Research Article

Conservation of metal-coordinating residues

Ioannis N. Kasampalidis [1], Ioannis Pitas [1*], Kleoniki Lyroudia [2]

[1]Department of Informatics, School of Applied Sciences, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

[2]Department of Endodontology, School of Dentistry, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

email: Ioannis Pitas (pitas@aiia.csd.auth.gr)

[*]Correspondence to Ioannis Pitas, Department of Informatics, Box 451, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

Funded by:

- EU; Grant Number: 508803

KEYWORDS

metallo-proteins • metal-binding site • residue conservation • substitution matrix • sequence entropy

ABSTRACT

As a result of rapid advances in genome sequencing, the pace of discovery of new protein sequences has surpassed that of structure and function determination by orders of magnitude. This is also true for metal-binding proteins, that is, proteins that bind one or more metal atoms necessary for their biological function. While metal binding site geometry and composition have been extensively studied, no large scale investigation of metal-coordinating residue conservation has been pursued so far. In pursuing this analysis, we were able to corroborate anecdotal evidence that certain residues are preferred to others for binding to certain metals. The conservation of most metal-coordinating residues is correlated with residue preference in a statistically significant manner. Additionally, we also established a statistically significant difference in conservation between metal-coordinating and noncoordinating residues. These results could be useful for providing better insight to functional importance of metal-coordinating residues, possibly aiding metal binding site prediction and design, metal-protein complex structure prediction, drug discovery, as well as model fitting to electron-density maps produced by X-ray crystallography. Proteins 2007. © 2007 Wiley-Liss, Inc.

Received: 29 August 2006; Accepted: 21 November 2006

DIGITAL OBJECT IDENTIFIER (DOI)

10.1002/prot.21384 About DOI

ADDITIONAL MATERIAL

The supplementary material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

ARTICLE TEXT

INTRODUCTION

Proteins are known to recruit one or more metal atoms in a number of processes involving electron-transfer, or to better stabilize their structure. Together, metal and protein form a distinct biological unit, often referred to as a metallo-protein. Such proteins play a fundamental role in numerous biological processes, as evidenced by the fact that about one-third of known protein structures contain metal-binding sites, as shown by a simple Protein Data Bank (PDB) search.[1] The study of metal- protein interactions is one of the domains of interest of bioinorganic chemistry, a field dealing with the crucial interactions between inorganic metals and biological molecules. This field has attracted relatively little attention, compared with its biological significance.[2]

Recently, a couple of tools specifically devoted to metalloprotein research have been developed. The Metalloprotein Database (MDB)[3] and MSDsite[4] have been designed to analyze structural data of metalloproteins in the PDB, and to provide key information on the geometry and composition of the metal-binding site (MBS). Other notable efforts include analysis of the geometry and composition of MBS by Harding[5] and analysis of MBS geometry and chemical properties by Tainer et al.[6] The extended

environment of mononuclear metal centers is investigated by Karlin et al.,[7] whereas a related study on first-second shell interactions of MBS has been performed by Dudev et al.[8] The flexibility of metal binding sites, that is the conformational change between the apo and the metal-bound state has been the topic of discussion by Babor et al.[9] Such analysis can have significant impact on the functional characterization of metal-binding proteins, drug design, database search for metal-binding-proteins, as described in Andreini et al.,[10] or MBS prediction. Regarding MBS prediction, a number of recent approaches have been proposed, including an energy-based method by Laurie et al.,[11] a support-vector machine predictor of cysteine binding state by Passerini and Frasconi,[12] and a recursive neural network predictor of disulfide bridge connectivity by Vullo and Frasconi.[13]

A key piece of information in some of these methods[12][13] is MBS conservation. However, no large-scale analysis has been performed on conservation of metal-coordinating residues. Such a study could provide justification for using conservation information as a feature for MBS prediction. Additionally, it could provide further insight to the functional importance of certain metal-residue combinations by comparison of the extent of residue conservation coordinating different metal atoms and protein families. Large-scale conservation analysis could also be useful for MBS and drug design, by identifying the residues that are key to maintaining MBS properties.

In this study, we focus on conservation analysis of residues coordinating with some of the metals most commonly found in the PDB, namely: Ca, Cu, Fe, K, Mg, Mn, Na, and Zn. Proteins coordinating with some of these metals have recently been analyzed based on the composition and geometry of the metal-binding site.[5] Here, we distinguish between residues coordinating with a a metal through their side-chain atoms, and those coordinating through the main-chain carbonyl O. We name the former category of residues as side-chain-coordinating and the latter as main-chain-coordinating.

