Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The architecture of metal coordination groups in proteins

Marjorie M. Harding

Institute of Cell and Molecular Biology, University of Edinburgh, Michael Swann Building, Mayfield Road, Edinburgh EH9 3JR, Scotland

Correspondence e-mail: marjorie.harding@ed.ac.uk

Received 12 September 2003
Accepted 20 February 2004

A set of tables is presented and a survey given of the architecture of metal coordination groups in a representative set of protein structures from the Protein Data Bank [Bernstein *et al.* (1977), *J. Mol. Biol.* **112**, 535–542; Berman *et al.* (2000), *Nucleic Acids Res.* **28**, 235–242]. The structures have been determined to a resolution of 2.5 Å or better; the metals considered are Ca, Mg, Mn, Fe, Cu, Zn, Na and K, with particular emphasis on Ca and Zn and the exclusion of haem groups and Fe/S clusters; the proteins are a representative set in which none has more than 30% sequence identity with any other. In them the metal is coordinated by several donor groups from different amino-acid residues in the protein chain and often also by water or other small molecules. The tables, for ∼600 metal coordination groups, include information on the conformations of the protein chain in the region around the metal and reliability indicators. They illustrate the wide variety of coordination numbers, chelate-loop sizes and other properties and the different characteristics of different metals. They show that glycine has a particular significance in the position adjacent to a donor residue, especially in Ca coordination groups. They also show that metal coordination does not appear to lead to significant distortions of the torsion angles φ, ψ from their normally allowed values. Very few metal coordination groups occur more than once in the representative set and when they do they are usually related in fold and function; they have similar but not necessarily identical conformations. However, individual chelate loops, for example Zn(−C−X−X'−C−), in which both cysteines are coordinated to Zn through S, and X and X' are any amino acids, are repeated frequently in many different and unrelated proteins. Not all chelate loops with the same composition have the same conformation, but for smaller loops there are usually one or two strongly preferred and well defined conformations. Quite frequently more than one metal coordination group is associated with one protein chain; these proteins are identified.

1. Introduction

Metal atoms or ions occur widely in association with proteins and have a variety of functions. In some cases the metal is part of the active site for a catalytic process; in others the metal appears to play a role in maintaining structure. Knowledge and understanding of the architecture of protein molecules (see, for example, Lesk, 2001) play a key role in understanding their function. In a similar way, knowledge of the architecture of different metal coordination groups within proteins is important in addition to an understanding of the different chemical behaviour of the metals. *The Biological Chemistry of*

the Elements (Frausto da Silva & Williams, 1991) provides an excellent account of both the chemical behaviour of the different metals and the biological significance of metal coordination groups.

Two metal coordination groups are illustrated in Fig. 1. The first aspects of interest are the number and nature of the donor groups around the metal atom or ion, the metal-to-donor atom distances and the angles between metal–donor bonds. In very many cases the protein molecule is, in the coordination chemist's terminology, a multidentate ligand, so we are also interested in the number and nature of the amino-acid donor groups from the protein chain, their relative positions in the amino-acid sequence and the size and conformation of the resulting chelate rings. How does the protein-chain conformation adapt to the requirements of the metal coordination? We want to know what generalizations, if any, can be made about these different properties and how far they can be predicted. Vallee & Auld (1990) commented on the significance of the spacing between donor residues in 12 zinc enzyme structures and suggested how the observed long and short spacings contributed to effectiveness in catalytic function; a wealth of additional data is now available.

This paper presents a set of tables which allow comparisons of donor groups, chelate-loop sizes and conformations in ∼600 metal coordination groups. They are for the metals Ca, Mg, Mn, Fe, Cu, Zn, Na, K, the most commonly occurring metals in biological chemistry and the commonest in the Protein Data Bank [PDB (Bernstein *et al.*, 1977); available through the RCSB (Berman *et al.*, 2000)], which is the primary source of the information used here. For Na and K the borderline between a 'coordination compound' and an electrostatic association of ions is certainly debatable, but regardless of the description of the bonding it is useful to describe the geometric situation around these ions as found in protein crystals. This study concentrates on coordination groups where amino-acid side chains or the main-chain carbonyl group provide donor atoms: haem groups, iron–sulfur clusters and chlorophyll derivatives have been excluded (there are some specialized articles about these, *e.g.* Parisini *et al.*, 1999; Maher *et al.*, 1999; Chong *et al.*, 1999; see also Huber *et al.*, 2001). In previous articles (Harding, 2000, 2001), data on coordination numbers and on metal-to-donor atom distances and angles in proteins have been gathered for these eight metals. For all the coordination groups in the tables presented here, the distances, angles and coordination group shape may be found at http://tanna.bch.ed.ac.uk/; for a more extensive range of metal coordination groups the Metalloprotein Database (MDB) is valuable (Castagnetto *et al.*, 2002).

The definition of a coordination group used here requires the donor atoms to be within specified target distances of the metal atom; this definition is objective, but it is narrow and it excludes the second and third coordination shells around a metal atom that are generally considered to be important in enzyme activity (see, for example, Duda *et al.*, 2003). Similarly, in building a library of structural motifs of metal coordination sites with catalytic activity, MacArthur & Thornton (2002) include functional groups substantially further from the metal than a simple bond distance; their motifs, including the three-dimensional coordinates of donor atoms, can be used as templates or probes for a systematic classification of sites.

The results and discussion given here are based entirely on a 'representative set' of proteins, a set within which none has more than 30% sequence identity with any other. There are various difficulties in comparing the frequency of occurrence of different coordination groups or donor patterns in proteins using the PDB. The proteins whose structures have been deposited in the PDB are far from a random sample. The use of a 'representative set' of proteins is a simple expedient to obtain a fairly diverse sample, but is based on sequence similarity of the whole protein chain, not just the part in the immediate vicinity of and more relevant to the metal coordination group. Furthermore, many protein crystals contain two or more copies of the protein molecule in the crystal asymmetric unit; allowance for this has been made in several different ways at different stages of this project. Even with these allowances, the set of proteins in the PDB is by no means a random selection; small differences in statistics of distributions should not be thought to be significant, only broad trends.

2. Some definitions

'Target distances' for different types of metal–donor atom bond were based on the distances observed in accurately determined small-molecule crystal structures (Harding, 2001, 2002). The metal coordination number is the number of donor atoms within the target distance + 0.75 Å; some of these are normally donor atoms from amino-acid side chains within the protein or the O atoms of main-chain carbonyl groups, but

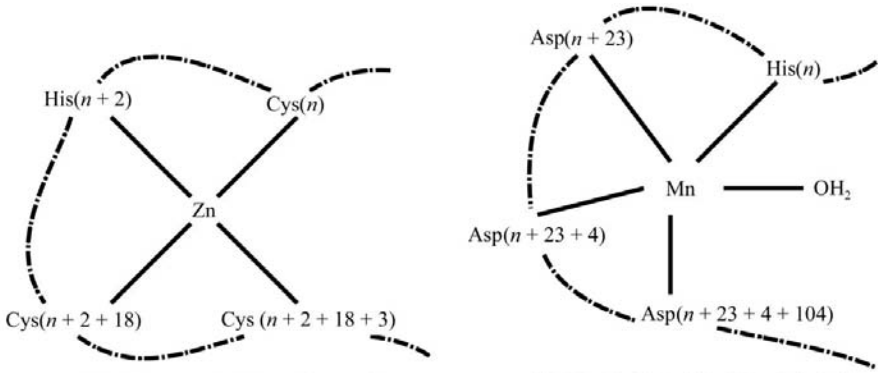

Figure 1
Schematic illustration of coordination groups; see text for definitions.

they may also include water-molecule O atoms or atoms from other non-protein small molecules present at the metal site.

