Inorganic Chemistry

Article

pubs.acs.org/IC

Zinc Coordination Spheres in Protein Structures

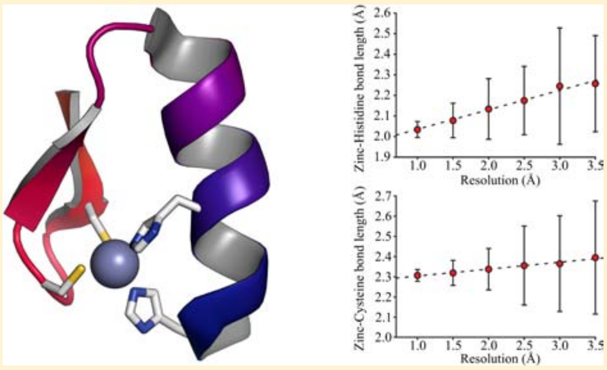
Mikko Laitaoja,[†] Jarkko Valjakka,[‡] and Janne Jänis*,[†]

[†]University of Eastern Finland, Department of Chemistry, P.O. Box 111, FI-80101 Joensuu, Finland

[‡]University of Tampere, Institute of Biomedical Technology, FI-33014, Tampere, Finland

Supporting Information

ABSTRACT: Zinc metalloproteins are one of the most abundant and structurally diverse proteins in nature. In these proteins, the Zn(II) ion possesses a multifunctional role as it stabilizes the fold of small zinc fingers, catalyzes essential reactions in enzymes of all six classes, or assists in the formation of biological oligomers. Previously, a number of database surveys have been conducted on zinc proteins to gain broader insights into their rich coordination chemistry. However, many of these surveys suffer from severe flaws and misinterpretations or are otherwise limited. To provide a more comprehensive, up-to-date picture on zinc coordination environments in proteins, zinc containing protein structures deposited in the Protein Data Bank (PDB) were analyzed in detail. A statistical analysis in terms of zinc coordinating amino acids, metal-to-ligand bond lengths, coordination number, and structural classification was performed, revealing coordination spheres from classical tetrahedral cysteine/histidine binding sites to more complex binuclear sites with carboxylated lysine residues. According to the results, coordination spheres of hundreds of crystal structures in the PDB could be misinterpreted due to symmetry-related molecules or missing electron densities for ligands. The analysis also revealed increasing average metal-to-ligand bond length as a function of crystallographic resolution, which should be taken into account when interrogating metal ion binding sites. Moreover, one-third of the zinc ions present in crystal structures are artifacts, merely aiding crystal formation and packing with no biological significance. Our analysis provides solid evidence that a minimal stable zinc coordination sphere is made up by four ligands and adopts a tetrahedral coordination geometry.



■ INTRODUCTION

Zinc is one of the most abundant metals is biology, and it is estimated that about one-tenth of proteins may contain a zinc ion as a cofactor.[1,2] The chemical properties of zinc render it distinct from other transition metals, such as copper and iron. Unlike copper and iron, which display several different oxidation states in biological systems, zinc exists as a redox-inert Zn(II) cation with an electron configuration of [Ar]$3d^{10}$. The completely filled d-orbital renders it diamagnetic and thus invisible in EPR spectroscopy. Out of its three stable isotopes, ^{67}Zn is NMR active, but due to its low natural abundance and its low receptivity, only solid-state, low temperature NMR studies of small zinc compounds are practically feasible. In addition, zinc complexes have no absorbance in the UV−vis and microwave spectral regions, therefore greatly limiting available analytical methods for their analysis. Zinc is considered as a borderline metal, being coordinated by both the sulfur atom of cysteine and nitrogen atom of histidine (soft base ligands) or by carboxylate anions of aspartate and glutamate (hard base ligands).[3] These properties with the lack of ligand field effects make zinc an excellent metal for different coordination numbers and binding geometries in different biological systems.

Zinc coordination environments in proteins have been defined into four main categories: (1) catalytic, (2) cocatalytic, (3) structural, and (4) interface.[4] In many proteins, zinc ions are also required for correct folding of the polypeptide chain, such as in zinc finger proteins. Zinc can be found as an active site metal (cofactor) in all six IUBMB enzyme classes. In enzymes, there is almost invariably one coordination site occupied by a water molecule, which can easily be displaced to create a catalytically active species for an incoming substrate/inhibitor molecule. For example, in carbonic anhydrase, the binding of a water molecule to the positively charged zinc center reduces the pK_a of water from 15.7 to ∼7, generating a hydroxide ion that attacks carbon dioxide and further converts it into bicarbonate. In some enzymes, like D-hydantoinase (dihydropyrimidinase), also binuclear, cocatalytic sites exist where two metal ions act in concert to catalyze the reaction.[5] Zinc ions at protein interfaces can also affect the formation of a stable quaternary structure; the best example is hexameric insulin, which is assembled from three insulin dimers and two zinc ions. Zinc coordination is diverse in proteins, although the binding is generally occurring *via* side chains of histidine,

