# 國立交通大學

## 電機與控制工程學系

## 碩士論文

# 基於類神經網路之
# 蛋白質金屬鍵結胺基酸預測

# Protein Metal-Binding Residue Prediction Based on Neural Networks

研 究 生：楊志賢

指導教授：王啟旭 博士

林進燈 博士

楊裕雄 博士

中華民國 九十三 年 六 月

# 基於類神經網路之蛋白質金屬鍵結胺基酸預測

# Protein Metal-Binding Residue Prediction Based on Neural Networks

| | |
|---|---|
| 研究生：楊志賢 | Student：Chih-Hsien Yang |
| 指導教授：王啟旭 博士 | Advisor：Dr. Chi-Hsu Wang |
| 林進燈 博士 | Co-Advisor：Dr. Chin-Teng Lin |
| 楊裕雄 博士 | Co-Advisor：Dr. Yuh-Shyong Yang |

國立交通大學

電機與控制工程學系

碩士論文

A Thesis
Submitted to Department of Electrical and Control Engineering
College of Engineering and Computer Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of Master
in
Electrical and Control Engineering
June 2004
Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

# 基於類神經網路之

# 蛋白質金屬鍵結胺基酸預測

學生：楊志賢

指導教授：王啟旭 博士

共同指導教授：林進燈 博士

共同指導教授：楊裕雄 博士

國立交通大學電機與控制工程研究所

## 摘要

傳統上，結構生物學家利用物理實驗的方式去了解金屬蛋白 (與金屬產生鍵結的蛋白質) 的特性，並藉由其立體結構與實驗上的觀察去推論酵素產生功能的原因及其反應機制。然而，大多數的蛋白質都已知其一級結構 (胺基酸序列組成) ，立體結構資訊卻不如序列資訊來的普遍。再者，目前由序列資訊預測蛋白質立體結構的技術，也仍尚未達到絕對可靠的階段。

因此，本論文中提出，純粹以蛋白質序列的資訊，使用類神經網路為主要核心的預測器，搭配滑動框架式特徵擷取以及生物化學式特徵編碼的技術，對蛋白質金屬鍵結胺基酸進行預測。針對生命系統中的四種主要金屬 (鈣、鉀、鎂與鈉) 鍵結蛋白質中，在五等份的交叉驗證中，均可達到九成以上的鍵結偵測敏感度且兼具極佳的準確度。

**關鍵字**：生物資訊學，類神經網路，金屬蛋白，酵素。

# Protein Metal-Binding Residue Prediction Based on Neural Networks

Student: Chih-Hsien Yang Advisor: Dr. Chi-Hsu Wang
Co-Advisor: Dr. Chin-Teng Lin
Co-Advisor: Dr. Yuh-Shyong Yang

Department of Electrical and Control Engineering
National Chiao Tung University

## Abstract

Traditionally, structural biologists used to investigate properties of metalloproteins (proteins which bind with metal ions) by physical means and interpret the function formation and reaction mechanism of enzyme by their structures and observation from experiments in vitro. Most of proteins have primary structures (amino acid sequence information) only; however, the 3-dimension structures are not always available. Moreover, the prediction from protein sequence to structure is still not completely reliable so far.

Consequently, a direct analysis method is proposed to predict protein metal-binding amino acid residues only from its sequence information by neural network with sliding window-based feature extraction and biochemical feature encoding techniques in this thesis. In four major bulk elements (Calcium, Potassium, Magnesium, and Sodium) in life system, the metal-binding residues are identified with a binding sensitivity > 90% and nearly 100% accuracy under five fold cross validation.

**KEYWORD**：Bioinformatics, Neural Networks, Metalloprotein, Enzyme

# 致謝

獻給所有陪我一路走來的人。

# Acknowledgements

To these who go along with me through this way.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Introduction to Bioinformatics[1]

With rapid growth in computer and information science in recent years, most things in daily life have changed the way they were including biology — the study of living things. From last several decades, biologists have collected and accumulated data from interaction of spices and populations, the function of tissues and cells within an individual organism, and even the structure and function of molecules (such as protein, DNA, RNA, etc.) inside or outside the cell. Sophisticated laboratory technology today helps biologists collect data faster, but it can't speed up the interpretation of these massive and divergent biological data.

For instance, we have huge volume of human DNA sequences after Human Genome Project (HGP)[2], but how do we know which parts of DNA sequence can control which kinds of chemical processes or reactions in human body (Gene annotation or labeling)? We have many outstanding structural biologists spent great effort on determining protein structures by Nuclear Magnetic Resonance (NMR) or X-ray crystallography, and figuring out the structure of some proteins, but how do we determine the structure and function of other proteins and even a whole new protein (protein structure prediction and function analysis)? **Table 1.1.a** show the exponential data growth in GenBank[3] and Protein Data Bank[4] respectively.

Consequently, it is necessary for biologists to use current computational and

[1] Use of computers in biological research, such as the use computerized databases for genomes, proteins, etc.

[2] It is an international research effort to identify sequence and map all genes in human DNA. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[3] GenBank is the NIH (National Institutes of Health, http://www.nih.gov/) genetic sequence database, an annotated collection of all publicly available DNA sequences. http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

[4] It is a single worldwide repository for the processing and distribution of 3-D biological macro-molecular structure data. http://www.rcsb.org/pdb/

internet technologies to help them store, share and analyze their biological data on computer or world-wide-web instead of their hands and eyes so as to yield "*high throughput*" biology, and accelerate the discovery in life science and development of biomedical products, such as drug, and therapy for cancers or other currently unsolvable diseases.

But unfortunately, bioinformatics have become a buzzword with the hype about mapping the human genome. All of these wonderful dreams should be based on every small pieces of understanding about the whole organism from top to toe, and not based on the exaggeration of newspaper or people's short passion and day dream.

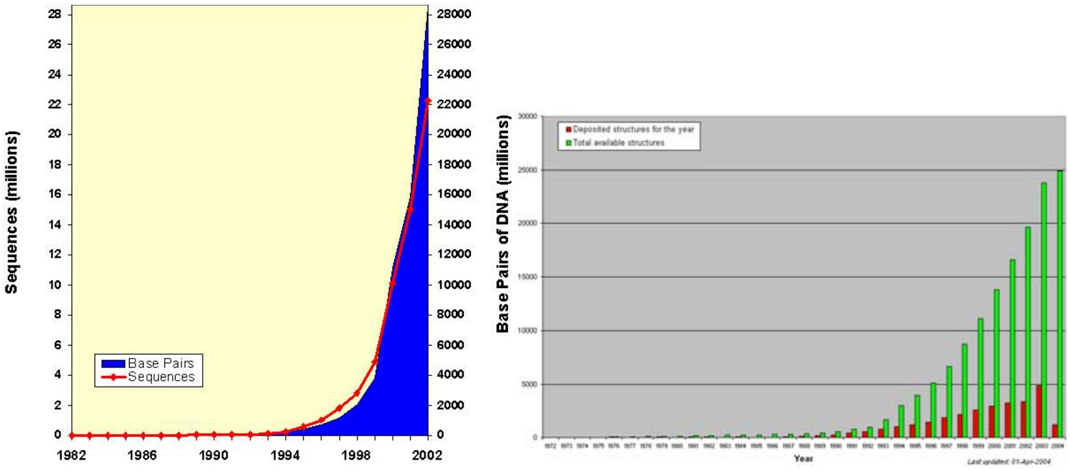


**Table 1.1.a** Growth in GenBank (left) and Protein Data Bank (right)

## 1.2 Metalloprotein and Motivation

Metalloproteins are proteins capable of binding one or more metal ions, which are required for their biological function or for regulation of their activities or even for structure purposes. It is very interesting and amazing that more than one-quarter of the elements in periodic table are required for life (as shown in **Figure 2.3.b**) , and most of them are metal ions.

