Structure, Vol. 12, 1–20, June, 2004, ©2004 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.str.2004.03.029

The Linked Conservation of Structure and Function in a Family of High Diversity: The Monomeric Cupredoxins

Julian Gough[1,*] and Cyrus Chothia[2]
[1]RIKEN Genomic Sciences Centre
1-7-22 Suehiro-cho, Tsurumi-ku,
Yokohama 230-0045
Japan
[2]MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH
United Kingdom

Summary

The monomeric cupredoxins are a highly divergent family of copper binding electron transport proteins that function in photosynthesis and respiration. To determine how function and structure are conserved in the context of large sequence differences, we have carried out a detailed analysis of the cupredoxins of known structure and their sequence homologs. The common structure of the cupredoxins is formed by a sandwich of two β sheets which support a copper binding site. The structure of the deeply buried core is intimately coupled to the binding site on the surface of the protein; in each protein the conserved regions form one continuous substructure that extends from the surface active site and through the center of the molecule. Residues around the active site are conserved for functional reasons, while those deeper in the structure will be conserved for structural reasons. Together the two sets support each other.

Introduction

The monomeric cupredoxins bind a copper ion that is used for electron transport in photosynthesis and respiration. Domains homologous to these are also found in multimeric enzymes. Both forms occur with very diverse sequences (see Adman, 1991; Murphy et al., 1997). In the monomeric forms of known structure small differences in the copper binding sites do occur (and these can be of functional importance), but, overall, they have very similar core structures in spite of the protein sequences being very diverse. Here we analyze the structure and sequences of the monomeric members of the cupredoxin family of proteins to determine how the binding site is conserved in the context of large sequence changes.

We do not consider the multidomain members here. The subunit-subunit interactions play a role in their structural evolution and this can obscure and complicate the points we wish to make in this paper. Also, in some cases, their cupredoxin domains have lost the ability to bind copper. Previous work has been carried out on the analysis of the structures of the cupredoxin family (Guss and Freeman, 1983; Walter et al., 1996; Hart et al., 1996; Petratos et al., 1988; Guss et al., 1996; Baker, 1988; Durley et al., 1993). These will be discussed later in the paper after we have presented our results.

*Correspondence: gough@gsc.riken.jp

Cupredoxins of Known Structure

At the time this work was carried out there were seven very different monomeric members of the cupredoxin family which had had their structure experimentally determined. (1) Plastocyanin (Guss and Freeman, 1983; Colman et al., 1978; Guss et al., 1992; Garrett et al., 1983), which is part of the photosynthetic apparatus in plants and algae, carries an electron between photosystems II and I receiving an electron from cytochrome b6/f and donating it to the P700+ reaction center. (2) Amicyanin (Durley et al., 1993), which is part of the respiratory chain of certain methylotrophic bacteria, accepts an electron from methylamine dehydrogenase and transfers it to C-type cytochromes. (3) Pseudoazurin (Petratos et al., 1988, 1995), which is part of the aerobic respiratory chain frequently found in denitrifying bacteria, transfers an electron from various donors to nitrite reductase, and possibly others (Leung et al., 1997). (4) Azurin (Baker, 1988; Adman et al., 1978; Shepard et al., 1993), which is in the respiratory system of bacteria, together with cytochrome c551, transfers an electron from the membrane-bound bc1 complex to a soluble nitrite reductase. (5) Rusticyanin (Walter et al., 1996; Ryden, 1984), which is present in acidophilic bacteria, is thought to play a role in the electron transport chain following oxidation of Fe^{2+} (Djebli et al., 1992). (6) Stellacyanin (Hart et al., 1996) has no known function. (7) Cucumber Basic Protein (Guss et al., 1996), which with stellacyanin is part of the Phytocyanin subgroup of cupredoxins (Ryden, 1984), may be associated with photosystem II particles in chloroplasts a/d.

Subsequent to the completion of most of this work, an eighth member, auracyanin (Bond et al., 2001; van Driessche et al., 1999) was published. We discuss this structure near the end of the paper.

The size of these seven proteins varies between 96 and 155 residues. Orthologs of the first four of these proteins are widely distributed and have had structures determined from a variety sources, e.g., plants, fungi, and/or bacteria. These orthologs have sequence identities of about 70% and structures that are much more similar to each other than they are to the different family members listed above. In this work, therefore, we use the seven structures described above to represent the diversity of members of the family. Information on their PDB files, size, and references for the structure determinations are given in Table 1. The structures have all been determined at high resolution (1.31–1.90) and have low R factors (14%–19%).

The cupredoxins have a conserved β sheet sandwich structure (Chothia and Lesk, 1982; Guss and Freeman, 1983; Baker, 1988) made up of seven or eight parallel and antiparallel strands, a variable α helix region at one side, and the copper binding site situated mostly atop

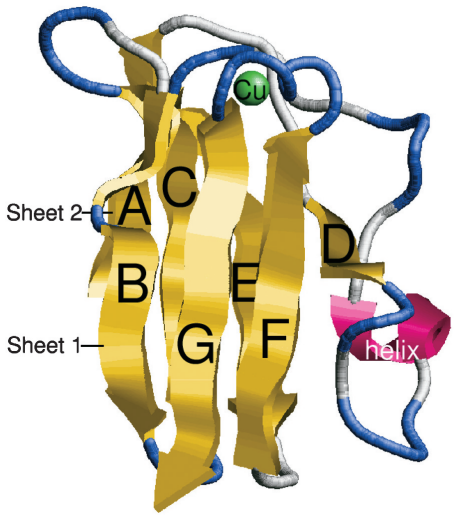

Figure 1. The Structure of Plastocyanin Viewed with the Second β Sheet behind the First

Note the copper atom bound atop strands G and F of the first sheet, and the α helix and irregular loop region to right the edge of the two sheets.

one sheet. Here, plastocyanin is used as a standard for the family and the others are compared to it. The structure of plastocyanin is shown in Figure 1. The secondary structures common to all cupredoxins are the two β sheets that pack face to face: one with strands A, C, and E, and the second with strands G, F, and D. The N-terminal half of strand B is part of the first β sheet in five of the structures, and the C-terminal half is part of the second sheet in all seven structures. Regions equivalent to that around the α helix and interstrand loops in plastocyanin can have quite different conformations and sizes in the other cupredoxins (Figure 2) (Colman et al., 1978; Adman et al., 1978; Durley et al., 1993; Petratos et al., 1988; Walter et al., 1996; Guss et al., 1996; Hart et al., 1996). Although the two β sheets are conserved in all cupredoxins of known structure, their positions relative to each other, and the exact length of their strands, can vary. To determine the regions of structure common to all structures we examined their hydrogen bonds, residue contacts, and residue accessible surface areas and carried out structural superpositions.

