Contents lists available at ScienceDirect

Journal of the Taiwan Institute of Chemical Engineers

journal homepage: www.elsevier.com/locate/jtice

Structural bioinformatics analysis of free cysteines in protein environments

Sheau Ling Ho[a,*], Andrew H.-J. Wang[b]

[a] *Department of Chemical Engineering, Chinese Culture University, Taipei 111, Taiwan*
[b] *Institute of Biological Chemistry, Academia Sinica, Taipei 115, Taiwan*

ARTICLE INFO

Article history:
Received 23 October 2007
Accepted 25 July 2008

Keywords:
Hydrophobic
Free cysteine
SH
SS
Proteases
Spatial neighborhood
Structural preference
Structural bioinformatics

ABSTRACT

Cysteine has been considered as a “hydrophilic” amino acid because of its pK_a and its ability to form (weak) hydrogen bonds. However, cysteines are found mostly in hydrophobic environments, either in S–S (disulphide) form or in free cysteine form. When free cysteines are found on the surface of proteins, they are often involved in catalytic residues, as in cysteine proteases, P-loop phosphatases, *etc.* Additionally, a unique property of cysteines is that their side-chain volume is different from all other amino acids. This study is focused on the discrimination between structural versus active free cysteines based on a local environment analysis which does not appear to have been attempted previously. We have demonstrated the corresponding structural positions associated with free cysteines in their three-dimensional localization environment. We examined protein samples including nine, sequenced, coronavirus proteases and cysteine-rich non-membrane proteins. Our present study shows that the sequential environments of free cysteines of coronavirus proteases are rather hydrophobic and that the free cysteines of non-membrane proteases have a higher amount of contacts to hydrophobic residues and lower amount of contacts to polar or charged residues.

1. Introduction

Cysteine is only incorporated into proteins at a level of 1.7% relative to the other amino acids (Mccaldon and Argos, 1988; Nilsson *et al.*, 2005). Cysteine’s importance in protein structure is related to the presence of a sulfur-containing thiol group in its side-chain. The thiol group of one cysteine residue is capable of combining with the thiol group of another to form a disulfide bridge, either linking two peptide chains together, as in the case of insulin, or causing a single peptide chain to fold back on itself, making a loop. The unique thiol side-chain (volume 219.43 Å^3) of this amino acid is often heavily involved in maintaining the proper configuration of both structural proteins and enzymes. Because of its pK_a and ability to form (weak) hydrogen bonds, cysteine has been considered to be a “hydrophilic” amino acid. It has a sulfhydryl (SH) group and is found in most proteins.

When free cysteines are found on the surface of proteins, they are often involved as catalytic residues, *e.g.*, in cysteine proteases, P-loop phosphatases that assist in the metabolism of various biochemicals, *e.g.*, heparin, biotin, coenzyme A and glutathione. In addition, they can be used in chelating metal ions, *e.g.*, in metallo enzymes, iron–sulfur clusters (F_2S_4), which is important in enzymatic and nucleophilic functions. When two cysteines are bonded by an S–S bond, the resulting molecule between the two protein chains is called cystine. The formation of disulfide bonds between cysteines present within proteins is important to the formation of active structural domains in a large number of proteins. However, they are found mostly in hydrophobic environments, either in free cysteine form (SH) or in disulphide form (SS). Recently, our lab has been involved in studying 3CL proteases which employ cysteine and histidine residues as catalytic dyads in the catalytic site. Around 4% of cysteine residues were observed in our protein sequence which showed a much higher fraction of amino acids within proteins than was expected.

The presence of a hydrophilic cysteine residue within the active site is what drew our attention. Thus we were particularly interested in having a better understanding of the roles played by free cysteines and in which they interact with other molecules. This project aimed at discriminating between structural, as opposed to, free cysteines based on analysis of their local environments.

In this study, we used computational approaches to annotate the free cysteines of individual cysteine-rich proteins, with criteria for determining the extent of other cysteine proteases. An analysis

[*] Corresponding author.
E-mail address: hxl3@faculty.pccu.edu.tw (S.L. Ho).

Abbreviations: Å, RMSD from the crystal structure of the complex; aa, amino acid; C, cysteine; SH, free cysteine form; SS, disulphide from; HSSP, a database of homology-derived secondary structure of proteins; PDB, Protein Data Bank; PERL, Practical Extraction and Report Language, a dynamic programming language; RMSD, the square deviations; pro, main proteinase.

1876-1070/$ – see front matter © 2008 Taiwan Institute of Chemical Engineers. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jtice.2008.07.015

of cysteines has enabled us to define their association in a sequence alignment by grouping residues into families. Supplementary research included a consensus approach to cysteine residues within the 3CLpro and extended it to a number of other proteins. The identification of the free cysteines, combined with classification based on functional features and spatial schematics, provides a basis for experimental validation and association of new molecules involved in cysteine activities. Since neighboring residues share physical characteristics (Zvelebil *et al.*, 1987), we have undertaken a more detailed study of the surroundings of cysteine residues in protein structures. The patterns discerned in the distribution of various residues and their constituent around cysteine should be useful in improving our understanding of protein stability, molecular recognition and binding.

2. Materials and methods

2.1. Database search and sequence analysis

2.1.1. Cysteine-rich proteinases

Sequences of cysteine-containing proteases were retrieved from the public 3D structural databases (Berman *et al.*, 2000) using combinations of sequence, and conserved motif searches to choose identifiable categories. All of the selected proteins employed in this study were characterized by X-ray crystallography with 3.5 Å or better resolution. Membrane proteins are known to be complicated targets for structure determination and were excluded. Therefore, according to their environmental differences, a total of 21 cysteine-rich proteins were randomly collected. This further resulted in a sample of 15 distinct types of non-membrane proteins. The PDB codes and the corresponding protein names for each are listed in Table 1. Moreover, based on the literature data available for each, the 15 non-membrane proteins were classified according to common functions into nine different categories. These included cysteine proteases, phosphatases, metabolic enzymes, kinases, interleukins, transcription factors, motility, virus capsid proteins, and ribosomes.

