Study of Mining Protein Structural Properties and its Application

A Dissertation Proposal
Presented to the
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

In Partial Fulfillment of
the Requirements for the Degree
Doctor of Philosophy

by
Yu-Feng Huang
Dr. Chien-Kang Huang, Dissertation Supervisor
Dr. Yen-Jen Oyang, Dissertation Supervisor
December 11, 2007
# Table of Contents

Table of Contents ............................................................................................................i  
List of Tables ......................................................................................................................iv  
List of Figures .......................................................................................................................v  
Abbreviations .......................................................................................................................vi  
1. Introduction .......................................................................................................................... 1  
   1.1. Current Status of Structural Genomics ................................................................. 1  
   1.2. Sequences, Structures, and Functions ................................................................. 4  
       1.2.1. Protein Structure .................................................................................. 4  
       1.2.2. Sequence, Structure, and Function ...................................................... 6  
   1.3. Tackled Issues in this Dissertation ......................................................................... 6  
       1.3.1. Study of Local Structure Representation ........................................... 6  
       1.3.2. Study of Conserved Structure for Functional Classification ............... 7  
       1.3.3. Mining General Protein Structural Properties ................................... 7  
       1.3.4. Involving the New Approaches of Fast Structure Mining ................. 8  
       1.3.5. Coordination of Sequence and Structural Conservation .................... 8  
       1.3.6. Apply Mining Results in Function/Structure/Sequence Prediction and Annotation .......................................................... 9  
   1.4. Overview ..................................................................................................................... 9  
2. Literature Reviews ........................................................................................................... 11  
   2.1. Sequence, Structure, and Function ........................................................................ 11  
   2.2. Sequence Motif and Structural Motif ................................................................... 11  
   2.3. Structural Property ............................................................................................... 12  
   2.4. Structural Database ............................................................................................... 13  
       2.4.1. Worldwide Protein Data Bank ............................................................... 13  
       2.4.2. Enzyme Data Bank .............................................................................. 14  
       2.4.3. Nucleic Acid Database ......................................................................... 14  
   2.5. Structural Classification ......................................................................................... 14  
       2.5.1. SCOP .................................................................................................. 14  
       2.5.2. CATH ............................................................................................... 15  
   2.6. Functional Classification ......................................................................................... 16  
       2.6.1. Enzyme Classification ........................................................................... 16  
3. Thesis Statement ............................................................................................................. 19  
   3.1. Motivation .................................................................................................................. 19  
   3.2. Framework of this Dissertation ............................................................................ 19  
       3.2.1. Study of Local Structure Representation ......................................... 19
3.2.2. Study of Conserved Structure for Functional Classification ..........20
3.2.3. Mining General Protein Structural Properties..............................21
3.2.4. Involving the New Approaches of Fast Structure Mining ..............21
3.2.5. Coordination of Sequence Conservation and Structural Conservation ........................................................................................................22
3.2.6. Apply Mining Results in Function/Structure/Sequence Prediction and Annotation ........................................................................................................22

4. Research Description .............................................................................................................23
4.1. Protein Local Structure Representation .................................................................23
        4.1.1. Introduction........................................................................................................23
                4.1.1.1. Motivation ................................................................................................24
        4.1.2. Local Conservation and Functional Site ........................................................24
        4.1.3. Local Structure Representation .................................................................25
                4.1.3.1. Alignment Result of Protein Structure Comparison ... 25
                4.1.3.2. Neighborhood Residues Sphere ...........................................................25
        4.1.4. Structure Conservation Detection ...............................................................26
                4.1.4.1. Pair-wise Protein Structure Comparison Approach .... 27
                4.1.4.2. NRS-based Conservation Mining Approach .........................27
        4.1.5. Experiments .......................................................................................................28
        4.1.6. Discussions ......................................................................................................29
                4.1.6.1. Pair-wise Protein Structure Comparison Approach .... 29
                4.1.6.2. NRS-based Conservation Mining Approach .........................30
                4.1.6.3. Summarization ......................................................................................31
        4.1.7. Conclusions ......................................................................................................32

4.2. Protein Structure Conservation Mining .................................................................33
        4.2.1. Introduction ......................................................................................................33
        4.2.2. Local Structure Representation .................................................................34
        4.2.3. Mining Conserved Patterns ...........................................................................36
                4.2.3.1. NRS Segmentation ..............................................................................37
                4.2.3.2. Sequence Conservation Grouping ......................................................37
                4.2.3.3. Representative Selection .................................................................38
        4.2.4. Template Library ............................................................................................38
        4.2.5. Enzyme Classification Prediction .................................................................39
        4.2.6. Comparison with other Template Libraries ................................................41
        4.2.7. Discussion ......................................................................................................45
        4.2.8. Conclusion ......................................................................................................47

4.3. Protein Structural Property Exploration .........................................................49
        4.3.1. Introduction ......................................................................................................49
4.3.2. Review of Protein Structural Property Exploration .....................49
4.3.3. Proposed Indexing Mechanism for Massive Structural Property
Exploration...................................................................................................50
  4.3.3.1. Residue Environmental Sphere and Indexing Mechanism 50
  4.3.3.2. Materials ..................................................................................51
  4.3.3.3. Database Design.......................................................................52
4.3.4. Statistical Analysis of Structural Properties on Protein Data Bank 53
  4.3.4.1. Residue-Residue Contacts ..................................................53
  4.3.4.2. Chemical Component Contacts.............................................53
4.3.5. Property Analysis on Disulfide Bond .............................................54
  4.3.5.1. Disulfide Bond .......................................................................54
  4.3.5.2. SSBOND................................................................................54
  4.3.5.3. Residue-Residue Contacts of Cysteine Pairs .....................54
4.3.6. Results............................................................................................55
  4.3.6.1. Residue-Residue Contacts and Chemical Component
           Contacts 55
  4.3.6.2. Disulfide Bond ......................................................................55
4.3.7. Discussion.......................................................................................56
  4.3.7.1. Difference of SSBOND and Cysteine Pairs ......................56
  4.3.7.2. File Parsing and Efficiency of Database Query ...................57
4.3.8. Conclusions....................................................................................58
  5. Summarization......................................................................................59
  5.1. Protein Local Structure Representation .......................................59
  5.2. Protein Structure Conservation Mining .......................................59
  5.3. Protein Structural Property Exploration.....................................60
6. Ongoing Status.......................................................................................61
  6.1. Structural Data Information Analysis .........................................61
  6.2. Protein Structure Conservation Mining base on Sequence-Structure
       Correlation .......................................................................................62
  6.3. Structure-based Mining Approach for Structure Conservation Discovery 62
  6.4. Protein Structural Property Exploration of Interaction Region ......63
  6.5. Summary ........................................................................................67
References....................................................................................................69
List of Tables

Table 1. A rough guide to the resolution of protein structure ........................................ 13
Table 2. List of protein chains for 6 randomly selected EC families.............................. 29
Table 3. Experimental results for local conservation discovery via pair-wise protein
structure comparison..................................................................................................... 29
Table 4. Description of assessment.............................................................................. 41
Table 5. Experimental results for enzyme classification prediction............................. 42
Table 6. Multiple EC label prediction........................................................................... 43
Table 7. Statistical result of SSBOND and Cysteine pair.............................................. 56
List of Figures

Figure 1. Yearly growth of released structures in Protein Data Bank ................................................. 2
Figure 2. 20 standard amino acids. ........................................................................................................ 3
Figure 3. Venn diagram grouping amino acids according to their properties. ........................................... 5
Figure 4. Position specific score matrix (PSSM) generated by PSI-BLAST ........................................... 12
Figure 5. The hierarchy of CATH .............................................................................................................. 15
Figure 6. The overall framework for mining conserved local structure ...................................................... 20
Figure 7. Neighborhood residues sphere .................................................................................................. 26
Figure 8. The flow chart for mining conserved structural patterns via pair-wise protein structure comparison .............................................................................................................................................. 27
Figure 9. The flow chart for mining conserved structural patterns via NRS-based conservation mining approach .............................................................................................................................................. 28
Figure 10. Protein PDB ID 1J9Z:A and its binding substrates ................................................................. 30
Figure 11. PDB ID 1SMI:A and the substrate is HEM ............................................................................... 32
Figure 12. Neighborhood Residues Sphere ............................................................................................... 35
Figure 13. Flow chart of mining conservation patterns ................................................................................ 36
Figure 14. Enzyme classification prediction ............................................................................................... 40
Figure 15. Conserved patterns of EC 3.2.1.17 ............................................................................................ 44
Figure 16. Conserved local structure and a ligand ...................................................................................... 46
Figure 17. Conserved pattern and ligand, SDK, of protein PDBID 1AU0 ............................................... 48
Figure 18. Residue environmental sphere .................................................................................................. 51
Figure 19. Database table schema for structural property exploration ....................................................... 52
Figure 20. Distribution between distance and its frequent ........................................................................... 56
Figure 21. Disulfide bond and ligand ......................................................................................................... 57
Figure 22. Comparison of latest version and previous version of 1UMR ................................................. 61
Figure 23. Encoding scheme for transforming structure information into binary signature ........................................... 63
Figure 24. Residue-residue contacts .......................................................................................................... 64
Figure 25. Protein-ligand contact ............................................................................................................... 64
Figure 26. Protein-protein interaction region ......................................................................................... 65
Figure 27. Protein-RNA interaction region .................................................................................................. 65
Figure 28. Protein-DNA interaction region ............................................................................................... 66
Figure 29. Intermolecular disulfide bond ................................................................................................. 66
Figure 30. Intramolecular disulfide bond ................................................................................................. 67
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>one-dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>three-dimensional</td>
</tr>
<tr>
<td>ASA</td>
<td>accessible surface area</td>
</tr>
<tr>
<td>CATH</td>
<td>CATH Protein Structure Classification – Class, Architecture, Topology, Homologous Superfamily</td>
</tr>
<tr>
<td>CSA</td>
<td>Catalytic Site Atlas</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>HGP</td>
<td>Human Genome Project</td>
</tr>
<tr>
<td>NDB</td>
<td>Nucleic Acid Database</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>RMSD</td>
<td>root mean square deviation</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RSA</td>
<td>relative solvent accessibility</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PSSM</td>
<td>Position Specific Score Matrix</td>
</tr>
<tr>
<td>SCOP</td>
<td>Structural Classification of Proteins</td>
</tr>
<tr>
<td>SH2</td>
<td>src homology 2</td>
</tr>
<tr>
<td>wwpdb</td>
<td>Worldwide Protein Data Bank</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

1.1. Current Status of Structural Genomics

The “Human Genome Project” (HGP) was a 13-year project coordinated by the U.S Department of Energy and the National Institutes of Health since 1990. This project was completed in 2003, and researches from Hong Kong, Japan, France, Germany, China, and others joined the HGP during the period. Project goals were to identify all the approximately 20,000-25,000 genes in human DNA, determine the sequences of the 3 billion chemical base pairs that make up human DNA, store this information in databases, improve tools for data analysis, transfer related technologies to the private sector, and address the ethical, legal, and social issues (ELSI) that may arise from the project (adopted from http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml).

With the huge growth of protein sequences, structures, and biological data, researchers have to face a huge scale of dataset for analysis. Bioinformatics can be defined as the study of two information flows in molecular biology [1]. He pointed out two information flows: the first is based on the central dogma of molecular biology: DNA sequences are transcribed into mRNA sequences and then mRNA sequences are translated into protein sequences, and the second is based on experimental information from observations to models. In the first flow, we use informatics methodology to analysis biological data of sequences and structures. In the second flow, we have to build a model to explain our observations and then use new experiments to test a model.

Beccari prepared the first protein of vegetable origin [2] in 1747, and the Protein Data Bank began to collect examined three-dimensional structural data from 1976. In the past three decades, the number of released structures grows exponentially as shown in . As of January 1, 2008, there are 48161 determined structures examined by X-ray or nuclear magnetic resonance (NMR) in Protein Data Bank (PDB) [3]. They include proteins, protein complexes, nucleic acids and protein nucleic acid complexes. Determined protein structures have been greatly increasing from 1976, since then protein functional analysis has become more and more important [13]. Accompanying with the fast growth of Protein Data Bank, protein functional analysis has become more important. Researches focused on functional classification have
been investigated for many years. Based on previous researches, if we attempt to understand the relationship between protein structure and function, data mining technique should be involved for massive protein structure analysis.

Figure 1. Yearly growth of released structures in Protein Data Bank.