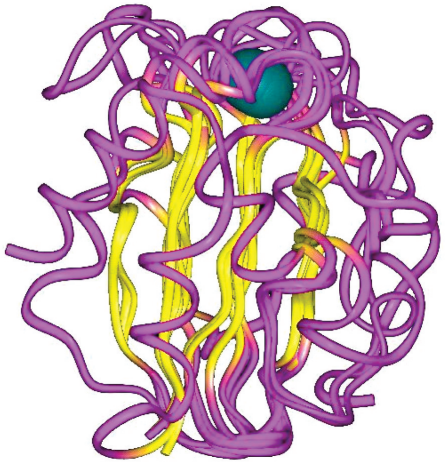

Figure 2. The Conserved and Variable Regions of the Cupredoxin Family

Three cupredoxin structures (plastocyanin, 1plc; rusticyanin, 1rcy; and pseudoazurin, 1paz) have been superposed. The yellow regions are aligned, and the bound copper atoms are green. The purple regions show the variation in structure of the nonconserved regions.

Structural Data for the Cupredoxins

These data were all calculated using Arthur Lesk's PINQ program.

Hydrogen Bonds

We identified in each structure all of the backbone hydrogen bonds which join the strands into parallel and antiparallel β sheets. The pattern of hydrogen bonds identifies the strands in the two sheets and suggests an initial set of equivalent positions between structures. The conservation of hydrogen bonds between structures gives an initial suggestion as to the extent of conservation of the secondary structure in the different proteins. The conserved β sheet hydrogen bonds can be seen in Figure 3.

Residue Contacts

Contacts between residues in protein interiors tend to be preserved and give useful conservation information about the residues that play equivalent roles in different homologs. We calculated the contacts made by residues in the protein, i.e., residues containing atoms whose distance apart is less than a specified threshold. The threshold used here is that a contact exists between two residues if they have atoms whose distance apart is less than the sum of their Van der Waals' radii plus 0.5.

Accessible Surface Area

The accessible surface area (Lee and Richards, 1971) was calculated for all residues in each structure (Figure 3). The accessible surface area is important conservation information that indicates which residues are on the surface, which are buried, and to what extent. For residues that conserve their conformation in different cupredoxin structures, the average value was determined (see Figure 4).

Structural Comparisons of the Seven Cupredoxins

The Regions of Similar and Different Conformation

When comparing hydrogen bond patterns of β sheets, a spatial shift of two residues either up or down in the plane of the sheet produces an alternative alignment. As a consequence, the hydrogen bond diagrams do not always give an unambiguous positional equivalence of stands between structures. The correct solution must be determined by use of hydrogen bonds, residue contacts, and three-dimensional superpositions. Once the backbone hydrogen bond patterns have been used to suggest an initial structural alignment, full superpositions can be carried out from this starting point. Figure 2 shows the result of full superpositions between three structures; the conserved strands are clearly aligned, whereas the loops and helices are not.

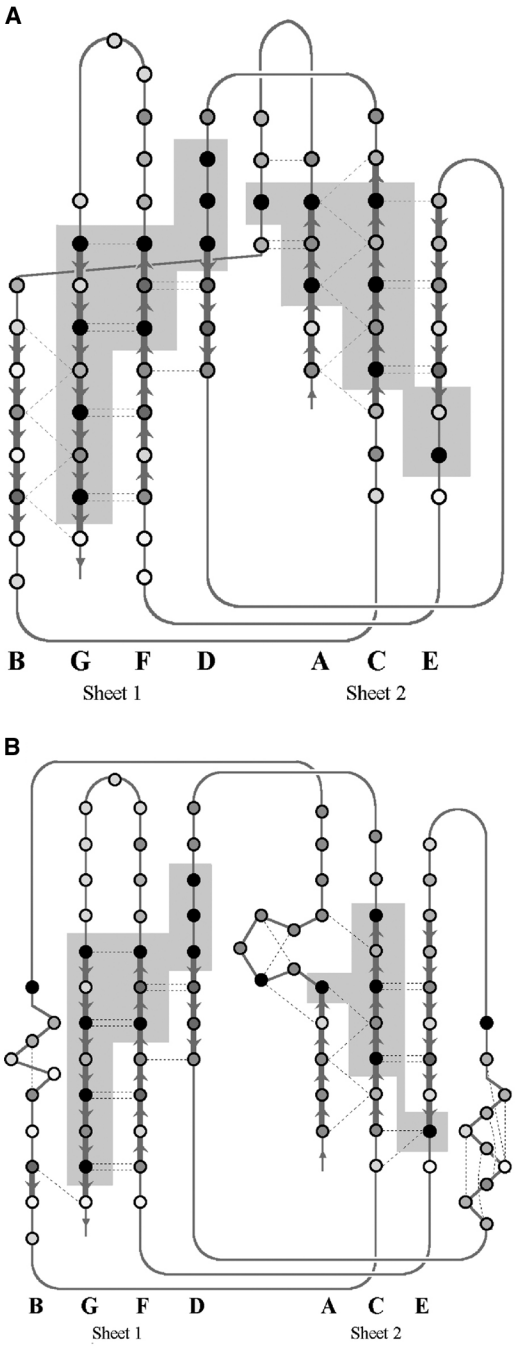

Figure 3. The Conserved Structure of the Members of the Cupredoxin Family

The dashed lines represent hydrogen bonds, the solid lines the backbone of the chain, with the thicker sections for β strands. Each circle is a peptide with the darkness representing the surface accessible area. The darker the circle, the less accessible area, going from maximum exposure (white) to maximum buriedness (black). Note in the rusticyanin-stellacyanin subfamily the disruption of the left-hand edge caused by a change in the contiguous region including the tops of strands A and B. Note also the diagonal shaded areas marking the inner surfaces of the sheets contacting each other where they twist around each other. See also Figure 7.