Metal coordination groups can be drawn very schematically, as in Fig. 1. In this work, a donor atom is defined entirely on geometric criteria: it must be within the already established target distance + 0.75 Å of the metal atom. Single-letter amino-acid codes are used to specify the donor groups (of the protein) and O indicates main-chain carbonyl O atom as a donor. We thus describe the coordination group shown in Fig. 1(*a*) as CHCC Zn 2 18 3, since Zn is coordinated to the sulfur of cysteine (*n*) with sequence number *n*, N of histidine (*n* + 2), S of cysteine (*n* + 2 + 18) and S of cysteine (*n* + 2 + 18 + 3). The total coordination number (CN) is 4. The sequence differences, seqdif, in the three chelate loops are 2, 18 and 3. In the first chelate loop there are seven backbone atoms of the donor residues and the residue between, as well as atoms from both side chains, making a ring of 14 atoms altogether (if N^{ε} of histidine coordinates). The relative sequence number of each donor amino acid in the coordination group is given by relseq; in this example there are cysteines at relseq = 0, 20 and 23, and histidine at relseq = 2. nspan is the sequence-number difference between the last and first amino-acid donors, which is the sum of all the seqdifs between them; nspan is 23 in this example. Many metal coordination groups also include water molecules or donor atoms from small non-protein molecules, for example as in Fig. 1(*b*). It has also been useful to look at the chelate loops, which are the building blocks of coordination groups, *i.e.* the adjacent pairs of donors, such as CH 2, HC 18, CC 3 in this example. For full identification of a particular coordination group or a chelate loop, the protein name and the residue number and chain letter of the first amino acid must be given, *e.g.* the above group occurs in 1a1i at A137 (and another at A165). In comparing the compositions of metal coordination groups it has been necessary to treat the carboxylate group as one donor whether it is monodentate or bidentate, since the distinction between these is unreliable in structures determined at lower resolutions.

3. Methods and procedures

The basis for generating the coordination-group tables is the program *MP* (Harding, 2001), which reads a PDB file, extracts the coordinates and occupancy of each metal atom and of all atoms within 3.6 Å of the metal atom and summarizes all the coordination information. Lists of PDB codes were obtained using the Jena Image Library search facility (http://www.imb-jena/ImgLibPDB/pages/hetDir/PSE2HET.shtml) for structures containing each of the metals. From these lists protein and protein–nucleic acid complexes were selected with structures determined by diffraction to a resolution $\leq$ 2.5 Å and the program *MP* run for all that were available in the RCSB release of July 2001 (except that the July 2002 release was used for potassium proteins in order to augment the very small number of available structures). Additional smaller programs then gave information on coordination group descriptions for the full lists or for selections from them. One such selection is a 'representative set' which excludes any structure which has more than 30% sequence identity with any other in the set; the culled PDB files of Dunbrack (2001) were used to make this selection.

3.1. Concerning coordination-group definition

An atom is identified here as a donor when its distance from the metal atom is within target distance + tolerance. The target distances have been carefully established using appropriate small-molecule compounds from the Cambridge Structural Database (CSD, Allen & Kennard, 1993*a*,*b*) and checking against high-resolution protein structures (Harding, 1999, 2000, 2001; the results of a check using 167 protein structures determined up to April 2003, with resolution of 1.25 Å or better, are given at http://tanna.bch.ed.ac.uk). Errors in determination of atom positions, especially in low-resolution structures, might result in incorrect decisions on whether or not an atom is within the metal coordination group. For this reason, structures determined at resolutions less than 2.5 Å are not included. The tolerance was set at 0.75 Å after examining the distribution of the differences between observed and target distances. When the resolution is <1.8 Å there should be no 'wrong decisions' about whether an atom is within the metal coordination group; when the resolution is poorer, but still <2.5 Å, a few 'wrong decisions' will inevitably be made, but their number should be well under 5% of the whole. Less reliable decisions are indicated by a high r.m.s. deviation from target distances and/or additional donor atoms within distances up to target + 0.95 Å. A few metal atoms in the coordination-group tables have coordination numbers lower than would normally be expected (*i.e.* <5 for Ca, <4 for Mg, Mn, Fe and Zn and <3 for Cu). Usually this is the result of a failure to identify a donor group such as a water molecule in the electron-density map, but in a few cases it could be the result of a shortcoming in the software, which does not (yet) detect when the metal atom is coordinated to a donor group in a neighbouring asymmetric unit of the crystal. Metal coordination groups in which any atoms are disordered or have occupancy less than 0.7 are omitted.

3.2. Redundant protein chains

There are frequently two or more identical protein chains within the crystal asymmetric unit. In comparisons of chelate loops these were all included initially and the r.m.s. difference in φ and ψ evaluated over the range relseq = -10 to a relseq of 10 beyond the end of the chelate loop; when the r.m.s. difference in φ and ψ over the range was less than 15°, the redundant chains were eliminated. In a few cases the r.m.s. difference was 20–25°, which probably represents uncertainties in interpretation of maps rather than true differences in conformation. Subsequently, whenever the PDB file included two or more protein chains with equivalent numbering, only the first was used. Even this does not work perfectly. There are a few cases, mostly with resolution in the range 2–2.5 Å, in which different coordination groups are identified for otherwise equivalent chains within the crystal asymmetric unit. (In

Table 1
A small part of the deposited Table 1D for Zn coordination groups illustrating some of the information stored.

Table 1D, with the complete tables for eight metals, has been deposited as supplementary material (and is also available at http://tanna.bch.ed.ac.uk/arch/). np is the number of donors from the protein chain; nw is the number of water molecules; nn is the number of non-protein donor groups; dons are the amino-acid donor groups in the order in which they occur in the polypeptide chain, using the normal single-letter codes for amino acids and O for the main-chain carbonyl O atom; sd1 to sd7 are the seqdifs (−99 signifies that the donors are from two different polypeptide chains, −1 is given when the second donor is water or another non-amino-acid donor); his indicates whether histidine coordination is by ND or NE; cn is the total number of donor groups, including water molecules and small-molecule ligands, always treating carboxylate as one group (the coordination number, as it would be defined by a chemist, is then number of donor groups + number of bidentate carboxylate groups); r.m.s. is the r.m.s. deviation of metal-to-donor atom distances within the coordination sphere from target distances, which is a useful indicator of quality (0 is good, 0.5 is poor); res is the resolution (Å) of the structure determination; carbi indicates bidentate carboxylate groups, *e.g.* ..b. indicates that the third of four donor groups appears to be a bidentate carboxylate. (For additional information stored, see §4.)

