Disulfide bonds, their stereospecific environment and conservation in protein structures

Rajasri Bhattacharyya, Debnath Pal[1] and Pinak Chakrabarti[2]

Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Calcutta 700 054, India

[1]Present address: UCLA–DOE Institute for Genomics and Proteomics, Box 951570, University of California at Los Angeles, Los Angeles, CA 90095-1570, USA

[2]To whom correspondence should be addressed.
E-mail: pinak@boseinst.ernet.in or pinak_chak@yahoo.co.in

We studied the specificity of the non-bonded interaction in the environment of 572 disulfide bonds in 247 polypeptide chains selected from the Protein Data Bank. The preferred geometry of interaction of peptide oxygen atoms is along the back of the two covalent bonds at the sulfur atom of half cystine. With aromatic residues the geometries that direct one of the sulfur lone pair of electrons into the aromatic π-system are avoided; an orientation in which the sulfide plane is normal or inclined to the aromatic plane and on top of its edge is normally preferred. The importance of the S···aromatic interaction is manifested in the high degree of its conservation across members in homologous protein families. These interactions, while providing extra overall stability to the native fold and reducing the accessibility of the disulfide bond and thereby preventing exchange reactions, also set the orientation of the conserved aromatic rings for further interactions and binding to another molecule. The conformational features and the mode of interactions of disulfide bridges should be useful for molecular design and protein engineering experiments.

Keywords: conservation of interaction/disulfide bond/protein stability/S···aromatic interaction/S···O interaction

Introduction

The specificity of folding is provided by non-bonded interactions, such as hydrogen bonding. Increasingly, other specific interactions, such as aromatic–aromatic, X–H···π (X = N, O or C), C–H···O, etc., are being recognized (Singh and Thornton, 1985; Burley and Petsko, 1988; Umezawa and Nishio, 1998; Samanta *et al*., 1999, 2000; Brandl *et al*., 2001; Steiner and Koellner, 2001; Bhattacharyya *et al*., 2002, 2003; Thomas *et al*., 2002; Bhattacharyya and Chakrabarti, 2003). Although hydrophobic interactions are known to contribute to the stability of the native fold (Dill, 1990), analysis of the geometry of interactions between Pro and aromatic residues shows that some specific geometries occur in numbers significantly greater than expected, indicating that it is not just the burial of non-polar surface but the existence of specific interactions which is also of importance (Bhattacharyya and Chakrabarti, 2003). The telltale sign of such interactions is their directional nature and, although weaker than the conventional hydrogen bonds, as they are numerous, they can contribute to the stability of the local structure (Bhattacharyya *et al*., 2002), and also the overall fold of the protein molecule (Burley and Petsko, 1988; Samanta *et al*., 2000).

The existence of covalent bonds in the form of disulfide linkages contributes to the stability and is especially important for extracellular proteins (Thornton, 1981). There have been analyses of residues around half cystines in protein sequences and methods to identify them (Fiser *et al*., 1992; van Vlijmen *et al*., 2004). The cross-linking of the polypeptide chain has been used to engineer additional conformational stability into proteins by site-directed mutagenesis (Matsumura *et al*., 1989; Clarke and Fersht, 1993; Mansfeld *et al*., 1997). However, the success of this strategy has not been universal (Mitchinson and Wells, 1989; Betz, 1993; van den Burg *et al*., 1993; Hinck *et al*., 1996), suggesting that some of the engineered bonds may have strained conformation (Katz and Kossiakoff, 1986) or the specific, but as yet not fully understood, environment of the bonds in wild-type proteins has an associated stabilizing role, which may be lacking at the engineered sites. The contribution of the environment in the form of the interaction of Cys with hydrophobic residues has been suggested to give rise to an intermediate with a single disulfide bond in the early stages of folding of BPTI (Dadlez, 1997), which contains a total of three such bonds (Darby and Creighton, 1993). A highly conserved disulfide–Trp interaction has been observed in the immunoglobulin fold (Ioerger *et al*., 1999).

The thiol group in Cys and the sulfide group in Met are not very adept at forming hydrogen bonds (Ippolito *et al*., 1990; Gregoret *et al*., 1991; Allen *et al*., 1997), but still sulfur can partake in a number of non-bonded interactions, such as that involving aromatic residues (Desiraju and Steiner, 1999; Meyer, *et al*., 2003). In particular, Rosenfield *et al*. (1977) have observed, from an analysis of the environment of a divalent S atom (Y–S–Z) in organic and inorganic crystals, that a nucleophilic O atom tends to approach the S atom from the backside of S–Y and S–Z bonds (i.e. along an antibonding orbital) to make the S···O interaction a stabilizing one. In spite of the presence of myriad other non-covalent interactions in proteins, the directional preference of the S···O interaction was essentially maintained in the environment of the S atom of Met (Pal and Chakrabarti, 2001; Iwaoka *et al*., 2002a). Moreover, the interaction has also been shown to regulate enzymatic function; for example, there exists an S···O interaction between the sulfur atom of *S*-adenosylmethionine and the carboxylate group of Asp118 in *S*-adenosylmethionine synthetase (Taylor and Markham, 1999). The concept of electrophile–nucleophile interaction was extended to include the π-electron system of aromatic rings acting as nucleophile and, again, there was marked directionality (Pal and Chakrabarti, 2001). In this paper, we address the question of whether a similar geometric relationship is retained in the interaction of the disulfide group with carbonyl oxygen atoms (of which we consider only the main-chain ones) and the aromatic rings. The importance of an

interaction, especially when it involves a side chain, can also be discerned by finding out the degree of conservation of the participating residues during evolution, something we also studied. Although the general characteristics of disulfide bonds have been enumerated in numerous studies (Richardson, 1981; Thornton, 1981; Srinivasan *et al*., 1990; Morris *et al*., 1992; Petersen *et al*., 1999), such features need to be constantly refined using larger databases. Therefore, while addressing our primary concern of the environment of disulfide moiety, we also analyzed the conformations, primary and secondary structural features of the half-cystines making up the disulfide bond and in this process identified some folding patterns.

Materials and methods

Atomic coordinates were obtained from the Protein Data Bank (PDB) at the Research Collaboratory for Structural Bioinformatics (RCSB) (Berman *et al*., 2000). A total of 1266 polypeptide chains were selected using PISCES (Wang and Dunbrack, 2003) from PDB files (as of August 2003) with an *R*-factor $\leqslant$20% and resolution of $\leqslant$2.0 Å and sequence identity <25%. Of these, 247 chains contain one or more disulfide bonds. Only those Cys and aromatic residues and carbonyl groups were considered for which the fractional occupancies and temperature (or *B*) factors (of S^{γ} atom of Cys and all ring atoms of aromatic residues and carbonyl oxygen atoms) were 1.00 and $\leqslant$30 $Å^2$, respectively. Disulfide bonds were identified using a cut-off distance of 2.3 Å between the S^{γ} atoms of two Cys residues. In PDB files disulfide bonds are also specified with the 'SSBOND' record. However, in five proteins [1cru_A: 338–345 (distance 2.89 Å), 1fjs_L: 89–100 (2.35 Å), 1p5u_A: 98–137 (2.31 Å), 1ubk_L: 84–549 (2.87 Å), 2sic_I: 35–50 (2.46Å)] the specified disulfide bonds have a larger S^{γ}–S^{γ} distance. These disulfide bonds were excluded from our analysis.

To ascertain the relative importance of the non-bonded contacts involving the sulfur atom with different kinds of atoms (aromatic and aliphatic carbon atoms, main-chain carbonyl oxygen atoms and all types of side-chain oxygen atoms), a density function, ρ, was determined for each type of contact at distances $r = 3.0, 3.1, \ldots, 5.5$ Å. A shell of width 0.1 Å was assumed at each distance and the number of atoms of a given type in it was found (for a residue in contact, even if more than one atom satisfied the criterion, only one was accepted). If the outer radius r_2 had N_2 occurrences and the inner radius r_1 had N_1, then

$$\rho = (N_2 - N_1)/\left[4/3\pi\left(r_2^3 - r_1^3\right)\right]$$

C^{β} atoms of aromatic residues and C^{β} onwards for aliphatic residues were considered as aliphatic carbon atoms.

For the calculation of the geometry, the centroids of aromatic residues were first determined (for His, Phe and Tyr, the center of mass of the five- or six-membered ring; for Trp, the mid-point of CD2 and CE2 atoms). A molecular axial system was defined with the origin at the centroid of the aromatic residue and the *z*-axis along the normal to the aromatic plane. The interplanar angle, *P*, and the angle θ made between the *z*-axis and the line joining the centroid of the aromatic residue to S_{γ} were computed (Figure 4a). For each interaction the geometry was placed in one of the elements in a 3×3 grid (each element spanning a range of 30° along *P* and θ) (Figure 4b). Each relative orientation is designated by a two-letter code (*fp*, *ot*, *en*, etc.). The first letter indicates if the aromatic residue is interacting with its face (*f*) or the S_{γ} of half cystine is near its edge (*e*) or located in an intermediate (offset or *o*) position. The second letter denotes if the sulfide plane (passing through C_{β}–S_{γ}–S_{γ}') is normal (*n*) or parallel (*p*) to the aromatic ring or has a tilted (*t*) orientation; these labels are slightly different from those in the earlier studies where both the interacting residues were aromatic (Samanta *et al*., 1999; Bhattacharyya *et al*., 2002, 2003). If O_{ij} is the observed frequency of occurrence in the grid element corresponding to the *i*th row and *j*th column (*i* and *j* varying from 1 to 3), the corresponding expected value (E_{ij}) can be calculated as the product of the sum of the observed numbers in the elements in the *i*th row and that in the *j*th column, divided by the total number of observations in all nine grid elements (Samanta *et al*., 1999).

To see the directionality of aromatic centroid or carbonyl oxygen atoms (X) relative to the sulfide plane, two additional parameters, θ_1 and φ were used (Figure 4c); θ_1 is the polar angle between the normal to sulfide plane and $S_{\gamma}\cdots X$ vector (if $\theta_1 > 90°$, θ_1 is made equal to $180° - \theta_1$, so that contacts above or below the plane are assumed to be equivalent; i.e., $0° \leqslant \theta_1 \leqslant 90°$). φ is the azimuthal angle between the extension of the bisector of the $\angle C_{\beta}$–S_{γ}–S_{γ}' and the projection of X in the disulfide plane.

The secondary structural elements were determined using the algorithm DSSP (Kabsch and Sander, 1983). Unless mentioned otherwise, we grouped all helix types (H, G, I) as H, β-strands (E, B) as E and turns (S, T) as T; C corresponds to non-regular structure. The solvent-accessible surface areas (ASA) of the Cys residues were computed with the program NACCESS (Hubbard, 1992), which implements the algorithm of Lee and Richards (1971). The relative accessibility of a Cys residue is the percentage ASA of the residue in the structure as compared with its ASA in an extended Ala–Cys–Ala tripeptide.

The degree of conservation of aromatic residues in a protein was found based on the amino acid usage at the same position in all homologous protein families, as delineated in the HOMSTRAD structural alignment database (Mizuguchi *et al*., 1998; Stebbings and Mizuguchi, 2004). Molecular diagrams were generated using MOLSCRIPT (Kraulis, 1991) and the scatterplot for the distribution of O atoms around S was drawn using the software IsoGen from the Cambridge Structural Database (Allen, 2002). In the text a PDB file is mentioned as the four-lettered code (in lowercase), with the chain identifier, if present, appended (in uppercase). The disulfide-linked Cys residues are designated Cys1 and Cys2 based on the sequential order.