Plastocyanin was chosen as a "master" structure, and the other six structures were each in turn aligned to it (Table 1). Ultimately the pairwise superpositions were combined to find the multiple structural alignment. Although within each of the β sheets there is very little difference in conformation, there can be more variation in the positions of the two sheets relative to one another. Thus the pairwise structural alignments were arrived at by first superposing sheet one only, then sheet two, and then the whole structure at once by combining the matching regions from sheets 1 and 2. The work on hydrogen bond patterns was used to define an equivalence between any two structures of a few residues on each strand of one of the two sheets. These equivalencies were used to fit the pair of structures to each other by minimizing the rms deviation between backbone atoms of these few equivalent residues on the chosen sheet. The number of residues used for the fit was then iteratively increased by extending the region of each strand included in the fit, inferring the equivalence from the previous fit. This was continued as long as no pair of residues differed in position by 3.0. Alignments of sheets 1 and 2 in plastocyanin and azurin were described in a previous paper (Chothia and Lesk, 1982). Our results indicate that this previous work made an error in the alignment of sheet two, probably because at the time it was carried out only C coordinates were available for the structure of azurin.

Once the pairwise alignments had been made, a comparison of the regions of plastocyanin which align to the other structures could be used to find the regions which are common to all. This gives the full multiple structural alignment between all members. It became clear during the analysis that cupredoxins fall into two sets. The phytocyanins (Cucumber Basic Protein and stellacyanin) are more similar to each other than to the rest and vice versa (Guss et al., 1996; Hart et al., 1996). The set of five structures we will call set I and the set containing the other two we will call set II. In set I, the five cupredoxins share 57 residues which have the same conformation. This common core comprises 58% of the residues in the smallest (plastocyanin) and 37% of those in the largest (rusticyanin). Cucumber Basic Protein and stellacyanin also have 57 residues in their common core: 59% and 52% of their respective total residues. The 57 sites for set I are shown later in Table 3, along with 35 sites in common with both sets (see below).

The difference between the sets is produced in part by the top of strand A and strand B having alternate bulging conformations in set II (because of an inserted region that significantly alters that edge of the structure [Guss et al., 1996; Hart et al., 1996]). Note that this edge of the protein and the loop which joins the strands are not involved in the binding site. In addition the conformation of six residues in strand C, and one to three residues in the other strands differ in the two sets. The effect is to reduce the number of residues with the same conformation in the ACE β sheet by a greater degree than in the BGFD β sheet. The remaining 35 sites that do have the same conformation are shown later in Figure 4.

Differences in the Relative Positions of the Two Conserved β Sheets

In addition to the difference of local conformation of the peripheral regions, there are differences in the relative

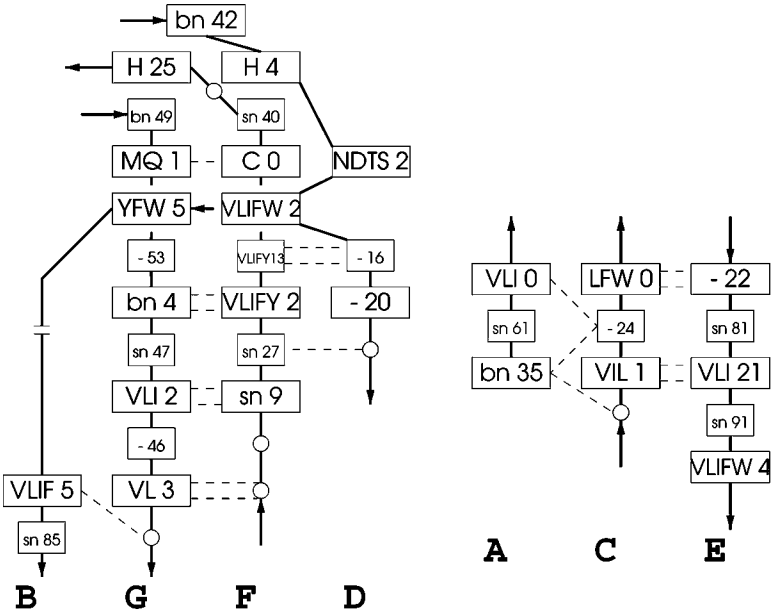

Figure 4. The Key Features of the Conserved Core of the Cupredoxin Family Corresponding to Table 3

Residues B3 (YFW) and CD4 (VLIFW) are shown packing under the binding site residues. Dashed lines show the conserved hydrogen bonds. Within each box we give the sequence residues conserved at the sites and the average accessible surface area of sites in the seven structures. The boxes represent key residues with their possible amino acids and observed accessible surface area inside; large boxes are inward-facing residues which pack into the core or around the binding site, whereas small boxes are outward-facing residues. The circles represent other positions which are not in the buried inner core. Note the topmost box is the fifth and weakest (backbone oxygen) binding residue, and the four (large) underneath it are the tetrahedral binding residues.

positions of the conserved β sheets. The procedure for calculating the differences in their relative positions is as follows. First, the master structure (1plc) is moved to a chosen orientation about the *z* axis. Then, the master structure is moved such that the backbone atoms of the conserved residues in the second sheet have a minimum rms deviation from the *x-y* plane. The master structure is then moved in the plane such that the center of mass of the backbone atoms of conserved residues in the second sheet lies at the origin. The second structure, which is to be compared to the master, is placed such that the backbone of the first sheets of both structures have a minimum rms deviation from each other. Finally, the translations and rotations necessary to move the second sheet of the master structure to the position of the second sheet of the second structure are calculated.

