University of Louisville
ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

8-2016

Structure-function analysis and characterization of metalloproteins.

Sen Yao
University of Louisville

Follow this and additional works at: https://ir.library.louisville.edu/etd

 Part of the Bioinformatics Commons

Recommended Citation

Yao, Sen, "Structure-function analysis and characterization of metalloproteins." (2016). *Electronic Theses and Dissertations.* Paper 2537.
https://doi.org/10.18297/etd/2537

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

STRUCTURE-FUNCTION ANALYSIS AND CHARACTERIZATION OF METALLOPROTEINS

By

Sen Yao
B.S., Guangzhou University of Chinese Medicine, China, 2008
M.S., University of Louisville, U.S., 2012

A Dissertation
Submitted to the Faculty of the
School of Interdisciplinary and Graduate Studies of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in
Interdisciplinary Studies: Specialization in Bioinformatics

School of Interdisciplinary and Graduate Studies
University of Louisville
Louisville, Kentucky

August 2016

© Copyright 2016 by Sen Yao

All Rights Reserved

STRUCTURE-FUNCTION ANALYSIS AND CHARACTERIZATION OF METALLOPROTEINS

By

Sen Yao
B.S., Guangzhou University of Chinese Medicine, China, 2008
M.S., University of Louisville, U.S., 2012

A Dissertation Approved On

July 20, 2016

by the following Dissertation Committee:

_____
Eric C. Rouchka, D.Sc. Advisor

_____
Hunter N.B. Moseley, Ph.D. Co-Advisor

_____
Jeffrey C. Petruska, Ph.D.

_____
Shesh N. Rai, Ph.D.

_____
Guy Brock, Ph.D.

DEDICATION

This dissertation is dedicated to

my parents, my husband, and all my family and friends,

for their endless love, support, and encouragement.

ACKNOWLEDGEMENTS

There are a number of people without whom this dissertation could not have been possible, and to whom I owe my deepest gratitude.

First and foremost, I would like to express my deepest gratitude to my advisor Dr. Hunter Moseley, for his excellent guidance and continuous support though my master's and doctoral journey. His broad knowledge and great passion towards research has guided me to explore various aspects of bioinformatics, ranging from computational and statistical skills, dissertation research, and paper writing, to presentation and communication skills. His perpetual energy and research integrity have motivated me to always work diligently. It is said that choosing a mentor is one of the most important decisions one need to make in the graduate years, and I feel so blessed to have such a wonderful mentor. Working with him has always been a delightful experience. Without his guidance and persistent help, this dissertation would not have been possible.

I am also especially grateful to my co-advisor, Dr. Eric Rouchka, for kindly taking me under his wing and for his continuous support, inspirational guidance, and optimistic encouragement throughout my research career at University of Louisville. He is also the director and one of the main forces in launching this bioinformatics Ph.D. program, which provided me with the great opportunity of training in this exciting new field. His has guided and helped me in so many different aspects, ranging from inspirational research idea, and paper/dissertation writing, to even technical problems

with the fellowship. Dr. Rouchka has played a vital role for me to survive through the program, and to develop my skills and expertise in becoming a qualified bioinformatician.

I want to thank Drs. Guy Brock, Shesh Rai, and Jeff Petruska for being my committee members for their unconditional assistance and insightful discussions, involving research ideas and professional and scientific suggestions. I took the statistical computing course with Dr. Brock, and learned from him the most valuable tool, R, which I have been using extensively ever since. He also helped me with developing the statistical algorithms in my research in a more rigorous and professional manner. Dr. Rai shared his expertise in statistics with me and made several valuable suggestions to my research. Discussion with him is always a pleasure and learning experience on statistical rigorousness. Dr. Petruska provided me with many valuable advices in the biological aspect of my research. He inspired me to develop a broader and deeper understanding on the biological problem of my research, and also helped me structure and deliver the research proposal and dissertation in a much better and stronger fashion.

I am very grateful to Dr. Robert Flight for his support, suggestion, and encouragement. He has been actively participating in this project. His insightful ideas, suggestions, and experience were crucial to perform my project successfully and eventually developed into this dissertation. He was always supportive and available to discuss any problems I encountered in my work.

A special thank you goes to Dr. Mark Mashuta, for the informative discussion on the crystallographic symmetry. It was at a critical time of my research, and only with the help from him, I was able to finish this dissertation in time.

I would like to thank my former and current lab members: Andrey Smelter, Joshua Mitchell, Xi Chen, Dr. Abdallah Eteleeb, Ernur Saka, Mohammed Sayed, and many others. They are always warming, supportive, and encouraging, which makes the labs interesting and enjoyable.

Last but not least, I would like to thank my parents, for the great education opportunity they provided me. Without their love and support, I could not have come this far both physically and professionally, and could not have achieved what I achieve today. I would like to thank my husband, who has been the sweetest surprise I got in my graduate years. He has been nothing but supportive to my research career. I would also like to thank all my family and friends, for their support, encouragement, and friendship. They made my life in the graduate years much more enjoyable.

ABSTRACT

STRUCTURE-FUNCTION ANALYSIS AND CHARACTERIZATION OF METALLOPROTEINS

Sen Yao

July 20, 2016

Metalloproteins are proteins that can bind at least one metal ion as a cofactor. They utilize metal ions for a variety of biological purposes, and are essential for all domains of life. Due to the ubiquity of metalloprotein's involvement across these processes across all domains of life, how proteins coordinate metal ions for different biochemical functions is of great relevance to understanding the implementation of these biological processes. One of the most important aspects of metal binding is its coordination geometry (CG), which often implies functional activities.

Most of the current studies are based on the assumption of previously reported CG models founded mainly in a non-biological chemical context. While this general procedure provides us with great measures on the closest CG model a metal site adopts, it also biases and limits the binding ligand selection and coordination results to the canonical CG models examined. Thus, if a CG model exists that has never be reported previously or is not accounted for in a study, instances from the CG would either be

misclassified into an expected model and cause a high in-class variation or considered as outliers.

To solve this problem, we have developed our analysis, where the less-biased low-variation measure, bond-length, was used determine the binding ligands and the higher-variation measure, angle, was used to cluster the metal shells into canonical or novel CGs with functional associations. This methodology is model-free, and allows us to derive the CG models from the data itself. Thus, we can handle unknown CGs that may cause problems to the classification methods. This new methodology has enabled the discovery of several previously uncharacterized CGs for zinc and other top abundant metalloproteins. By recognizing these novel/aberrant CGs in our clustering analyses, high correlations were achieved between structural and functional descriptions of metal ion coordination.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Metalloproteins

Metalloproteins are proteins that can bind at least one metal ion as a cofactor. They play various structural, functional, and signal transductional roles in proteins, and are essential for all domains of life [1]. Many proteins depend on metals to help hold their structures together [2, 3], while others require metals to directly participate in the biochemical reactions they catalyze [4, 5]. However, most transition metals, while essential in their bound state, are highly toxic in their free ionic form, which requires tight regulations [6]. Therefore, there are many proteins involved in the sensing, transporting, and storing of metal ions in biological systems in order to maintain their appropriate forms at levels [7]. It is estimated that roughly 30-40% of whole proteomes across the biosphere are metalloproteins [8, 9]. Malfunction of metalloproteins, like misfolding or lack of incorporation of the proper metal in its proper form, may result in many human diseases, such as neurodegeneration disease [10-12], diabetes [13, 14], and cancer [15, 16]. Individual metalloproteins have been shown to be the key factors of many diseases, and therefore are also important targets for drug designs [17, 18], providing yet another reason for their systematic study. To our knowledge, studies of metalloprotein's involvement in certain types of human disease on a system level have

not yet been performed and reported, which will become an important extension of the results of this dissertation.

Metal ions generally bind to proteins via coordination by electronegative atoms from the protein, such as nitrogen, oxygen, and sulfur. One of the most important aspects of metal binding is its coordination geometry (CG), which often relates to the metalloprotein's functional activities [9, 19]. In inorganic chemistry, a metal ion can bind to its ligands matching to a canonical CG almost perfectly. And in this context, metal ions are observed and verified to adopt only a few of different canonical CGs according to their physiochemical properties [20, 21]. While in biology, the chemical environment around metal ions can be much more diverse, leading to additional novel or aberrant CGs [22].

1.2 Motivation

With the help of modern analytical technologies, more and more sequential and structural data are available, with the ultimate goal in understanding their biological function in vivo. There has be significant efforts in developing computational tools to better utilize these data to facilitate this purpose. As for metalloproteins, several methods have been developed to analyze the coordination environment and functional implication of the metal binding sites [23-26]. However, most of the studies are based on the assumption of previously reported CG models founded mainly in a non-biological chemical context. They also all tend to follow the same general procedure. First, the metal coordination shells are acquired using a simple distance cutoff. Second, the ligands in the shell are compared to the known ideal CGs to compute a score. Finally, the metal

site is classified to the CG that gives the highest score. Studies mainly differ in the selected canonical CG models and the way a score is computed. While this general procedure provides us with great measure on the closest CG model a metal site adopts, it also limits the results to the examined known CG models. Thus, if a CG model exists that has never be reported previously or is not accounted for in a study, instances from the new CG would either be misclassified into an expected model and cause high in-class variance or considered as outliers. The standard of classifying outlier is also arbitrary, lacking a rigid validation of the results.

In several studies [8, 22, 24] including our own initial analysis with zinc metalloproteins, we observed a similar trend of either a significant number of outliers or abnormally high variance in classified CGs (more details can be found in Section 3.2). We attempted to directly handle and account for the high variability in zinc CG. As we explored the factors that could cause such high variance in ligand-zinc-ligand angles, we detected the existence of significant numbers of compressed angles due to coordination by bidentate ligands (i.e. two binding atoms are from the same amino acid residue), which prompted us to develop a new method for classifying zinc coordination geometries. Given this, we hypothesized that the high variability observed in zinc metalloproteins were due to the existence of a significant number of aberrant CGs, which are prevalent across all metalloproteins, and have distinct functional relationships.

1.3 Dissertation Contributions

In contrast to the angles, the bond-length showed very low variance in classified canonical CGs, which is consistent in several studies[22, 27, 28]. These observations and

hypothesis lead us to design new methods in analyzing the structure of metalloproteins, where we use the highest quality and less biased measure, bond-length, to determine the binding ligands and then cluster the metal binding sites using the resulting angles. This methodology is model-free, and allows us to derive the CG models from the data itself and assign each CG into these derived CG models. Thus, we can handle unknown CGs that may cause problems to biased classification methods. These new analysis methods have enabled the discovery of several previously uncharacterized CGs for zinc and other top abundant metalloproteins.

1.4 Dissertation Outline

Chapter 2 reviews the relevant concepts of structural biology and metal coordination in both inorganic and organic context. It begins with basic concepts such as primary structure, secondary structure, tertiary structure, and quaternary structure. The chapter then explains the theory of metal binding in inorganic chemistry, and the application of metal binding in biology. It also reviews most popular databases and tools of protein and metalloprotein structure.

Chapter 3 reviews some recent achievements and limitations on metalloprotein structure study, suggesting ways to overcome the obstacles. It lays out general principles used in conducting this project.

Chapter 4 describes the analysis on the structure of Zn metalloproteins. It includes a brief review of current studies of zinc metalloprotein structure and their limitations. The chapter then introduces the design, rationale, and implementation of each step of our

methodology. It finally shows the results for evaluating the performance of the methods, ending with major conclusions.

Chapter 5 describes the analysis on the structure of the five most prevalent metalloproteins. It starts with a brief review of the five metalloproteins. It then introduces the design, rationale, and implementation of each steps of the methodology. The chapter emphasizes major modifications to the methodology made since Chapter 4, allowing their application to other metalloproteins with multiple coordination numbers. Improvements of the methodology are highlighted, validated, and demonstrated with corresponding superior results. The chapter ends with major conclusions interpreted from the results.

Chapter 6 is devoted to discussions and conclusions of the whole analysis, with potential pitfalls and possible solutions. Time and memory usage are analyzed. Plans for future steps in the application of this project to improve the functional understanding on metalloproteins are put forth.

CHAPTER 2

BIOLOGY AND BIOINFORMATICS BACKGROUND

2.1 The Structure of Proteins

Proteins are large biological macromolecules composed of one or more chains of amino acid residues. They are the fundamental components in cells, and carry out most of the essential functions required for all living organisms[29-31]. Proteins typically consist of a sequence(s) of amino acid residues that fold into one or more specific conformations to perform its biological activity. A fundamental principle of structural biology is that sequence dictates structure, which enables function. The diversity of a protein's sequence and structure allows the unique arrangements of chemical groups at specific locations, which facilitate the occurrence of particular enzymatic reactions [32].

In order to fully understand protein function on a molecular level, the three-dimensional structures of a protein are often required. X-ray crystallography, nuclear magnetic resonance (NMR), and electron microscopy are the most common analytical methods providing experimental data for deriving protein structure [33]. Decades of structural studies have determined tens of thousands of protein structures [34] organized into a couple thousand structural families [35-37], providing great examples for understanding the principles of protein structures and how they are utilized to achieve protein function. Though protein structures exhibit large diversities due to the wide range

of functions they perform, they follow a coherent set of principles that can be ordered into four levels of structural complexity: primary, secondary, tertiary, and quaternary (Figure 2.1).

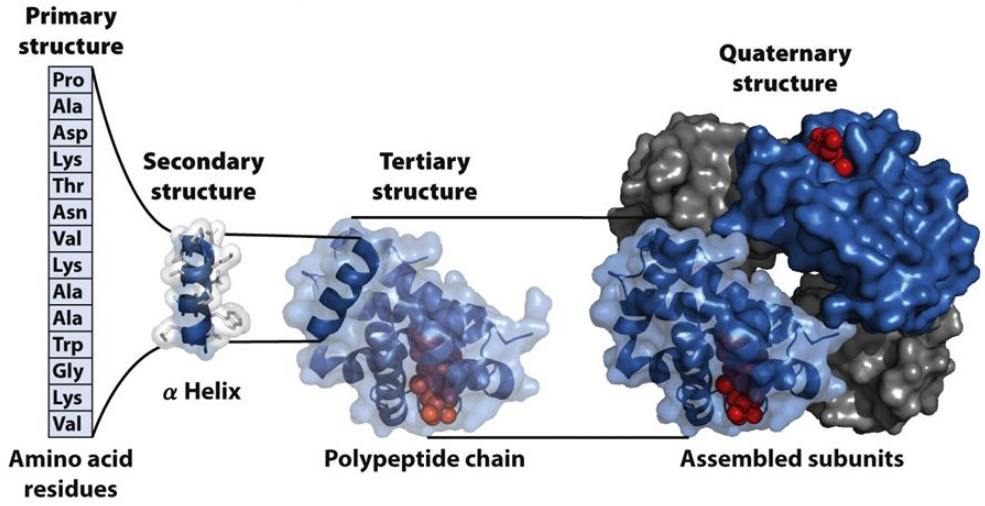


Figure 2.1 The four levels of protein structural complexity [38]. From *Lehninger Principles of Biochemistry,* 6th Edition, By David L. Nelson and Michael M. Cox, Copyright 2013 by W.H. Freeman and Company. Used with permission by the publisher.

2.1.1 Primary Structure

2.1.1.1 Amino Acid Residues

Amino acids are the building blocks in constructing a protein. The sequence of a distinct linear amino acid combination determines the ultimate three-dimensional structure of protein. Figure 2.2 shows the 20 standard amino acids. They have a similar structure containing an α-carboxyl group, an α-amino group, and a characteristic R group. They can be roughly grouped into three categories based on its R group chemical

property: polar, non-polar, and charged. Proline is special in its formation of a cyclic side chain structure, which hampers the flexibility of the amino acid residue and in turn the overall structure of the protein. Except for glycine, the α-carbon atoms of all amino acids are chiral. Chirality describes the geometric property of a molecule that is non-superposable on its mirror image[39, 40]. And the L-stereoisomer is preferred in most proteins.

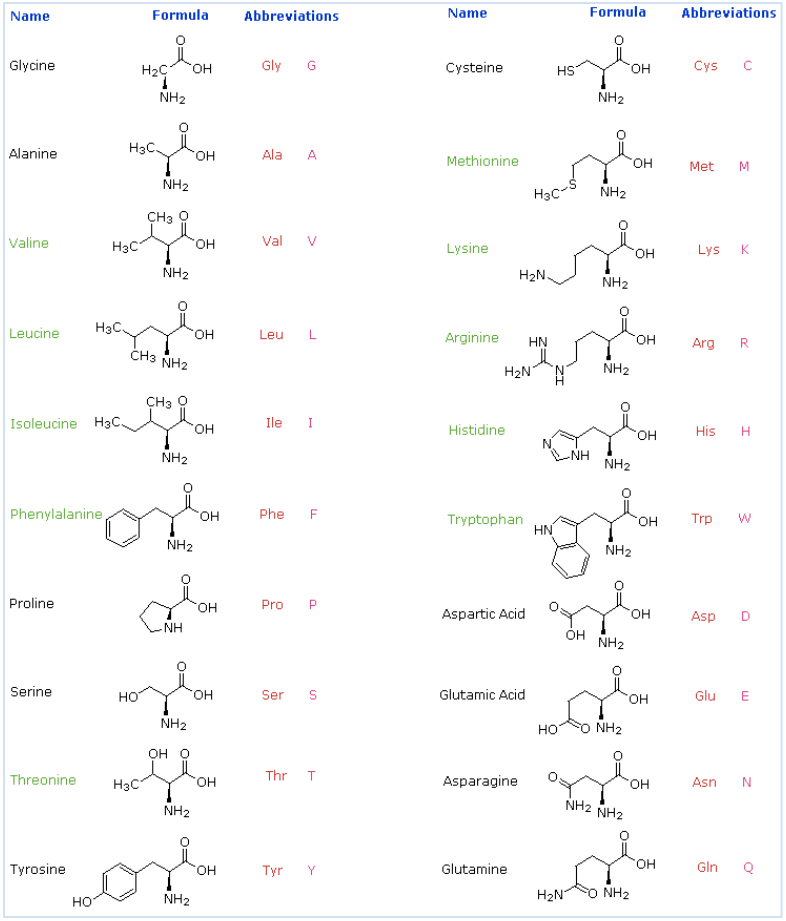


Figure 2.2 The 20 standard amino acid [41]. Used with permission.

2.1.1.2 Primary Structure

The primary structure of a protein is the linear sequence of amino acid residues known as a polypeptide chain. By ad hoc definition, proteins contain 50 or more amino acid residues. Two adjacent amino acid residues are linked together through forming a covalent peptide bond between the carboxyl group and amino group from one another, as shown in Figure 2.3 A. On the ends of a protein chain, there is an unbounded carboxyl group (C-terminal) and amino group (N-terminal), illustrated as 5-residue peptide in Figure 2.3 B. Determination of the protein sequence always proceeds from its N-terminal. The primary structure of a protein is translated from its corresponding messenger ribonucleic acid (mRNA), which is transcribed from its gene. It essentially determines all follow-up structures and ultimately the functions of a protein [42, 43]. Thus, the primary structure is a rich source for protein structure and function analysis.

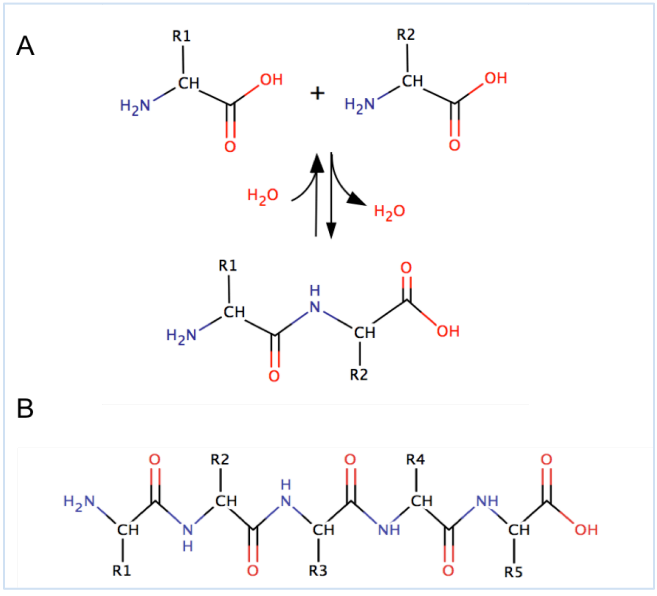


Figure 2.3. (A) Formation of a peptide bond by condensation. (B) Structure of a 5-residue oligopeptide. R1 - R5 can be the same or different amino acid residues.

2.1.2 Secondary Structure

2.1.1.1 Alpha-helix and Beta-sheet

Secondary structure is the local spatial arrangement of backbone atoms for a segment of protein sequence. The two most common secondary structures are α-helix and β-sheet, as illustrated in Figure 2.4. Figure 2.5 shows that most secondary structures have distinctive Φ and Ψ backbone angles, which are defined by the dihedral angle of C-N-C_α-C and N- C_α-C-N respectively. An α-helix has approximately 3.6 residues per turn [44], while a β-sheet is normally 3-10 residues per strand [45]. Both the α-helix and β-sheet are held together by the hydrogen bond interaction between the amino acid residues. Their stability is greatly affected by their primary structure. Beside these two common secondary structures, there are also some non-repeatable and less regular structures, such as β-turns and loops, which are often categorized as coil. They can be the connection between helices and sheets, and they are often subject to more conformational flexibility.

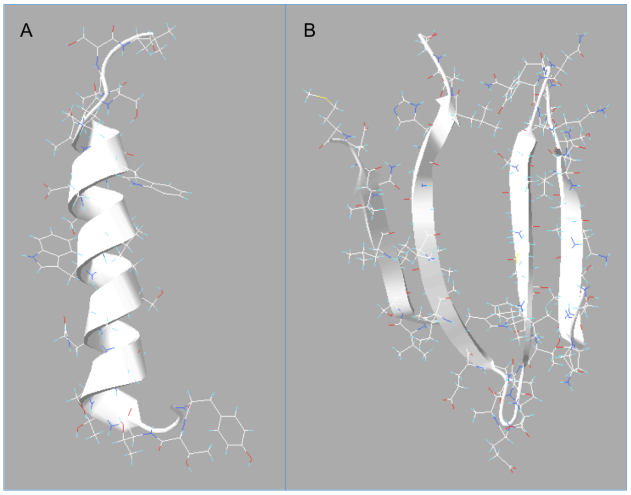


Figure 2.4 Example of α-helix (A) and β-sheet structures (B). PDB ID: 1RTQ. Structure figure generated in Swiss-PDBviewer (http://www.expasy.org/spdbv/) [46].

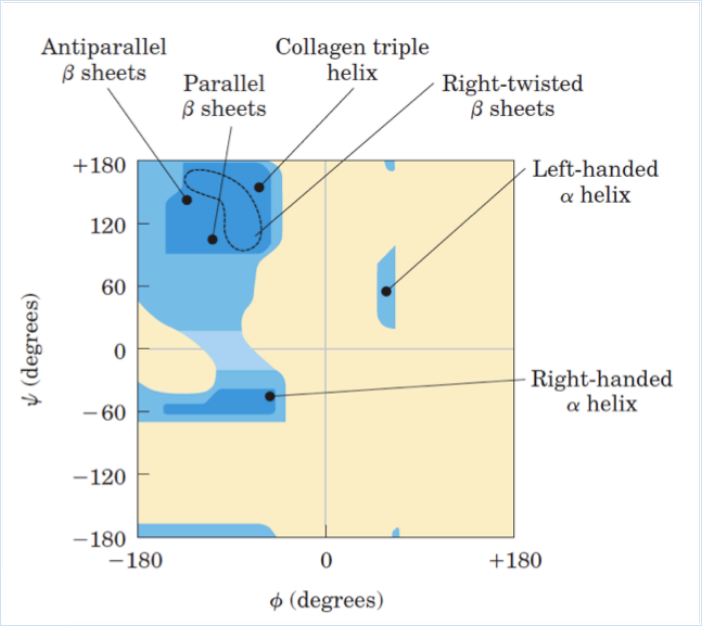


Figure 2.5 Ramachandran plot for general amino acids[38]. From *Lehninger Principles of Biochemistry,* 6th Edition, By David L. Nelson and Michael M. Cox, Copyright 2013 by W.H. Freeman and Company. Used with permission by the publisher.

2.1.1.2 Motif and Domain

Motif and domain are two terms often used to characterize a collection of secondary structural elements with specific 3D relationships [32]. A motif is the distinguishable three-dimensional pattern of a few secondary structures together. It can be recurrently found in many protein structures. It does not necessarily have structural independency, or functional integrity[47, 48]. In contrast, a domain is an independent stable portion of a protein that is stable and exhibits specific functions even within different proteins. It normally can fold into and maintain its structure enabling specific function even if it is separated from rest of the protein [49].

2.1.3 Tertiary Structure

Tertiary structure is defined as the overall three-dimensional structure of one polypeptide chain. It is a combination of protein secondary structures, motifs, and domains. The process of a protein converting from extended linear chains into its tertiary structure is often referred to as protein folding. In comparison to secondary structure, which is mainly defined by the local backbone interaction, the physiochemical properties of amino acid side chains (R-groups) greatly influence protein folding via main types of interactions including non-specific hydrophobic interactions with the solvent, specific hydrogen bonding, ionic interaction such as salt bridges, and disulfide bonds between cysteine residues. The folding process goes under a lot of physiochemical and biological regulations to achieve the final one or limited number of conformations, albeit that the possible folding space is almost infinite[50, 51]. Thus protein folding is a complex process, and has not been complete understood and modeled yet. A lot of progress has been made, but it is still unfeasible to predict the tertiary structures for all available protein sequences to the same level of accuracy as experimental methods [51-53].

There are three major biochemical types of tertiary structures: globular, membrane-associated, and fibrous[32]. Globular proteins are often composed of different types of secondary structure, and fold into a compact globular structure. They are often found in aqueous environment, and are the predominant proteins in cytoplasm and body fluid. Due to its easy accessibility, their structures are well studied and consist of the majority portion structures in wwPDB[34], the central repository of biological macromolecular structures. Membrane-associated proteins are very similar in the overall shape as globular proteins, except the surface of the protein is often hydrophobic so as to

fit in its environment, the membrane. They are much harder to determine experimentally, and therefore are poorly represented in the wwPDB. Fibrous proteins are drastically different from the above two proteins. They have simple repetitive component of secondary structures, and often function to create and sustain gross structural biological components. These structures are also poorly represented in the wwPDB due to their difficulty to be experimentally derived.

2.1.4 Quaternary Structure

Quaternary structure is an assembly of multi-subunits or multi-polypeptides of a protein[38]. It can range from a simple dimer to a very large multi-mer. The subunits are often individually folded, and can be either identical or different. They are normally held together through non-covalent bond between amino acid residues[54, 55].

Proteins structures are complex, but there are basic principles that can guide us in better understanding them. The rapid developing experimental technology and computational tools are now available to help us determine protein structures. There are also projects like Protein Structure Initiative (PSI) [56] that help to make protein structures easily obtainable from knowledge of their corresponding DNA sequences. With more and more data available, more questions can be asked now from a bioinformatics perspective, and assist us to move forward with a deeper understanding of protein structures, which biochemical and biological functions they enable, and ultimately how these protein structures enable their specific functions.

2.2 Crystal Field Theory and Metal Coordination Geometry

2.2.1 Crystal Field Theory (CFT)

CFT is a model that explains the breaking of orbital degeneracy in transition metal coordination complexes due to the presence of ligands. Hans Bethe and John Hasbrouck van Vleck first developed this theory in the 1930s[57]. It successfully accounted for coordination geometry, magnetic properties, and colors of transition metal complexes, but it did not attempt to describe bonding itself. It was then developed further as ligand field theory, which explained the chemical bonding of transition metal complex.

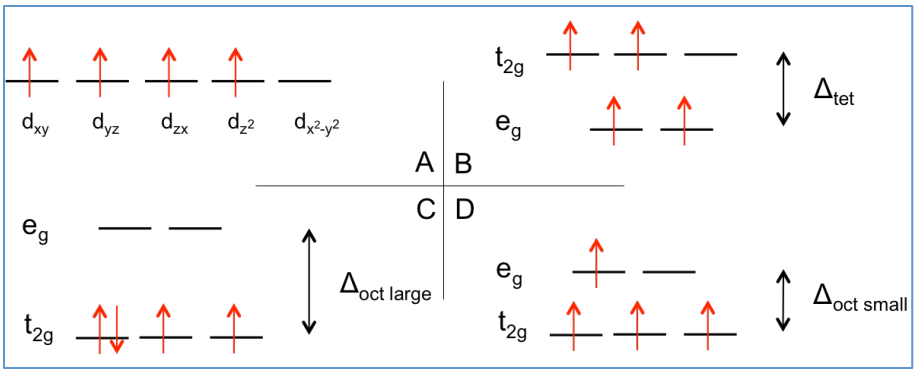


Figure 2.6 d-orbital energy splitting diagrams.

Transition metals normally possess five d orbitals with the same energy level as shown in Figure 2.6 A. According to CFT, when there are ligands around the metal, the electron-electron repulsion between the metal and the ligands will cause the splitting between the five d orbitals[58]. The splitting pattern depends on the metal-ligand geometry. Figure 2.6 B shows the splitting pattern for tetrahedral, while panel C and D are both for octahedral. The energy difference of the splitting (Δ_{tet} and Δ_{oct}) has to do

with the binding strength of the ligand. Strong field ligands create large splitting, as in Figure 2.6 C, while weak field ligands produce small splitting, as in Figure 2.6 D. The order of some common ligands is as follows:

$$I^- < Br^- < S^{2-} < Cl^- < F^- < OH^- < H_2O < NH_3 < NO_2^- < CN^- < CO.$$

Weak-field ligand is on the left hand side. They create small splitting, and the electrons are prone to be in the high-spin orbital. The energy gap between the splitting then determines the position of the electron pairs.

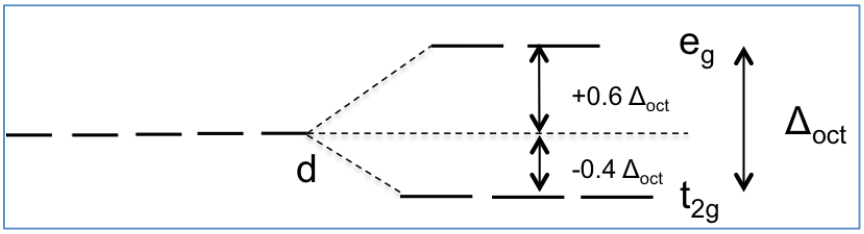


Figure 2.7 Energy splitting diagram for octahedral.

2.2.2 Crystal Field Stabilization Energy and Metal Coordination Geometry

The Crystal Field Stabilization Energy (CFSE) is defined as the energy of the electron configuration in the ligand field minus the energy of pairing electrons in one orbital:

$$CFSE = \Delta E = E_{ligand\ field} - E_{isotropic\ field} \qquad (2.1)$$

In an energy splitting diagram of octahedral (shown in Figure 2.7), the CFSE of a 4-electron low-spin (Figure 2.6.C) is:

$$CFSE_{low} = (4 * -0.4\ \Delta_{oct}) + P = -1.6\Delta_{oct} + P,$$

where P represents the Spin Pairing Energy. And a similar high-spin (Figure 2.6.D) is:

$$CFSE_{high} = (3 * -0.4\ \Delta_{oct}) + (1 * 0.6\ \Delta_{oct}) = -0.6\Delta_{oct}.$$

The preference of high verse low spin depends on the energy difference between Δ_{oct} and P, as well as the number of d-orbital electrons.

Similarly for the splitting systems of other CGs, the favored high/low spin and its CFSE can be calculated based on the CG's specific energy splitting diagram. The possible CGs can then be selected via the smallest CFSEs. In the factors that influence the CFSE, the energy difference of the splitting, Δ_{cg}, is affected by the binding strength of the ligand; the pairing energy, P, is a property of the metal and affected by the energy splitting level; and the number of d-orbital electrons is determined by the metal and its oxidation state. All these factors together define which CGs a metal can adopt.