cngpname	nspan	np	nw	nn	dons	met	sd1	sd2	sd3	sd4	his	cn	r.m.s.	res	carbi
1dsz_A 1135	20	4	0	0	CCCC	Zn	3	14	3	−1	....	4	0.1	1.7	....
1dcq_A 264	23	4	0	0	CCCC	Zn	3	17	3	−1	....	4	0.1	2.1	....
1ee8_A 238	23	4	0	0	CCCC	Zn	3	17	3	−1	....	4	0.1	1.9	....
1a8h_ 127	20	4	0	0	CCCH	Zn	3	14	3	−1	...d	4	0.2	2.0	....
1vfy_A 176	27	4	0	0	CCCH	Zn	3	21	3	−1	...d	4	0.1	1.1	....
1ah7_ 55	67	4	1	0	DHHD	Zn	14	49	4	−1	.de.	5	0.2	1.5	....
1hxr_A 23	74	4	0	0	CCCC	Zn	3	68	3	−1	....	4	0.1	1.6	....
1psz_A 67	213	4	0	0	HHED	Zn	72	66	75	−1	ee..	4	0.3	2.0	..b.
1vhh_ 141	42	3	1	0	HDH	Zn	7	35	−1	−1	e.d	4	0.1	1.7	...
1lbu_ 154	43	3	1	0	HDH	Zn	7	36	−1	−1	e.d	4	0.2	1.8	...
1amp_ 117	139	3	1	0	DEH	Zn	35	104	−1	−1	..e	4	0.3	1.8	.b.
1cg2_A 141	244	3	1	0	DEH	Zn	35	209	−1	−1	..e	4	0.2	2.5	.b.
1hzy_A 201	29	2	2	1	HH	Zn	29	−1	−1	−1	de	5	0.2	1.3	

1kev for example, we find Zn CHD 22 91 at A353, but Zn CHED 22 1 90 at B353; the distance Zn—O of glutamate in the second is 2.43 Å, rather improbable for monodentate glutamate.)

3.3. Coordination group tables and comparisons of composition and conformation

The coordination-group tables, illustrated by a small selection of Zn coordination groups in Table 1 and given in full as supplementary Table 1D[1], were thus assembled. Furthermore, the program which generated the lists could also generate *MOLSCRIPT* input files, which allowed quick viewing of a coordination group (similar to the examples in Fig. 6).

Local programs were further developed (i) to select from the coordination-group lists particular sequences for comparison, for example all occurrences of a particular chelate loop, and (ii) to extract the requisite atomic positions from the PDB files, calculate and store the torsion angles φ, ψ, ω, χ_1, χ_2 and assign the φ, ψ angles to categories according to their positions in the Ramachandran plot (see §4.1 for categories used). In comparisons of chelate loops, additional output included conformation categories from relseq = −10 to a relseq of 10 beyond the end of the chelate loop, aligned amino-acid sequences over the same range (also extracted from the PDB) and protein names and resolution; this output was the basis for the files of Table 4W[1] (at http://tanna.bch.ed.ac.uk/arch/). Torsion angles could then be compared graphically or analytically, most conveniently by evaluating the r.m.s. difference between φ, ψ in all pairs of protein chains over any selected range in the (aligned) sequences; this allowed chelate loops with the same or similar conformations to be identified quickly. For a set of similar chelate loops, the mean and standard deviation of φ and ψ at each relseq position were then evaluated. Graphical superpositions of selected coordination groups and chelate loops were made with *INSIGHT*II, but since this is quite slow the preliminary analysis of torsion angles is essential. The versatile CSD program *VISTA* (Allen & Kennard, 1993*a*,*b*) was also used in some comparisons.

In the chelate-loop comparisons, fold families were obtained (manually) from SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/); in a few cases, the secondary-structure categories in chelate loops were examined [taken manually from PDBSUM (http://www.biochem.ucl.ac.uk/bsm/pdbsum), where they are established with the program *PROMOTIF*] and the immediate geometry around the Zn (bond angles, coordination shape, bond length from http://tanna.bch.ed.ac.uk). These details are in Table 4W (at http://tanna.bch.ed.ac.uk/arch/).

4. Results and discussion

Tables for all eight metals are deposited as supplementary Table 1D. Table 1 illustrates some of the data stored for a small selection of Zn coordination groups; not shown here but also stored in these files are (i) the increase in coordination number corresponding to an increase in coordination sphere radius of 0.2 Å, (ii) water molecules and other non-protein donors in the coordination group, (iii) the EC enzyme number when it is given in the PDB, (iv) part of the header name from the PDB file, (v) the names in the PDB file of the metal and the first donor atom and (vi) the sequence of residue conformations in each of the chelate loops (in full when the loops

[1] Supplementary data have been deposited in the IUCr electronic archive as a PDF file of Tables 1D, 2D, 4D and 5D, and Fig. 5D; and as a PDF file and zip archive of the website http://tanna.bch.ed.ac.uk/arch/ containing Tables 2W, 3W, 4W and 5W (Reference: AD0206). Details for accessing these data are given at the back of the journal.

Table 2
Constitution of metal coordination groups in the representative set of proteins.

(*a*) and (*b*) are for all eight metals; 372 structures (PDB codes) are included; duplicate chains within a structure are excluded. (*c*), (*d*) and (*e*) are for Zn and Ca only; for the other metals, this and further information can be found in supplementary Table 2D.

(*a*) Numbers of occurrences of different kinds of donor groups (from amino-acid side chains) in metal coordination groups with two or more protein donors. M.ch. O stands for main-chain carbonyl O atom is a donor.

	D, N	E, Q	S, T	H	C	M	K, R	Y	M.ch. O	All
Ca	339	127	34	3	—	—	—	1	309	813
Mg	88	42	38	3	—	—	1	2	54	228
Mn	51	30	3	22	1	—	—	—	6	113
Fe	12	30	—	60	18	3	—	5	7	135
Cu	2	3	3	77	26	10	—	1	4	126
Zn	63	50	1	179	206	1	3	—	10	517
Na	22	12	6	—	—	—	1	—	93	135
K	16	17	18	—	—	—	—	—	79	130

(*b*) Amino-acid types (%) for main-chain carbonyl oxygen donors, all metals combined, using categories based on those of Lesk (2001).

Glycine (G)	13
Other small amino-acids (A, S, T)	14
Medium and large hydrophobic amino acids	37
Acidic (D, E)	12
Basic (K, R)	14
Polar (N, Q, H)	11

(*c*) Coordination numbers of Ca and Zn in coordination groups with two or more protein donors.

Coordination No.	2	3	4	5	6	7	>8	All
No. Ca coordination groups	2	6	13	36	110	22	1	190
No. Zn coordination groups	7	19	89	31	3	—	—	149

(*d*) Numbers of protein donor groups interacting with Ca and Zn.