The released statistics was updated on December 11, 2007.
Structural bioinformatics is the subdiscipline of bioinformatics that focuses on the representation, storage, retrieval, analysis, and display of structural information at the atomic and subcellular spatial scales [4]. Protein structure determination and prediction, both have been investigated for many years. These issues in structural biology include secondary structure prediction [5-7], protein disorder region prediction [8-10], b-factor prediction [11], binding residue prediction [12, 13], RNA-binding residue prediction [14-17], DNA-binding residue prediction [18-24] and prediction of protein-protein interaction [25-27], protein-RNA interaction [17], or protein-DNA interaction [28]. Furthermore, researches on contact preferences also have been investigated in interaction regions of protein-protein [29], protein-RNA [30], and protein-DNA [31, 32].

![Figure 2. 20 standard amino acids. This diagram is adapted from http://matcmadison.edu/biotech/resources/proteins/labManual/images/amino_000.gif](http://matcmadison.edu/biotech/resources/proteins/labManual/images/amino_000.gif)
1.2. Sequences, Structures, and Functions

1.2.1. Protein Structure

Proteins are linear chains of amino acids and linked together by polypeptide bonds between the carboxyl and amino groups of adjacent amino acid residues in order. The sequence of the different amino acids is called a primary structure. In nature, there are 20 standard amino acids, but the residue in a protein would be chemically altered in post-translational modification. These 20 standard amino acids in Figure 2 are alanine (Ala, A), arginine (Arg, R), asparagine (Asn, N), aspartic acid (Asp, D), cysteine (Cys, C), glutamic acid (Glu, E), glutamine (Gln, Q), glycine (Gly, G), histidine (His, H), isoleucine (Ile, I), leucine (Leu, L), lysine (Lys, K), methionine (Met, M), phenylalanine (Phe, F), proline (Pro, P), serine (Ser, S), threonine (Thr, T), tryptophan (Trp, W), tyrosine (Tyr, Y), and valine (Val, V). Each amino acid has its own properties shown in .

In proteins, secondary structure can be recognized by DSSP software [33] according to the hydrogen bonds between backbone amide groups, and can be classified as α-helix and β-sheet. The secondary structure of a protein is nonlinear, localized to regions of an amino acid chain, and formed and stabilized by hydrogen bonding. The hydrogen bonding in these elements of structure provides much of the enthalpy of stabilization that allows the polar backbone groups to exist in the hydrophobic core of a folded protein [34]. In biochemistry, the tertiary structure of a protein is its three-dimensional structure with the atomic coordinates. However, in protein structure recognition, secondary structure is widely used to describe a three-dimensional form of local segments of biopolymers instead of atomic coordinates. Tertiary structure of a protein is nonlinear, formed and stabilized by hydrogen bonding, covalent bonding, hydrophobic packing toward core and hydrophilic exposure to solvent. A quaternary structure of a protein is formed by the folded chains which have more than one polypeptide chain. Protein assemblies composed of more than one polypeptide chain are called oligomers and the individual chains of which they are made are termed monomers or subunits [34]. Quaternary structure of a protein is nonlinear, global and across distinct amino acid polymers, formed by hydrogen bonding, covalent bonding, hydrophobic packing and hydrophilic exposure, and favorable, functional structures occur frequently and have been categorized.
In protein structure, residues interact with each other in three-dimensional space via covalent bonding or non-covalent bonding such as electrostatic, hydrogen bonds or Van der Waals forces. The covalent bonding is an induced dipole-dipole interaction that is characterized by the sharing of pairs of electrons between atoms, or between atoms and other covalent bonds. The covalent bonding is stronger than most non-covalent bonding. Disulfide bond is one kind of special bond connectivity in protein structure, which is linked via two $S$ atoms of cysteine residues in protein folding. Disulfide bond could be occurred inter-molecularly or intra-molecularly. Disulfide bond formation is a covalent modification; the oxidation reaction can either be intramolecular (within the same protein) or inter-molecular (within different proteins, e.g., antibody light and heavy chains). The reaction is reversible.
Van der Waals interactions contribute strong repulsion at short distances and weak attraction at distances just greater than the sum of the atomic radii. Salt bridges play important roles in protein structure and function, e.g., in oligomerization, molecular recognition, allosteric regulation, domain motions, flexibility, thermostability, and alpha-helix capping. The electrostatic contribution to the free-energy change upon salt-bridge formation varies significantly, from being stabilizing to marginal to being destabilizing [35]. A hydrogen bond occurs between an electronegative atom and a hydrogen atom bonded to another electronegative atom, which is a special type of dipole-dipole bond. The typical hydrogen bond is stronger than Van der Waals forces, but weaker than covalent, ionic and metallic bonds.

1.2.2. Sequence, Structure, and Function

With the increasing growth of sequence, structural, and biochemical data, evolution of protein function can be determined from sequence and/or structure. Homologous proteins can be determined via BLAST [36] or FASTA [37] alignment approach to identify the relation between proteins. Sequence alignment algorithm can tell us sequence similarity between protein sequences, and evolutionary information can be detected via alignment of aligned sequence fragments. With the help of multiple sequence alignment, sequence conservation also can be discovered to link with protein function. From a structural standpoint, protein function and protein structure are inherently linked [38], and structural template comparison can recognize protein function by comparing template against protein structures [39]. Neither sequence similarity nor structure similarity can directly infer protein function alone. They all tell us partial information about protein function or something about evolution [40].

1.3. Tackled Issues in this Dissertation

1.3.1. Study of Local Structure Representation

According to research recommendation from Najmanovich et al., predicting the function of a protein from its three-dimensional structure is a major intellectual and practical challenge [41]. They reveal that detecting local structure similarity can be applied to predict a function of a protein. The point mentioned by Orengo et al. is that sequence-based methods can fail to detect very distant relationships and these can
only be recognized from 3D structure, which is much more highly conserved during evolution [42]. Moreover, researchers make more effort on the study of protein functional site or ligand binding areas [39, 43, 44]. All these research findings give us an important hint on the study of relation between protein function and local structure. Hence, can we develop an appropriate representation to describe the connection between the dedicated local structure and corresponding function in proteins?

1.3.2. Study of Conserved Structure for Functional Classification

Based on the common assumption that proteins of the same function share common local regions, the concept of local region conservation comes from a motif, which is a fragment with biological or functional meaning. In sequence analysis, Campbell et al. [45] applied sequence alignment to discover sequence conservation, and then they map conserved regions into their three-dimensional space which are close to binding area. In structure analysis, the binding area of protein-ligand complex is widely used to identify protein function via local structure recognition. CSA (Catalytic Site Atlas) [39] and Protemot [44] use protein-ligand complexes to recognize protein function via local structure similarity. Based on research results of CSA and Protemot, the authors point out that non-homologous proteins may have the same function; in the other words, proteins have dissimilar global structures may have the same function, and the observations can be found that function may occur in protein local structure. Currently, we approach two directions to achieve, and one is protein structure comparison, and another is to use neighborhood residues sphere (NRS), a sphere with the radius of $d$ ($d=10$ as default), to describe local structure. In our experimental results, both approaches can discover conserved local structures for most enzyme family, and some of conserved local structures are close to ligands.

1.3.3. Mining General Protein Structural Properties

With the fast growth of protein structure, it provides more materials on the study of discovering local residue environment with/without chemical bond information. Residue environment has been studied and applied on protein threading and protein binding site characterization [46]. In the protein structure, a residue is the essential element for conformation, and residue-residue contacts will affect the overall framework of a protein structure. Protein folding is highly correlated to residue contacts with chemical bonds such as covalent bonds, ionic bonds, hydrogen bonds,
Van der Waals attractions, or disulfide bonds. For quick searching of residue environment, we use residue environmental sphere to describe environment information surrounding a residue. On the purpose of protein structural property exploration, we have to analyze different residue neighborhood in whole protein structure collection. Applying mining technique on protein structures is an interesting issue to discover residue environmental information inside protein structure, and to handle huge protein structure collection is also a great challenge to store entire structure and sphere information in database.

1.3.4. Involving the New Approaches of Fast Structure Mining

Because massive pair-wise sequence and structure comparison are time-consuming task, we still have to improve performance for fast structure mining. According to the definition of protein blocks [47] proposed by Breven et al., the authors try to use protein blocks to understand the sequence-structure relationship and structural alphabet [48] is an improved representation of protein blocks. Therefore, they encode a protein structure into a one-dimensional sequence and they can treat one-dimensional sequence as protein sequence and BLAST can be easily applied. They also proposed substitution matrix for structural alphabet based on statistics analysis of alphabet mutations. In contrast to structural alphabet, we propose to encode protein structure via signature and indexing technique for fast structure mining. The same as conserved structure mining, we use neighborhood residues sphere to describe protein local structure, transform each sphere as bit-string signature, and the indexing technique will be applied to provide fast database search. Furthermore, we encode each neighborhood residues sphere as environmental signature for protein structure indexing and quick database searching.

1.3.5. Coordination of Sequence and Structural Conservation

According to research results of MAGIIC-PRO [49] developed by Hsu et al., which is driven by homologues protein sequence analysis on detecting a functional signature, the authors approach sequence pattern mining to discover functional signatures of a query protein. Their experimental results reveal that sequence conservation has correlation to protein function according to ligand information. Based on our previous study on local conserved structures, we attempt to integrate sequence conservation and structure conservation for analyzing the relationship among sequences, structures, and functions in the future. Our original idea is to discuss the
relationship between sequence conservation and structure conservation for each enzyme family. In each enzyme family, proteins within an enzyme family have the same function derived from different species; therefore, it is a good start to discover sequence and structure conservation based on the relationship between sequences, structures and functions.

1.3.6. Apply Mining Results in Function/Structure/Sequence Prediction and Annotation

According to the experimental results of first three sub-topics, we plan to combine mining results and machine learning technique to improve prediction accuracy and annotation. Recent research has been applied structure properties in primary sequence prediction to improve prediction accuracy. Computer-aid annotation for protein sequences, structures, and functions has been studied based on protein global sequence and structure information. Our idea start from protein local sequence and structure to correlate with its function; therefore, we attempt to include protein structure properties of local region to study the correlation of sequence, structure, and function from the view of local region. In addition, we will also include structure information as feature information in primary sequence prediction of machine learning.

1.4. Overview

The sections of the paper are organized roughly according to the issues tackled in this dissertation. In the next section, we review previous researches related to structure mining and protein function. Section 3 considers the framework for mining conserved local structure and the study of local structure and protein function. Section 4 gives detail information about each part of overall framework. Section 5 discusses and summarizes experimental results for this dissertation. Finally Section 6 introduces our ongoing status and further study.
2. LITERATURE REVIEWS

2.1. Sequence, Structure, and Function

Sequence similarity is determined by aligning sequences according to percent identity. Homologous sequences derived from the same ancestral sequence can be examined under some identical residues at the corresponding positions in the sequence. In general, similar protein sequences can be implied that they have similar structures and similar functions. Therefore, protein function can be inferred by determining sequence similarity and structure similarity, but there are still some exceptions. For example of TIM-barrel proteins, they have eight $\beta/\alpha$ motifs folded into a barrel structure, and many functions [50]. Proteins that differ in sequence and structure may have converged to similar active site, catalytic mechanisms and biochemical function. Proteins with low sequence similarity but very similar overall structure and active sites are likely to be homologous [34].

2.2. Sequence Motif and Structural Motif

The term motif is used to represent a characteristic fragment which is biological significant to protein function. It can be represented as sequence motif, structural motif, and functional motif. A sequence motif refers to a particular amino acid sequence that is characteristic of a specific biochemical function. Zinc finger motif is an example of sequence motif which is found in a family of DNA-binding proteins, and the motif is formed as Cys-X$_{2,4}$-Cys-X$_3$-Phe-X$_5$-Leu-X$_2$-His-X$_3$-His $(C_2H_2)$ [51, 52]. Sequence motif can be evolution conservation which could be discovered by sequence alignment based evolutionary similarity. Researches related to discover sequence conservation has been found that discovered sequence motifs correlate to biological functions [53]. The structural motif refers to motif in three-dimensional space. Commonly, structural motif is a set of contiguous secondary structure elements that either have a particular functional significance or define a portion of an independently folded domain [34]. The helix-turn-helix is an example of structural motif found in DNA-binding proteins.
2.3. Structural Property

In sequence based prediction, the position-specific scoring matrix (PSSM) is used to improve their prediction accuracy for protein sequence analysis as shown in Figure 4. The PSSM gives the log-odds score for finding a particular matching amino acid against to a target sequence. Therefore, the prediction tools treat PSSM as sequence property for each amino acid. In protein structure prediction, amino acid property, secondary structure information, b-factor, accessible surface area (ASA), or relative solvent accessibility (RSA) are structural properties. Therefore, protein structure prediction from purely sequence information has been tried to encode biochemical properties relative to protein structure to improve prediction accuracy. In 1992, Singh and Thornton [54] discovered the atlas of protein side-chain interaction to understand sidechain-sidechain interactions. In this research, they revealed interactions for 20 * 20 amino acids, and counted the frequency for each amino acid pairs.

