Metal-binding proteins play important roles in structural stability and complex formation (Wintz et al., 2003; Cox and McLendon, 2000; Michel and Berg, 2002; de la Calle Guntinas et al., 2002; Yang et al., 2002), gene expression regulation and alteration (Wintz et al., 2003; Jensen et al., 2005; Wu et al., 2005; Hantke, 2001; Bouton and Pevsner, 2000), DNA processing (Feng et al., 2004), signaling processes and cellular events (Carafoli, 2002), transport (Hantke, 2001; Harris, 2000; O'Halloran and Culotta, 2000), metabolism control (Wintz et al., 2003; de la Calle Guntinas et al., 2002; Harris, 2000; Vallee and Auld, 1990), metal homeostasis (Cobbett and Goldsbrough, 2002; Papoyan and Kochian, 2004), antibody recognition (Zhou et al., 2005), and other events such as cellular respiration, muscle movement, and antioxidant defense (Lieu et al., 2001). Approximately 1/3 of structurally-determined proteins are metal-bound (Barondeau and Getzoff, 2004), and large percentages of metals present in human body are bound to proteins (de la Calle Guntinas et al., 2002; Sandier et al., 1997). Identification of metal-binding proteins and knowledge of metal-protein interactions is important for elucidating the function and functional mechanism of proteins and biological processes (Lin et al., 2006).

Several computational methods have been explored for identifying and characterizing metal-binding proteins. In many metalloproteins, the metal ions tightly bind to the proteins and their metal-bound structures could be accurately determined by x-ray crystallography. Thus structural information has been used for predicting metal-binding sites based on the detection of principal liganding residues and metal-ligand complex architectures (Gregory et al., 1993; Andreini et al., 2006), the use of common local structural parameters (Gregory et al., 1993), combination of sequence and structural profiles (Sodhi et al., 2004), analysis of bond strength contributions (Nayal and Di Cera, 1994), and the computation of force fields (Schymkowitz et al., 2005; Khalili et al., 2004). But for those proteins with loosely or temporarily bound metals, such as enzymes that use metal ions as cofactors, the specific metal binding sites are often poorly characterized or unknown (Jensen et al., 2005).
Moreover, combinatorial use of multiple structural, sequence alignments and annotation methods has been found to be highly useful for improving prediction accuracy of metal-binding proteins.

Metal-binding proteins are diverse in sequence, structure, and function. Nonetheless, metal cations generally bind to centers of high hydrophilicity and reduce the enthalpy of a system upon binding, and metal ions bind to a shell of polar hydrophilic residues surrounded by a shell of non-polar residues. The binding sites of some metal-ligand complexes have specific structural architectures. To some extent, these metal binding features are similar to those of other molecule-binding features of proteins such as RNA-binding proteins, DNA-binding proteins and transporters that are also diverse in sequence, structure and function whose binding capability are mediated by certain structural and physiochemical characteristics.

The use of metal compounds in medicine can be traced back to 5000 years ago, when copper was used by Egyptians to sterilize water. Over the past decades, extensive knowledge of the coordination and redox properties of metal ions has facilitated the development of metal-based drugs, especially after the discovery of the anti-tumor activity of cisplatin. Some metal complexes have been successfully used therapeutically, such as platinum, gallium and arsenic complexes in cancer therapy, gold complexes as anti-arthritis and asthma agents and bismuth complexes in anti-ulcer treatment. Enormous efforts have been made on the biocoordination chemistry of metal-based agents. Structures of several clinically used metal-based drugs are shown in figure 2.1.
Fig. 2.1: Structures of metal-based drugs, \textit{i.e.,} cisplatin, carboplatin, oxaplatin, myocrisin, NAMI-A, KP1019 and colloidal bismuth subcitrate (CBS). Note that myocrisin and CBS are polymeric complexes in solution.