All sequence data were downloaded, and PERL scripts (www.perl.org) were used to count the cysteines and the counts (length) were reported. Several methodical approaches have been proposed for such a study. They include analysis of amino acid characteristics of spatial neighbors to the target residue (free cysteine, Cys_SH) which is measured within a 3.7 Å radius sphere with the sulfur atom of the cysteine residue as the center point, analysis of the hydrophobicity distribution around the target residue (free cysteine, Cys_SH), and structure-based threading. When choosing the sphere radius of 3.7 Å, we took into account that spheres should, cover as much of the space between the atoms as possible but, then again, we would not want the spheres to overlap too strongly.

The secondary structure classes from the HSSP (Dodge *et al.*, 1998) files were grouped in the following ways: helices were defined by the class G, I and H, strands by B and E, turns by T and bends by S (Kabsch and Sander, 1983). Geometric alignments and backbone superposition processes of two protein structures were made using the log procedure of program O (Jones *et al.*, 1991).

2.1.2. Coronaviruses proteinases

The main coronaviruse proteinases were extracted from the public DDBJ/EMBL/GenBank database (abbreviations in parentheses): SARS 3CLpro coronavirus (SARS, PDB ID code 1UJ1), human coronavirus 229E (229E, PDB ID code 1P9S), transmissible gastroenteritis virus (TEGV, PDB ID code 1LVO), human coronavirus OC43 (OC43), bovine coronavirus (BCoV), murine hepatitis virus (MHV), porcine epidemic diarrhea virus (PEDV), avian infectious bronchitis virus (IBV), and feline infectious peritonitis virus (FIPV). Multiple sequence alignment of the nine coronavirus proteinases with their homolog was performed using the CLUSTAL W program (Thompson *et al.*, 1994). Those selected query sequences were characterized as: cysteine, identical, hydrophobic, small, and charge/hydrophilic similar amino acid residues. Accordingly, through comparative sequence analysis, we examined their identities, similarity and differences.

In addition, more precise distributions of residues around the sulfur atom of free cysteine were analyzed.

3. Results and discussions

3.1. Environments around free cysteines

In Table 1, we listed the PDB codes and names of 15, non-membrane proteins respectively as well as their concise clarifications

Table 1
Classification results for 13 non-membrane proteins.

PDB	Protein	#aa (total)	#aa (hydrophobic)	#C (total)	#C (free)	Secondary structure	Surrounding residues (3.7 Å)																			
							L	F	V	A	I	P	T	D	S	H	Y	K	G	W	E	R	C	Q	N	M
1UJ1	3CL protease (SARS coronavirus)	306	148	12	12	α/β	5	3	3	1	2	2	0	1	1	2	1	2	0	0	0	0	0	0	0	0
1P9S	3CL protease (229E)	300	143	8	8	β	2	0	2	1	1	2	0	0	0	1	0	1	1	0	0	1	0	1	0	0
1LVO	3CL proteass (TGEV)	302	136	6	6	β	1	0	1	1	1	2	0	1	0	0	0	0	0	0	1	0	0	0	1	0
1AAX	PTP1B (*Homo sapiens*)	321	143	5	5	α	3	1	0	1	0	2	0	0	1	0	2	0	0	0	0	0	0	0	0	0
1CWS	CDC25B (*Homo sapiens*)	211	97	7	7	α/β	0	1	0	0	0	0	0	0	2	0	1	0	1	0	1	0	0	0	0	0
1FRB	FR-1 (*Mus musculus*)	315	150	6	6	α/β	2	0	1	1	0	0	2	1	0	2	0	0	0	0	0	0	0	0	0	0
1PD5	Chloramphenicol acetyltransferase (*Escherichia coli*)	219	110	5	5	α/β	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1FGI	Fibroblast growth factor receptor 1 (*Homo sapiens*)	310	151	5	5	α/β	2	1	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
1MD6	Interleukin 1 (*Mus musculus*)	154	79	5	3	β	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0
1K1A	B-cell lymphoma 3-encoded protein (*Homo sapiens*)	241	115	4	4	α	0	0	0	1	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0
1FHG	TELOKIN (*Meleagris gallopavo*)	754	337	5	5	β	0	1	1	0	1	0	0	2	0	0	1	1	1	0	0	0	2	0	0	0
1MQT_A	Polyprotein/coat protein (Swine vesicular disease virus)	283	124	3	3	β	0	2	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
1MQT_B		261	131	8	8	β	1	0	1	4	2	1	2	1	0	0	0	0	0	0	0	1	1	0	2	0
1MQT_C		238	123	7	7	β	1	3	1	0	0	0	0	1	1	0	0	0	1	2	0	0	0	0	0	0
1VI6	3OS ribosomal protein	208	104	1	1	β	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total				87	85		20	13	12	11	11	10	8	7	7	6	5	5	4	2	3	3	3	3	3	1

that are used throughout the following discussion of our results. As shown in Table 1 columns, total aa, total hydrophobic aa, C total, C free, second structure (the locations of cysteine residues), and the residues surrounding cysteine in a 3.7 Å radius sphere are addressed. Table 1 shows a total of 87 cysteine residues collected from the non-membrane protein data sets, and those were further divided into two forms: 2 disulfide-bonding cysteines (Cys_SS), and 85 free cysteines (Cys_SH).

