Volume 279, number 1, 73–78 FEBS 09384 February 1991

ADONIS 001457939100049F

A structure-derived sequence pattern for the detection of type I copper binding domains in distantly related proteins

Christos Ouzounis and Chris Sander

EMBL, Meyerhofstrasse 1, D-6900 Heidelberg, Germany

Received 2 November 1990; revised version received 12 December 1990

A structure-based approach to the definition of sequence patterns characteristic of protein domains is presented by example. The approach requires a multiple sequence alignment of a family (or set of related families) as well as at least one three-dimensional structure. The pattern derived does not merely summarize the information in the known sequences but attempts to generalize the pattern specifications based on structural insight. In this example, the pattern-driven database search identified correctly most of the known type I copper-binding domains and detected the presence of a homologous domain in a previously unknown case (CopA protein). The significance of these results is discussed.

Copper binding; Small blue protein; Multicopper oxidase; Sequence motif; Sequence pattern; β-Sheet preference

1. INTRODUCTION

Blue copper proteins constitute a diversified class of proteins including small blue proteins and multicopper oxidases [1]. All members of this class contain a bound 'blue' or type I copper, which absorbs light around 600 nm [2].

The structurally best-known family in this class is the family of small blue proteins, which includes azurins and plastocyanins. It is a group of small, monomeric proteins which contain one copper ion per molecule. These small blue proteins are electron-transport proteins in bacteria and plants [1]. They are single-domain β-sheet sandwiches, composed of eight strands in two sheets, and have predominantly antiparallel β-strand topology. Sequence divergence is so large that current methods fail to align residues which have clearly equivalent positions in the three-dimensional structure [3].

It had been suggested on the basis of sequence similarities [4] and physical properties [2] that multicopper oxidases are remotely related to the small blue proteins. Multicopper oxidases, are large, complex proteins, with many copper atoms per molecule. All appear to have at least one type I copper site. This family includes ascorbate oxidase, laccase, ceruloplasmin, as well as coagulation factors, factor V/factor VII. Multicopper oxidases reduce molecular oxygen to water, with accompanying one-electron transfer from the substrate (reviewed in [1]). Evolutionary relationship of multicopper oxidases to the small blue proteins has been confirmed by the recent X-ray structure of ascorbate oxidase from zucchini [5].

Ascorbate oxidase is a three-domain protein, with the domains strongly diverging in sequence, but each with the same basic plastocyanin/azurin fold [6]. The single type I copper site is in domain III, and a multicopper binding site is formed between domains I and III. The latter has one type II (normal) and two type III (coupled binuclear) copper ions, similar to the copper sites of Cu/Zn-superoxide dismutases (type II) [7] and hemocyanins (type III) [8], respectively.

Searches for patterns are a powerful tool for identification of related molecules in databases [9], provided that the patterns used describe a protein class in a concise and unique way. A simple way of defining patterns is to extract invariant residues from multiple sequence alignments. When three-dimensional structural information is available, a pattern can be made more powerful by analysis and specification of structural requirements at particular positions [10].

The principal difficulty in generating a sequence pattern for the class of blue copper proteins is that residue identities between distant members are few, even within the small blue protein family [3]. The only conserved residues within the whole class are the type I copper ligands, spaced at variable distances, but these are too few to define a specific pattern. Fortunately, based on the fact that structure is more tenaciously conserved than sequence, it is possible to examine the known structures and identify important interactions in the construction of the common type I copper site that are not directly obvious from inspection of the family alignment.

Our approach combines information from multiple

Correspondence address: C. Ouzounis, EMBL, Meyerhofstrasse 1, D-6900 Heidelberg, Germany

sequence alignments and known structures with functional information (e.g. copper ligands) and approximate rules of protein folding. First, patterns characteristic for each family are derived. Subsequently, with the aid of structural data, these are merged into a generalized pattern descriptive of the entire class. The patterns are evolved in an iterative test-and-refine fashion. The final patterns are general enough to identify related sequences that did not play any role in pattern generation - without excessive overprediction - and thus have predictive power.