![Figure 4. Position specific score matrix (PSSM) generated by PSI-BLAST.](image)

In addition, Glaser et. al. [55] also studied structural property of residues at protein-protein interfaces. In order to realize the inside of protein structure conformation, protein structural property exploration is very important such as amino
acid interactions or residue-residue contact. Contact preference is another important
issue for structure environment analysis to discuss how residues interact with each
other [29-31]. Each residue has different tendencies to contact with other residues in
the structure environment. Furthermore, residue-residue contact in protein-protein
interaction region is another way to know residue environment while protein interacts
with another one. In addition, contact preference of residue and nucleic base pair is
another issue for structure environment analysis in interaction region.

Table 1. A rough guide to the resolution of protein structure

<table>
<thead>
<tr>
<th>Resolution (Å)</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 4.0</td>
<td>Individual coordinates meaningless. Secondary structure elements can be</td>
</tr>
<tr>
<td></td>
<td>determined.</td>
</tr>
<tr>
<td>3.0 - 4.0</td>
<td>Fold possibly correct, but errors are very likely. Many sidechains placed with</td>
</tr>
<tr>
<td></td>
<td>wrong rotamer.</td>
</tr>
<tr>
<td>2.5 - 3.0</td>
<td>Fold likely correct except that some surface loops might be mismodelled. Several</td>
</tr>
<tr>
<td></td>
<td>long, thin sidechains (lys, glu, gln, etc) and small sidechains (ser, val, thr, etc)</td>
</tr>
<tr>
<td></td>
<td>likely to have wrong rotamers.</td>
</tr>
<tr>
<td>2.0 - 2.5</td>
<td>As 2.5 - 3.0, but number of sidechains in wrong rotamer is considerably less. Many small errors can normally be detected. Fold normally correct and number of errors in surface loops is small. Water molecules and small ligands become visible.</td>
</tr>
<tr>
<td>1.5 - 2.0</td>
<td>Few residues have wrong rotamer. Many small errors can normally be detected. Folds are extremely rarely incorrect, even in surface loops.</td>
</tr>
<tr>
<td>&lt; 1.5</td>
<td>In general, structures have almost no errors at this resolution. Individual atoms in a structure can be resolved</td>
</tr>
</tbody>
</table>

Table is taken from Daniel (2007) and Blow (2002).

2.4. Structural Database

2.4.1. Worldwide Protein Data Bank

The Worldwide Protein Data Bank (wwPDB) [56] consists of organizations that act as
deposition, data processing and distribution centers for PDB data. The founding
members are RCSB PDB (USA) [3], MSD-EBI (Europe) and PDBj (Japan). Since
1747 Beccari discovered first protein of vegetable origin [2], and Protein Data Bank
(PDB) began to collect three-dimensional structure data in 1976. Now the PDB
contains 47625 protein structures on December 4, 2007. It is a worldwide repository
for three-dimensional structure data of proteins, protein complexes, nucleic acids, and
protein nucleic acid complexes. Typically, these data examined by X-ray crystallography, NMR spectroscopy, or electron microscopy. Most of structures are determined by X-ray crystallography, and then NMR spectroscopy. In, it is a rough guide to the resolution of protein structure that can help us how to utilize the structural data information. Materials in this table is taken from Blow [57] and Minor [58].

2.4.2. Enzyme Data Bank

The enzyme data bank [59] is a collection of information focused on all known enzymatic reactions defined by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). The EC (enzyme commission) number is given by International Union of Biochemistry and Molecular Biology. The EC number is designated by four numerals such as 1.6.2.4 similar to Internet Protocol address, and it represents the hierarchical classification of enzymes according to the type of chemical reactions catalyzed by enzymes. In enzyme data bank, entry corresponding to EC number consists of recommended name, alternative names, catalytic activity, cofactors, and protein sequences linked to SWISS-PROT [60]. The six classes in the top hierarchy are oxidoreductases (EC 1.-.-.-), transferases (EC 2.-.-.-), hydrolases (EC 3.-.-.-), lyases (EC 4.-.-.-), isomerases (EC 5.-.-.-), and ligases (EC 6.-.-.-).

2.4.3. Nucleic Acid Database

The Nucleic Acid Database [61] established in 1992 is a single archive to store three-dimensional crystal structures of nucleic acids including DNA (Deoxyribonucleic acid) and RNA (Ribonucleic acid). As of June 2007, the Nucleic Acid Database has collected 3557 nucleic acid structures are derived from both the Protein Data Bank and the literature.

2.5. Structural Classification

2.5.1. SCOP

The Structural Classification of Proteins (SCOP) database provides a detailed and comprehensive description of the relationships of all known proteins structures. It is a largely manual classification of proteins according to their structural domains based
on similarities of their amino acid sequence and three-dimensional structure. The class representation is on hierarchical levels: the first two levels, family and superfamily, describe near and far evolutionary relationships; the third level, fold, describes geometrical relationships. The leaf level is protein domain, the basic unit in the hierarchy. Under the domain, there are proteins PDB entries that reference to their own PDB description. Detail descriptions for SCOP hierarchy are:

1. Class - general structural architecture of the domain
2. Fold - similar arrangement of regular secondary structures but without evidence of evolutionary relatedness
3. Superfamily - sufficient structural and functional similarity to infer a divergent evolutionary relationship but not necessarily detectable sequence homology
4. Family - some sequence similarity can be detected.

![CATH Hierarchy Diagram](image)

Figure 5. The hierarchy of CATH.

### 2.5.2. CATH

The CATH Protein Structure Classification is a semi-automatic, hierarchical
classification of protein domains published in 1997 by Christine Orengo, Janet Thornton and their colleagues. CATH shares many broad features with its principal rival, SCOP, however there are also many areas in which the detailed classification differs greatly. The name CATH is an acronym of the four main levels in the classification. The four main levels of the CATH hierarchy are as follows:

1. Class - the overall secondary-structure content of the domain (automatic)
2. Architecture - a large-scale grouping of topologies which share particular structural features (orientation of secondary structures, manual)
3. Topology - high structural similarity but no evidence of homology. Equivalent to a fold in SCOP (topological connection and number of secondary structures)
4. Homologous superfamily - indicative of a demonstrable evolutionary relationship. Equivalent to the superfamily level of SCOP. (superfamily clusters of similar structures and functions)
5. Sequence family

CATH defines four classes according to the ratio of secondary structure elements: mostly-alpha, mostly-beta, alpha and beta, few secondary structures. The domains are automatically sorted into classes and clustered on the basis of sequence similarities. These groups form the H levels of the classification. The topology level is formed by structural comparisons of the homologous groups. Finally, the Architecture level is assigned manually. As shown in Figure 5, it is a CATH hierarchy of class, architecture, and topology levels.

2.6. Functional Classification

2.6.1. Enzyme Classification

Clearly, functional hierarchical classification classifies proteins into class according to protein function and reaction. Functional classifications derive groups on the basis of functional similarity in terms of enzyme reaction mechanism, participation in biochemical pathways, functional roles and cellular localization [62]. There are three reasons choosing functional hierarchical classification, (1) in order to provide a function, proteins should have stable structure in their functional area; (2) correlation between functional related structure region and protein function is easy to be verified via contact area of protein-substrate complex; (3) if proteins have the same function, they should have conservation in their functional areas.
The Enzyme Commission (EC) number is developed by the International Union of Biochemistry and Molecular Biology (IUBMB), which is used to classify enzyme based on the chemical reaction they catalyze. In enzyme, proteins with the same EC number have the same protein function or biochemical reaction; therefore, they may have similar functional area to react with other molecular to provide function. In enzyme hierarchical classification, they use four levels to classify enzyme into hierarchy. The top level, reaction type of the enzymes, is divided into six major classes including oxidoreductases (1.-.-.-), transferases (2.-.-.-), hydrolases (3.-.-.-), lyases (4.-.-.-), isomerases (5.-.-.-), and ligases (6.-.-.-), defined according to the reaction catalyzed. The second level is divided based on group specific action, the third level by substrate specificity and the forth level contains enzymes. Currently, Thornton et. al. extend from the Enzyme Data Bank [59] and the Protein Data Bank to build enzyme structures database (http://www.ebi.ac.uk/thornton-srv/databases/enzymes/).

Six major classes in enzyme.
Class 1. oxidoreductases   (1.-.-.-)
Class 2. transferases   (2.-.-.-)
Class 3. hydrolases   (3.-.-.-)
Class 4. lyases   (4.-.-.-)
Class 5. isomerases   (5.-.-.-)
Class 6. ligases   (6.-.-.-)

Besides, enzyme classification provides a good environment to realize protein structure and protein function. Proteins with the same EC number have same function or activate the same reaction would be grouped together. Enzyme active sites commonly occur in large and deep cavity on the protein surface, and they need significant favorable interactions between ligand and protein, which usually means that other small molecule ligand are embedded in surface depressions. If proteins provide the same function, they should have certain level of conservations on their structure conformation, and those conservations might be conserved by its conformation or function. Therefore, structure conservations might be reserved for structure conformations or protein functions. As the enzyme classification is one kind of functional classifications, and we try to find the relation of structure conservation and protein function.
3. THESIS STATEMENT

3.1. Motivation

In this dissertation, we focus on the study of discovering the relation between structure and function from a viewpoint of local structure. Based on the assumption that protein structure is more conserved for protein function, we try to discover conserved structural information from known protein functions. Therefore, the question would be to mine local structures shared among a group of proteins correlated to their function. But, another issue is that sequence and structure similarity will affect the quality of mined local structure. The reason is that if a group of proteins share highly both sequence and structure similarity, the mining result would be meaningless.

Currently, we focus on the following subtopics, and there are (i) study of local structure representation; (ii) study of conserved structure for functional classification; (iii) mining general protein structural properties; (iv) coordination of sequence conservation and structural conservation; (v) involving the new approaches of fast structure mining; and (vi) applying mining results in function/structure/sequence prediction and annotation.

3.2. Framework of this Dissertation

3.2.1. Study of Local Structure Representation

There are different types of representation could be applied to describe local structure such as protein blocks [47], structural alphabet [48, 63], structural motif [64, 65], or sequence motif with corresponding three-dimensional structure [65]. The original idea of protein blocks comes from N-gram in information retrieval. They use five consecutive Cα (“protein blocks”) as a block to describe protein local structure; therefore, a protein structure can be composited as several protein blocks [47]. Moreover, they use an unsupervised cluster analyzer to identify a local structural alphabet composed of 16 folding patterns from protein blocks. Yang et al. [66, 67] also apply structural alphabet to describe local structure, and they obtain 23 structural alphabets to represent 23 local structures. Jonassen et al. [65] use neighborhood sequence to discover sequence patterns and then check patterns in their corresponding
space. If the sequence pattern has k structure occurrences, this sequence pattern will be a local packing motif. In this dissertation, we adopt the concept of local packing motif proposed by Jonassen et al. as a local structure representation, a sphere with a distance of d Å from a central residue.

### 3.2.2. Study of Conserved Structure for Functional Classification

Based on the common assumption that proteins of the same function share common local regions, the concept of local region conservation comes from a motif, which is a fragment with biological or functional meaning. In addition, we also try to discover functional site without the help of protein-ligand complexes such as CSA (Catalytic Site Atlas) [39] and Protemot [44]. Therefore, our idea is to apply mining frequent itemset on a group of proteins, and these proteins should share the same function or reactions. Hence, if a protein structure can be decomposed as a set of local structures; frequent itemset mining can be easily applied to discover frequent local structures. The most important issue we should address is how the link could be made between protein function and discovered local structures. Because discovered local structure shares among a group of proteins, it can be viewed as conserved structure for a group. As shown in Figure 6, this is the overall framework for mining conserved local structure.