According to [5]PIR, the release 78.03 contains 283,336 entries in November 24, 2003. In contrast, in Protein Data Bank there are 24,358 structures are available in February 17, 2004. Transparently, the sequence material is greatly richer than the structure in proteins by 10 times or more. As this result, if a direct prediction method only based on sequences is practical, it will be very helpful in current status. The objective of this thesis is to build metal-binding model for protein by computer-based machine learning method so that it can be a reliable metal-binding residues predictor for proteins without actual coordinate information, and further be used to investigate and understand the formation of biochemical function of metalloprotein, and eventually offer "functional templates" as a guideline to design new protein with specified function.

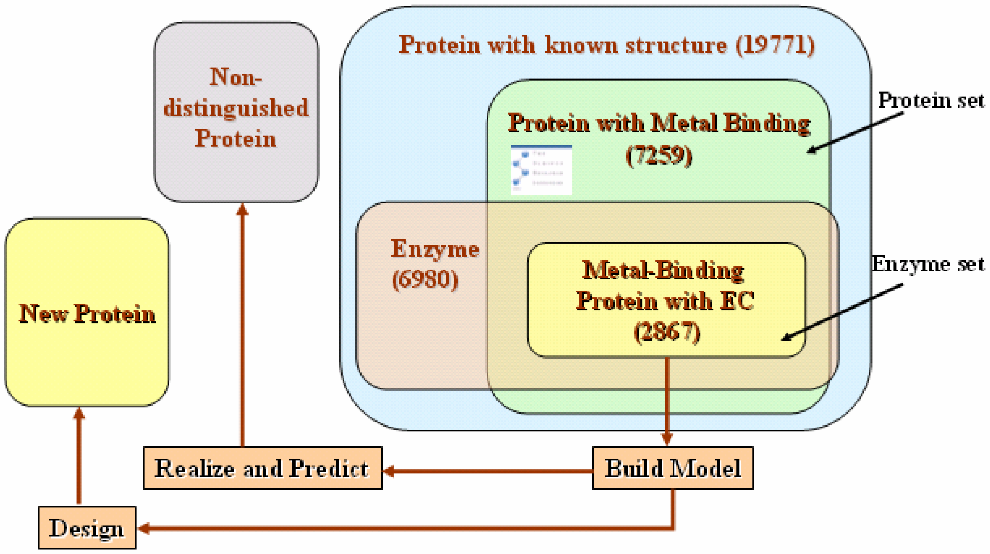

Protein with known structure (19771)
Non-distinguished Protein
Protein with Metal Binding (7259)
Protein set
Enzyme (6980)
Metal-Binding Protein with EC (2867)
Enzyme set
New Protein
Realize and Predict
Build Model
Design

**Figure 1.2.a** Overview of data resource and working map

[5] Protein Information Resource, an integrated public resource of protein informatics to support genomic and proteomic research and scientific discovery. http://pir.georgetown.edu/

In this thesis, two major data sets (protein set and enzyme set, as shown in **Figure 1.2.a**) are extracted from 19771 protein structures in PDB and all experiments are based on enzyme set. There are 7529 protein molecules with metal binding and 6890 protein molecules with EC number[6]. Besides, there are over one-third (36.72%) proteins containing metal ions in 19771 protein structures and nearly 40% of them are enzymes (**Figure 1.2.b and 1.2.c**).