2. DATA BASES AND COMPUTATIONAL METHODS

All protein sequences were taken from the EMBL/Swissprot collection [11], release 15. The small blue protein sequences in the current database are: 1 amicyanin (AMCY), 11 azurins (AZU1, AZU2, AZUR), 2 pseudoazurins (AZUP), 1 basic blue protein (cusacyanin - BABL), 2 bacterial H8 outer membrane azurin-like proteins (H8, H81), 25 plastocyanins (PLAS, PLAT), and 1 stellacyanin (STEL). The total number of known small blue proteins is 43. The multicopper oxidase sequences are: human ceruloplasmin (CERU - repeats I/II/III), coagulation factor VIII (FA8 - repeats A1/A3), three laccases (LAC1, LAC2) and an ascorbate oxidase (ASO). Factor V [12] and repeat A2 of factor VIII are known not to contain type I copper sites and are not used. The total number of known multicopper oxidases is 9.

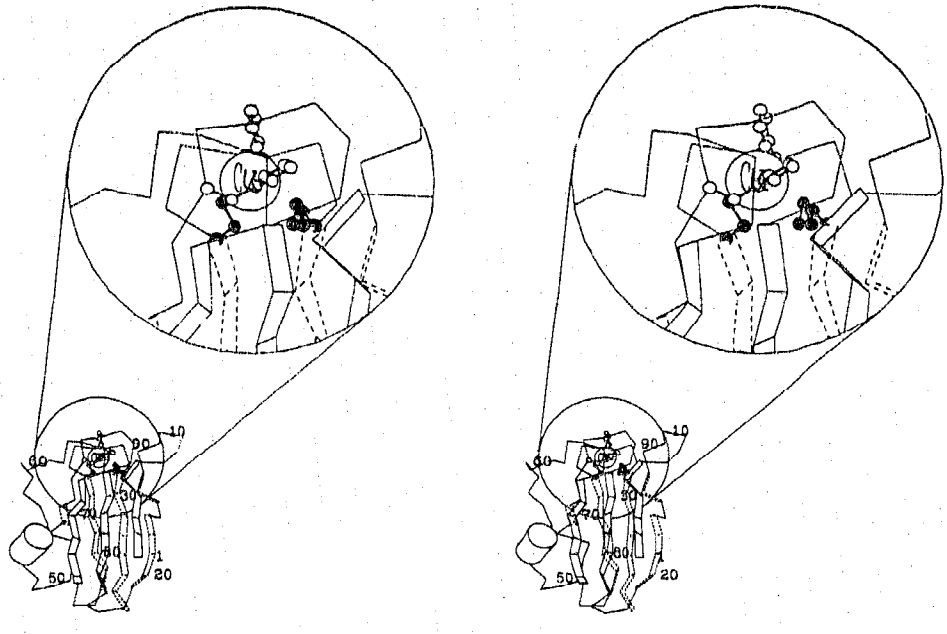
Three-dimensional coordinate sets were taken from the Protein Data Bank [13]. Structures analysed include the azurins from *Pseudomonas aeruginosa* (1AZU) [14] and *Alcaligenes denitrificans* (2AZA) [15], pseudoazurin from *Alcaligenes faecalis* (1PAZ, 2PAZ) [16] and poplar plastocyanin (1PCY, 6PCY) [17]. Experimental secondary structures were derived by the DSSP program [18] and multiple sequence alignments for the azurin/plastocyanin subfamilies were taken from the HSSP database [19].

Software: Various sequence analyses and database searches were performed with the Genetics Computer Group (GCG) Sequence Analysis Software Package [20], version 6.1, and structural analyses with the Insight molecular graphics program (BIOSYM Technologies, San Diego, CA). Pattern generation and searching was facilitated by the Scrutineer protein sequence motif analysis program [21].

3. RESULTS AND DISCUSSION

3.1. *Structural features in the type I copper-binding domains*

The type I copper-binding domains share characteristic features that can be used for the generation of a sequence pattern. The copper ligands are highly conserved, but located at variable distances along the sequences, as the proteins differ substantially in length. The first ligand, always a His, lies just before a β-strand, away in sequence from the other three ligands. The other ligands, Cys, His, and Met, are located in a loop between the last two β-strands in all known structures. All strands adjacent to the copper-binding site belong to one β-sheet, whose sequence variability is lower than that of the other sheet [3] (Fig. 1). The interaction of the residues in the antiparallel β-strands 4 and 7 appears to be important for the construction of the copper site. Thus, this part of the structure is a good candidate on which to base the generation of a sequence pattern (Fig. 2).