Received: April 30, 2013
Published: September 23, 2013

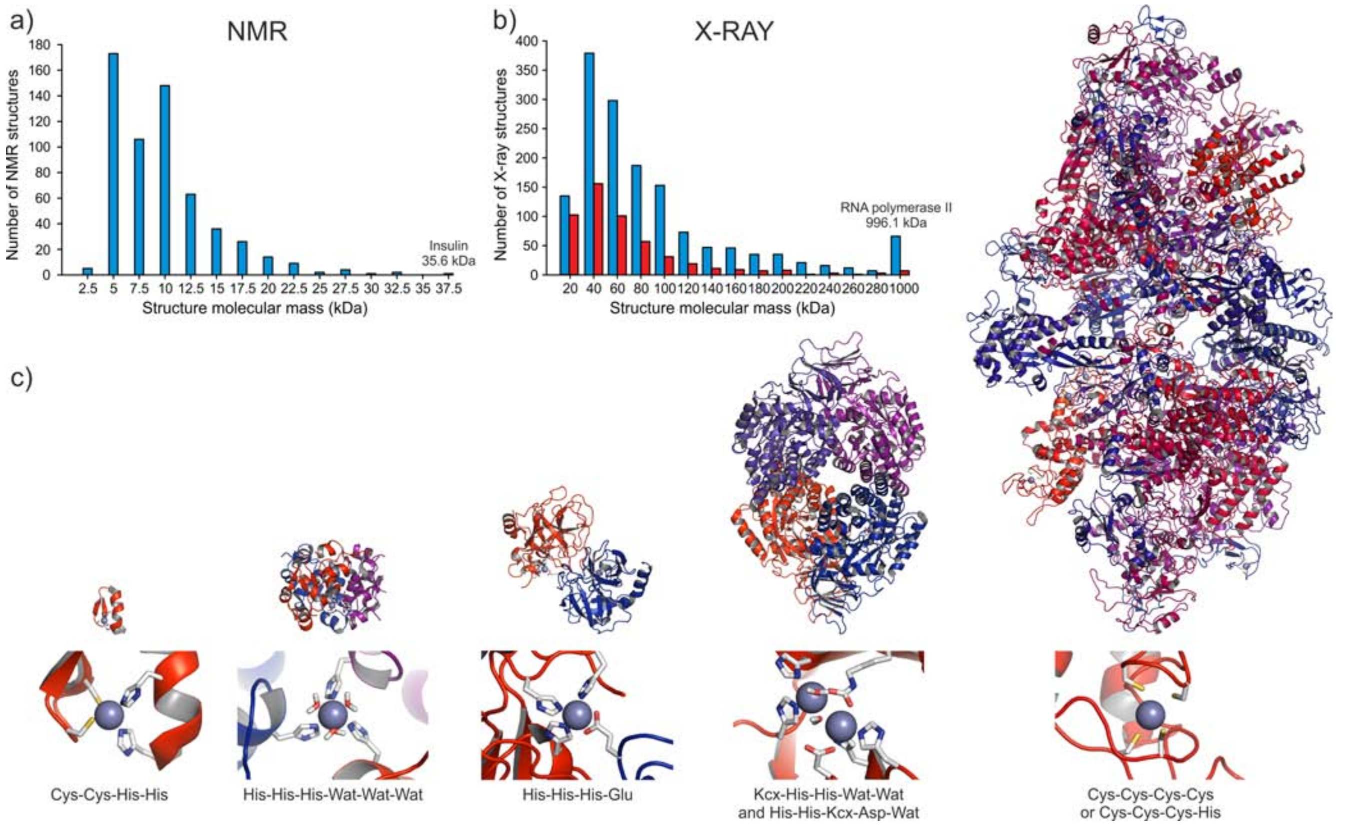


Figure 1. Molecular mass distributions of zinc proteins determined by (a) NMR and (b) X-ray crystallography. Red bars in the X-ray correspond to crystallization artifacts. (c) Representative zinc protein 3D structures. The insets show the actual coordination spheres. From the left to right: NMR structure of 31st zinc finger from *Xenopus laevis* zinc finger protein Xfin (PDB entry 1ZNF); crystal structure of hexameric human insulin in T6 state (PDB entry 1MSO); crystal structure of tonin, a serine protease from the rat (PDB entry 1TON); D-hydantoinase from *Thermus sp.*, a binuclear enzyme with carboxylated lysine residue (Kcx; PDB entry 1GKP); and RNA polymerase II from *Schizosaccharomyces pombe*, the largest zinc containing X-ray structure with a molecular mass of 996.1 kDa (PDB entry 3H0G). Wat = coordinated water molecule. See text for details.

cysteine, aspartate, and/or glutamate residues. Despite the variability in the coordination environments, the affinity toward zinc ions is usually very high (K_d values in the μM to pM range).[6]

A number of database surveys have been previously conducted on zinc coordination in proteins. However, these surveys have dealt with a limited number of (high-resolution) crystal structures only.[7−15] In addition, some of these studies greatly suffer from severe flaws and misinterpretations, which are discussed later in detail. In the present study, zinc coordination spheres in protein structures deposited in the Protein Data Bank (PDB at http://www.pdb.org) were analyzed in great detail.[16] More specifically, over 2600 structures with ~7800 individual zinc ions and ~31 000 coordination bonds were *manually* analyzed without the use of any computer algorithms. This analysis also took into account low-resolution X-ray structures, lacking electron densities for ligands, as well as the symmetry-related molecules in crystals. Thus, we present here an unbiased and comprehensive analysis of all unique zinc-containing proteins present in PDB. Some of these structures might contain non-native metal ions, whether by error or intentionally, for instance, probing enzyme active sites in nonactive metals.[17,18] Both X-ray and NMR structures were separately analyzed in this survey, with the X-ray structures being clearly dominant (77% out of all structures). In NMR structures, zinc position is defined by using experimental atom angle and distance constraints as well as geometry calculations.[19,20] Thus, zinc ions are "invisible" in NMR, and exogenous zinc ions (other than catalytic or structural zinc ions) are hard, if not impossible, to detect. The amount of zinc containing proteins in the PDB has doubled in the past three years, clearly showing the growing interest in the "zinc proteome" in structural biology studies.

EXPERIMENTAL PROCEDURES

The Protein Data Bank (http://www.pdb.org), as of January 18, 2012, was queried for structures containing zinc ions. [Currently, the database contains 8474 structures with zinc ions, i.e., a 19% increase in the number structures as compared to the time of the database query. After removal of highly similar structures, a total of 2866 (637 NMR and 2229 X-ray) structures remained, i.e., only a 9% increase as compared to the used data set (2616 structures), supporting the relevancy of the data.] The initial number of structures was reduced to contain only the structures without nucleic acid ligands. A search was conducted separately for NMR and X-ray structures. A given protein structure may be represented by several PDB entries of different amino acid mutants or substrate/inhibitor complexes, and their incorporation would cause considerable bias in the statistical analysis. A choice to remove the structures with 95% sequence identity was accomplished using the BLASTClust program. No other restraints, such as atom specific distance cut-offs or factor restraints, were used in collecting the working data set. Since most of the zinc-containing proteins are enzymes, a separate search was performed for each International Union of Biochemistry and Molecular Biology (IUBMB) enzyme class.

The structures were initially analyzed individually by using Ligand Explorer 3.9 available on the PDB Web site. All structures were *manually* inspected to assess zinc coordinating ligands, metal-to-ligand bond lengths, and coordination geometries. In cases where coordination showed largely distorted geometry or clearly missing ligand(s), resulting in unrealistic bond lengths and/or an electronically incomplete coordination sphere, original publications (when available) were inspected for clarification. For the analysis of crystallographic symmetry-related molecules, the structures were further inspected by using the PyMOL 1.3 software.[21]

Histidine can coordinate *via* either of its nitrogen atoms, but since the functional and structural significance of these two binding modes remains unknown, these modes were not differentiated.[6] The binding of carboxylate (aspartate and glutamate) residues and similar ligands to zinc ions is also problematic due to multiple binding modes.[22] The energy difference between monodentate and bidentate coordination is minor, however.[23] In this study, aspartate and glutamate residues were considered as monodentate, two-electron donors to a single zinc ion; hence only a shorter bond was used in the analysis. Upon bridging two separate metal ions, i.e. a binuclear site, a carboxylate group can also form coordination bonds with both of its oxygen atoms. A similar bridging mode can also exist with a sulfur atom of cysteine, as can be seen, for example, in the metallothionein protein. The zinc ions in these structures were analyzed separately. The coordination geometry of each zinc binding site was analyzed and broadly categorized by coordination number and bond angles into tetrahedral, trigonal bipyramidal/square pyramidal, octahedral, and incomplete geometries. Protein structures usually have some distortions, and the geometry lies somewhere in between these ideal geometries.[1,9] The coordinating amino acid residues are represented by their three-letter codes, e.g., Cys for cysteine and Kcx for carboxylated lysine. The other abbreviations used are as follows. Water is marked as Wat, and exogenous ligands, inhibitors, and solvent molecules are marked by their coordinating atom such as oxygen (O), nitrogen (N), sulfur (S), and chlorine/chloride (Cl).