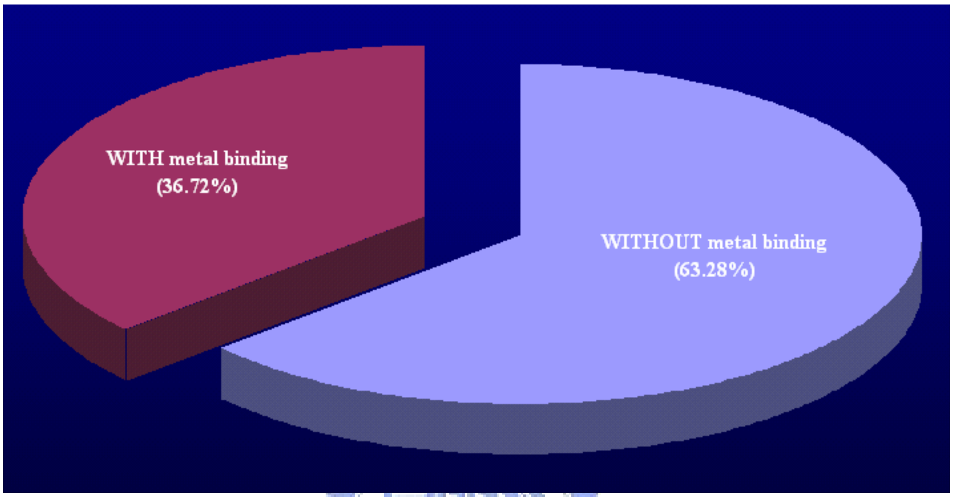


**Fig 1.2.b** Metal-binding proteins in 19771 proteins

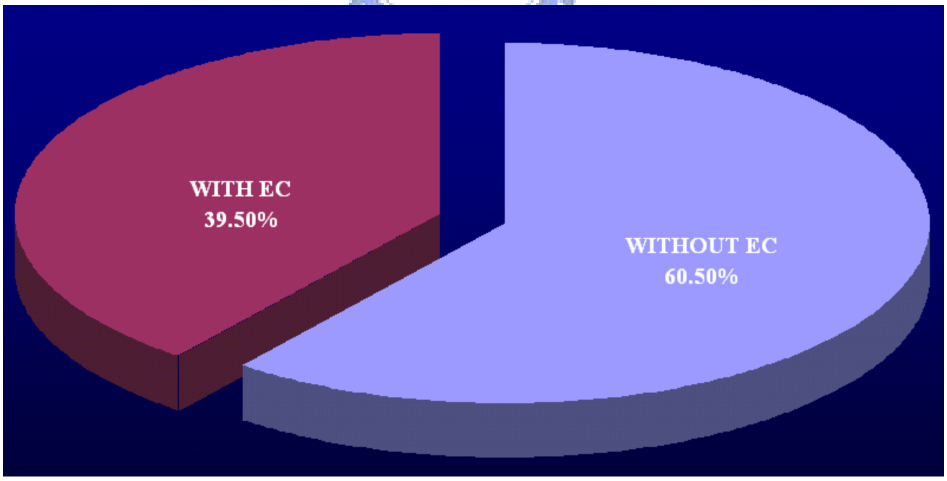
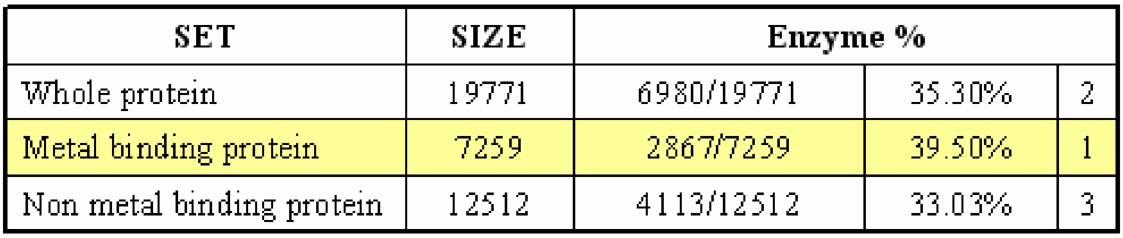


**Fig 1.2.c** Enzymes in 7250 metal binding proteins

In **Table 1.2.a** and **Figure 1.2.d** show the comparison of enzyme percentage

[6] Enzyme Commission number, a nomenclature for enzymes, developed by The International Union of Biochemistry and Molecular Biology, is described by a sequence of four numbers, preceded by "EC" in the form of "EC X.X.X.X."

between metal-binding protein set, non metal-binding protein set and entire protein set, a protein with metal binding is more likely to be an enzyme.



| SET | SIZE | Enzyme % | | |
|---|---|---|---|---|
| Whole protein | 19771 | 6980/19771 | 35.30% | 2 |
| Metal binding protein | 7259 | 2867/7259 | 39.50% | 1 |
| Non metal binding protein | 12512 | 4113/12512 | 33.03% | 3 |

**Table 1.2.a** Tabular comparison of enzyme percentage in different sets

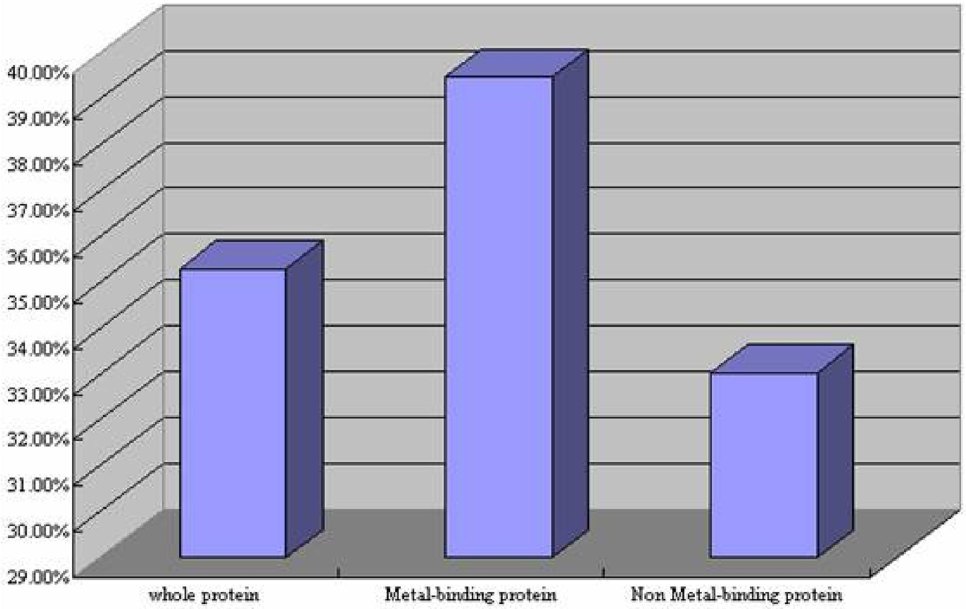


**Fig 1.2.d** Histogram of enzyme percentage comparison

An enzyme, in biology, is a special protein molecule whose function is to facilitate or accelerate most chemical reactions in cells. Many chemical reactions occur within biological cells, but most of them happen too slowly without catalysts in vitro (test tube) to be biologically relevant.

By common convention, an enzyme’s name is a description of what is does, with the word ending “-ase” added, such as alcohol dehydrogen**ase** and DNA polymer**ase**. The International Union of Biochemistry and Molecular Biology has developed a nomenclature for enzymes, the EC numbers; each enzyme is described by a sequence of four numbers, preceded by “EC” in the form of “EC X.X.X.X.” The first number broadly classified the enzyme based on its mechanism as below：

| EC | Classification | Function and Mechanism |
|---|---|---|
| 1.X.X.X | Oxidoreductases | catalyze oxidation (any electrochemical process which involves the formal oxidation state of an atom or atoms within a molecule being increased by removal of electrons) /reduction reactions |
| 2.X.X.X | Transferases | transfer a function group (in organic chemistry, it is a sub-molecular structural motif such as a methyl or phosphate group, characterized by specific composition and connectivity, that confer reactivity upon molecule that contains them) |
| 3.X.X.X | Hydrolases | catalyze the hydrolysis (a chemical process in which a molecule is cleaved into tow parts by insertion of a molecule of water) of various bonds |
| 4.X.X.X | Lyases | cleave various bonds by means other than hydrolysis and oxidation |
| 5.X.X.X | Isomerases | catalyze isomerization changes (conversion into another molecules with the same molecular formula but different arrangement of atoms, called isomers) within a single molecule |
| 6.X.X.X | Ligases | join two molecules with covalent bonds |