In the first part of our analysis, we investigate the correlation of residue preference, when binding certain metals, with residue conservation. In the second part, we compare the conservation level of side-chain- and main-chain-coordinating residues with that of nonmetal-coordinating residues, and we examine whether residues coordinating with some metals are more conserved than others.

For our analysis, we derived a nonredundant set of metal-coordinating proteins from PDB. Since this set did not contain enough members for a detailed large-scale analysis, we performed homology search via PSI-BLAST[14] to obtain additional putative metal-coordinating proteins, limiting our selection to one ortholog protein from each species. The known-structure proteins were grouped according to families, as defined in Structural Classification of Proteins (SCOP),[15] while their orthologs were also included in the same family. Protein grouping by family was preferred, since same family membership in SCOP indicates clear evolutionary relationship. A multiple sequence alignment (MSA) was performed on all members of each family.

Three main statistical measures were employed in this analysis. The first is the frequency of each coordinating residue per metal. The second is the residue identity ratio, which measures the fraction of residues in the MSA column identical to the residue of interest. The identity ratio is calculated only on known-structure sequences, since for unknown-structure sequences the metal-coordinating residues cannot be guaranteed. As a third measure, we used sequence entropy, a measure that was used recently to study conservation on protein-protein interfaces.[16] Intuitively, sequence entropy is lower when conservation is higher.

METHODS

Dataset

We created a dataset of metal-coordinating structures with the help of PDB[1] using an appropriate query, where we required structures to have resolution better than 2.5 Å and no mutant residues. The reasons for excluding mutant residues is that in a number of occasions, metal-coordinating residues are targeted for site-directed mutagenesis to determine the structural impact of such a mutation. From the structures, which meet the above criteria, we chose only the ones classified in the following SCOP classes: (1) all alpha proteins, (2) all beta proteins, (3) a+b proteins, (4) a/b proteins, and (5) membrane and cell surface proteins and peptides. We also required sequences to have a length greater than 40 residues, since protein sequences are shorter than that are usually dominated by metal-coordinating residues. A nonredundant set was derived from the proteins meeting these criteria with the help of the algorithm by Li et al.[17] at the 90% identity level, as implemented in PDB. The total number of PDB files per SCOP class and metal atom is shown in Table I.

Table I. Number of PDB Files Containing Each metal, Grouped by SCOP Class

	Alpha	Beta	a + b	a/b	Membrane
Ca	146	203	141	153	10
Cu	9	73	12	9	3
Fe	240	70	107	62	13
K	26	20	28	49	2
Mg	104	84	166	359	11
Mn	35	30	60	99	1
Na	47	89	62	95	6
Zn	105	165	177	174	3

Metal-Coordinating Residues

For each structure, the metal-coordinating residues were identified using a uniform distance cut-off of 3 Å from the metal atom. Although this criterion does not take into account the possible distance differences between different metal-residue pairs, it still serves as a good general upper empirical bound, as described in Harding.[5] In addition, this cut-off serves to identify mostly first-shell residues. Only the domains, as defined in SCOP, containing metal-coordinating residues were selected for multiple sequence alignment and these domains were afterwards grouped by SCOP family. A small number of domains from the original set were discarded because their species could not be identified based on the NCBI taxonomy database.[18][19]

Multiple Sequence Alignments

Initial multiple sequence alignments for the domains containing metal-coordinating residues were performed using PSI-BLAST[14] against the NCBI NR database,[18][19] with an e-value cutoff of 10^{-5}. To identify orthologs for each protein, we selected only the reciprocal best hit from each species in the PSI-BLAST reports. In the reports, sequences corresponding to the same species as the query sequence were discarded. Sequences were further filtered by discarding entries with the following keywords: "synthetic," "putative," "probable," "predicted," "hypothetical," "unnamed," "unknown," "unidentified," "designed," "vector." The resulting sequences were grouped with known-structure sequences into SCOP families and the multiple sequence alignments were further refined using MUSCLE (multiple sequence comparison by log-expectation).[20]

Residue Frequency and Identity Ratio

Some types of residues may be more suitable for coordinating with different metals than others. In addition, some residue types may be more important for the protein function, while others may be easily replaced by another with similar chemical properties. To explore the former hypothesis, we calculated the frequency of interactions of each metal with each residue type. For each residue, this frequency is calculated as the fraction of SCOP families containing a specific metal, interacting with that residue type.

