Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Pattern prediction and coordination geometry analysis from cadmium-binding proteins: a computational approach

R. Jesu Jaya Sudan[a] and C. Sudandiradoss[a,b]*

[a]Bioinformatics Division, VIT University, Vellore 632 014, India, and [b]Centre for Nanobiotechnology, VIT University, Vellore 632 014, India

Correspondence e-mail: csudandiradoss@vit.ac.in

Received 10 November 2011
Accepted 25 June 2012

Cadmium toxicity has been reported to have major health effects including carcinogenicity, respiratory disorders, kidney failure, neurotoxicity and liver dysfunction. Understanding the nature of the association of cadmium with biomolecules has thus become imperative and a key factor in predicting the phenomena behind predisposition to disease. Accordingly, a computational investigation of cadmium-binding characteristics was performed using about 140 cadmium-bound structures and 34 cadmium-binding sequences. The metal-coordinating architecture defining the chelate loops, residue arrangement, secondary-structural characteristics, distances and angles were analyzed. Binding patterns were predicted based on the probability of occurrence of residues within the coordination distance and were further corroborated with sequence patterns obtained from cadmium-binding proteins. About 56 different chelate loops were identified. Based on these chelate loops, putative cadmium-binding patterns were derived that resembled short-length motifs, namely Y-*X*-G-*X*-G, Q-X_9-E, E-X_2-E-X_2-E and T-X_5-E-X_2-E, which were observed within the conserved regions of the cadmium-binding proteins. The poorer conservation of residues around these motifs resulted in a deviating pattern against the coordination loops. These structure-based motifs are proposed to be an efficient tool in building chelators for the effective removal of cadmium.

1. Introduction

Metals are important constituents of life, driving economic activity and industry, but can also be a hazard to human health (Briner, 2010). Naturally occurring heavy metals are involved in much of human industry and many products; hence, exposure to heavy metals has become a common phenomenon owing to their environmental pervasiveness (Järup, 2003). 'Heavy metals' is frequently used as a group name for metals and semimetals (metalloids) that have been associated with contamination and potential toxicity (Duffus, 2002). Among the heavy metals, cadmium, lead and mercury are examples of toxic metals that are not essential for nutrition. The toxic effects of these metals may be mediated or enhanced by interaction with or deficiencies of nutritionally essential metals such as calcium, iron, zinc and selenium (Goyer, 1995; Thivierge & Frey, 2006; Soghoian & Sinert, 2009; Levander, 1978). Toxic metals serve no biological functions, therefore their presence in tissues reflects contact of the organism with its environment. These metals are cumulative poisons and are toxic even at low doses; they are also nonbiodegradable, with a very long biological half-life (Chowdhury & Chandra, 1987; Barbier *et al.*, 2005). In addition, individual differences which are caused by genetic, nutritional, hormonal, habitual and

many other factors may also contribute to the toxic or metabolic effects of a single metal (Tsuchiya, 1977).

Cadmium has been in industrial use for a long period of time. It is widely used in industrial processes as an anticorrosive agent, as a stabilizer in PVC products, as a colour pigment, as a neutron absorber in nuclear power plants and in the fabrication of nickel–cadmium batteries (Godt *et al.*, 2006). As a result, cadmium emissions have increased enormously, one reason being that cadmium-containing products are rarely recycled and are often dumped together with household waste (Järup, 2003). Hazards of cadmium absorption have been reported to include shortness of breath, lung oedema, destruction of mucous membranes in cadmium-induced pneumonitis, kidney damage, itai-itai disease, carcinogenicity and defects in the central nervous system (Godt *et al.*, 2006). Cadmium has been classified as a human carcinogen, affecting health through occupational and environmental exposure. The prostate is one of the organs with the highest levels of cadmium accumulation. Importantly, patients with prostate cancer appear to have higher levels of cadmium both in the circulation and in prostatic tissues (Golovine *et al.*, 2010). Cadmium acts in almost all stages of the oncogenic process and is thought to act through multiple nonexclusive mechanisms such as oxidative stress, oncogene activation, apoptotic bypass and altered DNA methylation. Recently, it has been proposed that cadmium acts as a metalloestrogen *via* interactions with oestrogen receptor α (ER-α), stimulating downstream oestrogen-related processes. As a result, cadmium acts as a xenoestrogen in oestrogen-related cancers such as breast cancer (Benbrahim-Tallaa *et al.*, 2009). Cadmium can also cause bone damage, either *via* a direct effect on bone tissue or indirectly as a result of renal dysfunction (Järup & Akesson, 2009). Cadmium is known to inhibit protein-synthesis, carbohydrate-metabolism and drug-metabolizing enzymes in the liver of animals, including humans (Nath *et al.*, 1984).

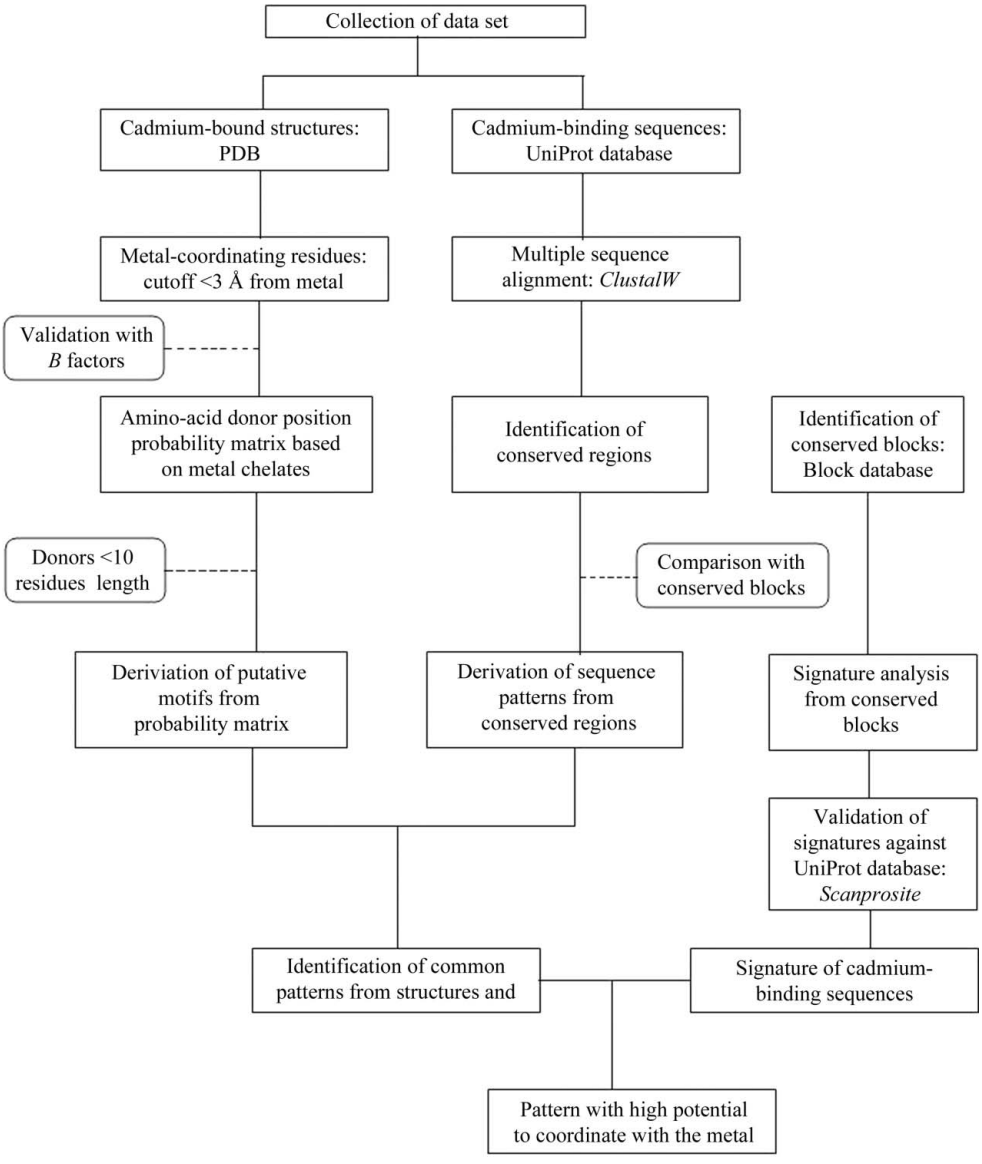

Figure 1
Workflow of the coordination study and pattern analysis.