Our data show that free cysteines of non-membrane proteins prefer a β strand environment. For non-membrane proteins: the hydrophobic residues such as leucine, valine, isoleucine and alanine were more frequently seen in the spatial neighborhood around free cysteines; the same was observed for the aromatic phenylalanine residue. Thus, the sequential differences in the positions between Cys_SH and its local neighborhood among the non-membrane protein data sets suggest that the surrounding residues are mostly hydrophobic (Table 1). Moreover, free cysteines, among non-membrane proteins, have a high number of leucine contacts.

3.2. *Relative position of the complement in the sequence among coronvirus proteases*

In an attempt to determine information about the uniqueness of cysteine within 3CL cysteine proteases, we compared SARS 3CLpro to other coronavirus proteases. Thus, we applied a structure–base sequence alignment of these nine coronavirus main proteinases to identify any spatial correspondences involving cysteine among them. Residues comprising the nine coronavirus main proteases are illustrated in Fig. 1. It can be observed that this multiple sequences alignment figure shows a strong conservation of hydrophobic residues (valine, leucine, isoleucine, alanine phenylalanine and proline) and small residues (serine and glycine) in proximity to cysteine residues. The presences of 28 cysteine residues (around 9%) in well-conserved positions were also noticed. This indicates that these homologous proteins had a higher proportion of cysteine residues than others. Moreover, our findings show that the environment of free cysteines is rather hydrophobic and are in fair agreement with the results reported by

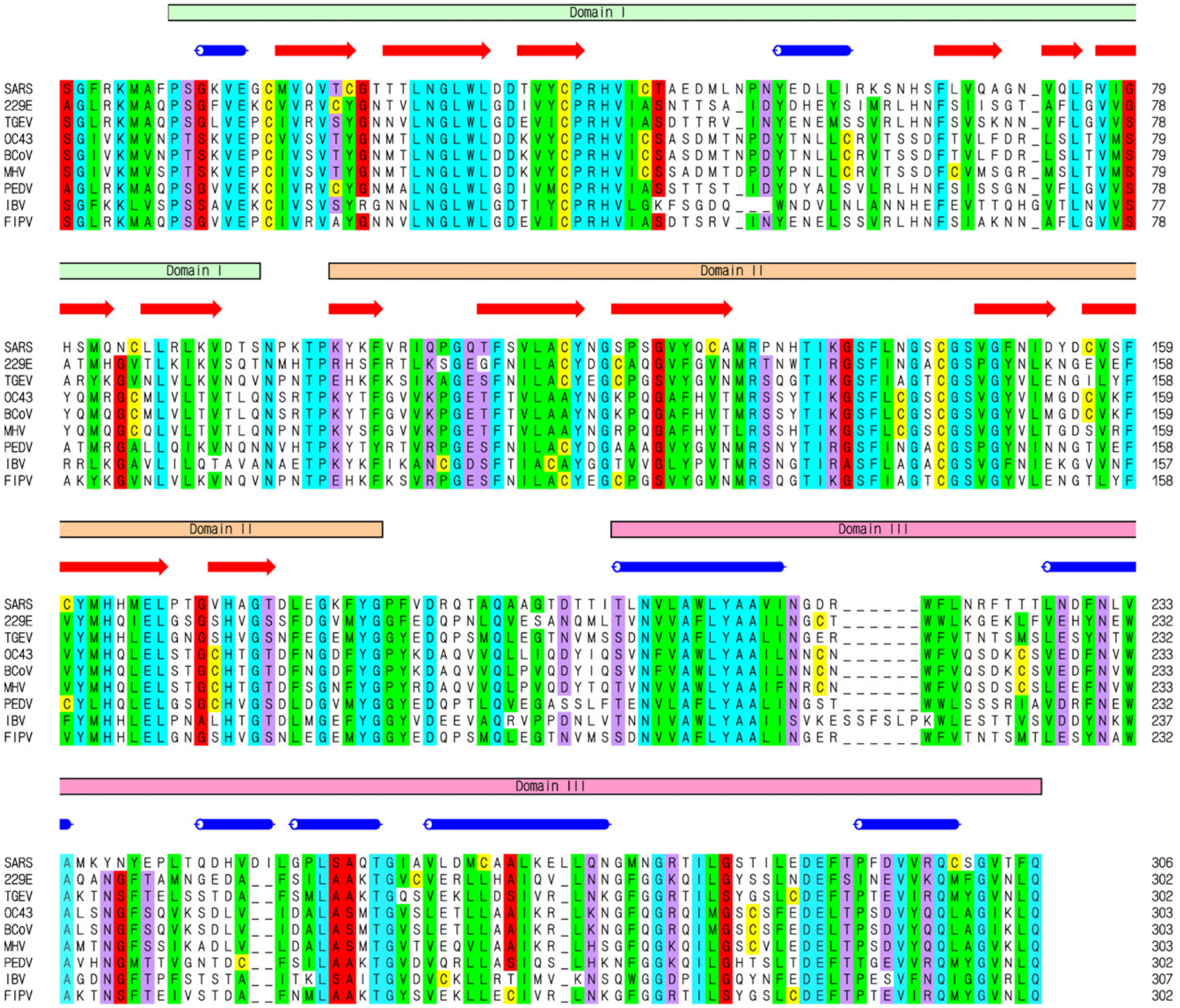


Fig. 1. Sequence alignment of nine coronaviruses main proteinases. The corresponding sequences of SARS (3CLpro), 229E, TGEV, OC43, BCoV, MHV, PEDV, IBV, FIPV were derived from the replicate polyproteins of the respective viruses whose sequences are deposited at the DDBJ/EMBL/GenBank database (accession nos: SARS, gi|29837498; 229E, gi|30024078; TGEV, gi|30146762; OC43, gi|50844478; BCoV, gi|26008084; MHV, gi|25121563; PEDV, gi|30138155; IBV, gi|25121547; FIPV, gi|37999875). The α-helices, β-strands and the domains as revealed in the SARS 3CLpro crystal structure are shown above the sequence alignment. The alignment was produced using CLUSTAL W (Thompson *et al.*, 1994). Colored outlines indicate cysteine, identical, hydrophobic, small, and charge/hydrophilic similar amino acid residues, respectively, (yellow, blue, green, red, purple). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