To explore the hypothesis that some residues are more important than others, we examine their conservation levels. For this, we initially use a quite simplistic statistical tool, the identity ratio, which is calculated as the ratio of identical residues and the length of the MSA column. Identity ratio was calculated only for residues of known-structure sequences, whereas sequence gaps in the MSA columns were not included in the calculation. The main advantages of identity ratio are that (a) it is computationally cheap to calculate and (b) it ignores chemically similar residue substitutions, thus providing an accurate measure of the "exact" conservation of each residue type. In our analysis, the identity ratios of all residues of known-structure sequences within each SCOP family were averaged, and an average identity ratio was calculated for each metal residue pair in each family.

Sequence Entropy

We have used the information theoretic concept of entropy introduced by Claude Shannon to measure the conservation of a particular residue of a protein sequence. This measure of conservation was recently used to compare conservation levels in protein-protein interfaces.[16] Sequence entropy is given by the following expression:

$$s(i) = \sum p(k)\ \ln(p(k)), \qquad (1)$$

where $p(k)$ is the probability that the ith position in the sequence is occupied by residue type k. A low value of sequence entropy in a position of the multiple sequence alignment indicated that the residues in

that position have been subjected to relatively higher evolutionary conservation than others with a higher value of sequence entropy. The advantage of this approach is that it is intuitive, as high conservation can be viewed as a comparatively ordered arrangement, and therefore one of low entropy.

RESULTS

We followed a multi-faceted approach in our analysis, trying to answer a number of conservation-related questions. First, we explore whether some types of residues interact with metals more frequently and are more conserved than others, while investigating the correlation between frequency and conservation. Second, we investigate whether metal-coordinating residues are more conserved than non-metal-coordinating residues. Expanding on that, we examine whether residues coordinating with some metals are more conserved than others. To answer the first question, we use the identity ratio statistical tool, while for the second question we used sequence entropy. To test statistical significance we use the two-tailed *t*-test.

Residue Type Frequency and Conservation

The frequencies of each of the residue ligands coordinating with each metal are shown in Table II for the side-chain-coordinating residues and in Table III for the main-chain-coordinating residues. The first conclusion that can be drawn from the residue frequency tables is that side-chain residue coordination is far more specific than main-chain binding. Side-chain coordination is dominated by D, E and H, while hydrophobic residues like F, L, and V are seldom found. The rare presence of interaction of A,R,I,L,K,F,W, and V with metals is attributed to possible coordinate errors in the PDB file, with all of these interactions occuring within the 2-3 Å range. On the other hand, main-chain-coordinating residues exhibit a uniform low frequency across most metals. A notable exception is G, which has higher frequency and conservation than the other residues, a fact that could be attributed to its small size and special role in protein structure.[21]

Table II. Side-Chain-Interacting Residues: Frequency of Residue Types Interacting With Each Metal

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ca	0	0.01	0.22	0.61	0	0.06	0.41	0	0.03	0	0.01	0	0	0	0	0.07	0.06	0	0	0
Cu	0	0	0	0.16	0.20	0.04	0.20	0	0.64	0	0	0	0.16	0	0	0.04	0	0	0.08	0
Fe	0	0	0.03	0.14	0.32	0	0.17	0	0.65	0.03	0	0	0.08	0	0	0.03	0.01	0	0.06	0.03
K	0	0	0.04	0.16	0	0.08	0.10	0	0.06	0	0	0	0	0	0	0.12	0.08	0	0.04	0
Mg	0	0.01	0.10	0.44	0	0.05	0.30	0	0.14	0	0.01	0.01	0	0	0.01	0.08	0.09	0	0.01	0.01
Mn	0	0	0.09	0.66	0.01	0.04	0.36	0	0.39	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0	0	0
Na	0.01	0.01	0.05	0.23	0.02	0.02	0.11	0	0.03	0.01	0	0.03	0.01	0	0.01	0.15	0.06	0	0.02	0
Zn	0.01	0.01	0.01	0.30	0.23	0.02	0.30	0	0.69	0	0	0.02	0.01	0	0	0.01	0.03	0.01	0.01	0
Mean	<0.01	<0.01	0.07	0.34	0.10	0.04	0.24	<0.01	0.33	<0.01	<0.01	<0.01	0.03	<0.01	<0.01	0.06	0.05	<0.01	0.03	<0.01

The last row shows the frequency of each residue type averaged over all interacting metals.