Toxic metal removal involves different methods of chelation employing synthetic, chemical and peptide chelators. The concept of chelation is based on simple coordination chemistry. Coordination of transition metals to peptides through the incorporation of either unnatural chelating groups or amino-acid-ligating side chains expands the utility of peptides for biological studies (Ma, 2010). Coordination chemical strategies have already paved the way to successful clinical applications. The study of the association of toxic metals with functionally vital biomolecules will eventually produce significant information on the type, geometry and structure of the residues that favour such coordination. These results, when combined with chelation therapy, will produce new insights into the design of effective peptide chelators. Many studies have been reported that employ coordination chemistry in the design of chelators (Jones, 1994; Archibald, 2011; Banerjee *et al.*, 2005; Flora & Pachauri, 2010). Although various chelation strategies have been developed for the removal of cadmium, the successful use of such chelators has yet to be achieved. A thorough

understanding of the mechanism of the association of cadmium with biological molecules is required to predict the behaviour of the metal in different environments, namely the structural folds, physiochemical properties and solvent preferences that govern the function of the proteins. This study is focused on understanding the coordination geometry of cadmium based on the description of the geometry of metal ion-binding sites within proteins of Harding (2004). The coordination study gives an account of the residue, residue position, donor atoms and distances that provides an insight into the choice of chelators. We also examined the distributions of bond lengths and coordination numbers together with the *B* factor (the displacement parameter, sometimes referred to as the 'temperature factor') and the relative occupancies of metal ions (Zheng *et al.*, 2008). The overall workflow of the current study is shown in Fig. 1.

2. Materials and methods

2.1. Construction of the data set

A search was made of the Protein Data Bank (PDB; Berman *et al.*, 2000) to identify all possible cadmium-bound protein structures. About 570 structures that contained bound cadmium ions were identified. To maintain uniformity in our analysis, we selected only X-ray crystallographic structures of cadmium-bound proteins. Any NMR models or DNA-associated proteins were excluded, as a result of which about 466 cadmium-bound proteins were obtained. Redundancy in the data set was removed using *PISCES* (Wang & Dunbrack, 2003) such that no two sequences shared >40% identity. The resultant data set comprised a diverse set of 195 cadmium-bound proteins. In order to study the differences in cadmium coordination as a factor of resolution, we subsequently categorized the data set into three classes: group *A* (0.5–1.9 Å resolution), group *B* (2.0–2.9 Å resolution) and group *C* (3 Å resolution and above). Consequently, there were 99 structures between 0.5 and 1.9 Å resolution, 94 structures in the 2.0–2.9 Å resolution range and only one structure in group *C*. Hence, group *C* was ignored for further study. To make the analysis more specific, we categorized the proteins based on their functional folds obtained from the SCOP and CATH databases (Csaba *et al.*, 2009).

2.2. Metal-coordinating groups and patterns from cadmium-bound structures

In the coordination of metals by most proteins, specific amino acids preferentially interact with the metal. This has been well documented for biologically important metals that act as cofactors in enzymatic reactions. The specificity of the coordinating residues and their geometry provides an insight into the favourable structural organization of metal coordination. In recent years, metal coordination has been emphasized for most toxic metals. Bioremediation and phytoremediation studies have led to the identification of various metal-resistant microbes and plants, and have resulted in the sequencing and analysis of metal-binding proteins. However, the binding patterns of such proteins are not yet fully understood. Therefore, we have determined the cadmium-binding patterns and their geometry using a set of cadmium-bound structures. The cadmium-coordination group was derived on the basis of the work of Harding (2004). The coordinating residues and their atom types, atomic orientations, atomic distances from the metal and sequence patterns were analyzed.

2.2.1. Determination of cadmium-coordinating residues from protein structures. All 195 cadmium-bound structures in the data set were subjected to analysis of their cadmium-coordinating residues. *ANAMBS* (Kuntal *et al.*, 2010), a standalone tool that predicts the microenvironment of a metal atom in a protein structure within a specified distance, was employed to predict the cadmium-coordinating residues. A distance cutoff of 3 Å was used in the present study. Although this factor does not account for any specificity in cadmium interaction, it is suggested to be a good upper experimental threshold. Additionally, this cutoff serves to identify mostly the first-shell residues (Kasampalidis *et al.*, 2007). For all structures, features such as residues, residue positions, atoms and the distance of the atoms from cadmium within the specified cutoff distance were analyzed. The results of the tool were cross-validated with *Swiss-PdbViewer* (Guex & Peitsch, 1997) for about 30 structures. To avoid redundancy of results owing to homodomain structures, only a single polypeptide chain from each structure with a bound cadmium ion was considered. If multiple cadmium ions were present, only a single ion in the first chain was considered. Metal coordinations with alternative conformers were excluded. The details of the coordination study include chelate loops, size, donor atoms, nonresidue donors such as water or any heteroatoms coordinating with the metal and the coordination number, which defines the total number of occupied coordination sites around the metal ion. The PDB entry and corresponding chain for each structure analyzed are also provided. The chelate loops were further analyzed for their secondary-structural characteristics, φ–ψ angles, structural folds and sequence conservation within ten residues downstream and upstream of the chelate loop. The chelate-loop conformations and their similarities were visualized using *PyMOL* (DeLano, 2002) and *Discovery Studio* (Accelrys Software Inc.).

2.2.2. Validation of metal-coordinating residues. The coordinating residues predicted from all structures in the data set were validated based on the displacement and occupancy parameter values. These values were obtained from the PDB files of the structures. The *B* factor for the metal environment was calculated as the mean *B* factor of all atoms within 3 Å of the metal. Individual structures with *B* factors as low as 2.0 Å^2 and occupancies outside the range 0.1–1 were considered to be incorrect and such entries were excluded. The correlation between the metal and the residue *B* factors was plotted as a factor of the average deviation. All outliers showing an average deviation of $\geq$6.0 were ignored in the present study.

2.2.3. Prediction of metal-coordinating patterns from cadmium-bound structures. Sequence patterns are defined as stretches of residues that represent structurally or

functionally important regions of a protein (Bork & Koonin, 1996). Putative binding patterns were analyzed using the positional frequencies of the residues within the coordination distance. The binding patterns were derived from chelate residues spanning a sequence length of <10 residues within coordinating distance. Since the residues of chelate loops are positioned at random distances, they are observed to mostly exceed the suitable length for pattern writing. As a result, structures with cadmium-binding sites comprising less than two residues and coordinating amino acids at a distance of >10 residues within the specified 3 Å cutoff were excluded. By observing the coordinating residues and their relative positions with respect to the preceding amino acids, we created a table with positions in the rows and preceding amino acids in the columns. We placed all the coordinating amino acids at the corresponding position at which they succeed their nearest predecessor in the coordination group. We derived possible patterns from the table and cross-validated them with the sequence motifs predicted from cadmium-binding proteins.

2.3. Sequence-motif prediction from cadmium-binding proteins

In most protein analyses, sequence motifs are used as representatives of the function of the protein (Bork & Koonin, 1996). As proteins exhibit a structure–function relationship, structural features of proteins become vital in order to elucidate their functions. Notably, in metal–protein interactions the geometry of the atoms in the binding site determines the efficacy of interaction. A structural motif is thus necessary to study the interactions of proteins with heteroatoms and ions. Since the structures of cadmium-binding proteins have not been well studied, we have adopted a method to determine their binding motifs by observing the coordination in experimentally bound structures and validating them using sequence motifs identified through multiple sequence alignment.

2.3.1. Multiple sequence alignment of cadmium-binding proteins. The protein sequences required for the prediction of metal-binding motifs were retrieved from the UniProt database (Apweiler *et al.*, 2004). The data set encompassed about 34 cadmium-binding proteins and excluded multi-metal-binding proteins. To facilitate the identification of conserved regions in the proteins, we performed a multiple sequence alignment of all cadmium-binding proteins using *ClustalW* (Thompson *et al.*, 1994) and the EBI database (Emmert *et al.*, 1994). A slow alignment method using a Gonnet matrix with gap open and extension penalties of 10 and 0.1 was used for pairwise alignment. For multiple sequence alignment the matrix remained the same; the gap was open and extension penalties of 10 and 0.2 were used. The minimum gap distance was set to 15; no end-gap values were used and no iterations were specified. The individual segments of the conserved regions were predicted using the Block database (Henikoff *et al.*, 1999).

2.3.2. Determination of cadmium-binding patterns. Functionally or structurally important regions in a protein family are well conserved across species. However, owing to extensive mutations and speciation, proteins belonging to a similar family tend to show a larger degree of variation (Vulić *et al.*, 1999). Therefore, the conserved regions of multiple sequence alignment predicted by the heuristic approach were chosen to write patterns of residues that could be involved in binding cadmium (Giri *et al.*, 2004). Conserved segments were identified as short regions without gaps and with higher similarities. To account for orthologues, conserved regions with identities, substitutions, semi-conserved substitutions and variations were included for pattern writing. The patterns were written in the regular PROSITE format and were further validated.