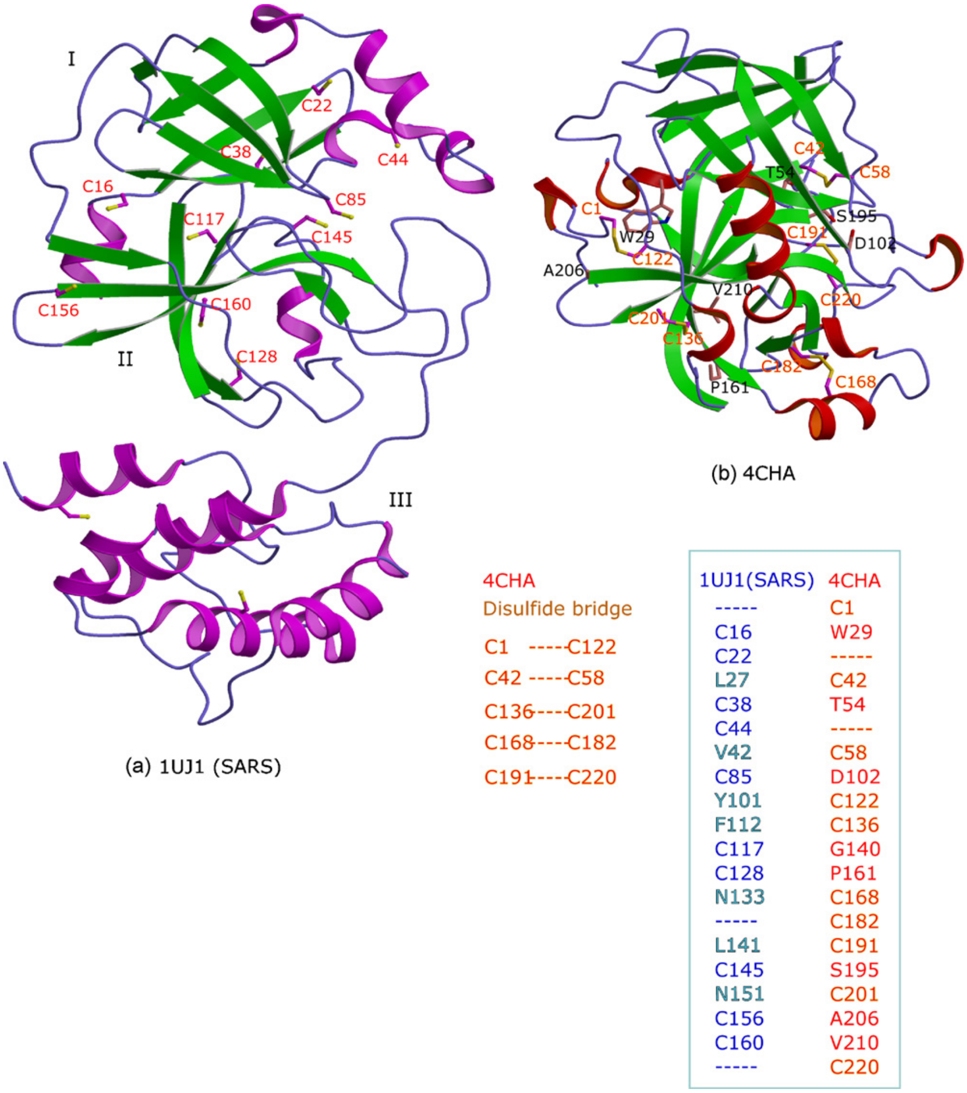
Fiser *et al*. (1992) and Muskal *et al*. (1990), who showed that hydrophobic residues accumulated in the vicinity of free cysteines.

3.3. Comparison of associate residues between SARS 3CLpro and α-chymotrypsin protein

SARS 3CLpro (PDB ID code 1UJ1) forms a dimer with the two promoters oriented almost at right angles to each others. Initially, we identified SARS 3CLpro domains that somewhat matched the α-chymotrypsin domains. Fig. 2 shows that each monomer is folded into three domains, the first two of which are antiparallel β-barrels and together resemble the architecture in serine proteinases of the chymotrypsin family (PDB ID code 4CHA). Domain II of SARS 3CLpro is smaller than domain I and also smaller than the homologous domain II of α-chymotrypsin. As the quantitative results show the RMSD among 110 atoms is 1.906, including domain I and domain II for these two structures with a limit of 3.8 Å. With a limit of 2.0 Å, the RMSD result among 48 atoms for domain I is 0.990, and the RMSD result among 40 atoms for domain II is 1.156. These results show that these two proteins are barely different in conformation. We have highlighted the side-chain of cysteines and the residues with the spatially equivalent residue positions of cysteine in the two proteins (Fig. 2(b)). α-Chymotrypsin has total eight residues (tryptophan, threonine, aspartic acid, glycine, proline, serine, alanine, valine) in the corresponding positions (cysteine residues in SARS 3CLpro), whereas SARS 3CLpro has total seven residues (leucine, valine, tyrosine, phenylalanine, asparagines) in the corresponding position (cysteine residues in α-chymotrypsin) Apparently, most of the spatially equivalent residues are subject to hydrophobic which may be somehow involved in catalysis.

