 |
| Overall quality factor | 98.077 | 61.905 | 95.349 | 52.475 | 94.845 | 70.297 | 79.381 | 67.327 | 87.755 |
| PROVE |          |              |                 |              |                 |                 |              |                 |                 |
| Z-score RMS | 1.216 | 56.758 | 1.459 | 1.428 | 1.659 | 1.479 | 1.705 | 1.550 | 1.895 |
| ProSA |          |              |                 |              |                 |                 |              |                 |                 |
| Z-Score | -4.73 | -4.69 | -3.35 | -4.49 | -3.93 | -3.93 | -3.76 | -4.78 | -4.31 |
| ANOLEA |          |              |                 |              |                 |                 |              |                 |                 |
| Non-local normalized energy Z-score | -1.67 | 1.20 | -2.92 | 0.86 | -2.08 | 0.30 | -2.77 | 1.48 | -2.33 |
| Molprobity |          |              |                 |              |                 |                 |              |                 |                 |
| % of Residues with bad bonds | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| % of Residues with bad angles | 0.00 | 0.88 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.92 | 0.00 |

Note: The initial and final structures correspond to the direct output of MODELLER and the refined post simulation structures.
Material) whereas 156LYS, 139LYS and 142GLN come up as important triadic HB centres (Table 4 and SV, Supplementary Material).

It is evident from the IE and ASA analysis (Tables SII and SIII, Supplementary Material) that cystatin M/E interacts differently with papain in comparison to S-type cystatins which shares a somewhat common pattern within them. For S-type cystatins, the conserved L1 (QXVXG) and L2 (PW) loop are mainly responsible for inhibition of papain but in case of cystatin M/E the N-terminal region contributes equally as seen from IE, HB and SASA analysis (Tables SII-SIV, Supplementary Material).

![Figure 6](image-url)

Figure 6. (A) Three-dimensional superimposed structure of docked complexes. Spatial details of binding sites of papain-cystatin M/E (B), papain-cystatin S (C), papain-cystatin SA (D) and papain-cystatin SN (E) complexes. From B-E, papain is in red, cystatin is in green, important residues of papain and cystatin are marked in blue and pink, respectively.
Material). In the L1 region residue 46, 47 and in L2 region the highly conserved Trp99/95 are key residues of binding in all four complexes and also take part in triadic HB formation (Table 4 and SII-SIV, Supplementary Material). From IE analysis (Table SII, Supplementary Material) we can say, cystatin M/E and cystatin SN interact much more strongly with papain than the others, which is also obvious from binding free energy results (Table 3). Cystatin S shows least number of hydrogen bonding (Table 3 and SIV, Supplementary Material) and interaction centres (Table SV, Supplementary Material), which goes well with its physiologically inactive and highly specific nature (Dickinson, 2002; Dickinson et al., 2002). Absence of N-terminal as well as C-terminal interactions and presence of less hydrophobic PHE46 instead VAL in L1 loop of cystatin S possibly make it a poorest inhibitor among S-type cystatins.

| Complex          | ASA (Å²) | Number of hydrogen bonds | Polar residues (%) | Non-polar residues (%) | Binding free energy (kcal/mol) |
|------------------|----------|--------------------------|-------------------|------------------------|-------------------------------|
| Papain-cystatin M/E | 1823.8   | 14                       | 70.4              | 29.6                   | −100.93                      |
| Papain-cystatin S   | 2190.2   | 11                       | 73.6              | 26.4                   | −64.62                       |
| Papain-cystatin SA  | 2181.0   | 14                       | 68.9              | 31.1                   | −69.85                       |
| Papain-cystatin SN  | 2474.8   | 14                       | 71.4              | 28.6                   | −86.02                       |

Note: Residues exhibiting SSHB are italicized and residues involved triadic hydrogen bonds are underlined. Residues ranked highly in IE, HB and SAS analysis are made bold. Chain A and I denotes Papain and Cystatin, respectively.

### Conclusion

In this work we have explored the inhibitory interaction between papain and human S-type cystatins and cystatin M/E. By comparative modelling, MD simulation and protein docking studies, we have predicted and validated the structures of the cystatins as well as their complexes with papain, with a noted propensity of cystatin S towards dimerization. The docked complexes elucidate structural details of interaction and indicate that some key residues are involved in interaction. The protein interfaces are characterized with ASA values distinctive for most stable complexes and numerous HBs, hydrophobic contacts that add specificity to the interaction. Four zones are identified in papain mainly responsible for cystatin binding and dissimilar approaches for papain binding between cystatin M/E and S-type cystatins are also recognized. Docked complexes exhibit diverse speci-
ificiencies towards papain and comparative inert nature of cystatin S is also elucidated. Our results are therefore consistent with the current knowledge and explain poorer inhibitory nature of cystatin S.