2.4. Validation of PROSITE patterns derived from cadmium-bound protein sequences

Validation of the PROSITE patterns involved identifying the best pattern that detects the respective protein family but shows dissimilarity to unrelated proteins. Validation was performed based on the degree of similarity to the protein family and matches to related proteins across species. For this purpose, the sequence patterns predicted from multiple sequence alignment of cadmium-binding proteins were validated against UniProt-TrEMBL database sequences (O'Donovan *et al.*, 2002) using *Scanprosite* (de Castro *et al.*, 2006), a tool for detecting pattern matches between protein sequences. The extent of matches showing related and unrelated proteins was analyzed. A phylogenetic tree of the sequences from scan results was constructed and the diversity of the taxonomical classes that match the patterns was studied. Patterns giving the best expected results were further cross-verified against binding patterns obtained from cadmium-bound structures. Cross-validation of the sequence pattern with the structural patterns required manual comparison of the residue occurrences in both motifs. This pattern matching was performed with low stringency to account for the variation among functionally distinct structures and sequences.

3. Results and discussion

3.1. Validation by *B* factors of the metal ion and coordinating residues

Cadmium-coordinating residues were predicted for all 194 structures in the data set. To obtain significant results, the data set was validated by its *B* factor and the occupancy values of the coordinating residues and metal. The observed *B* factor was noted to be >2.0 Å^2 in all structures and the occupancies were well within the range 0.5–1.0. The mean *B* factors for the metal environment ranged between 2 and 60 Å^2. However, the *B* factors of the metal and coordinating residues were poorly correlated. As stated by Zheng *et al.* (2008), the *B* factor of a properly determined and refined metal ion should be close to the *B* factor of its coordinating atoms. For this reason, outliers that exceeded an average deviation of 6.0 Å^2 were eliminated. Consequently, we obtained a statistically significant correlation of 0.86. A scatter plot showing the correlation between the metal and its environment is shown in Fig. 2. The resultant

data set was limited to 140 cadmium structures after the validation and was further subjected to metal-coordination analysis.

3.2. Cadmium-coordinating residues

In all of the proteins analyzed, the acidic amino acids glutamic acid and aspartic acid and the basic amino acid histidine were found to predominate within coordinating distance of cadmium, indicating their close association with the metal. The polar amino acid cysteine was found to be the next most dominant residue. The order of preference for the amino acids in cadmium-bound structures was predicted to be Glu > Asp > His > Cys. The nonpolar residues tryptophan, phenylalanine and alanine showed no preference for cadmium in either group. Other hydrophobic amino acids, namely methionine, glycine, leucine and proline, were observed in high-resolution structures, whereas valine was only observed in group *B*. However, both groups showed similar results for the dominantly coordinating residues. These results indicate that cadmium has a greater preference for charged and polar amino acids.

We further analyzed the atomic preference of cadmium by predicting the atoms closely associated with the metal at the specified cutoff distance. The preference for side-chain atoms over backbone atoms was also analyzed. The specific atom types and their numbers of interactions are plotted and shown in Fig. 3(*a*). Analysis of the data set revealed a preference for the ε oxygen (OE) and the δ carbon (CD) of glutamic acid. Similarly, several δ-oxygen (OD) and γ-carbon (CG) interactions were observed for aspartic acid. The histidines favoured side-chain interactions at OE, the ε nitrogen (NE) and the δ nitrogen (ND) through the pyrolidine ring, and the cysteine at the γ-sulfur (SG) atom. In all of these amino acids the backbone atoms did not contribute to the association with the metal. Unlike the amino acids Glu, Asp, His and Cys, the polar and hydrophobic residues, namely methionine, glycine, isoleucine, tyrosine, proline, leucine, serine and valine, showed a backbone-atom preference towards cadmium. Of the backbone atoms, the carbonyl O atom contributed most towards cadmium interaction, with the exceptions of serine and leucine which coordinated through the backbone N atom and C atom,

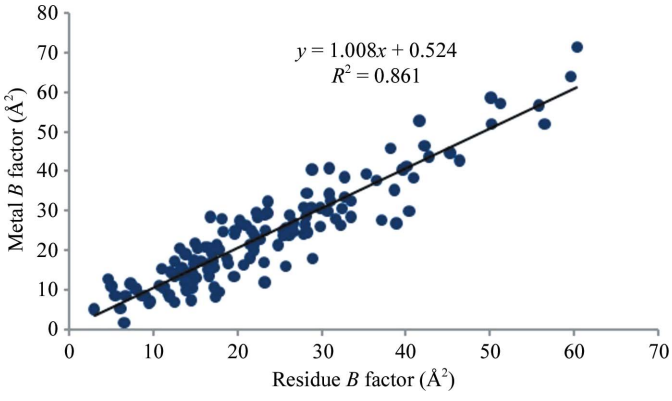

Figure 2
Correlation plot of the metal and the metal environment.

respectively. The basic amino acids asparagine and glutamine showed a preference for coordination at both backbone and side-chain atoms through the OE, OD, ND and backbone O atoms. Overall, the analysis suggested that cadmium has a higher preference for side-chain atoms over backbone atoms. The strongest interaction of the atoms with cadmium was predicted by analysing the atomic distances. From analysis of the atomic distances of all atoms within the specified cutoff, we found only six atoms, namely OE1 and OE2 of glutamic acid, OD1 and OD2 of aspartic acid and ND1 and NE2 of histidine, at distances of <2 Å from cadmium. The strongest association was observed with the δ-nitrogen of histidine, with a closest distance of 1.9067 Å to cadmium. The distance ranges for the atoms were observed to be between 1.9067 and 2.577 Å for nitrogen, between 1.9464 and 2.993 Å for oxygen, between 2.4391 and 2.9943 Å for carbon and between 2.2119 and 2.824 Å for sulfur. The overall interaction study revealed that side-chain N and O atoms prefer a coordinating distance of between 2.2 and 2.5 Å. Owing to the closer association of the ε- and δ-oxygens, the corresponding δ- and γ-carbon atoms usually are found at a larger distance range of 2.7–3.0 Å, while S atoms prefer 2.4–2.6 Å as indicated in Fig. 3(*b*). From all of the structures analyzed, we found that the presence of at least one of the atoms OE, OD, ND or NE is required for coordination to cadmium ion.

3.3. Chelate loops and structural properties

The chelate loops were predicted for all of the structures in the data set and are listed in Table 1.[1] The metal-coordination group is represented by the chelate loop and its size, the residue length of the chelate and the water atoms or heteroatoms that coordinate the metal. Additionally, the secondary-structural features of the chelate-loop residues and the structural classification of the proteins according to CATH and SCOP are provided. Different patterns of chelate loops were noted with differences in their lengths. Analyses of the chelate-loop patterns, sequence conservation, structural deviation and chelate-loop conformations are presented and discussed.

3.3.1. Chelate-loop patterns. A wide range of chelate-loop patterns was observed. A maximum of five protein donors were found to coordinate to the metal; however, the coordination number ranged up to seven. Despite the presence of many water and nonresidue donors within the coordination distance, the metals in most of the structures preferred the presence of at least a single residue donor within the coordination sphere. This may account for the stability of cadmium coordination to residues over other nonresidue donors. The most common donor pairs observed among the coordinating residues and their numbers of occurrences are given in Table 1(*b*). As can be seen from the table, the EH pattern remains the most dominant, followed by combinations of other residues, namely aspartic acid, histidine and glutamic

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: XB5048). Services for accessing this material are described at the back of the journal.

Table 1
Cadmium coordination.

(*a*) Cadmium-coordinating residues predicted from a cutoff distance of 3 Å. PID, PDB code followed by chain ID. Position represents the position of the first residue in the chelate loop. n_{span} represents the length span of the chelate loop. n_p represents the number of donor atoms; dons indicate the donor atoms specified as single-letter codes. Met is the metal ion and sd1–sd7 give the positioning of the residues from the first residue of the chelate. Only part of the table is shown here. The complete table has been deposited as Supplementary Material.

PID	Position	n_{span}	n_p	dons	met	sd1	sd2	sd3	sd4	sd5
1r0i_*A*	6	36	4	CCCC	CD	3	30	3	—	—
1rzm_*A*	102	207	4	CHED	CD	170	26	11	—	—
1cdp_*A*	90	11	5	DDDKE	CD	2	2	2	5	−1
1con_*A*	8	16	4	EDDH	CD	2	9	5	−1	—
3kbs_*A*	217	40	4	EHDD	CD	3	35	2	—	—

(*b*) The commonest donor pairs and their corresponding number of entries in the data set.