No. protein donor groups	1	2	3	4	5	6	7	All
No. Ca coordination groups	27	29	26	45	61	27	2	228
No. Zn coordination groups	33	21	51	76	—	—	—	184

(*e*) Distribution of chelate-loop sizes for Ca and Zn coordination groups.

seqdif	0	1	2	3	4	5	6–10	11–19	20–29	30–49	50–99	100–199	200–499	All
Ca	31	56	237	68	14	38	16	29	28	48	22	16	3	606
Zn	9	9	37	69	29	13	26	38	30	31	40	18	5	354

(*f*) Most commonly occurring chelate loops for Ca and Zn. (See supplementary Table 2D for other metals and Table 2W at http://tanna.bch.ed.ac.uk/arch/ for numbers of all chelate loops for each metal.) For commonly occurring Ca and Zn donor pairs, individual details are given in Table 4W (http://tanna.bch.ed.ac.uk/arch/), including amino-acid sequences through the chelate loop and before and after it, conformation described by Efimov type, name of protein from PDB header and resolution, together with a summary of the agreement found by analysis of the torsion angles.

Metal	Total No. coordination groups	No. chelate loops	Commonest donor pairs (number)				
Ca	190	606	DD 2 (35)	DO 1 (19)	OE 5 (27)	OD 0 (12)	
			DN 2 (16)	DO 2 (38)	OO 2 (38)	OD 2 (32)	[ON 2 (6)]
				NO 2 (15)	OO 3 (20)	OD 3 (12)	[ON 3 (6)]
Zn	149	354	HH 2 (11)	HH 4 (18)	CC 2 (9)	CC 3 (53)	CC 5 (9)

contain up to five residues; abbreviated for larger loops). The tables can be downloaded and searched for particular coordination groups or other features and sorted or otherwise manipulated in, for example, Microsoft *EXCEL*. For each coordination group the metal–donor atom distances and bond angles can be found at http://tanna.bch.ed.ac.uk (or at http://metallo.scripps.edu/).

There is much diversity in the coordination groups and different metals have very different characteristics. The preferences of different metals for different amino-acid donors are shown in Table 2(*a*) and 2(*b*). Oxygen donors (carboxylate, amide, water *etc.*) are almost never found in the same coordination group as cysteine, although either may occur alongside histidine. Tables 2(*c*), 2(*d*) and 2(*e*) summarize, for Ca and Zn coordination groups, metal coordination numbers and chelate-loop sizes and Table 2(*f*) lists the most commonly occurring chelate loops for each; fuller details are deposited for these and all the other metals (supplementary Table 2D).

In Ca proteins the EF-hand (see Pidcock & Moore, 2001; Nelson & Chazin, 1998; see also http://structbio.vanderbilt.edu/cabp_database/) is a very dominant structural motif, with 27 examples of the coordination group DDDOE 2225 or its close relatives in this set of representative proteins, and for Zn the pattern CCCC 3 *n* 3 with *n* = 10–20 is common in zinc fingers and related proteins. Apart from these, iden-

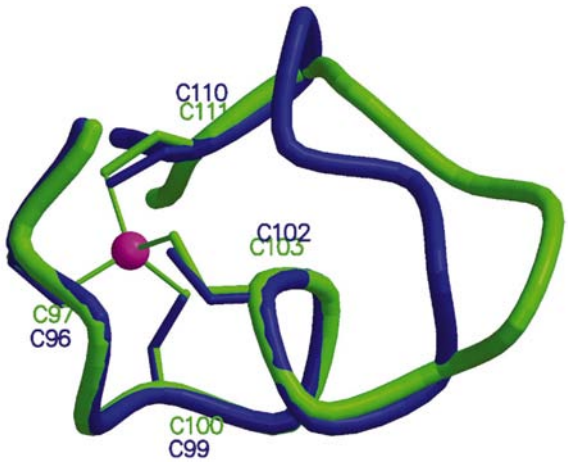

Figure 2
The coordination group Zn CCCC 3 3 8 showing its conformation in 1het_A 97 (green) and in 1e3j_A 96 (blue). The coordinating cysteines are labelled C96 in the blue chain, C97 in the green chain *etc*. 1het is an alcohol dehydrogenase; 1e3j is a ketose reductase. The backbone atoms of residues of the first two chelate loops, CCC 3 3, have been superposed using the program *LSQKAB* from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994); their conformations are the same, r.m.s. displacement 0.23 Å, whereas there are marked differences in the larger chelate loop, CC 8. This figure was prepared using *MOLSCRIPT* (Kraulis, 1991) and *RASTER3D* (Merritt & Murphy, 1994).

tical coordination groups (same donors, same residue separation) do not often recur in these tables; when they do the proteins usually have related folds and functions, but even then the conformations may differ, especially in the larger chelate loops. Fig. 2 shows an example. A detailed study was made of all the recurring Ca and Zn coordination groups, their conformations, amino-acid sequences *etc.* and these are available in Table 3W (at http://tanna.bch.ed.ac.uk/arch/).

While whole coordination groups are not often repeated here, except for the Ca EF-hands, some chelate loops that are their components occur frequently in different unrelated proteins, although other chelate loops are found only once or a few times. Small chelate loops, particularly seqdif = 2, are very common for calcium, whereas for zinc seqdif = 3 and larger loops are much more common. In coordination groups with only two protein donors these donors are rarely more than ten residues apart, which is understandable on simple stability grounds. When there are three or more protein donor groups it is common for there to be at least one large loop. Large chelate loops will usually serve the function of holding two parts of the polypeptide chain close to each other; this may be at the active site or simply to provide stability for the whole structure. It is common for long and short chelate loops to alternate in the protein-chain sequence and uncommon for a long loop to follow another long loop; a short loop following another short loop is uncommon in zinc coordination groups, but common in calcium groups.

4.1. Residue conformations and the significance of glycine

Within the chelate loops the nature of the amino acids which are not donors is very varied, even in small loops with the same conformation, but glycine plays an important part in many. The average glycine content over all proteins is 6.9% (evaluated using http://www.expasy.org/tools/pscale/A.A.SWISS-PROT.html for the whole SWISS-PROT database). For all the coordination groups studied there is a 10–15% probability that the amino acid following a donor, *i.e.* at relseq = +1, is glycine and there is a similar probability for the amino acid preceding a donor; in each position the probability is about twice that in a random sequence. In calcium coordination groups the probability is even higher than in complexes of other metals, rising to 18% in calcium coordination groups with small loops (seqdif = 1–3). High coordination numbers and/or small chelate loops lead to the greatest steric congestion; this should account for the higher frequency of glycine in positions adjacent to donors.

Residues containing donor atoms or adjacent to donor atoms have been examined to see whether any particular conformations are favoured in metal coordination. Conformations have been assigned to categories which are regions of a Ramachandran plot (*a*) following Hovmöller *et al.* (2002) and (*b*) in a way related to proposals of Efimov (1993), as shown in Fig. 3. The Efimov-type conformations are given in supplementary Table 1D for the residues in each coordination group. Their distributions are shown in Table 3.