Abbreviations

3D three dimensional
Å Angstrom
AHSG alpha 2HS glycoprotein
ASA accessible surface area
DS Discovery Studio
E_IE electrostatic interaction energy
HB hydrogen bond
HRG histidine-rich glycoprotein
IE interaction energy
K Kelvin
MD molecular dynamics
nm nano meter
ns nano second
PDB Protein Data Bank
pl Isoelectric pH
ps pico second
Rg radius of gyration
RMS root mean square
RMSD root mean square deviation
rSASA relative solvent accessible surface area
VDW van der Waals interaction energy

Acknowledgements:

The whole study was supported by grants from BTIS net programme of DBT, Ministry of Science & Technology, Government of India, New Delhi and University of Kalyani.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2012.706403.

References

Abrahamson, M., Alvarez-Fernandez, M., & Nathanson, C.M. (2003). Cystatins. In J. Saklatvala, H. Nagase, & G. Salvesen (Eds.), Biochemical Society Symposia: Vol. 70. Proteases and the Regulation of Biological Processes (pp. 179–199). London: Portland Press Limited. Retrieved from http://symposia.biochemistry.org/bssymposium/070/bssymp07001197.htm

Abrahamson, M., & Grubb, A. (1994). Increased body temperature accelerates aggregation of the Leu-68 Gln mutant cystatin C, the amyloid-forming protein in hereditary cystatin C amyloid angiopathy. Proceedings of the National Academy of Sciences of the United States of America, 91, 1416–1420. Retrieved from http://www.pnas.org

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Alvarez-Fernandez, M., Barrett, A.J., Gerhartz, B., Dando, P. M., Ni, J., & Abrahamson, M. (1999). Inhibition of mammalian legumain by some cystatins is due to a novel second reactive site. The Journal of Biological Chemistry, 274, 19195–19203. doi: 10.1074/jbc.274.27.19195

Apol, E., Apostolov, R., Berendsen, H.J.C., van Buuren, A., Bjellkmar, P., van Drunen, R., ... Lindahl, E. (2011). Gromacs User Manual (Version 4.5.4) [Computer software]. Retrieved from http://www.gromacs.org/

Baron, A., DeCarlo, A., & Featherstone, J. (1999). Functional aspects of the human salivary cystatins in the oral environment. Oral Diseases, 5, 234–240. doi: 10.1111/j.1601-0825.1999.tb00307.x

Barrett, A.J., Fritz, H., Grubb, A., Isemura, S., Järvinen, M., Katunuma, N., ... Turk, V. (1986). Nomenclature and classification of the proteins homologous with the cysteine-proteinase inhibitor chicken cystatin. Biochemical Journal, 236, 312. Retrieved from http://www.biochemj.org

Barrett, A.J., Rawlings, N.D., Davies, M.E., Machleidt, W., Salvesen, G., & Turk, V. (1986). Cysteine proteinase inhibitors of the cystatin superfamily. In A.J. Barrett & G. Salvesen (Eds.), Proteinase inhibitors (pp. 519–569). Amsterdam: Elsevier.

Berendsen, H.J.C., Grigera, J.R., & Straatsma, T.P. (1987). The missing term in effective pair potentials. Journal of Physical Chemistry, 91, 6269–6271. doi: 10.1021/j100308a038.

Blankenvoorde, M.F., Henskens, Y.M., van’t Hof, W., Veerman, E.C., & Nieuw Amerongen, A.V. (1996). Inhibition of the growth and cysteine proteinase activity of Porphyromonas gingivalis by human salivary cystatin S and chicken cystatin. Biological Chemistry, 377, 847–850. doi: 10.1515/bchm3.1996.377.12.847.