Table III. Main-Chain-Interacting Residues: Frequency of Residue Types Interacting With Each Metal

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ca	0.08	0.03	0.07	0.11	0.01	0.08	0.08	0.16	0.02	0.07	0.07	0.07	0.02	0.03	0.06	0.10	0.07	0.02	0.07	0.07
Cu	0	0	0	0	0	0.04	0	0.04	0	0	0	0.04	0	0	0	0	0	0	0	0
Fe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0.08	0.08	0.08	0.06	0.06	0.04	0.02	0.18	0	0.08	0.10	0.02	0.02	0.10	0.06	0.10	0.10	0	0.02	0.20
Mg	0.05	0.02	0.02	0.06	0.02	0.03	0.01	0.04	0	0.03	0.05	0.02	0.01	0.01	0	0.02	0.04	0.02	0.02	0.03
Mn	0	0.03	0.03	0	0	0.01	0.01	0	0	0.01	0	0.01	0.01	0	0	0	0.01	0	0.01	0.01
Na	0.11	0.05	0.07	0.07	0.02	0.05	0.07	0.17	0.02	0.08	0.12	0.03	0.02	0.07	0.07	0.15	0.07	0	0.07	0.13
Zn	0	0	0	0	0.00	0	0	0.00	0.00	0	0	0.00	0	0	0	0	0	0	0	0.00
Mean	0.04	0.03	0.03	0.04	0.01	0.03	0.02	0.07	0.01	0.03	0.04	0.02	0.01	0.03	0.02	0.05	0.04	<0.01	0.02	0.06

The last row shows the frequency of each residue type averaged over all interacting

metals.

Further investigation of metal-residue pair frequencies yields a number of interesting observations. Among residues interacting with Ca, N is found most frequently among side-chain residues, whereas G is found most frequently among main-chain residues. Cu is predominantly coordinated by side-chain residues, where, besides the acidic D and E, the sulphur-containing residues C and M are also quite abundant. The single most frequent side-chain ligand for Fe is H, followed by C, and the acidic D and E, whereas no main-chain coordination is observed. Mg and Mn seem overwhelmingly coordinated by the side-chains of the acidic residues and H, with very little main-chain coordination. This order is reversed for Zn, which is mainly interacting with H, followed by D and E. Extremely little main-chain coordination is observed. Finally, K and Na show a slight preference for the side chains of D, E, and S, as well as the main chain of G and L. However, none of these preferences is particularly prominent.

The identity ratios for all metal-residue coordination pairs are shown in Table IV for the for the side-chain-coordinating residues and in Table V for the main-chain-coordinating residues. Over all metals, identity seems to be correlated with frequency, as the side-chains of D, E, and H exhibit the highest identity, with G ranking first by a wide margin among the main-chain coordinating residues. The correlation coefficient between frequency and identity ratio for all metals is shown in Table VI. The coefficients are all positive, with the exception of the coefficients for the main-chains of Cu, Fe, and Zn, since FE interacts with no main-chain-coordinating residues, whereas Cu and Zn only interact with residue from three and five types, respectively, thus causing numerical errors in the calculation of the correlation coefficient. For side-chain-coordinating residues, the *P*-values, shown in Table VI, are below the statistical significance threshold of 0.05, with the exception of Mn and K. For main-chain residues though, the picture is mixed, with only the *P*-values of Ca, Mg, and Na meeting that threshold. For Ca, the highest identity ratios occur with side chains of D and E and N, whereas main-chain coordination shows high identity as well. For Cu, it is E, which is followed by H and C. For Fe, the highest identity occurs with the side-chains of D, H, R and E. The acidic residues also have some of the highest identity ratios for Mg and Mn, whereas for Zn, H, C, D, and E are the most highly conserved residues. Finally, for K and Na there are no specific residue side-chains with much higher identity, while main-chain interaction is also highly conserved. In general, it is observed that the acidic residues D and E, as well as the basic residue H, obtain the highest values for both identity ratio and frequency for side-chain-, while G ranks first for main-chain-interacting residues.