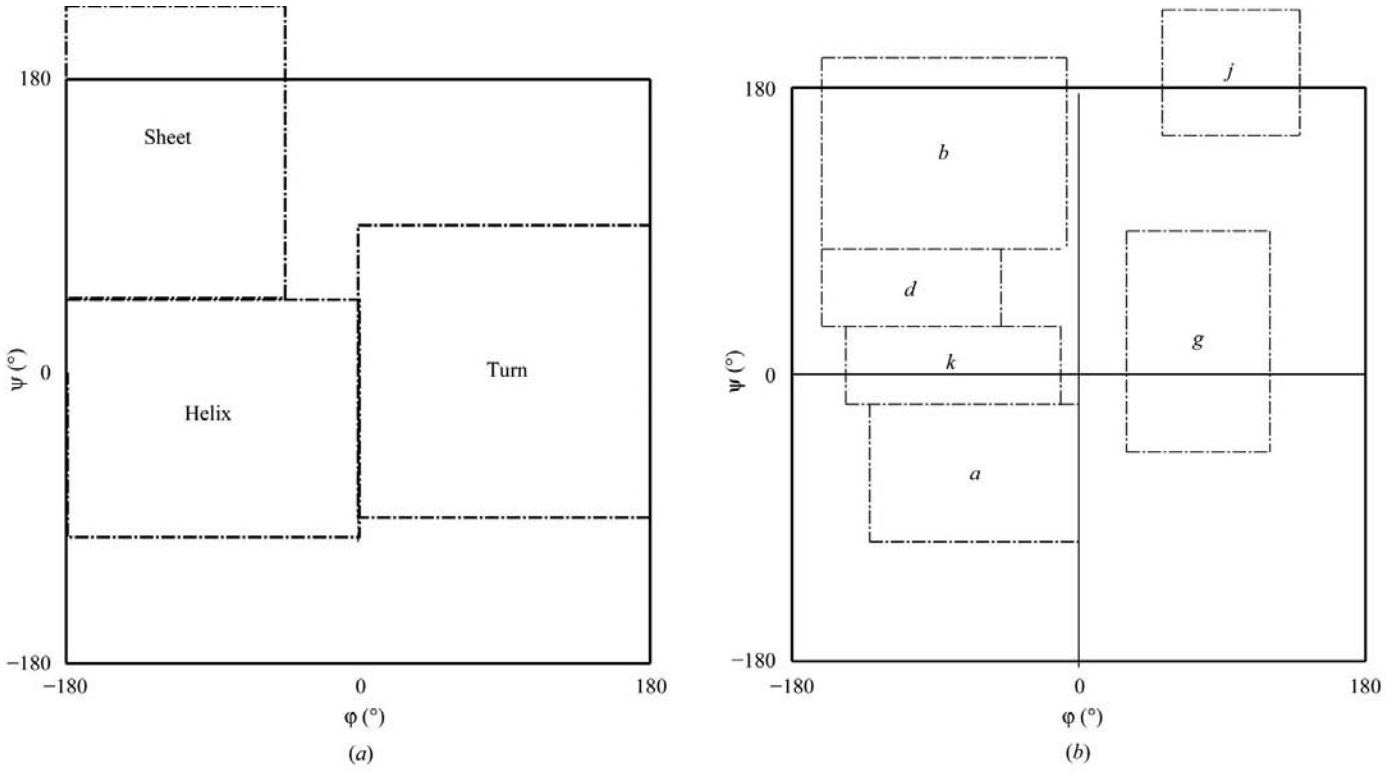

Figure 3
Definition of conformation categories (*a*) as used by Hovmöller *et al.* (2002); the areas are described as sheet, helix, turn and other and (*b*) based on those of Efimov (1993), but extended so that they are contiguous and cover nearly all the allowable conformation space.

Table 3
Distributions of conformations.

(*a*) Distribution of conformations of amino acids according to the categories helix, sheet, turn and 'other' defined by Hovmöller *et al.* (2002) in all the metal coordination groups treated here. The conformation definitions are shown in Fig. 3(*a*). For comparison, two distributions from Hovmöller *et al.* (2002) are given: the first is for all amino acids in their set of non-redundant and representative protein chains and the second for the subset of these which are classified (FAST) as random coil.

		Helix (%)	Sheet (%)	Turn (%)	Other (%)	No. observations
All metal coordination groups	Non-glycine donors	42	53	4	1	2042
	Non-glycine adjacent to donors	50	46	4	0	3417
All metal coordination groups	Glycine donors	14	22	19	44	77
	Glycine adjacent to donors	24	21	36	18	442
Compare whole PDB (Hovmöller *et al.*, 2002)	All	51	43	5	2	237384
	Classified as random coil	32	54	11	4	96442

(*b*) Distribution of conformations of donor amino acids and of amino acids adjacent to donors in all the metal coordination groups treated here. The categories are based on those of Efimov (1993) and are shown in Fig. 3(*b*). The comparison sample is for all amino acids in the structures of nine Ca-containing proteins, determined with resolution ≤1.4 Å.

	b (%)	*d* (%)	*k* (%)	*a* (%)	*g* (%)	*j* (%)	Other (%)	No. observations
Metal coordination groups								
Non-glycine donors	48	3	15	26	3	0	4	2042
Non-glycine adjacent to donors	43	2	16	32	3	0	3	3417
Glycine donors	16	0	5	8	18	30	23	77
Glycine adjacent to donors	17	1	7	17	35	11	12	442
Sample of all amino acids in nine Ca proteins	45	2	13	28	5	2	5	1846

(*c*) Distribution of conformations according to *PROCHECK* categories. The categories are 'core', 'allowed', 'generous' and 'not' (see text). The distributions are for all amino acids other than glycine and proline which provide donors in metal coordination groups or which are adjacent in the amino-acid sequence to one or more such donors; the recommendations are from the current *CCP*4 instructions for structures determined at a resolution of <2.0 Å (Collaborative Computational Projcet, Number 4, 1994).

		Core (%)	Allowed (%)	Generous (%)	Not (%)	Total No. amino acids
All metals here	Donor groups	82	17	0	0	2028
	Adjacent to donors	85	14	1	0	3207
Ca, Zn; resolution ≤ 1.8 Å	Donor groups	83	17	0	0	536
	Adjacent to donors	85	14	0	0	861
Recommended in *PROCHECK*		90	10	0	0	

4.1.1. Conformations in helix, sheet, turn and 'other' regions. The distributions (Table 3*a*) show that more than half the glycine residues have conformations in the turn or 'other' regions. Especially when chelate loops are small or coordination numbers are high, there must be bends in the protein chain at or near the residue coordinated to the metal; when present at one of these positions, glycine can obviously play a significant part in the bend. For 20% of donors in all calcium coordination groups the donor itself or one of the adjacent amino acids is glycine.

4.1.2. Conformations in categories based on those of Efimov. The distributions (Table 3*b*) are compared with the distribution for all the residues in a small sample of proteins. Metal coordination does not appear to affect the distribution of conformations. The tendency of glycine to have a conformation in the *g* or *j* region is evident, but the *g* region is just allowable for other amino acids, as also noted by Hovmöller *et al.* (2002).