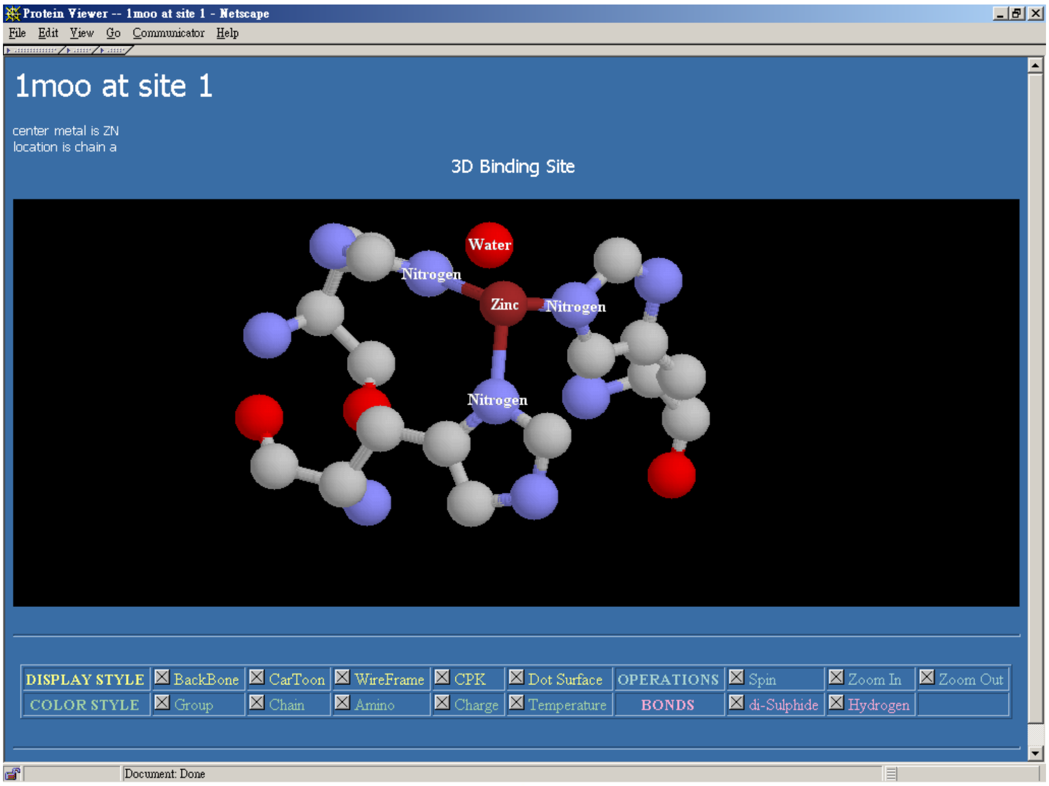
**Table 1.2.b** Major enzyme classes and functions

Enzymes are essential for the function of cells and are very specific as to the reactions they catalyze and the chemicals (substrates) that involved in the reactions. Substrates fit their enzymes like a key fits its lock. Many enzymes are composed of several proteins that act together as a unit. Most parts of an enzyme have regulatory and structural purposes. The catalyzed reaction takes place in only a small part of the enzyme called active site. Many enzymes incorporate metal divalent cations and transition metal ions within their structures to stabilize the folded conformation of protein or to directly participate in the chemical reactions catalyzed by the enzyme. **Figure 1.2.c** shows the major enzyme class distribution in metalloprotein in the thesis.

Metal also provides a template for protein folding, as in the zinc finger domain of nucleic acid binding proteins, the calcium ions of calmodulin (a protein molecule that is necessary for many biochemical process, including muscle contraction and the release of a chemical that carries nerve signals), and the zinc structural center of insulin. Metal ions can also serve as redox centers for catalysis, such as heme-iron centers, copper ions and non-heme irons. Other metal ions can serve as electrophilic reactants in catalysis, as in the case of active site zinc ions of the metalloprotease. For example, the enzyme carbonic anhydrase (**Figure 1.2.e**) typically forms 4 coordinate bonds in a tetrahedral arrangement about its metal ion.

| **Biological level** | **Element** | **Enzyme Class Distribution** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **EC 1 Oxidoreductases** | **EC 2 Transferases** | **EC 3 Hydrolases** | **EC 4 Lyases** | **EC 5 Isomerases** | **EC 6 Ligases** |
| **bulk element** | Ca | 125 | 74 | 571 | 19 | 4 | 1 |
| | Mg | 38 | 183 | 133 | 45 | 15 | 44 |
| | Na | 16 | 102 | 257 | 29 | 22 | 1 |
| | K | 27 | 32 | 24 | 18 | 0 | 11 |
| **trace element** | Zn | 101 | 87 | 312 | 125 | 12 | 19 |
| | Fe | 415 | 6 | 15 | 12 | 1 | 0 |
| | Mn | 39 | 65 | 41 | 6 | 12 | 14 |
| | Se | 34 | 33 | 36 | 17 | 9 | 5 |
| | Cu | 95 | 1 | 8 | 4 | 0 | 0 |
| | Co | 6 | 10 | 31 | 10 | 14 | 0 |
| | Ni | 10 | 2 | 35 | 3 | 2 | 0 |
| | I | 10 | 7 | 18 | 1 | 2 | 0 |
| | Mo | 20 | 0 | 1 | 0 | 0 | 0 |
| | V | 9 | 2 | 4 | 0 | 1 | 0 |
| | Cr | 0 | 4 | 0 | 0 | 0 | 0 |
| **possible trace element** | As | 26 | 7 | 9 | 1 | 1 | 0 |
| **N/A** | Hg | 8 | 7 | 16 | 55 | 0 | 0 |
| | Cd | 10 | 12 | 26 | 1 | 2 | 0 |
| | Al | 1 | 5 | 9 | 0 | 1 | 1 |
| | U | 3 | 2 | 2 | 0 | 0 | 1 |
| | Cs | 0 | 2 | 1 | 2 | 1 | 1 |
| | Pb | 0 | 2 | 1 | 4 | 0 | 0 |
| | Pt | 1 | 1 | 2 | 1 | 0 | 0 |
| | Tl | 0 | 1 | 3 | 0 | 0 | 1 |
| | Be | 0 | 2 | 2 | 0 | 0 | 0 |
| | Sm | 1 | 1 | 1 | 1 | 0 | 0 |
| | Sr | 0 | 2 | 2 | 0 | 0 | 0 |
| | W | 0 | 1 | 3 | 0 | 0 | 0 |
| | Yb | 0 | 1 | 2 | 1 | 0 | 0 |
| | Au | 0 | 0 | 2 | 0 | 1 | 0 |
| | Ho | 0 | 2 | 1 | 0 | 0 | 0 |
| | Ba | 0 | 2 | 0 | 0 | 0 | 0 |
| | Ag | 0 | 0 | 1 | 0 | 0 | 0 |
| | Eu | 0 | 1 | 0 | 0 | 0 | 0 |
| | Gd | 0 | 0 | 1 | 0 | 0 | 0 |
| | In | 0 | 0 | 0 | 1 | 0 | 0 |
| | Li | 0 | 0 | 0 | 1 | 0 | 0 |
| | Rb | 0 | 0 | 1 | 0 | 0 | 0 |
| | Te | 1 | 0 | 0 | 0 | 0 | 0 |
| | La | 0 | 0 | 0 | 0 | 0 | 0 |
| | Tb | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 1.2.c** enzyme distribution in metalloprotein



1moo at site 1

center metal is ZN
location is chain a

3D Binding Site

| DISPLAY STYLE | ☒ BackBone | ☒ CarToon | ☒ WireFrame | ☒ CPK | ☒ Dot Surface | OPERATIONS | ☒ Spin | ☒ Zoom In | ☒ Zoom Out |
|---|---|---|---|---|---|---|---|---|---|
| COLOR STYLE | ☒ Group | ☒ Chain | ☒ Amino | ☒ Charge | ☒ Temperature | BONDS | ☒ di-Sulphide | ☒ Hydrogen | |

**Fig 1.2.e** 3D metal-binding structure of carbonic anhydrase II

## 1.3 Thesis Organization

The organization of this thesis is described as follows. First, chapter 2 is involved in the biological resource building and sequence data processing. Chapter 3 is concerned about the algorithms on metal binding residue prediction and Chapter 4 is the results and discussion about prediction. Appendix is a simple manual for bioinformatics tool or protocol in the thesis.