Zinc, for example, is a transition metal, and binds to proteins in its +2 states, which means a stable full $3d^{10}$ and empty $4s^2$ and $4p^6$ orbitals. Since the d orbital is full, there is no difference in the electron pairing energy and the energy from the split ligand field is also canceled out for any CG. As a result, the CFSE of all CGs are the same, which means this electron configuration allows zinc to bind four, five, and six ligands with roughly equal stability [59].

2.2.3 Ligand Field Energy (LFT)

To further explain the bonding, orbital composition, and other properties of coordination complexes, the LFT was developed as an extension of the CFT. It is based on both the CFT and the molecular orbital theory. As shown in Figure 2.8, transition metal ion normally has nine outer layer orbitals, five nd, one (n+1)s, and three (n+1)p orbitals. In an octahedral coordination, when the ligands approach the metal from x, y, and z axes, some orbitals become higher in energy as anti-bonding orbitals and some

become lower in energy as bonding orbitals. The five d orbitals also degenerate in the center of the diagram as predicted by the CFT. The degenerated higher energy d orbitals, s orbitals and p orbitals form the M-L σ orbital together with the ligand orbitals. Six of them are bonding while the other six are anti-bonding. The degenerated lower energy d orbitals are un-affected, and become non-bonding d orbitals. They have the potential to form π bonding with the p orbitals from the ligands. The final interaction between the metal and ligands are a synergic combination of σ-bonding and π-bonding. The empty orbitals of the metal in Figure 2.8 can be filled with electrons according to the property of the metal, which follows all the rules introduced in the CFT.

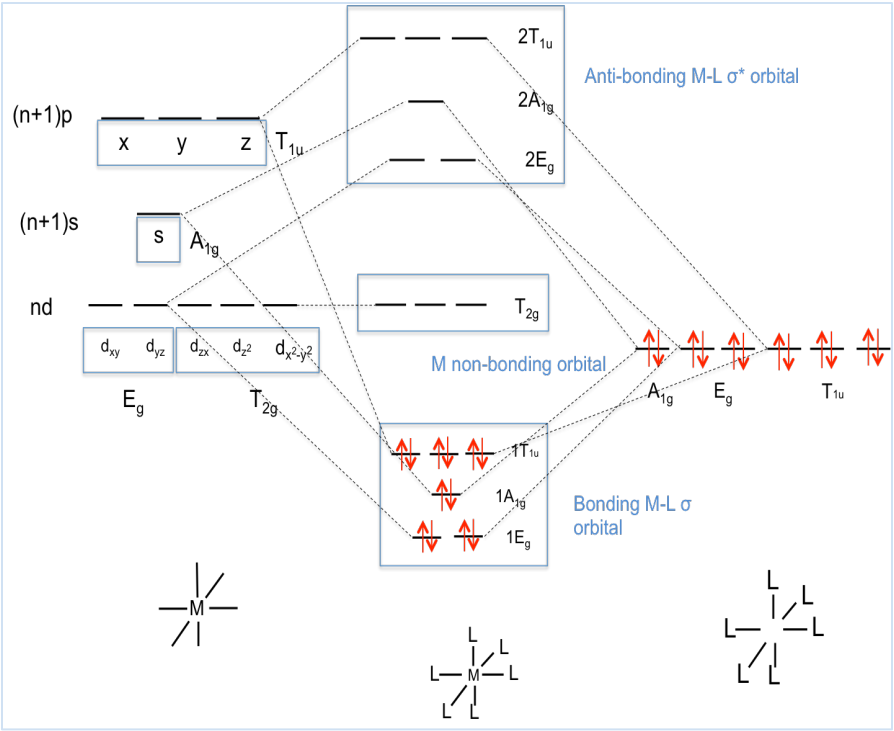


Figure 2.8 Ligand-Field scheme of σ-bonding in the octahedral ML_6 complex.

2.2.4 CGs in Metalloproteins

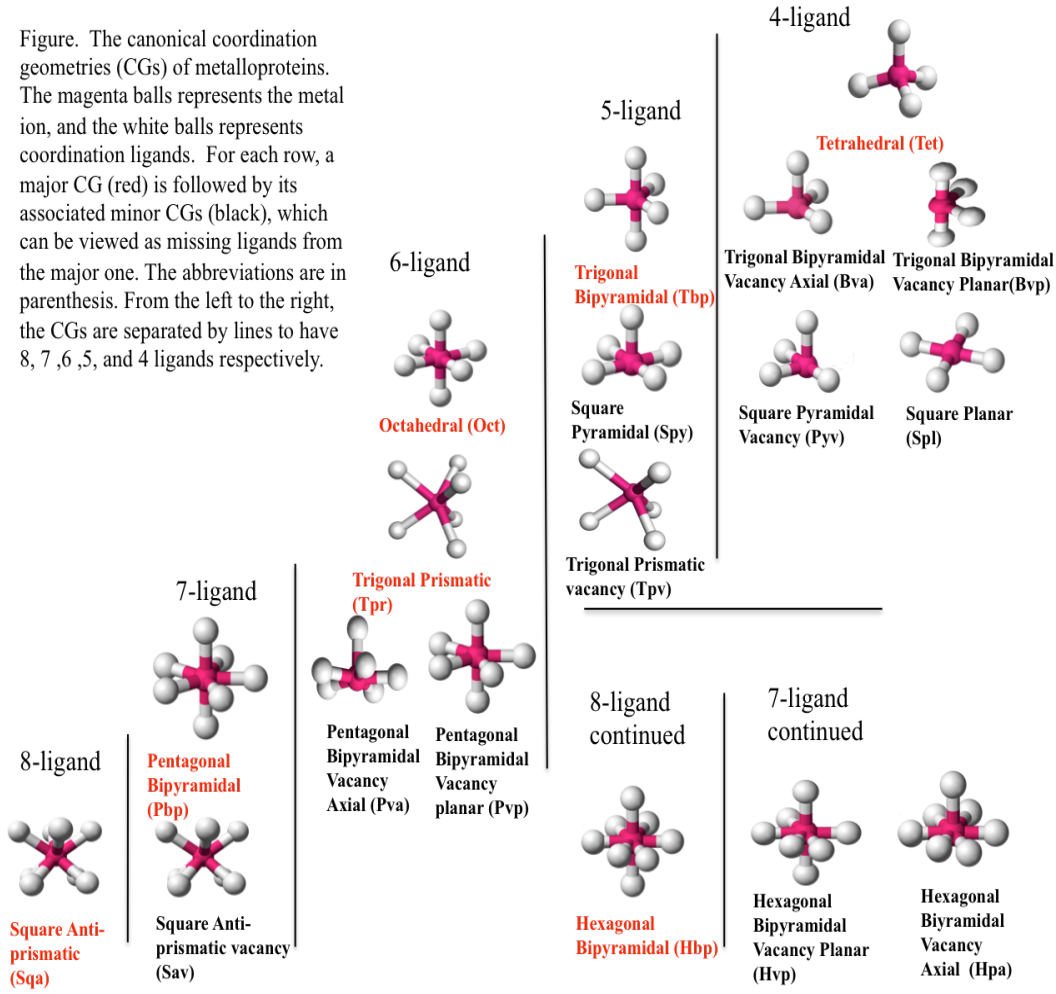


Figure 2.9 Structure of canonical CG models in metalloproteins

Metal ions bind to proteins via electronegative ligands like nitrogen, oxygen, or sulfur atoms of amino acids in the metalloprotein. CG is defined as the geometric pattern shaped by the metal ion center and its surrounding atoms, and is one of the most important aspects of metal structure. The binding atoms are also called ligands. A metal's CG identifies the set of proper ligands and their spatial orientation to the metal, and often

has functional implications. According to the CFT mentioned in Section 2.2.1 and 2.2.2, there are several major CGs that metal ions can adopt, including Tetrahedral (Tet), Trigonal Bipyramidal (Tbp), Octahedral (Oct), and others, as shown in Figure 2.9 (bold and red), where the magenta balls represent the metal ion and the white balls represent ligands. Due to biological variations or missing substrates, major-CG-associated minor CGs also have been reported [60], and structures and abbreviations of them are also shown in Figure 2.9. Studies have shown that different CGs exhibit very distinct ligand compositions and functional propensities [8, 25]. The two most important properties that could define a CG are ligand-metal-ligand angle (angle) and metal-ligand bond-length (bond-length). Also, the CGs can be classified into four-, five-, six-, seven, and eight-ligand based on the coordination number. Different CGs have distinctive set of angles, as shown in Table 2.1.

Table 2.1 Ideal angles of canonical CG models.

CG	Angles	CG	Angles
4-ligand:		7-ligand:	
Tet	109.5 x 6	Pbp	72 x 5, 90 x 10, 144 x 5, 180 x 1
Bva	90 x 3, 120 x 3	Hva	60 x 6, 90 x 6, 120 x 6, 180 x 3
Bvp	90 x 4, 120 x 1, 180 x 1	Hvp	60 x 5, 90 x 10, 120 x 3, 180 x 3
Pyv	90 x 5, 180 x 1	Sav	70.5 x 6, 82 x 6, 109.5 x 3, 143.6 x 6
Spl	90 x 4, 180 x 2	8-ligand:	
5-ligand:		Hbp	60 x 6, 90 x 12, 120 x 6, 180 x 4
Tbp	90 x 6, 120 x 3, 180 x 1	Sqa	70.5 x 8, 82 x 8, 109.5 x 4, 143.6 x8
Spy	90 x 8, 180 x 2		
Tpv	70.6 x 2, 90 x 4, 131.8 x 4		
6-ligand:			
Oct	90 x 12, 180 x 3		
Pva	72 x 5, 90 x 5, 144 x 5		
Pvp	72 x 4, 90 x 7, 144 x 3, 180 x 1		
Tpr	70.6 x 3, 90 x 6, 131.8 x 6		

2.3 Available Databases And Tools Of Protein And Metalloprotein Structure

2.3.1 Protein Primary Structure (Sequence) Databases

The application of rapidly improving genomic sequencing technologies is generating huge amounts of gene sequence information and expression data. These technologies are rapidly growing both gene sequence and derived protein sequence databases, ranging from plain sequence repositories, which has little or no manual intervention and no connection to other databases, to extensively curated databases that analyzed, classified, annotated with functional characterization, and cross-referenced to lots of other databases.

The National Center for Biotechnology Information (NCBI) Protein database (http://www.ncbi.nlm.nih.gov/protein/) is a simple sequences repository, with its record coming from several different sources, including RefSeq [61], Swiss-Prot [62], PIR [63], and PDB [64]. This database contains mainly the sequence information, with some reference to its source database, and various link to other NCBI gene and transcript sequence database.

The Universal Protein Resource (UniProt) (http://www.uniprot.org/) [62] is a comprehensive resource for protein sequence and annotation data. It is a collaborated work between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource (PIR). It is composed of two main parts, UniProtKB/Swiss-Prot [65], which is manually curated and annotated, and UniProtKB/TrEMBL [66], which is uncurated, computationally generated. Records

in TrEMBL will be added into Swiss-Prot after experts annotate them, while more records in Swiss-Prot will update the algorithm to generate more annotations in TrEMBL.

2.3.2 Protein Secondary and Super-Secondary Structure Databases

The Structural Classification of Proteins (SCOP) database [37] is a comprehensive classification of all proteins of known structure, according to protein domains and their structural similarities. It was created in 1994 and is still maintained by the Centre for Protein Engineering and the Laboratory of Molecular Biology. Protein domains in SCOP are hierarchically classified into families, superfamilies, folds and classes based mainly on the composition of secondary structures and the shape they form. In SCOP, families contain protein domains that share a common ancestor, for having the same shape and similar sequence and/or function. Superfamilies contain protein domains that are from distinct origins, where they share the same shape, but have little sequence or functional similarity. Folds contain a group of superfamilies that process a common core structure, like Globin-like, and Long alpha-hairpin. Finally, depending on the type of folds, they are grouped in to all-alpha, all-beta, etc. classes. The last version of SCOP, release 1.75 was built from 38,221 PDB Entries (23 Feb 2009), with 110800 Domains. The total number for folds, superfamilies, and families are 1195, 1962, and 3902 accordingly.

In 2014, SCOP was discontinued, and the prototype of a new Structural Classification of Proteins 2 (SCOP2) database [36] has been made publicly available. SCOP2 inherited from SCOP. Similarly to SCOP, SCOP2 starts from proteins and protein domains that are available in the PDB. It uses a new approach in the classification of proteins, where it constructs directed acyclic graph network of nodes instead of a tree-

like hierarchy. Each node represents a region of protein structure and sequence, and the network exemplifies a many-to-many relationship between protein structures.

CATH [35] is another systematic classification of protein structures and is short for its four main hierarchies, protein class (C), architecture (A), topology (T) and homologous superfamily (H). Its latest release CATH v4.0, which was built upon PDB dated March 26, 2013, contains 235,858 CATH domains, 2,738 CATH superfamilies, and 69,058 annotated PDBs. Class is the simplest level, and basically describes the secondary structure composition and packing, such as all-alpha, all-beta, alpha/beta. Architecture summarizes the orientations and shape formed by the secondary structure units, such as bundles, sandwiches, and barrels. At the topology level, both structure orientation and sequential connectivity is taken into account. For example, barrels can be further classified as αα barrel, αβ barrel, β barrels and others. Superfamily is assigned when structures belonging to the same T-level have similar functions, sequence or be evolutionary support of some lever of homology. Proteins are put into different categories through a semiautomatic manner, where the domains identification, sequence alignment, and structure comparison are all computed using various of programs, and any unclassified structures are then manually checked and assigned.

2.3.3 Protein Tertiary and Quaternary Structure Databases

2.3.3.1 The Worldwide Protein Data Bank (wwPDB)

wwPDB [34] was first launched in 1971 at Brookhaven National Laboratory as the Protein Data Bank archive (PDB). It has now developed into an organization that manages Protein Data Bank in Europe (PDBe) [67], Biological Magnetic Resonance Data Bank (BMRB) [68], Protein Data Bank Japan (PDBj) [69], and Research Collaboratory

for Structural Bioinformatics Protein Data Bank (RCSB PDB, PDB archive, or PDB) [64]. The RCSB PDB serves as the central repository of biological macromolecular structures, while all the other members also stores rich information about PDB entries, advanced services, and many bioinformatics tools to facilitate the global community. wwPDB is a collaboratory (center without walls) for bio-macromolecular structural research with data deposited by biologist and biochemist from all round the world. Data are self-deposited into the wwPDB by the scientists, and then become available to general public after a series of quality checks and annotation, and upon the consent of the depositor. Structural data are normally obtained by various experimental methods, such as x-ray crystallography, NMR, and electron microscopy. It is the key resource for structural biology, and many of other structural databases are built from it, such as SCOP and CATH. Structural data can be viewed on their website or in viewer programs individually, and can easily be downloaded in multiple formats. The database is updated weekly, and as of June 23, 2016, a breakdown of the PDB structures is shown in Table 2.2.

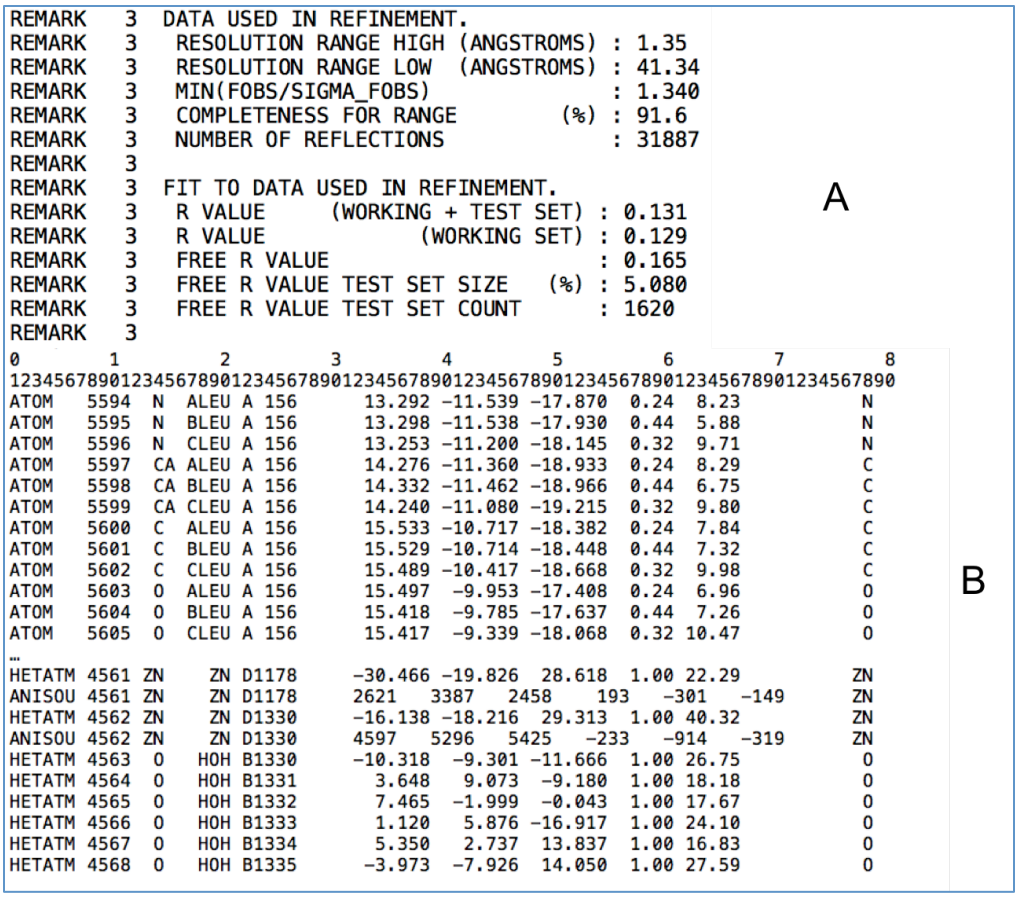
Table 2.2 PDB current holdings breakdown, June 23, 2016.

Exp.method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-ray	100,170	1,744	5,129	4	107,047
NMR	10,031	1,144	235	8	11,418
Electron Microscopy	768	30	267	0	1,065
Hybrid	90	3	2	1	96
Other	174	4	6	13	197
Total	111,233	2,925	5,639	26	119,823

The wwPDB entries contain information about the structure, including introduction, title, primary structure, heterogen (non-standard residues), secondary structure, connectivity annotation, miscellaneous features, crystallographic and coordinate transformation, coordinate, connectivity section, and bookkeeping sections. These major formats are provided: PDB, mmCIF, and PDBML (PDB XML) format. Different formats contain essentially the same information, and are presented to benefit different type of analyses and interpretors. PDB format was the first format developed, and has been serving as the standard in representing macromelecular structures for decades. The structures of some records has changed dramatically due to increasing bioinformatics challenges. mmCIF and PDBML format emerged recently to facilitate better capture and parsing of all data and metadata. They are both more modern formats, which are easier to parse, search, and mine in more systematic ways. The structures of some records between formats are dramatically different due to the changing bioinformatics use-cases and underlying computatation tools being utilized.

This project used mainly the PDB format, which is illustruated in Figure 2.10. Records in PDB format are assigned to a range of character postion. At the beginning of each line is the record name. Different types of records often have a distictive layout, which is machine parsable, while REMARK is the most free-styled record, so that it can present experimental details, comments, references, and any information that cannot fit in other records well (Panel A). ATOM and HETAEM records shown in panel B contain all the coordinate information of the structure. Information is arranged into columns. From left to right, the data fields are record name, atom serial number, atom name, alternate location indicator, residue name, chain identifier, residue sequence number, code for

insertion of residues, orthogonal coordinates for X in Å, orthogonal coordinates for Y in Å, orthogonal coordinates for Z in Å, occupancy, temperature factor, element symbol, and charge on the element.




```
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.35
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 41.34
REMARK   3   MIN(FOBS/SIGMA_FOBS)               : 1.340
REMARK   3   COMPLETENESS FOR RANGE        (%) : 91.6
REMARK   3   NUMBER OF REFLECTIONS               : 31887
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   R VALUE     (WORKING + TEST SET) : 0.131
REMARK   3   R VALUE           (WORKING SET) : 0.129
REMARK   3   FREE R VALUE                     : 0.165
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 5.080
REMARK   3   FREE R VALUE TEST SET COUNT      : 1620
REMARK   3
```

```
0        1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM   5594  N   ALEU A 156      13.292 -11.539 -17.870  0.24  8.23           N
ATOM   5595  N   BLEU A 156      13.298 -11.538 -17.930  0.44  5.88           N
ATOM   5596  N   CLEU A 156      13.253 -11.200 -18.145  0.32  9.71           N
ATOM   5597  CA  ALEU A 156      14.276 -11.360 -18.933  0.24  8.29           C
ATOM   5598  CA  BLEU A 156      14.332 -11.462 -18.966  0.44  6.75           C
ATOM   5599  CA  CLEU A 156      14.240 -11.080 -19.215  0.32  9.80           C
ATOM   5600  C   ALEU A 156      15.533 -10.717 -18.382  0.24  7.84           C
ATOM   5601  C   BLEU A 156      15.529 -10.714 -18.448  0.44  7.32           C
ATOM   5602  C   CLEU A 156      15.489 -10.417 -18.668  0.32  9.98           C
ATOM   5603  O   ALEU A 156      15.497  -9.953 -17.408  0.24  6.96           O
ATOM   5604  O   BLEU A 156      15.418  -9.785 -17.637  0.44  7.26           O
ATOM   5605  O   CLEU A 156      15.417  -9.339 -18.068  0.32 10.47           O
…
HETATM 4561 ZN    ZN D1178     -30.466 -19.826  28.618  1.00 22.29          ZN
ANISOU 4561 ZN    ZN D1178     2621   3387   2458    193   -301   -149       ZN
HETATM 4562 ZN    ZN D1330     -16.138 -18.216  29.313  1.00 40.32          ZN
ANISOU 4562 ZN    ZN D1330     4597   5296   5425   -233   -914   -319       ZN
HETATM 4563  O   HOH B1330     -10.318  -9.301 -11.666  1.00 26.75           O
HETATM 4564  O   HOH B1331       3.648   9.073  -9.180  1.00 18.18           O
HETATM 4565  O   HOH B1332       7.465  -1.999  -0.043  1.00 17.67           O
HETATM 4566  O   HOH B1333       1.120   5.876 -16.917  1.00 24.10           O
HETATM 4567  O   HOH B1334       5.350   2.737  13.837  1.00 16.83           O
HETATM 4568  O   HOH B1335      -3.973  -7.926  14.050  1.00 27.59           O
```

Figure 2.10 Example of PDB format. (A) An example of PDB REMARK record. Information about experimental data, such as resolution range, R-values are provided. Contents are moderately structured and can be text-mined with regular expression. (B) ATOM and HETAEM record. Information is column-oriented, and contains the coordinate information about the structure.

2.3.3.2 Quaternary Structure Databases

The coordinate data in PDB entries normally represent only one asymmetric unit of the structure detected in X-ray crystallography. A true biologically active quaternary structure may be a collection of several of the same unit together. So crystallographic symmetry operations and biologically assembly are required to obtain the full active form in biology. Information needed for this procedure, like space group and symmetry operations, can be found in each PDB file. Databases that compute the quaternary structures of protein are available, and serve as an extension of the PDB. PQS [70] is a protein quaternary structure file server. It performs the procedure in two main steps. The first step is to assemble any crystallographic and non-crystallographic symmetry based on information in the PDB file. The second step reexamines the quaternary complex determined by the first step to evaluate the likelihood of proposed protein-protein and protein-solvent interactions. This database later evolved into PDBePISA (Proteins, Interfaces, Structures and Assemblies) [71], which includes all features of PQS, while adding other function for searching pre-calculated results, such as accessible/buried surface area and presence/absence of salt bridges or disulphide bonds.

2.3.4 Metalloprotein Structure Databases and Tools.

Several databases are available with a focus on metalloprotein structures. Such databases typically use the available metalloprotein structures in wwPDB, then either analyze the metal structures in various different manners, or use them to build a predictive model and then apply to all other proteins. A non-inclusive list of these databases includes, Metal PDB [8], MetalS^2 [72], MetalS^3 [23], BioMe [27], CheckMyMetal [73], FindGeo [60], MetLigDB [74], MESPEUS [25], MDB [75], and

COMe [76]. This section will go over a couple of them in more details.

BioMe, biologically relevant metals, is a web server that can calculate different statistics of metalloprotein binding sites. The statistics include ligand composition of metal, distribution of coordination number, ligand combination percentage, distribution of mono- and bi-dentation, counts of bi-metal, distribution of coordination geometry, and mean and standard deviation of bond-length. They used 3Å as the distance threshold for defining the coordination shell, then a geometric root mean square deviation (gRMSD) for classifying the coordination geometry. On the website, one can select from a number of different parameters, such as metals, ligands, coordination numbers, experimental methods, etc., and conduct the calculation on the selected groups.

CHED is short for Cysteine (Cys), Histidine (His), Glutamate (Glu), and Aspartate (Asp), the four most common ligands of metal ions found in protein. Models were built from metal sites in PDB and their 95% sequence identity structures but without a metal. It then searches for any 3-ligand sets (triad) of amino acids composed of 4 residue types (Cys, His, Glu, Asp) having ligand atoms within a specific distance cutoff. Applying this algorithm to all triads detected in the Protein Data Bank and structural genomics initiative found a large number of previously unknown, putative metal-binding sites. People can also submit their own PDB formatted structure file to detect any potential metal binding sites.

Bertini and his colleagues have been one of the leading groups in studying metalloproteins, both experimentally and from a bioinformatics perspective. They have developed a series of databases/tools focusing on different aspects of metalloproteins. FindGeo is one of their first tools they developed, and is used for determining metal

coordination geometry. On this website, a user can select metal, distance cutoff, and specific element for detection, and a gRMSD value is computed to determine the metal's coordination geometry. Results can be viewed in list with detailed information and three-dimensional interactive structure. MetalPDB is a follow-up product of FindGeo. It first defines a unique term Minimal Functional Site (MFS) as the set of atoms including the metal ion, its ligands and any other atoms within 5Å from any ligands. MFS is designed to capture the local environment around the metal ion, which more than just the immediate binding ligands, while not have to consider the overall protein fold in which it is embedded. It then clusters all MFS detected in PDB into equivalent and equistructural sites based on sequence homology and structure similarity. Then when a user is searching a structure of interest, the results can return not only the characterization of the protein itself, but also similar proteins in the same group, which will hopefully provide functional implications. The database summarized over 17,000 structural clusters as of August 2012. MetalS^2 was then released as a tool that was designed specifically to compare any given two MFSs. It first overlaps the two metal centers, and orients the two structures by aligning the metal-binding ligands. It then will calculate a best score of the backbone atoms between the two structures while refining the alignments between the two structures. MetalS^3 combines both MetalPDB and MetalS^2. When given a PDB ID or a PDB formatted file, MetalS^3 could search through all structural clusters defined in Metal DB and use MetalS^2 engine to compare and find the most similar structures.

2.3.5 Protein Function Database

The ultimate goal in studying protein structures is to alleviate the huge imbalance between the explosive data on protein sequences and limited knowledge on their

functions. InterPro [77] is one of the database tools that can link protein sequence to its functions with help from several structure databases. It first classifies proteins sequence into families based on sequence homology and function, and then uses them to predict the important domains and sites. To achieve that, it uses predictive models, which are called signatures, provided by its member databases or consortium, including Pfam [78], PRINTS [79], PROSITE [80], ProDom [81], CATH-Gene3D [82], HAMAP [83], PANTHER [84], PIRSF [85], SMART [86], SUPERFAMILY [87], and TIGRFAMs [88]. The overall goal of InterPro is to unite individual databases and provide a single resource for comprehensive information about protein sequence, families, domains and functional sites.

CHAPTER 3

CURRENT STUDIES ON THE STRUCTURE OF METALLOPROTEINS AND THE HYPOTHESIS OF THIS PROJECT

3.1 Current Studies on the Structure of Metalloproteins

CG provides a bridge between the sequence space and functional space of metalloproteins, and therefore knowledge about them is rather valuable. The challenge is how to characterize a metal's CG given its xyz coordinates, which are available from structural databases such as wwPDB. The prevailing methodology is to first obtain a list of all possible CG models from the literature, and then define and score a metal coordination shell for how well it matches any known CG models. The model with the highest score will be classified as the metal's CG. Various studies differ in which sets of CG models they consider, and how they compute a comparison score of a specific CG to each CG model.

Alberts *et al*. [89] were among the first to classify the CGs of zinc metalloproteins. They manually analyzed 111 high-quality zinc sites, and only identified four CG models, including Tet, Tbp, Spy, and Oct. Due to the lack of description, it is unclear of how the classification was done. The authors summarized several of the detailed aspects in their study, including ligand combinations, bond-length and angle statistics. They observed high variance in the angle results, which is mainly due to

bidentate binding and multi-zinc sites. They also evaluated all zinc sites in two groups, structural and catalytic, using the criteria of whether or not there are solvent molecules as binding ligands.

Patel *et al*. [24] conducted a very similar study and examined 382 PDB entries using in-house programs. They classified CGs into four models, Tet, Tbp, Spy, and Oct. They observed high angle variance as well and believed that it was cause by electron pair repulsions, bidentation of carboxyl group, extraneous H-bonding to secondary coordination sphere or solvent, and bridging ligands in multi-Zn sites.

Liu *et al*. [26] developed a method to identify three-ligand and four-ligand major CG models of zinc by calculating a potential zinc center from the ligand coordinates and measuring its distance from the real zinc center. They compared the difference in sequence length, residue preference, secondary structures, and geometrical distance between 3- and 4- ligand sites. They found that the bond-length was tightly restricted, and can be used to predict potential zinc binding sites.

Harding *et al*. [20, 21, 25, 90, 91] conducted a series of studies on several different metalloproteins. The data was from Cambridge Structural Database (CSD) [92], which is an equivalent of wwPDB, but for small-molecule organic and metal-organic crystal structures. Bond-length statistics in several different conditions were summarized. They estimated that if properly modeled in X-ray, the M-O and M-N bond-length standard deviation should be in the range of 0.004-0.02 Å. The larger estimated standard deviations from the wwPDB clearly reflect the larger errors in determining atom positions in proteins. The difference of the bond-length statistics from wwPDB and CSD could indicate some systematic error or artifact in protein structure refinement. Root

mean square deviation of actual angles to their ideal values was calculated to classify CGs. They reasoned that the high angle variance after classification might come from several reasons: “(i) experimental uncertainties in determination of crystal structure, (ii) intramolecular effects, electronic or steric, (iii) intermolecular effects, sometimes called `packing forces', (iv) the existence of additional partial M--L bonds (this could be considered as a part of (ii) or (iii), but it is found here to be an important factor and is therefore listed separately).”

Andreini *et al*. [60]determined given PDB entries’ metal CGs by first superimposing the structure to ideal CG templates, and then calculating the root-mean-square-deviation (RMSD) value for each template. Out of all the studies, they considered the most complete set of CG models to compare to. They suggested a set of a uniform terminology of geometries with three-letter acronyms, which was what we adopted in this project as in Figure 2.3. They then further developed the coordination concept in to a Minimal Functional Site (MFS), which is defined as the combination of metal ion, its ligands and any atoms within 5Å from any ligand. They used MFS to reflect the local structural environment around the cofactor, regardless of the larger context of the protein fold. They then showed that this MFS has very strong functional implications.

3.2 Limitations

In all of these studies described above, only known major and some minor CG models were considered. Thus, if a previously unreported CG existed, specific instances from the new CG would either be misclassified into an expected model or considered as outliers and not classified at all. These methods all have a potential problem if there is not

a proper model to be classified into. There is also a tendency to underestimate the biological variance that may cause the average of actual structures to differ from ideal ones. To the best of our knowledge, no study has tried to explain the high variability after classification in terms of possible missing CGs. Most accepted variance and tried to explain it with bidentation or multi-metal, or simply remove a large number of outliers.

To illustrate this problem, MetalPDB conducted a summary on the CGs for zinc metalloproteins, and the category that got the highest count is irregular (outlier), as shown in Figure 3.1. That is, there exist many aberrant structures in real metalloproteins data, and they are classified as outliers when using a threshold to capture bad scores. On the other hand, if those structures are forcefully classified into one of the known CGs, it could cause a high variance in the class variance. It is illustrated later in Chapter 4.3.2 that the in-class angle variance is typically below 10 degrees, while the real angle variance can get up to 24 degrees (Figure 3.2).