3.4. Propensity to be clustered with hydrophobic residues

A superimposition (stereo image) of the structures of SARS 3CLpro demonstrates that cysteine residues not only favor positioning in a hydrophobic environment but also develop hunched posture in the surroundings of aromatic residues, see Fig. 3.



4CHA
Disulfide bridge

C1 -----C122
C42 -----C58
C136----C201
C168----C182
C191----C220

1UJ1(SARS)	4CHA
-----	C1
C16	W29
C22	-----
L27	C42
C38	T54
C44	-----
V42	C58
C85	D102
Y101	C122
F112	C136
C117	G140
C128	P161
N133	C168
-----	C182
L141	C191
C145	S195
N151	C201
C156	A206
C160	V210
-----	C220

Fig. 2. A MOLSCRIPT diagram showing the SARS 3CLpro monomer and α-chymotrypsin (PDB code 4CHA) structures (a) a monomer of SARS 3CLpro is presented as ribbons, and cysteines exposed. It contains two β-barrel domains and the α-helical C-terminal domain. The β-barrels of each I and II are composed of 6-standard β-sheets of domain. Domain III is composed mainly α-helices. The first two of which are antiparallel β-barrels reminiscent of those found in the chymotrypsin family (b) an α-chymotrypsin (PDB code 4CHA) presented as ribbon and side-chain of certain residues are highlighted to the corresponding cysteins of SARS 3CLpro. Domain II of SARS 3CLpro is not only smaller the domain I but also smaller than the homologous domain II of α-chymotrypsin.

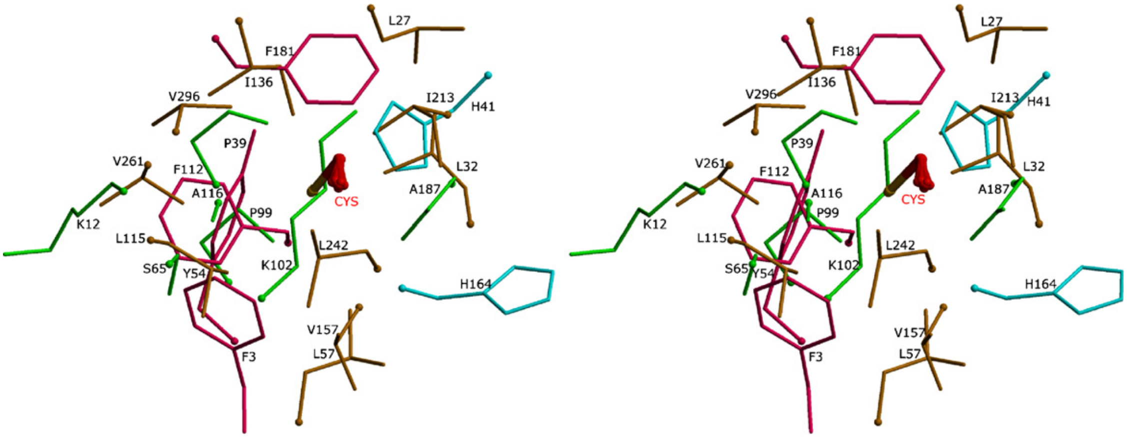
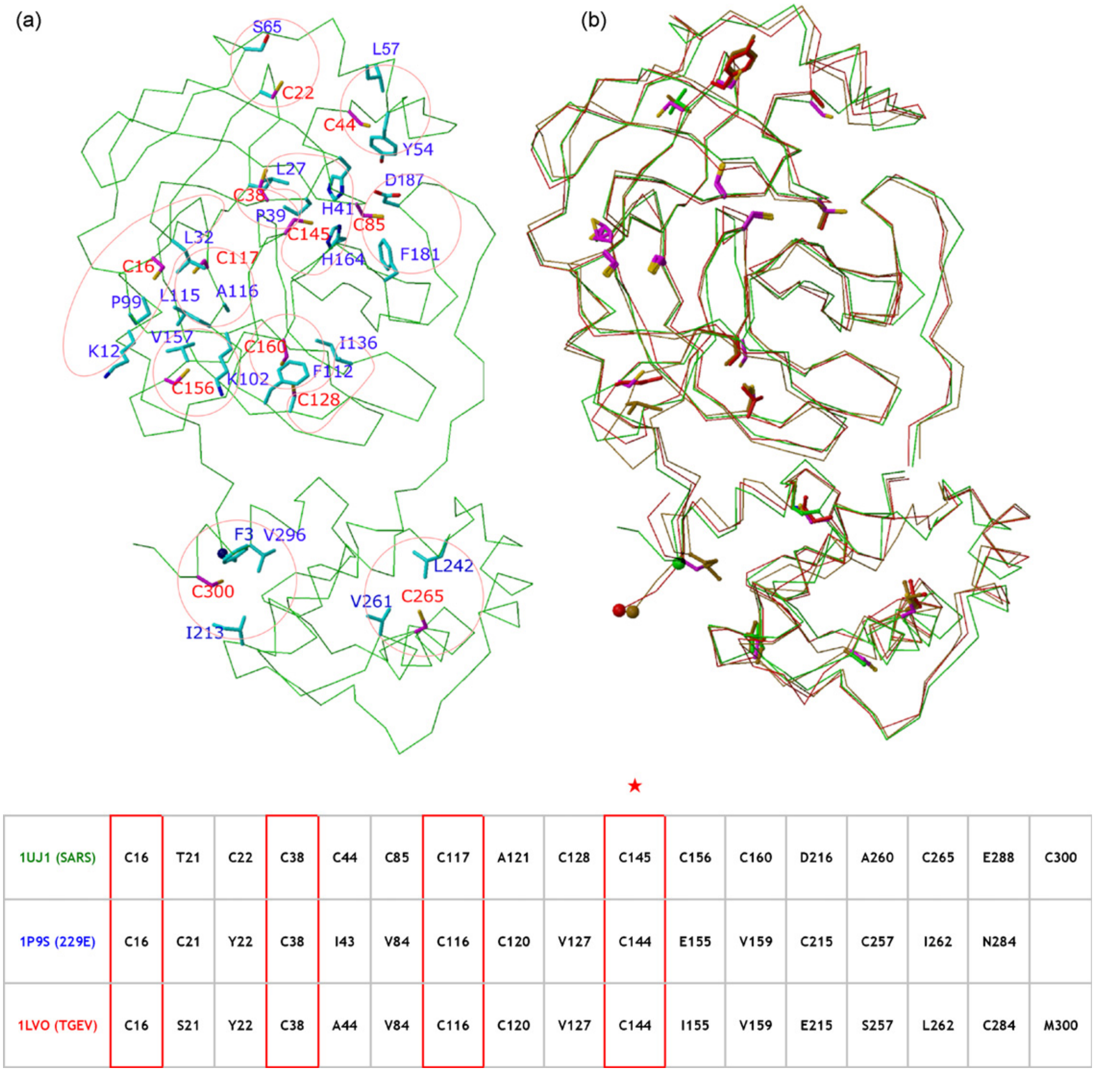