■ RESULTS AND DISCUSSION

At the time of the database search, the PDB contained around 76 000 protein structures. Zinc was present in about 7100 of them, somewhat in accordance with a number of zinc proteins estimated from genomic studies, although these values might be biased toward already known proteins with predicted homology and function as most PDB structures are determined from soluble human, *E. coli*, yeast and mouse proteins.[6,24] It must be noted that for creating a completely unbiased data set, a detailed analysis of protein sequence, classification, resolution, and function would be needed. Highly similar structures were removed from the initial data set as explained in the Experimental Procedures. Thus, the working data set contained about 35% of all deposited zinc protein structures (i.e., 2616 individual structures). As the search was conducted separately for NMR and X-ray structures (590 and 2026 structures, respectively), the same protein may be represented in both subdata sets if determined by both methods, e.g., histone lysine demethylase JARID1A-PHD finger (PDB entries 2KGI and 3GL6).[19]

NMR Structures. Unique NMR structures analyzed counted for around 85% of all zinc-containing NMR structures in the PDB (590 in total). It is noteworthy that 60% of the analyzed NMR structures are unpublished. This is due to the fact that a large number of these structures are from structural genomics/proteomics initiatives, determined e.g. at RIKEN.[25,26] Most of the NMR structures contain one or two zinc ions giving an indication of the two different classes of zinc fingers deposited in the PDB. The classical zinc finger contains only one zinc ion coordinated by two cysteine and two histidine residues, whereas PHD, LIM, and RING fingers contain two zinc coordination spheres.[27] This can clearly be seen from the molecular mass distribution (Figure 1a), which is bimodal, peaking around the molecular masses of 5 and 10 kDa. The higher molecular mass structures are metallothioneins and repeats of single zinc finger motifs. A total of 922 zinc ions were found in different NMR structures.

X-Ray Structures. Zinc containing X-ray structures deposited in the PDB count for over 7000 entries. A marked difference compared to the NMR structures is that only around 2000 structures remain following the BLASTClust identity removal (see Experimental Procedures for details), and the remaining unique structures represent then about 30% of all zinc containing X-ray structures (1945 structures in total). This indicates that many proteins had been determined more than once in different space groups, as different complex structures with substrates and/or inhibitors or as different mutant structures. For example, the database query for insulin results in about 190 different structures. If the structures are limited to enzyme classes before the identity removal, some additional structures are found compared to the previous search. This indicates that there are some higher resolution structures without enzyme classification number or highly similar structures without enzymatic function, which are excluded in the search and from the enzyme list. The addition of these enzyme structures increases the total number of structures to 2026. In addition, some structures have not been classified as enzymes even though the publications clearly indicate this. These structures were added to their corresponding enzyme classes if similarity with the other enzymes was found or if classified as "enzyme" in case the reaction catalyzed is uncertain. The resolution range for zinc proteins spanned from 0.79 Å to 4.30 Å with a weighted average of 2.06 Å. About 96% of the crystal structures deposited in the PDB have a resolution better than 3.00 Å. The rest of the structures are mostly from large protein complexes (~245 kDa on average) where the size sets an obvious limit to the achievable resolution (Figure S1 in the SI). The X-ray structures contained a total of 6950 individual zinc ions.

Structure Molecular Mass. Figure 1 shows the distribution of structure molecular mass of the analyzed zinc proteins with some representative structures shown.[28−32] The NMR structures have an average molecular mass of 8.6 kDa, and a majority of structures are small zinc fingers.[33] The apparent size limit of the NMR analysis can clearly be seen as the largest zinc-containing NMR structure was from hexameric insulin (PDB entry 1AI0) with a molecular mass of 35.6 kDa. This is in contrast to the largest X-ray structure of RNA polymerase II (PDB entry 3H0G), having a structure molecular mass of 996.1 kDa with 16 zinc ions present in the structure. In X-ray structures, the molecular mass distribution has an average around 80.6 kDa, although this represents the molecular mass of the asymmetric unit, and not the biological assembly or the functional unit.[34] Enzymes have an average molecular mass of 92.6 kDa, clearly beyond the range of NMR, further emphasizing the larger number of X-ray structures in the database.

Classification. The analyzed zinc ions were categorized according to the protein structure classification. Enzymes may contain catalytic as well as structural sites, where the zinc is not required for the enzymatic reaction, but for stability and correct folding of the protein. Summarized in Figure 2 is the distribution of zinc ions in different types of proteins. In the case of enzymes, the two numbers report the number of zinc ions in active or structural sites. In oxidoreductases, hydrolases, lyases, and isomerases, the majority of zinc ions have a catalytic role. In contrast, transferases and ligases mostly contain

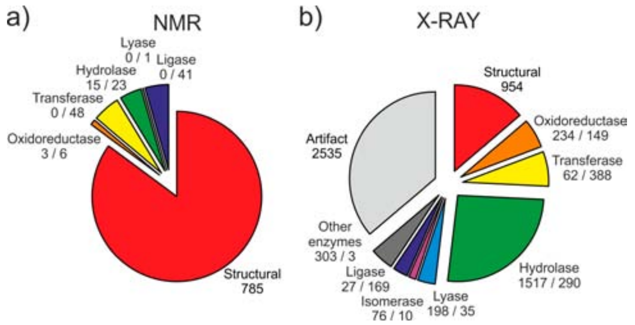


Figure 2. Distribution of zinc ions in structure classes determined by (a) NMR and (b) X-ray. In different enzyme classes, the two numbers represent catalytic/structural zinc ions. The class "other enzymes" represents those enzyme structures which could not be classified. "Artifact" in X-ray means crystallization artifacts with no catalytic/structural role for the bound zinc ions (see text for details).

structural zinc ions. About 300 enzyme structures could not be classified. Since most NMR structures are from different zinc fingers, zinc ions have a purely structural role. The total number of zinc metalloenzymes in NMR structures was only 84 (out of 590 structures). The results indicated also that many enzymes contain more than one zinc ion in their structure; in particular, most ligases contain two zinc ions. By analyzing the structure molecular mass of enzymes in NMR structures, it was noted that the majority of enzymes have a molecular mass less than 22 kDa. A few tricoordinated patterns were found in these structures, where the bond angles clearly indicated a tetrahedral coordination sphere. Inspection of the original publications (whenever available) indicated the plausibility of coordinated water molecules, increasing the actual coordination number to four. Noticeable was the low amount of coordinated water molecules in NMR structures. Some structures contained a coordinated inhibitor molecule, where zinc may be considered as the active site metal. These results suggest that in most NMR structures zinc is not the active site metal in these enzymes and plays only a structural role, stabilizing other parts of the structure. NMR structures of the database are highly focused on small structural metal domains, due to the apparent size limit and inability to determine water positions.