# Chapter 2

# Biological Resource

In this chapter, there are three issues are concerned. Section 2.1 shows how to obtain proper biological reference for demand. Further, section 2.2 gives an example to design a self-defined database for target data storage fitted to actual protein metal-binding model. Finally, section 2.3 shows the biological data processing and sampling.

## 2.1 Integration of Web Biological Databases

The main data resources come from two web sites; one is the metalloprotein database and browser (MDB) of metalloprotein structure and design program of the Scripps Research Institute (http://metallo.scripps.edu). Another one is Protein Data Bank (PDB, http://www.rcsb.org/pdb/), which provides general information about every protein structure. Hence, by combing these two data sets, the detail description of metalloprotein can be driven. For simplicity, the PDB information can be replaced by another compacted data － PDBFinder (http://www.cmbi.kun.nl/gv/pdbfinder/) released at September, 14, 2003.

In MDB, all proteins with binding metal can be entirely extracted and the binding site is also defined by nearby amino acid residues and compounds in order to develop a sufficient understanding about metalloproteins. To achieve this, it is needed to comprehend the set of structural, environmental, and functional requirements for metal-binding sites in existing metalloprotein or metalloenzymes. In structure, MDB has catalogued several important issues, such as what types of (metal) ions are bound to protein molecule, what types of ligands that bind these metal ions (i.e. the first-shell ligands), and what residues that contact the metal-binding ligands (i.e. the second-shell ligands) as illustrated in **Figure 2.1.a**.

As the result, there are three tables (As shown in **Table 2.1.a** to **Table 2.1.c**) — Protein Table, Site Table and Ligand Table needed to describe the structural relationship between protein and metal binding site in nature. That is to say, the objective can be translated as discovering the metal-binding second-shell ligands (residues) from protein primary structure (protein amino acid sequence).

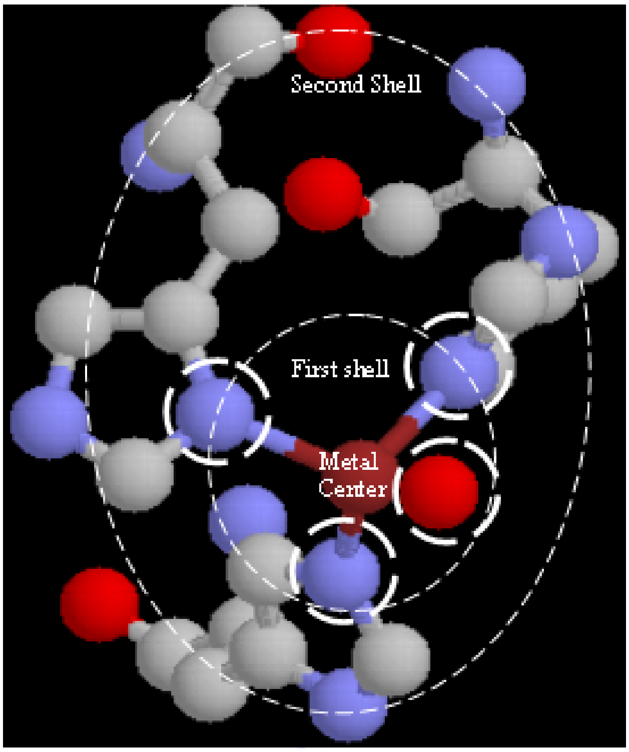
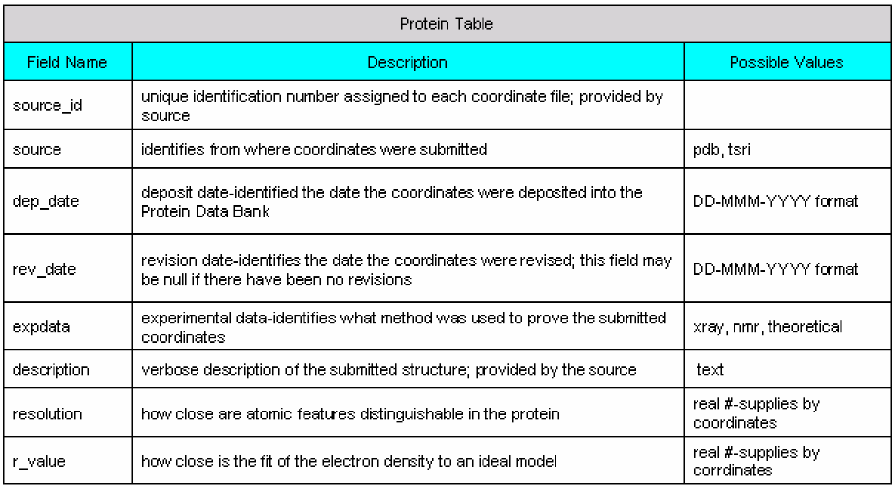
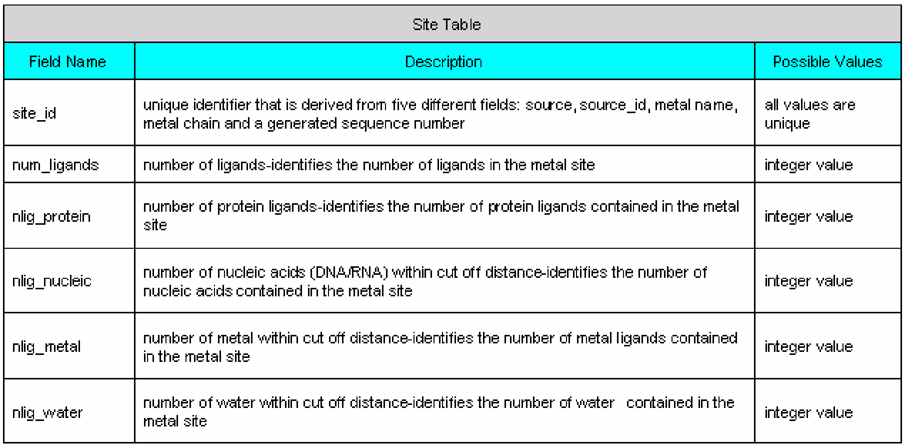