Table 2 shows the relative shifts of the sheets in the structures compared to the positions of the sheets in plastocyanin (1plc). The movements of the sheets relative to each other are large compared to the movements within sheets. This means that although there are very strong structural constraints holding the strands in their positions in the sheets, there is much less of a constraint on the whole sheets to move relative to each other during the course of evolution. This is discussed below.

Sequence Residues in the Conserved Core of the Monomeric Cupredoxins

To investigate the properties of the residues in the conserved core of the proteins, we collected sequences of cupredoxins whose structure is unknown but whose homology to one of the known structures is clear. By aligning other sequences to those of known structure, a more detailed view of the nature of residues allowed at positions in the core can be obtained.

The sequence of each of the structures was matched to sequences in a nonredundant database (Holm and Sander, 1998) using Fasta (Pearson and Lipman, 1988). The few sequences that were found by more than one of the searches were assigned only to that with the strongest similarity. For each search, partial sequences were eliminated; then, those remaining were aligned to the sequence of the structure used for the search using ClustalW (Thompson et al., 1994). The result was seven sequence alignments of nonredundant homologs. The accurate structure-based sequence alignment of the

Table 1. The Structural Alignment of Six Structures to Plastocyanin

Name	PDB	Residues	Aligned	Identical	% Identity	Rms Fit	R Factor	Resolution
Plastocyanin	1plc	99	99	99	100	0	0.15	1.33
Amicyanin	1aac	105	73	20	27	1.1	0.16	1.8
Azurin	2aza	129	63	13	31	2.0	0.17	1.8
Phytocyanin	2cbp	96	55	15	27	1.5	0.14	1.8
Stellacyanin	1jer	109	58	13	22	1.3	0.19	1.6
Pseudoazurin	1paz	123	81	21	26	1.1	0.18	1.55
Rusticyanin	1rcy	155	59	13	22	2.0	0.18	1.9

For each structure the PDB code is shown, followed by the number of residues of the domain, the number which align to plastocyanin, the number which are identical in sequence, the identical residues expressed as a percentage of those aligned, and the root-mean-square deviation of backbone residues in angstroms between the aligned parts of the pair of structures. The last two columns are extra information about the X-ray structure.

Table 2. The Movement in the Sheets Relative to Plastocyanin

Name	Structure	Translation (Å)	Rotation (°)
Plastocyanin	1plc	0	0
Pseudoazurin	1paz	0.3	1.6
Amicyanin	1aac	0.8	5.1
Cucumber protein	2cbp	1.9	6.4
Stellacyanin	1jer	2.0	10.6
Azurin	2aza	3.0	15.7
Rusticyanin	1rcy	3.8	9.3

seven proteins was used to align the seven alignments to each other, creating one large alignment of 77 sequences. This alignment was carefully studied and adjusted using the results of the structural analysis to correct the alignment. Where necessary, further investigation of the structures was carried out to solve specific uncertainties in the alignment.

The alignment of 77 sequences was analyzed to see what variations occurred in the residues of the conserved core and binding site. The occurrence of every amino acid in each column of the alignment was counted. A simple program was written to extract the salient features at each site based on the negative log odds probability, normalized by the composition of the alignment, of getting the observed frequency of amino acids at each site. The full alignment and residue analysis can be seen respectively as supplemental information at http://supfam.org/SUPERFAMILY/alignment_stats.txt; http://supfam.org/SUPERFAMILY/cgi-bin/alignment_stats.cgi?example=y.

Residue and Site Conservation in the Common Cores

The information gathered from the analysis of the structures described above showed that the common core consists of the binding site and parts of the two sheets which support it. In Table 3, we describe the nature and extent of the residue conservation at the sites in the common core. A schematic diagram of the common core structure is shown in Figure 4.

When examining residue conservation, it is convenient to consider not just individual residues but also classes of residues. A classification we have found useful is based on the two correlated properties of residues: (i) the extent to which they are distributed between the surface and interior of proteins and (ii) their free energy of their transfer between water and organic solvents (Miller et al., 1987). This classification put residues into one of three classes: hydrophilic (s), neutral (n), and hydrophobic (b). The residues in the three classes are "s": R. K, E, D, and Q; "n": P, H, Y, G, A, S, and T; and "b": C, V, L, I, M, F, and W.

First we will discuss residue conservation in the set I structures and their homologs. There are 57 residue sites in their common core. Significant residue conservation is found at 50 of these sites (see Table 3 and Figure 4). The four Cu binding residues are absolutely conserved and two residues associated with these at CD1 (P or G) and CD3 (N) are absolutely conserved in all but one or two sequences. There are 18 sites that conserve that have hydrophobic or aromatic residues in 86%–100% (on average 97%) of the known homologs. At most of these sites, the alternative residues are either similar in size, e.g., L or M (B1) and V or I (C4), or have a limited range of volumes, e.g., medium or large hydrophobic residues. On a less restricted level, there are 24 sites where there are limitations on classes from which residues are drawn. Of the 24, 21 are "sn" sites, with residues selected from the hydrophilic or neutral classes in 86%–100% of sequences, and three are "bn" sites, with residues selected from neutral or hydrophobic classes in 93%–100% of cases.

When all seven structures and their homologs are considered together, the size of the common core is reduced to 35 residues; see Table 3. In Table 3 we list the residues found at these sites in Cucumber Basic Protein, stellacyanin, and their nine known sequence homologs. Twenty-four sites have residues that are the same as those found in set I proteins. There are another six sites where the residues are similar to those in set I sequences, e.g., B3 which has YF in set 1 and W in set II. The remaining six sites have residues different to those found in set I sequences. At at least some of these sites, the differences are accommodated by the regions that differ in conformation in the two sets. For example, at site G4, set I sequences have GA residues while set II has VLI; the larger side chains are accommodated by the bulges in the B strand of set II structures.

Linked Conservation of the Active Site and Structure

We have described the extent to which structure and residues are conserved in seven cupredoxins that have known structures and their close homologs. Overall, the conserved structure comprises some 57 residues in the set I structures and some 35 residues in both set I and II. In the two sets, the structure of β sheet one is largely conserved whereas β sheet two conserves only a small central region. Absolute residue conservation occurs at three copper binding sites, and strong conservation occurs at some 16 sites that are in the binding site region or form the core of the protein.