In contrast, X-ray structures had a larger share of different enzymes. On average, each enzyme contained three zinc ions. Apart from the binuclear sites, this value indicates that enzymes have a tendency to form higher oligomers as their functional unit, such as tetrameric hydantoinase[31] (Figure 1) or trimeric γ-carbonic anhydrase.[35] Out of all 2026 structures, only 197 structures contained a multinuclear site. These multinuclear sites contained 981 zinc ions (from the total of 6950 ions) and were mostly present in hydrolases. The amount of enzymes determined by X-ray along with their higher molecular mass explains the higher occurrence of histidine coordination compared to the NMR structures; coordination spheres in enzymes more frequently have histidine and acidic glutamate or aspartate residues and a lower amount of cysteine residues. These differences are more a result of the limitations of the methods than the actual distribution of zinc ions in different proteins. A large number of structures contained zinc ions on the surface of proteins with a random amount of ligands attached. Upon inspection of these structures and corresponding publications, it became obvious that these zinc ions are merely crystallization artifacts (hereafter referred to as the "artifacts"), resulting from zinc-containing buffer/precipitant solutions used in crystallization experiments.[36] Various publications indicate that diffraction quality crystals could only be obtained in the presence of zinc buffers, zinc ions probably aiding the crystal packing.

The coordination of zinc was also observed to reduce side chain movement and conformational space.[7,12] Usually, coordination was accompanied by more than one water molecule. Most of the artifact ions show incomplete spheres resulting from a nearby solvent channel. A high concentration of free zinc can cause conformational changes or even induce protein oligomerization, which is reflected by the high amount of artifactual zinc coordination.[30,37] These nonspecific interactions could explain why free zinc concentration is tightly controlled in cellular environments; the total concentration of zinc in cells is about 200 μM, but the concentration of free zinc is only picomolar.[6,38] These artifacts were not included in the subsequent analysis. However, they have important roles in aiding crystal formation by stabilizing intermolecular crystal contacts, and their binding sites can be determined using anomalous signals.[20,39]

Coordinating Ligands, Geometry, and Coordination Number. In NMR structures, cysteine and histidine residues, representing for over 97% of all coordinating ligands, dominate coordination spheres. Other coordinating ligands are mainly aspartate and glutamate residues, both with a 1% share. Although zinc is "invisible" in NMR, it is generally determined using distance restraints, where the bond length is limited to a certain value and the models are calculated based on these values. The amounts of coordinating ligands in NMR and their average bond lengths are summarized in Table 1. The relative amounts of coordinating ligands are similar to previous surveys.[9,40] In NMR structures, the coordination sphere is essentially tetrahedral (98.0%), which is a reflection of a high number of zinc fingers with only a small number of enzyme structures deposited (Figure 3). Some trigonal bipyramidal geometries (1.0%) were found in the enzyme−inhibitor complexes.

Table 1. Coordinating Ligands in Zinc Proteins Determined by NMR

coordinating ligand	total	bond length (Å)[a]
cysteine (Cys)	2659	2.32 ± 0.16
histidine (His)	936	2.09 ± 0.14
glutamate (Glu)	36	1.83 ± 0.16
aspartate (Asp)	33	2.10 ± 0.24
other oxygen (O)	17	2.20 ± 0.13
water (Wat)	3	2.18 ± 0.03
other sulfur (S)	1	2.63
total	3685	

[a] Values reported as average ± 1 standard deviation.

In X-ray structures, cysteine and histidine residues are the most frequent coordinating residues followed by acidic side chains of aspartate and glutamate residues and water molecules (Table 2; Table S1 and Figure S2 in the SI). A bidentate mode of binding observed for carboxylates would form a four-membered chelate ring with two highly distorted orbitals (and geometry) and is represented by resonance structures of single coordination bonds.[7] In structures where the bidentate mode is observed, the bond angles of other ligands are still close to an ideal tetrahedron, thus supporting our analysis. In structural sites, the coordination is mostly by cysteine and histidine with

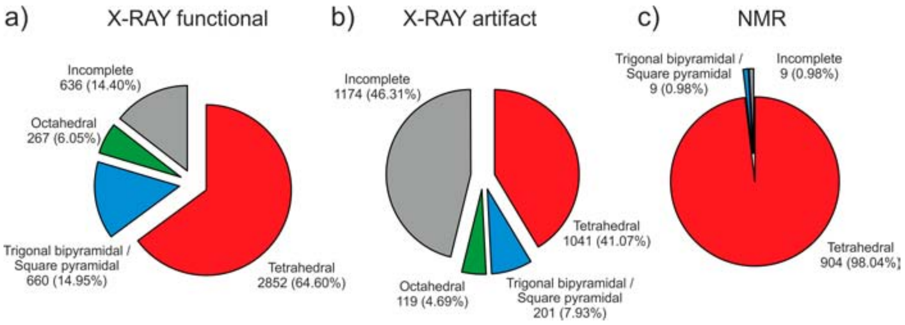


Figure 3. Geometries of zinc coordination spheres in proteins: (a) X-ray functional, (b) X-ray artifact, and (c) NMR structures.

Table 2. Coordinating Ligands of Zinc Proteins Determined by X-Ray Crystallography

coordinating ligand	functional	artifact	total
cysteine (Cys)	6102	122	6224
histidine (His)	5716	1810	7526
aspartate (Asp)	2026	1415	3441
water (Wat)	1753	2586	4339
glutamate (Glu)	1293	1952	3245
other oxygen (O)	974	546	1520
carboxylated lysine (Kcx)	161		161
other nitrogen (N)	137	193	330
chlorine (Cl)	65	115	180
lysine (Lys)	54	46	100
asparagine (Asn)	51	50	101
other sulfur (S)	49	1	50
serine (Ser)	24	42	66
threonine (Thr)	24	36	60
tyrosine (Tyr)	22	14	36
glutamine (Gln)	20	37	57
phosphoserine (Sep)	7	2	9
selenomethionine (Mse)	5		5
methionine (Met)	4	7	11
formylglycine (Fgl)	4		4
arginine (Arg)		6	6
bromine (Br)		2	2
tryptophan (Trp)		2	2
total	18491	8984	27475

some contribution of acidic residues. In enzymes, usually three side chains form a framework to which the substrate, water (hydroxide), or inhibitor molecule binds. This is reflected by a high number of water molecules and other nitrogen and oxygen ligands. A notable coordinating ligand is carboxylated lysine (Kcx), which has not been recognized previously, an important ligand in several binuclear zinc enzymes.[31] This modified amino acid residue can also be found in β-lactamases (e.g., PDB entry 1M6K) and rubisco (1RCO).[41,42] The other ligands, such as lysine, have an approximately 2% share in total. In X-ray structures, the coordination sphere is more diverse, but still mostly tetrahedral geometries are found (56.0%). Most structural proteins along with mononuclear enzymes are tetrahedral, and binuclear enzymes are trigonal bipyramidal in their native state. Enzyme−inhibitor complexes and solvent-bound artifacts increase the relative shares of trigonal bipyramidal and octahedral coordination geometries to 12.4% and 5.6% shares, respectively (Figure 3). A large part of inhibitors are based on sulfonamide, hydroxamic acid, or phosphonate functional groups. Interestingly, a very low amount of exogenous sulfur ligands were found, although cysteine is the most common coordinating ligand.