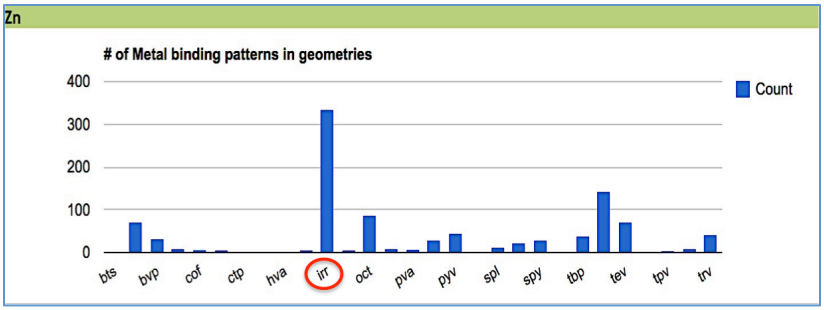


Figure 3.1 Example shown from MetalPDB [8]. When using a score cutoff for ensuring good classification into known coordination geometries, the most abundant coordination geometry category turns out to be irr (irregular, i.e. outliers).

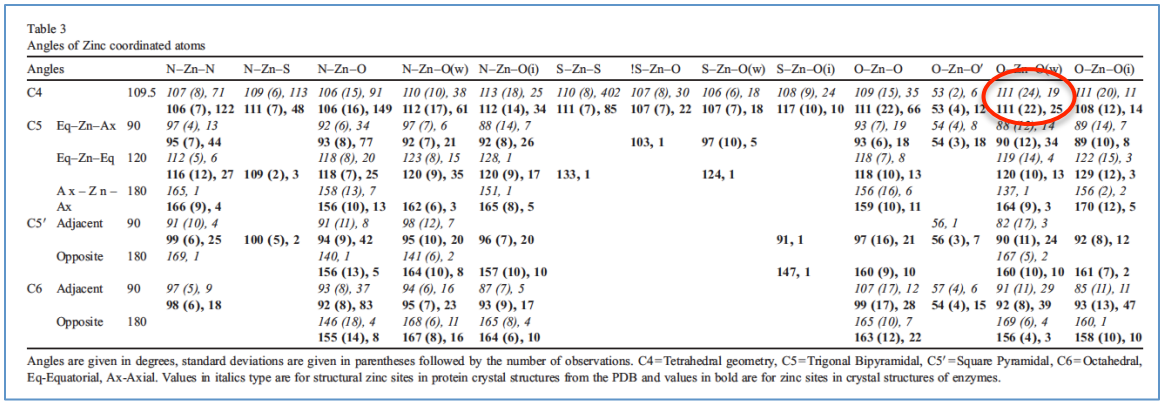


Table 3
Angles of Zinc coordinated atoms

Angles			N–Zn–N	N–Zn–S	N–Zn–O	N–Zn–O(w)	N–Zn–O(i)	S–Zn–S	!S–Zn–O	S–Zn–O(w)	S–Zn–O(i)	O–Zn–O	O–Zn–O′	O–Zn–O(w)	O–Zn–O(i)
C4		109.5	*107 (8), 71*	*109 (6), 113*	*106 (15), 91*	*110 (10), 38*	*113 (18), 25*	*110 (8), 402*	*107 (8), 30*	*106 (6), 18*	*108 (9), 24*	*109 (15), 35*	*53 (2), 6*	*111 (24), 19*	*111 (20), 11*
			106 (7), 122	**111 (7), 48**	**106 (16), 149**	**112 (17), 61**	**112 (14), 34**	**111 (7), 85**	**107 (7), 22**	**107 (7), 18**	**117 (10), 10**	**111 (22), 66**	**53 (4), 12**	**111 (22), 25**	**108 (12), 14**
C5	Eq–Zn–Ax	90	*97 (4), 13*		*92 (6), 34*	*97 (7), 6*	*88 (14), 7*					*93 (7), 19*	*54 (4), 8*	*88 (12), 14*	*89 (14), 7*
			95 (7), 44		**93 (8), 77**	**92 (7), 21**	**92 (8), 26**		**103, 1**	**97 (10), 5**		**93 (6), 18**	**54 (3), 18**	**90 (12), 34**	**89 (10), 8**
	Eq–Zn–Eq	120	*112 (5), 6*		*118 (8), 20*	*123 (8), 15*	*128, 1*					*118 (7), 8*		*119 (14), 4*	*122 (15), 3*
			116 (12), 27	**109 (2), 3**	**118 (7), 25**	**120 (9), 35**	**120 (9), 17**	**133, 1**		**124, 1**		**118 (10), 13**		**120 (10), 13**	**129 (12), 3**
	A x – Z n – Ax	180	*165, 1*		*158 (13), 7*		*151, 1*					*156 (16), 6*		*137, 1*	*156 (2), 2*
			166 (9), 4		**156 (10), 13**	**162 (6), 3**	**165 (8), 5**					**159 (10), 11**		**164 (9), 3**	**170 (12), 5**
C5′	Adjacent	90	*91 (10), 4*		*91 (11), 8*	*98 (12), 7*							*56, 1*	*82 (17), 3*	
			99 (6), 25	**100 (5), 2**	**94 (9), 42**	**95 (10), 20**	**96 (7), 20**				**91, 1**	**97 (16), 21**	**56 (3), 7**	**90 (11), 24**	**92 (8), 12**
	Opposite	180	*169, 1*		*140, 1*	*141 (6), 2*								*167 (5), 2*	
					156 (13), 5	**164 (10), 8**	**157 (10), 10**				**147, 1**	**160 (9), 10**		**160 (10), 10**	**161 (7), 2**
C6	Adjacent	90	*97 (5), 9*		*93 (8), 37*	*94 (6), 16*	*87 (7), 5*					*107 (17), 12*	*57 (4), 6*	*91 (11), 29*	*85 (11), 11*
			98 (6), 18		**92 (8), 83**	**95 (7), 23**	**93 (9), 17**					**99 (17), 28**	**54 (4), 15**	**92 (8), 39**	**93 (13), 47**
	Opposite	180			*146 (18), 4*	*168 (6), 11*	*165 (8), 4*					*165 (10), 7*		*169 (6), 4*	*160, 1*
					155 (14), 8	**167 (8), 16**	**164 (6), 10**					**163 (12), 22**		**156 (4), 3**	**158 (10), 10**

Angles are given in degrees, standard deviations are given in parentheses followed by the number of observations. C4=Tetrahedral geometry, C5=Trigonal Bipyramidal, C5′=Square Pyramidal, C6=Octahedral, Eq-Equatorial, Ax-Axial. Values in italics type are for structural zinc sites in protein crystal structures from the PDB and values in bold are for zinc sites in crystal structures of enzymes.

Figure 3.2 Example shown in Table 3 from Patel *et al.* 2007 [24]. Some angle variances can get at high as 24 degrees when forcefully classify every metal sites into one of the pre-assumed coordination geometry models. Used with permission.

3.3 New Hypothesis And Methodology to Characterize the CG of Metalloproteins

In our own initial analysis with zinc metalloproteins using only known CGs, we observed similar phenomenon of abnormally high variance in classified CGs (Table 3.1). We have tried to directly handle and understand the reasons for the high variability in zinc CG. As we explored the factors that could cause such high variance in angles, we detected the existence of significant numbers of compressed angles due to coordination by bidentate ligands (more details in Chapter 4.3.1). Thus, if forcibly classified into one of the known CGs, the metal sites with a compressed angle will cause the high variance observed in Table 3.1.

Table 3.1 Ligand-zinc-ligand angles statistics when forcibly classified into canonical CG models

Model	Count	Ideal Angle (degrees)	Mean Angle (degrees)	Standard Deviation (degrees)	Coefficients of variation
Tetrahedral (Tet)	10,077	109.5	109.1	8.66	0.079
Tetrahedral Vacancy (Tev)	493	109.5	105.2	10.9	0.104
Trigonal Bipyramidal (Tbp)	597	90	93.60	13.2	0.141
		120	116.2	13.8	0.119
		180	146.9	45.7	0.311
Trigonal Bipyramidal Vacancy Axial (Bva)	884	90	92.56	13.9	0.150
		120	115.7	19.5	0.169
Trigonal Bipyramidal Vacancy Planar (Bvp)	1,597	90	90.27	16.8	0.186
		120	120.8	10.7	0.089
		180	140.1	37.6	0.268
Octahedral (Oct)	325	90	89.96	6.66	0.074
		180	169.4	9.02	0.053
Square Planar (Spl)	18	90	89.80	6.30	0.070
		180	168.9	5.68	0.034
Square Pyramidal (Spy)	632	90a	91.84	7.23	0.079
		90p	90.97	11.0	0.121
		180	164.4	19.4	0.118
Square Pyramidal Vacancy (Pyv)	1,178	90a	95.02	7.86	0.083
		90p	92.71	10.1	0.109
		180	157.0	24.2	0.154
Trigonal Planar (Tpl)	51	120	117.1	12.1	0.103
Overall	15,852	-	-	10.4	

Given this, **we hypothesized that the high variability observed in metalloproteins are due to the existence of a significant number of aberrant or novel CGs, which are prevalent across all metalloproteins, and have functional implications.**

To the best of our knowledge, bond-length is strongly dependent on direct physicochemical properties of the metal and binding atom, while angles are dependent on both physiochemical properties and the biochemical function(s) of the bound metal. At the same time, the zinc bond-length showed very low variance in classified canonical

CGs (Table 3.2), which is consistent with several other studies. The initial results and other studies prompted us to develop a less biased method for classifying metalloprotein coordination geometries. Our methodology involves two main steps: first is to acquire the metal first coordination shells via statistical test using bond-length parameter only, and second is to cluster them based on their angle similarities, and then assign known and novel CGs to each cluster. This methodology is model-free, and allows us to learn and assign the final CGs from the data itself. Thus, we can handle unknown aberrant CGs that may cause problems in pure classification methods. Also, we can deal with the compressed group separately, so that they can be examined in more details without interfering normal group metal shells.

Upon using this less biased analysis, we discovered previously uncharacterized CG models. As to our best knowledge, no previous study has tried to explain the high variability after classification in terms of possibly unknown CGs. Our efforts also include analyses of the functional annotation of these new structural classifications, which indicate distinct functional relationships for these previously uncharacterized CGs.

Table 3.2 Zinc-ligand bond-length statistics when forcibly classified into canonical CG models

Zn–X	Count	Mean Bond Distance (Å)	Standard Deviation	Coefficients of variation
Zn-S	26,770	2.34	0.16	0.068
Zn-O	25,417	2.25	0.31	0.138
Zn-N	23,582	2.14	0.18	0.084
Zn-Cl	354	2.38	0.33	0.139
Zn-P	182	2.97	0.12	0.040

CHAPTER 4

THE COORDINATION GEOMETRY AND FUNCTIONAL PROPENSITIES OF ZINC METALLOPROTEINS

4.1 Introduction

Zinc metalloproteins are proteins that contain at least one zinc ion cofactor. They are the most abundant metalloproteins in living organisms composing an estimated 10% of the whole proteomes [19]. They participate in various biological processes and are crucial across all domains of life [7]. Zinc can play structural, functional, or regulatory roles, from holding protein structures together and participating in enzymatic reactions, to signaling and regulating other proteins' activity. Due to its prevalence and importance, the number of studies on zinc metalloproteins keeps increasing significantly. It has also been a popular target for drug designs [93-95].

As more and more data become available on zinc metalloproteins, the need for bioinformatics tools and methods with the aim of gaining any kind of global perspective of these zinc metalloproteins has also increased significantly [96, 97]. Traditional bioinformatics analyses of protein sequence have uncovered the ubiquity of zinc metalloproteins and many of its functional roles, while structural bioinformatics can provide even stronger connections between zinc metalloprotein sequence and function [98]. The exploration of zinc metalloprotein structure-function relationships requires

structure-based analyses that include adequate coordination geometry (CG) representations. Current methodologies in characterizing the CG of zinc metalloproteins, however, consider only previously reported CG models based mainly on non-biological context. Thus, if a previously unreported CG existed, specific instances would either be misclassified into a canonical model or considered as outliers and not classified at all. In this chapter, we present a method we developed that directly handles potential exceptions without pre-assuming any CG models.

4.2 Methods

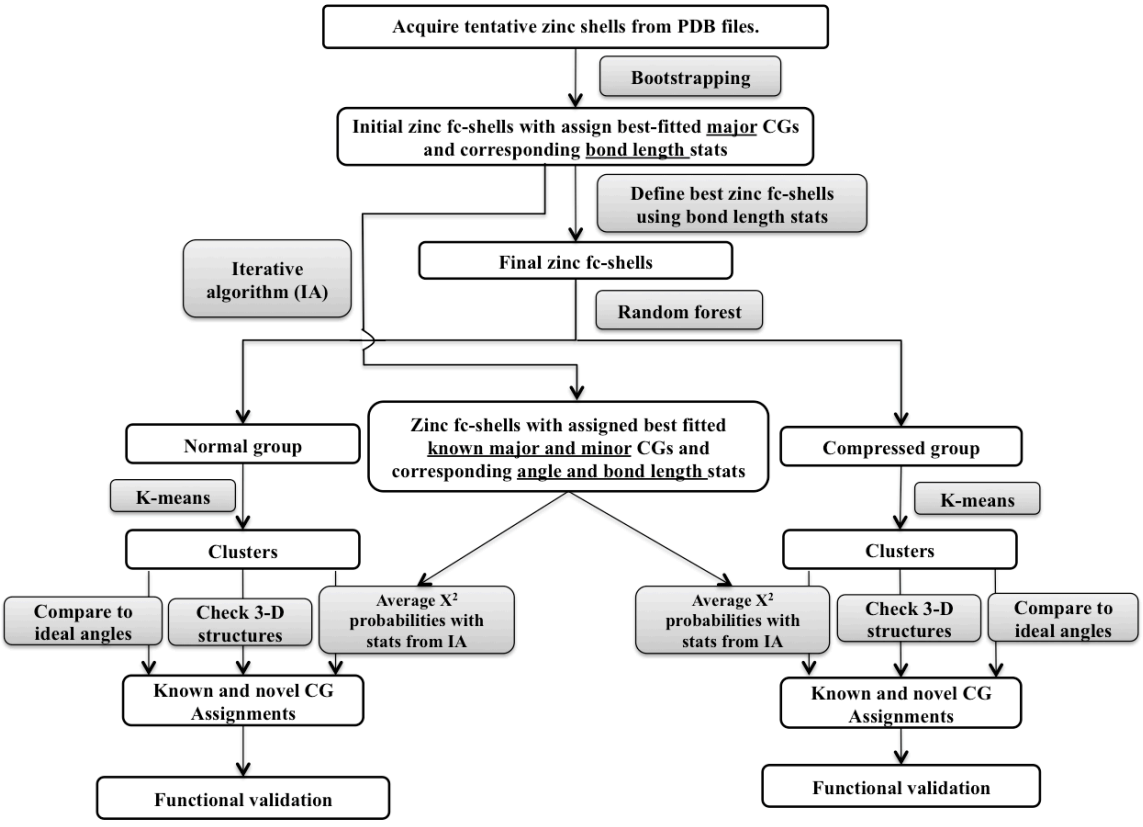


Figure 4.1 Workflow of Chapter 4.

4.2.1 Defining Zinc First Coordination Shells (Fc-Shells)

According to several other studies and our own initial results, the bond-length has more stable statistics compared to the angle. Thus we developed methods to define zinc fc-shells, i.e. the directly coordinating ligands, primarily from bond-length statistics.

4.2.1.1 Acquire Zinc Metalloproteins from PDB

Structural data was acquired from the wwPDB on Mar 13, 2013. Our initial data filtering tools identified all PDB entries with at least one zinc atom in the HETATM record and removed entries with fewer than 20 amino acids in the SEQRES record. Next, zinc clusters were identified and removed, using two zinc atoms closer than 3 Å as the filter. For each of the remaining zinc sites, we generated a potential zinc ligand list from the non-C/H atoms within a distance cutoff between 1.3 Å to 3.2 Å of the zinc atom.

4.2.1.2 Acquire Bond-length Statistics Via Empirical Bootstrapping (Step 1)

For a given zinc ion and its potential ligands, our CG evaluation tools were applied to bootstrapping the best-fitted canonical CGs, Tetrahedral (Tet), Trigonal Bipyramidal (Tbp), and Octahedral (Oct). To achieve this, our tools evaluated all possible permutations of four, five, and six ligands as an all-to-all mapping to the ligands of an ideal major CG. If the potential ligand list contained four or more atoms, all non-equivalent permutations of four were mapped to the ideal Tet four ligands, and the corresponding angles were compared to the ideal angles. The angle variance was computed as:

$$\sigma_s^2 = \frac{1}{I}\sum_{i=1}^{I}(a_{s,i} - e_{s,i})^2 \qquad (4.1)$$

where $a_{s,i}$ is the i-th observed ligand-zinc-ligand angle in the structure for CG model *s*, *I* is the total number of angles (6 for Tet, 10 for Tbp, and 15 for Oct), and $e_{s,i}$ is the i-th ideal (expected) angle of the corresponding CG model *s* (refer to Table 2.1 for ideal angle list). For each given zinc site, our tools calculated one variance for each permutation. The permutation with the smallest variance was then identified as the **initial zinc fc-shell**. The corresponding model *s* was assigned to the given zinc as an initial best-fitted major CG.

Angle statistics are CG specific. For each CG model *s* and its angle position *i*, our tools calculated the angle statistics (mean and variance),

$$\hat{\mu}_{s,i} = \frac{1}{M}\sum_{m=1}^{M} a_{m,s,i} \; ; \; \hat{\sigma}_{s,i}^2 = \frac{1}{M-1}\sum_{m=1}^{M}\left(a_{m,s,i} - \hat{\mu}_{s,i}\right)^2 \qquad (4.2)$$

where $a_{m,s,i}$ is the observed angle *i* of CG *s* for fc-shell m. Bond-length statistics are element specific. For each element t (O, N, S, …), our tools calculated element-specific bond-length statistics (mean and variance),

$$\hat{\mu}_t = \frac{1}{N}\sum_{n=1}^{N} b_{n,t} \; ; \; \hat{\sigma}_t^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(b_{n,t} - \hat{\mu}_t\right)^2 \qquad (4.3)$$

where $b_{n,t}$ is the *n*th Zn-*t* bond-length derived from all initial fc-shells and t is given ligand element (e.g. O, N, S, …).

4.2.1.3 Define Best Zinc Fc-Shells Using Bond-length Statistics (Step 2)

We then reexamined each zinc ion and its potential ligand list to define the final fc-shells. All non-equivalent combinations of potential ligands were considered. We calculated the term χ^2 probability (p-value) as 1 minus the cumulative distribution function of a χ^2 distribution. That is, P(B) = 1 - P($\chi^2(B) \leq \chi^2_{q,obs}$) and B is the degrees

of freedom which is the same as the number of ligands in combination. The χ^2 statistic was calculated using:

$$\chi^2_{obs} = \sum_{j=1}^{J} \left(\frac{b_j - \hat{\mu}_{t(j)}}{\hat{\sigma}_{t(j)}}\right)^2 \qquad (4.4)$$

where b_j is the jth observed bond-length with the ligand being element *t*, J is the number of ligands (4 for Tet, 5 for Tbp, and 6 for Oct), $\hat{\mu}_{t(j)}$ and $\hat{\sigma}_{t(j)}$ are the corresponding means and standard deviations of ligand *j*'s element *t* as calculated in the bootstrapping step. This χ^2 probability, P(B), was used as a goodness of fit measure in selecting the set of ligands to compose the final zinc fc-shell from the given potential-ligand list.

The ligand combination with the highest χ^2 probability P(B) was defined as the less biased best zinc fc-shell for later clustering analyses. While this approach identified four-, five-, and six-ligand fc-shells, we mainly explored four-ligand zinc fc-shells in this study, which represented the vast majority (95.7%) of the **final fc-shells** identified.

4.2.1.4 Determine Non-Redundant Set Of Zinc Sites

As the best fc-shell was defined, these ligands were first mapped to the corresponding SEQRES sequence by aligning ATOM record-based sequences to SEQRES sequences. Then for each zinc site, we defined the binding domain as a five-residue extension of the minimum sequence range that includes all ligands identified in the best fc-shell. For example, if the ligand residues positions are 11, 24, 45, and 123 on a protein sequence, the binding domain will be defined as residues 6-128 of the sequence. For ligands that are scattered over multiple chains, we extracted the sequence section of each chain, and consider them together. We removed all redundant domain-ligand combinations, and kept only one with either the best resolution or most recently deposited

date for each redundant group. Then for the non-redundant set, we kept those with a resolution better than (i.e. less than) 3 Å.

4.2.1.5 Iterative Algorithm (IA) for the Mixture Canonical CG Models (Step 3)

With the aim of both identifying the best fitting CG based on angle and bond-length statistics (means and variances) as well as refining those statistics via CG assignment, we performed the following iterative algorithm (IA). This algorithm is in the spirit of an Expectation-Maximization (EM) algorithm [99]. A workflow of this IA process is illustrated in Figure 4.2.

The bootstrapping step served as an initialization for the iteration process. It provided the first guess of the unknown parameters ($\hat{\mu}_s, \hat{\sigma}_s^2, \hat{\mu}_t, and\ \hat{\sigma}_t^2$). Tet, Tbp, Oct, and their minor CGs were used as mixture canonical models for zinc. Our IA algorithm employed a χ^2 probability, P(k), to determine the best fitting CG at each iteration, which was based on the following χ^2 statistic:

$$\chi_s^2 = \left(Y - \hat{\mu}_{s+t}\right)^T C_s^{-1} \left(Y - \hat{\mu}_{s+t}\right) \qquad (4.5)$$

where Y is the observed angle and bond-length vector of a given zinc site, $\hat{\mu}_{s+t}$ ($\hat{\mu}_s\ and\ \hat{\mu}_t$) is the mean vector of corresponding angles and bond-lengths generated from the initialization or previous iteration, and C_s is the covariance matrix of CG model *s*. This formula could handle dependency between variables, where the degree of freedom was derived from the defective rank of the covariance matrix term. And corresponding χ^2 probability was computed as P(k) = 1 - P(χ^2(k) $\leq$ χ^2_{p*s}), where the degrees of freedom k is the same as the rank of the covariance matrix. Again, all permutations of atoms in the initial fc-shell ligand list were considered for every zinc site.

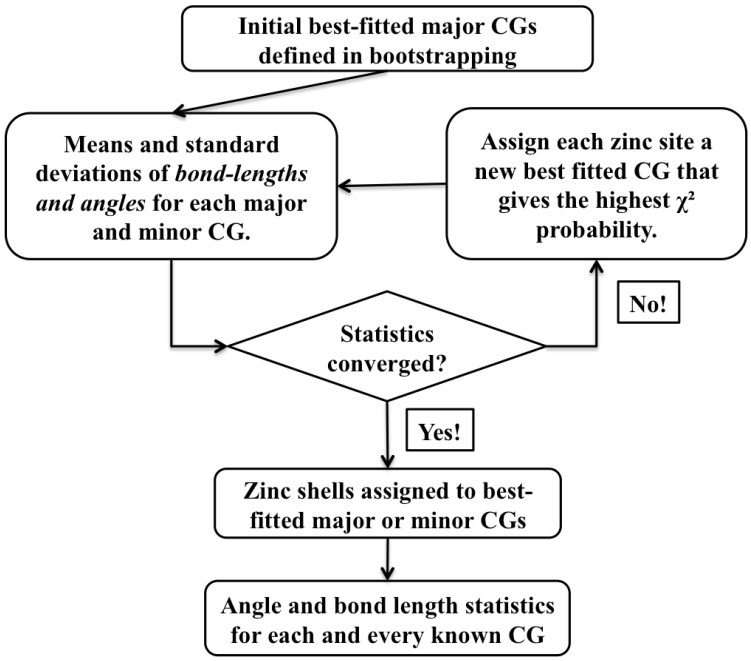


Figure 4.2 Workflow of the IA process

For each zinc site at every iteration, our IA tool defined a new fc-shell as the assigned best-fitted CG with the highest χ^2 probability. Then the IA tool updated the means and variances of angle based on estimates from those zinc fc-shells classified into that CG at the given iteration and using Equations 4.2. Similarly, bond-length statistics were updated for each element via Equation 4.3.

To prevent the actual CG models' angle means drifting dramatically from the ideal values over iterations, we used the means of major CG, $\hat{\mu}_{s,major}$, in the χ^2 calculation for all associated minor CGs. And to prevent any of the CG models to become statistically greedy and attract a large number of "outliers", a pooled angle variance

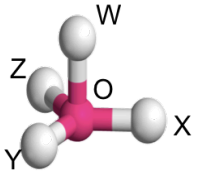
$$\hat{\sigma}^2_{pool} = \frac{\sum_{i=1}^{s} n_i * \hat{\sigma}_i^2}{\sum_{i=1}^{s} n_i} \qquad (4.6)$$

was used for all CG models instead of their individual angle variance, where *s* is the total number of different angles from all CG models, $\hat{\sigma}_i^2$ is the angle variance of angle i, and n_i is the corresponding number of instances for angle i. The angle portion of the C_s, was calculated by multiplying $\hat{\sigma}_{pool}^2$ with a fixed simulated correlation matrix Σ_s, representing the spatial restriction of the ideal CG model *s*. The bond-length portion of the matrix C_s was constructed using element-specific variance $\hat{\sigma}_t^2$ on the diagonal and 0 for everywhere else, because bond-lengths are independent of each other and are also independent from all angle variables. The covariance matrix C_s for each CG model *s* was updated as the $\hat{\sigma}_{po}^2$ and $\hat{\sigma}_t^2$ evolve after each iteration. Due to the existence of compressed angle, we restricted all ligand permutations to have a minimum angle of 68 degrees. Our IA tool repeated the iterative process until all statistics converged, providing each zinc fc-shell with a converging CG classification and final angle and bond-length statistics for later steps of the overall analysis.

The angle correlation matrix (Σ_s) was estimated before the IA process via simulation using an R script and remained the same through the iteration process. One correlation matrix was simulated for each CG separately. Figure 4.3 shows in illustration of angle correlation matrix simulation for Bva. As a starting point for the simulation, our R simulation script set the zinc atom at O (0,0,0), and the ligands at corresponding positions based on bond-lengths $\hat{\mu}_t$ from the bootstrapping and ideal angles μ_s for each CG (Chapter 4.2.1.2). A spherical normal distribution was assumed for each ligand with (0, $\hat{\sigma}_t^2$) on each of the x, y, and z dimensions, where variance $\hat{\sigma}_t^2$ was acquired from the bootstrapping as well. The simulation generated 1000 random and independent Euclidian points for each ligand, composing a sphere around the ideal ligand position W, X, Y, and

Z (Figure 4.3.B). The simulation R script then calculated correlations between angles from the simulated data, and arranged these correlations in a matrix. A raw correlation matrix is shown in Figure 4.3.C. Since angles of the form WOX/WOY/WOZ were structurally identical and were arbitrarily ordered in the matrix, we then smoothed all equivalent positions in the matrix by taking their average. The final correlation for Bva is shown in Figure 4.3.D.

A



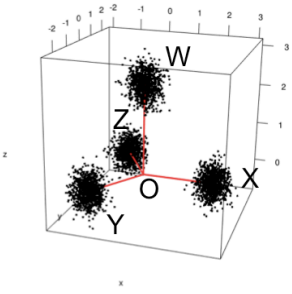
Trigonal Bipyramidal
Vacancy Axial (Bva)

Angle	Ideal Value
WOX	90
WOY	90
WOZ	90
XOY	120
XOZ	120
YOZ	120

C

	WOX	WOY	WOZ	XOY	XOZ	YOZ
WOX	1.00	-0.25	-0.25	-0.01	0.01	0.00
WOY	-0.25	1.00	-0.27	0.01	-0.02	0.00
WOZ	-0.25	-0.27	1.00	0.01	-0.01	0.00
XOY	-0.01	0.01	0.01	1.00	-0.49	-0.50
XOZ	0.01	-0.02	-0.01	-0.49	1.00	-0.50
YOZ	0.00	0.00	0.00	-0.50	-0.50	1.00

B



Angle	Mean	Standard deviation
WOX	90.1	9.0
WOY	90.0	9.0
WOZ	89.8	8.9
XOY	119.5	8.9
XOZ	119.8	9.0
YOZ	119.5	8.9

D

	WOX	WOY	WOZ	XOY	XOZ	YOZ
WOX	1	-0.25	-0.25	0	0	0
WOY	-0.25	1	-0.25	0	0	0
WOZ	-0.25	-0.25	1	0	0	0
XOY	0	0	0	1	-0.5	-0.5
XOZ	0	0	0	-0.5	1	-0.5
YOZ	0	0	0	-0.5	-0.5	1

Figure 4.3 A schematic view of angle correlation matrix simulation for trigonal bipyramidal vacancy axial (Bva). (A) Ideal structure and angles of Bva. (B) A three-dimensional view of 1000 simulated data at each ligand point and their angle statistics. (C) The raw correlation matrix. (D) The final correlation matrix after smoothing.

4.2.2 Separating Zinc Fc-shells into Normal and Compressed Angle Groups Using Random Forest (Step 4)

Upon calculating the final fc-shells in 4.2.1.3, we obtained the smallest angle for each fc-shell and plotted the histogram of the smallest angles (Figure 4.3). As shown in Figure 4.3, there exist a large number of abnormally compressed minimum angles. We denote these angles significantly below 90 degrees as compressed angles (and below 38 as super-compressed). Zinc sites with a compressed angle were treated separately to prevent interference between each other in clustering.

The randomForest package in R (randomForest 4.6-7 in R version 3.0.2) [100, 101]was employed to separate the defined final zinc fc-shells into normal, compressed and super-compressed groups. Features for the random forest analysis included angles, bidentation status, and ligands. Here is an example feature vector with elements of the vector separated by semicolons: "149.3; 85.8; 90.5; 103.6; 121.4; 86.7; 000100; CYS.SG.S; CYS.SG.S; CYS.SG.S; HIS.ND1.N." For four-ligand zinc CGs, the first six elements are angles, which are ordered in 'largest-sorted-middle-opposite' order: first is the largest angle of the six ligand-zinc-ligand angles; followed by the middle four angles, which share one of the two ligands composing the largest angle, sorted from smallest to largest; and last is the angle sharing no ligand with the largest angle. Ideal angles in this ordering of the four-ligand CGs can be found in Table 5.2. This ordering makes the largest angle and the opposite angle the discriminating angles. The next element is a string with six 0/1 digits corresponds to the bidentation status of the six angles, where 0 means no bidentation, and 1 means bidentation of that angle. Ligands make up the last four elements, and are represented as residue.atom.element. The first two ligands

comprise the largest angle, ordered alphabetically. And the rest two ligands are ordered alphabetically as well. We sorted angles and ligands in this way so that they are comparable through all zinc fc-shells without introducing any artificial scrambling.

The smallest angle was used to separate sites as super-compressed (<38 degrees), compressed (38-58 degrees) or normal (> 68 degrees) groups for training. The default settings of random forest were used to build the classifier that was then be applied to the overlapping part of the data, where the smallest angle is between 58° and 68°, as well as the training data itself.

4.2.3 Clustering Zinc Fc-Shells Using K-Means and Assigning Known and Novel CGs to Each Cluster

4.2.3.1 Determine Optimal Cluster Number K (Step 5)

K-means [102] is one of the most popular clustering methods, and is good at clustering numeric data. As with all clustering methods, determining the numbers of clusters (k) is crucial for achieving successful and meaningful clustering results. We approached this problem by testing the stability and biological relevance of the final cluster centers while varying k. The k-means function from the stats package in R was used with default settings, except that iter.max was set to 30. By default, the package uses the Hartigan-Wong algorithm. For each value of k from k=1 to k=30, we ran 500 repetitions of k-means clustering with different cluster initializations. For each value of k, we calculated the average of the sum of absolute differences of all pairwise best matching cluster centers:

$$D_k = \frac{1}{K * \binom{R}{2}} \sum_{q=p+1}^{R} \sum_{p=1}^{R-1} \sum_{j=1}^{K} \sum_{i=1}^{A} |ca_{pj,i} - ca_{qj,i}| \qquad (4.7)$$

where *i* is the angle position, *j* is the matching cluster numbers between two repetitions, *A* is the total number of angles (*A*=6 for four-ligand CGs), *K* is the number of clusters as the k in k-means, *p* and *q* are the repetition numbers, *R* is the number of repetitions (500), and $ca_{pj,i}$ is the cluster center angle at position *i* and clustered as cluster *j* in repetition *p*. The sum of absolute difference measures the distance of the cluster centers from each other between the *R* repetitions. We took the max(D_k) - D_k as the final measure so that a larger value is preferred.