4.2. Conformations in small chelate loops

The most commonly occurring chelate loops with Ca and Zn have been examined to see how closely the conformations are the same for all, or to what extent they may be affected by amino-acid sequence or be dictated by the overall protein fold

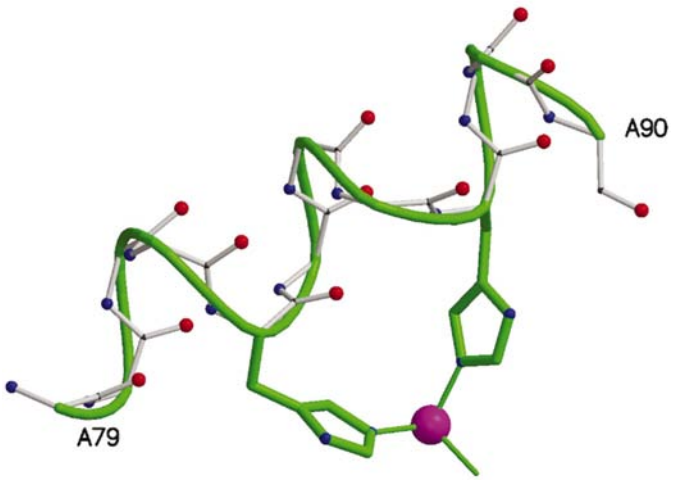

Figure 4
In the chelate loop Zn HH 4 both histidine residues belong to a helix and this helix usually extends in both directions. This example is in 1c7k_A 83. (Prepared in same way as Fig. 3.)

or possibly other factors. Can we predict that the conformation will be the same as that of another chelate loop with the same donors and residue separation? For each chelate loop all the conformations were found and all the amino-acid sequences from the first to the second donor residue and for ten residues before and after. Chelate loops with similar conformations were identified by comparison of the sequences of torsion angles (φ, ψ) within the chelate ring. The means were evaluated for each φ and ψ and their sample standard deviations, which give an indication of the spread of values.

The composition of a chelate loop (donors, residue separation) does not necessarily correspond to one conformation; often there are one or two strongly preferred and well defined conformations for the loop, together with one or a few outliers. Within and near the chelate loop amino-acid sequences can be very different, with no obvious simple relation to differences in conformation. Only two examples will be given here. In the chelate loop Zn HH 4 all 18 occurrences have the same conformation, with two histidine residues separated by one turn of α-helix, as illustrated in Fig. 4. The mean values of the φ, ψ angles within the loop have sample standard deviations between 5 and 14°, no more than would be expected from coordinate errors in the crystal structure determinations. Other chelate loops such as Zn DH 4, Zn ED 4 and Ca DD 4 have the same helix conformation, but some *M XX* 4 are quite different.

Of the 50 occurrences of the chelate loop Zn CC 3, all but three have conformations like those in Figs. 5(*a*) or (*b*). For the complete set of 47, the sample standard deviations of the mean φ, ψ angles in the loop are between 9 and 26°; thus, some of these angles differ by more than would be expected from coordinate errors. However, a subset of 14 are very close to Fig. 5(*a*) (sample s.d. = 4–11°, r.m.s. deviation of backbone atoms $\simeq$ 0.2 Å) and another 11 are similarly close to Fig. 5(*b*). The backbones of Figs. 5(*a*) and 5(*b*) are superposed in Fig. 5(*c*); the r.m.s. displacement between the backbone atoms is 0.55 Å. In all of these a bend in the protein chain is stabilized by the bonding of the two cysteines to Zn; the residue conformations (in Efimov categories) are *baaa* or *baak*; no conformations in the *g* or 'turn' region are involved. The proteins belong to many different fold types and the small differences within the chelate loops are associated with differences in backbone conformations outside the loops. The remaining three CC 3 loops all have the same quite different conformation, illustrated in Fig. 5(*d*), *agab* in Efimov categories. They are unlike all the other CC 3 loops in Zn coordination groups in that they each precede another small chelate loop.

Supplementary Table 4D summarizes the conformations found for all the common small chelate loops with Ca and Zn, giving sample standard deviations within sets of similar conformation and examples and comments on their relation to fold families and local protein-chain conformation.

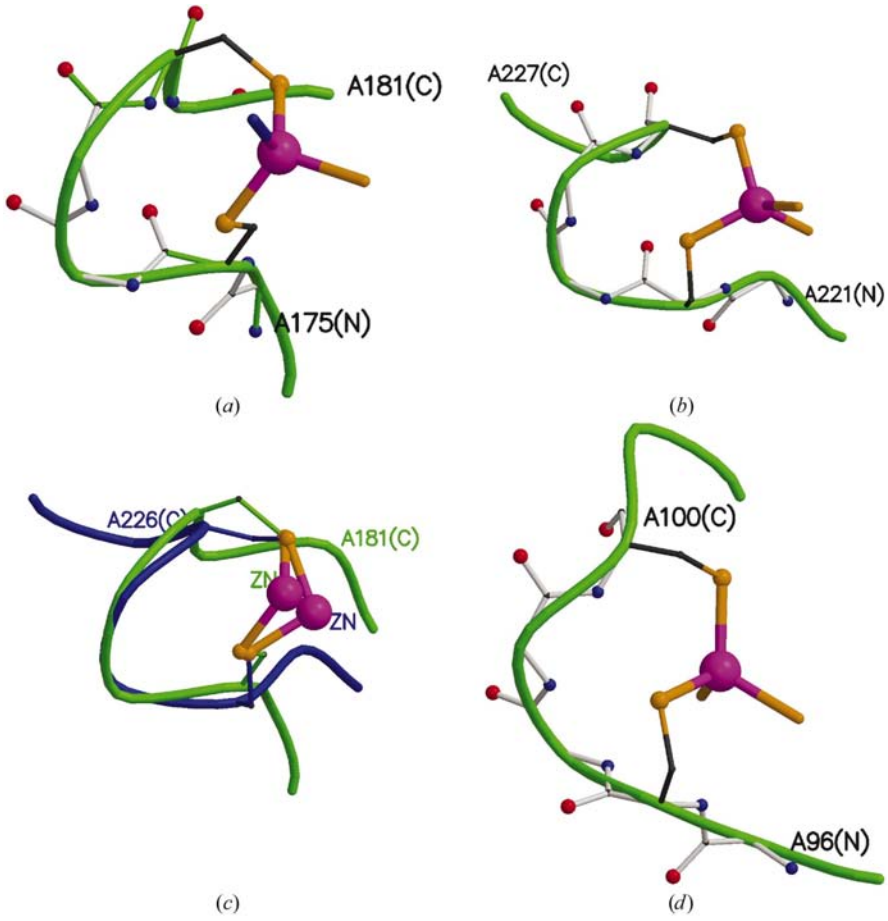

Figure 5
Three examples of the chelate loop Zn CC 3. The examples (*a*) 1vfy_A 176 and (*b*) 1vfy_A 222 are very similar, but not identical; their differences are shown in (*c*), where 1vfy_A 222 (blue) is superposed on 1vfy_A 176 (green) and the r.m.s. displacement is 0.55 Å. The example in (*d*) 1het_A 97 is quite different. (All figures prepared in the same way as Fig. 3.)