For zinc binding proteins, coordination numbers (CN) ranging from two to eight have been reported in the literature.[11,14] CNs of two or three are rare occasions in metal complexes. In the case of zinc, di- or tricoordinate zinc ions mainly exist in organozinc compounds, having covalent zinc−carbon bonds, but these complexes are electron-deficient and highly reactive under ambient conditions. Our analysis provides solid evidence that a minimal stable zinc coordination sphere requires the presence of four coordinating ligands in protein structures (satisfies an 18-electron rule, inherent for most stable transition metal complexes). On the other hand, our data do not show any coordination spheres with more than six ligands. Therefore, possible coordination numbers for zinc proteins are CN = 4 (tetrahedral), CN = 5 (trigonal bipyramidal/square pyramidal), and CN = 6 (octahedral). The small size of the Zn(II) cation (~74 pm for four-coordinate and ~88 pm for six-coordinate ion) prevents higher coordination numbers due to molecular repulsion and higher energy orbitals.[43] The general approach for determination of coordinating amino acid residues in zinc proteins is by mutation of each potential coordinating residue one at a time. Then, the mutation of each residue is monitored against the loss of a function and/or a correct fold or degradation of a protein.[44,45] A majority of transition metals are capable of forming compounds with less than 18 electrons, due to their unfilled d orbitals, but trigonal planar geometry is rare and usually requires especially bulky ligands such as triphenylphosphine. As the structures were manually checked for coordinating ligands without any computer algorithms, we found out that one-fourth of zinc containing structures displayed vacant coordination sites or otherwise incomplete spheres.[46,47] Most of these "empty" coordination sites are due to low crystallographic resolution where the refinement of ligands cannot be done with high accuracy. In these circumstances, CN might be wrongly assigned. Incomplete spheres found in NMR structures (1.0%) are due to missing ligands or unresolved binding conformations of amino acid residues. Thus, in NMR an incomplete coordination sphere was a rare occasion and was clearly more frequent in X-ray structures. The reasons leading to incomplete coordination spheres are summarized in Table 3. A complete list of these structures is given in Table S3. Example structures covering these factors are shown in Figure S3.

About 46.3% of the artifact zinc ions and about 14.4% of the catalytic/structural zinc ions have an incomplete coordination sphere. The main reasons for incomplete spheres are symmetry-related molecules in crystals or missing water/

Table 3. Reasons for Incomplete Sphere of Zinc Ions in Protein Structures

reason	functional	artifact	fraction of functional sites	fraction of artifact sites
symmetry-related molecules	90	760	10.7%	39.2%
missing solvent molecules	21	642	2.5%	33.1%
missing water from active site	467	132	55.5%	6.8%
symmetry with missing solvent		288		14.9%
missing side chain or ligand	116	8	13.8%	0.4%
metal placed to fit electron density		70		3.6%
side chain conformation	60	10	7.1%	0.5%
metal or ligand occupancy	58	11	6.9%	0.6%
unknown or missing metal	12	17	1.4%	0.9%
side chain flip (His/Asn/Gln)	17	2	2.0%	0.1%
total	841	1940	19.1%	76.5%

solvent or other ligand molecules. The asymmetric unit for a given crystal structure may contain only a single molecule, and excluding the symmetry-related molecules may result in an incomplete sphere (usually CN = 2 or 3).[15] However, in these cases, a detailed inspection of the binding geometry usually reveals the actual geometry and CN, e.g., clearly tetrahedral geometry with a single vacant coordination site (for example, see PDB entry 1OHT). The same is true in the case of missing electron densities for water or other nonprotein ligands. The large data set allows comparison between similar proteins, where higher resolution structures show additional ligands (usually water or other small molecules) compared to lower resolution structures. Table 3 shows that one-fifth of the functional sites and three-quarters of the artifact sites cannot be directly analyzed from the original coordinate file. These factors have not been considered in previous surveys and have led to many erroneous interpretations. Insulin provides an excellent example of this phenomenon. In many insulin crystal structures, the asymmetric unit contains a dimer (e.g., see PDB entry 1MSO for human insulin in the T6 state; zinc binding site shown in Figure S3) with zinc ions coordinated to a single histidine (HisB10) residue and an external water ligand.[29] This clearly results in an incomplete coordination sphere (CN = 2), if only the asymmetric unit is considered. However, the biological assembly of insulin is a hexamer when stored in pancreatic β-cells. In these structures, insulin is indeed hexameric with the coordination sphere of zinc being fulfilled by two more HisB10 residues from the other two dimers, residing along the 3-fold symmetry axis. This forms a biological assembly formed by six insulin monomers (three dimers) and two zinc ions. The zinc coordination sphere is further fulfilled by one (e.g., chloride) or three external (water) ligands, forming either tetrahedral or octahedral coordination depending on the state of the protein (R- or T-state, respectively).

The artifacts are more affected by symmetry-related molecules since coordinating ligands are on the surface of the protein. Missing electron densities, alternative conformations, or highly mobile areas often results in distorted binding geometries of incomplete spheres. This comes up largely in the case of enzymes which should have a catalytic water molecule bound to the zinc ion, which counts for half of the incomplete spheres in functional sites. Non-native metals are listed as artifacts. When metal identity is not known in advance and electron density corresponding to a heavy atom is found in the structure, metal can principally be identified based on the coordination number and geometry and confirmed using anomalous scattering experiments. Iron, copper, cobalt, and nickel have very similar properties and coordination environments in proteins.[12] Some structures show His, Asn, or Gln coordination erroneously with flipped side chains.[1] Although these are unintentional errors, it demonstrates the need for validation of the deposited structures in the database.[26,48]

Coordination Spheres. In NMR structures, the most common coordination sphere in tetrahedral zinc sites is Cys_2-Cys/His-Cys/His (with positional variations), representing over 92% of all structures. In X-ray structures, the same sphere is by far the most common one. The coordination is very diverse as over 500 different spheres were found, half of them occurring only once or twice (Table 4 and Table S2 in the SI).