Results

In 247 polypeptide chains, 572 disulfide bonds are present of which 556 are intra-chain and 16 are inter-chain. If one selects the cysteine and half-cystines after applying the temperature factor cut-off, the total number of Cys residues (free and half cystine) in 247 chains is 1269, which indicates that 90% of the Cys in these chains participate in disulfide linkages–hence free Cys residues are fairly rare in proteins containing disulfide bonds.

Of the 16 cases (in 11 PDB files) in which the disulfide bond links two chains, in five the molecule is a heterooligomer, and

except in three proteins [myeloperoxidase (length = 466 residues), renal dipeptidase (369) and serine carboxypeptidase (255)], the chain length is <100 residues.

A total of 305 Cys (27% of half cystines) from 222 disulfide bonds (i.e. 39% of the total) are involved in 357 Cys · · · Arom (Arom = aromatic residue) interactions. If we calculate the ratio of the observed to the expected number of Cys · · · Arom interactions (Bhattacharyya *et al*., 2003), the highest value, 1.5, is observed for Trp whereas for the other three aromatic residues, the values are 1.2 (Phe), 0.92 (Tyr) and 0.46 (His). Sixty-six aromatic residues [Phe (34), Tyr (16), Trp (9) and His (7)] interact with both of the half cystines.

A larger number (899) of Cys (79% of half cystines) are found to be involved in 1307 S · · · O (carbonyl) interactions, of which 1296 are intra-chain and 11 inter-chain. Out of 572 disulfide bonds, 538 (94%) are in interaction with carbonyl oxygens; 204 oxygen atoms are found to interact with both of the half cystines.

Sequence and structural features of disulfide bonds

In Figure 1, the number of intra-chain disulfide bonds per 100 residue long polypeptide chain is plotted against the chain length. Normalized in this way, one can fit a power function to the data. This number is high in small proteins and reaches a plateau when the chain is ∼200 residues long. Consideration of the relative accessibility of half cystines indicates that 50% of them are completely buried with relative accessibilities <5% (data not shown).

The distribution of the sequence difference between the half cystines (Δ) for intra-chain disulfide is presented in Table I. If the interactions with $\Delta \leqslant 5$ are considered local, only 7.9% of disulfide bonds are formed locally. Interestingly, there is no cystine with $\Delta = 2$. Some 49% of disulfide bonds are formed at higher values ($\Delta > 25$). Hence disulfide bonds stabilize the three-dimensional structure by holding together distant regions of the chains.

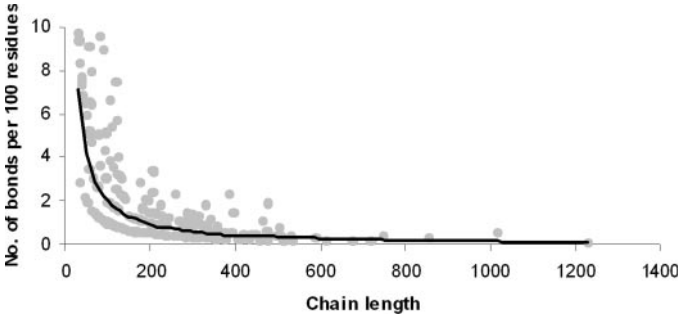

Fig. 1. Number of disulfide bonds per 100 residues plotted against the chain length. The line is a fit of a power function ($y = 294.7x^{-1.1}$, $R^2 = 0.62$) to the data.

Table I. The absolute value of the sequence difference, |Δ|, of the residues involved in intra-chain disulfide bond

	|Δ|					
	1–5	6–10	11–15	16–20	21–25	>25
No. observed[a]	44 (7.9)	74 (13.3)	71 (12.8)	61 (11.0)	34 (6.1)	272 (48.9)
χ_3 distribution[b]	17/27	31/43	30/41	29/32	24/10	151/121

[a]The percentage values are in parentheses.
[b]The numbers of cases with negative/positive values of χ_3 (Figure 2) are indicated.

A number of groups (Richardson, 1981; Thornton, 1981; Srinivasan *et al*., 1990; Morris *et al*., 1992; Petersen *et al*., 1999) have analyzed the $\chi_1 - \chi_3$ conformations of disulfide bonds. As is known, the distribution of the χ_3 torsion angle is not symmetric (Figure 2a), with clustering around the left-handed ($\chi_3 = -86.4 \pm 8.5°$) and right-handed conformers ($\chi_3 = 95.0 \pm 10.3°$). Whether the two half cystines are local or remote in sequence does not seem to have any effect on the sign of χ_3 (Table I). When the side-chain conformation angles (χ_1 and χ_2) of the half cystines are plotted, separated into two ranges in χ_3 and the secondary structural elements of the residues marked, one can see some interesting trends (Figure 2b and c). The most populated conformation has a negative value of χ_3 and those for χ_1 and χ_2 around $-60°$ for all the secondary structural states of the two residues. Between χ_1 and χ_2, the former is more restricted to a value around $-60°$ (the g^+ conformation) (Janin *et al*., 1978; Chakrabarti and Pal, 2001), while the latter can have values around 60 and 180° also. However, whereas χ_1 is centered on the canonical values of ± 60 and 180°, χ_2 shows considerable deviation. Considering Figure 2c, the average χ_2 values for the two half cystines taken together in the $\pm 60°$ regions are $+91(33)°$ and $-77(18)°$ [in Figure 2b, the most populated region has a χ_2 average of $-71(17)°$]. It is interesting that the preferred values of χ_2 torsion are very similar to the corresponding angle (C_α–C_β–S_γ–M) involving the cation (M) in metal-bound Cys residues, where angles are observed around ± 90 and 180° (Chakrabarti, 1989). The distribution of half cystines between Figure 2b and c is not different if the residues are in a turn or non-regular conformation (symbol: circle), but is roughly inverse if the secondary structural state is helical, as opposed to β-strand. The ratio of the observed numbers with a negative χ_3 value to that with a positive value is 1.41 when the secondary structure is helical (crosses), but 0.83 when it is β-strand (triangles). Another finding is that greater numbers of half cystines with χ_2 around 60° are observed (black symbols in Figure 2b and c) when χ_3 assumes a value that is on the outside edge of the distribution; this is especially noticeable in Figure 2b, indicating the possibility that such disulfide bonds have some inherent strains.

The combination of the secondary structural elements for the half cystines is presented in Table II. This shows that one of the elements being a β-strand is fairly common. In 62 out of 68 cases of 'EE' type, both the half cystines are in β-strands, which are mostly antiparallel. Also in 58%, the disulfide bond links two strands of a single β-sheet, whereas in others it connects two β-sheets. When the disulfide bond links two helices, the latter are antiparallel in 42% of cases, parallel in 16% and have intermediate orientations in the remaining cases. The highest number of observations is found not linking two regular secondary structures, but connecting a strand with a non-regular region. However, cases with two Cys residues with non-regular structure being linked by disulfide bonds are rare.

Distribution of aromatic residues and carbonyl oxygen atoms around half cystines

Figure 3 displays the density (defined in Materials and methods) of some atom types at different distances from S_γ of half cystine. The distribution for Cys · · · Arom interactions shows a peak at ∼3.7 Å, which falls to a minimum value at ∼4.3 Å, beyond which it rises slightly to reach a plateau. Similar

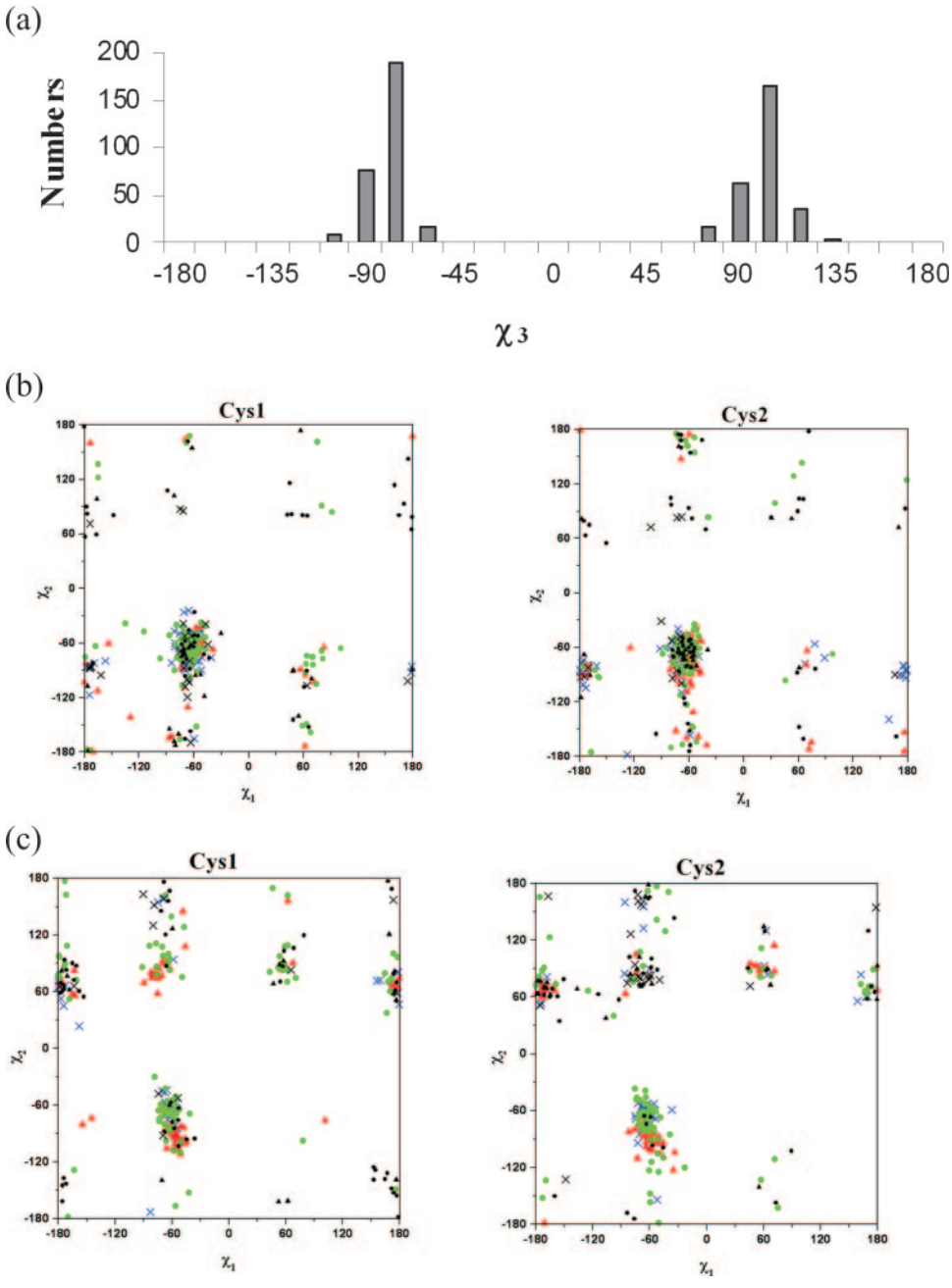

Fig. 2. (**a**) Distribution of χ_3 torsion angles of the disulfide bonds. χ_1, χ_2 plots for Cys1 and Cys2 when χ_3 has (**b**) a negative or (**c**) a positive value. The secondary structures of the residues are indicated as: helix (cross), strand (triangle) and others (circle), which are colored blue, red and green, respectively, if χ_3 is within the range -90 to $-70°$ or 90 to 120°, outside which the symbols are in black. In (a) the labels on the horizontal axis are the upper limits of the bins in the histograms.