Fig. 3. Stereo image of superimposition of cysteines of SARS 3CLpro (PDB code 1UJ1). The residues surrounding cysteine in a 3.7 Å radius sphere are Phe, Tyr (in purplepink color), His (color in cyan), Leu, Ile, Val (in gold color), Ala, Pro, Ser, Lys (in green color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)



1UJ1 (SARS)	C16	T21	C22	C38	C44	C85	C117	A121	C128	C145	C156	C160	D216	A260	C265	E288	C300
1P9S (229E)	C16	C21	Y22	C38	I43	V84	C116	C120	V127	C144	E155	V159	C215	C257	I262	N284	
1LVO (TGEV)	C16	S21	Y22	C38	A44	V84	C116	C120	V127	C144	I155	V159	E215	S257	L262	C284	M300

Fig. 4. Comparison of the structure SARS 3CLpro (PDB code 1UJ1) with 229E (PDB code 1P9S) and TGEV (PDB code 1LVO). Figures were generated using MOSCRIPT. (a) Cα-trace of SARS 3CLpro monomer. Residues of cysteine and those surrounding cysteine of 3.7 Å are indicated. (b) Cα-trace of SARS 3CLpro (green) superimposed on 229E (blue) and TGEV (red). The backbone structure was locally aligned with each other in two regions. Side-chains of certain residues are highlighted for 229E and TGEV that are identical or that are in the close proximity to the corresponding cysteines of SARS 3CLpro. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Cysteines and the residues surrounding cysteine in a 3.7 Å radius sphere were identified in a Cα-trace outline for SARS 3CLpro monomer (PDB ID code 1UJ1). See Fig. 4(a). This figure shows that as the hydrophobic residues such as leucine, valine, phenylalanine, and isoleucine were more frequently seen in the spatial neighborhood around free cysteines. SARS 3CLpro (PDB ID code 1UJ1) was then superposed with the other two 3CL proteases 229E (PDB ID code 1P9S) and TGEV (PDB ID code 1LVO). These three proteins are similar to each other and can be superposed well (RMSD 1.12–1.18). Fig. 4(b) shows the superimposed Cα profile of SARS 3CLpro, 229E and TGEV. The side-chains are also highlighted on the spatially conserved residues which correspond to the residue of the free cysteine. The spatial corresponding residues to free cysteines within each are serine, tyrosine, alanine and valine. These residues are well conserved according to a Risler matrix (Risler *et al.*, 1988).