Donor pairs	DD	DE	EE	DH	HH	EH	CC
No. of occurrences	10	15	15	17	15	23	6

(*c*) The diverse ranges of chelate lengths.

Chelate size	0	1	2	3	4	5	6–10	11–20	21–30	31–50	51–100	101–200	201–400
No. of chelates	2	43	4	2	10	6	7	15	11	10	13	8	9

(*d*) The minimum and maximum number of donors within the coordination distance.

No. of donors	0	1	2	3	4	5
No. of chelates	2	43	48	24	18	5

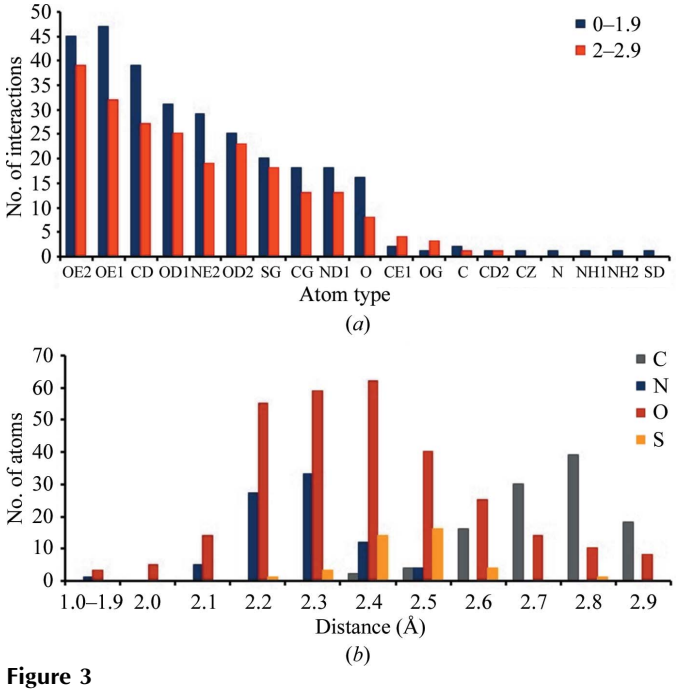

Figure 3
Cadmium-coordinating atoms and distances. (*a*) The extent of specific atomic interactions with cadmium for group *A* (blue) and group *B* (red) structures. Atom types are indicated by their chemical symbols: O, C, N and S represent oxygen, carbon, nitrogen and sulfur, respectively. E, D, G and Z represent the ε, δ, γ and ζ positions of the corresponding atoms. NH1 and NH2 represent the H atoms of the amino group. (*b*) The coordination distance ranges of the atoms. Colours indicate carbon (grey), nitrogen (blue), oxygen (red) and sulfur (yellow).

acid. However, the sequence length between the chelating residues varies among the different structures and ranges from a single residue to 361 residues. Identical chelates also vary in length, which may probably arise from differences in their structural folds. For instance, in CCCC chelates the residue positions were noted to be 3 *n* 1/3, where *n* = 2–30, and their corresponding lengths ranged from 10–36 residues. Similarly, in the biresidue chelate 'DD' the lengths varied widely. Thus, it is evident that there is no correlation between the number of donors and the chelate length. In contrast, in the similar but non-identical chelates DDD, DDDKE and DDTKE successive aspartates were noted at the *n*, *n* + 2 and *n* + 4 positions. In the CHED chelate, the residue positions were observed to be 170 26 11, 174 37 11 from the first residue. These details clearly indicate that cadmium coordination prefers proximity of the chelate residues and not their one-dimensional or three-dimensional profiles. Thus, the geometry or orientation of the residues within the interaction distance of the metal becomes the primary factor for its coordination. A statistical report on chelate size and number of donors is given in Tables 1(*c*) and 1(*d*).

3.3.2. Sequence conservation among chelate loops. For all the dominant patterns, the sequence conservation around the chelate was predicted by observing ten residues upstream and downstream of the coordinating residues. We noted that in all of the CCCC chelates every cysteine residue had a conserved glycine adjacent to it. The significance of glycine proximal to cysteine residues can be seen by flexibly placing cysteines at a favourable distance and angle for coordination with the metal. DD chelate loops also had glycines in their proximity but with no positional conservation; however, they did have a glutamic acid conserved at the ninth or tenth position downstream of the loop. The EE chelates had no glycine conservations but were rich in glutamic acids. A minimum of three glutamates were noted in each of the EE chelate loops and around seven glutamic acids were observed in some. The abundance of glutamates around the chelate loop creates a stronger acidic environment that eventually aids in a stronger affinity and effective chelation of the metal. Most of the glutamic acids in the chelates were observed to be polydentate, thus proving their chelating ability. No significant conservation was observed for the other chelates.

3.3.3. Structural deviation owing to cadmium coordination. The structural distortions in the protein structure arising from cadmium binding were predicted from the torsion angles of the residues within the coordination sphere. The empirical

distribution of the φ–ψ angles across five regions of the Ramachandran plot for the coordinating residues were validated. The five regions of the Ramachandran plot were defined based on the values reported by Deane & Blundell (1999). The regions A, B and E refer to α-helix ($\varphi = -180$ to 0°, $\psi = -120$ to 60°), β-sheet ($\varphi = -180$ to 0°, $\psi = 60$ to -240°) and left-handed α-helix ($\varphi = 90$ to 100°, $\psi = -20$ to 80°), respectively. Regions C and D correspond to the partially allowed region $\varphi = -180$ to -40°, $\psi = 0$ to -40°. We observed that group A structures had a lower deviation than group B structures. The α-helical residues were well confined within the core regions in both groups; however, most β-sheet residues showed larger deviations, as indicated in Fig. 4. This could probably account for the lower flexibility owing to planarity and hydrogen bonding between the strands. The residues in the coils were also well stabilized. However, these distortions were predicted to be negligible from the structures observed in the absence of the metal. These results indicated that metal binding had no effect on the structural fold of the chelate.

3.3.4. Chelate-loop conformations. Structures belonging to different functional folds such as metallohydrolase, orthogonal bundle, TIM barrel, Rossmann fold, EF-hand and many others were analysed using *PyMOL* and *Discovery Studio*. Chelators with identical folds showed a preference for specific residues, but varied in length. For instance, all of the metallohydrolases in the data set were observed to possess a histidine-dependent coordination in which the neighbouring histidines were separated by two residues. Also, the chelate residues of this family were located in a turn/coil region, indicating an increased flexibility for metal coordination. Immunoglobin β-like sandwich folds are mostly constituted of only a single residue within the first shell along with 2–3 water molecules merely as an attempt at space filling. In the absence of water molecules the coordination space is occupied by residues neighbouring the metal ion. The chelators of the TIM-barrel