Table II. Secondary structures of the cystines

	Secondary structure[a]									
	HH	HE	HT	HC	EE	ET	EC	TT	TC	CC
No. observed[b]	43 (7.7)	**64** (11.5)	43 (7.7)	**68** (12.2)	**68** (12.8)	38 (6.8)	**102** (18.3)	26 (4.7)	55 (9.9)	49 (8.8)

[a]The two letters indicate the secondary structure elements of the two half cystines.
[b]The percentage values of the observations are given in parentheses. Numbers that have >10% of the occurrences are in bold.

behavior has been seen in the interaction between aromatic rings (McGaughey *et al*., 1998) and can be interpreted as the manifestation of a binding interaction between S and aromatic residues inside the minimum in the distribution, beyond which any direct interaction is lost because of random thermal motion. Hence within a limiting distance of 4.3 Å, any preferential orientation between the interacting groups would be revealed. In contrast, the S···C(aliphatic) interaction does not

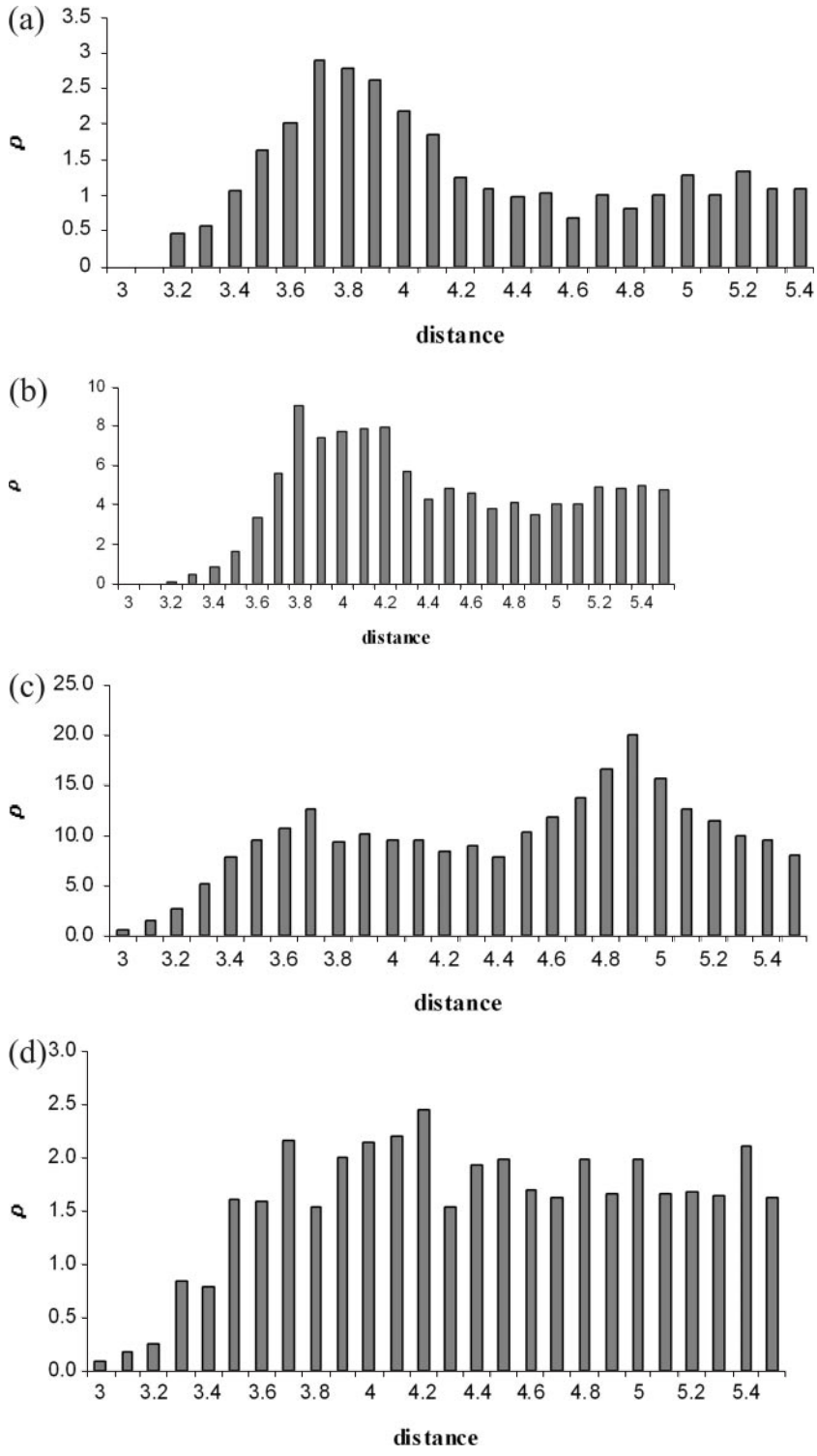

Fig. 3. Histograms for the density function, ρ (see Methods), involving atoms from (**a**) aromatic residues, (**b**) aliphatic groups, (**c**) main-chain carbonyl oxygen atoms and (**d**) side-chain oxygen atoms.

show any sharp peak or trough, suggesting the absence of any specific interaction between the groups. However, S···O(main-chain) interactions have features akin to Cys···Arom interactions and we have used a cut-off distance of 4.0 Å for identifying specific interactions of this type and then delineating their geometric features. In contrast, the distribution involving the side-chain oxygen atoms does not have any peak, but only a plateau beyond 3.5 Å, indicating that the higher thermal parameters of the side-chain atoms would mask any directional features of the S···O interaction. Within the cut-off distance used, the average distances of the centroids of different aromatic residues from S_γ are His, 4.5 (±0.5) Å; Phe, 4.8 (0.5) Å; Tyr, 4.8 (0.5) Å; Trp, 5.2 (0.8) Å. These distances are shorter than 6.0 Å, observed by Reid *et al*. (1985) to be the optimum distance between the centroid of aromatic residues and S_γ of Cys. A half cystine can have more than one aromatic

residue or carbonyl oxygen atom in contact. Of the 222 cystines involved in S · · · Arom interactions, 129 have one such contact, 59 two, 26 three and 8 four; the equivalent numbers for S · · · O interactions are 117 one contact, 195 two, 139 three, 58 four, 23 five and 6 six.

Geometry of Cys · · · Arom interactions relative to the aromatic ring

Two parameters, *P*, the interplanar angle and θ, the angular displacement of the S_γ atom relative to the aromatic ring (Figure 4a), have been used to study the geometry of interaction between half cystine and the aromatic residue. For visualization, the values (in the range 0–90°) have been grouped in bins of size 30° along the two variables, resulting in nine grid elements, and each geometry is designated by a two-letter tag (Figure 4b) as given in Materials and methods. The observed and expected numbers of occurrences in all the grid elements and that in which the two values have significant differences are shown in Figure 5. There are some minor differences in the preferred relative orientations depending on the type of aromatic residue. Tyr and Trp interact with half cystine in similar orientations with '*et*' and '*op*' geometries having more than the expected number of observations. Instead of '*et*', Phe shows a preference for the adjacent element '*en*'. Of the aromatic residues, the number of interactions is the highest (154) with Phe; with His the number is rather small (33) and it is not possible to make any definite statement on the preferred geometry. Examples of some geometric orientations are shown in Figure 6.

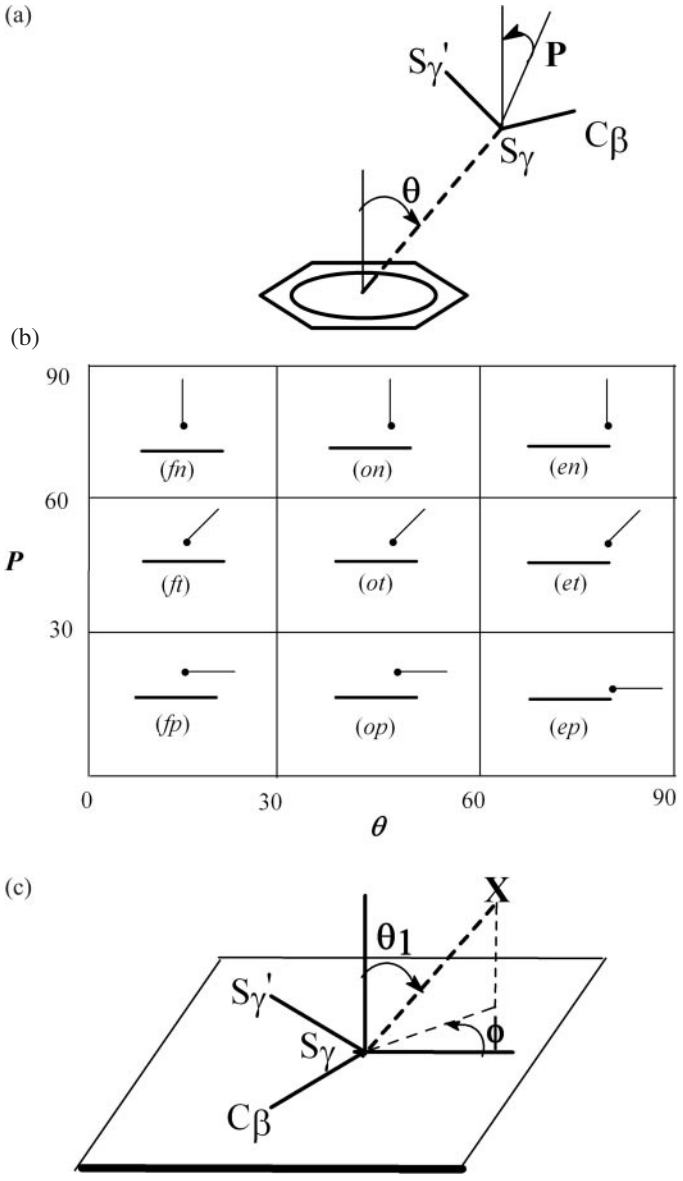

Fig. 4. (**a**) Parameters (*P* and θ) describing the orientation of the disulfide plane (defined by the S_γ atom having the contact and its two bonded neighbors) relative to the aromatic ring. The two normals passing through the aromatic centroid and the S_γ atom of the disulfide plane are shown. The interplanar angle, *P*, is the angle between the two vectors; θ is the angle between the normal to the aromatic ring and its center to the S_γ direction. (**b**) Schematic representations (for nomenclature, see Methods) for orientations corresponding to various combinations of *P* and θ values (in °). Lines signify planes (the longer one for the aromatic ring and the shorter one for the disulfide plane and the dot represents the S_γ that is in contact) perpendicular to the paper. (**c**) Spherical polar angles (θ_1, φ) describing the position of X (the aromatic centroid or the carbonyl oxygen atom) relative to the disulfide plane.