In the cupredoxin structures discussed here, the copper ion is completely buried and has bonds with the side chains of four residues. The binding site is mostly atop sheet one (see Figure 5). A cysteine at the end of the F strand and a methionine at the beginning of the G strand form sulfur bonds with the copper (except in the case of stellacyanin where the methionine is replaced by a glutamine [Hart et al., 1996]). Two histidines, one from the CD loop and one from the FG loop, form bonds to the copper with imidazole nitrogens. In azurin, the carbonyl oxygen of the residue at CD1 is close enough to the copper atom to form a weak fifth bond; in the other structures, this oxygen is somewhat further away (Baker, 1988; Guss et al., 1996).

The coordination geometry of these ligands is that of a distorted tetrahedron (see Figure 6) whose structure is intermediate between those optimal for the two different oxidation states of copper and as such facilitates the electron transport properties of the proteins (Baker, 1988; Colman et al., 1978; Guss and Freeman, 1983). In the different cupredoxins, only very small differences in geometry (<1) are observed (and these have a related difference in their redox potential [Bond et al., 2001;

Table 3. Residue Conservation in Regions that Have the Same Conformation in Monomeric Cupredoxins

	Homologous Set I		Homologous Set II	
Site	Residues	Extent (%)	Residues	Extent (%)
A1-1	bn	93	bn	100
A2-2	sn	89	sn	91
A3-3	VLI	96	VI	100
A4-4	–	–	X	
A5-5	bn	98	X	
A6-6	sn	86	X	
B1-12	LM	93	X	
B2-13	–	–	X	
B3-14	YF	98	W	100
B4-15	–	–	X	
B5-16	PT	88	X	
B6-17	sn	98	X	
B7-18	sn	91	X	
B8-19	VLIMF	98	X	
B9-20	sn	95	X	
B10-21	VLI	95	F	100
B11-22	sn	100	sn	100
C1-24	sn	100	X	
C2-25	–	–	X	
C3-26	sn	96	X	
C4-27	VI	98	L	100
C5-28	–	–	–	–
C6-29	LFW	98	F	100
C7-30	–	–	X	
C8-31	NHP	96	X	
C9-32	sn	93	X	
CD1-36	PG	95	bn	82
CD2-37	H	100	H	100
CD3-38	N	98	NDTS	100
CD4-39	VLIFW	100	V	100
D1-40	VLI	84	–	–
D2-41	VLIF	91	–	–
D3-42	sn	86	X	
E1-69	sn	95	X	
E2-70	sn	84	–	–
E3-71	sn	89	sn	91
E4-72	VLI	86	I	91
E5-73	sn	96	sn	82
E6-74	VLIFW	100	L	100
E7-75	sn	100	X	
F1-79	sn	93	X	
F2-80	Y	100	sn	81
F3-81	sn	89	YH	100
F4-82	VLIFY	100	FY	100
F5-83	VIFY	88	VLI	100
F6-84	C	100	C	100
F7-85	sn	100	sn	100
FG1-87	H	100	H	100
G1-91	G/LIMFY	100	G	100
G2-92	M	100	MQ	100
G3-93	–	–	K	100
G4-94	GA	96	VLI	100
G5-95	sn	88	sn	100
G6-96	VLI	100	VI	100
G7-97	–	–	sn	100
G8-98	VL	98	V	82
G9-99	sn	93	X	

This table shows 57 entries for set I, but no data is given for set II except for the 35 which overlap set I. "–" indicates no significant conservation. "X" indicates residues whose conformation in set II (*continued*)

Table 3. Continued

are different to those in set I. "s," "n," and "b" correspond to surface, neutral, and buried residues, respectively. The numbering in the table corresponds to 1plc; the equivalent regions of all seven structures are listed here using the PDB residue numbering in the "ATOM" records: 1plc: 1–6, 12–22, 24–32, 36–42, 69–75, 79–85, 87, 91–99; 1paz: 3–8, 16–26, 28–36, 39–45, 63–69, 73–79, 81, 85–93; 1aac: 21–26, 28–38, 40–48, 52–58, 77–83, 87–93, 95, 97–105; 2aza: 5–10, 13–23, 28–36, 45–51, 92–98, 107–113, 115, 120–128; 1rcy: 40–45, 52–56, 61–66, 71–79, 84–90, 122–128, 133–139, 141, 147–155; 2cbp: (3–6, 11, 24–25), 27–35, 38–44, 65–71, 73–80, 81, 88–96; 1jer: (5–7, 13, 31–32), 34–42, 45–51, 74–80, 83–92, 93, 98–106.

Baker, 1988; Durley et al., 1993; Guss et al., 1996]). This geometry very largely imposes on the structure of the protein. Apo-cupredoxins have structures that are virtually identical to that of the holoenzyme (Shepard et al., 1993; Garrett et al., 1983; Petratos et al., 1995; Durley et al., 1993).

How is the structure of the copper binding site conserved in the context of conformational differences that affect up to three-quarters of the structures in the different cupredoxins?

In Figure 7 we show for plastocyanin a space-fill drawing of the 20 side chains of (i) the residues that form the base of the active site (in green) and (ii) the conserved residues whose sites are on one diagonal in each of the two β sheets (in red and yellow). Inspection of the conserved residues in this figure shows that together these form a continuous column that runs through the center of the structure. The four copper binding residues are at the top of this column. The other strongly conserved residues run through to near the bottom of the structure.

The loops and residues around the copper ligands also have interlocking sets of hydrogen bonds that are largely conserved. These have been described in previous publications (Bond et al., 2001; Durley et al., 1993; Guss et al., 1996; Baker, 1988).