Table 4. Common Zinc Coordination Spheres in Different Classes of Proteins

function	common coordination spheres (share-%)[a]
structural	Cys_4 (31.4%), Cys_2-His-Cys (10.9%), Cys_2-His_2 (4.7%), His-Cys_3 (4.5%), His_3-Asp (4.0%), Glu_2-His-Glu (2.0%)
oxidoreductase	Cys_4 (27.7%), His_3-Asp (13.3%), Cys-His-Cys (12.5%), Cys-His-Asp (6.3%), Cys-His-Glu (3.9%), Asp-His-His (3.7%)
transferase	Cys_4 (45.8%), Cys_3-His (6.9%), His-Cys_3 (4.9%), Cys_2-His-Cys (3.8%), Cys_3 (3.6%), Cys-His-Cys_2 (3.3%)
hydrolase	Cys_4 (7.1%), His_3 (5.8%), His_2-Glu (5.7%), His_2-Kcx-Asp (4.3%), Kcx-His_2 (4.3%), His_3-Asp (4.2%), His_2-Glu (4.0%), Asp-His_2 (3.3%), His-Glu-His (3.1%), Asp-Glu-His (3.0%), His_2-Asp_2 (2.9%)
lyase	His_3 (28.6%), Cys-His-Cys (12.0%), Asp-His_2 (8.1%), Cys-Asp-His-Cys (7.3%), Glu_2-His_2 (5.1%), Cys_3 (3.8%)
isomerase	His-Asp-His-Asp (17.0%), Cys_4 (9.6%), Glu-Asp-His-Asp (8.5%), His_3 (6.4%), His_2-Glu-His (5.3%)
ligase	Cys_4 (42.9%), Cys-His-Cys_2 (13.8%), Cys_2-His-Cys (11.7%), Cys_3-His (6.1%), Cys-His_2 (3.6%), Cys_2-His_2 (3.1%)
unclassified enzyme	His_2-Glu (13.7%), His_3 (9.2%), His_2-Glu-Asp (5.9%), His_2-Kcx-Asp (5.2%), Kcx-His_2 (5.2%), Glu-Asp-His-Glu (5.2%), His_3-Asp (4.9%), His-Glu-His (4.2%), Asp-His_2 (4.2%)
artifact	His (8.3%), His-Glu (5.7%), Glu_2 (5.6%), Asp (4.4%), Glu (4.1%), His-Asp (4.1%), Asp_2 (4.0%), Glu-His (3.6%), His_2 (3.6%), Glu_3 (3.2%), Asp-His (3.1%)

[a] Note that these contain only the incorporated protein ligands and the actual structures also contain various external ligands, especially at low coordination numbers.

The coordination spheres of different enzyme classes were separately analyzed, which could help in determining enzymatic reactions catalyzed by putative enzymes, if similar coordination spheres were found. The most common spheres in oxidoreductases (EC 1.x.x.x) are Cys_4, followed by His_3−Asp, which is a result of a high number of Cu/Zn superoxide dismutases (structural sites), and Cys-His-Cys-O (active site) from alcohol dehydrogenases, where the fourth coordination site is completed by various external oxygen ligands. Transferases (EC 2.x.x.x) and ligases (EC 6.x.x.x) are dominated by classical zinc finger coordination spheres. In large RNA polymerases, the low crystallographic resolution prevents exact determinations, even though the sequence similarity between different RNA polymerases is high and the coordinating amino acid residues are conserved. Hydrolases (EC 3.x.x.x) have the most diverse set of coordinating ligands, likely due to the higher amount of structures than other enzyme classes. Most of the carboxylated lysines can be found in hydrolases as coordinating ligands. A coordination sphere formed of three protein ligands and various external ligands is the major one, and similar trend can be seen in lyases (EC 4.x.x.x) and isomerases (EC 5.x.x.x). In these enzymes, zinc is a part of the active site, which is also supported by the markedly lower cysteine content than in transferases or ligases.[14] This suggests that most ligases do not use zinc for enzymatic reactions but for stabilizing the protein structure. Other unclassified enzymes have very similar spheres as found in hydrolases. The artifacts have very diverse coordination spheres as they are generally found on protein surfaces where coordination is completed by water or other solvent molecules. A noticeable observation is also a low number of cysteines and high water content in the coordination spheres of artifacts.

Atom Specific Bond Lengths: NMR. In recent database surveys, atom based cutoff values have been used for determination of the coordinating residues.[11,40] Our analysis shows that the use of such (arbitrary of statistically derived) cutoff values results in false determination of coordination spheres in some cases. Most of the metal-to-ligand distances fall close to the average values, but some coordination spheres show noticeably longer bond lengths. In the case of crystal structures, this usually results from lower resolution or higher B-factors for zinc and the ligands.[1] The shortest zinc-to-ligand distance found in the NMR structures was only 1.23 Å for the Zn−Cys bond (PDB entry 2FUU). This is clearly unrealistic, as it results in the van der Waals radii of the metal and the ligand atoms clearly overlapping. In contrast, the longest distance found was 4.12 Å for the Zn−His bond (PDB entry 2JMI). In this NMR ensemble, the calculated models clearly show unrealistic coordination distances and geometries. Figure 4 shows the zinc-to-ligand bond lengths for the NMR structures. Cys and His are statistically more significant than the other ligands due to their larger amount. Interestingly, in the distribution for Zn−His bond lengths, two peaks at around 2.05 Å and 2.35 Å are observed; this could be due to distance restraints used in the structure determination, where coordinating residues were not treated separately. A bimodal distribution was also found for glutamate, and surprisingly the bond lengths are shorter for glutamate than for aspartate, although the coordination is identical. Peaks for Zn−Glu were found at around 1.85 Å and 2.05 Å, compared to the average Zn−Asp distance of 2.10 Å. A few structures (PDB entries 1U7J, 1U7M, 2KIK, and 2LFD) contain most of these shorter Zn−Glu bonds and cause this difference. The acidic residues in X-ray structures have almost equal distributions.

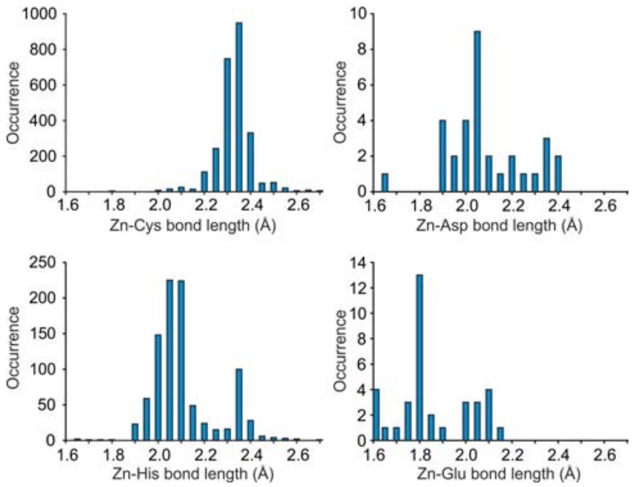


Figure 4. Zinc−ligand bond lengths in NMR structures.