We also measured the average Jaccard index of all the pairwise best matching cluster centers:

$$J_k = \frac{1}{K * \binom{R}{2}} \sum_{q=p+1}^{R} \sum_{p=1}^{R-1} \sum_{j=1}^{K} J(S_{jp}, S_{jq}) \qquad (4.8)$$

where S_{jp} is the set of zinc fc-shells clustered as cluster j in repetition p, and

$$J(S_{jp}, S_{jq}) = \frac{|S_{jp} \cap S_{jq}|}{|S_{jp} \cup S_{jq}|} \qquad (4.9)$$

The average Jaccard index measures how well the same set of zinc sites are clustered into the same cluster between repetitions. It can take a value between 0 and 1, with a smaller value indicating better performance.

Two other metrics were used for measuring the biological relevance of the clusters: structure-function correlation rho and p-value, with more detailed description in Chapter 4.2.4.

4.2.3.2 Clusters Assignment (Step 6)

After the optimal number of clusters was determined for the normal and compressed groups separately, we re-ran k-means with the optimal k to obtain the final cluster results. We assigned a best-fitted CG to each cluster by 1) comparing the cluster

centers with ideal angles of the CG models; 2) finding the representative zinc fc-shell that is the closest to the cluster center, and checking its 3D structure; and 3) calculating an average χ^2 probability for each cluster on each canonical CG model using Equation 5 and statistics acquired from the IA process. For zinc sites with a compressed angle, we left out the compressed angle in calculating the χ^2 probabilities to minimize the effect of the compressed angle in comparing to canonical CGs. The χ^2 probabilities were used as a mathematical characterization of each cluster to each canonical CG. Assignments of clusters were based on cluster centers, 3D structures, and χ^2 probabilities together.

4.2.4 Functional Analysis (Step 7)

4.2.4.1 Acquire Functional Annotations From InterProScan

We ran InterProScan 5.7.48.0 [103] using the current versions of its member databases on the non-redundant sequences previously determined. We retained only those results with an InterProScan (IPR) annotation mapping and overlapping at least one ligand.

4.2.4.2 Derive and Evaluate Consistency of CG-Based Structure and Sequence-Based Function Annotation Relationships Between K-Means Clusters

We calculated both CG-based structural and sequence-based functional distance matrices between pairwise k-means clusters and then compared these two matrices with respect to two different measures of consistency: hierarchical clustering and spearman correlation. To construct the CG-based structural distance matrix, we calculated a root-mean-square-deviation-like (RMSD-like) distance matrix between each cluster based on angles:

$$M_{struct} = \begin{pmatrix} m_{11} & \cdots & m_{k1} \\ \vdots & \ddots & \vdots \\ m_{1k} & \cdots & m_{kk} \end{pmatrix}, with\ m_{xy} = \sum_{q=1}^{s(y)} \sum_{p=1}^{s(x)} \sqrt{\frac{1}{A} \sum_{i=1}^{A} (a_{xp,,i} - a_{yq,i})^2} \qquad (10)$$

where k is the clustering number k in k-means, A is the number of angles (A=6 for four-ligand CGs), and s(x) and s(y) are the size of cluster x and y, $a_{xp,i}$ is the ith ($1 \leq i \leq A$) angle of fc-shell p in cluster x ($1 \leq p \leq s(x)$).

To construct the sequence-based function annotation distance matrix, we first calculated the proportional representation of functional annotation from each cluster:

$$prop_{tn} = \frac{number\ of\ entries\ in\ cluster\ n\ annotated\ as\ term\ t}{size\ of\ cluster\ n} \qquad (11)$$

$prop_{tn}$ is normalized across all clusters so that $\sum_n prop_{tn} = 1$. We then constructed a k*k (k being the clustering number k in k-means) matrix for each annotation t:

$$M_t = \begin{pmatrix} m_{11} & \cdots & m_{k1} \\ \vdots & \ddots & \vdots \\ m_{1k} & \cdots & m_{kk} \end{pmatrix}, where\ m_{xy} = \min\left(prop_{tx}, prop_{ty}\right)^2 \qquad (12)$$

Next, the inter-cluster values across all annotations *t* are summed to create the matrix M_{sim} and then normalized by the max value in M_{sim} to create M_{sim_norm}, representing functional similarity between clusters. Finally, we took $1 - M_{sim\_norm}$ as the distance matrix M_{func}. In other words, we represented functional annotations across cluster members as a rational vector space of proportional functional annotations, which we then transformed into a pseudo-continuous metric space represented by the resulting distance matrix M_{func}. This works much better than a covariance or correlation matrix, since the large number of zero proportions are ignored and not interpreted in terms of functional similarity or dissimilarity.

In our R script, we calculated Spearman correlations of the between-cluster structural and functional distances (m_{11} … m_{kk}) and computed rho's and p-values computed for k=3 to 30 as biological validation in selecting the optimal k. Ward's hierarchical agglomerative clustering was constructed using the standard hierarchical clustering function in the R stats package for structural and functional distance matrices separately. We then compared the two distance matrices' hierarchical dendrogram and Spearman's rank correlation.

4.2.4.4 Determine Functional Enrichment of Normal and Compressed Groups

Using the normal and compressed classification to designate a "group of interest" compared to all of the zinc sites with an annotation, we used a hypergeometric test to determine whether any of the InterProScan annotations or EC number annotations based on the mapping of InterProScan annotations to KEGG pathways[34] were enriched in either group. For EC numbers, any zinc site that did not have an EC number was assigned 0.

4.3 Results

4.3.1 Low Variability in Bond-lengths Versus High Variability in Bond Angles and the Existence of Compressed Angles

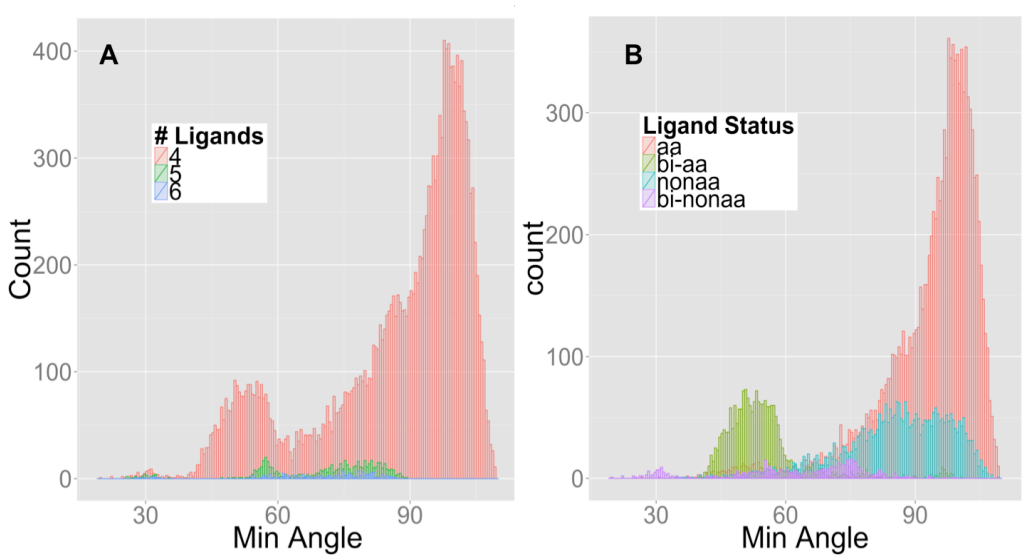


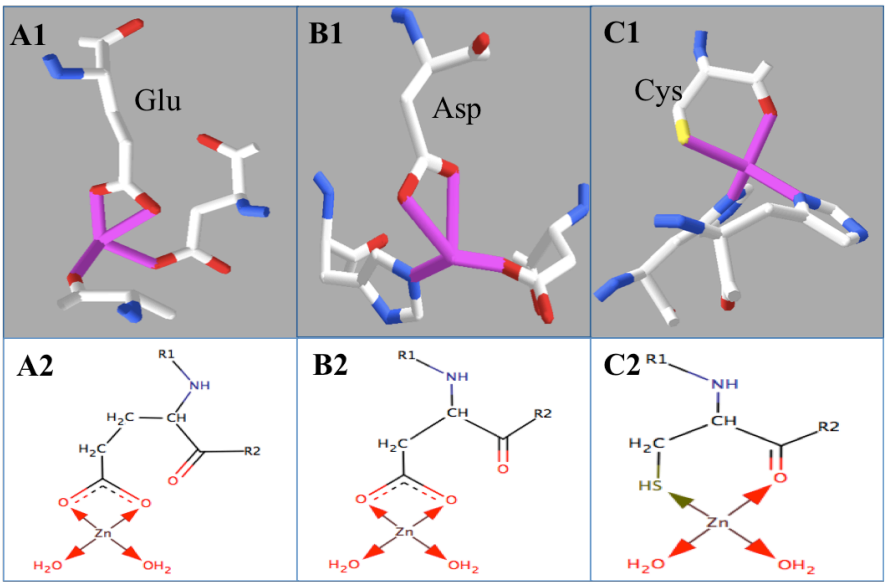
Figure 4.4. Histogram of minimum angles with respect to: A) the number of ligands in the fc-shells, and B) ligand type. aa represents standard amino acid, nonaa represents non-standard amino acid or any substrates from the protein, and bi represents bidentation.

7878 PDB entries were detected to have at least one zinc ion in the protein from wwPDB downloaded on Mar 13, 2013. From those entries, we identified a total of 17,135 four-ligand, 602 five-ligand, and 169 six-ligand non-cluster zinc fc-shells. In our initial attempt in analyzing zinc metalloproteins assuming Tet, Tbp, Oct, and their associated minor models, we observed abnormally high ligand-zinc-ligand angle variances and very low zinc-ligand bond-length variances in classified canonical CGs at the same time, which was consistent with several other studies [19, 28, 89]. From these high angle variances it appeared there are "outlier" CGs that do not belong to any known canonical CGs. Also, the histogram (Figure 4.4) of smallest angles from each zinc site revealed a significant number of sites with compressed (< 58 degrees) angles. The peak at 109 degrees is the contribution from Tet, and the shoulder peak at 90 degrees is from Tbp,

Oct and their associated minor CGs. However, none of the known CG models can account for the histogram peaks at 32 degrees and 53 degrees. The likelihood that these sites are artificial is low given that 1) there is a non-trivial number of zinc sites in this range, 2) the histograms around these peaks appear normally distributed, and 3) they occur in zinc fc-shells with 4, 5, and 6 ligands.

In an attempt to characterize the possible source of the compressed and super-compressed minimum angles, we characterized the two ligands comprising the smallest angle by bidentation status and inclusion/exclusion of the 20 standard amino acids (Figure 4.4.B). Bidentation occurs when two ligating atoms are from the same amino acid residue (e.g. the two oxygen atoms of one carboxylate from glutamate). Our analysis showed that 83.0% of the compressed angles could be explained by coordination by bidentate ligands and these bidentation patterns affect overall clustering ability with functional significance (Table 4.3, 4.4). Figure 4.5 pictorially shows the common bidentation patterns and their frequencies observed in zinc metalloproteins. Some of the bidentation patterns have been observed, such as ligation by carbonyl oxygens [104], or theorized to occur from simulation, such as bidentation by cysteine thiol and backbone carbonyl oxygen [105-107]; however, their frequency had not been systematically analyzed in the wwPDB.

Classifying a zinc fc-shell with a compressed angle into any of the previously canonical CG models will either create an outlier or add significant variance to subsequent analyses. Thus, we decided to separate zinc sites containing compressed angles from normal zinc sites.



	Ligand	Count	Percentage	Example
A	Glutamate	935	33.7%	2E4T
B	Aspartate	796	28.7%	1RTQ
C	Cysteine	153	5.5%	4FGL

Figure 4.5 Three most prevalent zinc bidentation of standard amino acids in the zinc metalloprotein, with real structures on top and schematic structures on the bottom.

4.3.2 Angle Correlation Matrix and IA Statistics

The main reason for using fixed angle correlation matrix with evolving angle variance instead of a simple evolving angle covariance matrix is because we needed to use pooled variance at each round so that there is no greedy CG or angle position that would take over the "outlier" structures. In addition, we had to represent the angle in a specific ordering while the angles in an ideal CG are identical and interchangeable, the correlation matrix could help make sure all angles are mathematically equivalent.

Since the mean and variance used for the angle matrix simulation was derived from bond-length statistics only, a by-product of the correlation matrices is an estimate of angle variance introduce by the bond-length variation mechanism. It turned out that the angle standard deviation is between 8 to 10 degrees, regardless of the CG and angle positions. The final angle statistics from the IA step are shown in Table 4.1. For most of the CGs, the angle standard deviations are within or below the 10 degrees limit, which means our IA tool effectively capture and separated the actual CGs. The only angle standard deviations that are a little higher is Bva and Bvp. It suggested that even after removing structure with angles smaller than 68 degrees (See methods 4.2.1.5), there were still some impurities in the data that caused a higher in-class variation. Those variations that cannot be explained by bond-length mechanism (intra-model variation) are most likely to be cause by new or aberrant CGs (inter-class variation).

Originally, this IA process itself (without removing angles smaller than 68 degrees) was what we used in classifying a metal's CG. But as we realized the existence of compressed angle that could greatly complicate the classification, we used this step to acquire angle and bond-length statistics that would provide us a useful mathematical guidance in assigning k-means clusters in subsequent steps. For IA process, we have to make assumptions on the CG models upfront for the probability calculation. Thus, the model-free methods was developed and applied in this study to better characterize the structures of zinc metalloproteins. K-means clustering algorithm was to truly learn the structures from the data itself.

Table 4.1 The final ligand-zinc-ligand angles statistics from IA.

Model	Count	Ideal Angle (°)	Mean Angle (°)	Standard Deviation
Tetrahedral (Tet)	13321	109.5	109.14	8.94
Trigonal Bipyramidal (Tbp)	564	90	90.16	7.98
		120	119.46	9.23
		180	167.39	8.18
Trigonal Bipyramidal Vacancy Axial (Bva)	1278	90	94.41	11.53
		120	118.55	14.98
Trigonal Bipyramidal Vacancy Planar (Bvp)	1261	90	91.83	10.74
		120	119.93	10.68
		180	159.06	16.34
Octahedral (Oct)	337	90	89.95	6.42
		180	170.07	5.07
Square Planar (Spl)	265	90	89.98	7.36
		180	161.06	7.62
Square Pyramidal (Spy)	784	90a	91.38	6.81
		90p	89.91	7.26
		180	168.90	6.32
Square Pyramidal Vacancy (Pyv)	779	90a	94.86	7.21
		90p	91.91	8.26
		180	164.86	7.82
Overall	18589	-	-	9.25

Table 4.2. The final zinc-ligand bond-length statistics from IA.

Zn–X	Count	Mean Bond Distance (Å)	Standard Deviation (Å)
Zn-S	30846	2.34	0.14
Zn-N	25910	2.12	0.15
Zn-O	24270	2.17	0.24
Zn-Cl	405	2.34	0.28
Zn-P	218	2.97	0.09

4.3.3 Separation Of Zinc Fc-Shells Into Normal And Compressed Groups

As demonstrated above, there exists a significant amount of compressed (and some super-compressed) angles between zinc fc-shell ligands. Due to the overlapping distribution of the normal and compressed angles and the ligand and bidentation

propensities of the ligands comprising these angles, we developed a random forest classifier to deconvolute this overlap. Then, we used this classifier to separate zinc sites into normal and compressed groups based on three key factors: angles, bidentation status, and ligand residue type. The training data consisted of 16,375 sites (14,210 normal, 2,087 compressed, 78 super-compressed) initially categorized based on the smallest angle. The out-of-bag error rate for the training data was 0.00 for the normal and compressed groups, and 0.06 for the super-compressed group. Importance measures showed the most important feature is angle 2 (with a score of 1836), followed by bidentation status (score 859) and angle 6 (score 279). The reason that angle 2 is the most important feature is because it is most likely to be the smallest angle due to the 'largest-sortedMiddle-opposite' ordering of angles used. Angle 1 is always the largest angle and is therefore nearly impossible to also be the smallest (except for special cases where all angles are exactly equal). Angle 6 is the angle opposite of angle 1 (e.g. has no ligand atoms in common with angle 1), which means the smallest angle could be in this position. So the smallest angle is the variable that has the most significant effect on importance, regardless of what position it is in. The bidentation status of ligands in the site also showed its importance as expected from the histogram Figure 4.3.

Sorting the six angles by 'largest-sortedMiddle-opposite' makes them comparable across all geometries without introducing artificial scrambling. This was necessary for robustness in many of the analyses. As shown in Table 4 of ideal angles in this ordering, angle 1 and angle 6 in combination are highly distinct for different CGs. The middle four angles should be very close to each other except in the case of Bva.

After the removal of redundant sites, 6,199 four-ligand zinc fc-shells were left for subsequent analyses. Applying the random forest classifier resulted in 4,845, 1,303, and 51 normal, compressed, and super-compressed fc-shells respectively.

4.3.4 K-Means Clustering

In an initial failed attempt to cluster zinc fc-shells using random forest (results not shown) the ligand type and bond-length showed very little influence in determining meaningful CGs, while the ligand-zinc-ligand angles and bidentation status were more important. Therefore, we applied k-means clustering to the angles to generate clusters of zinc sites. Note that clustering was performed on the normal and compressed zinc sites separately, otherwise the clustering was unstable to separate (Figure 4.8, Table 4.3, and 4.4).

Two measures were used to assess the stability of resulting clusters: sum of absolute differences and Jaccard index. The sum of absolute differences measures the differences between cluster centers over multiple times of clustering. The Jaccard index evaluates the agreement of the set of actual zinc fc-shells that are classified into the same cluster over multiple times of clustering. The other two measured biologically validate the optimal k: Spearman's rho and p-value between structural distances and functional distances of cluster pairs. In order to visualize the comparisons between all four values, we graphed the max sum of absolute differences minus each actual sum of absolute differences, and the negative log of the p-value. We expect the "true" k to have a local, simultaneous maximum for all four measures together. Figure 4.6.A shows how these four measures vary with respect to k values for the normal group. k=10 is consistent

local maximization of all four measures. Figure 4.6.B shows the same measures for the compressed group. In this case, k=8 is as the local maximization of all four measures.

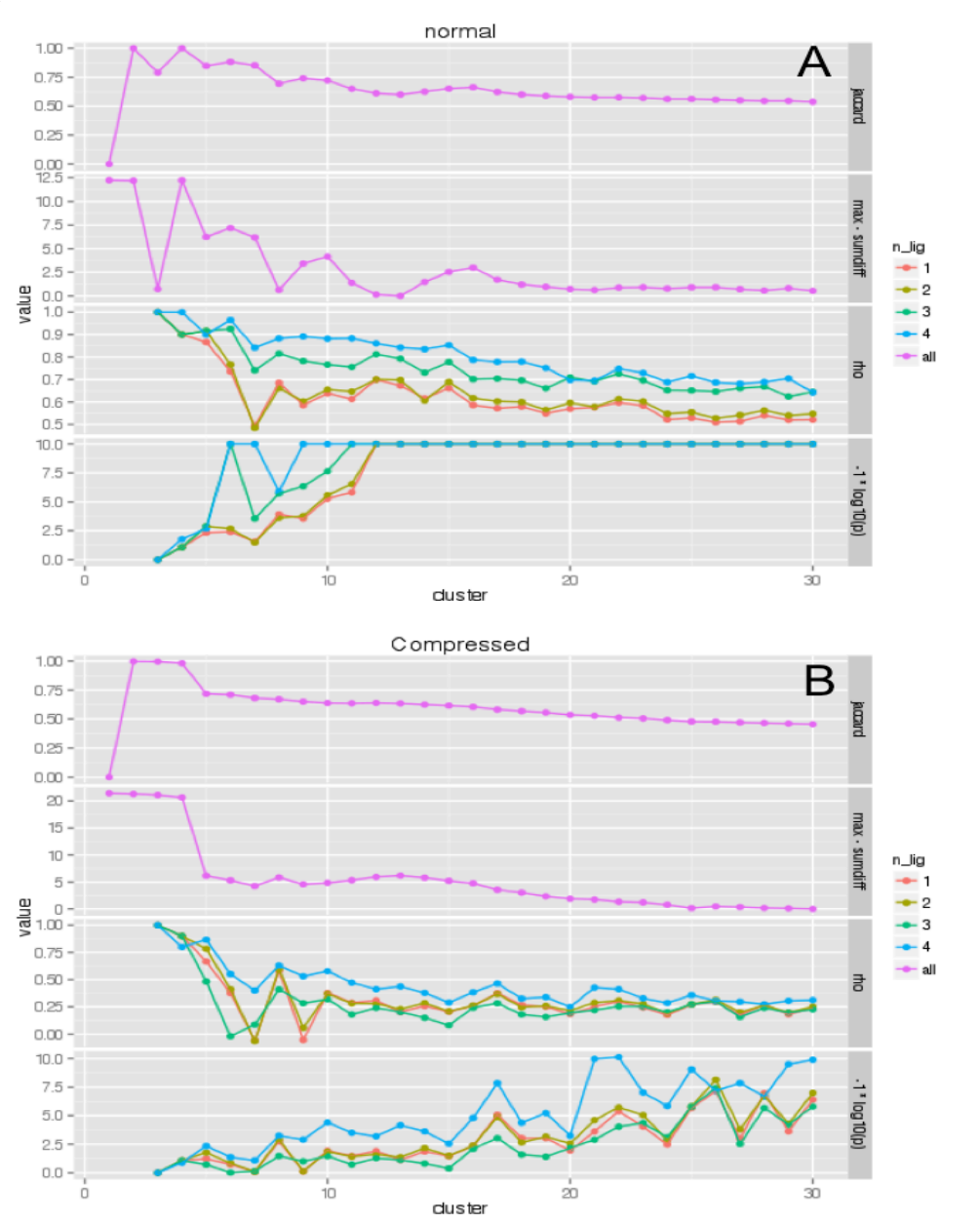


Figure 4.6 Four measures of k value in K-means clustering for the normal (A) and compressed (B) group.

Table 4.3 Mean/standard deviation, average χ^2 probability, and CG assignment for each cluster, normal group k=10.

Cluster	Size	Angle 1	Angle 2	Angle 3	Angle4	Angle 5	Angle 6	Tet	Bva	Bvp	Pyv	Spl	Assignment
1	331	150.0 ± 5.6	85.8 ± 7.0	93.8 ± 5.4	100.8 ± 4.4	109.2 ± 5.3	98.9 ± 7.1	0.028	0.090	0.193	0.265	0.125	Pyv distorted
2	741	123.4 ± 4.2	93.8 ± 4.9	101.8 ± 3.7	108.4 ± 3.9	115.2 ± 3.8	112.4 ± 4.6	0.543	0.042	0.015	0.006	0.000	Tet
3	213	135.5 ± 8.1	80.4 ± 7.1	91.1 ± 7.8	107.8 ± 8.2	122.3 ± 6.4	86.3 ± 9.6	0.033	0.197	0.193	0.125	0.017	Bva
4	381	167.4 ± 6.6	81.6 ± 6.0	87.4 ± 5.0	92.6 ± 4.5	99.0 ± 6.3	90.8 ± 8.6	0.004	0.044	0.399	0.683	0.445	Pyv
5	205	138.8 ± 6.7	84.6 ± 7.6	92.8 ± 7.1	102.5 ± 6.2	113.8 ± 8.1	120.5 ± 8.2	0.096	0.071	0.047	0.013	0.002	Tet distorted
6	1050	116.0 ± 2.9	103 ± 3.1	106.3 ± 2.1	108.9 ± 1.9	111.8 ± 2.1	110.5 ± 3.2	0.931	0.011	0.004	0.002	0.000	Tet
7	853	119.4 ± 3.0	100.8 ± 3.8	107.0 ± 2.9	111.2 ± 2.6	114.8 ± 2.5	101.3 ± 4.3	0.769	0.064	0.017	0.007	0.000	Tet
8	383	168.0 ± 6.7	80.4 ± 5.7	87.7 ± 4.1	93.2 ± 3.8	100.0 ± 5.6	116.9 ± 8.5	0.071	0.373	0.685	0.424	0.097	Bvp
9	165	166.8 ± 8.1	79.6 ± 5.6	87.1 ± 3.5	92.3 ± 3.2	99.7 ± 6.2	155.3 ± 11.0	0.009	0.063	0.461	0.585	0.564	Spl
10	523	131.1 ± 4.9	94.9 ± 5.4	102.3 ± 3.9	108.5 ± 4.2	115.7 ± 4.8	96.7 ± 6.3	0.218	0.149	0.082	0.070	0.009	Tet

Table 4.4 Mean/standard deviation, average χ^2 probability, and CG assignment for each cluster, compressed group k=8.

Cluster	Size	Angle 1	Angle 2	Angle 3	Angle4	Angle 5	Angle 6	Tet	Bva	Bvp	Pyv	Spl	Assignment
1	186	128.2 ± 8.2	53.7 ± 6.1	92.1 ± 8.7	105.6 ± 6.1	115.0 ± 6.1	90.8 ± 9.4	0.160	0.289	0.150	0.072	0.012	Bva with compressed 90
2	141	155.9 ± 8.6	57.9 ± 6.4	86.6 ± 7.5	98.8 ± 6.4	112.0 ± 9.6	134.0 ± 10.3	0.092	0.229	0.206	0.149	0.064	Spl with compressed 90
3	275	153.0 ± 7.0	55.2 ± 5.4	88.2 ± 5.8	98.3 ± 5.2	105.7 ± 6.0	103.2 ± 9.2	0.102	0.287	0.226	0.263	0.092	Distorted Pyv with compressed 90
4	84	128.5 ± 9.9	80.5 ± 7.6	92.3 ± 8.2	105.4 ± 9.5	116.4 ± 8.5	51.5 ± 4.8	0.074	0.159	0.090	0.062	0.015	Tet with compressed 109
5	126	130.8 ± 9.9	53.3 ± 6.3	75.2 ± 6.3	85.9 ± 6.7	100.7 ± 9.3	91.2 ± 11.9	0.031	0.146	0.154	0.184	0.060	New!
6	91	157.1 ± 10.6	54.8 ± 7.2	77.0 ± 8.2	105.1 ± 12	129.1 ± 11.1	92.4 ± 14.5	0.042	0.073	0.061	0.056	0.027	Pyv with compressed 90
7	53	159.8 ± 9.6	79.1 ± 9.0	86.7 ± 6.8	93.8 ± 6.8	103.1 ± 10.3	55.0 ± 6.3	0.022	0.313	0.313	0.362	0.330	Pyv with compressed 90
8	209	139.6 ± 8.2	52.7 ± 5.6	83.4 ± 7.7	96.8 ± 7.0	111.1 ± 9.1	118.8 ± 6.7	0.112	0.133	0.197	0.050	0.005	Distorted Bvp with compressed 90

The angle statistics and average χ^2 probabilities for the normal group are shown in Table 4.3, and one representation is chosen for each cluster that is closest to the cluster center (Figure 4.7). By comparing the angle means of each cluster to ideal angles, angle 1 of cluster 4, 8 and 9 appeared to be equivalent to 180 degrees, due to the folded normal distribution effect. Their angle 6 was equivalent to 90, 120, and 180 degrees, respectively. χ^2 probability serves as a mathematical characterization of the cluster with respect to specific canonical CGs, and the three-dimensional structure of the centroid zinc site is the visualization of the cluster. Based on all the evidences, cluster 4, 8 and 9 were assigned as Pyv, Bvp, and Spl. Similarly, cluster 1 was assigned as Pyv, but distorted. Cluster 3 was assigned as Bva. Clusters 2, 5, 6, 7, and 10 were all subclasses of Tet. In fact, all of the canonical CGs could be assigned to the same corresponding cluster(s) by considering only their maximal cluster average χ^2 probabilities for assignment.

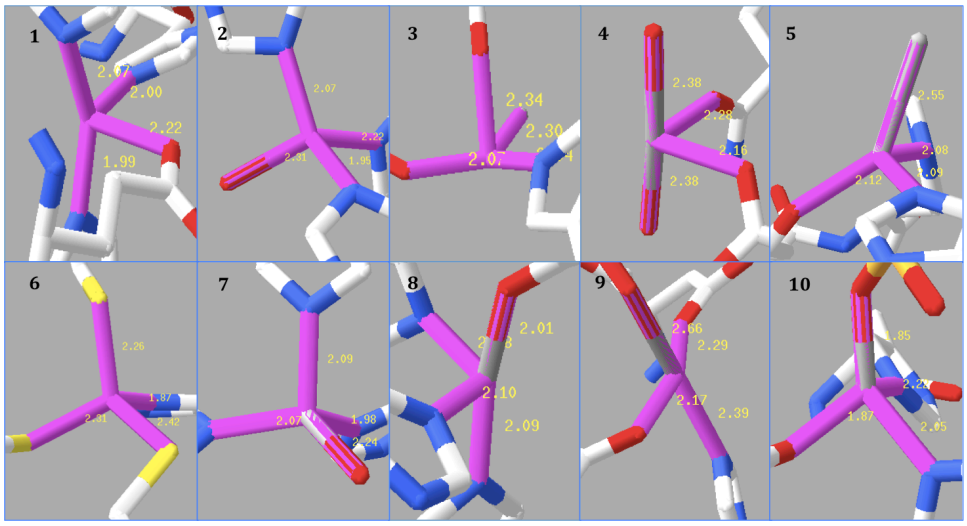


Figure 4.7 Three-dimensional structures of normal cluster representatives.

All angles have relatively tight standard deviations. According to the angle correlation simulation in Chapter 4.3.3, the angle standard deviation introduced by bond-length variation itself is about 8-10Å, which is similar or larger than most of the normal group angle standard deviation. It indicates that the k-means clustering method successfully separated different CGs, as no significant inter-class variation was observed. Also, since many the cluster standard deviations are only half of the value from the simulation, it suggests that some sub-class (sub-CG) clusters were being detected. That agreed with our assignment as well.

Table 4.4 shows the angle statistics and average χ^2 probabilities of the compressed group. χ^2 probabilities were assessed without considering the compressed angles, so that these aberrant structures could be related to canonical CGs with minimum effect from the compressed angles. Even by leaving out the compressed angle in calculating χ^2 probabilities, most of the average χ^2 probabilities are much lower than the normal group, which confirmed they should not be directly classified into any of the canonical CGs. In contrast to the normal group, canonical CG assignment cannot simply use the maximal cluster average χ^2 probabilities. In fact, such a simplistic assignment approach would have mis-assigned canonical CGs for five out of the eight compressed clusters. There is also no highest probability on Tet, because Tet is the most geometrically symmetric structure, and having a compressed angle tends to disrupt this balance more than others CGs. By using all three pieces of information, most clusters can be viewed as distorted forms of the canonical CGs with one of the angles compressed. As for cluster 5, it does not resemble any of the canonical CGs at all, except maybe a highly distorted Pyv, where

it has three ligands on the same plane but very close to each other, and the fourth ligand-zinc bond perpendicular to that plane.

When k-means is used on both normal and compressed group together instead of separately, stability tests show k=10 and k=14 are the potential optimal clustering numbers. On one hand, the Spearman rho starts from a negative number as shown in Figure 4.8, indicating a much weaker structure-function relationship through clusters if we were to combine everything together. On the other hand, angle statistics (Table 4.5, 4.6) show that all standard deviations, especially those with a compressed angle (cluster 4, 5, 9, and 10 in Table 4.3, and cluster 2, 3, 5, 9, and 12 in Table 4.4), are higher than when handling them separately. As shown in Table S6, the canonical CGs, Spv and Bvp, are very likely to be mixed together in cluster 8 when using k=10. Its discriminating position, Angle 6, is roughly the average of 90 degrees (Spv) and 120 degrees (Bvp) and the standard deviation is much higher compared to the other five angles. When using k=14 as shown in Table 4.4, Spv and Bvp can be separated into cluster 7 and 13 respectively. But the discriminating angle 6 of both clusters have their means further from their ideal angles and the associated standard deviations are relatively high compared to when handling them separately (Table 4.1, cluster 4 and 8). Restated, more zinc sites are misclassified and inappropriately associated if we cluster all zinc sites together rather than clustering zinc sites with all normal angles or with at least one compressed angle separately.