Under the binding site, two residues pack: B3 which is F, Y, or W in 99% of sequences and CD4 which has V, I, L, F or W in 100% sequences (Table 3). This conserved region then continues to the bottom of the structure with F5 (VILFY, 100%), G6 (VLI 100%), G8 VL (96%), and B10 (VILF 96%) from sheet one and A3 (VLI 96%), C6 (FLW 98%), C4 (VIL 98%), E4 (VIL 85%), and E6 (VILFW 100%) from sheet two (Table 3).

The constraints around the copper atom will, of course, be mainly functional, while those in the center of the protein will be mainly structural. The residues in the core are buried and pack tightly together. As a consequence, there are very strong structural constraints on the residues and only a limited variation is possible without disrupting the fold of the protein and destabilizing it. See also the comparisons, described below, that we make of the pattern of conserved residues in the variable domains.

The copper binding site is formed almost entirely by residues in sheet one and residues in loops that are part of this sheet. Only one residue in sheet two is likely to be important for the binding site: B3 which packs against

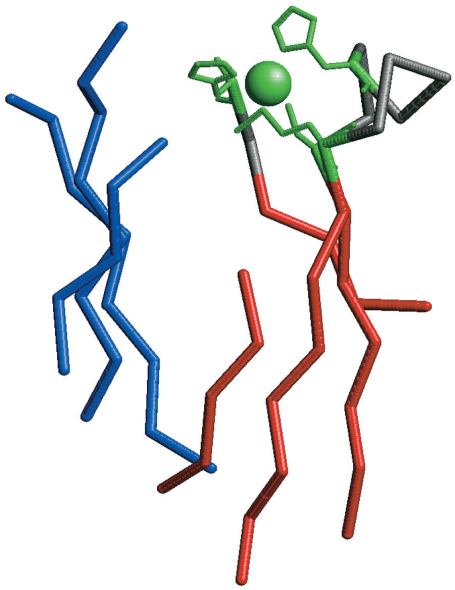

Figure 5. The Copper Binding Site of Plastocyanin

The copper atom and the residues which bind it are green. The backbone of sheet one is red and the backbone of sheet two is blue. The two base residues are in line, one in front of the other underneath the copper atom. The two histidines folding over above are side by side perpendicular to the two base residues.

the copper ligands at F6 C and G2 M. Residues, equivalent to those in strand B in set one, have a quite different conformation in set II structures, but they retain a residue in the same position of the same type and with the same role as set I B3 (Table 3).

Thus, except for B3, the major role of sheet two seems to be the stabilization of the structure of sheet one. This role is consistent with the variety of conformational differences in the regions around the center of the sheet. It also allows movements of sheet two relative to sheet one; for example, in azurin, sheet two it is shifted 3.1° and rotated 15.7° relative to its position in plastocyanin (see above and Table 2). Inspection of the two structures shows that this difference involves sheet two pivoting around the B strand corner of sheet two; this shifts the B strand about 1° and shifts the end of the E strand at the opposite corner 6°.

The Structure of Auracyanin

Subsequent to the work described above the structure of an eighth member of the Cupredoxin family was published: Auracyanin (Bond et al., 2001). This protein is an electron transfer agent in the photosynthetic pathway of certain bacteria. It has a structure close to that of azurin: 89 of the C atoms in auracyanin superpose on the equivalent positions in azurin with an rms difference of 0.8 (Bond et al., 2001).

The common core of the five set I cupredoxins described above contains 57 sites (Table 3). We superposed the main chain coordinates of the equivalent residues in auracyanin on those in plastocyanin. They fit with an rms difference of 1.9 . We also determined the differences in the relative position of the β sheets in the two proteins. Sheet two in auracyanin differs in position relative to sheet two in plastocyanin by a shift of 2.4° and a rotation of 12.4°: values a little smaller than the 3.0°/15.7° in azurin (Table 2).

The common core structure of the five set I cupredoxins considered above comprises 57 sites, and the pattern of residues found at these sites is described in Table 3. We examined the residues at the equivalent sites in auracyanin to determine how closely they fit this pattern. At 51 sites, an exact match is made. At three sites, that are on the surface in our set I structures, C1, D3, and F3, and have sn residues in 86%–100% of sequences (Table 3), auracyanin has hydrophobic residues V, V, and L. Inspection of the auracyanin structure shows that the three sites in auracyanin are buried by loop regions with novel conformations. The other three differences involve L in place of VI (98%) at C4 and two differences on the edge of the core structure: Q in place of PT (88%) at B5, and Q in place of PG (95%) at CD1.

The Pattern Formed by the Conserved Residues in the Cupredoxins Is Very Similar to that Found in Immunoglobulin Variable Domains

In Figure 3 we show a plan of the conserved structures found in set I and set II cupredoxins. Set I structures are formed by two standard β sheets packed face to face. In set II the structures are complicated by irregular conformations of the A and B strands. For the set I structures we see that the conserved buried hydrophobic residues lie along one diagonal of each β sheet. On the uppermost β sheet the conserved residues point down and cluster around a diagonal on the β sheet that runs from bottom left to top right. On the lowermost β sheet the conserved residues point up and cluster around a diagonal on the β sheet that runs from top left to bottom right. We have seen how the twist of the two β sheets brings these residues to the center of the protein (Figure 7).

The variable domains of the immunoglobulins also have two β sheets packed face to face like the set I cupredoxins. Sequences are available for over 5300 different variable domains and examination of these shows that the conserved hydrophobic residues cluster along one diagonal pattern of each β sheet in same manner as found here for the set I cupredoxins (Chothia et al., 1998). Mirny and Shakhnovich (1999) argued that proteins with the same fold have conserved sites at equivalent positions. However, variable domains and cupredoxins do not have the same fold: their chain topologies are quite different. What they both do share is the β sandwich packing. This packing places the residues on the opposite diagonal of each β sheet at the center of the structure where they form the “deep structure” of the protein. This deep structure is where it is most difficult to accommodate mutations and, apart from the active site, the most conserved part of the proteins (Chothia et al., 1998). Hill et al. (2002) made related observations for the residues conserved on two families of four helix proteins that have different chain topologies.