**Fig 2.1.a** Metal-Ligand diagram in metal-binding protein



| Protein Table | | |
|---|---|---|
| Field Name | Description | Possible Values |
| source_id | unique identification number assigned to each coordinate file; provided by source | |
| source | identifies from where coordinates were submitted | pdb, tsri |
| dep_date | deposit date-identified the date the coordinates were deposited into the Protein Data Bank | DD-MMM-YYYY format |
| rev_date | revision date-identifies the date the coordinates were revised; this field may be null if there have been no revisions | DD-MMM-YYYY format |
| expdata | experimental data-identifies what method was used to prove the submitted coordinates | xray, nmr, theoretical |
| description | verbose description of the submitted structure; provided by the source | text |
| resolution | how close are atomic features distinguishable in the protein | real #-supplies by coordinates |
| r_value | how close is the fit of the electron density to an ideal model | real #-supplies by corrdinates |

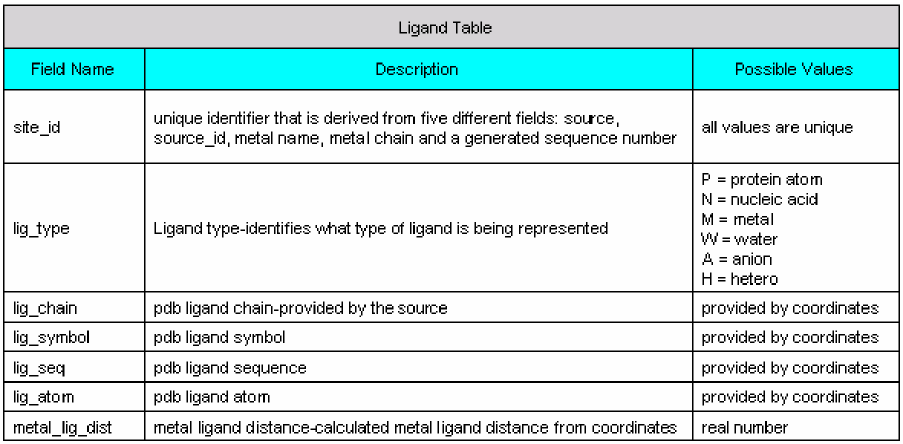
**Table 2.1.a** Protein Table in MDB



Site Table

| Field Name | Description | Possible Values |
|---|---|---|
| site_id | unique identifier that is derived from five different fields: source, source_id, metal name, metal chain and a generated sequence number | all values are unique |
| num_ligands | number of ligands-identifies the number of ligands in the metal site | integer value |
| nlig_protein | number of protein ligands-identifies the number of protein ligands contained in the metal site | integer value |
| nlig_nucleic | number of nucleic acids (DNA/RNA) within cut off distance-identifies the number of nucleic acids contained in the metal site | integer value |
| nlig_metal | number of metal within cut off distance-identifies the number of metal ligands contained in the metal site | integer value |
| nlig_water | number of water within cut off distance-identifies the number of water  contained in the metal site | integer value |

**Table 2.1.b** Site Table in MDB

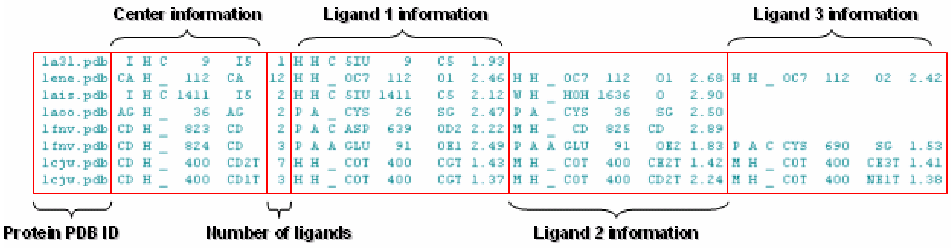


Ligand Table

| Field Name | Description | Possible Values |
|---|---|---|
| site_id | unique identifier that is derived from five different fields: source, source_id, metal name, metal chain and a generated sequence number | all values are unique |
| lig_type | Ligand type-identifies what type of ligand is being represented | P = protein atom<br>N = nucleic acid<br>M = metal<br>W = water<br>A = anion<br>H = hetero |
| lig_chain | pdb ligand chain-provided by the source | provided by coordinates |
| lig_symbol | pdb ligand symbol | provided by coordinates |
| lig_seq | pdb ligand sequence | provided by coordinates |
| lig_atom | pdb ligand atom | provided by coordinates |
| metal_lig_dist | metal ligand distance-calculated metal ligand distance from coordinates | real number |

**Table 2.1.c** Ligand Table in MDB

But these databases are not directly released to public. Only available information is formatted into 43 ligand text files with respect to 43 kind different binding metals. The latest version of file package is “18” and is updated at January, 17, 2003. The following **Figure 2.1.b** shows the format of ligand file.



```
1a31.pdb  I H C    9   I5   1 H H C 5IU    9   C5 1.93
1ene.pdb CA H _  112   CA  12 H H _ OC7  112   O1 2.46 H H _ OC7  112   O1  2.68 H H _ OC7  112   O2  2.42
1ais.pdb  I H C 1411   I5   2 H H C 5IU 1411   C5 2.12 W H _ HOH 1636   O   2.90
1aoo.pdb AG H _   36   AG   2 P A _ CYS   26   SG 2.47 P A _ CYS   36   SG  2.50
1fnv.pdb CD H _  823   CD   2 P A C ASP  639  OD2 2.22 M H _  CD  825  CD   2.89
1fnv.pdb CD H _  824   CD   3 P A A GLU   91  OE1 2.49 P A A GLU   91  OE2 1.83 P A C CYS  690   SG  1.53
1cjw.pdb CD H _  400 CD2T   7 H H _ COT  400  CGT 1.43 M H _ COT  400 CE2T 1.42 M H _ COT  400 CE3T 1.41
1cjw.pdb CD H _  400 CD1T   3 H H _ COT  400  CGT 1.37 M H _ COT  400 CD2T 2.24 M H _ COT  400 NE1T 1.38
```

**Fig 2.1.b** Format of ligand file in MDB package

Each line in file represents one binding site surrounded by one center atom in protein. The file format of each line in ligand file can be expressed as

$$[\textit{protein information}] + [\textit{center information}] + [\textit{number of ligands } (N)] + \sum_{i=1}^{N} [\textit{ligand information}]$$

The term protein information is the file name of PDB file, noted as unique PDB ID tailed with ".pdb." The second term is the information about metal center which is one text with 5 fields — type of central atom, recognition type (A or H) of central atom, protein chain identifier where central atom located, residue series number of central atom located, and symbol of central atom. In recognition type, if it shows "A" then the central element is recognized as atom of standard residue; else if it shows "H" then it is recognized as atom of non-standard group in protein.