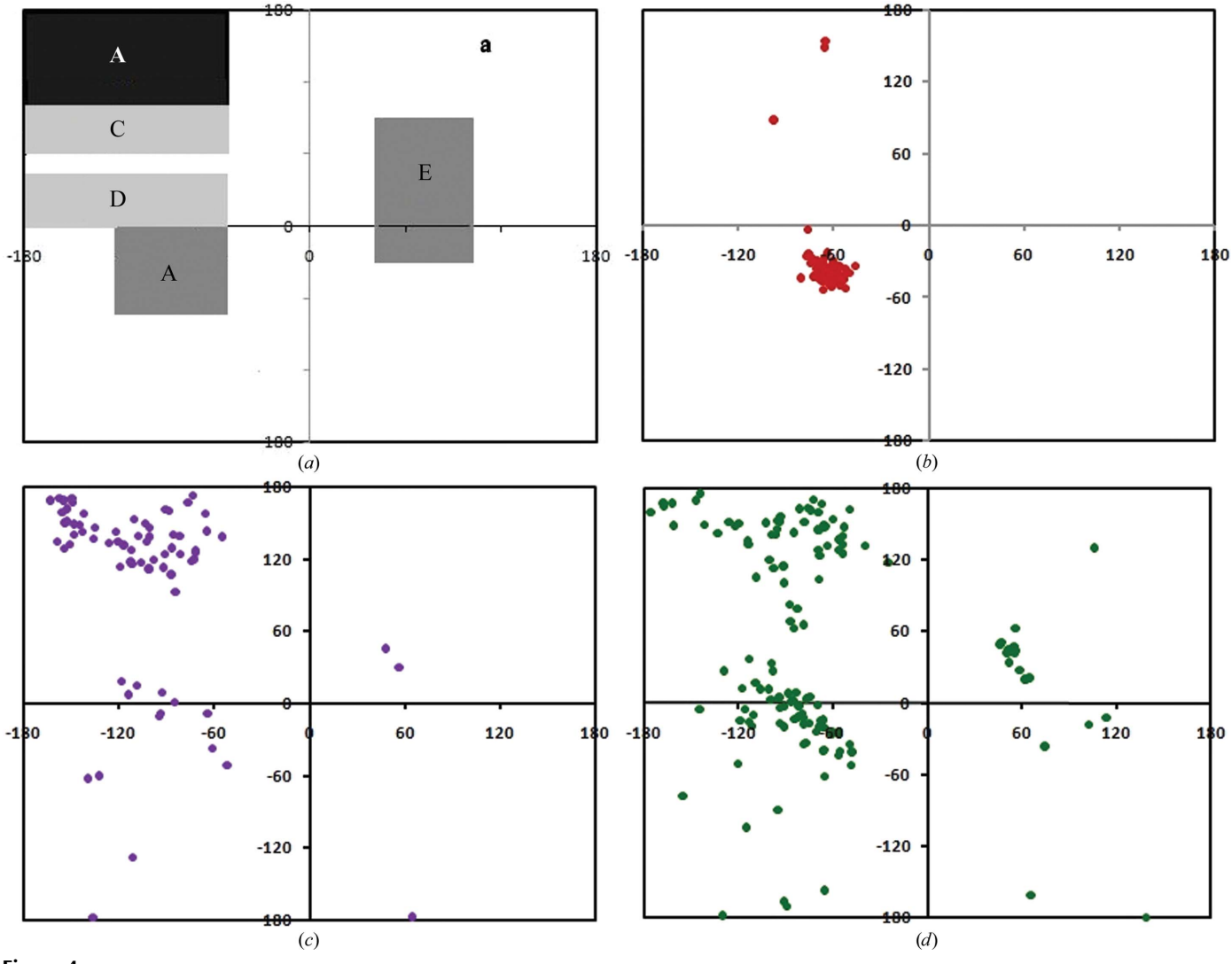

Figure 4
Ramachandran plot displaying the stability of coordinating residues. (*a*) Five regions of the Ramachandran plot: A, α-helices; B, β-sheets; C and D, additional allowed regions; E, left-handed α-helices. (*b*) Residues in α-helices in the core region. (*c*) Residues in β-sheets; these mostly occur in the additionally allowed regions. (*d*) Residues in coils and turns; these are mostly present in the allowed region of the Ramachandran plot.

domain are characterized by 2–4 donor atoms with an equal preference for histidine, cysteine, aspartate and glutamate. The structures with this fold showed no specificity for coordination number, which varied between one and five. The Rossmann fold was characterized by 2–3 donors, with histidine, aspartic acid and cysteine being highly preferred; however, the coordination number and chelate secondary structure varied drastically. The other folds exhibited no notable similarity and the chelation was limited to the choice of the amino acids proximal to the metal atom, which was preferably histidine, aspartic acid and glutamic acid. We observed that when structures shared the same number of donors, different chelates with the same fold and identical chelates with identical folds showed similar coordination patterns. Also, the chelation patterns of trypsin-like serine protease, although different, showed a perfect superposition of the chelate residues as shown in Fig. 5. Likewise, CHED chelates with the TIM-barrel fold also exhibited structural similarity. With respect to the number of donor atoms, the metal coordination of larger chelates such as CCCC and DDDKE tended to show a greater similarity than short chelates. As can be seen in Fig. 6(*a*), the γ-sulfur atoms of cysteine residues are oriented towards the metal such that the metal coordination holds the two peptides in an intact position. Similar arrangements are observed in other CCCC patterns from different functional classes. Fig. 6(*b*) shows the favourable structural arrangement of the glutamic acids such that they exhibit bidentate coordination through the δ-oxygen atoms; however, the difference in the orientation of the glutamates can be clearly observed. The secondary structures of the chelate residues also varied between identical chelates. For example, in CHED the preferences were THST and TTSH for identical folds, where T is turn, H is helix and S is sheet. However, chelates with a minimum of five donor atoms had an identical secondary-structural organization as observed in the EF-hand motif DDD[TK]E and the trypsin-like serine protease motif E[DN][VQ][ED]E despite the different donor patterns. These results indicated that cadmium-coordination geometry is only well established in chelates with a larger number of donor atoms.

3.4. Probabilities of amino-acid occurrence

The probability of a specific amino acid occurring within the coordination sphere was calculated as a factor of its preferred position succeeding and preceding the coordinating residues within the first shell. For this purpose, all structures with a single-residue interaction were omitted. Only chelate loops containing donors spanning a length of <10 residues were considered. The relative positions of these residues within the coordination distance were also analyzed and tabulated. The table was further used to predict probable pattern(s) of amino acids favourable for cadmium interaction based on their spatial arrangement. By observing the interacting residues, we classified the coordination patterns as single-amino-acid and multiple-amino-acid interactions for all of the structures in group *A* and group *B*.

Structures with single amino acids as donors within the coordination distance were mostly glutamic acid, aspartic acid, histidine and cysteine. Glutamic acid was often found in group *A* and histidine in group *B*. For coordination with multiple amino acids, a maximum of five and a minimum of two donors were observed in groups *A* and *B*. From each of the structures analyzed, the results showed that the predominant amino

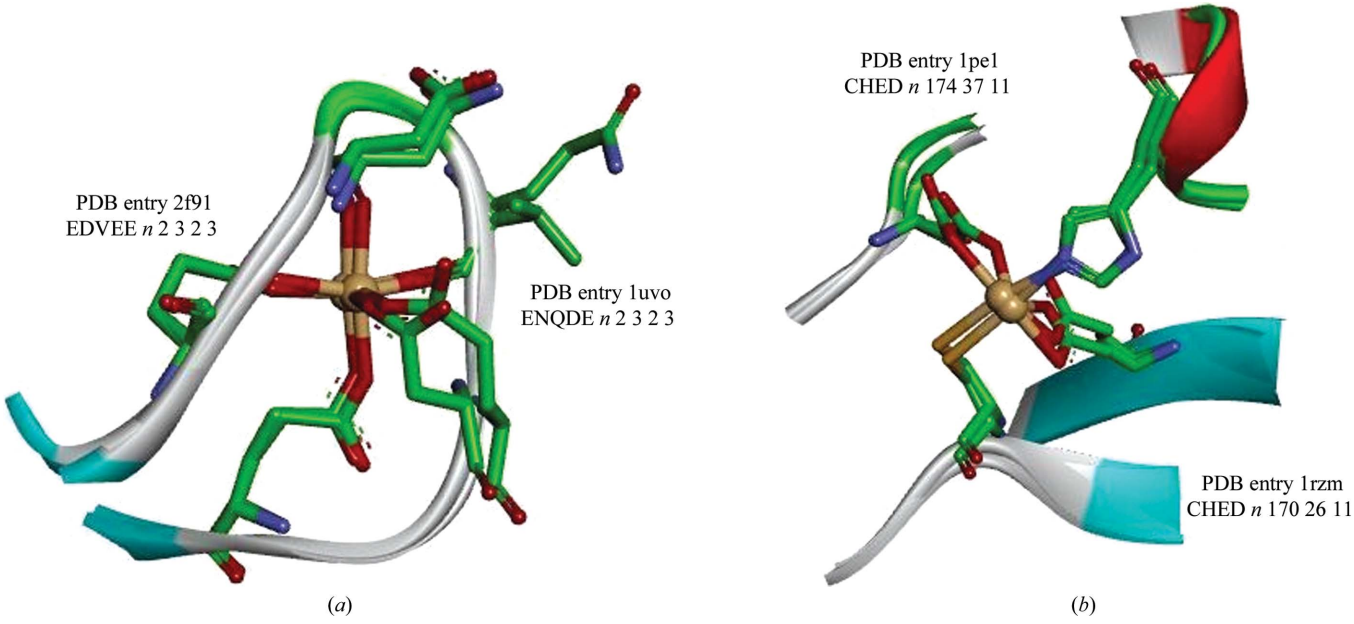

Figure 5
Chelate geometry of identical and different folds. (*a*) Superimposed structure of trypsin (PDB entry 2f91) and elastase (PDB entry 1uvo) chelates with the trypsin-like serine protease fold. Identity in the fold is noted despite the pattern variation. (*b*) Superimposed structures of 3-deoxy-D-*manno*-octulosonate 8-phosphate synthase (PDB entry 1pe1) and 3-deoxy-D-*arabino*-heptulosonate-7-phosphate synthase (PDB entry 1rzm), which are CHED chelates with a TIM-barrel fold.