Metal ion binding can stabilise a protein by cross-linking points in the protein which may lie distant from one another in the protein sequence. This is particularly useful for small domains which would not otherwise be stable. Calcium and zinc are the ions that are most commonly employed by proteins for this purpose (Petsko and Ringe, 2004). A metal binding site in a protein consists of one metal ion and all protein side chains and water molecules that participate in its first coordination sphere. The characteristic residue preferences and binding geometries (in terms of angles between residues) of different metals are well known (Glusker, 1999; Harding, 2001, 2004). Zinc ions playing a structural role are bound by sets of four cysteine and/or histidine residues with a tetrahedral geometry (example shown in Figure 2.2 a) (Dudev and Lim, 2003); calcium ions are typically bound by seven groups in a pentagonal
bipyramidal geometry comprising oxygen atoms from side chains, backbone carbonyl groups, and water molecules (example shown in Figure 2.2 b) (McPhalen et al., 1991).

The residue preferences of metals are largely determined by the polarisability of the metal ion. Less easily polarized ions such as Ca$^{2+}$ prefer to bind less easily polarized residue groups such as the carboxyl group in glutamate and aspartate. More easily polarized ions, such as Zn$^{2+}$, are able to bind to easily polarised residue groups, such as the sulphydryl group in cysteines (Pearson, 1963; Glusker, 1999).
The calcium binding sites employed a variety of motifs to bind the metal; only one of them used a standard EF-hand motif (calmodulin, PDB entry 1g4y). Most of these calcium binding sites served structural roles. The only exception was calmodulin, where the calcium is regulatory; calcium binding induces a structural change in calmodulin which enables it to bind to a variety of target proteins, altering their activity. There were four zinc binding site templates in the dataset that each consisted of four cysteines (PDB entries 1e7l, 1k3x, 1m2k and 1n8k). There were two zinc-binding templates that each consisted of three cysteines and one histidine (PDB entries 1btk and 1r5y). The remaining zinc binding site (PDB entry 1jk3b) consisted of three histidines and one glutamate. These zinc binding sites all served structural roles (Torrance, 2008).

Metal cations generally bind to centers of high hydrophilicity and reduce the enthalpy of a system upon binding (Khalili et al., 2004; Frausto da Silva and Williams, 1991), and metal ions bind to a shell of polar hydrophilic residues surrounded by a shell of non-polar residues (Gregory et al., 1993). The binding sites of some metal-ligand complexes have specific structural architectures (Gregory et al., 1993). To some extent, these metal binding features are similar to those of other molecule-binding features of proteins such as RNA-binding proteins, DNA-binding proteins and transporters that are also diverse in sequence, structure and function whose binding capability are mediated by certain structural and physiochemical characteristics (Lin et al., 2006, Fierro-Monti and Mathews, 2000; Perez-Canadillas and Varani, 2001).

2.1 Prediction of Metal and non-metal datasets

Sampathrajan in 2009 studies on prediction of metalloproteins and conclude that proteome of every organism requires a significant share of metal ions or metal containing cofactors to carry out its physiological function. Metalloproteins are proteins capable of binding one or more metal ions or metal containing cofactors, which are required for biological function or for the regulation of their activities or for structural purposes (Passerini et al. 2007). In in vitro condition, metal ions are observed to interact with unfolded polypeptide and may create local structure that initiates and directs the polypeptide folding process (Wilson et al. 2004). Metal-
binding capabilities are encoded in the amino acidic sequences and these primary sequences are related to the protein three-dimensional structure. Through genomic projects various organism genomic sequences have been annotated somehow along with metalloproteins contained in them (Andreini et al. 2004). Identification of metal binding through experimental methods is difficult and expensive. The use of bioinformatics has been extensively used to predict metal binding from amino acid sequences. Predictions of metal binding proteins are useful in structural genomics, to select proper growth medium for over-expression studies and for the easy interpretation of electron density maps. However, the available prediction methods are either based on the knowledge of the apoprotein structure or they are restricted to few specific cases, like the metal binding of histidines/cysteines.

All the protein sequences were downloaded from the UniProt database (Wu et al. 2006) available at http://www.uniprot.org/. The downloaded sequences, annotated as metal containing, were grouped into eight subsets. Each of the subsets, containing one of the metal species viz., calcium, cobalt, copper, iron, magnesium, manganese, nickel and zinc was considered to be metal-containing while all other entries were considered to be metal-free. Redundant sequences were removed with the cd-hit program (Li and Godzik 2006) at the 50% level of percentage of identity, analogous by the UniRef 50 list available at the UniProt database. This resulted in eight data sets containing 186 calcium-containing proteins, 69 cobalt containing proteins, 215 copper-containing proteins, 315 iron-containing proteins, 961 Magnesium-containing proteins, 386 manganese-containing proteins, 74 nickel containing proteins and 1716 zinc-containing proteins. All proteins containing calcium, cobalt, copper, magnesium, manganese, nickel or zinc were then subtracted from the UniRef50 list, resulting in a collection of 1,640,922 non-metalloproteins (Sampathrajan, 2009).