Plastocyanin: Cu binding site Plastocyanin: Cu binding site

Fig. 1. Structure cartoon (stereo) of poplar plastocyanin. The eight strands from an irregular, almost closed, barrel-like structure. A helical segment belongs to irregular strand 5. The copper site is magnified, showing the side chains involved in binding copper (Cu) [H37, C84, H87, M92]. Residue conservation in the sheet formed partly by these strands is significantly higher than the opposing one [3].

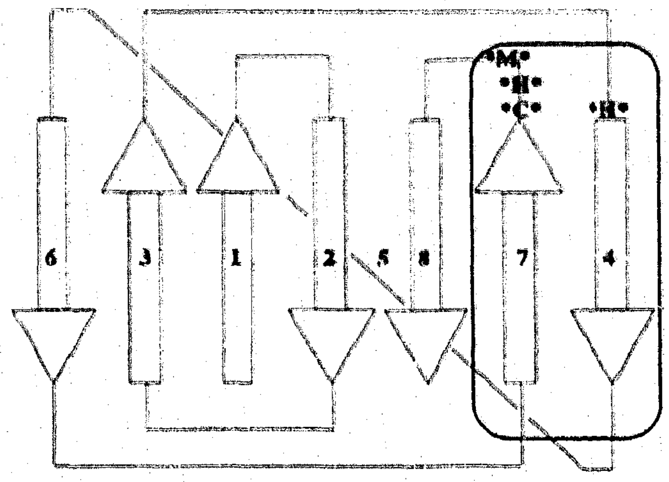


Fig. 2. A topological representation of the common elements in the crystal structures of plastocyanin, azurin, pseudoazurin and the domains of ascorbate oxidase. β-Strands are represented as arrows. Strands 1, 3, and 6 form one sheet, strands 2, 8, 7, and 4 the other. Strand 5 does not take part in β-sheet structure in some of the proteins. The copper-binding residues are clustered in space: the first His ligand lies just before strand 4 (which has at least three residues) and the second ligand, Cys, lies just after strand 7 (which has at least four residues in all known structures). The interaction of these two strands is essential for the formation of the type I copper site, bringing the remote His ligand close to the turn between strands 7 and 8, which contains the other three ligands. The boxed region is therefore taken as the invariant structural feature of the blue copper proteins and used for the generation of a characteristic pattern.

A detailed comparison of two known three-dimensional structures [3,15] had pointed out that, apart from the copper ligands, five other residues are invariant: N47 in azurin (N38 in plastocyanin, position no. 2 in Fig. 3), V49 (V40, no. 4), Y108 (Y80, no. 6), P115 (P86, no. 13/12, respectively) and G123 (G94, not shown). The first three of these residues are indeed in or near to strands 4 and 7. However, some of these residues are no longer invariant in all sequences in the family. Thus, Val at position no. 4 and Gly-123 in azurin are variable in a multiple alignment of more than 40 sequences (Fig. 3); moreover, in some multicopper oxidases Val (no. 4) is replaced by a His that participates in the trinuclear copper site [6] (Fig. 3). Pro (no. 13/12) is at variable positions and not conserved in multicopper oxidases. The Tyr residue (no. 6) at the beginning of strand 7 is highly conserved, but its role for the construction of the type I copper site is not well