Table 2
Putative amino-acid positions for cadmium coordination.

The table shows the probable combination of amino-acid occurrences within a 3 Å cutoff distance suitable for cadmium coordination spanning a length of ten residues. Amino acids that do not coordinate to cadmium are not shown in the table.

Position										
0	1	2	3	4	5	6	7	8	9	10
Asn		D	Q		D				E	
Asp	DEH	NDEHTK	E	DHKT	H	K	E		DEH	
Cys	C	CH	C			C	C	C	C	C
Gln		D	E		D E					
Glu		NDEYH	DEH	QE	QH	Q	DE			E
His	EH	DEH	EM	DEH	H	H	D		H	CH
Lys					E					
Pro			D							
Ser		S	D							
Thr		K					E			
Tyr		G		G						

acids (Glu, Asp, His and Cys) mostly occurred in combination with each other, as shown in Table 1(*b*), or as multiples of the same residue separated by a specific distance, as observed in the CCCC and CHED chelates. Analysis of the positional conservation of predominant chelate residues indicated that when an aspartic acid is at the zeroth position successive aspartates are largely observed at the second, fourth and ninth positions, glutamic acids at the seventh position and histidine at the first and fifth positions. After a glutamic acid, successive glutamates and histidines have an equal probability of occurring at the third position and aspartates at the second position. Histidine succeeding a histidine shows a high preference for the second position. Aspartates and glutamates occur at the second and third positions, respectively, from histidine. Strikingly, cysteines do not combine with any other amino acids within a ten-residue length and prefer to co-occur with cysteine with a maximum of five cysteine residues and a minimum of two consecutive cysteine residues within the cutoff distance. Residue preferences for other amino acids are also noted but with low confidence. Thus, by combining the positional patterns of the dominating residues we arrived at motifs such as D-H-D-*X*-D-H-*X*-E-*X*-D, E-*X*-D-[EH], H-D-H-E and C-C. Sub-patterns can be derived from these observed coordinating patterns and can be used to test for chelating ability. Table 2 displays the probable amino acids and their positions of occurrence with respect to other

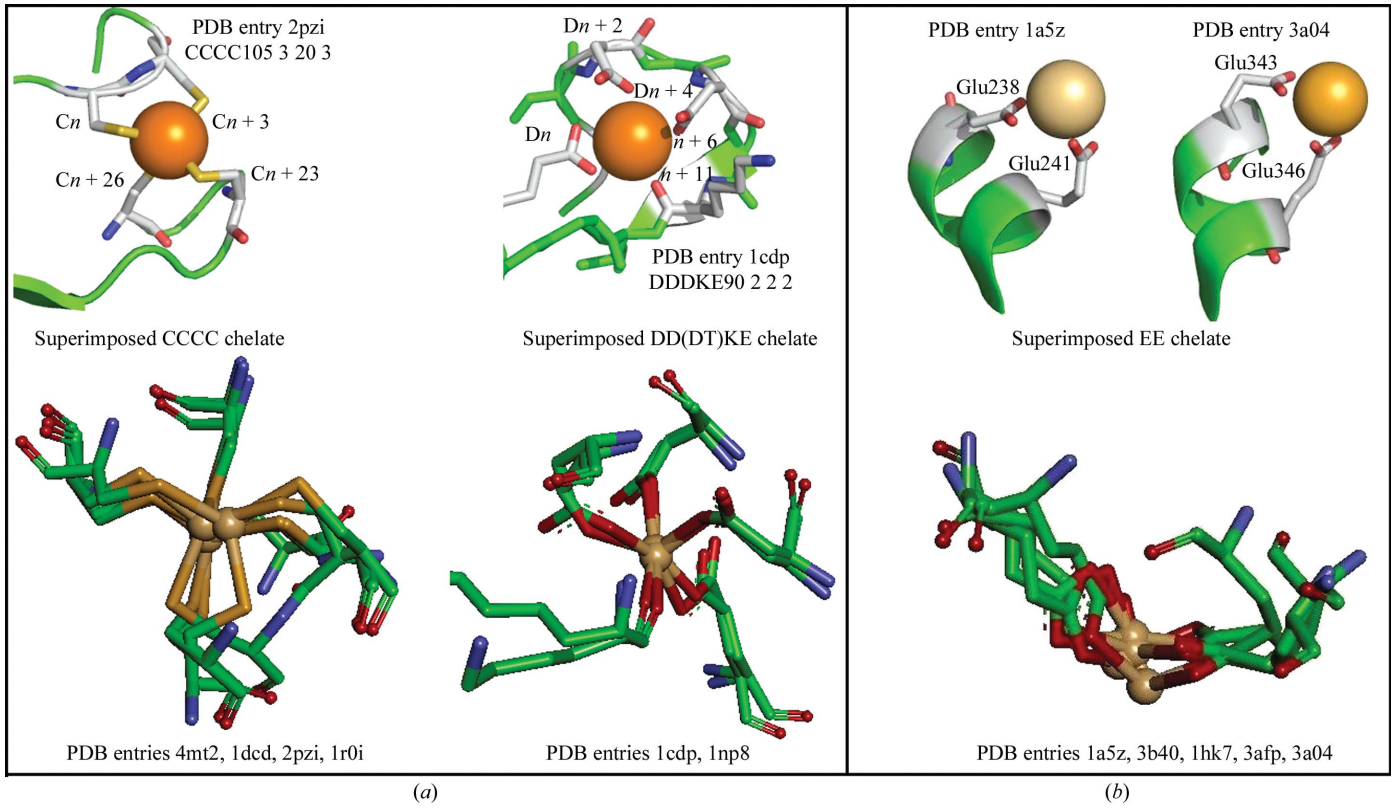

Figure 6
Cadmium coordination of large and small chelates. (*a*) Structures of protein kinase G (PDB entry 2pzi) showing CCCC chelate coordination and of parvalbumin (PDB entry 1cdp) showing DD[DT]KE chelate coordination. The superimposed structures of CCCC chelates are co-represented by metallothionine (PDB entry 4mt2), desulforedoxin (PDB entry 1dcd) and rubredoxin (PDB entry 1r0i). The superimposed structure of the DD[DT]KE chelate is co-represented by calpain. (*b*) EE chelate loops from different folds with different residue positioning represented by lactate dehydrogenase (PDB entry 1a5z) and tryptophanyl-tRNA synthetase (PDB entry 3a04). Superimposed EE chelates are co-represented by the structures of heat-shock protein 90 (PDB entry 1hk7), dipeptidase (PDB entry 3b40) and single-stranded binding protein (PDB entry 3afp). It can be clearly noted that all of the glutamic acids are oriented in the same fashion and are present at the turns of the helices. Most of the glutamates that are closely present are usually at a distance of $n + 3$ within a helix or located in different helices. However, the superimposed structures of the EE chelate clearly indicates the structural deviation.

Table 3
Probabilities of amino-acid occurrence.

The probabilities of residues at a specific position following an amino acid are displayed. Residues with high probabilities are highlighted in bold.

	Position										
First residue	1	2	3	4	5	6	7	8	9	10	Succeeding residues
Asn		0.007			0.007						Asp
			0.007								Gln
									0.007		Glu
Asp		0.007									Asn
	0.007	**0.067**		**0.037**					0.015		Asp
	0.007	0.007	0.007				0.022		0.007		Glu
	0.015	0.007		0.007	0.015						His
		0.007		0.007		0.007					Lys
		0.007		0.007							Thr
Cys	0.015	0.015	**0.052**			0.007	0.007	0.007	0.007	0.007	Cys
		0.007									His
Gln		0.007			0.007						Asp
			0.007		0.007						Glu
Glu		0.007									Asn
		0.022	0.007						0.007		Asp
				0.007		0.007					Gln
		0.007	**0.045**	0.007							Glu
		0.015									Gly
			0.037		0.015						His
His		0.022		0.007			0.007				Asp
										0.007	Cys
	0.015	0.007	0.015	0.007							Glu
	0.007	**0.075**		**0.022**	0.007	0.007			0.007	0.007	His
			0.007								Met
Lys					0.015						Glu
Pro			0.007								Asp
Ser			0.007								Asp
		0.007									Ser
Thr							0.007				Glu
		0.007									Lys
Tyr		0.007		0.007							Gly

residues. The overall probability and position-specific probabilities for all amino acids interacting with cadmium are given in Table 3. As discussed, only the commonest donor pairs shows the highest probabilities, indicating their greater preference for cadmium coordination.