Table IV. Side-Chain-Interacting Residues: Identity Ratios of Residue Types Interacting With Each Metal

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ca	0	0.04	0.42	0.64	0	0.25	0.63	0	0.22	0	0.14	0	0	0	0	0.31	0.48	0	0	0
Cu	0	0	0	0.34	0.55	0.01	0.75	0	0.70	0	0	0	0.63	0	0	0.48	0	0	0.46	0
Fe	0	0	0.34	0.84	0.82	0	0.70	0	0.82	0.10	0	0	0.64	0	0	0	0.23	0	0.77	0.12
K	0	0	0.41	0.39	0	0.51	0.45	0	0.10	0	0	0	0	0	0	0.66	0.53	0	0.42	0
Mg	0	0.25	0.73	0.73	0	0.37	0.69	0	0.54	0	0.03	0.26	0	0	0.11	0.44	0.54	0	0.20	0.15
Mn	0	0	0.77	0.89	0.50	0.61	0.80	0	0.87	0.47	0.50	0.33	0.47	0.45	0.49	0.45	0.61	0	0	0
Na	0.07	0.20	0.40	0.57	0.37	0.14	0.58	0	0.27	0.14	0	0.34	0.08	0	0.10	0.30	0.27	0	0.42	0
Zn	0	0.13	0.33	0.69	0.76	0.19	0.60	0	0.70	0	0	0.16	0.14	0	0	0.33	0.39	0.14	0.27	0
Mean	0.01	0.08	0.43	0.64	0.37	0.26	0.65	0.14	0.53	0.09	0.08	0.14	0.25	0.06	0.09	0.37	0.38	0.02	0.32	0.03

The last row shows the identity ratio for each residue type averaged over all metals.

Table V. Main-Chain-Interacting Residues: Identity Ratios of Residue Types Interacting With Each Metal

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ca	0.26	0.20	0.29	0.50	0.25	0.20	0.29	0.50	0.25	0.23	0.15	0.28	0.16	0.22	0.30	0.36	0.29	0.05	0.43	0.23
Cu	0	0	0	0	0	0.03	0	0.46	0	0	0	0.03	0	0	0	0	0	0	0	0
Fe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0.41	0.51	0.43	0.18	0.59	0.08	0.19	0.52	0	0.60	0.45	0.50	0.45	0.48	0.23	0.58	0.59	0	0.45	0.33
Mg	0.36	0.26	0.35	0.68	0.34	0.27	0.14	0.58	0	0.29	0.48	0.21	0.22	0.28	0	0.37	0.43	0.33	0.42	0.24
Mn	0	0.42	0.16	0	0	0.09	0.35	0	0	0.47	0	0.05	0.09	0	0	0	0.04	0	0.25	0.09
Na	0.43	0.21	0.48	0.54	0.37	0.24	0.40	0.66	0.34	0.26	0.53	0.16	0.18	0.33	0.44	0.42	0.52	0	0.46	0.40

Zn	0	0	0	0	0.03	0	0	0.14	0.00	0	0	0.10	0	0	0	0	0	0	0	0
Mean	0.18	0.20	0.21	0.24	0.20	0.11	0.17	0.36	0.07	0.23	0.20	0.17	0.14	0.16	0.12	0.22	0.23	0.05	0.25	0.16

The last row shows the identity ratio for each residue type averaged over all metals.

Table VI. Correlation Coefficient(CCF) and Corresponding *t*-test *P*-Values(p-v) Between Identity Ratio and Frequency per Metal

	Ca	Cu	Fe	K	Mg	Mn	Na	Zn	ALL
Side-chain CCF	0.86	0.55	0.62	0.31	0.79	0.85	0.71	0.81	0.86
Side-chain p-v	<0.01	0.16	0.04	0.46	<0.01	<0.01	<0.01	<0.01	<0.01
Main-chain CCF	0.75	N/A	N/A	0.24	0.79	0.29	0.62	N/A	0.73
Main-chain p-v	<0.01	N/A	N/A	0.35	<0.01	0.41	<0.01	N/A	<0.01

Metal-Coordinating Versus. Non-metal- Coordinating Residues

The mean sequence entropy of non-, side-chain-, and main-chain-coordinating columns in each family were calculated by simple averaging, and the families were further grouped according to the metals they interact with. The means and standard deviations of the sequence entropies are shown in Table VII. In our set, Fe was never coordinated by main-chain atoms, therefore no value for main-chain-coordinating residues is shown. The significance of these results was tested by pair-wise *t*-tests, with the *P*-values of these tests shown in Table VIII.