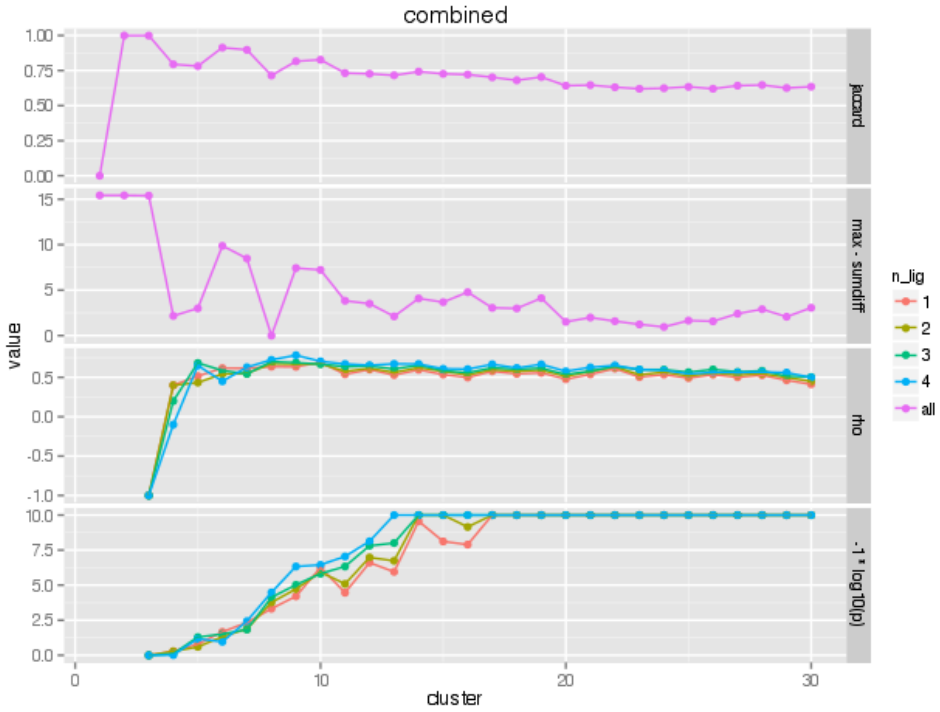


Figure 4.8 Four measures of the unstable K-means clustering of normal+compressed zinc fc-shells.

Table 4.5 Angle statistics of k-means clustering on normal+compressed zinc fc-shells, k=10.

Cluster	Size	Angle 1	Angle 2	Angle 3	Angle4	Angle 5	Angle 6
1	311	165.0 ± 9.0	77.0 ± 8.6	87.2 ± 4.7	93.5 ± 4.5	101.0 ± 7.3	145.5 ± 14.5
2	483	146.5 ± 6.3	86.1 ± 6.8	94.0 ± 5.4	102.2 ± 5.3	112.4 ± 6.8	98.5 ± 9.6
3	783	127.0 ± 5.3	95.6 ± 6.3	103.3 ± 4.8	110.3 ± 4.2	116.9 ± 4.6	95.8 ± 6.4
4	371	135.3 ± 12.7	53.3 ± 10.4	87.4 ± 12.0	104.9 ± 8.5	117.8 ± 10.8	92.4 ± 9.8
5	209	127.5 ± 13.1	51.9 ± 7.3	68.4 ± 11.1	84.8 ± 8.8	100.2 ± 11.1	88.6 ± 14.9
6	1840	117.3 ± 3.3	101.9 ± 3.7	106.4 ± 2.5	109.6 ± 2.3	112.8 ± 2.4	108.0 ± 4.6
7	684	126.7 ± 5.9	91.6 ± 6.2	100.0 ± 4.6	107.0 ± 4.6	115.1 ± 5.0	114.6 ± 6.0
8	694	167.9 ± 6.2	80.8 ± 6.9	87.9 ± 4.6	93.2 ± 4.3	99.6 ± 5.8	99.9 ± 12.4
9	589	149.5 ± 9.6	55.2 ± 8.0	85.0 ± 8.0	98.4 ± 7.0	110.7 ± 10.0	116.1 ± 10.7
10	235	141.5 ± 17.5	73.2 ± 13.8	87.2 ± 11.3	101.2 ± 10.5	113.9 ± 12.2	55.8 ± 10.0

Table 4.6 Angle statistics of k-means clustering on normal+compressed zinc fc-shells, k=14.

Cluster	Size	Angle 1	Angle 2	Angle 3	Angle4	Angle 5	Angle 6
1	349	149.1 ± 5.9	87.0 ± 6.8	94.6 ± 5.0	101.3 ± 4.2	109.4 ± 5.0	98.9 ± 7.0
2	285	130.5 ± 9.0	50.7 ± 9.3	90.4 ± 9.4	103.1 ± 6.9	113.3 ± 7.0	95.6 ± 9.8
3	524	148.7 ± 8.9	54.8 ± 7.1	85.4 ± 7.5	97.8 ± 6.7	109.6 ± 9.4	117.4 ± 10.7
4	257	134.4 ± 6.5	82.5 ± 7.6	94.5 ± 7.3	108.9 ± 7.0	122.1 ± 5.8	89.7 ± 8.2
5	189	126.4 ± 12.6	51.5 ± 7.0	67.2 ± 11.0	84.2 ± 8.8	99.5 ± 10.9	86.6 ± 16.8
6	773	126.3 ± 5.3	97.4 ± 4.5	104.3 ± 3.7	110.0 ± 3.9	115.7 ± 3.8	97.9 ± 5.5
7	442	168.1 ± 6.3	80.7 ± 5.9	88.1 ± 4.0	93.4 ± 3.8	99.8 ± 5.4	114.2 ± 9.1
8	951	120.9 ± 3.6	96.6 ± 4.4	103.2 ± 3.2	108.5 ± 3.4	114.0 ± 3.4	112.6 ± 3.9
9	162	135.1 ± 15.2	75.2 ± 13.3	90.8 ± 9.2	102.5 ± 9.2	115.2 ± 9.9	52.8 ± 8.4
10	300	134.6 ± 6.4	87.2 ± 6.8	96.3 ± 5.3	104.2 ± 5.0	115.0 ± 6.6	117.1 ± 7.6
11	1224	116.1 ± 2.8	103.4 ± 3.1	107.3 ± 2.1	110.0 ± 2.0	112.5 ± 2.1	106.7 ± 4.1
12	144	156.6 ± 10.0	54.6 ± 8.9	73.5 ± 11.0	106.1 ± 12.1	127.6 ± 11.5	86.8 ± 15.2
13	385	166.1 ± 7.5	79.3 ± 8.3	87.1 ± 5.3	92.6 ± 4.6	99.4 ± 6.6	86.3 ± 10.5
14	214	165.3 ± 8.7	76.4 ± 9.1	87.0 ± 4.7	93.6 ± 4.6	101.7 ± 7.7	152.7 ± 11.3

4.3.5 Functional Analysis

To assess how the CG structures might influence the functional characteristics of zinc sites, the distances between clusters were calculated from both the ligand-zinc-ligand bond angles and InterProScan annotations that overlap a zinc-ligand. These distances were compared using Spearman's rank correlation coefficient rho and p-value.

For k=10 normal group, the correlation ranged from 0.6 to 0.9 depending on the number of ligands required in the overlap between zinc binding sites and annotation sites identified by InterProScan. This high level of correlation implies there is a definite link between the coordination geometry and the functional properties of a given zinc size. This is expected based on the sequence-structure-function tenet of structural biology; however, it is still beautiful to see.

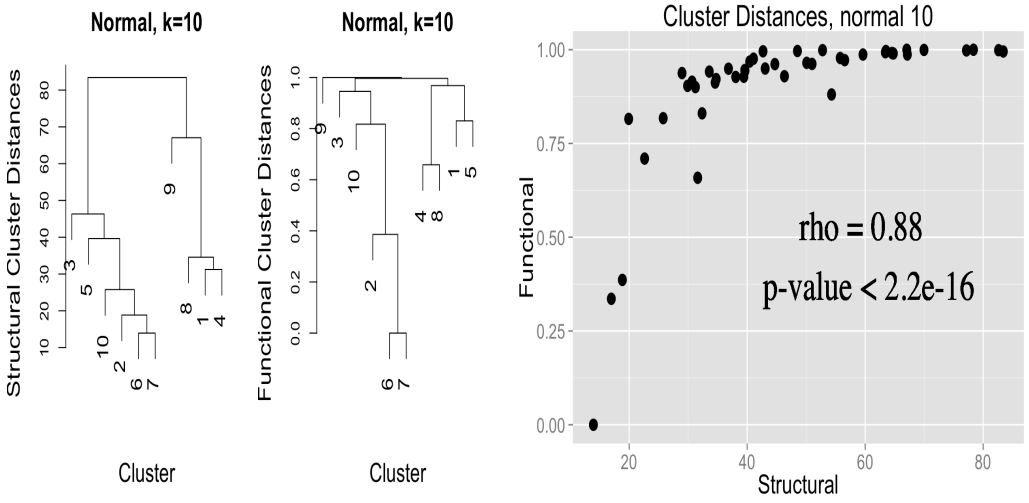


Figure 4.9. Hierarchical dendrogram (A, B) and Spearman's correlation (C) of structural and functional distances for k=10 in the normal group.

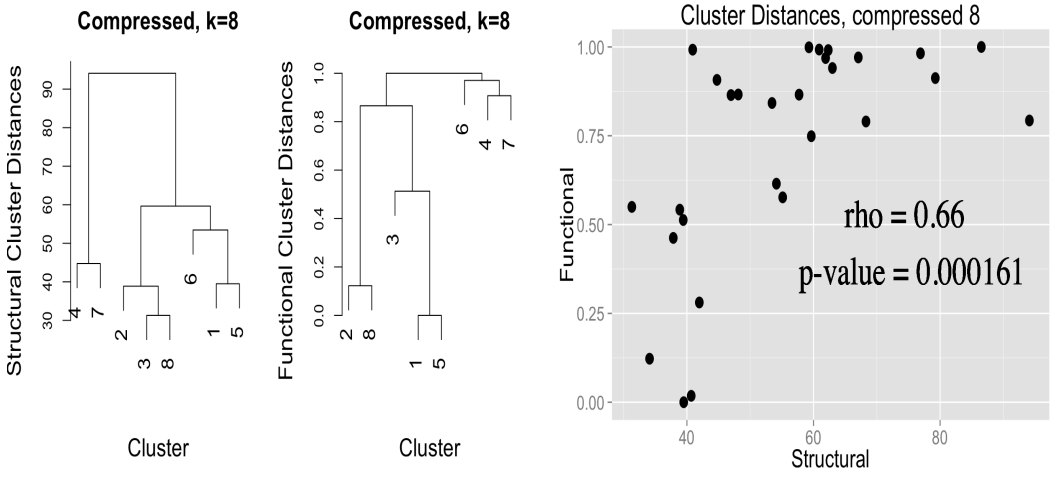


Figure 4.10 Hierarchical dendrogram (A, B) and Spearman's correlation (C) of structural and functional distances for k=8 in the compressed group.

Figure 4.9 shows the comparison of the dendrograms constructed from structural (Panel A) and functional (Panel B) distances for the normal group. Both structural and functional information created a hierarchical dendrogram cluster comprising normal k-means clusters 2 (nk2), nk5, nk6, nk7, and nk10 together, which are all Tet subclasses. Structurally, Bva (nk3) is the next closest k-means cluster to the Tet super-cluster; while functionally, Bva is closer to the core Tet super-cluster than distorted Tet (nk5), which shows a relationship with another distorted CG cluster (nk1). As for k-means clusters nk1, nk4, nk8 and nk9, distorted Pyv (nk1) and Pyv (nk4) are the first to cluster together in the structural dendrogram, closely followed by Bpv (nk8) and then Spl (nk9). Similarly in functional dendrogram, Pyv (nk4) and Bpv (nk8) are grouped together and then with Pyv (nk1).

Figure 4.10 shows the same comparison for compressed group. Compressed k-means cluster 4 (ck4) and ck7 are in a subgroup in both structural and functional dendrogram, and so are ck1 with ck5, and ck2 with ck8. These observations definitely indicate there are certain structure-function relations lying in these clusters that need to be further investigated. The 3D structure of ck1 looks like an inverted Tet or Bva, and ck5 is a completely new CG that does not even resemble any known CGs. They both worth further investigation as well.

In addition to comparing the structural and functional distances directly, functional annotation enrichment was performed for both the normal and compressed zinc sites. We used hypergeometric enrichment to compare the EC annotation and IPR annotations that overlap a zinc site in the normal and compressed groups relative to all of the annotated zinc sites.

We trimmed the EC numbers to the second digit as annotations for enrichment calculations. The EC numbers are enriched in either the compressed or normal group, but not both. The most enriched enzyme classes in the normal group are 4.2 (carbon oxygen lyases), followed by 2.1 (transferases transferring one-carbon groups), 3.4 (peptidases), and 4.4 (carbon sulfur lyases). Comparatively, in the compressed group, the most enriched enzyme classes are 1.7 (oxidoreductases acting on other nitrogenous compounds as donors), 0 (no EC number), 3.2 (glycosylases), 1.16 (oxidoreductases oxidizing metal ions), and 2.4 (glycosoltransferases).

Similarly, a number of InterPro annotations are enriched in either the normal or compressed group, but not both. In fact, many of the InterPro annotations in the normal zinc sites are not present at all in the compressed sites, but all sites are only in the normal group, including the most highly enriched annotations such as C2H2 zinc fingers (IPR015880, IPR007087) and glycoside hydrolase (IPR027291, IPR015341, IPR028995). Many of the other highly enriched annotations in normal have only a few sites in the compressed group, including carbonic anhydrase (IPR018338, IPR023561, IPR018443) and PHD-type zinc fingers (IPR013083, IPR019787, IPR019786).

The compressed-specific annotations included pollen allergen (IPR001778, IPR002914), as well as protein of unknown function (IPR010281). Other highly enriched annotations include immunoglobulin domains (IPR013783, IPR007110, IPR013106), ferritin (IPR009078, IPR012347), superantigens (IPR016091, IPR013307), staphylococcal/streptococcal toxins (IPR006126, IPR006173, IPR006177).

These results imply that although there are many functions that can be performed by both normal and compressed CGs, there are some that appear to be specific to one type or the other.

4.4 DISCUSSION

Previous works have attempted to characterize zinc binding in metalloproteins by considering only canonical zinc CGs that have been previously observed and explained by coordination chemistry. However, when these expectations of canonical CGs are applied to zinc ions bound by proteins, many zinc sites are classified as outliers or are misclassified with respect to CG [8, 24]. Our analysis of ligand-zinc-ligand bond angles, where the best fc-shell is determined from only previously characterized zinc-ligand bond-lengths, showed the presence of angles below 58 (compressed) and 38 (super-compressed) degrees. As these angles are incompatible with any previously characterized canonical CG, they implied the existence of unknown CGs. Many, but not all of the compressed and super-compressed angles appear to contain bidentate ligands (wherein two of the ligands to the zinc atom are from the same amino acid residue or molecule) or non-amino acid ligands. This points to the need for "less biased" methods for determining zinc CGs in proteins.

What is especially interesting is that it is not possible to organize all of the CGs using only the angle information. Clustering all of the zinc sites using only the sorted angles does not lead to stable clusters (Figure 4.8, Table 4.3, and 4.4). This aspect of the CG detection methodology (in combination with using known bond-length mean and standard deviations) leads to our method being "less biased" than previous methods,

however there is still a bias. The sites must still be classified as either normal or compressed prior to clustering on the angles. But this classification is based on direct observations of the angle distributions in the dataset and not on prior belief of what is in the dataset.

Following the clustering of the normal and compressed zinc sites, assignment to canonical CGs was made based on agreement with their expected angles. The normal sites fit canonical CGs very well, as is expected. An attempt was made to relate the compressed CGs to canonical CGs using a combination of criteria including χ^2 probability calculations after removing the compressed angle to remove that as a source of bias. The assignment to canonical CGs in this case is still a bit of a misnomer, as most of these severely compressed versions of canonical CGs have not been described in the literature. From this perspective they can be viewed as novel CGs. However, we took the conservative approach of simply describing them as large distortions of the canonical CGs. We have also labeled the compressed CG (cluster 5 of the compressed group) that appears completely distinct from all of the other canonical CGs as truly "novel".

To allay suspicions that these compressed angles are the result of experimental artifacts, such as whether or not it is just due to the uncertainty of the X-ray experiment, we calculated the average of the b-factors of the ligands composing the compressed angle versus normal angles. As shown in Figure 4.11, there is no significant difference in b-factors between different composing ligands. There is literature suggesting some of the compressed angles are a result of a phenomenon called a carboxylate shift [108], which is a thermodynamic mechanism enzymes employ to sustain the CG when binding and leaving a substrate. However no one has systematically examined this phenomenon in

terms of metal's CG in wwPDB. Also, a simple mechanism could not cover all instances, such as the bidentation caused by ligation of cysteine's backbone and side chain together.

The compressed and "novel" CGs beg the question: why have they not been previously reported? One answer is that until recently there have not existed enough example structures for them to be reliably observed even with our "less biased" characterization methods. Figure 4.12 shows how the number of compressed zinc sites has increased proportionately with the growth of wwPDB. It is only within the past 10 years that enough compressed sites existed in wwPDB for a rigorous study to observe and detect them. More importantly however, is the fact that even with a relatively large fraction of compressed sites, an analysis that considers only the canonical CGs from previously identified zinc coordinations and bonding structures, will remove compressed sites from the analysis as outliers. This is exemplified by the work of Andreini et al, MetalPDB[22], where the summary of zinc metal showed the "outlier" category had the largest number of instances. Figure 4.12 shows that there should have been more than enough compressed sites to be detectable; however, there were no compressed sites reported by Andreini et al. There were a number of outliers noted in their work. Some of the "outliers" reported by Andreini et al were likely zinc sites with compressed CGs, but because their analysis considered only "normal" zinc CGs, the compressed CGs were overlooked and not reported. This directly underscores the need for "less biased" analyses of metal CGs in proteins so that these previously described CGs are not overlooked or merely classed as "outliers" and completely removed from an analysis.

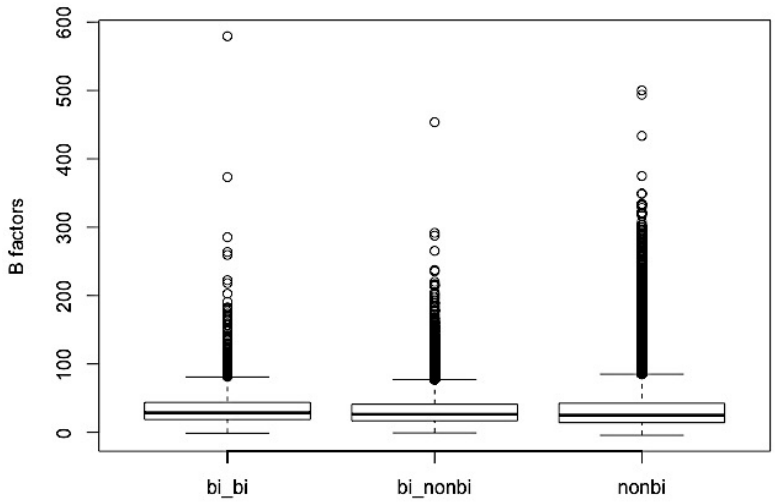


Figure 4.11 B-factors for different categories. bi_bi includes the bidentate ligands in the compressed zinc sites; bi_nonbi includes the non-bidentated ligands in the compressed zinc sites; nonbi includes all ligands of the normal zinc sites.

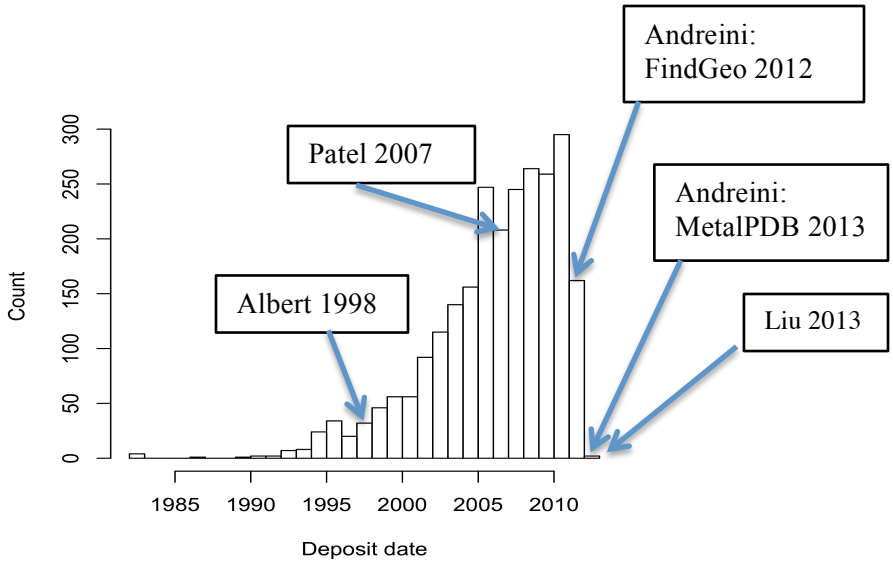


Figure 4.12 Analysis of the deposition history of the March 2013 wwPDB zinc metalloprotein entries with compressed angles. Publication date of the key references are indicated on the graph.

These compressed sites also show enriched functionality relative to all of the sites, suggesting there are particular functions or enzyme classes that are preferentially compressed. The correspondence between CG cluster distances from angles and cluster distances from functional annotation further emphasize the functional importance of the compressed and novel CGs. However, it should also be emphasized that it is difficult from this work to assign functionality to particular normal or compressed clusters, as multiple clusters seem to share functionality. We see two possible explanations: a) presence of false positives in associating function with the zinc sites and b) potential existence of zinc metalloproteins with multiple zinc-coordinating CG conformations, but where the x-ray crystal structure freezes out just one conformation. Improvements in functional annotation methods will be required to address these short-comings, including: i) the development of better annotating hidden Markov models (hmm) to better relate zinc binding site detected from protein sequence to specific protein functions ii) the development of better methods that relate overlapping protein regions with respect to protein functions. Dealing with the second explanation may only be addressed by NMR studies [109] and/or newer combined quantum mechanical, molecular mechanical, molecular dynamics (QM/MM-MD) simulations [110].

4.5 CONCLUSION

In this chapter, our "less biased" approach was presented for the classification of zinc binding sites with respect to CG that allows for the detection of novel CGs. From one perspective, we have detected eight novel CGs that contain compressed angles and

cannot easily be classified into one of the canonical CGs. From another perspective, seven of these eight novel CGs can be viewed as highly distorted versions of the canonical CGs; however, this perspective may be considered as simply trying to push a square peg into a round hole. From either perspective, one of the compressed CGs appears to be truly novel and distinct from all canonical CGs by every probabilistic, angle comparison, and visual inspection criteria we could use. As wwPDB continues to grow, additional distorted or novel CGs may become detectable; however, we will only be able to detect these previously undetected CGs by using an unsupervised clustering approach such as the one described in this chapter rather than applying a supervised classification method based on "known" CGs which has been the method of choice up to this point in time. In other words, we will only be able detect these previously undetected CGs if we stop assuming that we already know what a dataset contains before analyzing it.

CHAPTER 5

COORDINATION GEOMETRY AND FUNCTIONAL PROPENSITY OF TOP FIVE METALLOPROTEINS

5.1 Introduction

Chapter 4 demonstrated a proof of concept that our analysis pipeline works in characterizing zinc metalloproteins' coordination geometry (CG). It presented a general CG description of single zinc ion that could be constructed based on 3D-structure and has a high (0.88) correlation with function. Furthermore, we showed that a large number of aberrant 4-ligand CGs in zinc metalloproteins with significant deviations from canonical CGs existed due to structural constraints from the metalloprotein. These constraints, mostly in the form of bidentated ligands, and associated aberrant CGs revealed unique functional relationships. However, these results created several new questions:

1) Could similar functionally relevant structural descriptions of CG be constructed for other common metals, involving different numbers of ligands?

2) Would similar or even new structural constraints and aberrant/novel CGs be detected?

In order to address these questions, we greatly expanded our methodology to allow construction of CG structural descriptions with an arbitrary number of ligands. We also significantly improved our detection of metal binding ligands by adding several

quality control filters, compensating for crystallographic resolution, and by preventing false detection of ligands. These improvements helped to detect and structurally describe single metal ion CGs and their functional relationships across the five most abundant metalloproteins (see Table 5.3).

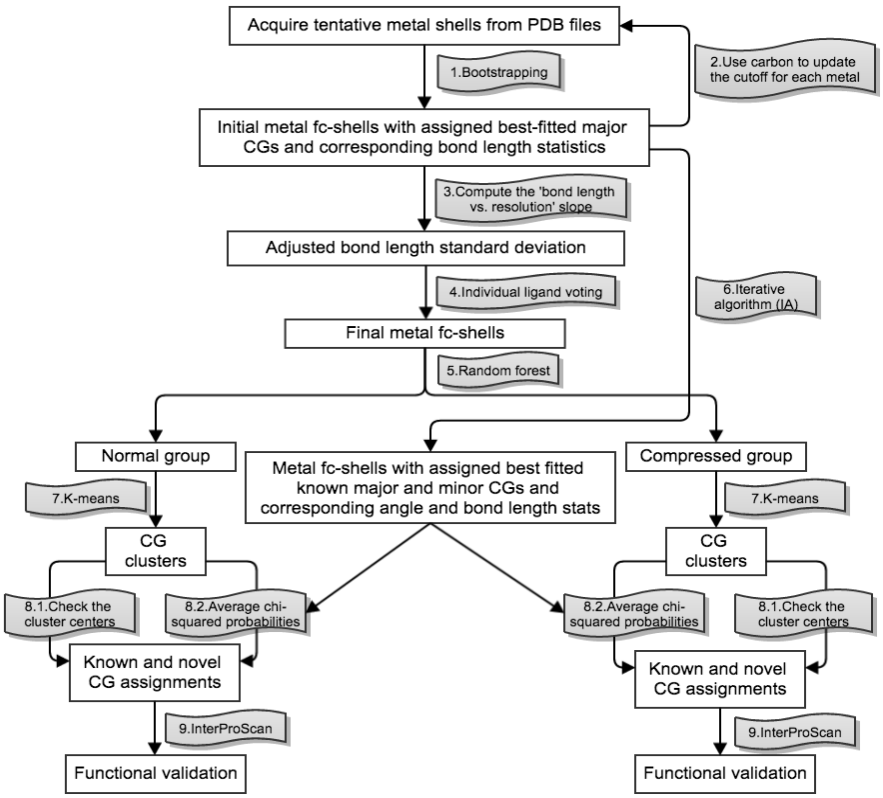


Figure 5.1Workflow for Chapter 5.

5.2 Methods

A general workflow for this chapter is shown in Figure 5.1. It kept the main steps of the pipeline in Chapter 4, while added several quality control steps to remove "bad"

data and improve the statistics in use. The basic idea holds that the less biased bond-length statistics was used in defining metal binding shells, while the angles were used for classifying the CG. Several improvements were made to reduce the error rate in ligand selecting. A collapsed angle space was used to increase the CG clustering efficiency.

5.2.1 Define Metal's First Coordination Shells (Fc-Shells)

All released structural entries were downloaded from wwPDB on Feb 25, 2015. Our metalloprotein filtering tool identified all PDB entries with at least one metal atom in the HETATM record and removed entries with fewer than 20 amino acids in the SEQRES record. Next, metal clusters were identified and removed, using two metal atoms within 3Å as the filter. The top five abundant metals, Zn, Mg, Ca, Fe, and Na, were kept for the rest of the analyses in this chapter. If not specified, all analyses were carried out for each metal separately at first and then combined together by ligand number. The overall workflow is shown in Figure 5.1. Since the general procedure is similar to what was performed in Chapter 4, we are mainly highlighting the extensive list of improvements here.

5.2.1.1 Acquire Initial Fc-Shells and Bond-length Statistics (Step 1)

For each metal site, we generated a list of potential non-H shell ligands (including carbon) within a certain distance of the metal atom. The initial lower cutoff is 1.3 Å for all metals, and the initial upper cutoff is based on the atomic radius of the metal as shown in Table 5.1. To avoid the inclusion of second shell atoms due to this generous upper cutoff, the bond-lengths between any atom and the metal must be smaller than 1.5 times the bond-length of the metal to any other atoms in the cutoff, and smaller than 1.5 times the bond-length between the two atoms. This 'triangular rule' can help exclude atoms that

do not directly bind to the metal but are still part of the metal's local chemical environment. We then used the CG evaluation tools to bootstraping the best-fitted canonical CGs in order to identify an initial set of binding ligands. To achieve that, all subsets and permutations of the ligands and the corresponding ligand–metal–ligand angles (angles) were computed and compared to the ideal angles of the canonical CGs, tetrahedral (Tet), trigonal bipyramidal (Tbp), octahedral (Oct), and pentagonal bipyramidal (Pbp). Several additional filters were applied to the set of atoms before checking against the canonical CGs: 1) only one of the alternate locations of an amino acid residue was allowed to be in the set; 2) if any two atoms are smaller than 1.5Å or greater than 6.0Å, they were marked as unreasonable atom-atom distance and eliminated; 3) if any of the atoms are symmetry-related, unless it is from author determined biological units or all symmetry-related atoms are water, the binding site would be excluded from further analysis; 4) we also excluded the metal site if the majority of its ligands were water. These filters limit the inclusion of metal binding sites that may represent non-specific binding or crystallographic artifacts. These filters limit the inclusion of metal binding sites that may represent non-specific binding or crystallographic artifacts. The canonical CG that passed all filters and had the smallest angle variance to the actual structure was classified as the structure's CG, and the set of atoms were considered the binding ligands.

5.2.1.2 Update the Upper Cutoff (Step 2)

As the initial binding ligands were identified, bond-lengths of each element type (O, S, N, …) were acquired. The inclusion of carbon as the binding ligands in Step 1 can be used to estimate the chance of having an atom accidently aligned as well as a

canonical CG in regard to other binding ligands. This is due to the increasing atom density with respect to angle space as a shell inclusion cutoff increases. A new upper cutoff was then set to the average between bond-length mean plus one standard deviation of the most abundant element and the main carbon peak. The updated upper cutoff is generous enough to include most of the actual binding ligands but still effective enough to exclude falsely detected ligand atoms. Take Zn for example, the most abundant ligand element is S, as shown in Figure 5.2, and the Zn-S bond-length mean and standard deviation are 2.341Å and 0.152Å accordingly. The main peak of the fictional Zn-C is 3.071Å, so the middle point between them is (2.341 + 0.152Å + 3.071)/2=2.782 (Å), which became the updated bond-length cutoff for the ligand detection of zinc ions.

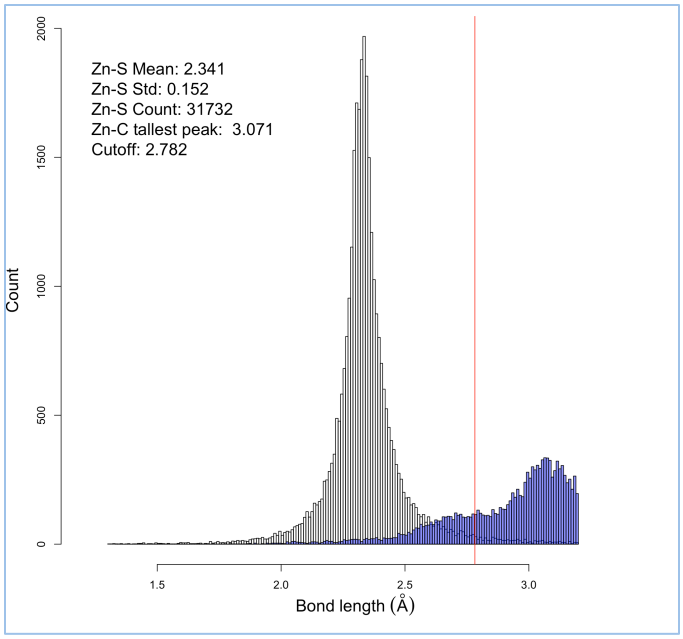


Figure 5.2 Updated bond-length cutoff for zinc.