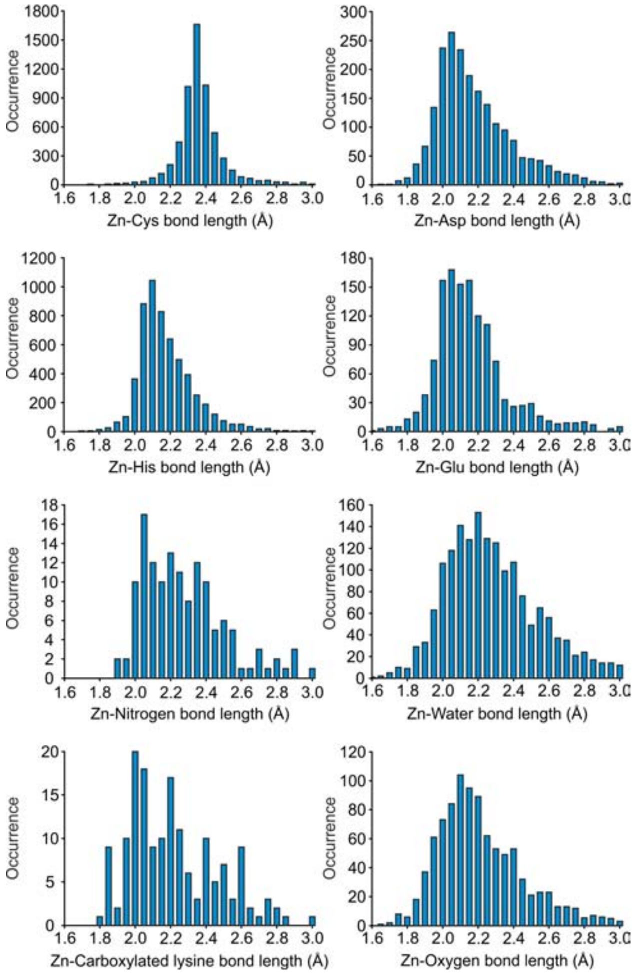


Figure 5. Zinc−ligand bond lengths in X-ray structures.

Atom Specific Bond Lengths: X-Ray. Similar bond lengths were seen in X-ray structures, as shown in Figure 5 for the eight most common ligands. Cysteine has a quite narrow and symmetrical distribution, but histidine and the others display distributions that are wider and broadened toward the higher bond lengths. The detailed analysis revealed a clear systematic increase in the average metal−ligand bond lengths as a function of crystallographic resolution (Figure S2 and Table S1 in the SI). This is a rather surprising observation. The average Zn−Cys bond length increases only by 0.1 Å on going from atomic to low resolution structures. However, this effect is more dramatic for the other ligands. The Zn−His bond length at <1.0 Å resolution is 2.03 Å on average. At 3.0 Å resolution, it is about 2.25 Å, and at resolution >4.0 Å it is already 2.86 Å. The average bond length corresponds to a resolution of ~2.2 Å. Also, the deviation increases noticeably after the resolution exceeds 2.5 Å. This dependence is also clearly seen with Glu and Asp ligands, and it is even more pronounced with water ligands. The high resolution values approach the bond lengths observed in small molecule complexes (Zn−His bond length is 2.00 Å on average for small molecule structures in Cambridge Structural Database).[1] The reason for increasing bond lengths with decreasing crystallographic resolution is not fully understood, but a similar trend has also been noted with other metals.[1,12,40] According to our analysis, the bond length is only dependent on the crystallographic resolution and is not affected by the date of data acquisition, structure molecular mass, or the refinement method applied. Interestingly, for zinc proteins the average bond lengths and the average resolution have remained roughly the same over the past 15 years. However, the low number of observations prevents statistical comparison with minor ligands.[1,47] Dependence of the bond lengths on the crystallographic resolution might pose a problem for computer-based coordination sphere analysis, if bond specific cutoff values are being applied. For example, the Zn−HisB180 bond length is 7.37 Å (PDB entry 1IRX, resolution 2.60 Å), clearly beyond normal cutoff values used, yet the publication clearly states HisB180 as the coordinating residue.[49] In contrast, the shortest bond found was to be only 0.27 Å for an artifactual zinc ion (PDB entry 1XOC, resolution 1.55 Å) and 0.72 Å for the structural zinc ion (PDB entry 2Y0S, resolution 3.80 Å). These results indicate some problems in atom placement and structure refinement in X-ray crystallography, which clearly result in arbitrarily long or short bond lengths. In these cases, however, *manual* inspection of the coordinate files along with the original publications results in unambiguous assignment of the coordination spheres present in these structures.

CONCLUSIONS

This survey represents the most comprehensive, up-to-date analysis on zinc protein X-ray and NMR structures present in the PDB. The NMR structures are mostly from small zinc fingers with tetrahedral coordination spheres, and infrequent occurrence of zinc enzymes explains higher cysteine content as compared to the X-ray structures. This is due to the fact that the average molecular mass of zinc enzymes is clearly beyond the mass range of NMR. Hence, most of the zinc protein structures have been determined by X-ray crystallography. A detailed analysis of protein structures deposited in the PDB has shown that examination of the zinc coordination sphere requires a deeper understanding of crystal structures, going beyond the asymmetric unit that is usually viewed by the molecular visualization programs. One should be cautious when interrogating metal atom coordination in proteins using automated algorithms. In X-ray structures, symmetry-related molecules and missing solvent or ligand molecules (resulting from poor crystallographic resolution) along with the actual function of zinc ion(s) should be taken into account. A high number of zinc ions were found to bind to the residues on the protein surface. These zinc ions are not required for folding or catalytic activity of the protein, but merely aid crystal packing. Thus, the amount of “real” zinc proteins is somewhat exaggerated in the database. Zinc has a marked preference for tetrahedral coordination geometry, dictated by the 18-electron rule. Five- and six-coordinate zinc ions were mostly found in enzymes with bound inhibitors or solvent molecules. A large number of structures showed electronically incomplete or geometrically distorted coordination. Poor or missing electron density or low crystallographic resolution in general results in missing water molecules or ligand atoms from the structure. Some of the structures have erroneously reported zinc coordination as the ligands can have multiple occupancies or flipped side chains or even a wrong metal ion, as determined by the authors.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Number of structures per resolution and molecular mass per resolution for zinc proteins determined by crystallography (Figure S1). Average zinc-to-ligand bond lengths by resolution (Table S1). Common enzyme coordination spheres (Table S2). Bond lengths and average bond lengths per resolution (Figure S2). Structures containing incomplete coordination spheres (Table S3). Example structures of incomplete coordination spheres (Figure S3). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +358-29-4453422. E-mail: janne.janis@uef.fi.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by funding from the Finnish National Doctoral Program in Organic Chemistry and Chemical Biology to M.L.

REFERENCES

(1) Harding, M. M.; Nowicki, M. W.; Walkinshaw, M. D. *Crystallogr. Rev.* **2010**, *16*, 247−302.