Table VII. Means and Standard Deviations of Sequence Entropy, per Metal, of Noninteracting(NI), Side-Chain-Interacting (SC), and Main-Chain-Interacting (MC) Residues

	Mean			STDEV		
Metal	NI	SC	MC	NI	SC	MC
Ca	0.33	0.26	0.32	0.11	0.22	0.21
Cu	0.33	0.23	0.48	0.13	0.29	0.33
Fe	0.38	0.15	N/A	0.10	0.19	N/A
K	0.35	0.29	0.31	0.11	0.25	0.2
Mg	0.34	0.20	0.23	0.13	0.24	0.21
Mn	0.34	0.1	0.38	0.10	0.17	0.25
Na	0.33	0.28	0.26	0.12	0.23	0.19
Zn	0.33	0.19	0.34	0.11	0.22	0.24

Table VIII. *t*-Test *P*-Values of Sequence Entropy Comparison of Noninteracting (NI) Versus Side-Chain-Interacting (SC) Versus Main-Chain-Interacting Residues (MC)

	Ca	Cu	Fe	K	Mg	Mn	Na	Zn
NI versus SC	<0.01	0.04	<0.01	0.22	<0.01	<0.01	0.05	<0.01
NI versus MC	0.68	0.14	N/A	0.25	<0.01	0.35	<0.01	0.87
SC versus MC	0.02	0.1	N/A	0.82	0.49	<0.01	0.54	0.14

From these data, we conclude that side-chain coordinating residues are more conserved than noninteracting and main-chain-interacting residues. The difference between side-chain- and noninteracting residues is statistically significant for all metals except K, while the difference between side-chain- and main-chain-interacting residues is not always statistically significant. Among all metals, Fe, Mg, Mn, and Zn exhibit the highest statistically significant difference between side-chain- and noninteracting residues. For most metals, there is no statistically significant difference between main-chain- and noninteracting residues, with the exception of Mg, where the main-chain-interacting residues

show a highly significant conservation compared with the noninteracting residues.

DISCUSSION

In this study, we consider the PDB to be representative of a complete genome, which is far from the truth. It is a fact that certain classes of proteins populate the PDB in far greater numbers than others. It is also true that the metal cofactors found in coordination with proteins in the PDB only indicate *in vitro* interactions. What percentage of these interactions also occur *in vivo* remains a topic for further research. In addition, we only considered conservation of first-shell residues, as defined by the 3 Å distance cutoff from the metal atom. This is partially justified on the basis that second-shell residues tend to interact mainly through the peptide backbone atoms, as shown in a recent survey of metal-binding sites of Mg, Mn, Ca, and Zn by Dudev et al.[8] Although, as is shown in our study, first-shell main-chain-coordinating residues are not significantly more conserved than noninteracting residues, it is possible that second-shell backbone-coordinating residues show higher conservation, possibly because of geometric constraints. Therefore, the study of second-shell conservation could provide ground for future research work.

In the first type of analysis in this article, we showed that side-chain coordination is far more specific and conserved than main-chain coordination, a finding that can be attributed to the presence of the carbonyl O atom in all residues. However, some differences in preference are observed even among main-chain-interacting residues, G being the most notable example, especially when interacting with Ca, K, and Na. The only route for G to donate electrons is through the main-chain O atom, and it is speculated that its relatively high incidence can attributed to its small size, which can accommodate possible special structural requirements. On the other hand, Fe, Cu, Zn, and Mn seem to avoid interaction with backbone atoms almost completely, a fact that could be attributed to special geometric or chemical constraints for the binding site of these metals, and it is a matter of further research. In addition, while most metals prefer to interact with the side-chains of coordinating residues, Na and K do not seem to exhibit any notable preference for side-chain or main-chain coordination. In fact, these two metals do not show any specific preference for any residue type whatsoever, while there exists an ongoing debate on whether their interaction with protein residues is electrostatic or not.[5]

On the other hand, the rest of the metals we analyzed seem to have specific preferences for certain residues. Most of side chain interactions occur through hydrophilic residues with available N, O, and S atoms. H is mostly prevalent in proteins contain Fe, Cu, or Zn, donating electrons usually either through $N^{\delta 1}$ or $N^{\varepsilon 2}$. Zn is also frequently coordinated by the acidic residues, as well as by C. Sulphur-containing C is observed as a ligand with the fourth highest frequency and fifth highest identity ratio is the fifth highest among all residue types, donating electrons through the S atom. The acidic bidentate residues can donate electrons via the $O^{\delta 1}$ or $O^{\delta 2}$ and the back-bone O atom. They are particularly dominant in Mg and Mn interactions. Mg and Mn exhibit very similar ligand frequencies and identity ratios, since these two metals often compete with each other in nature. Acidic residues are also the main preference for Ca, while, overall, we find that D, E, and H have the highest identity ratios and frequencies. N ranks fifth in frequency and fourth in identity ratio, donating electrons through the $N^{\delta 2}$ atom. Finally, the correlation coefficient between identity ratio and frequency was positive and statistically significant mostly for side-chain interacting residues. This important finding corroborates anecdotal evidence that residue types that interact more frequently with a certain metals tend to be more conserved, when coordinating with that metal.