![Figure 6. The overall framework for mining conserved local structure.](image)
3.2.3. Mining General Protein Structural Properties

As we know, protein folds by a series of interaction between amino acids. In the sphere model of local structure representation, residue environment information surrounding a residue can be easily detected. The interactions between amino acids consist of atom interactions and bond connectivity. Therefore, a sphere model is an appropriate representation to describe residue environment. Accompanying with the fast growth of protein structures, it provides more materials on the study of discovering local residue environment with/without chemical bond information. Residue environment has been studied and applied on protein threading and protein binding site characterization [5]. In the protein structure, a residue is the essential element for conformation, and residue-residue contacts will affect the overall framework of a protein structure. Protein conformation is highly correlated to residue contact with chemical bonds such as covalent bonds, ionic bonds, hydrogen bonds, Van der Waals attractions, or disulfide bonds. Protein structural properties could be discovered in a protein structure or the interaction regions of protein complexes.

3.2.4. Involving the New Approaches of Fast Structure Mining

Because massive pair-wise sequence and structure comparison are time-consuming task, we still have to improve performance for fast structure mining. According to the definition of protein blocks [6] proposed by Brevener et al., the authors try to use protein blocks to understand the sequence-structure relationship and structural alphabet [7] is an improved representation of protein blocks. Therefore, they encode a protein structure into a one-dimensional sequence and they can treat one-dimensional sequence as protein sequence and BLAST can be easily applied. In addition, substitution matrix for structural alphabet is also an issue should be addressed. Currently, our proposed approach applies signature and indexing technique for fast structure mining. The same as conserved structure mining, we use neighborhood residues sphere to describe protein local structure, transform each sphere as bit-string signature, and the indexing technique will be applied to provide fast database search. Furthermore, we encode each neighborhood residues sphere as environmental signature for protein structure indexing and quick database searching.
3.2.5. Coordination of Sequence Conservation and Structural Conservation

In accordance with MAGIIC-PRO [8] developed by Hsu et al., which is driven by homologues protein sequence analysis on detecting a functional signature, the authors approach sequence pattern mining to discover functional signatures of a query protein. The authors try to link the relationship between sequence patterns and protein function via the corresponding space information of sequence patterns. From this point of view, they use sequence conservation mining to discover functional motif relative to functional site. But another viewpoint we considered is from local conserved structures, we attempt to discover structure conservation with sequence information integration for analyzing the relationship among sequences, structures, and functions in the future. Functional classification would be a better choice to discover structure-function relation because of protein-ligand complex information. In each enzyme family, proteins within an enzyme family have the same function derived from different species; therefore, it is a good start to discover sequence and structure conservation based on the relationship between sequences, structures and functions.

3.2.6. Apply Mining Results in Function/Structure/Sequence Prediction and Annotation

Computer-aid annotation for protein sequences, structures, and functions has been studied based on protein global sequence and structure information. Recent research has been applied structure properties in primary sequence prediction to improve prediction accuracy. Therefore, our idea starts from protein local sequence and structure to inference its function; therefore, we attempt to include protein structure properties of local region to study the correlation of sequence, structure, and function from the view of local region. In order to annotate protein function, it is alternative to use mining results to predict protein function. This mining result discovered from a group of functional proteins should be significant to its protein function.
4. RESEARCH DESCRIPTION

4.1. Protein Local Structure Representation

4.1.1. Introduction

As protein function is activated in specific region of protein structure especially in local structure; therefore, local structure comparison plays an important role in detecting local structure similarity. Proteins with the same function should share similar local structure and provide binding area to contact with small molecule in order to activate their functions and these local structures are functional areas. In the past, molecular biologists examine lots of functional protein structures to understand the relationships between functionalities, amino acid sequences and protein structures [42, 68, 69]. These studies not only help molecular biologists understand more details about functional proteins but also provide helpful information while encountering unfamiliar proteins. With the help of fast computing machine and delicate algorithms, research staffs can mining more useful sequence and structure from hand-made protein database and further applied the mined knowledge in protein function prediction, active site prediction and other structure based researches.

With the fast growth of Protein Data Bank (PDB) [3, 56], protein functional analysis has become more important. Moreover, protein structure comparison among mass protein structure data is widely applied on protein structure analysis. According to researches and observations, protein function is highly correlated to its three-dimensional (3D) structure and researches are especially focused on special structure fragments which may connect to protein function or overall framework support [70-72]. Local structure similarity [41] can tell us similar local structure which may highly relate to protein function.

Currently, there are two major directions to analyze protein function; one is sequence-level analysis, and another is structure-level analysis. Mining the conservation area related to possible binding area is a hot issue to infer protein function from protein sequence or protein structure analysis. In sequence-level analysis, sequence alignment can be applied to detect conservation among protein sequence although the conservation is rough area [70]. They try to map sequence conservation region into their corresponding 3D space to link the relation between sequence, structure, and function [73]. Now, the question is that could we discover
local structure conservation related functional area, and how to discover. In structure-level analysis, the binding area of protein-ligand complex [39, 44] is widely used to identify protein functions via local structure comparison. Scientists first find protein pockets and voids [71, 72], which are possible binding regions of protein function. These regions can be further investigated in ligand docking and proved that discovered local structure conservations are conserved for protein function. Because homologous proteins may have different functions, it is hard to detect via sequence-based identification if evolution keeps the folding pattern far from sequence identity. Therefore, structure-based identification of homologues would succeed because of structure conservation for keeping protein functionality [74].

4.1.1.1. Motivation

In this study, our motivation is to discover local structure conservation via protein structure analysis. Therefore, we will discuss on local structure representation for structure conservation discovery and related miming approaches or algorithms. Based on the most believed assumption that proteins of same function share common local structure, we developed a different approach which mining the conserved region from the classified enzyme dataset [75]. Therefore, we try to detect or discover similar local structure via different approaches and local structure representations to mine local structure conservation and find the link between local structure and functional region. Beyond that, we will discuss local structure conservation discovery and relationships between local structures and functional regions.

4.1.2. Local Conservation and Functional Site

As found by Campbell and Jackson [53], Src homology 2 (SH2) family can be divided into two groups on the basis of similarity of binding site residues. In this research, it showed that proteins with the same family share similar local sequences and local structures closed to its binding area. The result also showed that sequence conservation would fall on whole sequence diversely but compact in 3D space. In this case, they observed that there exists conservation on local sequence and its corresponding 3D structure and has relationship between local structure and binding area. Moreover, according to MAGIIC-PRO developed by Hsu et al. [49] on detecting functional signature, they approach sequence pattern mining to discover functional signatures of a query protein. Their experimental results showed that gapped local sequence can be detected that its corresponding local structure might be
close to protein functional site.

The function often occurs in cavity, packets or voids of proteins. Therefore, the study of protein local structures is helpful for understanding the protein function. It is also a trend to discover relationship between function and protein local structures. In previous studies, CSA [39] extracts functional site information from research literatures manually; Protemot [44] uses computational approach to detect and extract all protein-ligand complexes in PDB automatically. Another trend on this topic is to discover possible functional areas on protein surface, such as CASTp [72] and pvSOAR [71].

4.1.3. Local Structure Representation

In the task of mining local structure conservation, local structure representation is the first consideration we should regard for. In this study, we first use the straightforward representation of the results derived from protein structure comparison. In addition, we adopt and modify the idea of structural motif of SPratt2 [64]. In SPratt2, they use sphere to describe local structure for discovering structural motif. We will illustrate details in the following sub-sections.

4.1.3.1. Alignment Result of Protein Structure Comparison

To use the alignment result generated by protein structure comparison is the first candidate to mine local structure conservation. While comparing a set of protein structure pair-wise, we can obtain a set of matched Ca points from each compared pair. And then we can apply simple clustering algorithm to group matched Ca points as local structure. Each group will be a representation of local structure for further investigation.

4.1.3.2. Neighborhood Residues Sphere

In order to depict local structure with an appropriate representation, our original idea comes from the NSr, called a neighbor string, developed by Jonassen et al. [65], which is used to mine structural motif. This string encodes all residues in the structure that are with a distance of d Å from r (d=10, as default), including r itself from N-terminal to C-terminal. We redefine NSr to be NRS, neighborhood residues sphere, which includes structure coordinate information therefore the NRS contains
local structure information with its sequence. As shown in Figure 7, if a central residue is colored in red and radius is 10 Å, residues within a blue part is neighborhood closed to central residue within 10 Å.

![Figure 7. Neighborhood residues sphere.](image)

A real case of protein (PDBID: 1AU0). Residues in blue are surrounded by central residue in red within 10 Å distance.

### 4.1.4. Structure Conservation Detection

In order to detect protein local conserved structure related to protein function or closed to protein binding area. In previous researches, the believed assumption is that proteins with the same function share similar local structure. Hence, to mining local structure region that have biochemical meaning will be very useful for identifying protein function. Given a set of protein chains, our goal is to extract local structure patterns shared among those protein chains which have the same function and apply the concept of mining frequent itemset to discover structure conservation [76]. In this section, we will introduce two methods of mining local structure patterns; one is using pair-wise protein structure comparison and another is sphere-based conservation mining approach, and will be illustrated in the following sub-sections.
4.1.4.1. Pair-wise Protein Structure Comparison Approach

In this approach, we use pair-wise protein structure comparison to obtain matched residue, group them as a substructure and check substructure similarity further. Our strategy is to describe local structure representation of matched residues via protein structure comparison and then detect frequent substructure. In addition, we use EMPSC [77] as protein structure alignment tool to compare protein structures pair-wisely. As shown in Figure 8, the overall framework contains three major parts: (I) local structure generation via pair-wise local structure comparison, (II) substructure comparison and similarity measurement, (III) similar substructure grouping and representative pattern selection.

![Flow Chart](image)

Figure 8. The flow chart for mining conserved structural patterns via pair-wise protein structure comparison.

4.1.4.2. NRS-based Conservation Mining Approach

In text mining, mining frequent itemset is often applied to find the frequent term in a corpus. But given a set of protein chains (e.g. 4HHB:A.), can we apply a concept of frequent itemset mining on protein chains? In the Figure 9, we illustrate an overall framework for pattern extraction. Given a set of protein chains, our goal is to extract representatives for a set. Those representatives are considered as conserved patterns.
which most of proteins share these substructures. Because the NRS contains sequence and structure information, we can apply analysis method on sequence and structure data. Our strategy is to apply sequence alignment for sequence conservation and then structure alignment for structure conservation. This framework is divided into three major steps to select conserved pattern for a set of protein chains: (I) NRS segmentation, (II) sequence conservation grouping, and (III) representative selection.

Figure 9. The flow chart for mining conserved structural patterns via NRS-based conservation mining approach.

4.1.5. Experiments

In order to compare two approaches on detecting structure conservation, we use enzyme classification as our data collection, and approach these two methods to figure out structure conservation in local region and find out the relationship between local structure regions and substrates or ligands. According to PDBSProtEC [13], we randomly select 6 EC families as our dataset to evaluate these two methods. In Table 2, we list all protein chains after removing identical protein sequences for these 6 EC families. In addition, substrate information is selected from PDBSum [12] (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/).
Table 2. List of protein chains for 6 randomly selected EC families.

<table>
<thead>
<tr>
<th>EC Numbers</th>
<th>List of Protein Chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.2.17</td>
<td>1FGS:_ 1JBV:A 1W78:A 2GC5:A 2GC6:A 2GCA:A 2GCB:A</td>
</tr>
</tbody>
</table>

Table 3. Experimental results for local conservation discovery via pair-wise protein structure comparison.

<table>
<thead>
<tr>
<th># of local conservation</th>
<th># of ligand contact</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSC based</td>
</tr>
<tr>
<td>1.6.2.4</td>
<td>13</td>
</tr>
<tr>
<td>1.14.99.3</td>
<td>5</td>
</tr>
<tr>
<td>2.3.1.74</td>
<td>16</td>
</tr>
<tr>
<td>4.1.2.17</td>
<td>0</td>
</tr>
<tr>
<td>5.3.1.9</td>
<td>7</td>
</tr>
<tr>
<td>6.3.2.17</td>
<td>4</td>
</tr>
</tbody>
</table>

4.1.6. Discussions

4.1.6.1. Pair-wise Protein Structure Comparison Approach

In Table 3, we list number of local conservation we found and number of substrate contacts within 10 Å between substrate and discovered local conservation. In the experimental results, not all EC family will discover local conservation because their global structures might be too similar or diversity. The experimental results reveal that we don’t detect in EC 4.1.2.17, and we find these sequences share above 90% sequence identity within this EC family, checked by BLASTCLUST [36]. Therefore, it is hard to use this approach to detect local conservation because above 90% sequence identity means that they have the same global structures. In addition, the
reason why we list the value of number of substrate, ligand, or metal ion is try to connect the relation between local conservations and substrates.