Conclusions

In particular, we identify the common core of the protein and examine the properties that the residues in the core

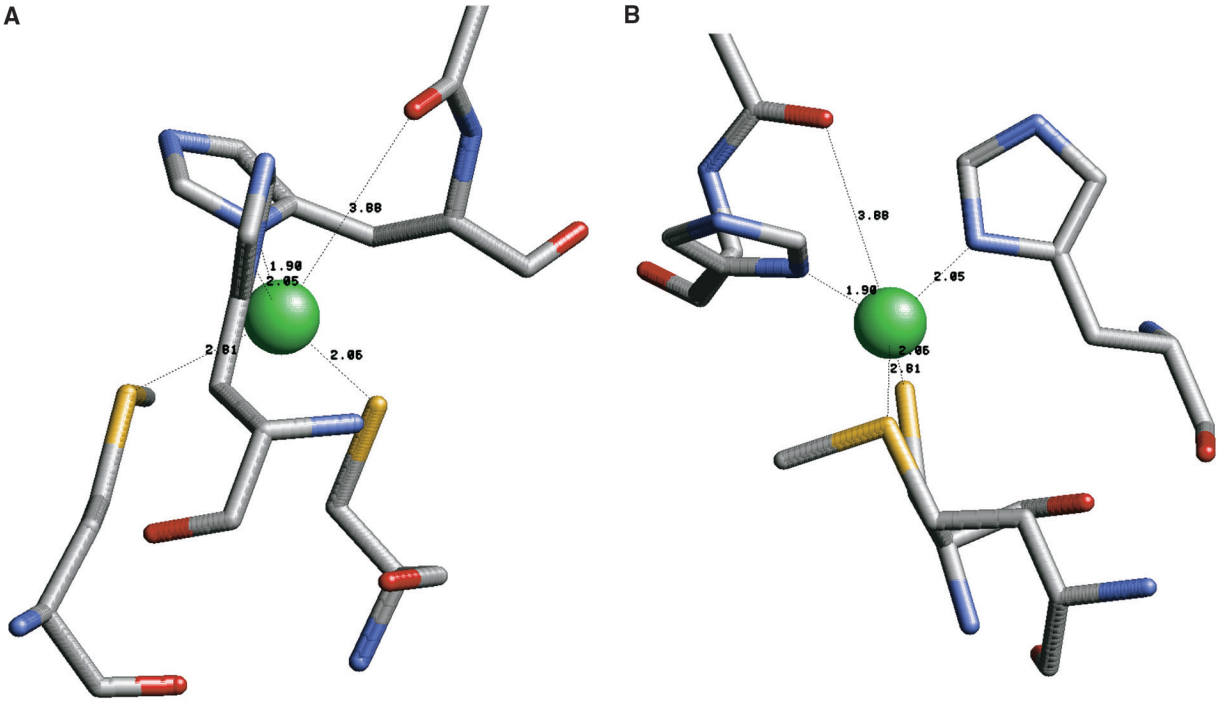

Figure 6. The Tetrahedral Binding Site in Azurin

(A) shows the side view in the same way as Figure 5 and (B) shows it rotated through 90°; now the histidines are in line and the base residues are side by side. The distances are marked in angstroms. Note there is a fifth additional, weaker bond formed by a backbone oxygen atom in this structure (Baker, 1988). In the other structures the oxygen is somewhat further away.

possess. We then relate the conservation at these residues to the residues that form the active site. The cupredoxins share the same diagonal pattern of residue conservation as those of another sandwich fold, the immunoglobulins. Unlike the immunoglobulin variable domains which make use of the variability of their loops for their function, the cupredoxin family has a strong constraint on the binding site required to maintain the function of binding the single copper atom for electron transport. To maintain the tetrahedral binding site, there is strong conservation of the residues in the loops which actually bind the copper atom. Residues in the loops surrounding the binding residues are also important to hold them in the right position; these are also conserved.

The binding site is supported by the rest of the structure. It sits mainly atop one sheet consisting of residues in the loops between strands of the sheet, and as a consequence, this sheet supporting the binding site is also required to maintain the function. Hence the tight-packing, hydrophobic, inward-facing residues of the sheet have a limited number of allowed sequence variations. Although the other sheet provides the second half of the sandwich and complimentary inward-facing residues of the buried core of the structure, it has little involvement in the binding site. Thus, there are some constraints on the second sheet to compliment the first and maintain the fold; freedom to move relative to the binding site requires less conservation in the residues not at the center of sheet two. Furthermore, the position of the sheet relative to the first sheet which supports the binding site is not conserved.

The example of the cupredoxin family shows that the constraints on the core of the protein are intimately coupled to the functional site on the surface of the pro-

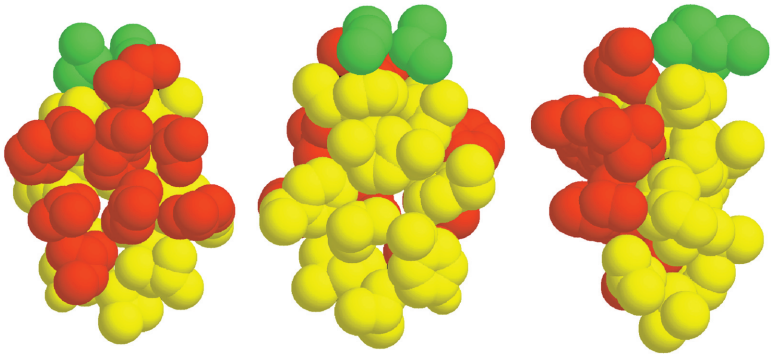

Figure 7. The Side Chains of Residues at the 20 Sites That Are Strongly Conserved

The side chains are of two types: first, functional (Cu binding) residues and their immediate neighbors and, second, residues at sites that pack at the center of the sheet sheet contacts. The core structure of the cupredoxin family (plastocyanin shown) are rendered in space-fill. Atoms which are part of sheet one are yellow, those in sheet two are red. The binding site residues are green; only the two base residues are shown.

tein. Large variations are seen throughout the protein with the exception of key residues which are either directly involved in the active site or, equally importantly, critical in maintaining the tertiary structure supporting the active site.