The next valuable bit of information from this study is the evidence showing that side-chain-coordinating residues are significantly more conserved than noncoordinating residues, whereas this is not true for main-chain-coordinating residues. The only exceptions are K, where there exists no significantly higher conservation for interacting residues and Mg and Na, where main-chain interacting residues are also significantly conserved.

In our analysis, residues coordinating with Fe, Zn, Mg, and Mn showed significantly higher conservation than noninteracting residues of proteins with those metals. Incidentally, two of these metals, Zn and Fe, are the ones most abundant in biological systems.[22] Zinc is a crucial metal, used, among others, to stabilize DNA-binding domains in a number of transcription factors, p53 being a most important example.[23] Another main family of proteins utilizing zinc is the zinc-finger family, which includes several hundreds of members, including Krox-26, a transcription factor, which may be involved in the molecular regulation of tooth development and amelogenesis.[24] On the other hand, Fe is probably the most conspicuous example of metal in bioinorganic chemistry, used in hemoglobin[25] and a number of other electron-transfer functions. Mg is a necessary element in DNA repair, since all downstream activities of major base excision repair proteins, such as APE, DNA polymerase, and DNA ligases utilize this metal.[26] Finally, Mn seems to play an important part in photosynthesis, among other roles, where strong evidence suggests the existence of cubane-like Mn_3CaO_4 cluster linked to a fourth Mn by a mono-μ-oxo bridge within the oxygen-evolving catalytic site of photosystem II.[27]

Among metals with relatively lower, but still statistically significant, conservation, Cu stands out as participating mainly in enzymes such as oxido-reductases, with cytochrome c oxidase being a prominent example, as the terminal enzyme of the respiratory chains of mitochondria and aerobic bacteria.[28] Ca proteins participate in cell signaling, with the S100 family of proteins forming the largest group of EF-hand signalling proteins in humans.[29] Finally, Na and K, participate in a number of proteins with diverse functions, including Na^+,K^+-ATPase.[30]

CONCLUSION

Life evolved on earth's crust, thus having access to abundant mineral supplies, which were recruited when organic compounds could not satisfy the ever increasing complexity of biological processes. Proteins evolved in a similar manner when amino acids, despite their remarkable chemical diversity, could not meet the demands of processes such as those involving electron transfer or complex structure stabilization. We have examined the conservation of the most abundant of metals found in coordination with proteins and reach a number of important conclusions. The two most important findings are (a) for residues coordinating through their side-chain, the correlation between residue conservation and residue preference is highly significant for most metals and (b) the highly significant conservation of side-chain-coordinating residues versus the conservation of noncoordinating residues for most metals.

Finally, we like to mention that the main impact of our results lies in the fact that metal-binding is still not a fully understood process; therefore, conservation information could provide valuable information for a number of tasks, including metal binding site prediction and design, metal-protein complex structure prediction, drug discovery, as well as model fitting to electron-density maps produced by X-ray crystallography.

Acknowledgements

We thank Prof. Frasconi at the University of Firenze for introducing us to this topic. Finally, we also thank the anonymous reviewers for their excellent and constructive comments.

REFERENCES

1 Berman HM,Westbrook J,Feng Z,Gilliland G,Bhat TN,Weissig H,Shindyalov IN,Bourne PE. The protein data bank. *Nucleic Acids Res* 2000; **28**: 235-242. Links

2 Bertini I,Rosato A. Bioinorganic chemistry in the postgenomic era. *Proc Natl Acad Sci USA* 2003; **100**: 3601-3604. Links

3 Castagnetto JM,Hennessy SW,Roberts VA,Getzoff ED,Tainer JA,Pique ME. Mdb: the metalloprotein database and browser at the scripps research institute. *Nucleic Acids Res* 2002; **30**: 379-382. Links

4 Golovin A,Dimitropoulos D,Oldfield T,Rachedi A,Henrick K. Msdsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins* 2005; **58**: 190-199. Links