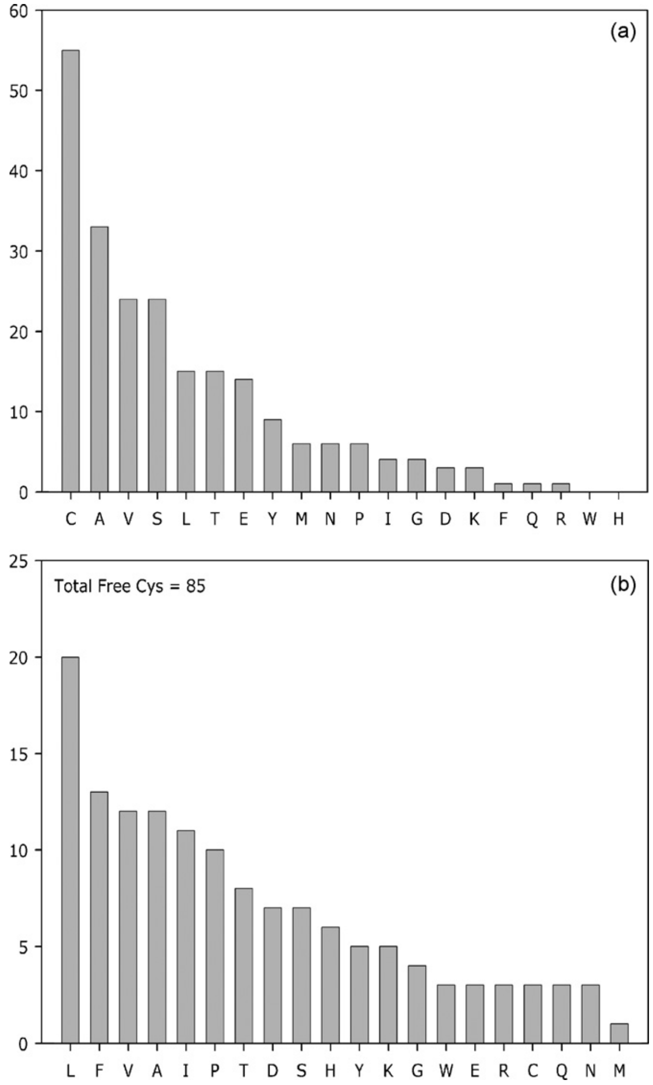


Fig. 5. Plots of occurrences of residues. (a) 28 cysteines conserved locations were observed of the sequence alignment of 9 coronaviruses (3CLpro) main proteinases. An occurrence of alternative residues in the cysteine conservation is revealed. This indicates cysteines are highly conserved in hydrophobic region, and smaller residues are favorite substitution. (b) Discrepancy in occurrences of residues which embedding cysteine (a distance less than 3.7 Å, sidechain-S of cysteine is as the center) for various classes of proteins (cysteine proteases, phosphatases, kinases, interleukins, transcription factors, motility proteins, metabolic enzymes, virus capsid proteins, and ribosomes). This indicates that either hydrophobic residues or small residues were cluster with cysteines.

3.5. Analysis of free cysteine distributions of SARS 3CL proteases, and non-membrane protein environments

Proteins having similar functions but from different sources can be identified by their sequences. A statistical analysis of the amino acids frequencies associated with nine 3CLpro alignment sequences has revealed the results shown in Fig. 5(a).

Fig. 5(a) shows the frequency of occurrences of alternative residues in the cysteine conservation. Our findings reveal that alanine, valine, serine, leucine, and threonine were more frequently observed among these nine coronaviruses (3CLpro) main proteinases. Thus, we can conclude:

(1) the cysteines are frequently found in hydrophobic regions containing alanine, valine, and leucine;
(2) small residues (serine and threonine) are favorable to be substituted with cysteine residues. These residues are relative smaller than the existing cysteine, thereby producing conservative changes that would not disrupt the native structure.

We have analyzed the distribution of residues around the sulfur atom of cysteine. It has been verified that the top eight occurrences of residues embedding cysteine (in a radius of a 3.7 Å sphere) in non-membrane proteins were leucine, phenylalanine, valine, alanine, isoleucine, proline, threonine and serine. These are classified as either hydrophobic or small residues (see Fig. 5(b)).

3.6. Interaction with aromatic rings of non-membrane proteins

It has been reported that in order to stabilize interactions involving cysteine residues that the free sulfhydryl group prefers to interact closely with the face of aromatic rings (Klingler and Brutlag, 1994; Pal and Chakrabarti, 1998). Thus, our data show that, based on the 3D spatial orientation, the aromatic rings of a selected residue, *i.e.* phenylalanine are most likely to be in contact with the sulfur atom of cysteines. We carefully examined some other non-membrane proteins. Approximately 62% and 60% of the aromatic ring faces of phenylalanine and tyrosine, respectively were clustered contiguously to the sulfur atom of cysteins, contrary to the way tryptophan (W) behaved. See Fig. 6. This might be because tryptophan not only has the largest nonpolar accessible surface area, but also observed less frequently in our

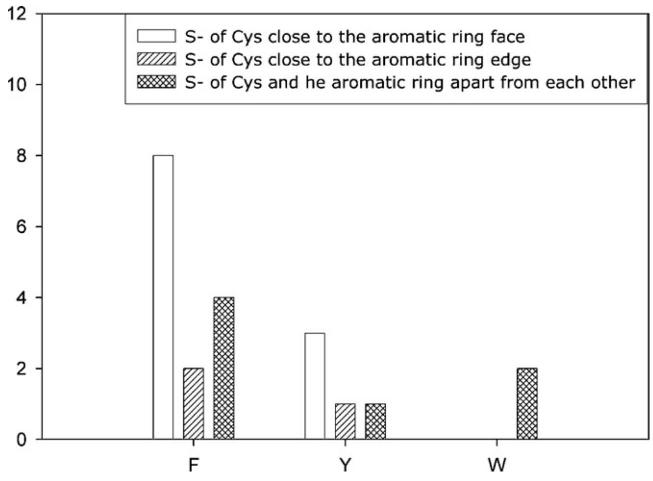


Fig. 6. Distribution of the relative position between sulfur atom of cysteine and the aromatic ring of F, Y, W in a 3.7 Å radius sphere (non-membrane proteins). The aromatic ring faces of phenylalanine and tyrosine respectively were clustered contiguously to the sulfur atom of cysteins, contrary to the way tryptophan (W) behaved.

selected proteins. In this regard, further investigation to provide a quantitative explanation is needed.

3.7. Relative location of free cysteine residues

From the data collected, we observed that approximately 23% of free cysteine residues of non-membrane proteins were located in the α-helix and 77% were found in the β-strand (Table 1). This observation correlates positively with a tendency for cysteine to occur preferentially in the β-strand (Williams *et al*., 1987; Wilmot and Thornton, 1988).