Although we only test few cases on discovering conserved structure patterns of proteins with same function, the result reveals that local structure conservation region could be detected under functional classification. We select all possible substrates information related to protein chains. In Figure 10, the picture shows the relationships between conserved patterns and substrates, and the protein PDBID is 1J9Z:A and substructures are areas colored in yellow, aqua, or lime and the ball colored in red, blue, and navy are substrates (Navy: FAD, Red: NAP, Blue: FMN). Moreover, we also find that local conservations discovered in proteins of PDBID 1BVY:A, 1AMO:A, 1BU7:A, 1SMI:A, 1B1C:A have substrate/ligand contacts such as FMN, EDO, FAD, HEM, and NAP.

4.1.6.2. NRS-based Conservation Mining Approach

For each EC family, we apply NRS-based conservation mining approach to mine local
conservation. Because of large amount of spheres, we first apply sequence alignment to group similar sequence and further check their structure similar within a group via geometric hashing. In Table 3, we also list the values of number of local conservation and number of substrate, ligand, or metal ions respectively. We still have two EC families, EC 1.14.99.3 and EC 2.3.1.74, that local structure conservation could not be detected. In EC 2.3.1.74, their sequences share above 90% sequence identity. And in EC 1.14.99.3, there are still 3 protein chains while the cut-off of sequence identity is below 50%.

As shown in our experimental results, conserved patterns are mined from protein chains with the same EC labels sharing highly conservation in local structure and conserved patterns have high capacity to identify. In addition, we also find that protein chains within the same EC labels can be grouped into more than two sub-groups. For example, while applying this approach on whole EC families, in EC 3.2.1.17, there are totally 895 protein chains, and we mined two conserved patterns. However, 326 protein chains share one of them, and 417 protein chains share another one, but these two conserved patterns have no overlapping region. According to our observation, number of conserved patterns has relation to the number of protein chains. In general, the more in the number of protein chains within the same EC labels, the lower in the number of conserved patterns, if protein chains within an EC label have diversity.

4.1.6.3. Summarization

As shown in Figure 11, this is PDBID 1SMI:A and the substrate is HEM (PROTOPORPHYRIN IX CONTAINING FE). The area colored in blue is the local conservation discovered by NRS-based conservation mining approach and the central residue is colored in red, and the area color colored in yellow are two local conservation discovered by pair-wise protein structure comparison approach. In addition, the area in pink is the area the overlapping area discovered by these two approaches. Comparing with these two approaches, local conservation detected by pair-wise protein structure comparison approach will be more fragment than NRS-based conservation mining approach. The reason is that NRS is more suitable to describe residue environmental information, but a group of matched residue points just provides local similar area and it is not a well-organized structure representation.
Figure 11. PDB ID 1SMI:A and the substrate is HEM.

The areas colored in yellow and blue are conserved local structure by protein structure comparison approach and NRS-based approach respectively. The area colored in pink is the overlapping area that both approaches discovered.

4.1.7. Conclusions

In this study, we try to find out relationships between local conservations and functional area via mining frequent itemset. Our purpose is to use different local structure representations as itemset and then apply mining frequent itemset to discover local structure conservation. Although the alignment results as local structure representation are not well-organized representation, it still provides us examples to realize how conservation could be formed in protein structure. Furthermore, we use neighborhood residues sphere as local structure representation to describe local structure. We use EC family to verify our purpose because of the ease of substrate/ligand verification. Therefore, we can use ligand contact to explain what we discovered. In our experiments, conserved local structure can be discovered and the observations show contact areas but not all elements of substrate contact with a substructure. We can discover conserved local structure region from functional hierarchical classification because proteins have the same function will
share some attributes reflect on their structures.

4.2. Protein Structure Conservation Mining

4.2.1. Introduction

Molecular biologists examine many functional protein structures to understand the relationship among functions, amino acid sequences and protein structures [42, 68, 69, 78, 79]. These analyses not only help molecular biologists understand more details about functional proteins, but also provide helpful information when encountering unfamiliar proteins. Now with the help of fast computing machines and delicate algorithms, research staffs can mine more useful sequence and structure information from a hand-made protein database, and then can apply the mined knowledge in protein function prediction, binding site prediction, protein fold prediction, and other researches which are based on protein structure information.

Based on the common assumption that proteins of the same function share common local regions, the concept of local region conservation comes from a motif, which is a fragment with biological or functional meaning. Both sequence motif and structure motif can be deduced from the discovered sequences and structures. Currently, there are two major directions to analyze protein function; one is sequence analysis, and another is structure analysis. In sequence analysis, multiple sequence alignment or pair-wise sequence alignment can be applied to detect conservation among protein sequences, although the conservation would be a rough area [70]. This analysis tries to map sequence conservation region into their corresponding 3D space to link the relation among sequence, structure, and function [73]. Campbell and Jackson found that Src homology 2 (SH2) family can be divided into two groups on the basis of binding site residues similarity [45, 53]; thus sequence conservations, which is related to their binding area, could be discovered. Moreover, according to MAGIIC-PRO developed by Hsu et al. [49], which is driven by homologues protein sequence analysis on detecting a functional signature, the authors approach sequence pattern mining to discover functional signatures of a query protein.

On the other way, researchers try to discover local structure conservation related to a functional area. In structure analysis, the binding area of protein-ligand complex [80] is widely used to identify protein function via local structure recognition. CSA, Catalytic Site Atlas [39], is a manually curated template library of protein-ligand
templates from literatures. Protemot [44] is another web service using protein-ligand complexes via computational advantage. Template is used to find binding residues in a protein surrounding a ligand within 6.5 Å distance; therefore, the template can be extracted automatically. Scientists first find protein pockets and voids [71, 72], which are possible binding regions of protein function. These regions can be further investigated in ligand docking, and scientists have proved that discovered local structure conservations are conserved for protein function. Because homologous proteins often have different functions, they are hard to detect via sequence-based identification if evolution keeps the folding pattern far from sequence identity. Therefore, structure-based identification of homologues would succeed because of structure conservation for keeping protein functionality [74].

Because proteins provide the same function, they may share some degree of folded conformation to express their function. Thus, in this paper, our motivation is to develop an approach of mining technique in functional families [76] without the help of protein-ligand complex information. In previous research [81], the authors point out that non-homologous proteins may have the same function; in the other words, proteins have dissimilar global structures may have the same function, and the observations can be found that function may occur in protein local structure. Comparing with protein local structures can be used to predict protein function. The local structures are usually assembled by shorter sequence segments i.e. protein binding sites, and they have some kind of conservation on sequence-level. Although there may be mutations in part of the sequence, we can also find conservation in local sequence segments. Thus, we believe that local sequence similarity has both higher sensitivity than global sequence similarity and higher significance for inferring function. In this paper, we adopt sphere based representation to describe local structure, and then apply mining technique to discover conservation regions which conserved in both local sequence and local structure.

4.2.2. Local Structure Representation

In data mining, feature extraction/selection is very important for classification or prediction. Hence, we have to define local structure representation for protein three-dimensional structure. Our original idea comes from the neighbor string (NSr,) developed by Jonassen et al. [64]. This string encodes all residues in the structure that are with a distance of d Å from r (d=10, as default), including r itself from N-terminal to C-terminal. This distance cut-off of 10 Å [82] is Van der Waals contribution and it dominates for less then 3 Å but is insignificant at 10 Å. The
origin of NSr is used to mine structure motif in Protein Data Bank (PDB). The authors use NSr to represent structure motif and use support $k$ of structure occurrences to decide which NSr is a significant structure motif. In addition, NSr is represented in regular expression encoded in gap information. In this paper, we redefine NSr to be, neighborhood residues sphere (NRS) to include structure coordinate information; therefore, the NRS contains local structure information with its sequence. Thus, the NRS has compact spatial conformation and gapped sequence information. As shown in Figure 12, if $G$ is a central point and the radius is 10 Å, residues within the gray part are a neighborhood closed to central residue with 10 Å. The sequence from N-terminal to C-terminal is ACWILYGT. The local structure representation is then used to mine local region conservation.

Figure 12. Neighborhood Residues Sphere.
4.2.3. Mining Conserved Patterns

In order to detect protein local conserved structure related to protein function or closed to protein binding area, we apply mining technique to discover conserved regions in protein structure. In previous researches, the believed assumption is that proteins with the same function may share similar local structure. Hence, to mine local structure region that have biochemical or functional meaning will be very useful for identifying protein function. Given a set of protein chains, our goal is to extract local structure patterns shared among those protein chains which have the same function. We use neighborhood residues sphere (NRS) as local structure representation, an itemset which contains both sequence and structure information, and then approach mining technique to discover conserved pattern [78]. During the mining process, we have to cluster the similar NRSs rather than just check the pattern frequency, as there are tiny differences between conserved NRSs from two different proteins.

Figure 13 illustrates an overall framework for mining frequent itemset in Protein Data Bank. Given a set of protein chains, our goal is to extract representatives for a set. Those representatives are considered as conserved patterns, and most of proteins have these substructures. Because the NRS contains sequence and structure information, we can apply an analysis of NRS for both sequence and structure data. To avoid a
huge local structure similarity comparison, we further apply dynamic programming of the Smith-Waterman algorithm and geometric hashing for NRS sequence and structure analysis respectively. Both two approaches are time consumed because of fully pair-wise comparison. This framework is divided into three major steps to select conserved patterns for a set of protein chains: (I) NRS segmentation, (II) sequence conservation grouping, and (III) representative selection.

4.2.3.1. NRS Segmentation
In NRS segmentation, we sequentially segment neighborhood residues spheres for a protein chain from N-terminal to C-terminal, residue by residue. If we have \( l \) residues in a protein, \( l \) NRSs will be outputted. Each NRS contains sequence and atom coordinates information for the next step. While applying NRS segmentation, we use a grid-based segmentation approach to speed up the performance. According to whole NRSs, the distribution of NRS length and frequency ranges from 13 to 23.

4.2.3.2. Sequence Conservation Grouping
At the step of sequence conservation grouping, we separate into two sub-steps: (a) sequence alignment, and (b) sequence clustering. In the sub-step of sequence alignment, the Smith-Waterman algorithm is applied to identify sequence identity. In order to keep flexibility in sequence alignment, we use PAM250 as the amino acid substitution matrix to keep positive mutation. Hence we can have an advantage by filtering out dissimilar sequences and reserving higher levels of tolerance. Each alignment score, SWscore, is normalized as NScore defined in equation (1), where NRS1 and NRS2 are derived from different protein chains.

In the sub-step of sequence clustering, we are going to group similar sequence segments according to the NScore of each pair. Sequence segments derived from the same protein chain are not taken into account; so the score will be zero. Then we use the average-link clustering approach, hierarchical agglomerative clustering algorithm [83], to cluster all pairs of sequence segments, and the threshold is set at 3.5 by experimental evaluation. After the threshold cut, we leave the largest cluster(s) as candidate set(s). In a candidate set, we can guarantee that sequence segments within a cluster share high conservation. The reason we group similar local sequences is that pair-wise structure comparison is more time-consumed than pair-wise sequence comparison; therefore, pair-wise sequence comparison can help us to filter out dissimilar sequences before checking structure similarity.
\[
N\text{Score} = \frac{SW\text{score}(NRS_1,NRS_2)}{\text{maxlength}(NRS_1,NRS_2)}
\]  
\[
GH\text{-score} = \frac{\# \text{ of aligned residues}}{\text{min}(NRS_1,NRS_2)}
\]

4.2.3.3. Representative Selection

In the step of representative selection, in order to keep sequence-structure consistent, we have to identify the structure confirmation within a candidate set. We use modified geometric hashing which adopts the characteristic of NRS that a central point should be superimposed while comparing two NRSs. Then the GH-score is defined as equation (2) to recognize structure similarity, where NRS\(_1\) and NRS\(_2\) are derived from different protein chains. If the average structure similarity within a cluster passes the threshold of GH-score, this candidate set is considered a significant set. Therefore, we select a representative NRS for a significant set by finding the one that is nearest to others. Currently, the threshold for GH-score is 0.8 by experimental evaluation.