Geometry of Cys · · · Arom interactions relative to the disulfide group

A number of parameters have been calculated to visualize the interacting group from the perspective of the cystine moiety. One of them is the angle, C_β–S_γ–Ar_{cen} or S_γ'–S_γ–Ar_{cen}, shown in Figure 4c. Whichever is larger is used to draw the histogram, Figure 7a. The highest number of occurrences is observed in the range 121–150° for both the angles, suggesting that the aromatic centroid is positioned nearly at the rear of the C_β–S_γ or S_γ'–S_γ bond. A scatter diagram of θ_1 vs φ (Figure 4c), presented in Figure 7b, indicates the same feature. More points

(a) Phe (number observed, 154)

10 (9.3)	*15* (23.0)	**43** (35.8)
6 (7.4)	19 (18.2)	29 (28.4)
5 (4.4)	**18** (10.8)	*9* (16.8)

(b) Tyr (106)

5 (5.1)	20 (21.7)	23 (21.2)
4 (3.6)	*12* (15.4)	**18** (15.0)
2 (2.3)	**15** (9.9)	*5* (9.7)

(c) Trp (66)

4 (4.2)	10 (10.3)	17 (16.4)
3 (3.3)	*7* (8.0)	**14** (12.7)
2 (1.5)	**5** (3.7)	*4* (5.8)

(d) His (33)

3 (2.1)	*2* (3.9)	**5** (3.9)
3 (3.6)	**8** (6.7)	6 (6.7)
1 (1.3)	3 (2.4)	2 (2.4)

Fig. 5. The observed and expected (in parentheses) numbers of interactions corresponding to the relative orientations shown in Figure 4b. The observed number exceeding the expected number by 1σ (where σ is the r.m.s.d. value, 5.08 for Phe, 2.89 for Tyr, 0.97 for Trp, 0.99 for His) is shown in bold; when less it is shown in italics.

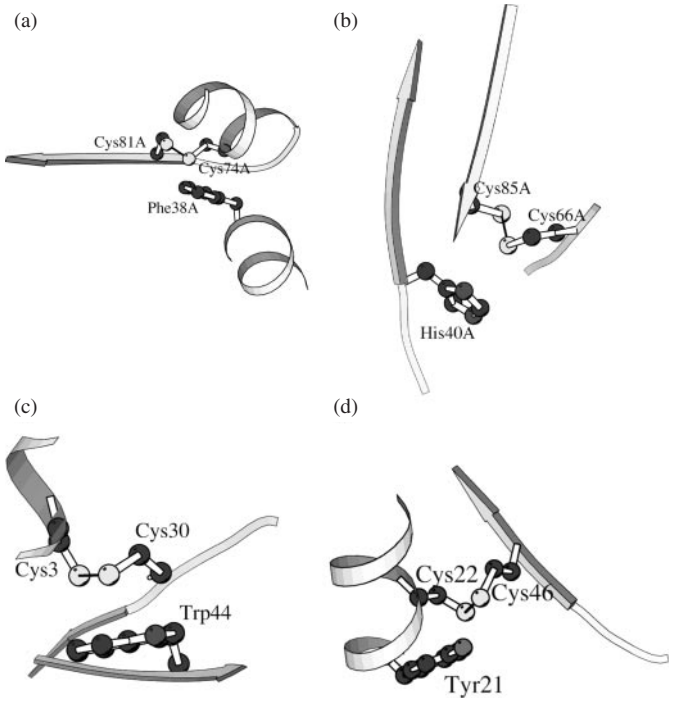

Fig. 6. Typical examples of the geometry of interaction of cystine and aromatic residues. A part of secondary structures involving the residues is shown. The protein names, PDB codes and the geometries are: (**a**) β-mannanase from *Thermomonospora fusca*, 1bqc_A, *en*; (**b**) activin receptor type II, 1bte_A, *ot*; (**c**) tick-borne encephalitis virus glycoprotein, 1svb, *op*; (**d**) scorpion toxin II, 1aho, *et*.

are closer to the disulfide plane (θ_1 in the range 45–90°) than perpendicular to it (θ_1: 0–45°) and ϕ values span the range −60 and 60°.

Sequence difference and secondary structural features of Cys · · · Arom pair

The sequence difference (Δ) between the residues involved in intra-chain Cys · · · Arom interactions is given in Table III. Only 24% of total interactions with $|\Delta| \leqslant 5$ may be termed local interactions. In 76% of interactions, the aromatic residue is more than five residues away from the half cystine, suggesting that Cys · · · Arom interactions contribute to the stability of the tertiary structure. Among the local interactions, $\Delta = -4$, −2, −1 and 2 are observed in higher numbers than the rest, but there is no interaction with $\Delta = 1$.

Some 47% of the interacting Cys · · · Arom pairs belong to three secondary structural motifs, 'HH', 'EE' and 'CE' (in the last category, 'C' corresponds to Arom). Some representative examples, along with the relative orientation of the residues, are given in Figure 8. When both the half cystine and the

Table III. Number of observations of cystine · · · aromatic interactions (intra-chain) with various sequence differences; Δ = (Arom – half cystine)

	Δ											
	<−5	−5	−4	−3	−2	−1	1	2	3	4	5	>5
Obs.	146	6	21	6	13	16	0	13	1	7	3	121

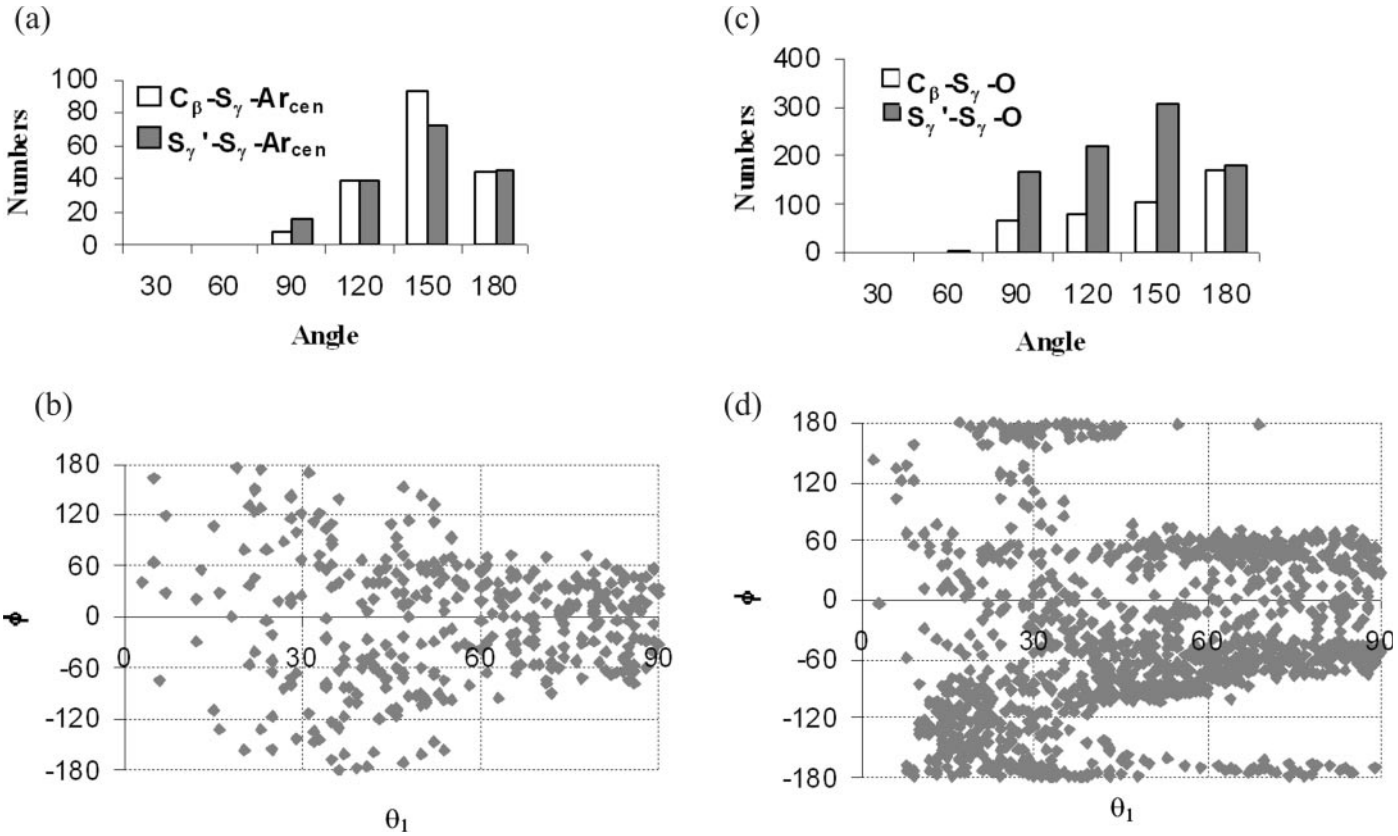

Fig. 7. Distribution of some parameters for characterizing cystine · · · Arom/oxygen interactions. (**a**) C_β–S_γ–Ar_{cen} and S_γ'–S_γ–Ar_{cen} angles (°) and (**b**) θ_1, ϕ angles (°); (**c**) and (**d**) are the equivalent figures for the cystine · · · oxygen interactions. In (a) and (c) the labels on the horizontal axis are the upper limits of the bins in the histograms.

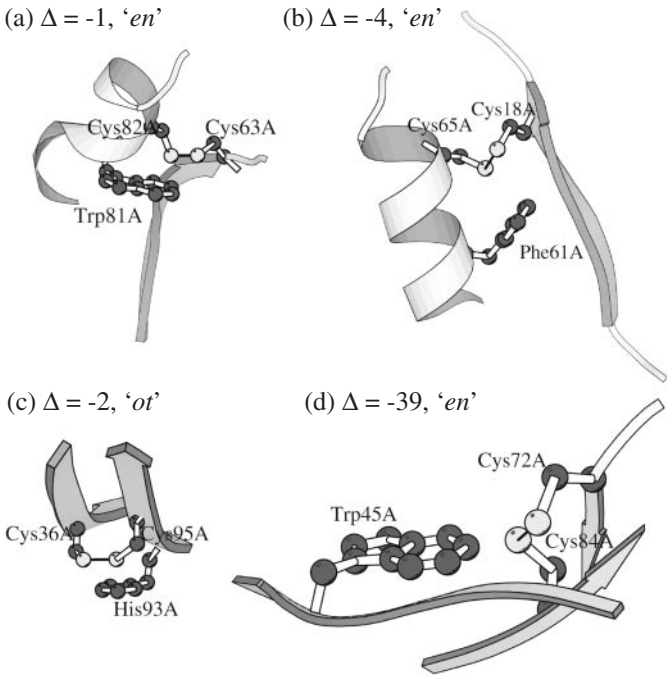

Fig. 8. Examples of the geometry of interaction between cystine and aromatic ring (with sequence difference, Δ, and geometry). The secondary structures and the labels of the interacting residues are shown. In (**a**) pectin lyase B, PDB code: 1qcx_A and (**b**) *Ustilago maydis* killer toxin, 1kp6_A, the residues are in α-helices; in (**c**) serum amyloid P component, 1sac_A, these are within a β-strand; in (**d**) activin receptor type II, 1bte_A, these are in adjacent strands of a β-sheet.

interacting aromatic residue are in the same helix, the observed sequence difference is invariably Δ = − 1 and −4 (24 cases) and not 1 and 4 (Figure 8a and b). This indicates the stereospecificity of the interaction between the half cystine and the aromatic residue, which is not possible in the reverse order.

The local interactions observed in a β-strand involve an aromatic side chain two residues preceding the half cystine (seven cases; the reverse order is found in four cases only) (Figure 8c). In 52 cases (81% of the 'EE' motifs), the half cystine and the interacting aromatic residues are in two different β-strands; of these, in 67% cases they are located in two adjacent strands of an antiparallel β-sheet (Figure 8d) (in the rest, the strands belong to two different β-sheets).