The third term is an integer number which indicates the number of binding ligands (assume to be N for illustration) with respect to this central atom. After this term, there are N ligands information follow. In ligand (binding atom) information, there are 7 fields involved — type of binding atom (P, N, M, W, A or H), recognition type of binding atom (A or H, the same as central atom recognition rule), protein chain identifier where binding atom located, residue name where binding atom located, residue series number of binding atom located, symbol of binding atom and distance (in angstrom) between central atom and binding atom. In binding atom type, if it shows "P", "N","M", "W", "A" or "H' then the type of binding atom is classified as atom of protein, atom of nucleic acid, metal atom, atom of water molecule, anion (negatively charged ion) or hetero atom. It is very useful when searching for metal binding residue (recognition type is 'A') in database.

In the similar way, PDB or PDBFinder database is released in the form of text files or accessed from html browser. For efficiency, it is necessary to build stand-alone database on local machine by parsing these released files. Therefore, each field in text file must be identified and clarified. In PDBFinder, there are three major levels of information about — (1) entire protein, (2) each one chain in protein and (3) hetero group of entire protein.

Level (1) **PROTEIN** provides several important messages about whether this protein is enzyme or not, the experimental details in determining this protein structure and statistics on total number of aligned sequences in HSSP (database of **H**omology-derived **S**econdary **S**tructure of **P**roteins), fraction of helix or beta sheet (major secondary structures), total umber of amino acid residues (standard and non-standard), total number of nucleic acids in protein, and total number of water molecules.

Level (2) **CHAIN** offers detailed description of each chain in protein, such as statistics about secondary structures (helix, 3/10 helix, pi helix, beta sheet, beta bridges, extended bridges, number of parallel and anti-parallel strand hydrogen bonds), amino acids (number of standard amino acids, number of non-standard amino acids, number of backbone-missing amino acids, number of sidechain-missing amino acids, number of only-Ca-given amino acid, number of unknown amino acids, number of Cystine residues, and number of chain-break which is larger than 4.5 angstrom) number of nucleic acids, number of enzyme substrate, number of water molecule, and primary structure sequence in this chain.

Level (3) **HET-Groups** show hetero group information as records in PDB file headed by HET. **Table 2.1.d**, **2.1.e** and **2.1.f** shows tree organization of records in level (1), (2), and (3) respectively. By combining these databases, the binding site of each protein can be extracted and it is possible to classify all proteins into enzyme or non-enzyme groups. **Figure 2.1.c** shows one example in PDBFinder released file. Because the details about PDB file format is too verbose so that it is described in appendix.

| Layer 1 | Layer 2 | Layer 3 | Location | Meaning/Note | Possible value/Data type |
|---|---|---|---|---|---|
| ID | | | | 4-letter PDB code | 4 letters |
| Header | Header | | PDB Header | PDB Header | Text |
| | Date | | PDB Header | Deposition date | Text |
| Compound | Compound | | PDB COMPND | PDB COMPND | Text |
| | Enzyme-code | | PDB COMPND | EC number | X.X.X.X, (X is one integer) |
| Source | | | PDB SOURCE | PDB SOURCE | Text |
| Author | | | PDB Author | Authors' names | Text |
| Exp_Method | Exp_Method | | PDB EXPDTA | Experiment methods | {X, NMR, FIBER, MODEL, NEUTRON, OTHER} |
| | Resolution | | PDB REMARK | X-ray only | Text |
| | R-factor | R-factor | REMARK3, 4 | X-ray only | Real number |
| | | Free-R | REMARK3 | X-ray only | Real number |
| | N-Models | | PDB MODEL | # of NMR models | Integer |
| | SF-Type | SF-Type | Structure factors file type | | PDB, CIF, Unknown |
| | | N-refl | # of reflections | | Integer |
| | | H-min, H-max | H index of reflection | | Integer |
| | | K-min, K-max | K index of reflection | | Integer |
| | | L-min, L-max | L index of reflection | | Integer |
| Ref-Prog | | | Names of refinement programs | | Text |
| HSSP-N-Align | | | # of aligned sequences in HSSP | | Integer |
| T-Frac-Helix, Beta | | | Total fraction of helix or beta | | Integer |
| T-Nres_Prot | T-Nres_Prot | | Total # of AA residues within the protein, including non standard | | Integer |
| | T-non-std | | Total # of non-standard residues | | Integer |
| T-Nres-Nucl | | | Total # of NA residues | | Integer |
| T-Water-Mols | | | Total # of water molecules | | Integer |

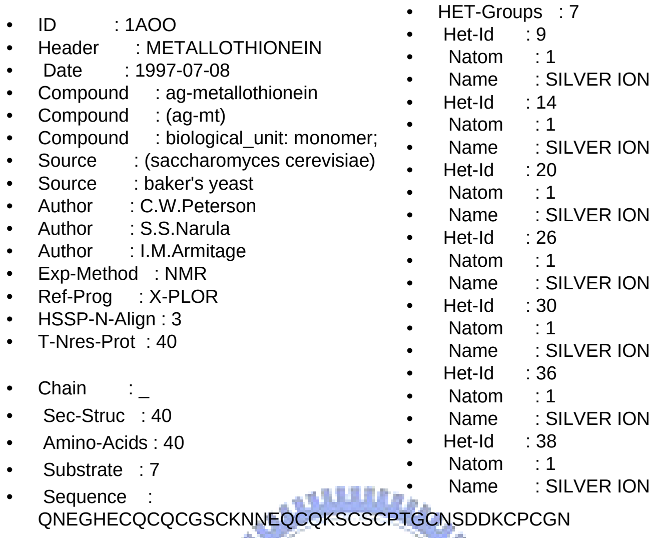
**Table 2.1.d** Record organization in level (1) PROTEIN

| Layer 1 | Layer 2 | Layer 3 | Layer 4 | Location | Meaning/Note | Possible value/Data type |
|---|---|---|---|---|---|---|
| Chain | Chain | | | | Protein polymer Chain ID | Text |
| | Sec-Struc | Sec-Struc | | From DSSP | # of residues which has SS | Integer |
| | | Helix | Helix | | # of residues which are helix | Integer |
| | | | i, i+3 | | # of residues which are 3/10 helix | Integer |
| | | | i, i+5 | | # of residues which are pi helix | Integer |
| | | Beta | Beta | | # of residues which are beta | Integer |
| | | | B-Bridge | | # of residues which are beta bridge | Integer |
| | | | E-beta | | # of residues which are extended bridge | Integer |
| | | | Para-Hb | | # of parallel strand Hydrogen bonds | Integer |
| | | | Anti-Hb | | # of anti-parallel strand Hydrogen bonds | Integer |
| | Amino-Acids | Amino-Acids | | | # of AA residues, including non-standard | Integer |
| | | Non-Std | | | # of non-standard AA | Integer |
| | | Miss-BB | | | # of backbone-missing AA | Integer |
| | | Miss-SC | | | # of sidechain-missing AA | Integer |
| | | Only-Ca | | | # of only-CA-given AA | Integer |
| | | UNK | | | # of unknown type AA | Integer |
| | | CYSS | | | # of Cys residues, it's about SS bond | Integer |
| | | Break | | | # of chain breaks (> 4.5 A) | Integer |
| | Nucl-Acids | | | | # of NA | Integer |
| | Substrate | | | | # of substrate atoms | Integer |
| | Water-Mols | | | | # of water molecules | Integer |
| | Sequence | | | | 1-letter code of AA or NA | Text |