		0 1	2345		6789		1 0	1234	5	6789	2 0	
			eee		eeee							
(a)												
Amcy$Paesp	47:	H	NVHF	-----	YDYI	86:	C	TP	H	PF	M	
Azu1$Metj	46:	H	NLVI	-----	YTFF	112:	C	SFPG	H	ATM	M	
Azu2$Metj	46:	H	NWVL	-----	YTYF	112:	C	SYPG	H	FEM	M	
Azup$Alcfa	63:	H	NVES	-----	YLVK	101:	C	TP	H	YAMG	M	
Azup$Paesp	40:	H	NVET	-----	YGFK	78:	C	AP	H	YHMG	M	
Azur$Alcde	46:	H	NWVL	-----	YAYF	112:	C	SFPG	H	WAM	M	
Azur$Alcfa	45:	H	NVVV	-----	YAFF	111:	C	SFPG	H	WSI	M	
Azur$Alcsp	46:	H	NWVL	-----	YAYF	112:	C	SFPG	H	FAL	M	
Azur$Borbr	46:	H	NWVL	-----	YTFF	112:	C	SFPG	H	GAL	M	
Azur$Paeae	66:	H	NWVL	-----	YGFT	132:	C	TFPG	H	SAL	M	
Azur$Paefb	46:	H	NLVI	-----	YGFF	112:	C	SFPG	H	ISM	M	
Azur$Paefc	46:	H	NWVL	-----	YMFT	112:	C	SFPG	H	IAM	M	
Azur$Paefd	46:	H	NWVL	-----	YEFT	112:	C	SFPG	H	NSM	M	
Azur$Paesp	46:	H	NWVL	-----	YAFT	112:	C	SFPG	H	SAM	M	
Bab1$Cucsa	39:	H	NVVV	-----	SYFI	79:	C	NFPG	H	CQSG	M	
H8$Neime	102:	H	NLVI	-----	YIGFA	166:	C	TFPG	H	GAL	M	
H81$Neigo	102:	H	NLVI	-----	YIGFA	166:	C	TFPG	H	GAL	M	
Plas$Anasq	73:	H	NVVF	-----	YSFY	123:	C	EP	H	RGAG	M	
Plas$Anava	39:	H	NVVF	-----	YTFY	89:	C	EP	H	RGAG	M	
Plas$Arath	109:	H	NVVF	-----	YGFY	156:	C	AP	H	QGAG	M	
Plas$Capbu	37:	H	NVVF	-----	YSFY	84:	C	AP	H	QGAG	M	
Plas$Chlfu	38:	H	NIVF	-----	YGYF	83:	C	EP	H	QGAG	M	
Plas$Cucpe	37:	H	NVVF	-----	YSFY	84:	C	SP	H	QGAG	M	
Plas$Cucsa	37:	H	NVVF	-----	YSFY	84:	C	SP	H	QGAG	M	
Plas$Entpr	38:	H	NIVF	-----	YGVY	83:	C	DP	H	SGAG	M	
Plas$Horvu	95:	H	NVVF	-----	YGFY	140:	C	EP	H	AGAG	M	
Plas$Lacsa	37:	H	NVVF	-----	YSFY	84:	C	AP	H	QGAG	M	
Plas$Lycas	108:	H	NVVF	-----	YSFY	155:	C	SP	H	QGAG	M	
Plas$Merpe	37:	H	NVVF	-----	YSFY	84:	C	SP	H	QGAG	M	
Plas$Pea	106:	H	NVVF	-----	YKFY	153:	C	SP	H	QUAG	M	
Plas$Petcr	37:	H	NIVF	-----	YKFY	82:	C	EP	H	AGAG	M	
Plas$Phavu	37:	H	NVVF	-----	YSFY	84:	C	SP	H	QGAG	M	
Plas$Popni	37:	H	NIVF	-----	YSFY	84:	C	SP	H	QGAG	M	
Plas$Rumob	37:	H	NIVF	-----	YSFY	84:	C	SP	H	QGAG	M	
Plas$Samni	37:	H	NVVF	-----	YKFY	84:	C	SP	H	QGAG	M	
Plas$Silpr	103:	H	NVVF	-----	YKFY	150:	C	AP	H	AGAG	M	
Plas$Solcr	37:	H	NVVF	-----	YSFY	84:	C	SP	H	QGAG	M	
Plas$Soltu	37:	H	NVVF	-----	YTFY	84:	C	AP	H	QGAG	M	
Plas$Spiol	37:	H	NVVF	-----	YKFY	84:	C	SP	H	QGAG	M	
Plas$Ulvar	38:	H	NIVF	-----	YGVY	83:	C	EP	H	AGAG	M	
Plas$Vicfa	37:	H	NVVF	-----	YKFY	84:	C	SP	H	QGAG	M	
Plat$Popni	37:	H	NVLF	-----	YTFY	84:	C	SP	H	QGAG	M	
Stel$Rhuve	46:	H	NVDK	-----	KYYI	87:	C	GVPK	H	CDLG	Q	<
(b)												
Aso$Cucsa	480:	H	PWHL	-----	WAFH	543:	C	HIEP	H	LHMG	M	
Ceru$Human	295:	H	AAFF	-----	HMLS	338:	C	QNLN	H	LKAG	L	<
Ceru$Human	656:	H	GIYF	-----	FNVE	699:	C	LTTD	H	YTGG	M	
Ceru$Human	994:	H	TVHF	-----	WLLH	1040:	C	HVTD	H	IHAG	M	
Copa$Paesy	542:	H	PIHL	-----	HAYH	591:	C	HLLY	H	MEMG	M	
Fa8$Human	286:	H	SIFL	-----	FLLF	329:	C	HISS	H	QHDG	M	
Fa8$Human	1973:	H	SIHF	-----	WRVE	2019:	C	LIGE	H	LHAG	M	
Lac1$Aspni	508:	H	PIHK	-----	SILH	586:	C	HIAS	H	QMGG	M	
Lac1$Neucr	477:	H	PIHL	-----	WLMH	549:	C	HIAW	H	VSGG	L	<
Lac2$Neucr	477:	H	PIHL	-----	WLMH	549:	C	HIAW	H	VSGG	L	<
		H	NF		F F		C		H		M	
			GI		W I						L	
			TL		Y L							
			PM		M							
			SV		V							
			W		W							
			Y		Y							