4.2.4. Template Library

For the purpose of functional prediction, we build a template library of enzymes for EC family (or label) prediction. Because proteins in enzyme classification are classified by their functionality or reaction, we try to predict enzyme function via discovered conserved patterns. Based on PDBProtEC [84], a resource links PDB chains with Swiss-Prot codes and EC numbers, and we can gather protein structures with their corresponding EC labels. From 13,373 enzymes distributed over 563 four level EC labels, we randomly select 1,000 non-redundant protein chains as testing samples with a sequence identity less than 60\%, and the others are training samples. All training samples will be used to extract conserved patterns.

As illustrated in section 3, we extracted conserved patterns as the template library for all EC labels, and we try to verify our assumption and the effectiveness of these templates with enzyme classification prediction experiments. We only select EC labels with more than two proteins in order to extract conserved patterns; so, we have 563 EC labels and 12,373 training samples. Unfortunately, not all EC labels have
conserved patterns; hence, we only have 456 EC labels with conserved patterns. Because of consideration of both local sequence and structure conservation, not all EC labels have significant conserved patterns. According to experimental observations, the reason is that NRS shared higher global sequence similarity but lower structure similarity or lower local sequence similarity. Currently, we obtain 56,164 NRSs among 456 EC labels of conserved patterns out of total 646 EC labels where 563 EC labels have more than two proteins, and the average size of these NRSs is 20.5. By comparing with NRSs of conserved patterns and overall NRSs, NRSs of conserved patterns (18–25 residues) have more residue numbers than overall NRSs (13–23 residues).

4.2.5. Enzyme Classification Prediction

Prediction by similarity, i.e. predicting function using similarity at the sequence level, is a very strong theme in genome annotation, and recent years have seen much discussion of the precise nature of the relationship of protein similarity at the sequence, structure, and functional levels. Recent researches reported that analysis of protein structure provides insightful ideas about the biochemical functions and mechanisms of proteins (e.g. active site, catalytic residues, and substrate interaction) [70-72]. Observations on the relationship among local sequence, spatial structure and protein function have been discovered. The enzyme classification, published by the International Union of Biochemistry and Molecular Biology in 1992, is in its sixth edition. This hierarchy is built by grouping enzymes with protein functions or reactions. Therefore, the hierarchy is a good source to observe the relationships between proteins at the sequence, structure, and functional levels. Given an unknown function protein as a query protein; our prediction procedure will give a predicted EC label. Because we have to test all EC labels, every query protein has to be compared with all patterns in the template library. The overall predication framework is showed in Figure 14 and detailed information is illustrated later.
First, given a query protein, we segment NRSs for the query protein. Next, we apply sequence alignment on query NRSs against conserved patterns in the template library to obtain alignment scores and threshold cut-offs to filter out dissimilar NRSs. If the pair-wise alignment score is higher than the threshold, structure alignment is applied to verify structure similarity. In order to keep sequence-structure consistent, after the procedure of sequence conservation grouping, structure level verification is necessary. In order to compare with CSA and Protemot, we adopt the assessment and evaluation defined by Protemot. In Table 4, detail description of the assessment is illustrated. We also use two evaluation equations defined in Protemot, (3) and (4), to evaluate EC label prediction. In the equation, A means “in lib” correct, B means “in lib” incorrect, C means “in lib” no prediction, D means “out lib” incorrect, and E means “out lib” no prediction. In Table 2 (a), we use mined conserved patterns as prediction patterns to predict EC labels and the prediction result shows 83.45% Confidence and 67.02% Accuracy in the 1,000 protein chains randomly selected from 13,373 enzymes among the 563 EC labels which are not in the training data set by ourselves.

\[
    \text{Confidence} = \frac{A}{A+B+D}
\]  

(3)
\[ \text{Accuracy} = \frac{A}{A + B + C + D} \]  

(4)

Table 4. Description of assessment.

<table>
<thead>
<tr>
<th>Condition(^a)</th>
<th>Assessments</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>in lib</td>
<td>Correct (A)</td>
<td>Answer EC label matches at least one predicted EC label(s), predicted EC label may be more than one.</td>
</tr>
<tr>
<td></td>
<td>Incorrect (B)</td>
<td>Answer EC label doesn’t match any of prediction EC label(s), predicted EC label may be more than one.</td>
</tr>
<tr>
<td></td>
<td>No prediction (C)</td>
<td>No predicted EC label output.</td>
</tr>
<tr>
<td>out lib</td>
<td>Incorrect (D)</td>
<td>Answer EC label doesn’t exist in our training EC labels, but we predict.</td>
</tr>
<tr>
<td></td>
<td>No prediction (E)</td>
<td>No predicted EC label output.</td>
</tr>
</tbody>
</table>

\(^a\) If the EC label of testing protein belongs to 465 EC labels, a testing protein is “in lib” (template library) prediction, otherwise “out lib” prediction.

### 4.2.6. Comparison with other Template Libraries

This section compares our built template library with other template libraries. It has been observed that enriched collections can improve prediction accuracy. Therefore, in constructing a template library, we iteratively extract conserved patterns for all EC labels. In our template library, our conserved patterns cover over 456 EC labels and the coverage are about 80% of 563 EC labels with more than two proteins.

The evaluation has been conducted with comparisons against the prediction power of template libraries based on CSA-based web server and Protemot web server. CSA-based web server is located at http://www.ebi.ac.uk/thornton-srv/databases/CSA/, in which CSA, Catalytic Site Atlas, is a manually-curated collection from literatures. This contains two types of entries, the original of the enzyme from hand-annotation and a homologous set by PSI-BLAST. Protemot is also a web server located at http://protemot.csbb.ntu.edu.tw/. Inside the Protemot, their template library is constructed by protein-ligand complexes. The template is extracted from residues surrounded by ligand within 6.5 Å scope; so, only EC labels with protein-ligand complex have templates. As Protemot emphasizes, the template library is automatically collected by extracting all possible protein-ligand complexes.

As shown in Table 5 (a), we randomly select 1000 protein chains which we exclude
from the training dataset and the experimental results reveal that Confidence is 83.45% and Accuracy is 67.02%; (b), our template library has doubled Confidence level than CSA and Protomot, and our performance is better than CSA and Protomot in 20% better than CSA and 10% better than Protomot in Accuracy with the same dataset tested by CSA, Protomot, and our proposed approach. The dataset is generated by Protomot, and these three approaches use the same dataset. Comparing the number of templates and the coverage rate, we have 56,164 templates while CSA has 147 templates and Protomot has 1051 templates, and our coverage rate is about 80% while CSA covers around 30% and Protomot covers 55%.

Table 5. Experimental results for enzyme classification prediction.

(a) The experimental result of 1,000 random protein chains selection.

<table>
<thead>
<tr>
<th>Conserved patterns (Proposed approach, NRS)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>in lib</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (A)</td>
<td>575</td>
<td></td>
</tr>
<tr>
<td>Incorrect (B)</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>No prediction (C)</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>out lib</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect (D)</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>No prediction (E)</td>
<td>142</td>
<td></td>
</tr>
<tr>
<td>Testing samples</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>83.45%</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>67.02%</td>
<td></td>
</tr>
</tbody>
</table>

(b) The experimental result of 1,000 random protein chains selected by Protomot for evaluating the performance of CSA, Protomot, and NRS.

<table>
<thead>
<tr>
<th></th>
<th>CSA(^a)</th>
<th>Protomot</th>
<th>NRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>in lib</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (A)</td>
<td>75</td>
<td>408</td>
<td>424</td>
</tr>
<tr>
<td>Incorrect (B)</td>
<td>8</td>
<td>310</td>
<td>46</td>
</tr>
<tr>
<td>No prediction (C)</td>
<td>63</td>
<td>14</td>
<td>274</td>
</tr>
<tr>
<td>out lib</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect (D)</td>
<td>77</td>
<td>14</td>
<td>56</td>
</tr>
<tr>
<td>No prediction (E)</td>
<td>777</td>
<td>254</td>
<td>200</td>
</tr>
<tr>
<td>Testing samples</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Confidence</td>
<td>46.88%</td>
<td>41.98%</td>
<td>80.61%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>33.63%</td>
<td>41.38%</td>
<td>53%</td>
</tr>
</tbody>
</table>

\(^a\) (highly probable + probable)

However, we may predict more than one EC label for testing a protein. From our observation, we find that only 78 out of 1,000 proteins have multiple predicted EC
labels. There are 53 proteins match one of predicted EC labels, 6 incorrect predicted EC labels in lib, and 19 incorrect predicted EC labels out lib. In Table 6, we list 4 sample protein structures with predicted EC labels and answer labels. According to this prediction results, we have capability to detect multiple EC labels via discovered local structure, but we still can’t distinguish major or minor conserved regions under functional hierarchical classification. However, there is no explicit description of a major or minor functional area, it is hard to evaluate multiple label prediction even though we can detect all possible multiple labels.

<table>
<thead>
<tr>
<th>PDBID</th>
<th>EC labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PJT</td>
<td>1.3.1.76, 2.1.1.107, 4.99.1.4 (predicted)</td>
</tr>
<tr>
<td></td>
<td>1.-.-.-, 2.1.1.107, 4.99.1.- (PDB)</td>
</tr>
<tr>
<td>1V3T</td>
<td>1.3.1.48, 1.3.1.74 (predicted)</td>
</tr>
<tr>
<td></td>
<td>1.3.1.48, 1.3.1.74 (PDBSum) / 1.3.1.48 (PDB)</td>
</tr>
<tr>
<td>1RBM</td>
<td>2.1.2.2, 6.3.3.1, 6.3.4.13 (predicted)</td>
</tr>
<tr>
<td></td>
<td>2.1.2.2 (PDB), 6.3.3.1, 6.3.4.13 (PDBsum)</td>
</tr>
<tr>
<td>1YV5</td>
<td>2.5.1.1, 2.5.1.10 (predicted)</td>
</tr>
<tr>
<td></td>
<td>2.5.1.10 (PDB / PDBsum)</td>
</tr>
</tbody>
</table>
326 proteins share (a) is a representative (PDBID: 1GBW), and 417 proteins shares (b). The red one is central residues, and the blue part is the area surrounding central residue.
4.2.7. Discussion

Our experimental results reveal that conserved patterns discovered from protein chains with the same EC labels share high conservation in local structure and that conserved patterns have a high capacity to be identified. In addition, we also find that protein chains within the same EC labels can be grouped into more than two sub-groups, and each sub-group can have different conserved patterns. In our experiment, proteins within the same EC label have also observed sub-groups. For example, in EC 3.2.1.17, there are totally 895 protein chains, and we mined two conserved patterns. However, 326 protein chains share one of them, and 417 protein chains share another one, but these two conserved patterns have no overlapping region as shown in Figure 15.

In the overall framework, we have threshold cut-off for sequence alignment, sequence clustering, structure similarity evaluation, and representative selection; the values are decided by experimental testing. In EC family prediction, we find that we have many “incorrect” predictions, and the reasons are threshold setting, and the relationship of sequence-structure consistency. If we increase the threshold value for sequence clustering, we can reduce the rate of “incorrect” prediction. Hence, we infer that conservation in both sequence and structure level can improve Confidence and Accuracy rate in predicting EC labels. Additionally, from our observation on ligand HEM (PROTOPORPHYRIN IX CONTAINING FE, C34H32N4O4Fe) as shown in Figure 16 (a), the 3D structure of HEM is flat. If a protein structure wants to contact with this ligand, we guess that it will be an area like a bed to support HEM. Figure 16 (b) is one of real cases that it is the discovered conserved local structure surrounding a ligand, HEM, and we observe that there exists a supporting area to bolster up a ligand in this case. In addition, we also observe many cases of conserved local structures surrounding a ligand, HEM. Fortunately, we find that our conserved structures have this kind of characteristics across multiple EC families.
Figure 16. Conserved local structure and a ligand.

(a) Crystal structure of HEM (PROTOPORPHYRIN IX CONTAINING FE, C34H32N4O4Fe). (b) Discovered conserved local structure surrounding the ligand, HEM.
4.2.8. Conclusion

The threshold value of sequence similarity and GH-score significantly affects the “no prediction” rate of prediction. In enzyme classification prediction, the experimental results show that the coverage rate of a template is correlated to the confidence level of classification prediction. Although there are still many cases that have “no prediction,” this results in the threshold of sequence or structure similarity which reflects the level of conservation. For example, in enzyme classification, we can find some conserved regions in protein chains within the same EC labels, and those conserved regions have higher sequence similarity and have similar conformation in spatial structures.