3.5. Multiple sequence alignment of cadmium-binding proteins

Metal-binding patterns were predicted from conserved regions of the cadmium-binding proteins as identified from multiple sequence alignment. Despite belonging to the same functional group, the sequences showed a larger degree of variation. Extensive mutations were observed even among intra-genus sequences. In order to obtain meaningful conservation, we culled the data set to remove the most distinct sequences. About four conserved blocks were identified from the block database. The block positions along the entire sequence are given as ––AA–––––BBB–––––CCCCC–DDD––, where ABCD are the block names and the number of characters indicates the length of the conserved segment in each block. Hydrophobic residues were strongly conserved, with alanine, leucine, isoleucine and valine dominating, whereas the acidic amino acids aspartate and glutamate were the most conserved hydrophilic resiudes. Glycines were widely present in all cadmium-binding sequences and were located in the proximity of aspartates and glutamates or embedded within hydrophobic residues. The occurrence of glycines around the coordinating residues indicates their flexibility in folding, making the residues well exposed for cadmium coordination. Sequence patterns exhibiting such flexibility include DDCEGE and GDSDEG. Histidines and cysteines, which were observed to largely coordinate to the metal, were sparingly noted in the cadmium-binding protein sequences. The sequence diversity and the negligible occurrence of key residues such as histidine and cysteine suggests that the mechanism of cadmium coordination differs between cadmium-binding proteins and cadmium-bound structures. However, short-length motifs such as Y-X-G-X-G, Q-X_9-E, E-X_2-E-X_2-E and T-X_6-E-X_2-E (where X denotes any amino acid) were noted which resembled the results of structural analysis as given in Table 2. Fig. 7 shows the conserved regions of cadmium-binding proteins, with the patterns highlighted in rectangular boxes. Structures partially showing such chelating patterns included 1m8r, 1uvo, 3a02, 3a04, 3afp, 3b40, 2f91, 1a5z, 1np8 *etc*. Only the EE chelates represented by the motif E-X_2-E-X_2-E were noted to be well established between the cadmium-bound structures and cadmium-binding sequences.

Most of these sequence motifs were found to be located within the C and D blocks. Although not noted for all of the sequences, these patterns are suggested to have binding potential for cadmium ions. Among the four conserved blocks the pattern-prediction space was limited to blocks A and C, as glutamates expected to coordinate to the metal did not show preferable positional conservation. Fig. 8 shows blocks A and C with the strongly conserved aspartic acid and weakly conserved glutamic acid. Since the C-terminal end of the protein sequences was more similar to the motifs obtained from the structures, cadmium-binding pattern(s) were preferably predicted from the C block. We found that the residues of block C were more ideal owing to their high conservation compared with other blocks. Also from the various patterns predicted from all of the blocks we found that the block C pattern gave the best results. The pattern [FVIL][VIL][TS][VIF]A[SMN][**CG**]**G**[AG]**DN**[LIV]G was able to match 106 cadmium-binding proteins against UniProtTrEMBL database sequences using *Scanprosite*. The strong conservation of aspartates and asparagines and the association

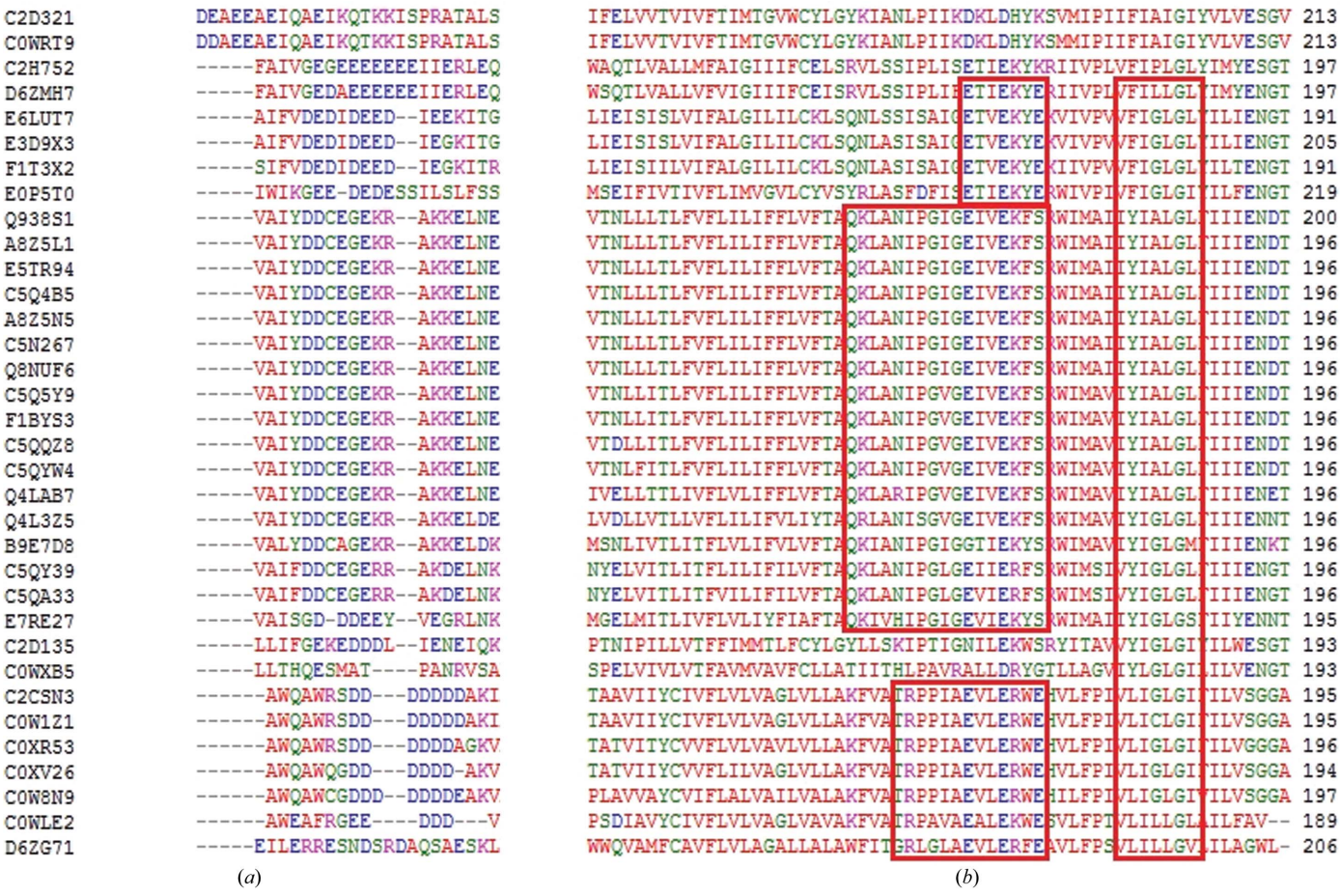

Figure 7
Conserved patterns observed in cadmium-binding proteins. (*a*) Block rich in acidic amino acids. Variations among the sequences are also observed. (*b*) The C-terminal conserved regions. The patterns resembling the putative cadmium-binding motifs are highlighted in rectangular boxes

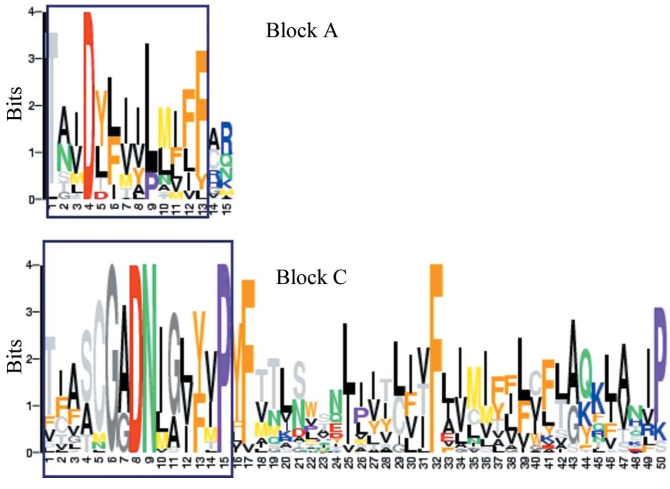

Figure 8
Conserved blocks predicted from cadmium-binding proteins. The numbers on the *x* axis represent the columns of the alignment; the characters are the single-letter codes of the amino acids for the corresponding column. The size of the letter indicates the extent of conservation of the residue. The blue rectangular box indicates the region that was chosen for pattern writing.

of cysteine and glycine noted in the pattern indicate the higher probability of these residues coordinating to the metal. The taxonomic distribution of the pattern was observed to represent different species of *Neisseria*, *Staphylococcus*, *Lactobacillus* and *Macrococcus*. The patterns predicted from the glutamate-rich region returned nonredundant results. This indicates that aspartic acids play a more important role than glutamic acids in cadmium-binding proteins. From our sequence analysis, we suggest that the short-length motifs can serve as binding motifs for cadmium coordination and the pattern representing the cadmium-binding/resistance protein family. The wide dispersion of the key binding residues within the sequence indicates the need for model building in order to understand the mechanism of cadmium binding by these proteins and to identify the association of these structural motifs with cadmium metal.