Bode, W., Engh, R., Musil, D., Thiele, U., Huber, R., Karshikov, A., ... Turk, V. (1988). The 2.0 A X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. The EMBO Journal, 7, 2593–2599. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/

Brown, W.M., & Dziegielew ska, K.M. (1997). Friends and relations of the cystatin superfamily—new members and their evolution. Protein Science, 6, 5–12. doi: 10.1002/prot.5560060102

Brown, W.M., Saunders, N.R., Møllgård, K., & Dziegielew ska, K.M. (1992). Fetuin—an old friend revisited. BioEssays, 14, 749-755. doi: 10.1002/bies.950141105

Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. The Journal of Chemical Physics, 126, 014101. doi: 10.1063/1.2408420

Case, D.A., Darden, T.A., Cheatham, T.E.III, Simmerling, C.L., Wang, J., Duke, R.E., ..., & Kollman, P.A. (2010). AMBER (Version 11) [Computer software]. San Francisco, CA: University of California.

Chen, R., Li, L., & Weng, Z. (2003). ZDOCK: An initial-stage protein-docking algorithm. Proteins: Structure, Function, and Bioinformatics, 52, 80–87. doi: 10.1002/prot.10389

Chen, R., & Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins: Structure, Function, and Bioinformatics, 47, 281–294. doi: 10.1002/prot.10092

Chen, H., & Zhou, H.X. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. Proteins: Structure, Function, and Bioinformatics, 61, 21–35. doi: 10.1002/prot.20514
Cheng, T., Hitomi, K., van Vlijmen-Willems, I.M., de Jongh, G.J., Yamamoto, K., Nishi, K., ... Zeeuwen, P.L. (2006). Cystatin M/E is a high affinity inhibitor of cathepsin V and cathepsin L by a reactive site that is distinct from the legumain-binding site. A novel clue for the role of cystatin M/E in epidermal cornification. The Journal of Biological Chemistry; 281, 15893–15899. doi: 10.1074/jbc.M600694200

Cheng, T., van Vlijmen-Willems, I.M., Hitomi, K., Pasch, M.C., van Erp, P.E., Schalkwijk, J., & Zeeuwen, P.L. (2008). Colocalization of cystatin M/E and its target proteases suggests a role in terminal differentiation of human hair follicle and nail. Journal of Investigative Dermatology; 129, 1232–1242. doi: 10.1038/jijd.2008.353

Chu, J.W., Trout, B.L., & Brooks, B.R. (2003). A super-linear method of protein structure modeling with MODELLER. Journal of Chemical Physics; 119, 12708–12717. doi: 10.1063/1.1627754

Colovos, C., & Yeates, T.O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Science; 2, 1511–1519. doi: 10.1002/pro.5560020916

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. The Journal of Chemical Physics, 98, 10089–10092. doi: 10.1063/1.464397

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., ... Richardson, D.C. (2007). MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research, 35, W375–W383. doi: 10.1093/nar/gkm216

Delage, G., Blanchet, C., & Gourjon, C. (1997). Protein structure prediction. Implications for the biologist. Biochimie, 79, 681–686. doi: 10.1016/S0006-3495(97)83524-9

Dickinson, D.P. (2002). Salivary (SD-type) cystatins: Over one billion years in the making—but to what purpose? Critical Reviews in Oral Biology & Medicine, 13, 485–508. doi: 10.1177/154411130201300606

Dickinson, D.P., Thiesse, M., & Hicks, M.J. (2002). Expression of type 2 cystatin genes CST1-CST5 in adult human tissues and the developing submandibular gland. DNA and Cell Biology, 21, 47–65. doi: 10.1089/10445490252810311

Drenth, J., Jansonius, J.N., Koekkoek, R., Swen, H.M., & Wolthers, B.G. (1968). Structure of papain. Nature; 218, 929–932. doi: 10.1038/218929a0

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp: Computed atlas of surface topology of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Research, 34, W116–W118. doi: 10.1093/nar/gkm282

Elias, P.M. (2005). Stratum corneum defensive functions: An integrated view. Journal of Investigative Dermatology; 125, 183–200. doi: 10.1111/j.0022-202X.2005.23668.x

Esvar, N., Eramian, D., Webb, B., Shen, M.Y., & Sali, A. (2008). Protein structure modeling with MODELLER. Methods in Molecular Biology; 426, 145–159. doi: 10.1007/978-1-59745-108-8_8

Frishman, D., & Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Engineering, 9, 133–142. doi: 10.1093/protein/9.2.133

Garnier, J., Gibart, J.F., & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. Methods in Enzymology, 266, 540–553. doi: 10.1016/S0076-8879(96)66034-0

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy Server. In J.M. Walker (Ed.), The proteomics protocols handbook (pp. 571–607). New York, NY: Humana Press. doi: 10.1385/1-59259-890-5:571