5 Harding M. The architecture of metal coordination groups in proteins. *Acta Crystallogr D* 2004; **60**: 849-859. Links

6 Tainer JA,Roberts VA,Getzoff ED. Protein metal-binding sites. *Curr Opin Biotechnol* 1992; **3**: 378-387. Links

7 Karlin S,Zhu Z,Karlin KD. The extended environment of mononuclear metal centers in protein structures. *Proc Natl Acad Sci USA* 1997; **94**: 14225-14230. Links

8 Dudev T,Lin Y,Dudev M,Lim C. First-second shell interactions in metal binding sites in proteins: a pdb survey and dft/cdm calculations. *J Am Chem Soc* 2003; **125**: 3168-3180. Links

9 Babor M,Greenblatt HM,Edelman M,Sobolev V. Flexibility of metal binding sites in proteins on a database scale. *Proteins* 2005; **59**: 221-230. Links

10 Andreini C,Bertini I,Rosato A. A hint to search for metalloproteins in gene banks. *Bioinformatics* 2004; **20**: 1373-1380. Links

11 Laurie AT,Jackson RM. Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 2005; **21**: 1908-1916. Links

12 Passerini A,Frasconi P. Learning to discriminate between ligand bound and disulfide bound cysteines. *Protein Eng Des Sel* 2004; **7**: 367-373. Links

13 Vullo A,Frasconi P. Disulfide connectivity prediction using recursive neural networks and multiple alignments. *Bioinformatics* 2004; **20**: 653-659. Links

14 Altschul SF,Madden TL,Schoffer AA,Zhang J,Zhang Z,Miller W,Lipman DJ. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25**: 3389-3402. Links
15 Murzin AG,Brenner SE,Hubbard T,Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995; **247**: 536-540. Links
16 Guharoy M,Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci USA* 2005; **102**: 15447-15452. Links
17 Li W,Jaroszewski L,Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001; **17**: 282-283. Links
18 Benson DA,Karsch-Mizrachi I,Lipman DJ,Ostell J,A RB,DL W. Genbank. *Nucleic Acids Res* 2000; **28**: 15-18. Links
19 Wheeler DL,Chappey C,Lash AE,Leipe DD,Madden TL,Schuler GD,Tatusova TA,Rapp BA. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2000; **28**: 10-14. Links
20 Edgar RC. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**: 1792-1797. Links
21 Branden C,Tooze J. *Introduction to protein structure*, 2nd ed. New York; Garland Publishing: 1999.
22 Coleman JE. Zinc proteins: enzymes storage proteins transcription factors and replication proteins. *Annu Rev Biochem* 1992; **61**: 897-946. Links
23 Cho Y,Gorina S,Jeffrey PD,Pavletich NP. Crystal structure of a p53 tumor suppressor-dna complex: understanding tumorigenic mutations. *Science* 1994; **265**: 346-355. Links
24 Gao Y,Kobayashi H,Ganss B. The human krox-26/znf22 gene is expressed at sites of tooth formation and maps to the locus for permanent tooth agenesis (he-zhao deficiency). *J Dent Res* 2003; **82**: 1002-1007. Links
25 Fermi G,Perutz MF,Shaanan B,Fourme R. The crystal structure of human dcoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol* 1984; **175**: 159-174. Links
26 Adhikari S,Toretsky JA,Yuan L,Roy R. Magnesium, essential for base excision repair enzymes, inhibits substrate binding of n-methylpurine-dna glycosylase. *J Biol Chem* ; 2006; **281**: 29525-29532. Links
27 Ferreira KN,Iverson TM,Maghlaoui K,Barber J,Iwata S. Architecture of the photosynthetic oxygen-evolving center. *Science* 2004; **303**: 1831-1838. Links
28 Michel H,Behr J,Harrenga A,Kannt A. Cytochrome c oxidase: structure and spectroscopy. *Annu Rev Biophys Biomol Struct* 1998; **27**: 329-356. Links
29 Santamaria-Kisiel L,Rintala-Dempsey AC,Shaw GS. Calcium-dependent and -independent interactions of the s100 protein family. *Biochem J* 2006; **396**: 201-214. Links
30 Matsui H,Homareda H. Interaction of sodium and potassium ions with Na+, K+-atpase. ouabain-sensitive alternative binding of three Na+ or two K+ to the enzyme. *J Biochem* 1982; **92**: 193-217. Links