The computational bootstrapping step was then carried out again using the updated cutoffs to obtain the list of potential shell ligands. The same triangular and other rules were applied. Though this time we only kept elements with a high occurrence (>5%), and also dismissed the carbon. After the second round of bootstrapping, the new tentative metal binding shells were defined. The element-specific bond-length statistics (means and variances) were calculated for each metal.

5.2.1.3 Adjust Bond-Length Standard Deviation Based on X-Ray Resolution (Step 3)

It has been known that the bond-lengths scatter more as the crystallographic resolution gets worse (17). Our data shows that the relationship between a bond-lengths standard deviation and resolution is similar regardless of the metal or the element type (Figure 5.3). Resolutions with more than 30 data points were kept in calculating the standard deviation specific to resolution. A resolution cutoff of 3.5Å was used to ensure a reasonable quality of the data in this step. Considering all metal-element pairs together, we were then able to compute the combined slope of bond-length standard deviation versus resolution. Then for each individual metal site, an adjusted bond-length standard deviation was calculated as:

$$sd_x = c\,(R_x - R_{avg}) + sd_{avg} \qquad (5.1)$$

where c is the combined slope, sd_{avg} and R_{avg} are the overall bond-length standard deviation and the average resolution of a given metal-element type, R_x is the resolution of the metal site to be calculated. The resulting adjusted bond-length standard deviation, sd_x, was used for the next step.

5.2.1.4 Acquire the Final Metal Fc-Shell (Step 4)

With this adjusted bond-length standard deviation, all atoms within the updated cutoff (5.2.1.2) were revisited and only atoms that are within 2.5 adjusted standard deviations of its expected value were kept. The same set of filters as in step 1 was employed again to check the quality of the remaining atoms. The atoms that passed all filters composed our final metal fc-shells for the rest of the analyses. Finally, a non-redundant set of metal list with a resolution better than (smaller than) 3Å and an occupancy greater than 0.9 was also derived for clustering analysis.

5.2.1.5 The Iterative Algorithm (IA) Process (Step 5)

At each iteration, a χ^2 probability was calculated for each CG model at each metal site, using a combined angles and bond-lengths vector. All combinations of the atoms within the updated cutoff defined in 5.2.1.2 were considered. We excluded the combinations if there are angles between the atoms that are below a cutoff specific to each metal based on its smallest angle histogram (Table 5.1). The same set of filters as in 5.2.1.1 was also applied. All CG models in Figure 2.9 were considered. An angle correlation was estimated for each CG to calculate the χ^2 statistics as in Chapter 4.2.1.5. However for Tpr, Sqa, Hbp, and their associated minor CGs, the angle correlation matrix has a large size and big inequality in the numbers. Thus, the inversion that is required for the χ^2 statistics calculation is incapable of capturing the angle's influence over each other. So we treated the angles as independent variables, but with a 1.5 multiplier on the variance to counter the effect of dependency in the χ^2 statistics calculation. The CG model that possesses the highest χ^2 probability was classified as the metal site's CG. Both angle statistics of each CG and bond-length statistics of each element were

calculated from at the end of an iteration, and were then used in the χ^2 probability calculation of the next iteration. The iteration continued until all statistics converged.

5.2.2 K-Means Cluster And Assignment

5.2.2.1 Random Forest (Step 6)

Random Forest was used to separate the normal and compressed groups. Training data were composed of the main angle peaks from the smallest angle histogram. The cutoff for the normal and compressed training data was specific to each metal (Table 5.1). The smallest angle, the two ligands composing the smallest angle, and the bidentation status of the smallest angle are the features for training the classifier.

Table 5.1. Top 5 metals and their derived distance cutoffs defining the coordination shell

Metal	Atomic radius (pm)	Initial distance upper cutoff (Å)	The most abundant element	Bond-length mean of the most abundant element (Å)	Bond-length standard deviation of the most abundant element (Å)	Carbon mean peak (Å)	Element included	Updated distance upper cutoff (Å)	IA small angle removal cutoff (Å)	Random forest cutoff (degrees)
Zn	135	3.20	S	2.340	0.152	3.071	S, O, N	2.782	68	58/68
Mg	150	3.35	O	2.350	0.368	3.067	O, N	2.892	65	52/68
Ca	180	3.65	O	2.481	0.271	3.432	O	3.092	60	55/65
Fe	140	3.25	N	2.063	0.134	3.081	N, O, S	2.639	68	63/73
Na	180	3.65	O	2.697	0.369	3.568	O	3.317	60	45/60

Table 5.2. 6-angle space for all CGs

CG	Largest	Sorted middle (Smallest-middle, 33-quantile-middle, 66-quantile-middle, largest-middle positions are in red)	Smallest opposite
Tet	109.5	109.5, 109.5, 109.5, 109.5	109.5
Bva	120	90, 90,120, 120	90
Bvp	180	90, 90, 90, 90	120
Pyv	180	90, 90, 90, 90	90
Spl	180	90, 90, 90, 90	180
Tbp	180	90, 90, 90, 90, 90, 90, 120, 120	120
Spy	180	90, 90, 90, 90, 90, 90, 90, 180	90
Tpv	131.8	70.6, 90, 90, 90, 90, 131.8, 131.8, 131.8	70.6
Oct	180	90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 180, 180	90
Pva	144	72, 72, 72, 72, 90, 90, 90, 90, 90, 144, 144, 144, 144	72
Pvp	180	72, 72, 90, 90, 90, 90, 90, 90, 90, 90, 144, 144, 144	72
Tpr	131.8	70.6, 70.6, 90, 90, 90, 90, 90, 90, 131.8, 131.8, 131.8, 131.8, 131.8	70.6
Pbp	180	72, 72, 72, 72, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 144, 144, 144, 144, 144	72
Hva	180	60, 60, 60, 60, 60, 90, 90, 90, 90, 90, 90, 120, 120, 120, 120, 120, 120, 180, 180	60
Hvp	180	60, 60, 60, 60, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 120, 120, 120, 180, 180	60
Cuv	180	70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 180, 180	70.5
Sav	143.6	70.5, 70.5, 70.5, 70.5, 70.5, 82, 82, 82, 82, 82, 82, 109.5, 109.5, 109.5, 143.6, 143.6, 143.6, 143.6, 143.6	70.5
Hbp	180	60, 60, 60, 60, 60, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 120, 120, 120, 120, 120, 120, 180, 180, 180	60
Cub	180	70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 109.5, 180, 180, 180	70.5
Sqa	143.6	70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 70.5, 82, 82, 82, 82, 82, 82, 82, 109.5, 109.5, 109.5, 109.5, 143.6, 143.6, 143.6, 143.6, 143.6, 143.6, 143.6, 143.6	70.5

5.2.2.2 K-Means Clustering (Step 7)

K-means clustering was employed to cluster the metal sites based on their ligand-metal-ligand angles. All angles were first ordered as largest angle, opposite angle, and the smallest opposite angle. The opposite angles are those that do not share any ligands with the largest angle, and the middle angles are all angles except the largest and the smallest-opposite angle. To enable metal sites with different number of ligands to be comparable with each other, we reduced the all-angle space to a 6-angle space via selecting the following angles from all angles of a given metal site: largest angle, smallest-middle angle, 33-quantile-middle angle, 66-quantile-middle angle, largest-middle angle, and smallest-opposite angle. Shown in Table 5.2, the selected middle angles are in red. This reduced angle space can preserve the key information needed for separating each CG, while reducing the redundancy of the repeated angles. Four measures were used in determining the optimal number of clusters (k). The Jaccard index computes how well matching clusters overlap between iterations. The sum of difference indicates how close the cluster centers are to each other between iterations. And the rho and p-value indicate an average of functional propensity between clusters. For all four measures, a larger value denotes a better performance.

5.2.2.3 Cluster Assignments (Step 8)

To characterize the clusters, we check the cluster centers and calculated a χ^2 probability of each CG model for each metal site. The model that had the highest cluster-average probability was then characterized as the cluster's CG.

5.2.2 Functional Validation of The K-Means Clusters (Step 9)

We ran InterProScan (21, 22) 5.18-57 using the current versions of its member databases on the non-redundant sequences previously determined. We retained only those results with an InterProScan (IPR) annotation mapping and overlapping at least one ligand residue. We derived and evaluated the consistency of CG-based structure and sequence-based function annotation relationships between k-means clusters. We also determined functional enrichment of k-means cluster.

Table 5.3 Numbers of metalloproteins in wwPDB as of Feb 2015.

Metal	Number of PDB entries	Number of total metal sites	Metal	Number of PDB entries	Number of total metal
Zn	9,360	26,788	Pb	48	152
Mg	9,145	53,896	Gd	42	197
Ca	7,762	24,335	Tl	40	261
Fe	6,359	27,514	Rb	37	153
Na	4,888	16,527	Sm	33	111
Mn	2,266	8,138	Ir	31	48
K	1,673	5,306	Pr	22	55
Cu	1,134	4,397	Rh	20	46
Ni	935	2,252	Eu	19	61
Co	915	2,087	Pd	19	85
Cd	758	4,289	Ag	18	75
Hg	528	1,923	Os	14	33
Pt	191	629	Lu	13	56
Mo	176	664	Ho	12	35
Al	158	351	Tb	11	32
V	120	364	Cr	9	21
Ba	118	311	Ga	8	10
Sr	118	3,551	La	8	18
Ru	99	134	Sb	5	10
Cs	88	393	Ce	4	7
W	76	1,443	Er	2	6
Yb	72	177	In	2	3
Au	64	322	Bi	1	1
Y	53	202	Dy	1	30
Li	52	88	Total	47,527	187,587

5.3 Results and Discussion

5.3.1 Defining the Metal Fc-Shells

The wwPDB contains a total of 106,427 structures as of Feb 25, 2015, and 47,527 of which are metalloproteins. The number of metalloproteins and metal sites can be found in Table 5.3. The five most abundant metals, Zn, Mg, Ca, Fe, and Na are primarily considered in this chapter.

5.3.1.1 Defining the Metal Coordination Shell Cutoff

Determining a metal's binding ligand is not as straightforward as one would anticipate, as first- and second-coordination atoms from a protein are often crowded together around the metal ion. In this situation, there is no simple rule in deciding whether an atom is metal binding or not. This is partly due to the limitations in structural resolution, crystallographic artifacts, and to phenomena such as carboxylate shift (23) that smear the metal-ligand bond-lengths. The determination is often achieved simultaneously with a metal binding site's CG classification. The most common approach is to use a simple distance cutoff and then select a ligand subset that best fits one of the canonical CG models (15), but also sometimes takes into account the bond valence model (14). The dilemma of choosing the cutoff is, if it is too generous, extra atoms will be included, which will increase the demand for a more accurate CG fitting method. But if it is too strict, some of the loosely bound ligands will be excluded in the first step, which will hinder the fitting to the correct CG model. This methodology also precludes the assumption of non-canonical, aberrant CGs.

As Chapter 4 showed, simply matching to canonical CG models is problematic, which makes the accurate detection of metal binding ligands even more critical for detecting and analyzing CG. In this chapter, we first used an initial shell cutoff based on the metal's atomic radius to detect potential ligands to canonical CGs in order to derive metal-ligand bond-length statistics for use in later steps. This first round of bootstrapping can capture the general distribution of bond-length for each ligand element. However, Figure 5.2 clearly shows that if this raw shell cutoff is the only criteria used, significant numbers of non-ligand second-shell atoms (represented by carbon) will be included due to the atom-angle density issue. To get rid of these non-ligand second shell atoms, we used carbon to estimate the false ligand metal distance distribution and then identified where the false ligand atoms start to appear with high probability (i.e. the highest carbon atom mode). In other words, we used the ubiquitous presence of carbon in protein structures to estimate 'accidental' angle alignment with other ligands to fit any canonical CGs. In order to make sure that most of the actual binding ligands were included, the new shell cutoff was also set to ensure the inclusion of the majority of the most abundant ligand element. The red lines in Figure 5.2 shows the cutoffs used for Zn, which were the middle points between the main carbon peak and mean plus one standard deviation of sulfur, the most abundant element of zinc. We also included a triangular test to filter out any atoms that is connected to the metal through more than one bond. With these improved shell cutoffs (Table 5.1) and additional heuristics, we were able to generate improved bond-length statistics for each metal-ligand elemental combination.

5.3.1.2 Bond-length Standard Deviation Adjustment

For accurately detecting the proper set of ligands, our next major improvement involved adjusting the bond-length standard deviation based on crystallographic resolution. With accurate bond-length statistics, the detection of the proper set of ligands can be performed independently, a single ligand at a time, via a statistical test. However, the bond-lengths tend to scatter (vary) more as the structure resolution gets worse (i.e. larger resolution value) for a specific metal-element type (17). Rather than greatly restricting our analyses to structure entries with only high resolution like <1.5Å, we are able to safely extend our analyses to structure entries with lower resolutions down to 3.0Å by taking the crystallographic resolution into consideration in the statistical test. In order to do this, we shifted all the bond-length standard deviation (bl-std) to resolution data points along the y-axis by its own overall metal-element bl-std. Figure 5.3 shows that regardless of the metal and binding element, the bl-std and resolution relationship is of the same proportion. Therefore, a combined slope can accurately describe this relationship and be used to adjust an individual metal-atom pair's standard deviation according to the entry's resolution as shown in Equation 5.1. We also tested deriving similar standard deviation adjustments based on R-factor and R-free and combinations of R-factor, R-free, and resolution (data not shown). Combinations did not work well since the low density of entries prevented accurate calculation of metal-ligand bond-length standard deviations. However, in the future, we may have enough structural examples to re-examine combinations. As of currently, the resolution-corrected bl-std provides the highest correlation for the resulting clusters.

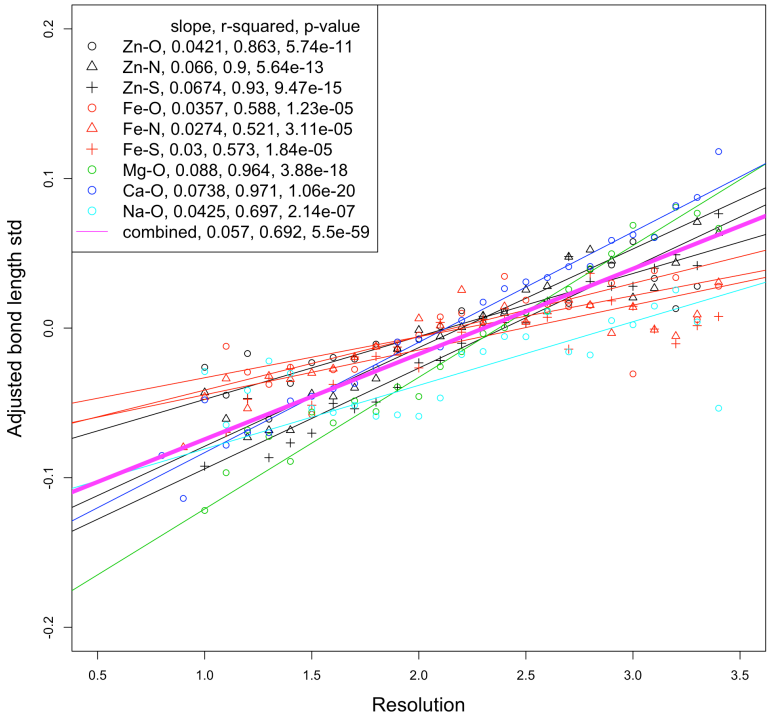


Figure 5.3 Bond-length standard deviation vs. resolution. The trend is similar regardless of the bond type. A combined regression line is determined based all data points together, weighted on the member counts composing each data points.

5.3.1.3 Voting Individual Ligand Via Statistical Test

The bond-length histograms show an approximate normal distribution for most of the metal-ligand bond types. Therefore, a simple parametric test is used to detect ligands based on bond-length means and resolution-adjusted standard deviations. We tested a range of ligand detection standard deviation cutoffs from 2 to 3 bl-stds. When the stricter cutoffs (i.e. 2 bl-stds) are used, all downstream cluster measures tend to be higher and

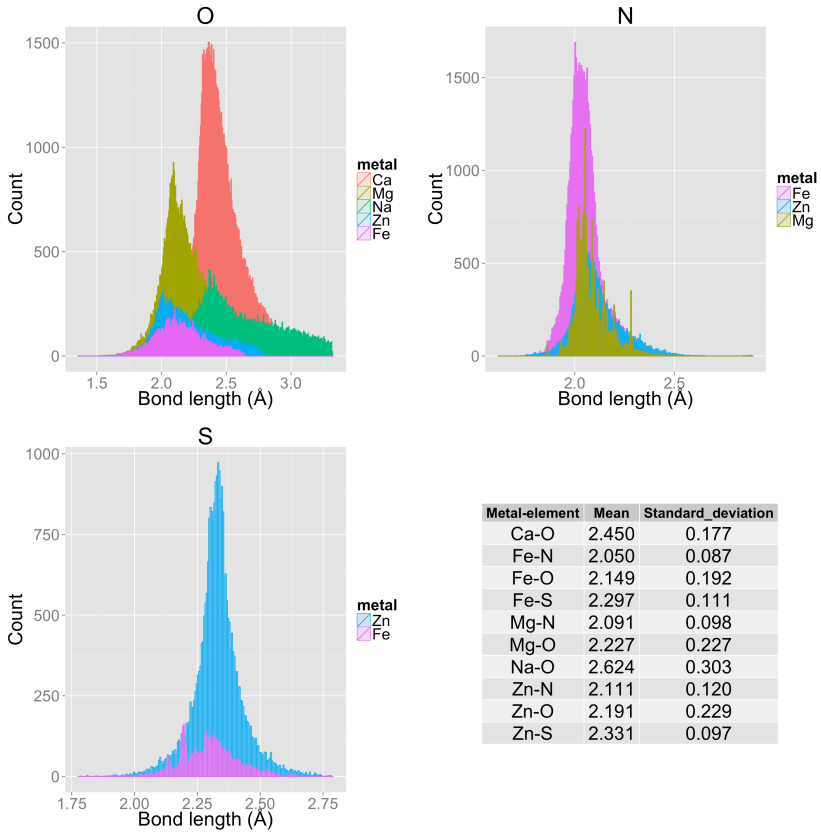
more stable. But on the other hand, fewer ligands will be counted as binding ligands. And due to deviations from normality, ligands in compressed angles are disproportionately lost, which leads to insufficient number of compressed angle for clustering. Therefore, a 2.5 bl-std cutoff was used for this study to compromise between the two situations. Thus, 2.5 standard deviations ensure that approximately 98.8% of the suitable ligands will be included. The bond-length distributions of the final fc-shell ligands are shown in Figure 5.4.

Another possible way of determining the binding ligands is to use chi-squared probability testing for the set of potential ligands together. Compared to the chi-squared method, the single ligand testing does a much better job in identifying a higher number of ligands, as it could correctly characterize the most common number of ligands of Fe, Mg, and Na as six and Ca as seven, while our previously published chi-squared probability method tended to favor 4-ligand structures for all metals.

5.3.1.4 Filters Used In Getting The Metal Fc-Shell Ligands

Several additional filters were employed in several steps throughout our analysis. A symmetry-related filter was one of the major filters added. In Chapter 4, all structures were only considering what was directly reported in the PDB coordinates file, while we overlooked how x-ray crystallography actually works. In x-ray crystallography, the protein form crystals, repeated symmetric components, before being detected in the instrument. Especially for homomultimeric proteins, only a representative single unit will be deposited in the PDB file. So when considering an interaction between metal and ligand, especially when the metal is on the interface between the multimers of a protein, we need the adjacent subunits to decide what is truly binding the metal. Crystallographic

symmetry information within a unit cell can be found in REMARK 290 and REMARK 350. Symmetry information of adjacent unit can be calculated from CRYST1, ORIGX, and SCALEX record. However, we need to be careful to distinguish a biological multimer from a crystallographic multimer. To accomplish that, we used the author provided information in PDB (REMARK 350) for the distinction. Water is an exemption here, as it is often the solvent, and does not belong to any multimeric protein unit even though symmetry-related water molecules can be calculated for different unit cells.



Metal-element	Mean	Standard_deviation
Ca-O	2.450	0.177
Fe-N	2.050	0.087
Fe-O	2.149	0.192
Fe-S	2.297	0.111
Mg-N	2.091	0.098
Mg-O	2.227	0.227
Na-O	2.624	0.303
Zn-N	2.111	0.120
Zn-O	2.191	0.229
Zn-S	2.331	0.097

Figure 5.4 Bond-length distribution and statistics of different metal-ligands.

Alternate locations may happen for an atom when there are more than one possible conformations captured in X-ray detection. It is often associated with the occupancy parameter, where one of the alternate locations is 0.6 and the other one is 0.4. Only one of the possible alternate locations was used for every amino acid residue. This was a sensible filter, but was never mentioned in any previous studies. The 'triangular' rule was employed to ensure a reasonable ligand-ligand contact. This prevented the selection of atoms where another ligand is intervening between the potential atom and the metal ion. The rational of removing metal sites that have ligand distance smaller than 1.5 Å or greater than 6.0 Å was that any abnormal ligand-ligand distances indicate a poor quality of the data, and thus should be eliminated. And after the non-redundant test, a resolution of 3 Å or better and occupancy of 0.9 or higher filter was also employed.

All these filters helped to eliminate non-specific binding and ensure a high quality of the structural data being analyzed. Table 5.4 shows the count of different number of final fc-shell ligand for each metal. The reason for using Tet, Tbp, Oct, and Pbp in bootstrapping was because they are the major CG for 4- to 7-ligand CGs, which are the most abundant ligand number being detected. As we analyze other metals in the future, more major CG may be added in bootstrapping according to the natures of the metal.

Table 5.4 shows the count of different number of ligand for each metal after step 4. Based on this data and the physiochemical bonding capacity of a metal ion (i.e. the number of ligands a metal ion can bond) [91, 111, 112]), we could estimate an error rate for our ligand detection analysis. The error rate was calculated as the number of ligands not physiochemically expected (e.g. 7 and 8 for zinc) divided by the total number of detected ligands for sites with the largest number of expected ligands. For example, the

zinc estimated ligand detection error rate is (2*2+22*1)/(2*8+22*7+1404*6) ≈ 0.00302. For four of the five metals, the estimated ligand detection error rate ranges from 0.018% to 0.50%, with an overall error rate of 0.23% across these metals. Error estimates for Na ions were not included since we could not find a reliable source indicating which coordination numbers for Na are not aqueously, and by inference, biologically relevant for Na ion coordination. Thus, our analyses provide both a false positive rate (0.23%) and a false negative rate (~1.2%) for ligand detection, indicating a very robust method. No prior protein metal binding site analysis methodology has undergone this level of statistical evaluation nor demonstrated this level of rigorous performance.

Table 5.4 Ligand counts for the metals with an estimated error rate.

Metal	4-ligand	5-ligand	6-ligand	7-ligand	8-ligand	9-ligand	Total	Error rate
Zn	12,797	3,192	1,404	22	2	-	17,417	0.00302
Mg	5,187	4,350	6,760	191	7	2	16,497	0.00503
Ca	1,494	2,122	5,281	6,029	1,495	24	16,445	0.00197
Fe	1,547	5,182	6,577	7	-	-	13,313	0.000177
Na	1,908	2,435	2,673	450	70	1	7,537	0.00176
Overall								0.00262

5.3.2 The Universal Existence Of Compressed Angles Among Metalloproteins

Upon identifying the binding ligands, the smallest ligand-metal-ligand angle of individual metal sites can be computed. The smallest angle histograms (Figure 5.5) show that there exists two types of angles, normal angles expected from canonical CGs, and compressed angles, the majority of which cannot be explained by expected canonical CGs. All metals contain both normal and compressed angles. Different metals have a

different number of compressed angles. Ca has the highest fraction of compressed angles partly due to its ability to bind 7 or 8 ligands, which increases atom density, resulting in increased numbers of compressed angles. Hexagonal bipyramidal and its associated minor CGs have expected angles of 60 degrees, but they only compose of a small portion of calcium's CGs (15). Mg and Na have a much smaller proportion of compressed angles.

The reason may be due to the fact that a large amount of their ligands are H_2O, which cannot form a bidentation with the metal. Though water may not be a causal factor, the high percentage of H_2O could limit the amount of the other possible ligands that could develop bidentation with the metal. Among the normal angles, the peaks around 72 degrees of Mg, Ca, and Fe can be justified by the Pentagonal bipyramidal (Pbp) CG, or its associated minor CGs. The peak around 90 degrees of Fe, Mg, Ca, and Na can be explained by Octahedral (Oct), Trigonal bipyramidal (Tbp), or their associated minor CGs. The 109-degree peak of Zn is from the Tetrahedral (Tet) as shown in Figure 5.5, which matches a similar graph Figure 4.3 generated from data that is two years older. Whereas the compressed angles are normally less than 60 degrees, and cannot be explained by any known 4-, 5-, and 6-ligand canonical CGs, which are the majority ligand numbers for Zn, Mg, Fe, and Na. All five metals contain significant numbers of compressed angles and they form a normal-like distribution. If we associate the smallest angle based on its binding ligand's type, like whether it is one of the 20 standard amino acids, water, or something else, or whether it is bidentated or not, most of the compressed angles consist of bidentated standard amino acid ligand residues.

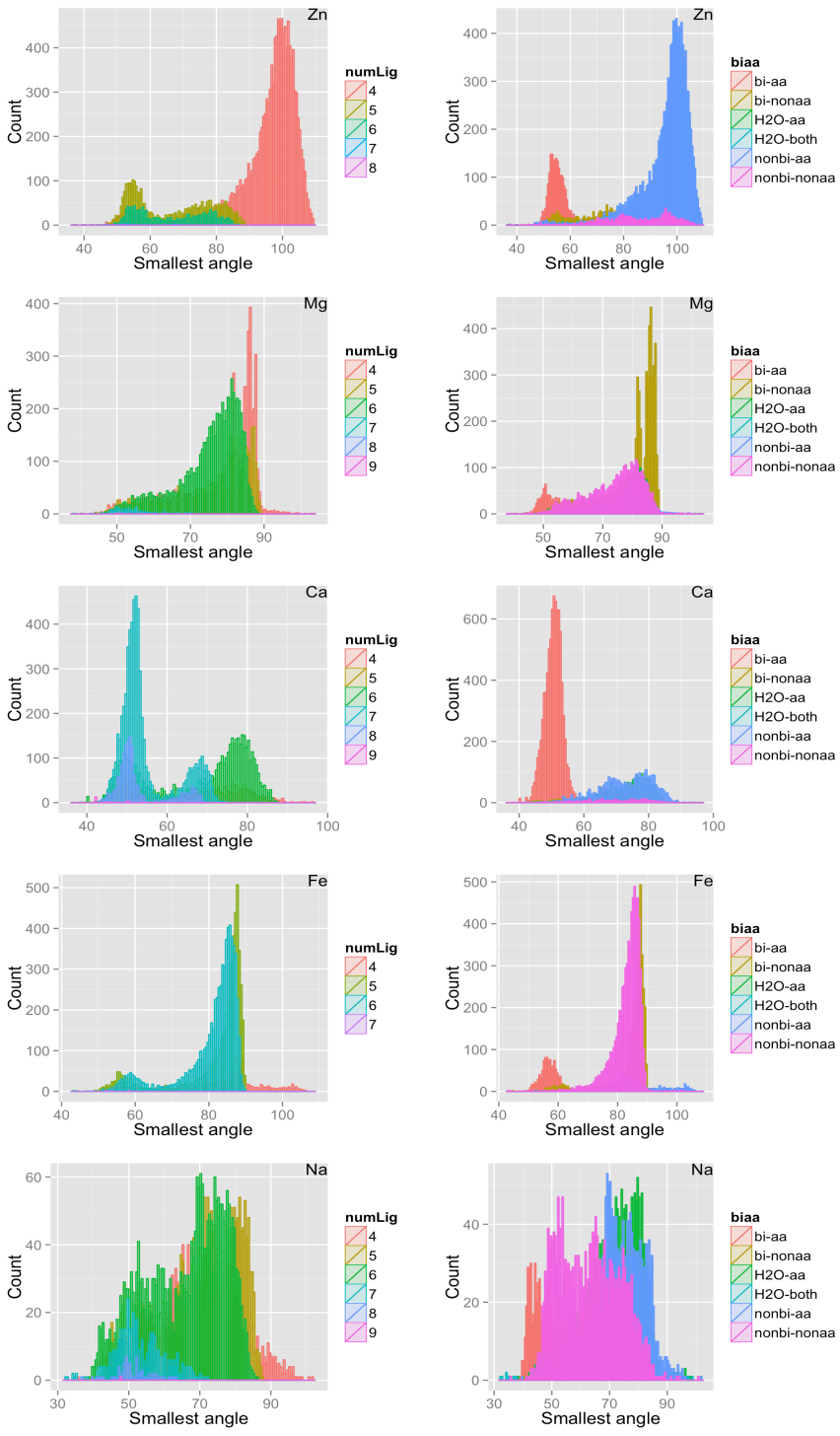


Figure 5.5 Histogram of minimum angles for different metals, broken down by ligand number (left) and ligand type (right).

5.3.3 Angle-Space Descriptions of CG

Instead of an all-to-all mapping of ligands followed by comparing all corresponding angles, we first ordered the angles by finding the largest and smallest opposite angles so that the basic orientation of the metal structure was anchored at the ends of the ordered tuple. Then the middle angles were sorted from small to large to prevent any scrambling that may be introduced by ligand positioning. This ordering allows us to compare individual metal fc-shell not only to canonical CGs, but also to other metal fc-shells. Thus, we were able to explore the similarity between metal structures. Moreover, different CG models process very distinct ordered angles and are easily separable by clustering algorithms. We then further reduced the full-angle space to a 6-angle space so that metal sites with different number of ligands are comparable to each other and can be analyzed together. As shown in Table 5.2, this ordered angle selection method tends to capture a discriminating angle profile for each CG. The largest angle and its smallest opposite angle are kept. The middle angles are evenly sampled based on their position in the ordering to preserve the key information needed for separating each CG while reducing the redundancy.

In the test of using full-angle space instead of 6-angle space, we observed very little difference in the performance in terms of the functional tendency, especially in 5- and 6- ligand structures. This suggested this angle space reduction was effectively picking up the functional relevant angle information, while removing the noisy redundancy coming from the structurally equivalent repeating angles. However, as the ligand number goes above 6, the collapsed 6-angle space represents less and less of the total angle information present. This is not surprising since it is harder to capture 21 (7-

ligand) and 28 (8-ligand) angles worth of information in just 6 representative angles. It is a fundamental problem associated with this angle representation scheme. We observed a slightly unstable correlation for 7- and 8-ligand Ca, which could be a synergic contribution from both small data size and inadequate angle space representation. Since the ligand number for majority of the metals in this study are 4 to 6, this effect needs further investigation as more high-ligand-number metals are analyzed.

To overcome this, we could either use higher or full angle-space representation but giving up the ability to combined different number of ligands together, or find other ways in representation angles. A possible way is to plot the all angles for each metal, and use the angle modes to represent its angle space. If this works, this could overcome the fundamental issue with the current angle representation, that is, different ligand numbers have drastically different angle numbers, which makes the comparison between them rather difficult.

5.3.4 K-Means Clustering and Assignment

K-means clustering was conducted with respect to each metal and each number of ligands separately, and on combined metals and combined number of ligands as well. We particularly analyzed the clustering results in comparison to its counterpart in Chapter 4 (Figure 4.5). An additional criterion was used to pick out the optimal k from Figure 5.6, i.e. all known canonical CGs should have at least one cluster representation. It turns out that when k=5, k=9, and k=11 (k that maximizes the four measures overall), all of them actually mixed the Bvp (120 at the last angle position), Spl (180) and Spv (90) members in different ways while capturing the Tet and Bva members very well. As we go up on the number k, it is only until k=24, the Bvp, Spl, or Spv clusters start to separate well.