According to the experimental results, we believe that proteins with the same function have conservations; however, not all of them have conservation on sequence and structure. In our template library, we have about 80% coverage in enzyme classification. From our observations, predefined classification is very important for prediction; thus, in enzyme classification, we find that functional classification is significantly beneficial to mine conserved patterns significantly to identify EC label. Comparing with CSA and Protemot, our approach tries to apply the concept of “mining frequent itemset” to identify conserved region for recognizing EC family without using protein-ligand complexes. According to our observation, we suggest that it is possible to have different levels for sequence and structure thresholds to achieve different levels of conservation in sequence or structure.

To evaluate the property of conserved region is still hard to recognize structural conservation and functional conservation. From our observations, we find that some conserved regions are neighbors to ligand or substrate but some are not. Figure 17 is an example of the relationship of conserved pattern and ligand, where (a) and (b) are different views of protein and the PDBID is 1AU0. There are two conserved patterns inside the protein. The red residue and green residue are the central point of each pattern. The blue area is the NRS of the red residue, and the yellow area is the NRS of green residue. The pink one is the ligand named SDK (1,3-BIS[[N-[(PHENYLMETHOXY)CARBONYL]L-LEUCYL]AMINO]-2-PROPANONE). According to these two pictures, the NRS in blue has contact to the ligand. We assume that our conserved regions may have structural or functional properties related to binding area. Hence, discussion on the relationship between conserved pattern and ligand is necessary in the future. In addition, substrate is also a subject, and we can study relationships between substrate and conserved pattern.
Figure 17. Conserved pattern and ligand, SDK, of protein PDBID 1AU0.

There are two conserved patterns inside the protein. The red residue and green residue are the central point of each pattern. The blue area is the NRS of the red residue, and the yellow area is the NRS of green residue. The pink one is the ligand named SDK.
4.3. Protein Structural Property Exploration

4.3.1. Introduction

As of July 3, 2007, there are 44,476 determined protein structures examined by X-ray or nuclear magnetic resonance (NMR) in Protein Data Bank (PDB) [85]. They include proteins, protein complexes, nucleic acids and protein nucleic acid complexes. Applying mining technique on protein structures is an interesting issue to discover residue environmental information inside protein structure [86-88]. Residue environment has been studied for many years and applied on protein threading and protein binding site characterization [89, 90]. In the protein structure, a residue is the essential element for conformation, and residue-residue contacts will affect the overall framework of a protein structure. Therefore, residue environment can help us to comprehend protein structure conformation. In addition, binding site environment analysis is also a good starting point to understand how residue contacts affect protein binding and protein function [43, 73].

In previous researches, residue-residue contact is an important issue to be investigated for protein structure fold, protein structure conservation, and protein function [89, 91-94]. With the fast growth of protein structure, it provides more materials on the study of discovering local residue environment with/without chemical bond information. Furthermore, protein conformation is highly correlated to residue contact with chemical bonds such as covalent bonds, ionic bonds, hydrogen bonds, Van der Waals attractions, or disulfide bonds. For quick searching of residue environment, we use residue environmental sphere to describe environment information surrounding a residue. On the purpose of protein structural property exploration, we have to trace residue neighborhood on whole protein structure collection. Furthermore, to handle huge protein structure collection is also a great challenge to store entire structure and sphere information in database.

4.3.2. Review of Protein Structural Property Exploration

In sequence based prediction, the position-specific scoring matrix (PSSM) is used to improve their prediction accuracy for protein sequence analysis. The PSSM gives the log-odds score for finding a particular matching amino acid against a target sequence. Therefore, the prediction tools treat PSSM as sequence property for each
amino acid. In protein structure prediction, amino acid property, secondary structure information, b-factor, accessible surface area (ASA), or relative solvent accessibility (RSA) are structural properties. In 1992, Singh and Thornton [54] discovered the atlas of protein side-chain interaction to understand sidechain-sidechain interactions. In this research, they revealed interactions for 20 * 20 amino acids, and counted the frequency for each amino acid pairs. In addition, Glaser et. al. [55] also studied structural property of residues at protein-protein interfaces. In order to realize the inside of protein structure conformation, protein structural property exploration is very important such as amino acid interactions or residue-residue contact.

4.3.3. Proposed Indexing Mechanism for Massive Structural Property Exploration

4.3.3.1. Residue Environmental Sphere and Indexing Mechanism

In order to describe residue environment of protein local structure, our original idea comes from the neighbor string (NSr,) developed by Jonassen et al. for mining structure motif [64]. This string encodes all residues in the structure that are with a distance of d Å from r (d=10, as default), including r itself from N-terminal to C-terminal. The protein structure is folded by the interactions between amino acids to connect with each other; therefore, amino acid plays an important role on protein folding. Therefore, each 10 Å sphere representation, residue environmental sphere (RES), can describe environmental information inside a protein. This distance cut-off of 10 Å [82] is Van der Waals contribution and it dominates for less then 3 Å but is insignificant at 10 Å. And we know that residue-residue interaction will affect protein structure conformation so that the residue environmental sphere should be a good candidate to extract residue environment to understand residue-residue contact for each protein structure. Figure 18 is an example to illustrate residue environmental sphere as indexing unit. Now, we use RES to identify each local structure surrounding a residue, and it is also a index unit to index protein structure residue by residue for quick database search, and this sphere is the essential/abstract form to record environmental information such as nearest neighbor residues, secondary structure information, biochemical property, and so on. With the great help of database, we store all structure information and index entire residue environmental sphere for analyzing residue-residue contacts.
4.3.3.2. Materials

In this work, we analyze entire protein structures in Protein Data Bank, and all structure information will be considered, such as coordinate information, connectivity annotation, heterogen information, physicochemical properties, and secondary structure information. In coordinate information, both ATOM and HETATM will be considered for protein structures, DNA/RNA structures, and hetero-atom structures respectively. The heterogen information is extracted from pdb file with HET and HETATM tags, which describe non-standard residues, such as prosthetic groups, inhibitors, solvent molecules, and ions for which coordinates are supplied. In our database implementation, DNA/RNA structures could be viewed as special chemical components. In connectivity annotation, SSBOND is the most important information to observe disulfide bonds both intra-molecularly and inter-molecularly. The fundamental physicochemical properties will be also concerned include hydrophobic, hydrophilic, charge (negative and positive), polar, etc. Currently, we select whole protein structures of 43427 as our data collection from Protein Data Bank in early 2007. In this collection, there are 40303 protein structures, 1152 protein/DNA complexes, 465 protein/RNA complexes, 28 DNA/RNA hybrid structures, 43 protein/DNA/RNA complexes, 892 DNA structures, and 544 RNA
structures.

Figure 19. Database table schema for structural property exploration.

4.3.3.3. Database Design

For the purpose of quick search on residue environment, we use residue environmental sphere as indexing unit to speed up table lookup and mine residue-residue contacts. Cooperating with atom coordinate table, and ligand/substrate table, it can be easy to mine residue environment surrounding a residue. In Figure 19, we illustrate database table schema for atom, hetatom, ligand, and residue environmental sphere. In database design, the great challenge is to put huge scale of protein structure into tables includes residue environmental sphere, coordinate information, substrate/ligand/DNA/RNA information, and bone connectivity. As we know, each PDB ID has 4-character code that uniquely defines an entry in the Protein Data Bank. The first character must be a digit from 1 to 9, and the remaining three characters can be letters or numbers. Therefore, we use middle two characters as table identifier; for example, if the PDB IDs are 4hhb, 2hhb, and 3hhb, their atom coordinates will be stored together in the database with table identifier “hh”. At last, we have 4 kinds of database tables to store protein structure information, and they are atom coordinate table, ligand/substrate table, and residue environmental sphere table. Unlike data cube structure, we don’t use grid structure.
to describe a protein structure, and a residue environmental sphere is used to describe neighborhood information surrounding a residue.

4.3.4. Statistical Analysis of Structural Properties on Protein Data Bank

4.3.4.1. Residue-Residue Contacts

In protein structure, residue-residue interactions make a protein to fold as a stable conformation. If two residues are considered to be in contact with each other provided the distance between their alpha carbon atoms (Cα) below a certain cutoff. Therefore, we collect residue-residue contacts from whole protein structures and extract all residue pairs and its neighbor residues to understand how interactions help protein folding. Moreover, each residue can have multiple properties on it such as biochemical property (hydrophobic, hydrophilic, charge, etc), physicochemical property, and secondary structure element type (α-helix, β-sheet, or coil). Inside the residue environmental sphere, we first use Cα in backbone to represent geometry information, but in order to describe detail residue contact with chemical bond, therefore, atom level residue-residue contacts will be also considered.

4.3.4.2. Chemical Component Contacts

In this sub-section, we try to observe residue environment surrounding a chemical component to understand the interaction environment between protein and ligand or substrate. We also use residue environmental sphere to observe chemical component close to a residue contacts. According to PDB format, HET records are used to describe chemical components or non-standard residues, such as prosthetic groups, inhibitors, solvent molecules, and ions for which coordinates are supplied. Groups are considered HET if they are not part of a biological polymer described in SEQRES and considered to be a molecule bound to the polymer, or they are a chemical species that constitutes part of a biological polymer that is not one of the following: (a) not one of the standard amino acids, and (b) not one of the nucleic acids (C, G, A, T, U, and I), and (c) not an unknown amino acid or nucleic acid where UNK is used to indicate the unknown residue name. Because we focus on residue-residue contacts to realize how they interacts with chemical component, and chemical component information is used to understand how interaction begins.
4.3.5. Property Analysis on Disulfide Bond

4.3.5.1. Disulfide Bond

In general, disulfide bonds are suggested to stabilize protein folding which has been reviewed [95-98]. In biochemistry, disulfide bond or disulfide bridge is connected between Cβ-Sγ-Sγ-Cβ (Sγ is a SG atom in PDB, and Cβ is a beta carbon) which can occur intra-molecularly (i.e. within a single polypeptide chain) and inter-molecularly (i.e. between two polypeptide chains). Disulfide bond in intra-molecular stabilize the tertiary structures of proteins while those that occur inter-molecularly are involved in stabilizing quaternary structure. In this paper, we focus on SSBOND section which identifies each disulfide bond in protein and polypeptide structures by identifying the two residues involved in the bond. Furthermore, we also use residue environmental sphere to detect residue-residue contacts of cysteine pairs intra-molecularly.

4.3.5.2. SSBOND

In PDB, the connectivity annotation section is used to allow the depositors to specify the existence and location of disulfide bonds and other linkages. The bond between two Sγ atoms is disulfide bond annotated as SSBOND by Protein Data Bank. We separate this collection into two groups, intra-molecular and inter-molecular; therefore, we have 48152 pairs in intra-molecular group and 2115 pairs in inter-molecular group. While applying secondary structure information, we observe that SSBOND tends to grasp at β-sheets and coils.

4.3.5.3. Residue-Residue Contacts of Cysteine Pairs

Unlike SSBOND discovery, not all protein structures contain disulfide bonds; therefore, we observe all cysteine pairs in whole PDB to distinguish the difference between SSBOND and residue-residue contacts of cysteine pair. In this work, we only collect all cysteine pairs in both Cα and atom level (Sγ) intra-molecularly to observe their environment. The reason to use atom level discovery is that we will miss some cysteine pairs if we only count Cα atom level. Therefore, we have 114,777 residue-residue contacts intra-molecularly for further analysis.
4.3.6. Results

Although we detect all possible residue-residue contacts among whole protein structures in PDB; according to previous studies, we select SSBOND annotation in PDB and residue-residue contacts of cysteine pair as example to explore protein structural property because of well-studied topic on disulfide bond.

4.3.6.1. Residue-Residue Contacts and Chemical Component Contacts

We detect all pairs of amino acid combination to discuss relationship among residue interaction and secondary structure property. In our experimental result, the top-10 residue-residue contacts contain Glycine, and the pairs are Gly-Gly, Gly-Ala, Gly-Ser, Gly-Pro, Gly-Asp, Gly-Glu, Gly-Lys, Gly-Leu, Gly-Thr, and Gly-Val ranked by their occurrence frequency. According to amino acid property, the amino acid glycine tends to contact with small or tiny amino acids such as Ala, Ser, Asp, Thr, and Pro. Focusing on cysteine pairs, we observe that Cys-Cys occurs in $\beta$-sheet and loop frequently. Moreover, the chemical component is defined as hetID in PDB; thus we totally extract about 6827 different hetIDs from PDB. The top-5 hetIDs are SO4, _CA, _ZN, _MG, and MSE.