4.3. Conformations in whole coordination groups

There is a very wide range of composition and stereochemistry which must await further comparison apart from a few brief comments here. On the basis of composition three main patterns can be seen in Zn coordination groups, one for coordination groups with two or three protein donors and two for coordination groups with four protein donors. In the first pattern, the donors are predominantly histidine, aspartate and glutamate (cysteine is found in only nine out of 56) and the proteins are predominantly enzymes, mostly hydrolytic; additional water molecules or non-protein small molecules may be present. Many of those with three protein donors and nspan < 30 have one or two helices with pyramidal Zn exposed on one face, *e.g.* Fig.

6(*b*). Zinc coordination groups with four protein donors fall into two patterns: the first is CCCC, or with one or two of these C residues replaced by H; all of these have an overall span less than 75 and have one or two short chelate loops with seqdif less than 5 and a longer loop, usually the middle one. Many of these are zinc fingers or DNA-related proteins; in Fig. 6(*c*) there are three examples of this type of coordination group, all in the structure 1rmd; they overlap each other. In the second pattern there are no cysteine donors and a greater overall span (all but two have nspan > 67) and all the groups have one or two long chelate loops (seqdif up to 200) as well as very short loops (seqdif < 4). Among Ca coordination groups, short overall spans of fewer than 20 residues, *e.g.* Fig. 6(*a*), are much commoner than in Zn coordination groups, even when there are more donors within the groups; they are made up of a series of short chelate loops, mostly with seqdif of 2 or 3, but 0 and 1 are also quite common. There are also many Ca coordination groups with longer spans (>40) and in these short and long loops often alternate. In their discussion of structural characteristics of Ca-containing proteins, Pidcock & Moore (2001) divide Ca sites into three general types: in the first all the ligands belong to a continuous short sequence of amino acids, in the second one the ligand is supplied by a part of the amino-acid sequence far removed from the main binding sequence and in the third the binding amino acids are remote from one another in the sequence. The first type corresponds well to the coordination groups with small values of nspan, the most obvious examples being the EF-hand type, and there are a good number which fit the second pattern, *i.e.* one longer chelate loop (say, >20 residues) preceded or followed by one or more short ones. However, inspection of supplementary Table 1D shows that it is very rare for all the chelate loops to be long ones as in the third type of Pidcock & Moore (2001); there are almost always one or more short chelate loops adjacent to a long one.

Some brief speculation on the reasons for these architectural patterns is possible. Where the function of a Zn coordination group is the maintenance of tertiary structure (rather than as an enzyme active site) an α-helix can be tied in place by a single coordinate link to Zn, but to hold its orientation firmly two points of attachment are essential; coordination to Zn fulfils this role in Zn HH 4, Zn H*X* 4 or Zn *X*H 4, where *X* is H, D or E. To hold two non-helical sections of protein chain together, including some constraint of their relative orientations, two Zn CC 3 groups are good, resulting in the coordination group Zn CCCC 3 *n* 3 (observed with $n > 14$). There are some variations: the replacement of C by H or seqdif = 2 or 4 in the short chelate loops. Ca complexes are much more labile than Zn complexes and are probably too labile to provide much stabilization of tertiary structure. In a Ca complex, for any stability at all several donor groups close to each other in the chain sequence are desirable or essential. For Ca transport or signalling, precise control of the lability is required and the EF-hand configuration may allow 'fine tuning' by the interchange of D, E, S and T in Ca DDDOE 2 2 2 5.

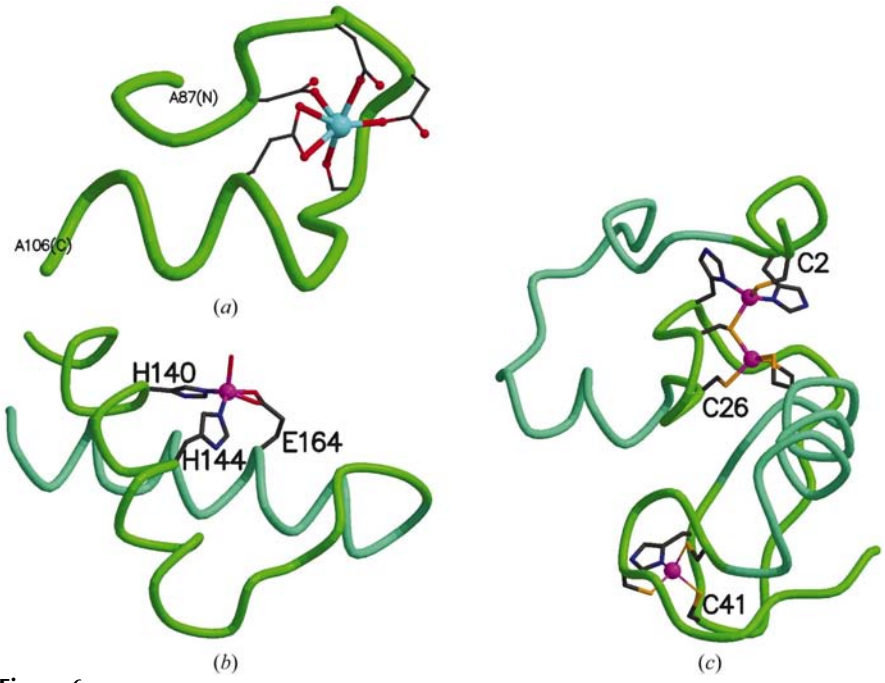

Figure 6
(*a*) The coordination group Ca DDDOE 2 2 2 5 in 2pvb_A 90. This is typical of Ca coordination groups in EF-hand proteins. (*b*) The coordination group HHE 4 20 in 1ezm_ 140. Zn is coordinated by three donor groups from the protein and by a water molecule. The donor amino acids belong to two helices. The protein is a hydrolase. (*c*) The protein chain of 1rmd of residues 1–70 and the three zinc coordination groups associated with it. The first group, CHCH, coordinates to Zn through residues 2, 6, 29 and 31, the second group, CCCC, through residues 26, 29, 46 and 49, and the third, CHCC, through 41, 43, 61 and 64. Note that these groups overlap. The sharing of a cysteine S between the first and second groups brings these two Zn atoms to 3.9 Å. The long chelate loops of the first and third coordination groups are shown in a slightly different shade of green. A fourth Zn is coordinated through residues 91, 96, 108 and 112, but is remote from any of the first three and is not shown here. (All figures prepared in the same way as Fig. 3.)

4.4. Is protein conformation distorted by metal coordination?

Conformation angles have been examined to see whether binding to the metal of several residues in the protein chain induces any distortions from normal geometry. The donor bond from an N, O or S atom to metal has an energy much greater than that of a hydrogen bond, although not quite as great as a simple C—C bond. Formation of such a bond could justify distortion of the protein geometry to allow movement of the donor atom to an optimum position in relation to the metal. The most easily distortable parts of the protein geometry for this purpose are the torsion angles around single bonds. Torsion angles around peptide bonds are less readily distortable, followed by bond angles such as C—C—C and then by the covalent-bond distances (the bond distances to metal atoms are

inflexible, but the angles between them are fairly flexible).