What is interesting is there is a clear peak at k=24 on the graph, which means our method is pretty sensitive to the structure-function correlation signal. One explanation is that the analysis in Chapter 4 over-estimated the size of 4-ligand zinc metal binding sites, because of its ligand detecting methods and lack of the extensive quality-control filters. That is, bad data and members of higher ligand CGs may have differentiated the old results. And in order to detect a smaller size of Bvp cluster here, k needs to be bigger so that the larger CG clusters can be broken down into smaller sub-clusters to match the size of Bvp. This unequal density of clusters is a fundamentally hard problem to solve for clustering algorithms [113]. We picked k=9 in this category for its moderate separation of all CGs. This leaded to a rho value of 0.93, which is an improvement of the results (0.88) from Chapter 4.

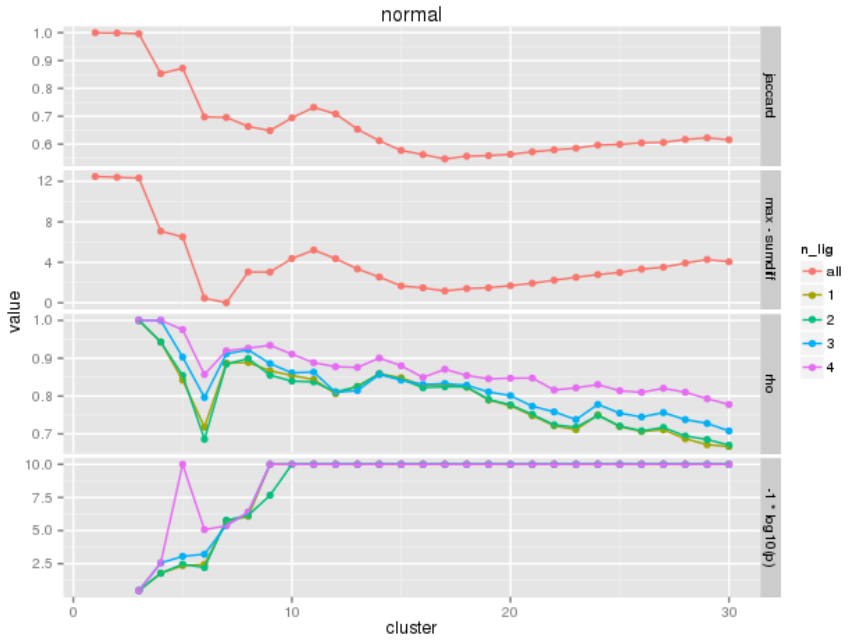


Figure 5.6 Four measures for 4-ligand normal group zinc.

An optimal cluster number k was manually picked for each group to maximize all four measures and to ensure a p-value less than 0.01. For all other categories, we computed the cluster centers, a characteristic average probability of each cluster, as well as structural vs. functional dendrogram.

Figure 5.7 illustrates that the ability to obtain good functional relevant (high rho) clusters is largely influenced by the size of the data to be clustered. The rho increases dramatically at lower counts and plateaus at higher counts. In other words, to achieve a stable high value of rho (~0.8), the data size should be at least 1000. Therefore, in some of the groups, like 4-ligand compressed zinc with a size of 241, the lack of data could greatly hinder our ability to detect a sensible structure-function relationship.

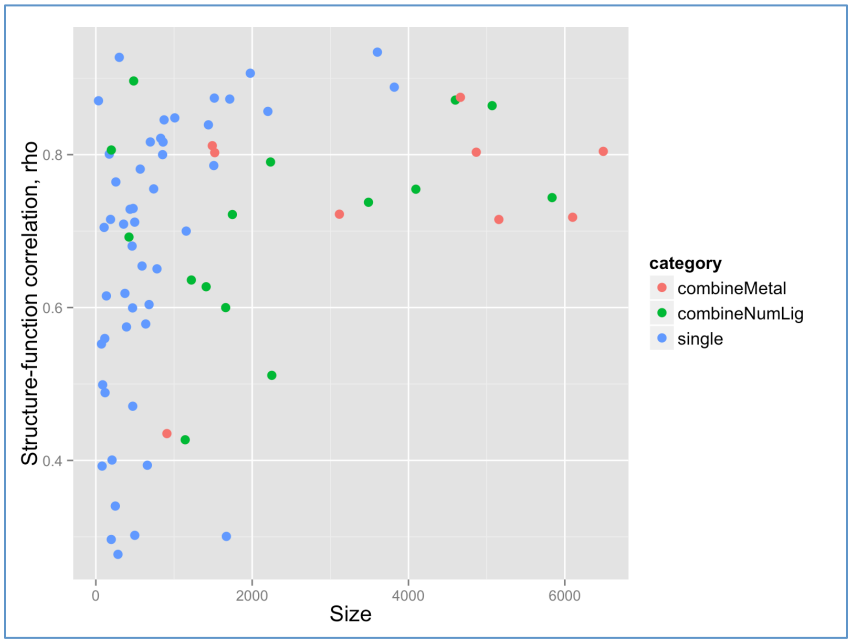


Figure 5.7 The structure-function Pearson's rank correlation coefficient (rho) as a function of the size for real data.

In general, combining different metals with the same number of ligand (combineMetal) shows a better performance than combining different ligand numbers of the same metal (combineNumLig), even though they both enlarge the size of the group. We believe this is partially due to how the 6-angle space collapses angle information from full angle spaces of different dimensionality. Also for a given number of ligand, there are only a fixed number of possible canonical CGs and thus less diversity, even with different metals together. It is interesting though that these different metals exhibit similar functional trends as long as they have similar sets of CGs. This may imply that different metals are somewhat interchangeable as long as the structure remains the same, and that the structures have higher impact on functions than the metal itself. It also provides evidence that we can combine metals with the same ligand numbers in analyzing the less abundant metals and thus have enough data to determine full structure-function correlations (rho's).

A simulation on the 4-ligand normal zinc sites exhibits the same trend. A series of subsets of the data were sampled without replacement. The sizes of the subset sequence were selected as 0.5/20, 0.75/20, 1/20, 1.25/20, 1.5/20, 1.75/20, 2/20, 3/20, 4/20, … of the original data, and each size was repeated for 20 times. k= 9, 11, and 13 were used for all subsets to acquire the rho. As shown in Figure 5.8, the average rho increases as the size grows regardless of the selected k. That is, in order to detect a decent correlation between structural and functional distance metrics, at least 1000 non-redundant metal binding sites is required. That is why only until the last few years, there was adequate structural data available to reliably detect the existence of compressed angles in CGs (16).

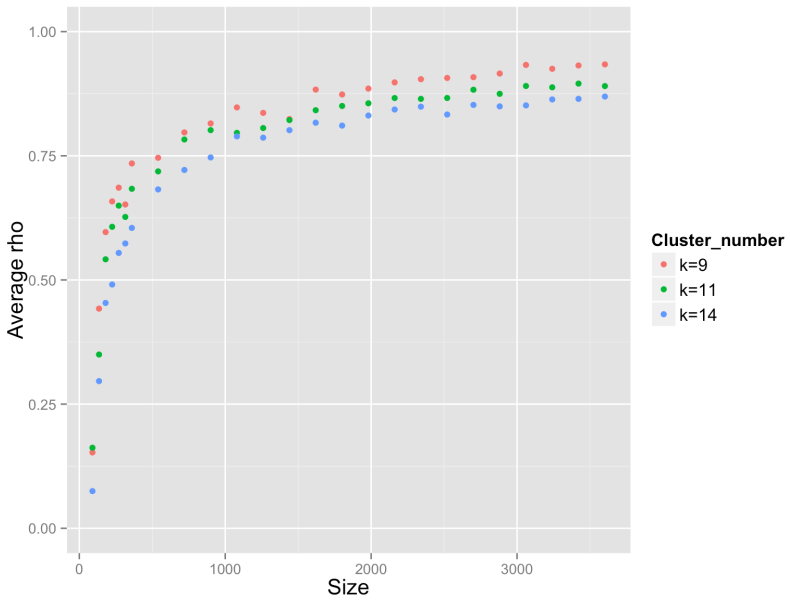
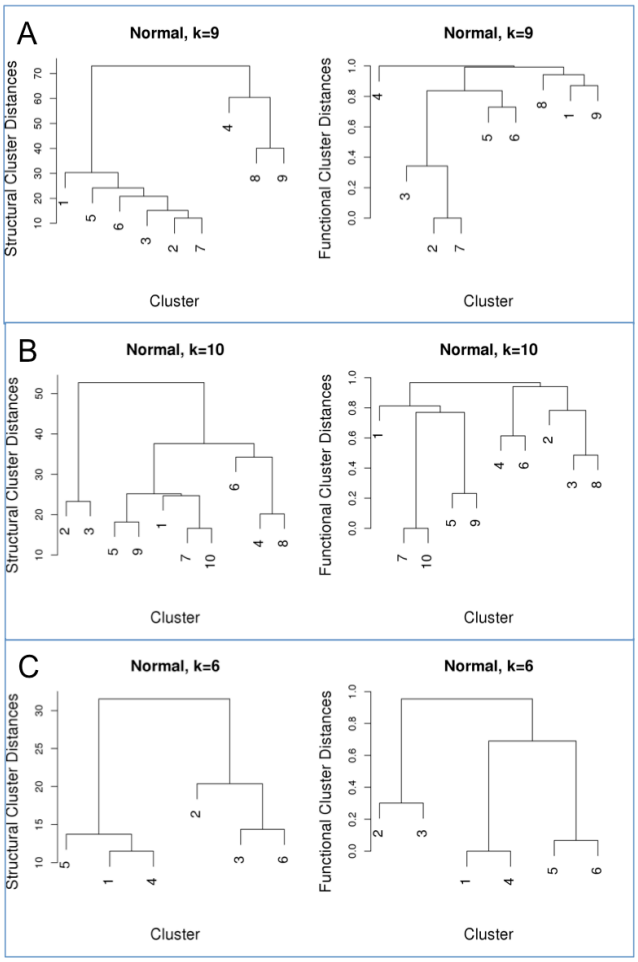


Figure 5.8 Simulation of rho verse size relation on 4-ligand normal-group zinc.

For all categories, the cluster centers and a characteristic average probability of each cluster are computed, together with the metal ids in each cluster. Figure 5.9.A-C uses normal zinc metalloproteins as examples to illustrate the structural vs. functional dendrogram comparison. The average probabilities for each cluster with respect to appropriate canonical CG models (Figure 5.9.D-F) provides a characterization of each cluster with respect to canonical CG models, with the highest canonical CG model probability for each cluster shaded. According to the χ^2 probabilities for 4-ligand zinc (Figure 5.9, panel A and D), clusters 2, 3, 5, 6, and 7 are all sub-classes of tetrahedral CG, which are also well clustered together based on both structural and functional distances. Cluster 1 is slightly distorted Tetrahedral CG due to its low probability in Tet. It is structurally closer to the other Tet clusters, while functionally closer to clusters 8 and 9, which is a mixture of Bvp and Spv, and distorted Bva respectively. Cluster 4 shows the

highest probability for Bvp, and is the furthest from other clusters both structurally and functionally. As for 5-ligand Zn metalloproteins (Figure 5.9, panel B and E), clusters 1, 5, 7, 9, and 10 are all classified as Tbp according to the χ^2 probability, and they are also grouped together in both structural and functional dendrograms nicely. Clusters 2 and 3 are both identified as square pyramidal (Spy), and clusters 4 and 6 are both identified as trigonal prismatic vacancy (Tpv). They all show very high similarities in the dendrograms. Similarly for 6-ligand Zn metalloproteins (Figure 5.9, panel C and F), clusters 1, 4, 5, and 6 are all identified as Oct with high probability. And the lower Oct probability cluster 3 is grouped together with cluster 2 as it is showing a slightly bigger separation from the other Oct clusters, especially in functional dendrogram. All these figures demonstrate that our CG cluster representations have very strong functional implications, as the structural and functional distances were calculated independently from different sources of information. And it is only through the CG clusters that this level of similarity is observed in the dendrograms. Likewise, similar dendrograms and patterns can be found for the rest of the metals.



D 4-ligand Zn metalloproteins

	Size	Tet	Bva	Bvp	Spv	Spl
1	196	0.185	0.008	0.006	0.004	0
2	791	0.773	0.012	0	0	0
3	771	0.686	0.054	0	0	0
4	46	0.001	0.002	0.165	0.009	0.059
5	390	0.23	0.113	0.002	0.004	0
6	310	0.4	0.067	0	0	0
7	765	0.936	0.008	0	0	0
8	184	0	0.003	0.209	0.283	0.002
9	149	0.006	0.095	0.002	0.011	0

E 5-ligand Zn metalloproteins

	Size	Tbp	Spy	Tpv
1	32	0.062	0	0.047
2	51	0.005	0.707	0.006
3	36	0.006	0.122	0.037
4	126	0.058	0.022	0.262
5	84	0.294	0.002	0.197
6	46	0.028	0.001	0.341
7	103	0.564	0.018	0.094
8	91	0.119	0.279	0.089
9	69	0.165	0.019	0.123
10	102	0.764	0	0.118

F 6-ligand Zn metalloproteins

	size	Oct	Tpr	Pva	Pvp
1	53	0.924	0	0	0.002
2	20	0.041	0.059	0	0.138
3	33	0.194	0.001	0	0.06
4	43	0.747	0	0	0.013
5	57	0.577	0.001	0	0.023
6	50	0.441	0.002	0	0.082

Figure 5.9 Three examples of structural versus functional dendrograms of clusters. (A, D) 4-ligand normal zinc metalloproteins. (B, E) 5-ligand normal zinc metalloproteins. (C, F) 6- ligand normal zinc metalloproteins.

Table 5.5. Instances of highly aberrant clusters of the compressed group for different metals.

Metal	Ligand Number	Cluster Number	Size	Angle 1	Angle 2	Angle 3	Angle 4	Angle 5	Angle 6	Tet	Bva	Bvp	Spv	Spl
Zn	4	4	7	158.8+/-11.6	57.2+/-4.6	71.8+/-7.9	117.3+/-11	133.4+/-8.6	85.6+/-5.9	0	0.002	0.004	0	0
Mg	4	7	13	153.5+/-9.7	55.2+/-4	74.2+/-9.4	105.2+/-14.8	131.3+/-7.3	116.6+/-10.8	0	0	0	0	0
Fe	4	4	10	143.9+/-11.4	60.8+/-5	88.1+/-7.8	103.3+/-8.1	125.1+/-6.9	93.3+/-10.5	0	0	0	0	0
										Tbp	Spy	Tpv		
Zn	5	2	75	164.5+/-5.5	55.7+/-3.9	87.9+/-4	104.4+/-3.5	124.4+/-5.5	102.8+/-5.5	0.096	0.003	0.063		
Mg	5	2	16	149.5+/-5.4	54.9+/-4	76.8+/-4.4	90+/-5.4	136.6+/-5.7	66.3+/-3.8	0	0	0.015		
Ca	5	11	17	140.9+/-7.7	55+/-5.6	73.7+/-4.4	86.3+/-5.5	129.5+/-5.3	65.3+/-6.6	0.003	0.008	0.062		
Fe	5	7	29	147.2+/-5.1	56.8+/-3.6	91.5+/-4.2	101.1+/-3.9	138.4+/-4.7	94.1+/-4.3	0	0	0.001		
										Oct	Tpr	Pvp	Pva	
Mg	6	1	10	148.1+/-3.7	65+/-4.5	78+/-5.2	108+/-10.1	140+/-5	54.8+/-3.1	0	0.004	0.002	0.01	
Ca	6	11	12	152.2+/-11.1	53.5+/-5.4	73.1+/-5.3	99.3+/-5.6	135.1+/-4.8	61.1+/-5.8	0.007	0.007	0.036	0.068	
Fe	6	2	18	165.8+/-4.9	61.7+/-4.2	89.5+/-2.6	102.7+/-3.2	159.1+/-5.1	66.4+/-4.2	0.001	0	0	0.004	
Na	6	7	3	168.8+/-3.9	68.9+/-2.2	96.1+/-2.5	107+/-3.9	138.1+/-4.9	46.5+/-4.6	0.054	0.131	0.001	0.096	

5.3.5 Aberrant CG Clusters

Of the 18756 non-redundant metal binding sites analyzed, roughly 24% contain compressed angles. While coordination geometries that contain unexpected compressed angles would be considered aberrant, some CG clusters are clearly highly aberrant with low similarity to any canonical CGs. Table 5.5 is a compilation of such highly aberrant CG clusters from the full CG cluster description tables in Supplemental Material. They can be found in all 4- to 6-ligand metals. Some of the clusters may have a small cluster size, which it is mainly due to the nature of small size compress group in general. As more and more data is accumulated, it can be expected that an increasing amount of these aberrant metal sites will be detected as well, processing similar structural distortions and functional propensities. 7- and 8-ligand metal sites tend to be less distorted from canonical CGs. This is primarily due to that 7- and 8-ligand CGs have small ideal angles naturally because of the crowed ligand space around the metal. And thus, compressed angle is expected for such metal sites.

5.4 Conclusion

In this chapter, we further developed the methods from Chapter 4, and successfully applied the methodology to top five abundant metals. The compressed angle and aberrant CG phenomenon we observed in chapter 4 remained true for all five metalloproteins and all numbers of ligands. The using of a series of additional filters greatly benefited us in improving data quality, and resulted in an increase in detected structure-function correlation from 0.88 to 0.93 for 4-ligand normal zinc sites. This chapter showed a great effect of the data size over the detectable correlation. As we start

analyzing less abundant metal elements, fewer and fewer sits will be available. On one hand, this is a warning that as we might have to combine metals with the same number of ligands in future analysis. On another, this further supported our statements that no such analysis was feasible before simply due to inadequate data. And only until recently, as the structural data has been rapidly growing, we can start to evaluate the compressed angle from a boarder perspective, and to study its influence on proteins.

CHAPTER 6

DISCUSSION

6.1 Evaluation of the Program Performance

The mainstream analyses are composed of a sequence of individual programs. The bootstrapping and IA process were written in Perl, and the rest were in R. Most of the calculations were conducted on machines with Intel® Core™ i7-4930K Processor, 32g RAM, and Linux system. All programs are designed to take parameters such as metal or distance cutoff from command line arguments, therefore the entire process can be accomplished via a bash script.

6.1.1 Bootstrapping and IA Analysis

The bootstrapping and IA steps are implemented together in an object-oriented style. A simplified diagram to show the relationship between the entities is in Figure 6.1. *MPCGanalysis* is the process object, and is called by the main program to fulfill the entire analysis. The general procedure is to first parse the PDB entries and convert each ATOM record into an *Atom* object, followed by creating the *MetalShell* object of the initial ligand set for each metal site. The collection of *MetalShell* is stored in *MPCGanalysis* as 'shells' attribute and are the object to be analyzed for the rest of the analysis. They are then calculated as *Coordination* object for bootstrapping or IA by

considering all possible combinations of the ligand sets and picking the best combo from them. Each CG is an individual object as a child of *Coordination*, so that it is easy to add extra CGs into the analysis.

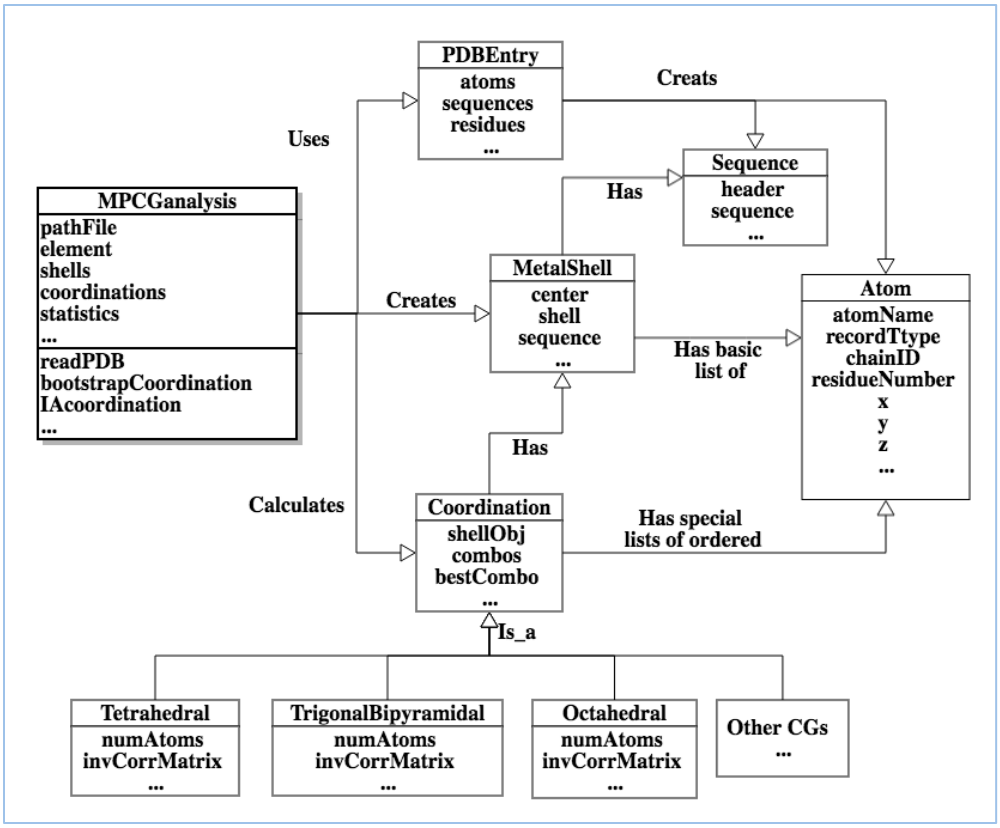


Figure 6.1 The objected-oriented diagram for bootstrapping and IA analysis.

For time complexity, if we denote *M* as the number of PDB entries, *N* as the number of total metal sites, *A* as the average symmetry calculation for each PDB entries, *B* as the total number CG calculations need to be computed for each metal site, and *C* as the number of iterations in the IA step before it converges.

The time complexity is $\Theta(A*M)$ for the parsing of the PDB files and the calculation of all symmetry-related coordinates. Similarly, it is $\Theta(A*N)$ time complexity for obtaining an initial fc-shell, as all atoms including symmetry-related ones needs to be scanned in order to find the set of ligands that are within a distance cutoff. A large A value would greatly increase the search space, so the larger the A value is, the longer both processes will take. M and N can be found in Table 5.3. Different metals have different A values, but generally are between 3 and 5. For the symmetry-related calculation in Chapter 5, we only considered the crystallographic symmetry if it overlaps with the metal shell. Often A would be over 100 instead of 3 to 5 if all possible crystallographic symmetries were considered. Thus, pre-filtering crystallographic symmetries before fully calculating them is a huge computational time saving, since factor A needs to be applied to all the X-ray structures of PDB entries, which is ~90% of the whole database. A can be considered as a coefficient in this step, but due to the uncertainty of the crystallographic symmetry of other metalloproteins, it is kept in the big Θ notation.

The bootstrapping step has a time complexity of $\Theta(B * N)$. As shown in Table 6.1, if only considering the all-to-all mapping from potential ligands to ideal CG ligand positions, the number of calculations is a permutation of the number of ligands, i.e. $P(n,n)$ for n-ligand CGs. Due to the internal structural symmetry of the CG, a non-redundant calculation required for actual calculation is shown in the third column, which can be denote as T. Then for a potential k-ligand shell (column), the number of different ways to get n ligands (row) out is $C(k, n)$, which makes the total $T*C(k, n)$. Column 4-8 shows the actual number of calculation for each number of potential ligands to compute each CG. In the bootstrapping step, only major CGs are considered, i.e. Tet, Tbp, Oct, and Pbp. So for

example, if there are seven potential ligands in the shell, the calculation number is 35+210+105+252 = 602, according to the shaded cells in the table.

Table 6.1 Number of calculations for each CGs given different potential ligand numbers in the metal fc-shell.

CG	All-to-all mapping	Non-redundant equivalent	4 potential ligands	5 potential ligands	6 potential ligands	7 potential ligands	8 potential ligands
4-ligand:							
Tet	24	1	1	5	15	35	70
Bva	24	4	4	20	60	140	280
Bvp	24	6	6	30	90	210	420
Pyv	24	6	6	30	90	210	420
Spl	24	3	3	15	45	105	210
5-ligand:							
Tbp	120	10		10	60	210	560
Spy	120	15		15	90	315	840
Tpv	120	60		60	360	1260	3,360
6-ligand:							
Oct	720	15			15	105	420
Pva	720	72			72	504	2,016
Pvp	720	180			180	1,260	5,040
Tpr	720	60			60	420	1,680
7-ligand:							
Pbp	5,040	252				252	2,016
Hva	5,040	420				420	3,360
Hvp	5,040	1,260				1,260	10,080
Sav	5,040	2,520				2,520	20,160
8-ligand:							
Hbp	4,0320	1,680					1,680
Sqa	4,0320	2,520					2,520

The time complexity of the IA step is Θ ($B * C * N$). It is similar to the bootstrapping step, except the B value is generally much larger as we consider all of the

CGs in Table 6.1. Also, there a *C* portion as the number of iteration the IA process went through before the statistics converged. The *C* values for Zn, Ca, and Na are 7, 10, and 12 respectively. As for Mg and Fe, the iteration failed to converge within 15 rounds, which is the maximum we allowed.

The actual time for different types of metalloproteins varied drastically at each step. For the PDB parsing and ligand identification, it took Zn about five hours, while took Mg roughly 24 hours on the machine we used. That is mainly due to the number of actual Mg ions is much large and the crystallographic symmetry number is also higher for Mg. For the bootstrapping and IA steps, the higher the number of potential ligands, the larger *B* will be, and the longer it takes. For Ca metalloproteins, which have a higher number of ligands than others (Table 5.4), it took a couple of days to run one iteration of IA at first. So we parallelized the IA step and optimized some of the calculation, and now it takes Ca about six hours per iteration.

As for the memory usage, it is greatly affected by the number of PDB entries *M*, the number of metal sites *N*, and the average crystallographic symmetry *A*. The relation of objects (Figure 6.1) creates some redundancy when calculating over the same collections of ligands repeatedly. Yet the program can normally run on a computer with 32 gigabytes (GB) of RAM without any problems.

Only a minimal amount of optimizations have been performed both on the time and the memory usage, as long as the program finished in a reasonable time with 32 GB of RAM. That is mainly because the focus of this project so far has been on the discovery of a novel phenomenon rather than providing a service. For analysis on the rest of the metalloprotein species, the number of PDB entries *M* and number of metal sites *N* are

both smaller than the top five metals. But there are the unknown factor of the crystallographic symmetry *A* and the number of potential ligands *B* that may cause the other metalloproteins to take longer time to finish. Also if we re-examine the top five metalloproteins using a newer set of data, *M* and *N* will be larger. Many optimization actions can be done to enhance the performance, such as reducing the redundancy of copying and storing the *MetalShell* object, which may become more crucial when we want to re-examine the top five metalloproteins after more data become available or when converting our work into a web-server or stand-alone tools for others to use.

6.1.2 Other Analysis

All other steps were based on the list of metal fc-shells bootstrapping step defined, which is much smaller after the additional filters and the non-redundant test. If we denote the number in the final non-redundant metal list to be *P*, most of the time complexities are simply $\Theta(P)$. For k-means stability test step, there is also a factor of iteration, which is set to be 500 in this study. For the rest of the metalloproteins with lower counts, these steps should be easily feasible. Though the processes that took a relatively long running time and large memory space can be considered for further optimization in order to minimize the requirement of both time and space.

6.2 Future Directions

6.2.1 Other Metalloproteins

The most obvious and immediate next step is to analyze all metalloproteins in Table 5.3. Chapter 5 already showed the applicability of our methodology to different metals. The presence of a compressed angle and its influence on the complication of CG

classification are expected in all metalloproteins. Parameters such as the bond-length cutoff need to be tuned for each metalloprotein species individually to reduce the false positive and false negative ligand detection rates. The potential problem is that for the some low count metals, the data size is not large enough to achieve a detection limit needed for structure-function analyses. We have shown in Chapter 5 that it is possible to combine the same number of ligands of different metals, and they revealed strong structure-function correlation as well. Though this requires further assessment. We also need to develop more efficient filters to account for potential errors like non-specific binding that is more common for lower-count metalloproteins. For example, some heavy metal compound, involving Hg, Pt, or Au, are very common additives in helping protein phasing and crystallizes [114-116], and thus can be found in many protein structures but is not physiologically bound. So special attention is required when analyzing heavy metals such as Hg to make sure our filters can differentiate binding specificity. It is also beneficial to revisit the already analyzed metals when significantly more data is available.

6.2.2 Functional Application

6.2.2.1 Nucleotide Polymorphisms Affect Metalloproteins

A Nucleotide Polymorphism (NP) is a variation in the genomic DNA sequence. Insertions and deletions are typically referred to as indels. The most common NP are single nucleotide polymorphisms (SNPs), which involve a variation in only a single DNA nucleotide base pair (bp), with an estimation of one SNP in about every 1200 bp [117]. Benign SNPs are responsible for various phenotypic differences in humans [118]. Yet some SNPs can cause harmful changes to proteins [119], and may lead to complex disorders or increase the susceptibility to disease [120-122]. SNPs that result in amino

acid changes in proteins are called non-synonymous SNPs (nsSNPs). They are a main type of SNP that can cause disease. One of the seminal tasks in NP analyses is to find the causal relationship between NPs and diseases.

Since sequence dictates structure, which enables function, disease-associate NPs often cause changes in the manifestation of gene function via structural perturbations in a protein gene product. However, interpreting NPs with respect to structural perturbations has been very challenging. Thus, our metalloprotein analyses could provide an enhanced structure-function interpretation of NPs that occur near metal binding sites. Given the ubiquity of metalloproteins in the human proteome, NPs in metalloproteins can help explain many disease-associated mutations, especially when they happen near the metal binding area. For example, the tumor suppressor protein p53 is crucial in regulating the cell cycle, and nsSNPs on p53 are found at an abnormally high frequency in many cancers [123-125]. A summarization of the fifty most frequent p53 cancer mutants-from positions of the IARC TP53 Mutation Database (Release 18) [126] is shown in Table 6.2. Based on our initial analysis, p53 (PDB ID: 2OCJ) contains a zinc ion with coordinating ligands C176, H179, C238, and C242 (red in Table 6.2), which are all found to have known mutations as shown in red. Moreover, if we define a binding domain as five additional residues around binding ligands, which is 171-184 & 233-247 (shaded), the binding domain of p53 may account for up to 20.13% of the total SNP occurrence while it only made up of 10% of the sequence length (289). Moreover, two positions immediately next to our defined range, R248 and R249, rank 1 and 6 respectively, and they count for another 10% of the SNP occurrences by themselves.

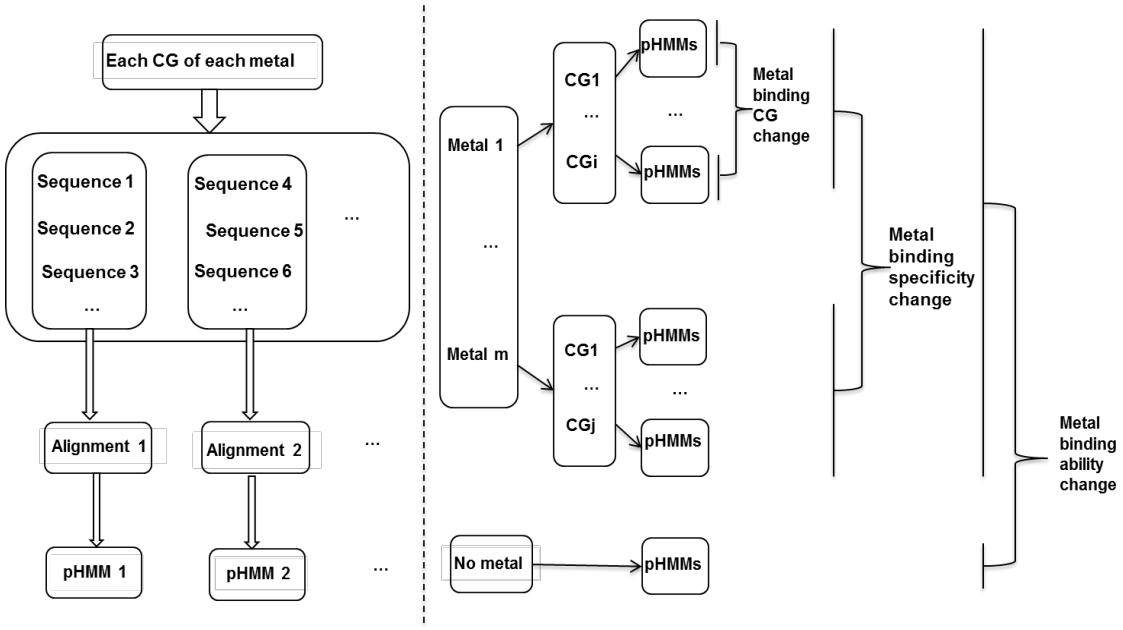


Figure 6.2 The construction and interpretation of pHMMs.