2.2 Biochemical Studies of Metal Binding to Proteins

Gonick (Gonick, 2011) review lead-binding proteins and said Lead is known to displace physiologically relative metal ions, such as calcium and zinc, in proteins. Kirberger and Yang (Kirberger and Yang, 2008) reported that approximately 1/3 of the lead binding sites were identified as due to zinc or calcium ionic displacement, whereas two-thirds were opportunistic. Oxygen atoms from amino acids or water
represent the major ligand for lead, followed by sulfur and nitrogen. Sulfur acts as the ligand in the case of displacement of zinc by lead at the zinc-binding sites in delta-aminolevulinic acid dehydratase (ALAD). Studies of calmodulin, a calcium binding protein with a propensity for displacement of calcium by lead, showed an initial activation followed by inhibition in response to increasing concentrations of lead. The latter was thought to result from more pronounced conformational changes resulting from additional opportunistic binding. Hanas et al. (Hanas et al., 1999) explored whether the association of lead with chromatin might suggest that the deleterious effects may in part be mediated through alterations in gene function. They specifically examined whether lead altered DNA binding of cysteine-rich zinc finger proteins. It was found that inhibition of Cys2His2 zinc finger transcription factors by lead ions at concentrations near those known to have deleterious physiological effects was suggestive for a new molecular mechanism for lead toxicity. However, the changes were seen predominantly at lead concentrations varying from 100 to 400 ug/dL, above the industrial or pathophysiological range. Becker et al. (Becker et al., 2008) analyzed naturally occurring metal-binding proteins in rat liver and kidney, utilizing nondenaturing gel electrophoresis together with laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). The mass range of proteins separated was from 45 to 120 kDa. Lead was found only in small quantities in the 65 kDa protein band, whereas zinc and copper were scattered throughout the protein bands. If this technique could be extended to the study of lead-treated as well as native rats and the protein mass separated down to 5 kDa, confirmation (or lack thereof) of the results reported earlier by Fowler and associates (Oskarsson et al., 1982; Mistry et al., 1985; Goering and Fowler, 1984; Mahaffey and Fowler, 1977; Fowler and DuVal, 1991) might be anticipated.

2.3 Databases on Metalloproteins

2.3.1 MetLigDB

MetLigDB (http://silver.sejong.ac.kr/MetLigDB) is a publicly accessible web-based database on which the interactions between a variety of chelating groups and various central metal ions in the active site of metalloproteins can be explored in detail (Choi et al., 2011). Additional information can also be retrieved including protein and inhibitor names, the amino acid residues coordinated to the central metal ion, and the
binding affinity of the inhibitor for the target metalloprotein. Although many metalloproteins have been considered to be a promising target for drug discovery, it is difficult to discover a new inhibitor due to the difficulty in designing a proper chelating moiety to impair the catalytic activity of the central metal ion. Because both common and specific chelating groups can be identified for varying metal ions and the associated coordination environments, MetLigDB is expected to enable the users to get insight into designing the new inhibitors of metalloproteins for drug discovery.

2.3.2 MDB

The Metalloprotien Databese and Browser (MDB; http://metallo.scripps.edu) at The Scripps Research Institute is a web-accessible resource for metallo-protein research (Castagnetto et al., 2002). It offers the scientific community quantitative information on geometrical perameters of matal-binding sites in protein structures available from the protein Data Bank PDB. The MDB also offers analytical tools for the examination of trend or pattern indexed metal-Binding sites. A user can perform interactive searches, metal-sites structure visualization (via a java applet), and analysis of the quantitative data by accessing the MDB throught a web browser without requiring an external application or platform-dependent plugin. The MDB also has a non-interactive interface with which other websites and network-aware application can seamlessly incorporate data or statistical analysis result from metal-binding sites. The information contained in the MDB is periodically updated with automated algorithms that find and index metal sites from new protein structures released by the PDB.