S · · · O interactions and geometric features

The average S · · · O distance involving the main-chain carbonyl oxygen atoms is 3.6(2) Å. As in Cys · · · Arom interactions (Figure 7a and b), the geometry of S(Cys) · · · O interactions has also been characterized using angular parameters (Figure 7c and d). Of the two angles in Figure 7c, S_γ'–S_γ–O occurs in a greater number of cases (883) than C_β–S_γ–O (423), indicating, as has also been noted earlier (Iwaoka *et al.*, 2002a), that the interaction along the backside of the S–S bond may be stronger than that along the rear of the S–C bond. However, the peak of the C_β–S_γ–O angle is at 180°, indicating a more linear interaction than the S_γ'–S_γ–O interaction, which peaks at ∼150°. This can also be inferred from Figure 7d, in which the angles θ_1 and φ (Figure 4c) are plotted. Although there are more points with negative values of φ, these are also scattered more. Overall, however, compared with Cys · · · Arom contacts, S(Cys) · · · O interactions are more numerous with distinct clustering. For example, points are distributed in the bands with

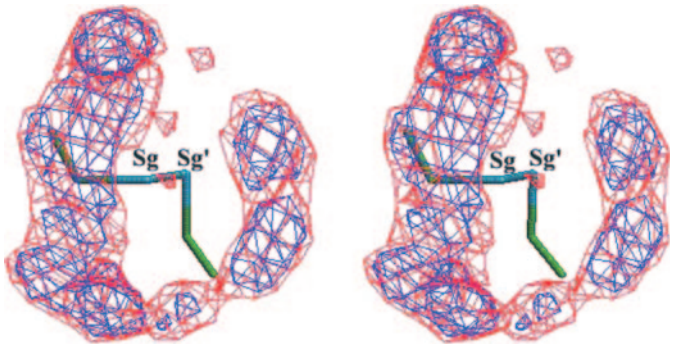

Fig. 9. Contour plot (in stereo) of the distribution of carbonyl oxygen atoms around the disulfide bond. S_γ is the atom of half cystine which is in shortest contact and the other sulfur is denoted S_γ'. C_β–S_γ–S_γ' is the common frame around which the oxygen coordinates are expressed; the three other side-chain atoms displayed are from a particular PDB file.

Table IV. The sequence difference, Δ = (carbonyl residue – half cystine), for intra-chain S · · · O interactions

	Δ												
	< −5	−5	−4	−3	−2	−1	0	1	2	3	4	5	>5
Obs.	257	12	107	64	38	173	328	21	15	9	13	13	246

φ ≈ ±60°, with points avoiding the region around φ ≈ 0°. This can be clearly seen in Figure 9, which shows no density along a plane perpendicular and bisecting the sulfide plane.

Sequence difference and the secondary structural features of S · · · O interactions

The sequence difference (Δ) of intra-chain S · · · O interactions is given in Table IV. In 25% of cases the half cystines are interacting with their own carbonyl oxygen atom. Excluding these, 48% of the remaining interactions are with $|\Delta| \leqslant 5$ and 52% with $|\Delta| > 5$, showing that the S · · · O interactions can be local, as well as long range (Figure 10).

Two structural motifs involving S · · · O interactions are conspicuous. One is the interaction between sulfur and the carbonyl oxygen present at four residues before the half cystine within an α-helix (Figure 10a); 64% of interactions of Δ = −4 occur in α-helices. The second is the interaction between half cystine and the carbonyl group of the preceding residue in a β-strand (Figure 10b); 51% of interactions of Δ = −1 are of this type. There are 82 examples of S · · · O interactions occurring between two antiparallel β-strands. Of these, in 45 cases the two strands belong to the same β-sheet (Figure 10c). Of the interactions involving antiparallel β-strands, in 35% of cases the half cystine occupies the first or the last position of a strand and 43% of the interacting carbonyl groups are also found at these two positions. In 10% of cases of long-range interactions, the residues involved do not possess any regular secondary structure.

Conservation of aromatic residues in contact with disulfide moieties

Abkevich and Shakhnovich (2000) observed a strong correlation of cystine content with polar residue content in proteins, indicating the possibility that certain amino acid classes may

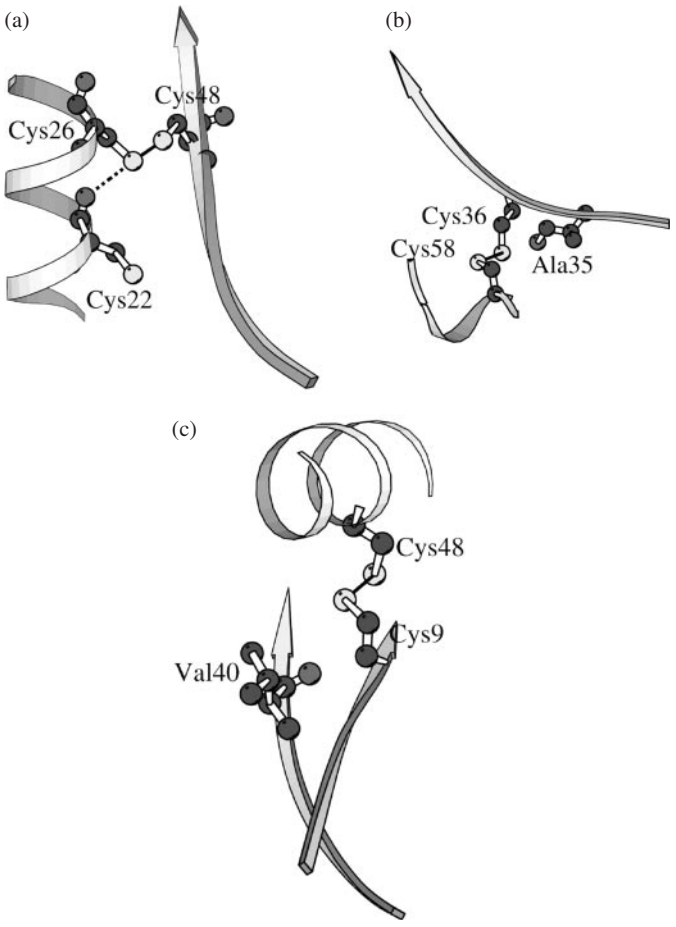

Fig. 10. Examples of S···O interactions. The protein name, PDB code, sequence difference, Δ, and secondary structures are: (**a**) scorpion toxin II, 1aho, Δ = −4, in an α-helix; (**b**) achromobacter protease I, 1arb, Δ = −1, in a β-strand; and (**c**) serum transferrin, 1a8e, Δ > 5, between two adjacent β-strands of a β-sheet.

influence the folding kinetics and stability of disulfides, although how the effect is exerted is unclear. The pronounced directionality observed in Cys···Arom interactions suggests that, just like the location of two half cystines, the presence of a neighboring aromatic residue at an optimum orientation may also be the hallmark of the existence of a disulfide bond in a protein family. Using the HOMSTRAD database (Mizuguchi *et al*., 1998; Stebbings and Mizuguchi, 2004), which has the structure-based alignments of all homologous protein families, one can study the conservation of Cys···Arom interactions in aligned sequences. The families corresponding to our PDB files showing Cys···Arom interactions were considered only if they had at least four members; 52 PDB files satisfied this condition (Table V). Of these, 41 disulfide bonds are fully conserved all across the families. Among these conserved moieties, 22 cystines are found to interact with 26 conserved (the criterion used being the degree of conservation >75%) aromatic residues, i.e. ∼50% of the conserved cystines have at least one conserved aromatic residue in specific contact. The break-up of the conserved aromatic residues interacting with 22 conserved disulfide bridge is Phe: 10, Tyr 12, Trp 4 and His 0 and the geometry in the majority of the cases is '*en*' or '*et*'.

It has been noted that non-conserved disulfide bridges lie on or near the surface of globular proteins (Thornton, 1981). Interestingly and from an opposite perspective, in 17 of these 22 bridges, the relative accessibility of both the half cystines is <7%. In all the remaining (19) conserved cases, the relative accessibility of half cystines is >7%. Thus the conserved aromatic residues have two roles—bury the disulfide bridge and provide stereospecific interactions—both contributing to the stability. The importance of an aromatic residue is further borne out by the fact that in 33 cases where the disulfide bond is not conserved amongst all the members, but the conservation is partial, it is found that for the members which show the conservation there is an aromatic residue in contact which is also conserved. Hence some sub-groups within families have a greater contribution from Cys···Arom interactions towards the evolutionary conservation of disulfide bonds. A few individual cases of conservation are discussed below.

In the papain family of cysteine proteinase (PDB file: 1me4; Table V) cysteine at the active site acts as a nucleophile for the peptide bond cleavage. An important structural motif in this family is the presence of three disulfide bonds, of which two have conserved, interacting aromatic residues. The half cystines along with the aromatic residues in contact form the hydrophobic core of the protein (Kamphuis *et al*., 1985). However, cathepsins B (1the) represents a family with a different topology of disulfide bonds (Jia *et al*., 1995). Within this family all the half cystines and the associated aromatic residues are conserved. Of these, His110 also constitutes the active site and, interestingly, Trp30 is a residue which is conserved in the papain family.

Porcine pancreatic spasmolytic polypeptide (2psp) belongs to the trefoil family of loop peptides and acts as a naturally occurring healing factor for various diseases of the gastrointestinal tract. The basic characteristic of these proteins is a domain of 38 or 39 residues, which includes six bridged Cys residues. 2psp contains two such domains and six disulfide bonds and all conserved residues are reported to be in the vicinity of the cleft areas (Petersen *et al*., 1996). Interestingly, these conserved aromatic residues (Phe36, Phe47, Phe85 and Phe96) at the cleft areas are found to interact with conserved disulfide bonds.

The members of serine proteinase inhibitors have three disulfide bonds, of which two are found to interact with conserved aromatic residues (1g6x). The disulfide bond, 14C–38C, in this PDB file was excluded from our analysis as the sulfur atom of Cys38 had fractional occupancy. However, there have been considerable solution studies involving the bridge. For example, a series of 24 mutants of bovine pancreatic trypsin inhibitor showed that the 14C–38C disulfide bond was formed at an early stage of protein folding and the rate of formation of this disulfide bond was affected 2-fold if Tyr35 was mutated by Ala (Dadlez, 1997). This Tyr35 is conserved in serine proteinase family. We found a homologous PDB file, 5pti, in which the atoms in Cys38 have full occupancy, but in this structure Tyr35 (at 6.4 Å) is beyond the contact distance from S_γ of Cys38. Yet another study found that the mutation of Tyr35 and Tyr23 affected the folding of bovine pancreatic trypsin inhibitor (Goldenberg *et al*., 1989) and Tyr23 is in interaction with the 30C–51C bridge.