Fig. 3. Sequences from structurally and functionally important elements in all blue copper proteins. All small blue proteins (a) and multicopper oxidases (b) from Swissprot Release 15 are included. The notation of Swissprot entry identifiers is 'protein$species'. The three repeats of ceruloplasmin and two (A1 and A3) repeats of factor VIII are listed as separate entries. The sequence patterns were based on two regions that contain the copper ligands and the adjacent strands (boxed region in Fig. 2). At the top, positions in the pattern are numbered sequentially 1-20. The first region (pos. 1-3), contains the first copper ligand followed by strand 4 (pos. 3-5). The second region (pos. 6-20), is strand 7 (post. 6-9) followed by a loop with the remaining three copper ligands (pos. 10-20). Both strands are indicated by the letter 'e' (extended). The numbers in vertical columns refer to the sequence positions for the first (His) and second (Cys) ligands. Amino acids occurring in structurally and functionally important positions are summarized at the bottom. In defining a sequence pattern, focus is applied to residues which are invariant and residues that appear to have a structural role. The type I ligands are the only strictly conserved residues (H, C, H, M at positions 1, 10, 15, 20) – except for the Met ligand in stellacyanin, in the first repeat of ceruloplasmin, and the two fungal laccases (<). Asp (N) at position 2 is an invariant residue in all small blue copper proteins, but can be another small residue (AGTPS) in the multicopper oxidases. Tyr (Y) at position 6 is also highly conserved in all small blue copper proteins, but is conservatively substituted in multicopper oxidases. The ligand loop (pos. 10-20) is so variable that only the ligand residues and length restrictions are considered as common.

understood [22]; it is conservatively replaced by Trp or Phe in multicopper oxidases (Fig. 3). These examples illustrate the difficulty of identifying precisely structural requirements with predictive power on the basis of a limited set of sequences. It is essential to add understanding of structural principles so that alternative residues at particular sites be predicted.

3.2. *Separate sequence patterns for the two blue copper protein families*

The separate sequence pattern for the first of the two families is based on the comparative analysis of three plastocyanin/azurin structures and more than 40 sequences [3,15]. Less information is available for the multicopper oxidases with only one known structure and less than 10 known sequences. Yet, for both families the separate patterns derived (example in Fig. 4) are simple and perform well in terms of identifying all known protein domains of this type with no false positives.

There are, however, two problems with such simple sequence patterns derived primarily on the basis of sequence consensus. First, the expected conservation might be violated by future sequence data if the current set of sequences has not exhausted the limits of what is permissible; so the patterns may be too limited in scope, leading to future false negative predictions. Second, the sequence patterns are different for each family although the basic protein fold as well as the copper site are clearly common to both.

3.3. *Structural invariance in the type I site and a generalized pattern*

In order to generate a general pattern that describes the entire blue copper protein class the ligand residues alone do not suffice - the pattern would be too permissive. Instead, either a more complicated sequence profile would have to be constructed or the structural characteristics of the type I copper site must be taken into account carefully. We opt for the latter, as consensus sequence profiles merely reflect the sequences they are based on and lack generality or predictive power.