Table 6.2 Top 50 somatic mutation positions of p53 protein based on IARC TP53 Mutation Database (Release 18).

Mutate from	Rank	Count	%	Mutate from	Rank	Count	%
R248	1	1,925	7.55	Y234	26	228	0.89
R273	2	1,826	7.16	C238	27	226	0.89
R175	3	1,328	5.21	S241	28	220	0.86
G245	4	869	3.41	K132	29	213	0.84
R282	5	751	2.95	V272	30	203	0.8
R249	6	700	2.75	Y205	31	198	0.78
Y220	7	469	1.84	C275	32	187	0.73
H179	8	447	1.75	C141	33	183	0.72
R213	9	439	1.72	D281	34	183	0.72
C176	10	400	1.57	I195	35	179	0.70
P278	11	320	1.26	E286	36	177	0.69
R158	12	312	1.22	R306	37	171	0.67
R280	13	302	1.18	Y236	38	171	0.67
R196	14	298	1.17	P152	39	166	0.65
V157	15	286	1.12	E258	40	161	0.63
G266	16	268	1.05	M246	41	158	0.62
Y163	17	260	1.02	A159	42	146	0.57
G244	18	258	1.01	A161	43	145	0.57
E285	19	255	1.00	Q192	44	141	0.55
H193	20	255	1.00	L194	45	134	0.53
C135	21	249	0.98	W146	46	133	0.52
P151	22	243	0.95	N239	47	127	0.5
M237	23	240	0.94	P250	48	122	0.48
C242	24	234	0.92	T155	49	122	0.48
V173	25	228	0.89	V216	50	121	0.47
				Sum		17,377	68.15

6.2.2.2 Interpret Sets of Disease-Associated SNPs via Profile Hidden Markov Models

A Profile Hidden Markov Models (pHMM) is a position-specific probabilistic scoring model that captures uniformity at each sequence position [127-129]. Methods for building pHMMs have been well-established and used by major protein domain sequence databases [130, 131], many of which are included in the aggregate pHMM protein family database Pfam [78]. As shown in Figure 6.2, with the direct structural and functional associations of CG clusters and sub-clusters, we can build pHMMs for every CG

associated with specific metal ions. Sequences of each CG structural/functional cluster and sub-cluster for each metal will first be aligned based on sequence similarity. We can use HMMER [130] to build pHMMs for each multi-alignment. We will then organize the constructed pHMMs by metal type, CG cluster, and functional CG sub-clusters. By comparing pHMMs from different categories, we can predict when nsNPs at various positions in metal-binding sites produce structural alterations that lead to changes in coordination geometry, to change of metal binding specificity, or to loss of metal binding ability. We can also compare results of pHMMs applied to specific disease associated metal binding NPs and their corresponding common NPs. These predicted, functionally significant positions could be further validated by biological experiments for their metal binding or specificity related functions.

CHAPTER 7

CONCLUSION

To sum up this research, a novel methodology was developed to analyze the structure of metalloproteins, especially the coordination geometry, and its relationship to biochemical and biological functions. It was designed in a way where no prior assumptions of existing CG models were needed after the initial bootstrap step, allowing the CG models present to be derived from the data itself. The core of the methodology relies on the low variance of bond-length, enabling the determination of binding ligands via a statistical test with several additional refinements that improve the overall performance of the method. Our ligand detection method is statistically rigorous, producing a low estimated false positive rate of 0.26% and false negative rate of 1.2%. Metal shells were separated into normal and compressed groups, and clustered by k-means independently. Individual clusters were examined and assigned via probabilistic comparison to characteristic canonical CG models. The overall clustering results were evaluated based on the strength (correlation) of their function-structure relationships. Many clusters were easily associated with a particular canonical CG via high probabilistic matching to the canonical CG models. Whereas, there were other clusters that represented either slightly distorted canonical CGs or highly aberrant or novel CGs. By recognizing these aberrant CGs in clustering, high correlations were achieved

between structural and functional descriptions of metal ion coordination. As the wwPDB continues to grow, additional aberrant or novel CGs may become apparent; however, we will only be able to detect these previously uncharacterized CGs if we stop assuming that we already knew what CGs a dataset contains before analyzing it. These CG clustering results also points the way to a future examination of the impact of SNPs on structural perturbations that lead to changes in protein function.

REFERENCE

1. Bertini I, Sigel A, Sigel H: **Handbook on metalloproteins**. New York: Marcel Dekker; 2001.
2. Miller J, McLachlan AD, Klug A: **Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes**. *The EMBO journal* 1985, **4**(6):1609-1614.
3. Tan D, Zhou M, Kiledjian M, Tong L: **The ROQ domain of Roquin recognizes mRNA constitutive-decay element and double-stranded RNA**. *Nature structural & molecular biology* 2014.
4. Rowlett RS: **Structure and catalytic mechanism of beta-carbonic anhydrases**. *Sub-cellular biochemistry* 2014, **75**:53-76.
5. Bellomo E, Massarotti A, Hogstrand C, Maret W: **Zinc ions modulate protein tyrosine phosphatase 1B activity**. *Metallomics : integrated biometal science* 2014, **6**(7):1229-1239.
6. Maret W: **Metalloproteomics, metalloproteomes, and the annotation of metalloproteins**. *Metallomics : integrated biometal science* 2010, **2**(2):117-125.
7. Maret W: **Zinc biochemistry: from a single zinc enzyme to a key element of life**. *Advances in nutrition* 2013, **4**(1):82-91.

8. Andreini C, Cavallaro G, Lorenzini S, Rosato A: **MetalPDB: a database of metal sites in biological macromolecular structures**. *Nucleic acids research* 2013, **41**(Database issue):D312-319.

9. Andreini C, Bertini I, Rosato A: **Metalloproteomes: a bioinformatic approach**. *Accounts of chemical research* 2009, **42**(10):1471-1479.

10. Levi S, Finazzi D: **Neurodegeneration with brain iron accumulation: update on pathogenic mechanisms**. *Frontiers in pharmacology* 2014, **5**:99.

11. Miller Y, Ma B, Nussinov R: **Zinc ions promote Alzheimer Abeta aggregation via population shift of polymorphic states**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(21):9490-9495.

12. Du X, Zheng Y, Wang Z, Chen Y, Zhou R, Song G, Ni J, Liu Q: **Inhibitory Act of Selenoprotein P on Cu/Cu-Induced Tau Aggregation and Neurotoxicity**. *Inorganic chemistry* 2014.

13. Pearson E: **Zinc transport and diabetes risk**. *Nature genetics* 2014, **46**(4):323-324.

14. Montonen J, Boeing H, Steffen A, Lehmann R, Fritsche A, Joost HG, Schulze MB, Pischon T: **Body iron stores and risk of type 2 diabetes: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study**. *Diabetologia* 2012, **55**(10):2613-2621.

15. Stevens RG, Cologne JB, Nakachi K, Grant EJ, Neriishi K: **Body iron stores and breast cancer risk in female atomic bomb survivors**. *Cancer science* 2011, **102**(12):2236-2240.

16. Kim YR, Kim IJ, Kang TW, Choi C, Kim KK, Kim MS, Nam KI, Jung C: **HOXB13 downregulates intracellular zinc and increases NF-kappaB signaling to promote prostate cancer metastasis**. *Oncogene* 2013.

17. Surade S, Blundell TL: **Structural biology and drug discovery of difficult targets: the limits of ligandability**. *Chem Biol* 2012, **19**(1):42-50.

18. Puerta DT, Schames JR, Henchman RH, McCammon JA, Cohen SM: **From model complexes to metalloprotein inhibition: a synergistic approach to structure-based drug discovery**. *Angew Chem Int Ed Engl* 2003, **42**(32):3772-3774.

19. Andreini C, Bertini I: **A bioinformatics view of zinc enzymes**. *Journal of inorganic biochemistry* 2012, **111**:150-156.

20. Harding MM: **The geometry of metal-ligand interactions relevant to proteins. II. Angles at the metal atom, additional weak metal-donor interactions**. *Acta crystallographica Section D, Biological crystallography* 2000, **56**(Pt 7):857-867.

21. Harding MM: **The geometry of metal-ligand interactions relevant to proteins**. *Acta crystallographica Section D, Biological crystallography* 1999, **55**(Pt 8):1432-1443.

22. Yao S, Flight RM, Rouchka EC, Moseley HN: **A less-biased analysis of metalloproteins reveals novel zinc coordination geometries**. *Proteins* 2015, **83**(8):1470-1487.

23. Valasatava Y, Rosato A, Cavallaro G, Andreini C: **MetalS(3), a database-mining tool for the identification of structurally similar metal sites**. *Journal of*

biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry 2014, **19**(6):937-945.

24. Patel K, Kumar A, Durani S: **Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures**. *Biochimica et biophysica acta* 2007, **1774**(10):1247-1253.

25. Harding MM, Hsin KY: **Mespeus--a database of metal interactions with proteins**. *Methods in molecular biology* 2014, **1091**:333-342.

26. Liu Z, Wang Y, Zhou C, Xue Y, Zhao W, Liu H: **Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins**. *Biochimica et biophysica acta* 2014, **1844**(1 Pt B):171-180.

27. Tus A, Rakipovic A, Peretin G, Tomic S, Sikic M: **BioMe: biologically relevant metals**. *Nucleic acids research* 2012, **40**(Web Server issue):W352-357.

28. Harding MM: **Small revisions to predicted distances around metal sites in proteins**. *Acta crystallographica Section D, Biological crystallography* 2006, **62**(Pt 6):678-682.

29. Werden SJ, McFadden G: **The role of cell signaling in poxvirus tropism: the case of the M-T5 host range protein of myxoma virus**. *Biochimica et biophysica acta* 2008, **1784**(1):228-237.

30. Lebiedzinska M, Suski J, Duszynski J, Wieckowski MR: **Role of the p66Shc protein in physiological state and in pathologies**. *Postepy Biochem* 2010, **56**(2):165-173.

31. Kim AK, DeRose R, Ueno T, Lin B, Komatsu T, Nakamura H, Inoue T: **Toward total synthesis of cell function: Reconstituting cell dynamics with synthetic biology**. *Sci Signal* 2016, **9**(414):re1.

32. Bourne PE, Weissig H: **Structural bioinformatics**. Hoboken, N.J.: Wiley-Liss; 2003.

33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic acids research* 2000, **28**(1):235-242.

34. Berman H, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank**. *Nature structural biology* 2003, **10**(12):980.

35. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG *et al*: **CATH: comprehensive structural and functional annotations for genome sequences**. *Nucleic acids research* 2015, **43**(Database issue):D376-381.

36. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG: **SCOP2 prototype: a new approach to protein structure mining**. *Nucleic acids research* 2014, **42**(Database issue):D310-314.

37. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *Journal of molecular biology* 1995, **247**(4):536-540.

38. Nelson DL, Cox MM, Lehninger AL: **Lehninger principles of biochemistry**, 4th edn. New York: W.H. Freeman; 2005.

39. McNaught AD, Wilkinson A, Jenkins AD, International Union of Pure and Applied Chemistry.: **IUPAC compendium of chemical terminology the gold book**. In., 1.0.0. edn. Online corrected version: International Union of Pure and Applied Chemistry; 2006: "Chirality".

40. Klein DR: **Organic chemistry as a second language : first semester topics**, 3rd edn. Hoboken, NJ: Wiley; 2012.

41. Reusch W: **Virtual Textbook of Organic Chemistry**. In.; 1999: Proteins and Amino Acids.

42. Roy A, Kucukural A, Zhang Y: **I-TASSER: a unified platform for automated protein structure and function prediction**. *Nat Protoc* 2010, **5**(4):725-738.

43. Piccoli S, Suku E, Garonzi M, Giorgetti A: **Genome-wide Membrane Protein Structure Prediction**. *Curr Genomics* 2013, **14**(5):324-329.

44. Hargittai B, Hargittai In: **Culture of chemistry : the best articles on the human side of 20th-century chemistry from the archives of the Chemical Intelligencer**. New York: Springer; 2015.

45. Kim S, Kim JH, Lee JS, Park CB: **Beta-Sheet-Forming, Self-Assembled Peptide Nanomaterials towards Optical, Energy, and Healthcare Applications**. *Small* 2015, **11**(30):3623-3640.

46. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling**. *Electrophoresis* 1997, **18**(15):2714-2723.

47. Burkhard P, Stetefeld J, Strelkov SV: **Coiled coils: a highly versatile protein folding motif**. *Trends Cell Biol* 2001, **11**(2):82-88.

48. Scott MS, Thomas DY, Hallett MT: **Predicting subcellular localization via protein motif co-occurrence**. *Genome Res* 2004, **14**(10A):1957-1966.

49. Wriggers W, Schulten K: **Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates**. *Proteins* 1997, **29**(1):1-14.

50. Kellogg EH, Lange OF, Baker D: **Evaluation and optimization of discrete state models of protein folding**. *J Phys Chem B* 2012, **116**(37):11405-11413.

51. Dill KA, MacCallum JL: **The protein-folding problem, 50 years on**. *Science* 2012, **338**(6110):1042-1046.

52. Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popovic Z, Baker D, Players F: **Algorithm discovery by protein folding game players**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(47):18949-18953.

53. Chang CC, Tey BT, Song J, Ramanan RN: **Towards more accurate prediction of protein folding rates: a review of the existing Web-based bioinformatics approaches**. *Briefings in bioinformatics* 2015, **16**(2):314-324.

54. Sheng Y, Qiu X, Zhang C, Xu J, Zhang Y, Zheng W, Chen K: **Quad-PRE: a hybrid method to predict protein quaternary structure attributes**. *Comput Math Methods Med* 2014, **2014**:715494.

55. Poupon A, Janin J: **Analysis and prediction of protein quaternary structure**. *Methods in molecular biology* 2010, **609**:349-364.

56. Berman HM, Westbrook JD, Gabanyi MJ, Tao W, Shah R, Kouranov A, Schwede T, Arnold K, Kiefer F, Bordoli L *et al*: **The protein structure initiative**

structural genomics knowledgebase. *Nucleic acids research* 2009, **37**(Database issue):D365-368.

57. Vleck JHV: **Theory of the variations in paramagnetic anisotropy among different salts of the iron group**. *Physical Review* 1932, **41**:208.

58. Shriver DF, Atkins PW, Salvador P: **Inorganic chemistry**, 4th edn. New York: W.H. Freeman; 2006.

59. Zastrow ML, Pecoraro VL: **Designing hydrolytic zinc metalloenzymes**. *Biochemistry* 2014, **53**(6):957-978.

60. Andreini C, Cavallaro G, Lorenzini S: **FindGeo: a tool for determining metal coordination geometry**. *Bioinformatics* 2012, **28**(12):1658-1660.

61. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic acids research* 2007, **35**(Database issue):D61-65.

62. UniProt C: **UniProt: a hub for protein information**. *Nucleic acids research* 2015, **43**(Database issue):D204-212.

63. Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF *et al*: **The protein information resource (PIR)**. *Nucleic acids research* 2000, **28**(1):41-44.

64. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S *et al*: **The Protein Data Bank**. *Acta crystallographica Section D, Biological crystallography* 2002, **58**(Pt 6 No 1):899-907.

65. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I: **UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View**. *Methods in molecular biology* 2016, **1374**:23-54.

66. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic acids research* 2003, **31**(1):365-370.

67. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P *et al*: **PDBe: improved accessibility of macromolecular structure data from PDB and EMDB**. *Nucleic acids research* 2016, **44**(D1):D385-395.

68. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z *et al*: **BioMagResBank**. *Nucleic acids research* 2008, **36**(Database issue):D402-408.

69. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley DM, Nakagawa A *et al*: **Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format**. *Nucleic acids research* 2012, **40**(Database issue):D453-460.

70. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server**. *Trends in biochemical sciences* 1998, **23**(9):358-361.

71. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state**. *Journal of molecular biology* 2007, **372**(3):774-797.

72. Andreini C, Cavallaro G, Rosato A, Valasatava Y: **MetalS2: a tool for the structural alignment of minimal functional sites in metal-binding proteins and nucleic acids**. *Journal of chemical information and modeling* 2013, **53**(11):3064-3075.

73. Zheng H, Chordia MD, Cooper DR, Chruszcz M, Muller P, Sheldrick GM, Minor W: **Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server**. *Nat Protoc* 2014, **9**(1):156-170.

74. Choi H, Kang H, Park H: **MetLigDB: a web-based database for the identification of chemical groups to design metalloprotein inhibitors**. *Journal of Applied Crystallography* 2011, **44**(4):878-881.

75. 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 research* 2002, **30**(1):379-382.

76. Degtyarenko K, Contrino S: **COMe: the ontology of bioinorganic proteins**. *BMC Struct Biol* 2004, **4**:3.

77. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S *et al*: **The InterPro protein families database: the classification resource after 15 years**. *Nucleic acids research* 2015, **43**(Database issue):D213-221.

78. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic acids research* 2014, **42**(Database issue):D222-230.

79. Attwood TK: **The PRINTS database: a resource for identification of protein families**. *Briefings in bioinformatics* 2002, **3**(3):252-263.

80. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation**. *Nucleic acids research* 2010, **38**(Database issue):D161-166.

81. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D: **The ProDom database of protein domain families: more emphasis on 3D**. *Nucleic acids research* 2005, **33**(Database issue):D212-215.

82. Sillitoe I, Lewis T, Orengo C: **Using CATH-Gene3D to Analyze the Sequence, Structure, and Function of Proteins**. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2015, **50**:1 28 21-21.

83. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cuche BA, Bougueleret L, Poux S *et al*: **HAMAP in 2015: updates to the protein family classification and annotation system**. *Nucleic acids research* 2015, **43**(Database issue):D1064-1070.

84. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD: **PANTHER version 10: expanded protein families and functions, and analysis tools**. *Nucleic acids research* 2016, **44**(D1):D336-342.

85. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH: **PIRSF family classification system for protein functional and evolutionary analysis**. *Evol Bioinform Online* 2006, **2**:197-209.

86. Letunic I, Doerks T, Bork P: **SMART: recent updates, new developments and status in 2015**. *Nucleic acids research* 2015, **43**(Database issue):D257-260.

87. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny**. *Nucleic acids research* 2009, **37**(Database issue):D380-386.

88. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E: **TIGRFAMs and Genome Properties in 2013**. *Nucleic acids research* 2013, **41**(Database issue):D387-395.

89. Alberts IL, Nadassy K, Wodak SJ: **Analysis of zinc binding sites in protein crystal structures**. *Protein science : a publication of the Protein Society* 1998, **7**(8):1700-1716.

90. Harding MM: **The architecture of metal coordination groups in proteins**. *Acta crystallographica Section D, Biological crystallography* 2004, **60**(Pt 5):849-859.

91. Harding MM: **Geometry of metal-ligand interactions in proteins**. *Acta crystallographica Section D, Biological crystallography* 2001, **57**(Pt 3):401-411.

92. Groom CR, Bruno IJ, Lightfoot MP, Ward SC: **The Cambridge Structural Database**. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 2016, **72**(Pt 2):171-179.

93. Zastrow ML, Pecoraro VL: **Influence of active site location on catalytic activity in de novo-designed zinc metalloenzymes**. *Journal of the American Chemical Society* 2013, **135**(15):5895-5903.

94. Sgrignani J, Magistrato A, Dal Peraro M, Vila AJ, Carloni P, Pierattelli R: **On the active site of mononuclear B1 metallo beta-lactamases: a computational study**. *J Comput Aided Mol Des* 2012, **26**(4):425-435.

95. Lopez M, Kohler S, Winum JY: **Zinc metalloenzymes as new targets against the bacterial pathogen Brucella**. *Journal of inorganic biochemistry* 2012, **111**:138-145.

96. Azia A, Levy R, Unger R, Edelman M, Sobolev V: **Genome-wide computational determination of the human metalloproteome**. *Proteins* 2015, **83**(5):931-939.

97. Kawai K, Nagata N: **Metal-ligand interactions: an analysis of zinc binding groups using the Protein Data Bank**. *Eur J Med Chem* 2012, **51**:271-276.

98. Shi W, Chance MR: **Metalloproteomics: forward and reverse approaches in metalloprotein structural and functional characterization**. *Curr Opin Chem Biol* 2011, **15**(1):144-148.

99. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society Series B (Methodological)* 1977, **39**(1):1-38.

100. Breiman L: **Manual–setting up, using, and understanding random forests V4. 0. 2003** http://oz.berkeley.edu/users/breiman. 2003.

101. Breiman L: **Random forests**. *Machine learning* 2001, **45**(1):5-32.

102. Hartigan JA, Wong MA: **Algorithm AS 136: A k-means clustering algorithm**. *Applied statistics* 1979:100-108.

103. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G: **InterProScan 5: genome-scale protein function classification**. *Bioinformatics* 2014, **30**(9):1236-1240.

104. Harding MM, Nowicki MW, Walkinshaw MD: **Metals in protein structures: a review of their principal features**. *Anglais* 2010, **16**(4):247-302.

105. McCall KA, Huang C, Fierke CA: **Function and mechanism of zinc metalloenzymes**. *The Journal of nutrition* 2000, **130**(5S Suppl):1437S-1446S.

106. Onoa B, Moreno V: **Nickel (II) and copper (II)–l-cysteine, l-methionine, l-tryptophan-nucleotide ternary complexes**. *Transition Metal Chemistry* 1998, **23**(4):485-490.

107. Roe RR, Pang Y-P: **Zinc′ s exclusive tetrahedral coordination governed by its electronic structure**. *Journal of molecular modeling* 1999, **5**(7-8):134-140.

108. Sousa SF, Fernandes PA, Ramos MJ: **The carboxylate shift in zinc enzymes: a computational study**. *Journal of the American Chemical Society* 2007, **129**(5):1378-1385.

109. Huang YJ, Montelione GT: **Structural biology: proteins flex to function**. *Nature* 2005, **438**(7064):36-37.

110. Hofer TS, Randolf BR, Rode BM: **Molecular dynamics simulation methods including quantum effects**. In: *Solvation effects on molecules and biomolecules.* Edited by Canuto S: Springer Netherlands; 2008: 247-278.

111. Dokmanic I, Sikic M, Tomic S: **Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in**

the coordination. *Acta crystallographica Section D, Biological crystallography* 2008, **64**(Pt 3):257-263.

112. Harding MM: **Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium**. *Acta crystallographica Section D, Biological crystallography* 2002, **58**(Pt 5):872-874.

113. Semagn K, Magorokosho C, Vivek BS, Makumbi D, Beyene Y, Mugo S, Prasanna BM, Warburton ML: **Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers**. *BMC genomics* 2012, **13**:113.

114. Lu J, Sun PD: **A rapid and rational approach to generating isomorphous heavy-atom phasing derivatives**. *The FEBS journal* 2014, **281**(18):4021-4028.

115. Yamashita K, Pan D, Okuda T, Sugahara M, Kodan A, Yamaguchi T, Murai T, Gomi K, Kajiyama N, Mizohata E *et al*: **An isomorphous replacement method for efficient de novo phasing for serial femtosecond crystallography**. *Sci Rep* 2015, **5**:14017.

116. Parks JM, Smith JC: **Modeling Mercury in Proteins**. In: *Methods in enzymology.* Academic Press.

117. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic acids research* 2001, **29**(1):308-311.

118. Garte S: **Human population genetic diversity as a function of SNP type from HapMap data**. *American journal of human biology : the official journal of the Human Biology Council* 2010, **22**(3):297-300.

119. Agola LE, Steinauer ML, Mburu DN, Mungai BN, Mwangi IN, Magoma GN, Loker ES, Mkoji GM: **Genetic diversity and population structure of Schistosoma mansoni within human infrapopulations in Mwea, central Kenya assessed by microsatellite markers**. *Acta tropica* 2009, **111**(3):219-225.

120. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J *et al*: **A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1**. *Nature genetics* 2007, **39**(2):207-211.

121. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W *et al*: **A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21**. *Nature genetics* 2007, **39**(8):984-988.

122. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU *et al*: **A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants**. *Science* 2007, **316**(5829):1341-1345.

123. Sina M, Pedram M, Ghojazadeh M, Kochaki A, Aghbali A: **P53 gene codon 72 polymorphism in patients with oral squamous cell carcinoma in the population of northern Iran**. *Med Oral Patol Oral Cir Bucal* 2014, **19**(6):e550-555.

124. Li Y, Chang SC, Niu R, Liu L, Crabtree-Ide CR, Zhao B, Shi J, Han X, Li J, Su J *et al*: **TP53 genetic polymorphisms, interactions with lifestyle factors and**

lung cancer risk: a case control study in a Chinese population. *BMC Cancer* 2013, **13**:607.

125. Abdel Hamid TM, El Gammal MM, Eibead GT, Saber MM, Abol Elazm OM: **Clinical impact of SNP of P53 genes pathway on the adult AML patients**. *Hematology* 2015, **20**(6):328-335.

126. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M: **Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database**. *Human mutation* 2007, **28**(6):622-629.

127. Taylor WR: **Identification of protein sequence homology by consensus template alignment**. *Journal of molecular biology* 1986, **188**(2):233-258.

128. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 1987, **84**(13):4355-4358.

129. Barton GJ: **Protein multiple sequence alignment and flexible pattern matching**. *Methods in enzymology* 1990, **183**:403-428.

130. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic acids research* 2011, **39**(Web Server Issue):W29-W37.

131. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

APPENDIX A

LIST OF ABBREVIATIONS

MP	Metalloprotein
CG	Coordination geometry
Tet	Tetrahedral
Bva	Trigonal bipyramidal vacancy axial
Bvp	Trigonal bipyramidal vacancy planar
Pyv	Square pyramidal vacancy
Spl	Square planar
Tbp	Trigonal bipyramidal
Spy	Square pyramidal
Tpv	Trigonal prismatic vacancy
Oct	Octahedral
Pva	Pentagonal bipyramidal vacancy axial
Pvp	Pentagonal bipyramidal vacancy planar
Tpr	Trigonal prismatic
Pbp	Pentagonal bipyramidal
Hva	Hexagonal bipyramidal vacancy axial
Hvp	Hexagonal bipyramidal vacancy planar
Sav	Square antiprismatic vacancy
Hbp	Hexagonal bipyramidal
Sqa	Square antiprismatic
FC	First-Coordination

PDB Protein Data Bank

IPR InterProScan

CFT Crystal Field Theory

LFT Ligand Field Theory

NP Nucleotide Polymorphism

SNP Single Nucleotide Polymorphism

HMM Hidden Markov Model

https://s100.copyright.com/AppDispatchServlet?publisherName=Wiley&publication=PROT&title=A%20less-b

Copyright Clearance Center RightsLink®

Home | Create Account | Help | Live Chat

Title: A less-biased analysis of metalloproteins reveals novel zinc coordination geometries

Author: Sen Yao,Robert M. Flight,Eric C. Rouchka,Hunter N. B. Moseley

Publication: Proteins: Structure, Function and Bioinformatics

Publisher: John Wiley and Sons

Date: Jun 13, 2015

© 2015 The Authors. Proteins: Structure, Function, and Bioinformatics Published by Wiley Periodicals, Inc.

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials. Already **a RightsLink user** or want to learn more?

Welcome to RightsLink

This article is available under the terms of the Creative Commons Attribution License (CC BY) (which may be updated from time to time) and permits use, distribution and reproduction in any medium, provided that the Contribution is properly cited.

For an understanding of what is meant by the terms of the Creative Commons License, please refer to Wiley's Open Access Terms and Conditions.

Permission is not required for this type of reuse.

Wiley offers a professional reprint service for high quality reproduction of articles from over 1400 scientific and medical journals. Wiley's reprint service offers:

- Peer reviewed research or reviews
- Tailored collections of articles
- A professional high quality finish
- Glossy journal style color covers
- Company or brand customisation
- Language translations
- Prompt turnaround times and delivery directly to your office, warehouse or congress.

Please contact our Reprints department for a quotation. Email corporatesaleseurope@wiley.com or corporatesalesusa@wiley.com or corporatesalesDE@wiley.com .

CLOSE WINDOW

Copyright © 2016 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

CURRICULUM VITAE

Sen Yao

Interdisciplinary Studies, Bioinformatics

University of Louisville

Louisville, KY, 40292

Email: sen.yao@louisville.edu

Education

- B.S., Pharmaceutical Science with a focus on Chinese Medicine, 2004-2008
  Guangzhou University of Chinese Medicine, Guangzhou, Guangdong, China.
- M.S., Chemistry with a focus on Biochemistry, 2009-2012
  University of Louisville, Louisville, KY, US
  Advisor: Dr. Hunter Moseley.
- Ph.D., Interdisciplinary studies with a focus on Bioinformatics (Expected August 2016)
  University of Louisville, Louisville, KY, US
  Advisor: Dr. Eric Rouchka and Dr. Hunter Moseley.

Publications

- Yao S, Flight RM, Rouchka EC, Moseley HN. "Aberrant coordination geometries discovered in the most abundant metalloproteins." Proteins: Structure, Function, and Bioinformatics, (Submitted).

- Yao S, Flight RM, Rouchka EC, Moseley HN. “Who should decide between novel versus aberrant metal coordination geometry models?” Proteins: Structure, Function, and Bioinformatics, (Submitted).
- Yao S, Flight RM, Rouchka EC, Moseley HN. “A less biased analysis of metalloproteins reveals novel zinc coordination geometries.” Proteins: Structure, Function, and Bioinformatics, 83(8): 1470-87.
- Koo I, Yao S, Zhang X, Kim S. (2014) “Comparative analysis of false discovery rate methods in constructing metabolic association networks.” Journal of Bioinformatics and Computational Biology, 12(4):1450018. (PMID: 25152043; PMCID: PMC4144070)

Positions and Honors

- 2009-2012, Research/Teaching Assistant, University of Louisville, Louisville, KY
- 2012-2016, University Fellowship, University of Louisville, Louisville, KY

Poster presentations

- Yao S, Flight RM, Rouchka EC, Moseley HN. (2016) “Aberrant coordination geometries models discovered in top abundant metalloproteins”, UT-KBRIN Bioinformatics Summit, Cadiz, KY
- Yao S, Flight RM, Rouchka EC, Moseley HN. (2015) “A less biased analysis of metalloproteins’ coordination geometries”, ACM-BCB '15, September 09-12, 2015, Atlanta, GA, USA
- Yao S, Flight RM, Rouchka EC, Moseley HN. (2015) “A less biased analysis of metalloproteins reveals novel zinc coordination geometries” UT-KBRIN Bioinformatics Summit, Buchanan, TN
- Yao S, Flight RM, Moseley HN. (2014) “Coordination characterization of zinc metalloproteins.” UT-KBRIN Bioinformatics Summit, Cadiz, KY

- Yao S, Cook TD, Moseley HN. (2013) “Coordination characterization and function annotation of zinc metalloproteins” UT-KBRIN Bioinformatics Summit, Buchanan, TN
- Yao S, Cook TD, Moseley HN. (2012) “Coordination characterization and function prediction trend of zinc metalloproteins”, UT-KBRIN Bioinformatics Summit, Louisville. KY

Professional Training Experience

- Essential Skills for Next Generation Sequencing and Data Analysis Workshop, July 20-24, 2015, University of Kentucky