4. Conclusions

Knowledge of the details of protein structures offer a range of possibilities for investigating their biological functions (Baker *et al*., 2003; Eisenstein *et al*., 2000; Teichmann *et al*., 2000). Conserved residues mapping on certain characteristics may likewise identify a key site or perhaps a sensible function. Accordingly, the approaches we have presented were based on the structures. Although spatial contacts have been studied to derive contact potentials for the different amino acid interactions (Brocchieri and Karlin, 1995; Miyazawa and Jernigan, 1996, 1999), the common strategy is to study the number of contacts within a given distance cut-off. To our knowledge structural bioinformatics detection of the cysteines has not been attempted previously. This study has revealed, for the first time, the discrimination of structural versus active free cysteines based on local environment analysis. The computational prediction and annotation of free cysteines in the protein environments has been described through analysis of 3D spatial correspondences.

Essentially we have demonstrated the corresponding structural positions associated with free cysteines in their three-dimensional environment and the frequency of occurrence of the residues surrounding the free cysteines in selected proteases. The types of residues involved in spatial contacts with free cysteines of non-membrane proteins found in the present study indicated that free cysteines have a higher capacity for contacts to hydrophobic residues and lower capacity for contacts to polar/charged residues.

We also examined nine sequenced coronavirus proteases including three primarily coronavirus proteases (SARS 3CL, 229E, TGEV) whose structures have been solved. For these it was shown that the sequential environments around free cysteines were rather hydrophobic.

The use of combined 3CL main proteases and cysteine-rich proteins (membrane/non-membrane proteins) database mining approaches allowed for the classification of free cysteines in proteins. The validity of this approach was supported by the identification of some known proteases. The identification and functional characterization of the free cysteines will have implications in many aspects of biology. Moreover the sets of these proteins and the knowledge-based methods used to identify them will form the foundation in the algorithms used for detection, in particular, within the protein sequence.

Acknowledgement

The authors are truly grateful to the Academic Sinica, R.O.C. for supporting this study.

References

Baker, E. N., V. L. Arcus, and J. S. Lott, "Protein Structure Prediction and Analysis as a Tool for Functional Genomics," *Appl. Bioinform.*, **2**, S3 (2003).

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, **28**, 235 (2000).

Brocchieri, L. and S. Karlin, "How Are Close Residues of Protein Structures Distributed in Primary Sequence?" *Proc. Natl Acad. Sci. U.S.A.*, **92**, 12136 (1995).

Dodge, C., R. Schneider, and C. Sander, "The HSSP Database of Protein Structure-Sequence Alignments and Family Profiles," *Nucleic Acids Res.*, **26**, 313 (1998).

Eisenstein, E., G. L. Gilliland, O. Herzberg, J. Moult, J. Orban, R. J. Poljak, L. Banerjei, D. Richardson, and A. J. Howard, "Biological Function Made Crystal Clear—Annotation of Hypothetical Proteins via Structural Genomics," *Curr. Opin. Biotechnol.*, **11**, 25 (2000).

Fiser, A., M. Cserzö, E. Tüdos, and I. Simon, "Different Sequence Environments of Cysteines and Half Cysteines in Proteins: Application to Predict Disulfide Forming Residues," *FEBS Lett.*, **302**, 117 (1992).

Jones, T. A., J.-Y. Zou, S. W. Cowan, and M. Kjeldgaard, "Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in These Models," *Acta Crystallogr.*, **A47** (Pt 2), 110 (1991).

Kabsch, W. and C. Sander, "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features," *Biopolymers*, **22**, 2577 (1983).

Klingler, T. M. and D. L. Brutlag, "Discovering Structural Correlations in Alpha-Helices," *Protein Sci.*, **3**, 1847 (1994).

Mccaldon, P. and P. Argos, "Oligopeptide Biases in Protein Sequences and Their Use in Predicting Protein Coding Regions in Nucleotide Sequences," *Proteins*, **4**, 99 (1988).

Miyazawa, S. and R. L. Jernigan, "Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading," *J. Mol. Biol.*, **256**, 623 (1996).

Miyazawa, S. and R. L. Jernigan, "Evaluation of Short-Range Interactions as Secondary Structure Energies for Protein Fold and Sequence Recognition," *Proteins*, **36**, 347 (1999).

Muskal, S. M., S. R. Holbrook, and S.-H. Kim, "Prediction of the Disulfide-Bonding State of Cysteine in Proteins," *Protein Eng.*, **3**, 667 (1990).

Nilsson, B. L., M. B. Soellner, and R. T. Raines, "Chemical Synthesis of Proteins," *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 91 (2005).

Pal, D. and P. J. Chakrabarti, "Different Types of Interactions Involving Cysteine Sulfhydryl Group in Proteins," *Biomol. Struct. Dyn.*, **15**, 1059 (1998).

Risler, J. L., M. O. Delorme, H. Delacroix, and A. Henaut, "Amino Acid Substitutions in Structurally Related Proteins. A Pattern Recognition Approach. Determination of a New and Efficient Scoring Matrix," *J. Mol. Bio.l*, **204**, 1019 (1988).

Teichmann, S. A., C. Chothia, G. M. Church, and J. H. Park, "Fast Assignment of Protein Structures to Sequences Using the Intermediate Sequence Library Pdb-Isl," *Bioinformatics*, **16**, 117 (2000).

Thompson, J. D., D. G. Higgins, and T. J. Gibson, "Clustal W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice," *Nucleic Acids Res*, **22**, 4673 (1994).

Williams, R. W., A. Chang, D. Juretić, and S. Loughran, "Secondary Structure Predictions and Medium Range Interactions," *Biochim. Biophys. Acta.*, **916**, 200 (1987).

Wilmot, C. M. and J. M. Thornton, "Analysis and Prediction of the Different Types of Beta-Turn in Proteins," *J. Mol. Biol.*, **203**, 221 (1988).

Zvelebil, M. J., G. J. Barton, W. R. Taylor, and M. J. Sternberg, "Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences," *J. Mol. Biol.*, **195**, 957 (1987).