Geourjon, C., & Deleage, G. (1995). SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Computer Applications in the Biosciences: CABIOS, 11, 681–684. doi: 10.1093/bioinformatics/11.6.681

Grandis, V.D., Bizzarri, A.R., & Cannistraro, S. (2007). Docking study and free energy simulation of the complex between p53 DNA-binding domain and azurin. Journal of Molecular Recognition, 20, 215–226. doi: 10.1002/jmr.840

Guermeur, Y., Geourjon, C., Gallinari, P., & Deleage, G. (1998). Improved performance in protein secondary structure prediction by inhomogeneous score combination. Bioinformatics, 15, 413–421. doi: 10.1093/bioinformatics/15.5.413

Guex, N., & Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis, 18, 2714–2723. doi: 10.1002/elps.1150181505

Hess, B., Bekker, H., Berendsen, H.J.C., & Fraaije, J.G.E.M. (1997). LINCS: A linear constraint solver for molecular simulations. Journal of Computational Chemistry, 18, 1463–1472. doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Hess, B., Kutzner, C., van der Spoel, D., & Lindahl, E. (2008). GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. Journal of Chemical Theory and Computation, 4, 435–447. doi: 10.1021/ct070310q

Hooft, R.W., Vriend, G., Sander, C., & Abola, E.E. (1996). Errors in protein structures. Nature, 381, 272. doi: 10.1038/381272a0

Ishida, T., & Kinoshita, K. (2007). PRDOS: Prediction of disordered protein regions from amino acid sequence. Nucleic Acids Research, 35, W460–W464. doi: 10.1093/nar/gkm363

Janowski, R., Kozak, M., Jankowska, M., Grzonka, Z., Grubb, A., Abrahamson, M., & Jaskolski, M. (2001). Human cystatin C, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. Nature Structural Biology, 8, 316–320. doi: 10.1038/86188

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577–2637. doi: 10.1002/bip.360221121

Kim, K.S., Kim, D., Lee, J.Y., Tarakeshwar, P., & Oh, K.S. (2002). Catalytic mechanism of enzymes: Preorganization, short strong hydrogen bond, and charge buffering. Biochemistry, 41, 5300–5306. doi: 10.1021/bi0255118
Spassov, V.Z., & Yan, L. (2008). A fast and accurate computational approach to protein ionization. Protein Science, 17, 1955–1970. doi: 10.1110/ps.056335.108

Srinivasan, J., Cheatham, T.E., Cieplak, P., Kollman, P.A., & Case, D.A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate–DNA helices. Journal of American Chemical Society, 120, 9401–9409. doi: 10.1021/ja981844+

Stubbs, M.T., Laber, B., Bode, W., Huber, R., Jeral, R., Lenarcic, B., & Turk, V. (1990). The refined 2.4 Å X-ray crystal structure of recombinant human stefin B in complex with the cysteine proteinase papain: A novel type of proteinase inhibitor interaction. The EMBO Journal, 9, 1939–1947. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., & Higgins, D.G. (1997). The CLUSTAL_X windows interface. Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Turk, V., Stoka, V., & Turk, D. (2008). Cystatins: Biochemical and structural properties, and medical relevance. Frontiers in Bioscience, 13, 5406–5420. doi: 10.2741/3089

Yang, Z.R., Thomson, R., McNeil, P., & Esnouf, R.M. (2005). RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics, 21, 3369–3376. doi: 10.1093/bioinformatics/bti534

Zeeuwen, P.L. (2004). Epidermal differentiation: The role of proteases and their inhibitors. European Journal of Cell Biology, 83, 761–773. doi: 10.1078/0171-9335-00388

Zeeuwen, P.L., Cheng, T., & Schalkwijk, J. (2009). The biology of cystatin M/E and its cognate target proteases. Journal of Investigative Dermatology, 129, 1327–1338. doi: 10.1038/jid.2009.40

Zeeuwen, P.L., Van Vlijmen-Willems, I.M., Jansen, B.J., Sotiropoulos, G., Curls, J.H., Meis, J.F., ... Schalkwijk, J. (2001). Journal of Investigative Dermatology, 116, 693–701. doi: 10.1046/j.1523-1747.2001.01309.x

Zhang, J., Shridhar, R., Dai, Q., Song, J., Barlow, S.C., Yin, L., ... Keppler, D. (2004). Cystatin M: A novel candidate tumor suppressor gene for breast cancer. Cancer Research, 64, 6957–6964. doi: 10.1158/0008-5472.CAN-04-0819