The disulfide connectivity of tick anticoagulant peptide (TAP, 1d0d) is similar to the kunitz family of serine proteinase inhibitors, although its amino acid sequence identity with this family is much less (Antuch *et al*., 1994). We have found that there are two aromatic residues (Tyr1 and Tyr49) in contact of the 5C–59C bridge. The conserved Tyr49 in TAP (or Phe found

Table V. Conservation of the half cystines involved in disulfide bonds and the aromatic residues interacting with them

PDB file	Family name	No. of structures[a]	Cystine			Aromatic			Geometry[f]
			Residue[b]	Relative ASA (%)	Conservation (%) [c]	Residue[d]	Conservation (%)[e]		
							Same residue	Any other aromatic residue	
2dnj	AP endonuclease family 1	4	173C–209C	37.3, 2.1	25, 25	211Y	25		–, *op*
1hx0	α-Amylase catalytic domain	23	378C–384C	10.0, 0.9	17, 17	19W	22		–, *et*
						386H	17	4	*en*, *et*
2mcm	Antibacterial protein	4	88C–93C	6.6, 2.3	100, 100	72F	75	25	*ep*,*en*
						111F	100		*op*, –
1qdd	C-type lectin	8	14C–25C	20.8, 0.3	50, 63	142F	63		–, *ot*
			42C–140C	0.4, 1.0	100, 100	27Y	25		–, *ot*
						41Y	25	13	–, *ep*
						138F	38	13	*en*, *en*
			115C–132C	0.5, 24.2	100, 100	35W	50	38	*ft*, –
2msb	C-type lectin	8	195C–209C	0.0, 18.9	100, 100	121F	25	63	*on*, –
1en2	Chitin binding domain	5	17C–31C	0.0, 0.6	100, 100	40W	20		–, *fp*
			49C–64C	4.4, 0.0	100, 100	84Y	20		–, *fp*
1edm	Epidermal growth factor-like domain	12	51C–62C	15.4, 30.3	100, 100	69Y	58	8	–, *ot*
			56C–71C	11.3, 20.6	100, 100	69Y			–, *en*
1ju2	GMC oxidoreductase	4	399C–450C	2.5, 0.0	25, 25	390Y	25	50	*on*, –
1bqc	Glycosyl hydrolase family 5	6	74C–81C	0.0, 0.1	50, 17	38F	17		*en*, –
1qnr	Glycosyl hydrolase family 5	10	26C–29C	0.0, 0.0	10, 10	28W	10		*fn*, *op*
						42F	40		–, *ep*
			172C–175C	0.5, 19.4	10,10	135Y	10		*ot*, –
			265C–272C	0.0, 0.0	10, 10	226F	10		*en*, *en*
						240F	30	10	–, *et*
1tml	Glycosyl hydrolase family 6	4	232C–267C	19.0, 62.3	100, 25	231W	100		*ot*, –
1cnv	Glycosyl hydrolase family 18	5	41C–93C	63.5, 27.7	20, 20	97Y	20	20	*on*, –
1jnd	Glycosyl hydrolase family 18, TIM barrel domain	10	322C–405C	16.2, 33.9	50, 10	318Y	90	10	–, *on*
3lzt	Glycosyl hydrolase family 22	12	30C–115C	1.4, 0.0	100, 100	34F	33	67	*et*, *et*
						123W	67	33	*on*, –
			64C–80C	0.0, 1.4	100, 100	53Y	100		*on*, *op*
			76C–94C	14.5, 1.6	100, 100	63W	92	8	*et*, –
1k5c	Glycosyl hydrolase family 28	6	175C–191C	1.1, 3.1	67, 67	186F	50		*en*, *ep*
			300C–303C	7.4, 0.0	83, 83	275F	17	17	–, *op*
						307W	50	17	*en*, *et*
3sil	Glycosyl hydrolase family 33	4	42C–103C	4.1, 0.0	25, 25	52F	50		–, *et*
1hxn	Hemopexin-like	5	338C–380C	13.5, 7.2	20, 20	336F	40		*et*, –
1k5n	Histocompatibility antigen binding domain	5	101C–164C	0.0, 0.1	100, 100	159Y	100		–, *on*
1qfo	Immunoglobulin domain-vset-variable non-immunoglobulin	9	22C–79C	1.2, 0.0	67, 11	24F	44	22	–, *et*
						40W	89		*en*, –
						96F	11		*fn*, *fp*
1mso	insulin	6	7(B)C–7(A)C		100, 100	5(B)H	33		*on*, –
1b3a	Interleukin 8-like protein	11	10C–34C	36.4, 10.7	27, 100	12F	27		*en*, –
1i71	Kringle domain	9	50C–73C	0.0, 0.2	100, 100	13F	11	11	*et*, –
			1C–78C	28.3, 19.7	100, 100	3H	22	22	–, *ot*
1beb	Lipocalin	15	66C–160C	27.7, 19.2	80, 7	61W	7	13	–, *en*
1e5p	Lipocalin	15	57C–149C	40.1, 8.4	80, 80	50Y	20	13	–, *op*
1i4u	Lipocalin family	15	12C–121C	15.6, 15.6	13, 26	129H	20		*ot*, *et*
			117C–150C	0.0, 0.0	7, 7	115Y	47	33	*et*, *en*
			51C–173C	3.1, 9.1	13, 80	175Y	7		–, *op*
1j8e	Low-density lipoprotein receptor domain class A	5	12C–30C	2.2, 5.7	80, 100	10F	80		–, *et*
1k07	Metallo-β-lactamase superfamily	6	256C–290C	10.6, 0.2	33, 33	259F	17		–, *op*
1i0v	Microbial ribonucleases	8	6C–103C	0.8, 11.3	75, 50	11Y	50		*et*, –
2psp	P or trefoil or TFF domain	4	8C–35C	6.8, 0.5	100, 100	47F	100		–, *op*
						106Y	25	50	*et*, –
			29C–46C	0.0, 5.3	100, 100	36F	100		*en*, –
			68C–83C	8.0, 11.7	100, 100	97F	25	25	*fp*, *on*
			78C–95C	0.0, 1.2	100, 100	85F	100		*en*, –
			84C–58C	0.0, 5.1	100, 100	96F	100		–, *op*
1h03	Pair of complement control protein molecules (sushi domain)	5	7C–48C	4.5, 0.0	100, 100	57W	100		*ft*, –
1dy5	Pancreatic ribonuclease	6	26C–84C	0.0, 0.0	100, 100	25Y	100		*en*, –
						46F	100		–, *en*
			40C–95C	9.6, 17.7	100, 100	92Y	50	17	–, *et*
			58C–110C	7.1, 0.0	100, 100	73Y	33	17	–, *en*
1me4	Papain family cysteine proteinase	13	56C–95C	12.4, 32.7	100, 100	86Y	77		*en*, –
			153C–200C	6.8, 0.0	77, 92	141W	8	85	*et*, *en*

Table V. *Continued*

PDB file	Family name	No. of structures[a]	Cystine			Aromatic			Geometry[f]
			Residue[b]	Relative ASA (%)	Conservation (%) [c]	Residue[d]	Conservation (%)[e]		
							Same residue	Any other aromatic residue	
1the	Papain family cysteine proteinase	13	63C–67C	8.7, 1.0	15, 15	30W	100		–, *et*
			108C–119C	4.7, 23.8	15, 15	110H	15		*ot*, *ft*
			100C–132C	0.0, 21.9	20, 15	136Y	15		–, *en*
1qcx	Pectin/pectate lyase	4	63C–82C	2.2, 7.7	50, 50	81W	50	25	*ep*, *en*
			303C–311C	0.0, 4.6	50, 50	292F	50	25	*ot*, *et*
						307F	25		–, *fn*
1aru	Peroxidase	5	12C–24C	4.8, 20.6	80, 80	27F	60		*et*, –
1mc2	Phospholipase A2	18	1027C–1125C	4.0, 22.7	100, 100	1120H	6	67	*en*, *en*
			1029C–1045C	0.8, 1.3	100, 100	1025Y	100		*en*, –
						1102F	6		–, *en*
			1050C–1134C	14.0, 48.6	100, 72	1046F	61	11	*ot*, –
			1084C–1096C	11.2, 0.0	100, 100	1073Y	100		–, *on*
			1051C–1098C	21.8, 1.4	100, 100	1052Y	100		–, *en*
1fk5	Plant lipid transfer proteins	5	50C–89C	1.2, 6.3	80, 80	81Y	80		*ot*, –
1ioo	Ribonuclease T2	6	16C–21C	26.7, 20.4	83, 83	15F	17	17	*en*, *ep*
			169C–180C	1.6, 0.8	83, 83	4Y	50	17	*on*, –
			153C–186C	11.3, 4.8	100, 100	134F	17		–, *et*
1uca	Ribonuclease T2	6	151C–184C	12.3, 5.0	100, 100	188F	17		*ot*, *ot*
			168C–179C	4.5, 9.4	83, 83	5W	17	17	*on*, –
1psr	S-100/IcaBP type calcium binding domain	6	46C–95C	0.3, 6.2	17, 17	42F	33		*fn*, –
						53Y	17		*op*, –
1l9l	Saposin-like type B	4	7C–70C	2.2, 0.9	50, 75	48F	25	25	*en*, –
			34C–45C	11.4, 0.0	50, 75	41W	25		–, *on*
1aho	Scorpion toxin	8	22C–46C	0.0, 0.0	100, 100	21Y	50	25	*en*, *et*
			16C–36C	2.2, 4.7	100, 100	38W	38	13	–, *on*
1qq4	Serine proteinase	5	101C–111C	7.9, 52.1	20, 20	109Y	20		*op*, *on*
1c2a	Serine proteinase inhibitor Bowman–Birk type	9	95C–110C	12.4, 5.7	89, 89	66W	22		–, *en*
1sgp	Serine proteinase inhibitor kazal type	6	8C–38C	22.7, 0.0	100, 100	37F	50		*et*, *en*
						11Y	33		–, *ft*
			16C–35C	34.8, 10.9	100, 100	11Y			–, *en*
			24C–56C	3.8, 59.9	100, 100	52H	50		*ot*, –
1d0d	Serine proteinase inhibitor	10	5C–59C	0.0, 6.5	100, 100	1Y	10		*op*, –
						49Y	10	90	*et*, *on*
1g6x	Serine proteinase inhibitor	10	5C–55C	0.0, 0.0	100, 100	4F	20	20	*fp*, *on*
						45F	90	10	*en*, –
			30C–51C	14.6, 0.0	100, 100	45F			*on*, –
						23Y	60	40	*et*, –
3seb	Staphylococcal/streptococcal toxin	6	93C–113C	17.1, 0.7	67, 67	95F	33	17	*en*, *en*
1aba	Thioredoxin	6	14C–17C	7.6, 0.0	50, 33	7Y	17		*en*, –
1a8e	Transferrin	7	118C–194C	0.2, 0.1	100, 100	185Y	100		–, *fn*
2tgi	Transforming growth factor-β	8	7C–16C	2.8, 0.0	38, 100	6Y	38		*en*, *ep*
			48C–111C	4.7, 1.1	100, 100	52W	25		–, *en*
1ezm	Zinc metalloproteinase, thermolysin like	4	30C–58C	7.6, 43.8	25, 25	56F	25		*fn*, *op*

[a]In the same family in which the given PDB chain is present. The family name given in the previous column is according to HOMSTRAD.
[b]Residue number and the one-letter code of Cys of disulfide bond.
[c]The two values are for the two half cystines.
[d]Residue number and one-letter code of the interacting aromatic residue.
[e]If the same aromatic residue has already occurred in the table, the information on conservation is not repeated a second time.
[f]Two-letter tag of the geometric orientation of the sulfide plane relative to the aromatic residue (Figure 4b). A dash is given if the aromatic residue is not interacting with the corresponding half cystine in a given disulfide bridge.

in some members) is responsible for the structural integrity of the 3_{10}-helix (Asn2–Leu4) which is responsible for the binding of the molecule to the secondary site of factor Xa (Charles *et al*., 2000). Tyr1 is crucial for TAP, as it renders the peptide highly specific for factor Xa (by binding to the S1 specificity pocket of Xa), exhibiting little inhibitory activity towards other serine proteases, such as trypsin, thrombin and other blood proteases (Waxman *et al*., 1990; Charles *et al*., 2000). This is an example where an aromatic residue is positioned by the disulfide bond (in the *op* geometry) for its functional role.

Yet another example of the functional importance associated with aromatic residues is provided by phospholipase A2, which hydrolyzes phospholipids to fatty acids and lysophospholipids. These are abundant in snake venoms of various species and share similar three-dimensional structures. Acutohaemolysin (1mc2) is a lipase that lacks catalytic and hemolytic activity. Three conserved disulfide bonds interact with conserved aromatic residues, some of which have catalytic functions. Among these, Tyr1025 lies in the calcium-binding loop, which is one of the most conservative regions in the structure. Tyr1052 is one

of the constituents of the invariant His–Tyr–Asp catalytic triad. Phe1102 (interacting with 1029C–1045C, *en* geometry), which exists only in acutohaemolysin, blocks the substrate binding to this catalytic triad, resulting in the loss of hemolysis (Liu *et al*., 2003).