The following considerations went into the construction of the generalized pattern. The solvent accessible surface areas of residues at positions no. 3 and no. 5 (strand 4) and no. 6 and no. 8 (strand 7) (see Fig. 3) are small and these residues make important hydrophobic contacts on the interior face of the β-sheet [3,19]. The observed sequence variation at positions no. 3 (AILVW) and no. 8 (FLMVY), is extended to the set of large hydrophobics (FILMVWY), anticipating further variation, but dropping A at position no. 3, at the cost of not detecting repeat I of ceruloplasmin. Position no. 5 (FILVSTK) is not used at all, as the unusual presence of S, T (two pseudoazurins) and K (stellacyanin, one laccase) is not understood. The observed variation at position no. 6 (FWY;S in blue basic protein and a laccase, and K in stellacyanin) is kept as restricted to the large aromatics, dropping S and K at the cost of losing these three proteins. As an additional structural constraint we require that the sequences of the two antiparallel strands 4 and 7 (positions nos 3-5 and nos. 6-9, respectively), should have an average value of the preference parameter for antiparallel beta strands [23] greater than 1.00. This lower limit was set according to the results of a simple parameter search (not shown). For position no. 2, which is invariant in small blue proteins, any small, non-hydrophobic residue is allowed. Finally, the pattern specifications for ligands were as for the simple sequence patterns (Fig. 4), except that Gln in place of Met, as in stellacyanin, was dropped, while Leu, as in two laccases and in a ceruloplasmin repeat, was allowed. Interestingly, mutation experiments on azurin have shown that Met actually is not an essential component of a blue copper site [24].

Although further restrictions based on the multiple sequence alignment could be added, especially in the ligand loop (positions 11-14 and 16-19; Fig. 3), we chose to keep residue occurrence constraints to a minimum, in order to retain predictive power. The general pattern is shown in Fig. 5.

The validity of a pattern is tested by its ability to identify all sequences it was derived from, and at the same time not identifying any irrelevant sequences. There are

Small Blue Copper Protein Pattern	Multicopper Oxidase Pattern
HN{30 70}Y{2 3}C{1 2}F{0 1}H{2 4}[MQ]	H[PSGTA]{4 4}G{30 70}C{4 4}H{3 3}G[ML]
43 True positives (SWISSPROT 15):	9 True positives (SWISSPROT 15):
Amicyanin (1)	Ceruloplasmin repeats (I/II/III) (3)
Azurins (11)	FactorVIII repeats (I/III) (2)
Pseudoazurins (2)	CopA copper resistance protein (1)
Basic blue protein (1)	Laccases (2)
H8 Outer Membrane Proteins (2)	AOase (domainIII) (1)
Plastocyanins (25)	
Stellacyanin (1)	
No true negatives or false positives	

Fig. 4. Separate sequence patterns for the small blue copper protein and the multicopper oxidases. The sequence patterns are expressed in terms of residue occurrence at specific positions, separated by gaps of fixed or variable length. The patterns are characteristic for the family they were derived from. Pattern searches in the Swissprot-15 database yield no false positives. An unexpected and probably 'true' positive for the multicopper oxidase pattern is the CopA copper resistance protein from *Pseudomonas syringae*.