4.3.6.2. Disulfide Bond

In Table 7, number of pairs and chemical component contacts are listed in both intra-molecular and inter-molecular for SSBOND and cysteine pair. We also measure min, max and average distance between two S$_\gamma$ atoms of SSBOND and cysteine pairs. In Figure 20, we also report distance distribution for SSBOND and cysteine pair. We collect 50627 SSBOND entries to analyze the connection between two amino acids of cysteine from PDB. In our discovered collection, we find the following problematic points: (1) extreme long bond length between two S$_\gamma$ atoms exists intra-molecularly or inter-molecularly (e.g. > 10 Å); (2) a residue in SSBOND would be a missing residue; (3) a residue in SSBOND would be heterogen, and most of them are modified residues. According to Protein Data Bank content guide, if S$_\gamma$ of cysteine is disordered then there are possible alternate linkages. PDB’s practice is to put together all possible SSBOND records. This is problematic because the alternate location identifier is not specified in the SSBOND record.
Table 7. Statistical result of SSBOND and Cysteine pair.

<table>
<thead>
<tr>
<th></th>
<th>Intra-molecular</th>
<th>Inter-molecular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSBOND</td>
<td>(A) 48152</td>
<td>2115</td>
<td>50267</td>
</tr>
<tr>
<td></td>
<td>(B) 3333</td>
<td>95</td>
<td>3429</td>
</tr>
<tr>
<td>Cysteine Pairs</td>
<td>(A) 114777</td>
<td>-</td>
<td>114777</td>
</tr>
<tr>
<td></td>
<td>(B) 12847</td>
<td>-</td>
<td>12847</td>
</tr>
</tbody>
</table>

(A) Number of pairs; (B) Chemical component contacts.

Residue-residue contacts of SSBOND and Cysteine pairs

Figure 20. Distribution between distance and its frequent.

In X-axis, for example, the annotation of 0-1 represents the measured distance is larger or equal to 0 Å and smaller than 1 Å. Most frequent distance between two SG falls in 2-3 Å.

4.3.7. Discussion

4.3.7.1. Difference of SSBOND and Cysteine Pairs

Based on disulfide bond analysis between SSBOND and residue-residue contact of cysteine pairs, the most frequent bond length between two cysteines ranges from 2 to 3 Å, and in general the disulfide bond length is around 2.8 Å. Due to disulfide bond conformed by two Sγ atoms, atom level analysis within a sphere is necessary rather than Cα atom. The problematic points we detected are minimum distance and
maximum distance between two $S\gamma$ atoms in SSBOND. The condition of zero distance between two $S\gamma$ atoms comes from the same coordinates of Cysteines annotated in SSBOND. Furthermore, we also find that the distance between two $S\gamma$ atoms larger than 10 Å (e.g. 149.663 Å in intra-molecular of 1RHG:C 64 and 74, and 77.881 Å in inter-molecular of 1UMR:B 135 and 1UMR:D 203), and it’s might be incorrect annotation of bond connectivity. In addition, because the size of chemical component will affect the result, we only select large-size chemical component for structure similarity evaluation. In whole residue environmental spheres containing SSBOND, we have 16 residue environmental spheres containing BEN and 66 residue environmental spheres containing FAD. We find that residue environmental spheres of SSBOND surrounded by two chemical components of FAD and BEN respectively have highly conserved region of spheres. In Figure 21, atoms in yellow are $S\gamma$ atoms and the chemical component in CPK mode is FAD. Unlike previous researches, we try to index whole PDB dataset to analysis all residue-residue contacts and bond connectivity inside a protein structure while previous researches focus on only analyze special pair preference and residue frequencies [55].

![Figure 21. Disulfide bond and ligand.](image)

Atoms in yellow are $S\gamma$ atoms that build the disulfide bond annotated as SSBOND in PDB ID 1BHY and the chemical component is FAD (FLAVIN-ADENINE DINUCLEOTIDE) in CPK mode.

### 4.3.7.2. File Parsing and Efficiency of Database Query

To extract atom coordinate information, we have to parse pdb file to obtain structure information, but file parsing is the worst way for data mining because of information reusable. Besides, the use of sphere can gain the advantage of neighborhood information. Thus, for the purpose of structure mining on PDB, we try to simplify
our mining procedure, and then we parse PDB raw files, index whole protein structures with residue environmental sphere and deposit all information into database. To avoid database connection I/O, we use database dump technique to prepare dump file for database restore instead of row-by-row insertion. Comparing with file parsing and database query, without consideration of preprocessing, we spend about 1 hour to select sphere information from database for detect residue-residue contact of cysteine pairs via database query while spending 17 hours via file parsing without database utilization. Therefore, we can gain more benefit from indexing mechanism and database query.

4.3.8. Conclusions

In summary, sphere-based neighborhood searching is an appropriate local structure representation for structure mining on PDB. Consequently, we obtain huge scale collection of residue environmental sphere for describing protein local environment based on the believed assumption of protein function interacted with local structure. In order to searching and mining among this collection, indexing mechanism is very important; therefore, the residue environmental sphere is local structure representation and indexing unit for the reason of information reuse. Focusing on disulfide bond, the observation can be put on both SSBOND and cysteine pairs. Although there is some problematic information in SSBOND, they still provide useful information to compare with SSBOND and cysteine pair. In the future, further analysis on different residue-residue contacts and discussion on structural property should be scrutinized.
5. **SUMMARIZATION**

5.1. **Protein Local Structure Representation**

In this study, we try to find out relationships between local conservations and functional area via mining frequent itemset. Thus our first step is to discuss protein local structure representation. Neighborhood residue sphere is a well-organized representation because a key issue of force field will be considered in a sphere. The sphere also has flexibility to adjust and could be encoded into a binary encoding. In order to link mined local structure with protein, we use EC family to verify our purpose because of the ease of substrate/ligand verification. Therefore, we can use ligand contact to explain what we discovered. In our experiments, conserved local structure can be discovered and the observations show contact areas but not all elements of substrate contact with a substructure. We can discover conserved local structure region from functional hierarchical classification because proteins have the same function will share some attributes reflect on their structures.

5.2. **Protein Structure Conservation Mining**

According to our previous study on local structure representation, we adapt sphere model to describe local structure, which contains both sequence and structure information of local region. Our experimental results reveal that conserved patterns discovered from protein chains with the same EC labels share high conservation in local structure and that conserved patterns have a high capacity to be identified. In EC family prediction, we find that we have many “incorrect” and “no” predictions and the reasons are threshold setting, and the relationship of sequence-structure consistency.

In enzyme classification prediction, the experimental results show that the coverage rate of a template is correlated to the confidence level of classification prediction. In addition, predefined classification is very important for prediction; thus, in enzyme classification, we find that functional classification is significantly beneficial to mine conserved patterns significantly to identify EC label. Comparing with CSA and Protemot, our approach tries to apply the concept of “mining frequent itemset” to identify conserved region for recognizing EC family without using protein-ligand
complexes.

The critical issue and also difficulty we meet is similarity either sequence similarity or structure similarity. Proteins with the same EC label mean that they have the same function or biochemical reaction. While evaluating sequence identity of proteins from enzyme classification, the observation we found is that sequences share higher sequence identity within the same EC label (~80%). Higher sequence identity also implies similar protein structure. Therefore, mining frequent itemset will suffer from this difficulty. Resolution of protein structure determination is another issue should be addressed. Different level of resolution gives us different quality of a protein structure. If we want to obtain precise information from protein structure, we can select protein structures with resolution lower than 3.0 Å.

5.3. Protein Structural Property Exploration

As we know, interactions between residues will reflect on protein structure when protein folds. Therefore, we attempt to understand contact preference of residue interactions. In order to explore protein structural property, we use sphere model, residue environmental sphere, to describe environment surrounding a residue. Residue environmental sphere has its own advantage of space neighbor residue identification. Protein structural property we mentioned in this dissertation will be defined as contact preference, residue environment, interaction preference, etc. For the purpose of identifying structure neighbor residues in protein structure, if we don’t have a well-organized representation, we have to parse structural data repeatedly. It would be a huge scale collection of local structure information if we decompose protein structure into spheres. In order to searching and mining among this collection, indexing mechanism is very important; therefore, the residue environmental sphere is local structure representation and indexing unit for the reason of information reuse.
6. ONGOING STATUS

6.1. Structural Data Information Analysis

Accompanying with the growth of structural data, PDB updates structural data information frequently. In addition, content guide for file format illustration has also been updated twice after 2006. As reported in section 4.3.7.1, there are some problematic annotations of SSBOND in PDB according to our observation on residue environment analysis. For instance of PDB 1UMR, while comparing previous version PDB with current released version, we find that the PDB corrected some problematic annotation. In Figure 22, we show the difference between current released version and previous version. Moreover, resolution of examined protein structure is also a critical point should be considered in protein structure determination. In another word, resolution means the quality of protein structure. Therefore, data preprocessing based on resolution is necessary for residue environment analysis and conservation mining.

![Figure 22. Comparison of latest version and previous version of 1UMR.](image)
6.2. Protein Structure Conservation Mining base on Sequence-Structure Correlation

According to experimental results, conserved local structure can be discovered via mining frequent itemset on a group of proteins sharing the same function from hierarchical functional classification. This approach will meet the problem of higher sequence identify and structure similarity because of checking sequence-structure consistency. If a group of proteins share higher sequence and structure similarity, the mining results will be redundant. Hence, we have to discover the correlation between sequence and structure from global and local of views.

6.3. Structure-based Mining Approach for Structure Conservation Discovery

According to experiences on mining conserved local structure based on sphere model, we choose alternative to discover structure conservation via structure-based mining. The reason to apply purely structure approach is that protein function is more conserved in structure than in sequence. Therefore, geometric matching with sequence constraint will be considered in our proposed approach. Based on sphere model, we encode there-dimensional space information into one-dimensional binary signature. In Figure 23, it is diagram to illustrate the encoding scheme. Indexing and hashing techniques will be also applied for distinguishing different kinds of space patterns. In order to evaluate meaningful structure conservation, we apply this approach on functional group of proteins, and enzyme classification is our first choice. In addition, whole protein structures in PDB will be took into account.
6.4. Protein Structural Property Exploration of Interaction Region

Since 1992, researchers had been investigated on structural analysis of interaction region of residue-residue [54], protein-protein [29], protein-RNA [30], and protein-DNA [31]. The essential issues for protein folding and protein function is to discuss how amino acids interact with amino acids, base pairs, or ions. The contact preference would be the topic for this essential issue. Structural analysis is used to help us to understand why proteins are folded and how protein functions are activated in specific environments. In Figure 24 and Figure 25, there are examples of residue-residue contact and protein-ligand contact. From Figure 26 to Figure 28, there are examples of interaction regions of protein-protein, protein-RNA, and protein-DNA respectively.

On structural analysis of chemical bond connectivity, disulfide bonding (or disulfide bridge) is an interesting case that disulfide bond is formed by two cysteines via an attraction of two $S_\gamma$ atoms. The disulfide bond plays the role to stabilize protein structure in both protein tertiary structure and protein quaternary structure. In Figure 29 and Figure 30, there are examples of intermolecular disulfide bond and intramolecular disulfide bond respectively. Therefore, residue environment analysis
surrounding disulfide bond is another issue to discuss the role cysteine plays in the interaction region.

Figure 24. Residue-residue contacts.

Figure 25. Protein-ligand contact.
Figure 26. Protein-protein interaction region.

Figure 27. Protein-RNA interaction region.
Figure 28. Protein-DNA interaction region.

Figure 29. Intermolecular disulfide bond.
6.5. Summary

Based on the study of local structure conservation and residue environment analysis, we know that protein structure provide more clues to represent protein function. Through local structure conservation mining, we can discover the relationship between sequence, structure, and function. The protein-ligand complexes help us to distinguish structural conserved structures and functional conserved structures although it is not significant. This is the first step to understand correlation of sequence, structure, and protein function from the view of local structure. Furthermore, global similarity and local similarity of protein sequence and protein structure is a key to comprehend the relation of sequence, local structure, and function.

Protein structure is a complicated model in living cell because it consists of the knowledge of biology, chemistry, physics, etc. Protein structure determination is the problem of protein folding, and protein folding reflects the relation between residues in three-dimensional space. In residue environment analysis, we try to summarize conservation information inside protein structure. The conservation information in residue environment analysis would be contact preference of residue-residue,
residue-ligand, and residue-nucleic base pair, environment preference of bond connectivity, interaction preference, and so on.
REFERENCES