(2) Coleman, J. E. *Annu. Rev. Biochem.* **1992**, *61*, 897−946.

(3) Vahrenkamp, H. *Dalton Trans.* **2007**, 4751−4759.

(4) Auld, D. S. *BioMetals* **2009**, *22*, 141−148.

(5) Maret, W. *J. Inorg. Biochem.* **2012**, *111*, 110−116.

(6) Maret, W.; Li, Y. *Chem. Rev.* **2009**, *109*, 4682−4707.

(7) Alberts, I. L.; Nadassy, K.; Wodak, S. J. *Protein Sci.* **1998**, *7*, 1700−1716.

(8) Auld, D. S. *BioMetals* **2001**, *14*, 271−313.

(9) Patel, K.; Kumar, A.; Durani, S. *Biochim. Biophys. Acta* **2007**, *1774*, 1247−1253.

(10) Dokmanić, I.; Šikić, M.; Tomić, S. *Acta Crystallogr.* **2008**, *D64*, 257−263.

(11) Sousa, S. F.; Lopes, A. B.; Fernandes, P. A.; Ramos, M. J. *Dalton Trans.* **2009**, 7946−7956.

(12) Zheng, H.; Chruszcz, M.; Lasota, P.; Lebioda, L.; Minor, W. *J. Inorg. Biochem.* **2008**, *102*, 1765−1776.

(13) Lee, Y.; Lim, C. *J. Mol. Biol.* **2008**, *379*, 545−553.

(14) Andreini, C.; Bertini, I.; Cavallaro, G. *PLoS ONE* **2011**, *6*, e26325.

(15) Hsin, K.; Sheng, Y.; Harding, M. M.; Taylor, P.; Walkinshaw, M. D. *J. Appl. Crystallogr.* **2008**, *41*, 963−968.

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

(17) Maret, W. *Metallomics* **2010**, *2*, 117.

(18) Sommerhalter, M.; Lieberman, R. L.; Rosenzweig, A. C. *Inorg. Chem.* **2005**, *44*, 770−778.

(19) Wang, G. G.; Song, J.; Wang, Z.; Dormann, H. L.; Casadio, F.; Li, H.; Luo, J.-L.; Patel, D. J.; Allis, C. D. *Nature* **2009**, *459*, 847−851.

(20) Shi, W.; Zhan, C.; Ignatov, A.; Manjasetty, B. A.; Marinkovic, N.; Sullivan, M.; Huang, R.; Chance, M. R. *Structure* **2005**, *13*, 1473−1486.

(21) *The PyMOL Molecular Graphics System*, version 1.3r1; Schrödinger, LLC: Cambridge, MA, 2010.

(22) Harding, M. M. *Acta Crystallogr.* **2001**, *D57*, 401−411.

(23) Ryde, U. *Biophys. J.* **1999**, *77*, 2777−2787.

(24) Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. *J. Proteome Res.* **2006**, *5*, 196−201.

(25) Yee, A.; Gutmanas, A.; Arrowsmith, C. H. *Curr. Opin. Struct. Biol.* **2006**, *16*, 611−617.

(26) Chruszcz, M.; Domagalski, M.; Osinski, T.; Wlodawer, A.; Minor, W. *Curr. Opin. Struct. Biol.* **2010**, *20*, 587−597.

(27) Laity, J. H.; Lee, B. M.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2001**, *11*, 39−46.

(28) Lee, M. S.; Gippert, G. P.; Soman, K. V.; Case, D. A.; Wright, P. E. *Science* **1989**, *245*, 635−637.

(29) Smith, G. D.; Pangborn, W. A.; Blessing, R. H. *Acta Crystallogr.* **2003**, *D59*, 474−482.

(30) Fujinaga, M.; James, M. N. G. *J. Mol. Biol.* **1987**, *195*, 373−396.

(31) Abendroth, J.; Niefind, K.; Schomburg, D. *J. Mol. Biol.* **2002**, *320*, 143−156.

(32) Spåhr, H.; Calero, G.; Bushnell, D. A.; Kornberg, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 9185−9190.

(33) Krishna, S. S.; Majumdar, I.; Grishin, N. V. *Nucleic Acids Res.* **2003**, *31*, 532−550.

(34) Krissinel, E.; Henrick, K. *J. Mol. Biol.* **2007**, *372*, 774−797.

(35) McCall, K. A.; Huang, C.; Fierke, C. A. *J. Nutr.* **2000**, *130*, 1437S−1446S.

(36) Tamames, J.; Ramos, M. *J. Mol. Model.* **2011**, *17*, 429−442.

(37) Miura, T.; Suzuki, K.; Kohata, N.; Takeuchi, H. *Biochemistry* **2000**, *39*, 7024−7031.

(38) Maret, W. *BioMetals* **2001**, *14*, 187−190.

(39) Ascone, I.; Strange, R. *J. Synchrotron Rad.* **2009**, *16*, 413−421.

(40) Tamames, B.; Sousa, S. F.; Tamames, J.; Fernandes, P. A.; Ramos, M. J. *Proteins* **2007**, *69*, 466−475.

(41) Sun, T.; Nukaga, M.; Mayama, K.; Braswell, E. H.; Knox, J. R. *Protein Sci.* **2003**, *12*, 82−91.

(42) Taylor, T. C.; Andersson, I. *J. Mol. Biol.* **1997**, *265*, 432−444.

(43) Miessler, G. L.; Tarr, D. A. *Inorganic Chemistry*, 4th ed.; Pearson Prentice Hall: Upper Saddle River, NJ, 2011.

(44) Viiri, K. M.; Jänis, J.; Siggers, T.; Heinonen, T. Y. K.; Valjakka, J.; Bulyk, M. L.; Mäki, M.; Lohi, O. *Mol. Cell. Biol.* **2009**, *29*, 342−356.

(45) Nomura, A.; Sugiura, Y. *Inorg. Chem.* **2004**, *43*, 1708−1713.

(46) Rulíšek, L.; Vondrášek, J. *J. Inorg. Biochem.* **1998**, *71*, 115−127.

(47) Peters, M. B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M. N.; Merz, K. M. *J. Chem. Theory Comput.* **2010**, *6*, 2935−2947.

(48) Read, R. J.; Adams, P. D.; Arendall, W. B., III; Brunger, A. T.; Emsley, P.; Joosten, R. P.; Kleywegt, G. J.; Krissinel, E. B.; Lütteke, T.; Otwinowski, Z.; Perrakis, A.; Richardson, J. S.; Sheffler, W. H.; Smith, J. L.; Tickle, I. J.; Vriend, G.; Zwart, P. H. *Structure* **2011**, *19*, 1395−1412.

(49) Terada, T.; Nureki, O.; Ishitani, R.; Ambrogelly, A.; Ibba, M.; Söll, D.; Yokoyama, S. *Nat. Struct. Mol. Biol.* **2002**, *9*, 257−262.