Lesk and Chothia (1982) first identified a high degree of conservation of Cys–Cys and Trp residues in the Fab molecule. Later, this Cys–Cys and Trp structural triad was found to be conserved in almost all immunoglobulin domains, such that Trp is packed against the disulfide bond (Ioerger *et al*., 1999). These residues, however, do not show up against the entry 1k5n for the heavy chain of histocompatibility antigen binding domain in Table V, as the distance (4.8 Å) between the half cystine and Trp is beyond our cut-off value and indeed it has been reported that the triad geometry and the distance between Cys–Cys and Trp is to some extent different in major histocompatibility complex antigen, as compared with other immunoglobulin domains (Ioerger *et al*., 1999). Nevertheless, in this molecule we have identified another conserved disulfide bond (101C–164C) and interacting Tyr159 (in *on* geometry). Interestingly, this Tyr, in turn, interacts with Phe3 of the antigen peptide, giving rise to a disulfide–aromatic–aromatic triad in the complex structure.

Discussion

The tertiary folds of native proteins are determined by a large number of weak interactions, viz., hydrogen bonding, hydrophobic interactions, salt bridges and weakly polar interactions, such as aromatic–aromatic, X–H $\cdots\pi$ (where X is N, O or C) and C–H $\cdots$ O (Singh and Thornton, 1985; Burley and Petsko, 1988; Umezawa and Nishio, 1998; Samanta *et al*., 1999, 2000; Brandl *et al*., 2001; Steiner and Koellner, 2001; Bhattacharyya *et al*., 2002, 2003; Bhattacharyya and Chakrabarti, 2003). The interactions other than the hydrophobic ones are characterized by pronounced directionality. In addition to these non-covalent forces, disulfide bridges formed by uniquely paired Cys residues in the folded state also stabilize certain proteins covalently. The question arises of whether the stability can be further enhanced by the involvement of the disulfide bridges in stereospecific interactions. A rather reluctant participant in hydrogen bonds (Ippolito *et al*., 1990; Gregoret *et al*., 1991; Allen *et al*., 1997), the sulfur atom can partake in attractive electrophile–nucleophile pairing. Non-bonded atomic contacts with divalent sulfur in crystals of small molecules show that electrophiles (such as cations) tend to approach sulfur roughly 20° from the perpendicular to the plane through atoms Y–S–Z, whereas nucleophiles (such as oxygen atoms) tend to approach approximately along the extension of one of the covalent bonds to S, indicating the preferred directions of electrophilic and nucleophilic attack on divalent sulfur (Rosenfield *et al*., 1977; Guru Row and Parthasarathy, 1981). Such interactions are also exhibited by the sulfur atom in free Cys and Met residues in protein structures (Chakrabarti, 1989; Chakrabarti and Pal, 1997; Pal and Chakrabarti, 1998, 2001; Iwaoka *et al*., 2002a) and in this paper we show a similar feature involving the Cys residues linked by disulfide bridges. Seen from this perspective, the S(Cys) $\cdots$ Arom interactions (Morgan *et al*., 1978; Reid *et al*., 1985; Pal and Chakrabarti, 1998, 2001) can also be rationalized. Based on the distribution of contacts observed at various distances (Figure 3), we have used cut-off distances of 4.0 and 4.3 Å for identifying S $\cdots$ O and S $\cdots$ Arom contacts, respectively. Delineating the geometric features of the former was restricted to main-chain oxygen atoms only, as the distribution involving the side-chain O atoms did not reveal any peak indicative of a region within which specific interactions between the atoms would stand up against the background of other non-bonded interactions.

Features of the disulfide bonds

For proteins containing disulfide bonds, one can fit a power function to the number of bonds per 100 residues plotted against the chain length (Figure 1); after an initial fall, the number flattens out at about 150 residues. The left-handed spiral structure (Richardson, 1981) with $\chi_1 \approx \chi_2 \approx -60°$ and $\chi_3 \approx -90°$ (Figure 2) is the most populated conformation. Unlike χ_1, the χ_2 torsion (C_α–C_β–S_γ–S_γ') angle is closer to $\pm 90°$ and this aspect is similar to what is exhibited by χ_2, the equivalent torsion involving metal ion, when Cys residues act as ligands (Chakrabarti, 1989). The residues with χ_3 values in the fringes of the distribution usually have χ_1 and/or χ_2 around +60°, a nominally higher energy conformation. Such bonds may be strained and it will be worthwhile to study the stability conferred by such bonds to protein structures.

Stereospecific interactions

The contour plot showing the distribution of oxygen atoms around the interacting sulfur of the disulfide bond is shown in Figure 9. There is a general avoidance of the circular region around the extension of the bisector to the C_β–S_γ–S_γ' angle and the points are clustered at the backside of C_β–S_γ and S_γ'–S_γ bonds, suggesting that the carbonyl oxygen atom tends to approach cystine along the antibonding orbital of the bonds. The stabilizing nature of the interaction when the molecular orbitals of an nucleophile (oxygen) and electophile (sulfur) interact lead to the directional properties of S $\cdots$ O non-bonded interactions (Rosenfield *et al*., 1977; Guru Row and Parthasarathy, 1981). This interaction (examples in Figure 10) has been noted earlier in protein structures, especially in the interaction involving the sulfur atom of Met residues (Pal and Chakrabarti, 2001; Iwaoka *et al*., 2002a). Although there are more points interacting at the backside of the S_γ'–S_γ bond, there is more diffusion of points away from the sulfide plane as compared with the region behind the C_β–S_γ bond. These geometric features are also revealed from an analysis of angular parameters (Figure 4c) displayed in Figure 7c and d. *Ab initio* calculations using model systems indicate that the S $\cdots$ O interaction can contribute 2.5–3.2 kcal/mol to the stability (Dixit *et al*., 1995; Iwaoka *et al*., 2002a,b).

Like the S $\cdots$ O interaction, the S $\cdots$ Arom interaction also has preferred geometries. From the perspective of the sulfide plane the distribution of the aromatic centroid is similar to that of the oxygen atoms, as revealed by the observed values (Figure 7) of θ_1 and ϕ angles (Figure 4c). However, as the π-electron cloud is spread over the whole aromatic ring, the clustering is less prominent and a contour map similar to Figure 9 for oxygen atoms could not be obtained. The orientation of the sulfide plane relative to the aromatic ring is also found to be restricted, the preferred geometry (Figure 4b) being *en* or *et* (Figures 5 and 6). The orientations having repulsive interaction between the face of the aromatic ring and the sp^3 lone pair of electrons in sulfur are avoided. For example, in the stacked *fp* geometry one lone pair orbital of sulfur points directly to the π electron cloud of the ring, whereas in *fn* both orbitals interact

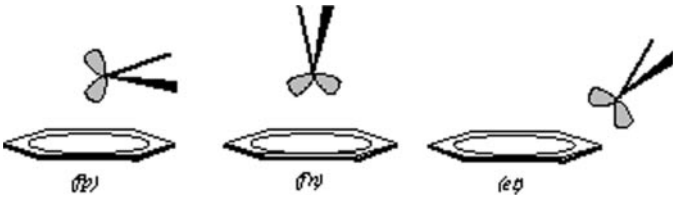

Fig. 11. Disposition of the lone pairs of electrons on the sulfur atom relative to the aromatic ring in three representative geometries in Figure 4b.

unfavorably (Figure 11). In contrast, the repulsive interaction is the minimum in *et* and the rear (i.e. the antibonding orbital) of the C_β–S_γ or S_γ'–S_γ bond can interact favorably with the π-electrons. Although Met · · · Arom interactions were not analyzed in terms of similar geometric parameters, the preferred geometry of sulfur appears to be on top of the aromatic ring, close to the periphery (Pal and Chakrabarti, 2001). Energy calculations also suggest an interaction of the sulfur atom with the aromatic face, but all of the possible geometries encountered in protein structures have not been investigated (Némethy and Scheraga, 1981; Pranata, 1997). It is normally assumed that the environment of disulfide bond is hydrophobic. However, the stereospecific location of O atoms and aromatic rings, whose geometry is dictated by the electronic interactions, suggests that electrostatic factors also control the local structure around the bond.

Conservation of Cys · · · Arom interactions and their functional role

When a Cys · · · Arom interaction occurs in a structure, it is found to be an invariant feature in at least 50% of the families to which the protein belongs, as long as the disulfide bond itself is fully conserved (Table V). Such bridges are also fully buried in the structure. In families in which the S–S bond is not fully conserved, constituent members showing the conservation of the bond usually have a Cys · · · Arom interaction conserved. When they are conserved, Cys · · · Arom interactions can be considered as the signature motif for a particular family of proteins and may be used as a fingerprint to annotate the function based on the three-dimensional structure of an unknown protein.

Although one normally associates disulfide bridges to provide extra stability, these may also be needed for function (Chang *et al*., 2003) and the engineered disulfide bond can also modulate the functional attribute of a molecule (Sauer *et al*., 1986). For example, human insulin contains two interchain and one intra-chain disulfide linkages, which make different contributions to the structure formation of insulin and are formed sequentially in the order A20–B19, A7–B7 and A6–A11 (the letter corresponds to the chain label) in the folding pathway of proinsulin; but all three are essential for receptor binding activity (Chang *et al*., 2003). Interestingly, in many of the proteins in Table V, the aromatic residues interacting with disulfide bonds have a unique role in the function, especially in binding or even preventing the substrate binding (see Results, last section) and, expectedly, these Cys · · · Arom interactions are evolutionarily conserved. Given the importance of Arom · · · Arom interactions (Singh and Thornton, 1985; Burley and Petsko, 1988; Samanta *et al*., 1999, 2000; Bhattacharyya *et al*., 2002, 2003) and the prevalence of aromatic residues at protein–protein interfaces (Chakrabarti and Janin, 2002), one can encounter a Cys · · · Arom · · · Arom triad stabilizing protein–protein interactions.

Transcription growth factor β (TGF_β) has four disulfide bonds formed by eight conserved half cystines. The hydrophobic clusters at the binding surface are constituted mostly by aromatic residues, which are highly conserved in this family of proteins. Among these conserved aromatic residues, His40, Trp45 and Phe83 are within 4.5 Å of cystine (Greenwald *et al*., 1999). The mutual exclusiveness of disulfides and aromatic residues in protein families suggests the importance of their interaction in the protein structure and function and this fact can be used in molecular design.

Stability of disulfide bonds and the sequential pattern of residues

The disulfide bonds are susceptible to thiol/disulfide exchanges, provided a nearby thiolate anion can attack it (Gilbert, 1984). The direction of this attack will be along the backside of the S_γ'–S_γ bond. As long as this approach is obstructed by the presence of an incipient electrophile–nucleophile interaction (sulfur with oxygen or aromatic) within the protein structure, with the Cys residues having very low accessibility, the disulfide bond is prevented from undergoing any reaction. This is also the reason why such disulfide bonds have survived the evolutionary pressure of mutation. However, when the cystine is at the enzyme active site, such as 58C–63C in glutathione reductase (Karplus and Schulz, 1987), the half cystines are more exposed with relative accessibilities 14 and 39% (in the PDB file, 3grs), respectively; the only carbonyl group close by being that of Cys58 (at 3.5 Å) with θ_1 and ϕ (Figure 4c) being 18 and $-121°$, a location nearly perpendicular to the sulfide plane, such that the rear side of the S_γ'–S_γ bond is available for nucleophilic attack by a thiolate anion.