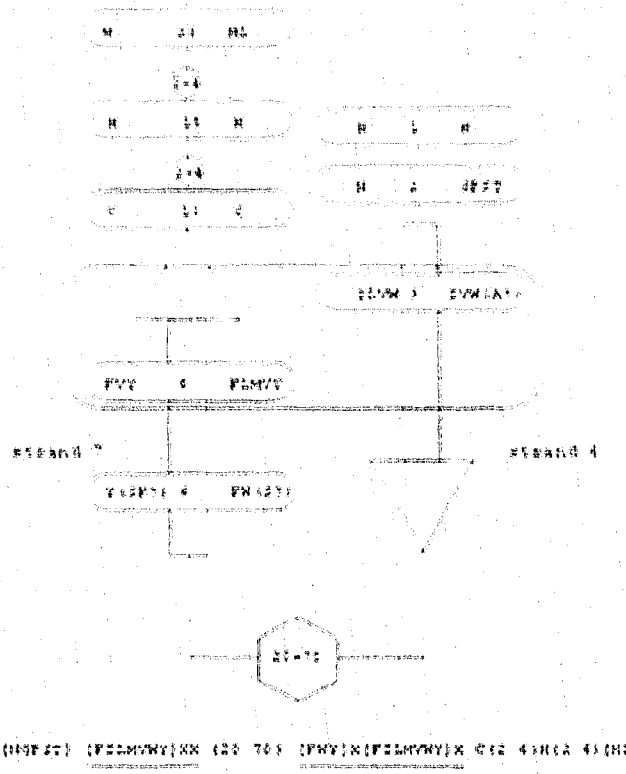


Fig. 5. The generalized pattern for the combined class of blue copper proteins incorporates structural features. The notation in the search pattern (bottom) is: [NP] = N or P at this position; X- = any residue at this position; {20 70} = gap of length 20–70 residues; underlined region = restrictions on average residue properties specified separately (see below). The structural cartoon above the pattern covers the region of strands 4 and 7. Residue occurrences in the plastocyanins/azurins are to the left, and those in the multicopper oxidases to the right of the pattern position numbers (1–20, as in Fig. 3) in the elongated oval buttons. Variable gap lengths are in hexagons. The boxed region encloses the two positions (3 and 8) where the occurrence of any large hydrophobic amino acid is allowed. The question mark in position 6 and 3 means that these amino acids are not included in the pattern. In the underlined regions a minimum antiparallel β preference is applied as an additional constraint. Thus, structural information is included in two ways: (i) by noting that the hydrophobic character is conserved at inside (removed from solvent) β-strand positions; the alternating inside/outside character of the β-strands is apparent in the residue solvent accessibilities (pos/access: 3/0.0, 4/16.0, 5/3.8 for strand 4 and 6/5.8, 7/17.5, 8/1.0, 9/31.0 for strand 7 – calculated as Å^2 of accessible surface area averaged over the six known structures of 2AZA/1AZU/1PAZ/2PAZ/1PCY/6PCY), and (ii) by requiring a minimum antiparallel β-strand preference over each strand (single residue preferences, derived statistically [19] are as follows: W:1.75, Y:1.68, I:1.54, V:1.53, T:1.30, L:1.26, C:1.24, F:1.23, Q:1.18, H:1.12, M:1.09, R:1.02, A:0.90, S:0.87, K:0.74, E:0.62, N:0.62, G:0.56, D:0.47, P:0.42).

three true negatives in the database searches, the basic blue protein (BABL) and stellacyanin (STEL) and a laccase (LAC1). The positions along the sequence of these proteins that prevent their detection are, in BABL, position 6 and, in STEL, positions 6 and 20, as well as low β-strand preferences in positions 3–5, and in LAC1, position 6. Repeat I of ceruloplasmin does not match the pattern at positions 2 and 3, but this domain may have lost its copper binding site and therefore allow larger shifts in tertiary structure. Such cases raise new questions about the structure and function of the particular protein.

Unexpected positives in a database search can mean new discoveries. Two more protein sequences were detected in the latest Swissprot database release: two 'probable' RNA-directed RNA polymerases (sequence code rrpo$crv, sequence position 546–611; and rrpo$tbsvc, 546–611). They belong to a group of five such sequence-related polymerases from plant viruses (the others are rrpo$cnv, rrpo$carmv, and rrpo$tcv) which thus may contain a domain structurally homologous to the type I copper binding proteins. Alternatively, these hits may be false positives.

In another such case, both the simple sequence multicopper oxidase pattern and the generalized sequence pattern detected, unexpectedly, the CopA copper resistance protein encoded by a plasmid in *Pseudomonas syringae* [25]. Closer examination of the CopA sequence revealed not only conservation of the type I and the trinuclear copper site ligands, but also significant sequence similarity to multicopper oxidases, e.g. ascorbate oxidase [26]. These facts had not been reported by the original authors [25]. In a subsequent release of the Swissprot database, the CopA entry was annotated to reflect this similarity (A. Bairoch, personal communication). Details of the biological function of CopA in plasmid-mediated copper resistance are not well understood [27].

4. CONCLUSION

The strongest evidence for an evolutionary connection between the two families of blue copper proteins comes from the discovery of plastocyanin-like domains in the crystal structure of ascorbate oxidase [28]. The fact, demonstrated here, that a copper-binding domain with such wide sequence variation can be successfully described by a succinct sequence pattern is consistent with such an evolutionary relationship.