The program *PROCHECK* is widely used for validating protein structures (Collaborative Computational Project, Number 4, 1994; Morris *et al.*, 1992) and defines the areas 'core', 'allowed', 'generous' and 'not'. For residues other than glycine and proline, 90% of the torsion angles (φ, ψ) in a protein are normally found within the core area if the structure has been well refined with high-resolution data (*e.g.* 1–1.5 Å); the remaining 10% are found within the allowed area. Some torsion angles are likely to be found in the other two categories, 'generous' and 'not', when structures have been incompletely refined or where the resolution is poor; they would also be found here if there were significant distortions from the normal range. The results in Table 3(*c*) show that there is no evidence for a higher than normal proportion of conformations in the 'generous' and 'not' regions; there is just possibly a slightly higher proportion in the allowed region at the expense of core, representing small but allowable distortions of conformation from the optimum in the absence of metal.

4.5. Number of metal atoms per protein chain

In about half the structures examined the stoichiometry is simple, with one metal atom coordinated by donor groups from one protein chain. In a small proportion of structures (<15%) the metal coordination group includes donor groups from more than one protein chain within the crystal asymmetric unit (listed in Table 5W at http://tanna.bch.ed.ac.uk/arch/). (In a very small number of cases the coordination group may include donor atoms which are not in the asymmetric unit listed in the PDB file, but are related to it by crystal symmetry; such links have not been taken into account in any of the descriptions of coordination groups here, although supplementary Table 1D does include a marker when the metal atom may lie on a crystallographic two-, three- or fourfold rotation axis.)

In many structures a single protein chain provides the donors for two or three metal coordination groups, occasionally for several more, and not necessarily all involving the same metal. In about one third of the metalloproteins here, one protein chain provides donor groups for two or three metal atoms and in about 15% for four or more metal atoms; the maximum found so far is eight Ca and one Zn in 1kap. Table 5W (http://tanna.bch.ed.ac.uk/arch/) provides a list of these proteins and coordination groups. Those with Ca and Zn have been examined a little further. In half of them the metal coordination groups are well separated in space and in the amino-acid sequence and can reasonably be regarded as independent in geometry, but in some they are close. Details of the type of interaction are given in supplementary Fig. 5D for Zn$\cdots$Zn approaches between 3.0 and 6.0 Å and for Ca$\cdots$Ca approaches between 3.6 and 7.5 Å. Overlap of coordination groups or close approach of the metal atoms does not appear to substantially affect the conformations of these small chelate loops, although there may well be small distortions.

5. Concluding remarks

This survey has shown the diversity of architecture in metal coordination groups. Bond lengths from metal to donor atoms are very predictable and are in line with those in simple molecules known to coordination chemists, as are the coordination numbers and angles at the metal atom. As the listings of coordination groups show, there is a very wide variety of composition and geometry in the chelate loops which make up the coordination group. The composition (nature of amino-acid residues and their separation in the sequence) is not sufficient to predict the conformation either for a whole coordination group or for its constituent chelate loops, although for each such loop there will be one or two likely conformations. Glycine is found adjacent to donor residues more frequently than random statistics would predict (but in no more than 20% of these positions); sometimes it provides a 'turn' in the protein-chain direction, but elsewhere its small size may be helpful in allowing the protein chain to make several coordinate links to a metal atom. Coordination to a metal ion from several positions in the protein chain does not appear to require distortion of the conformation angles φ, ψ, ω from their normally allowed range of values. With the exception of Ca coordination groups in EF-hand proteins, very few whole coordination groups occur more than once in this 'representative' set (30% cull) of proteins and those that do are usually related in overall fold or function; even so, they are not necessarily identical in conformation, although the smaller chelate loops (say <5 residues) usually are. Quite frequently one protein chain provides the donors for two or more metal coordination groups and sometimes these are quite close to each other.

These are some of the observations which emerge from this attempt to look at the architecture of metal coordination groups and look for patterns of behaviour which might help in the understanding of biological function or the prediction of structure from sequence, as well as in the interpretation of electron-density maps. Much remains to be done, including looking at coordination groups in more recently determined structures and in proteins which have appreciable similarity to those in the present selection; the latter include some quite different metal coordination groups as well as those already recognized here.

I am very grateful to Professor Malcolm Walkinshaw and to the University of Edinburgh for computing facilities, to Dr Paul Taylor for computational support, and to them and Drs Dmitriy Alexeev and Dietlind Gerloff for advice and helpful discussions. I am also grateful to the referees for helpful suggestions about presentation and to Tom Ellison who performed a preliminary exploration of some parts of this topic.

References

Allen, F. H. & Kennard, O (1993*a*). *Chem. Des. Autom. News*, **8**, 1.
Allen, F. H. & Kennard, O (1993*b*). *Chem. Des. Autom. News*, **8**, 31–37.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A. & Pique, M. E. (2002). *Nucleic Acids Res.* **30**, 379–382.
Chong, K. T., Miyazaki, G., Morimoto, H., Oda, Y. & Park, S.-Y. (1999). *Acta Cryst.* D**55**, 1291–1300.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Duda, D., Govindasamy, L., Agbandje-McKenna, M., Tu, C., Silverman, D. N. & McKenna, R. (2003). *Acta Cryst.* D**59**, 93–104.
Dunbrack, R. (2001). *Culling the PDB by Resolution and Sequence Identity*, http://www.fccc.edu/research/labs/dunbrack/culledpdb.html.
Efimov, A. V. (1993). *Prog. Biophys. Mol. Biol.* **60**, 201–239.
Frausto da Silva, J. J. R. & Williams, R. J. P. (1991). *The Biological Chemistry of the Elements.* Oxford: Clarendon Press.
Harding, M. M. (1999). *Acta Cryst.* D**55**, 1432–1443.
Harding, M. M. (2000). *Acta Cryst.* D**56**, 857–867.
Harding, M. M. (2001). *Acta Cryst.* D**57**, 401–411.
Harding, M. M. (2002). *Acta Cryst.* D**58**, 872–874.
Hovmöller, S., Zhou, T. & Ohlson, T. (2002). *Acta Cryst.* D**58**, 768–776.
Huber, R., Wieghardt, K., Ponlos, T. & Messerschmidt, A. (2001). Editors. *Handbook of Metalloproteins.* Chichester: Wiley.
Kraulis, P. (1991). *J. Appl. Cryst.* **24**, 946–950.
Lesk, A. M. (2001). *Introduction to Protein Architecture*, p. 21. Oxford University Press.
MacArthur, M. W. & Thornton, J. M. (2002). Private communication.
Maher, M. J., Xiao, Z., Wilce, M. C. J., Guss, J. M. & Wedd, A. G. (1999). *Acta Cryst.* D**55**, 962–968.
Merritt, E. A. & Murphy, M. E. P. (1994). *Acta Cryst.* D**50**, 869–873.
Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins Struct. Funct. Genet.* **12**, 345–364.
Nelson, M. R. & Chazin, W. J. (1998). *Biometals*, **11**, 297–318.
Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.
Pidcock, E. & Moore, G. R. (2001). *J. Biol. Inorg. Chem.* **6**, 479–489.
Vallee, B. L. & Auld, D. S. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 220–224.