When a specific interaction occurs between two residues at a particular sequence difference, it is observed that the order of the two residues in the sequence is also important. For example, the location of a His four residues following a Phe in an α-helix can give rise to an N–H · · · π (or C–H · · · π) interaction stabilizing the helix; the interaction is not attainable if the order of the pair is reversed (Bhattacharyya *et al*., 2002). Likewise in a helix, an interacting aromatic residue preceding a half cystine by four (or one) residues is found to occur more frequently than when the order is the reverse; a similar order is also observed with an interacting carbonyl group (Tables III and IV; Figures 8a and b and 10a) and in the S(Met) · · · Arom interaction (Pal and Chakrabarti, 2001).

Finally, we have shown in this paper that there are preferred geometries for interaction of half cystines with the peptide oxygen atoms and aromatic side chains and there are compelling reasons to believe that the S · · · Arom interaction is distinct from the hydrophobic interaction that one expects when the disulfide bond is buried in an environment of aliphatic groups. There is a high degree of conservation of the aromatic residues in contact with the bond and many of these are endowed with a functional role.

Acknowledgements

A Senior Research Fellowship from the Council of Scientific and Industrial Research, India, supported R.B. and the Department of Biotechnology provided some computational facilities.

References

Abkevich,V.I. and Shakhnovich,E.I. (2000) *J. Mol. Biol*., **300**, 975–985.

Allen,F.H. (2002) *Acta Crystallogr. B*, **58**, 380–388; http://www.ccdc.cam.ac.uk/

Allen,F.H., Bird,C.M., Rowland,R.S. and Raithby,P.R. (1997) *Acta Crystallogr. B*, **53**, 696–701.

Antuch,W., Guntert,P., Billeter,M., Hawthorne,T., Grossenbacher,H. and Wuthrich,K. (1994) *FEBS Lett.*, **352**, 251–257.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne, P.E. (2000) *Nucleic Acids Res*., **28**, 235–242.

Betz,S.F. (1993) *Protein Sci.*, **2**, 1551–1558.

Bhattacharyya,R. and Chakrabarti,P. (2003) *J. Mol. Biol*., **331**, 925–940.

Bhattacharyya,R., Samanta,U. and Chakrabarti,P. (2002) *Protein Eng*., **15**, 91–100.

Bhattacharyya,R., Saha,R.P., Samanta,U. and Chakrabarti,P. (2003) *J. Proteome Res.*, **2**, 255–263.

Brandl,M., Weiss,M.S., Jabs,A., Sühnel,J. and Hilgenfeld,R. (2001) *J. Mol. Biol*., **307**, 357–377.

Burley,S.K. and Petsko,G.A. (1988) *Adv. Protein Chem*., **39**, 125–189.

Chakrabarti,P. (1989) *Biochemistry*, **28**, 6081–6085.

Chakrabarti,P. and Janin,J. (2002) *Proteins*, **47**, 334–343.

Chakrabarti,P. and Pal,D. (1997) *Protein Sci.*, **6**, 851–859.

Chakrabarti,P. and Pal,D. (2001) *Prog. Biophys. Mol. Biol.*, **76**, 1–102.

Chang,S.G., Choi,K.D., Jang,S.H. and Shin,H.C. (2003) *Mol. Cells*, **16**, 323–330.

Charles,R.St., Padmanabhan,K., Arni,R.V., Padmanabhan,K.P. and Tulinsky,A. (2000) *Protein Sci.*, **9**, 265–272.

Clarke,J. and Fersht,A.R. (1993) *Biochemistry*, **32**, 4322–4329.

Dadlez,M. (1997) *Biochemistry*, **36**, 2788–2797.

Darby,N.J. and Creighton,T.E. (1993) *J. Mol. Biol.*, **232**, 873–896.

Desiraju,G.R. and Steiner,T. (1999) *The Weak Hydrogen Bond in Structural Chemistry and Biology*. Oxford University Press, Oxford.

Dill,K.A. (1990) *Biochemistry*, **29**, 7133–7155.

Dixit,A.N., Reddy,K.V., Rakeeb,A., Deshmukh,A.S., Rajappa,S., Ganguly,B. and Chandrasekhar,J. (1995) *Tetrahedron*, **51**, 1437–1448.

Fiser,A., Cserzö,M., Tüdös,É. and Simon,I. (1992) *FEBS Lett.*, **302**, 117–120.

Gilbert,H.F. (1984) *Methods Enzymol.*, **107**, 330–351.

Goldenberg,D.P., Frieden,R.W., Haack,J.A. and Morrison,T.B. (1989) *Nature*, **338**, 127–132.

Greenwald,J., Fischer,W.H., Vale,W.W. and Choe,S. (1999) *Nat. Struct. Biol.*, **6**, 18–22.

Gregoret,L.M., Rader S.D., Fletterick,R.J. and Cohen,F.E. (1991) *Proteins*, **9**, 99–107.

Guru Row,T.N. and Parthasarathy,R. (1981) *J. Am. Chem. Soc.*, **103**, 477–479.

Hinck,A.P., Truckses,D.M. and Markley,J.L. (1996) *Biochemistry*, **35**, 10328–10338.

Hubbard,S.J. (1992) *NACCESS: A Program for Calculating Accessibilities*. Department of Biochemistry and Molecular Biology, University College London, London.

Ioerger,T.R., Du,C. and Linthicum,D.S. (1999) *Mol. Immunol.*, **36**, 373–386.

Ippolito,J.A., Alexander,R.S. and Christianson,D.W. (1990) *J. Mol. Biol*., **215**, 457–471.

Iwaoka,M., Takemoto,S., Okada,M. and Tomoda,S. (2002a) *Bull. Chem. Soc. Jpn*., **75**, 1611–1625.

Iwaoka,M., Takemoto,S. and Tomoda,S. (2002b) *J. Am. Chem. Soc*., **124**, 10613–10620.

Janin,J., Wodak,S., Levitt,M. and Maigret,B. (1978) *J. Mol. Biol.*, **125**, 357–386.

Jia,Z., Hasnain,S., Hirama,T., Lee,X., Mort,J.S., To,R. and Huber,C.P. (1995) *J. Biol. Chem.*, **270**, 5527–5533.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Kamphuis,I.G., Drenth,J. and Baker,E.N. (1985) *J. Mol. Biol*., **182**, 317–329.

Karplus,P.A. and Schulz,G.E. (1987) *J. Mol. Biol.*, **195**, 701–729.

Katz,B.A. and Kossiakoff,A. (1986) *J. Biol. Chem.*, **261**, 15480–15485.

Kraulis,P.J. (1991) *J. Appl. Crystallogr*., **24**, 946–950.

Lee,B. and Richards,F.M. (1971) *J. Mol. Biol*., **55**, 379–400.

Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol.*, **160**, 325–345.

Liu,Q., Huang,Q., Teng,M., Weeks,C.M., Jelsch,C., Zhang,R. and Niu,L. (2003) *J. Biol. Chem.*, **278**, 41400–41408.

Mansfeld,J., Vriend,G., Dijkstra,B.W., Veltman,O.R., Van den Burg,B., Venema,G., Ulbrich-Hofmann,R. and Eijsink,V.G.H. (1997) *J. Biol. Chem.*, **272**, 11152–11156.

Matsumura,M., Signor,G. and Matthews,B.W. (1989) *Nature*, **342**, 291–293.

McGaughey,G.B., Gagné,M. and Rappé, A.K. (1998) *J. Biol. Chem.*, **273**, 15458–15463.

Meyer,E.A., Castellano,R.K. and Diederich,F. (2003) *Angew. Chem. Int. Ed.*, **42**, 1210–1250.

Mitchinson,C. and Wells,J.A. (1989) *Biochemistry*, **28**, 4807–4815.

Mizuguchi,K., Dean,C.M., Blundell,T.L. and Overington,J.P. (1998) *Protein Sci*., **7**, 2469–2471.

Morgan,R.S., Tatsch,C.E., Gushard,R.H., McAdon,J.M. and Warme,P.K. (1978) *Int. J. Pept. Protein Res.*, **11**, 209–217.

Morris,A.L., MacArthur,M.W., Hutchinson,E.G. and Thornton,J.M. (1992) *Proteins*, **12**, 345–364.

Némethy,G. and Scheraga,H.A. (1981) *Biochem. Biophys. Res. Commun.*, **98**, 482–487.

Pal,D. and Chakrabarti,P. (1998) *J. Biomol. Struct. Dyn.*, **15**, 1059–1072.

Pal,D. and Chakrabarti,P. (2001) *J. Biomol. Struct. Dyn*., **19**, 115–128.

Petersen,T.N., Henriksen,A. and Gajhede,M. (1996) *Acta Crystallogr. D*, **52**, 730–737.

Petersen,T.N., Jonson,P.H. and Petersen,S.B. (1999) *Protein Eng.*, **12**, 535–548.

Pranata,J. (1997) *Bioorg. Chem.*, **25**, 213–219.

Reid,K.S.C., Lindley,P.F. and Thornton,J.M. (1985) *FEBS Lett*., **190**, 209–213.

Richardson,J.S. (1981) *Adv. Protein Chem*., **34**, 167–339.

Rosenfield,R.E.,Jr, Parthasarathy,R. and Dunitz,J.D. (1977) *J. Am. Chem. Soc.*, **99**, 4860–4862.

Samanta,U., Pal,D. and Chakrabarti,P. (1999) *Acta Crystallogr. D*, **55**, 1421–1427.

Samanta,U., Pal,D. and Chakrabarti,P. (2000) *Proteins*, **38**, 288–300.

Sauer,R.T., Hehir,K., Stearman,R.S., Weiss,M.A., Jeitler-Nilsson,A., Suchanek,E.G. and Pabo,C.O. (1986) *Biochemistry*, **25**, 5992–5998.

Singh,J. and Thornton,J.M. (1985) *FEBS Lett.*, **191**, 1–6.

Srinivasan,N., Sowdhamini,R., Ramakrishnan,C. and Balaram,P. (1990) *Int. J. Pept. Protein Res*., **36**, 147–155.

Stebbings,L.A. and Mizuguchi,K. (2004) *Nucleic Acids Res*., **32**, D203–D207.

Steiner,T. and Koellner,G. (2001) *J. Mol. Biol*., **305**, 535–557.

Taylor,J.C. and Markham,G.D. (1999) *J. Biol. Chem.*, **274**, 32909–32914.

Thomas,A., Meurisse,R., Charloteaux,B. and Brasseur,R. (2002) *Proteins*, **48**, 628–634.

Thornton,J.M. (1981) *J. Mol. Biol*., **151**, 261–287.

Umezawa,Y. and Nishio,M. (1998) *Bioorg. Med. Chem*., **6**, 2507–2515.

van den Burg,B., Dijkstra,B.W., van der Vinne,B., Stulp,B.K., Eijsink,V.G.H. and Venema,G. (1993) *Protein Eng*., **6**, 521–527.

van Vlijmen,H.W.T., Gupta,A., Narasimhan,L.S. and Singh,J. (2004) *J. Mol. Biol.*, **335**, 1083–1092.

Wang,G. and Dunbrack,R.L.,Jr (2003) *Bioinformatics*, **19**, 1589–1591.

Waxman,L., Smith,D.E., Arcuri,K.E. and Vlasuk,G.P. (1990) *Science*, **248**, 593–596.

Received September 2, 2004; revised November 2, 2004; accepted November 27, 2004

Edited by P.Balaram