Our approach goes beyond the definition of sequence profiles or consensus sequences based on multiple sequence alignments in that it attempts to generalize structural observations. The risk is possible overprediction, the possible gain is predictive ability. By identifying structurally important regions in related proteins, allowed substitutions and residue properties averaged over a region can be predicted and incorporated in appropriately designed patterns. This structure-based approach might result in a simplified and consise description of structural and functional domains in proteins.

Acknowledgements: We thank Peter Sibbald for an introduction to his research software Scrutineer, Arthur Lesk for the plastocyanin stereo diagram and Matti Saraste for stimulating discussions.

REFERENCES

[1] Ryden, L. (1988) in: Oxidases and Related Redox Systems (King, T.S., Mason, H.S. and Morrison, M. eds) pp. 349–366, Alan R. Liss.
[2] Malmström, B.G. (1982) Annu. Rev. Biochem. 51, 21–59.
[3] Chothia, C. and Lesk, A.M. (1982) J. Mol. Biol. 160, 309–323.
[4] Ryden, L. (1982) Proc. Natl. Acad. Sci. USA 79, 6767–6771.
[5] Messerschmidt, A., Rossi, A., Ladenstein, R., Huber, R., Bolognesi, M., Gatti, G., Marchesini, A., Petruzzelli, R. and Finazzi-Agro, A. (1989) J. Mol. Biol. 206, 513–529.
[6] Huber, R. (1989) Angew. Chem. Int. Ed. Engl. 28, 848–869.
[7] Tainer, J.A., Getzoff, E.D., Richardson, J.S. and Richardson, D.C. (1983) Nature 306, 284–286.
[8] Volbeda, A. and Hol, W.G.J. (1989) J. Mol. Biol. 206, 531–546.
[9] Hodgman, T.C. (1989) Comput. Applic. Biosci. 5, 1–13.
[10] Taylor, W.R. (1988) Protein Engineering 2, 77–86.
[11] Bairoch, A. and EMBL Data Library Staff (1989) Swiss-Prot Release Notes and User Manual, Releases 12 and 13, EMBL Heidelberg, Germany.
[12] Church, W.R., Jernigan, R.L., Toole, J., Hewick, R.M., Knopf, J., Knutson, G.J., Nesheim, M.E., Mann, K.G. and Fass, D.N. (1984) Proc. Natl. Acad. Sci. USA 81, 6934–6937.
[13] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[14] Adman, E.T., Stenkamp, R.E., Sieker, L.C. and Jensen, L.H. (1978) J. Mol. Biol. 123, 35–47.
[15] Norris, G.E., Anderson, B.F. and Baker, E.N. (1983) J. Mol. Biol. 165, 501–521.
[16] Petratos, K., Banner, D.W., Beppu, T., Wilson, K.S. and Tsernoglou, D. (1987) FEBS Lett. 218, 209–214.
[17] Colman, P.M., Freeman, H.C., Guss, J.M., Murata, M., Norris, V.A., Ramshaw, J.A.M. and Venkatappa, M.P. (1978) Nature 272, 319–324.
[18] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[19] Sander, C. and Schneider, R. (1990) Proteins (in press).
[20] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[21] Sibbald, P.R. and Argos, P. (1990) Comput. Applic. Biosci. 6, 279–288.
[22] Guss, J.M. and Freeman, H.C. (1983) J. Mol. Biol. 169, 521–563.
[23] Lifson, S. and Sander, C. (1979) Nature 282, 109–111.
[24] Karlsson, B.G., Aasa, R., Malmström, B.G., and Lundberg, L.G. (1989) FEBS Lett. 253, 99–102.
[25] Mellano, M.A. and Cooksey, D.A. (1988) J. Bacteriol. 170, 2879–2883.
[26] Ohkawa, J., Okada, N., Shinmyo, A. and Takano, M. (1989) Proc. Natl. Acad. Sci. USA 86, 1239–1243.
[27] Silver, S. and Misra, T.K. (1988) Annu. Rev. Microbiol. 42, 717–743.
[28] Messerschmidt, A. and Huber, R. (1990) Eur. J. Biochem. 187, 341–352.