3.6. Norms for cadmium chelate design

The cadmium-bound structures analysed in this study include proteins with varying functionality. The mechanism of cadmium binding by these proteins may not be the native functionality of the structures. However, analysis of cadmium

coordination in these structures proved to be useful in predicting the residues that potentially prefer coordinating to the metal. In contrast, the set of cadmium-binding proteins considered in the analysis are characterized as being primarily involved in cadmium binding, although the structures of these proteins were not elucidated. Comparison of the coordinating residues with the conserved regions of cadmium-binding proteins further validates the results by allowing elimination of true negatives and false positives predicted to coordinate with the metal.

The cadmium coordination predicted from several cadmium-bound proteins provides a clear insight into the choice for the design of effective cadmium chelates. The comparison of the cadmium-bound structures and cadmium-binding sequences indicated a preference for specific residues such as glutamic acid, aspartic acid and cysteine residues. The absence of histidine conservation in cadmium-binding proteins indicates that histidine may not be a vital residue for consideration in chelator design. The cadmium coordination derived from the structures projects a random positioning of these residues, although in most cases a length span of 3–5 residues is observed between the coordinating residues. The probable positioning and preference of amino acids given in Table 2 suggest the combination of patterns suitable for chelator design. A coordination number of up to seven and a maximum of four donors specify the choice of the length to consider in cadmium chelator design. The larger the chelate length, the better the chelation, even between dissimilar patterns and folds. Furthermore, the secondary-structural features provide an insight into the architecture of a specific residue in the chelate. Some specific choices include the loop regions for EE chelates with a distance of $n + 3$ between the coordinating residues and DDD chelates with a specific distance of $n + 2$ between the residues. Similarly, fold-specific chelates and their geometry enable the design of chelates for a protein family. The concatenation of these specific features will provide useful information for the design of effective cadmium chelates.

4. Conclusion

The analysis of the specificity of cadmium binding resulted in contradictory results for cadmium-binding sequences and cadmium-bound structures. The sequence variations among cadmium-binding proteins even from similar species stands as a hurdle to their correlation with the results obtained from the structures. Yet, as observed from the structures, cadmium coordination requires a minimum of a single aspartic acid, glutamic acid, histidine or cysteine residue within the coordination sphere irrespective of chelate length, position and geometry. The sparse occurrence of histidine and cysteine in cadmium-binding proteins further reduces the chelators to aspartic acid and glutamic acid. The coordination patterns of similar and dissimilar folds were observed to be identical when the length of the chelates and the number of donor atoms increased. In chelates with one to three donor atoms, the loops, the fold and the geometry of chelates and residue positioning varied widely and corroborate the finding that cadmium coordination is random in shorter chelates. Despite these variations, we predicted short-length motifs in the sequences that resembled the structural patterns and thus these patterns are believed to have a higher probability of coordinating to cadmium ions. We also present a cadmium-binding signature that represents cadmium-binding proteins from a wide range of species with a conserved aspartic acid and aspargine which are suggested to equally contribute to cadmium coordination along with the other potential residues that have been identified. These findings should be useful in understanding cadmium coordination and in the design of chelators for therapeutic or bioremediation strategies.

The authors thank the management of VIT University for providing the facilities and encouragement to carry out this work. The authors also thank the editor and anonymous reviewers for their valuable comments and suggestions that helped to strengthen the quality of the work.

References

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L.-S. L. (2004). *Nucleic Acids Res.* **32**, D115–D119.
Archibald, S. J. (2011). *Inorg. Chem.* **107**, 274–296.
Banerjee, S. R., Maresca, K. P., Francesconi, L., Valliant, J., Babich, J. W. & Zubieta, J. (2005). *Nucl. Med. Biol.* **32**, 1–20.
Barbier, O., Jacquillet, G., Tauc, M., Cougnon, M. & Poujeol, P. (2005). *Nephron Physiol.* **99**, 105–110.
Benbrahim-Tallaa, L., Tokar, E. J., Diwan, B. A., Dill, A. L., Coppin, J. F. & Waalkes, M. P. (2009). *Environ. Health Perspect.* **117**, 1847–1852.
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.
Bork, P. & Koonin, E. V. (1996). *Curr. Opin. Struct. Biol.* **6**, 366–376.
Briner, W. (2010). *Int. J. Environ. Res. Public Health*, **7**, 4278–4280.
Castro, E. de, Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A. & Hulo, N. (2006). *Nucleic Acids Res.* **34**, W362–W365.
Chowdhury, B. A. & Chandra, R. K. (1987). *Prog. Food Nutr. Sci.* **11**, 55–113.
Csaba, G., Birzele, F. & Zimmer, R. (2009). *BMC Struct. Biol.* **9**, 23.
Deane, C. M. & Blundell, T. L. (1999). *Perspectives in Structural Biology*, edited by M. Vijayan, N. Yathindra & A. S. Kolaskar, pp. 197–207. Hyderabad: Indian Academy of Sciences.
DeLano, W. L. (2002). *The PyMOL Molecular Graphics System*. http://www.pymol.org.
Duffus, J. H. (2002). *Pure Appl. Chem.* **74**, 793–807.
Emmert, D. B., Stoehr, P. J., Stoesser, G. & Cameron, G. N. (1994). *Nucleic Acids Res.* **22**, 3445–3449.
Flora, S. J. S. & Pachauri, V. (2010). *Int. J. Environ. Res. Public Health*, **7**, 2745–2788
Giri, A. V., Anishetty, S. & Gautam, P. (2004). *BMC Bioinformatics*, **5**, 127.
Godt, J., Scheidig, F., Grosse-Siestrup, C., Esche, V., Brandenburg, P., Reich, A. & Groneberg, D. A. (2006). *J. Occup. Med. Toxicol.* **1**, 22.
Golovine, K., Makhov, P., Uzzo, R. G., Kutikov, A., Kaplan, D. J., Fox, E. & Kolenko, V. M. (2010). *Mol. Cancer*, **9**, 183.
Goyer, R. A. (1995). *Am. J. Clin. Nutr.* **61**, 646S–650S.
Guex, N. & Peitsch, M. C. (1997). *Electrophoresis*, **18**, 2714–2723.

Harding, M. M. (2004). *Acta Cryst.* D**60**, 849–859.
Henikoff, J. G., Henikoff, S. & Pietrokovski, S. (1999). *Nucleic Acids Res.* **27**, 226–228.
Järup, L. (2003). *Br. Med. Bull.* **68**, 167–182.
Järup, L. & Akesson, A. (2009). *Toxicol. Appl. Pharmacol.* **238**, 201–208.
Jones, M. M. (1994). *Coordination Chemistry*, edited by G. B. Kauffman, pp. 427–438. Washington DC: American Chemical Society.
Kasampalidis, I. N., Pitas, I. & Lyroudia, K. (2007). *Proteins*, **68**, 123–130.
Kuntal, B. K., Aparoy, P. & Reddanna, P. (2010). *Protein Pept. Lett.* **17**, 765–773.
Levander, O. A. (1978). *Environ. Health Perspect.* **25**, 77–80.
Ma, M. T. (2010). Thesis. School of Chemistry, The University of Melbourne.
Nath, R., Prasad, R., Palinal, V. K. & Chopra, R. K. (1984). *Prog. Food Nutr. Sci.* **8**, 109–163.
O'Donovan, C., Martin, M. J., Gattiker, A., Gasteiger, E., Bairoch, A. & Apweiler, R. (2002). *Bioinformatics*, **3**, 275–284.
Soghoian, S. & Sinert, R. H. (2009). *Heavy Metal Toxicity*. http://emedicine.medscape.com/article/814960-overview.
Thivierge, B. & Frey, R. (2006). *Gale Encyclopedia of Medicine*, 3rd ed. Michigan: Gale Group.
Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). *Nucleic Acids Res.* **22**, 4673–4680.
Tsuchiya, K. (1977). *Sangyo Igaku*, **19**, 471–478.
Vulić, M., Lenski, R. E. & Radman, M. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 7348–7351.
Wang, G. & Dunbrack, R. L. (2003). *Bioinformatics*, **19**, 1589–1591.
Zheng, H., Chruszcz, M., Lasota, P., Lebioda, L. & Minor, W. (2008). *J. Inorg. Biochem.* **102**, 1765–1776.