Acknowledgments

J.G. thanks Veronica Morea for much help and advice about protein structure, and Arthur Lesk for the use of PINQ.

Received: November 26, 2003
Accepted: March 4, 2004
Published: June 8, 2004

References

Adman, E.T. (1991). Copper protein structures. Adv. Protein Chem. *42*, 145–197.

Adman, E.T., Stenkamp, R.E., Sieker, L.C., and Jensen, L.H. (1978). A Crystallographic model for azurin at 3 Å resolution. J. Mol. Biol. *123*, 35–47.

Baker, E.N. (1988). Structure of azurin from Alcaligenes denitrificans refinement at 1.8 (Å) resolution and comparison of the two crystallographically independent molecules. J. Mol. Biol. *203*, 1071–1095.

Bond, C.S., Blankenship, R.E., Freeman, H.C., Guss, J.M., Maher, M.J., Selvaraj, F.M., Wilce, M.C., and Willingham, K.M. (2001). Crystal structure of auracyanin, a "blue" copper protein from the green thermophilic photosynthetic bacterium Chloroflexus aurantiacus. J. Mol. Biol. *306*, 47–67.

Chothia, C., and Lesk, A.M. (1982). The evolution of proteins formed by β-sheets: I. plastocyanin and azurin. J. Mol. Biol. *160*, 309–323.

Chothia, C., Gelfand, I., and Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. J. Mol. Biol. *278*, 457–479.

Colman, P.M., Freeman, H.C., Guss, J.M., Murata, M., Norris, V.A., Ramshaw, J.A.M., and Venkatappa, M.P. (1978). X-ray crystal structure analysis of plastocyanin at 2.7 Å resolution. Nature *272*, 319–324.

Djebli, A., Proctor, P., Blake, R.C., II, and Shoham, M. (1992). Crystallization and preliminary X-ray crystallographic studies of rusticyanin from Thiobacillus ferrooxidans. J. Mol. Biol. *227*, 581–582.

Durley, R., Chen, L., Lim, L.W., Mathews, F.S., and Davidson, V.L. (1993). Crystal structure analysis of amicyanin and apoamicyanin from Paracoccus denitrificans at 2.0 Å and 1.8 Å resolution. Protein Sci. *2*, 739–752.

Garrett, T.P.J., Clingeleffer, D.J., Guss, J.M., Rogers, S., and Freeman, H.C. (1983). The crystal structure of poplar apoplastocyanin at 1.8 Å resolution. J. Biol. Chem. *259*, 2822–2825.

Guss, J.M., and Freeman, H.C. (1983). Structure of oxidized poplar plastocyanin at 1.6 Å resolution. J. Mol. Biol. *169*, 521–563.

Guss, J.M., Bartunik, H.D., and Freeman, H.C. (1992). Accuracy and precision in protein structure analysis: restrained least-squares refinement of poplar plastocyanin at 1.33 Å resolution. Acta Crystallogr. B *48*, 790–811.

Guss, J.M., Merritt, E.A., Phizackerley, R.P., and Freeman, H.C. (1996). The structure of a phytocyanin, the basic blue protein from cucumber, refined at 1.8 Å resolution. J. Mol. Biol. *262*, 686–705.

Hart, P.J., Nersissian, A.M., Herrmann, R.G., Nalbandyan, R.M., Valentine, J.S., and Eisenberg, D. (1996). A missing link in cupredoxins: crystal structure of cucumber stellacyanin at 1.6 Å resolution. Protein Sci. *5*, 2175–2183.

Hill, E.E., Morea, V., and Chothia, C. (2002). Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. J. Mol. Biol. *322*, 205–233.

Holm, L., and Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics *14*, 423–429.

Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. *55*, 379–400.

Leung, Y.C., Chan, C., Reader, J.S., Willis, A.C., van Spanning, R.J., Ferguson, S.J., and Radford, S.E. (1997). The pseudoazurin gene from Thiosphaera pantotropha: analysis of upstream putative regulatory sequences and overexpression in Escherichia coli. Biochem. J. *321*, 699–705.

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. (1987). Interior and surface of monomeric proteins. J. Mol. Biol. *196*, 641–656.

Mirny, L.A., and Shakhnovich, E.I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J. Mol. Biol. *291*, 177–196.

Murphy, M.E., Lindley, P.F., and Adman, E.T. (1997). Structural comparison of cupredoxin domains: domain recycling to construct proteins with novel functions. Protein Sci. *6*, 761–770.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*, 2444–2448.

Petratos, K., Dauter, Z., and Wilson, K.S. (1988). Refinement of the structure of pseudoazurin from Alcaligenes faecalis S6 at 1.55 Å resolution. Acta Crystallogr. B *44*, 628–636.

Petratos, K., Papadovasilaki, M., and Dauter, Z. (1995). The crystal structure of apo-pseudoazurin from Alcaligenes Faecalis S6. FEBS Lett. *368*, 432–434.

Ryden, L.G. (1984). Structure and evolution of the small blue proteins. In Copper Proteins and Copper Enzymes Volume 1, R. Lontie, ed. (Boca Raton FL: CRC Press), pp. 183–214.

Shepard, W.E.B., Kingston, R.L., Anderson, B.F., and Baker, E.N. (1993). Structure of apo-azurin at 1.8 Å resolution. Acta Crystallogr. D Biol. Crystallogr. *49*, 331–342.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). 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*, 4673–4680.

van Driessche, G., Hu, W., van der Werken, G., Selvaraj, F., McManus, J.D., Blankenship, R.E., and van Beeumen, J.J. (1999). Auracyanin, A from the thermopholoc green gliding bacterium Choroflexus aurantiacus represents an unusual class of small blue copper proteins. Protein Sci. *8*, 947–957.

Walter, R.L., Ealick, S.E., Friedman, A.M., Blake, R.C., II, Proctor, P., and Shoham, M. (1996). Multiple wavelength anomalous diffraction (MAD) crystal structure of rusticyanin: a highly oxidizing cupredoxin with extreme acid stability. J. Mol. Biol. *263*, 730–751.