**Table 2.1.e** Record organization in level (2) CHAIN

| Layer 1 | Layer 2 | Location | Meaning/Note | Possible value/Data type |
|---|---|---|---|---|
| HET-Groups | HET-Groups | PDB HET | # of HET groups | Integer |
| | Het-Id | | HET residue series number | Integer |
| | Natom | | # of atoms within each HET group | Integer |
| | Name | | Full name of HET group | Text |

**Table 2.1.f** Record organization in level (3) HET-Groups



- ID : 1AOO
- Header : METALLOTHIONEIN
- Date : 1997-07-08
- Compound : ag-metallothionein
- Compound : (ag-mt)
- Compound : biological_unit: monomer;
- Source : (saccharomyces cerevisiae)
- Source : baker's yeast
- Author : C.W.Peterson
- Author : S.S.Narula
- Author : I.M.Armitage
- Exp-Method : NMR
- Ref-Prog : X-PLOR
- HSSP-N-Align : 3
- T-Nres-Prot : 40

- Chain : _
- Sec-Struc : 40
- Amino-Acids : 40
- Substrate : 7
- Sequence : QNEGHECQCQCGSCKNNEQCQKSCSCPTGCNSDDKCPCGN

- HET-Groups : 7
- Het-Id : 9
- Natom : 1
- Name : SILVER ION
- Het-Id : 14
- Natom : 1
- Name : SILVER ION
- Het-Id : 20
- Natom : 1
- Name : SILVER ION
- Het-Id : 26
- Natom : 1
- Name : SILVER ION
- Het-Id : 30
- Natom : 1
- Name : SILVER ION
- Het-Id : 36
- Natom : 1
- Name : SILVER ION
- Het-Id : 38
- Natom : 1
- Name : SILVER ION

**Fig 2.1.c** an example of released text file in PDBFinder

## 2.2 Metal Binding Model and Database Design

Since in section 2.1 all fields in each released text file of target have been identified. Next step is to build a “container” for these biological data. **Figure 2.2.a** and **Figure 2.2.b** shows the DSD (Data Structure Definition) schematics of PDBFinder and MDB. Abstractly, the data hierarchy can be defined as 4 layers ordered by their size. They are PROTEIN, CHAIN, SITE and LIGAND. The top level PROTEIN may contain one or several chain (s), and each chain is represented as one polypeptide chain belonged to one protein in nature. Every site contains the coordinate information about entire metal center binding site, just like shown in **Fig 1.2**. The environment information describes about how many binding atoms (ligands) participate in the site, which residue the ligand located, and what these binding ligands are. The binding hierarchy model and ERD (Entity-Relationship Diagram) is shown in **Figure 2.2.c** and **Figure 2.2.d**.

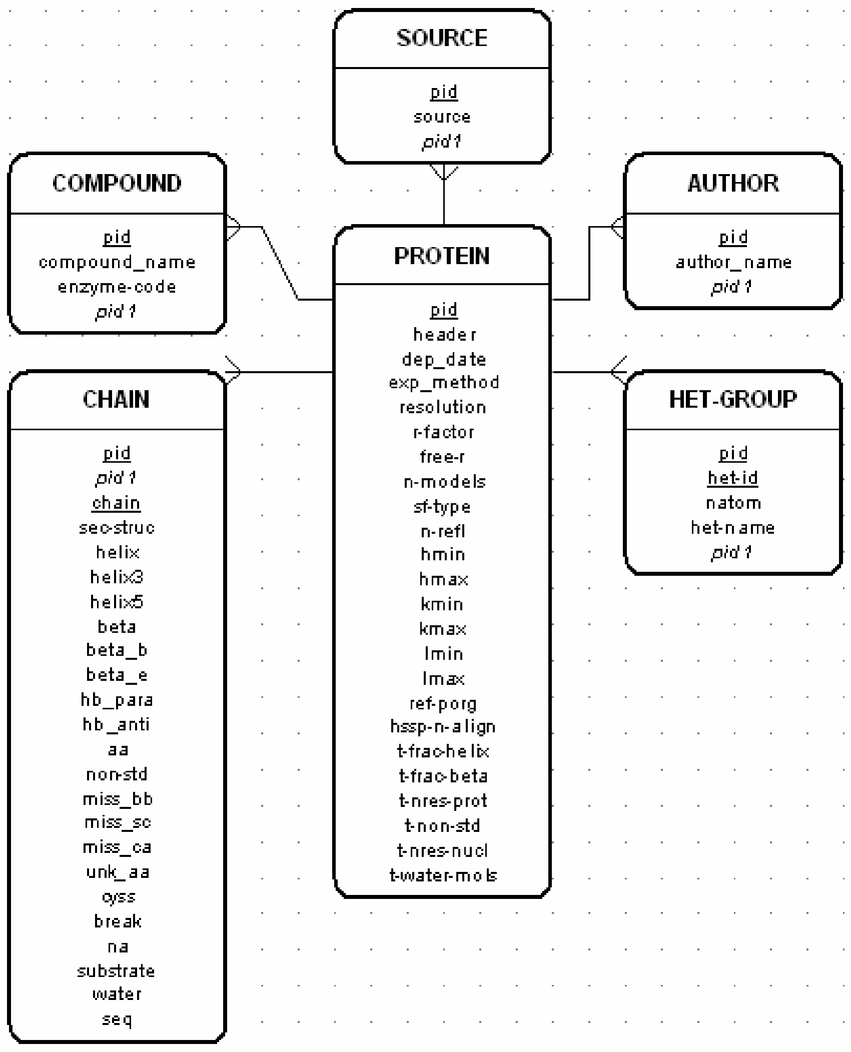


**Fig 2.2.a** DSD schematic of PDBFinder

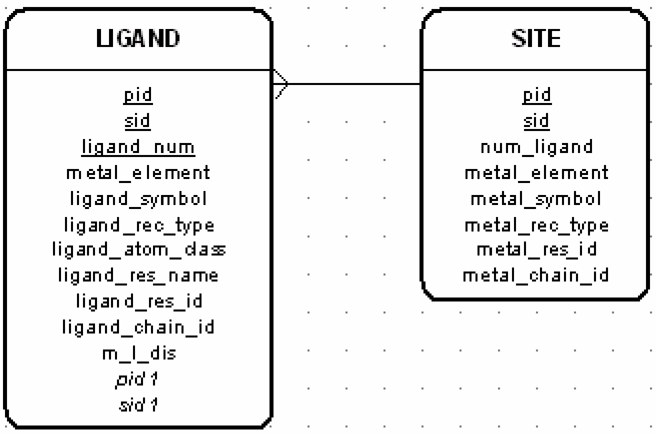


**Fig 2.2.b** DSD schematic of MDB

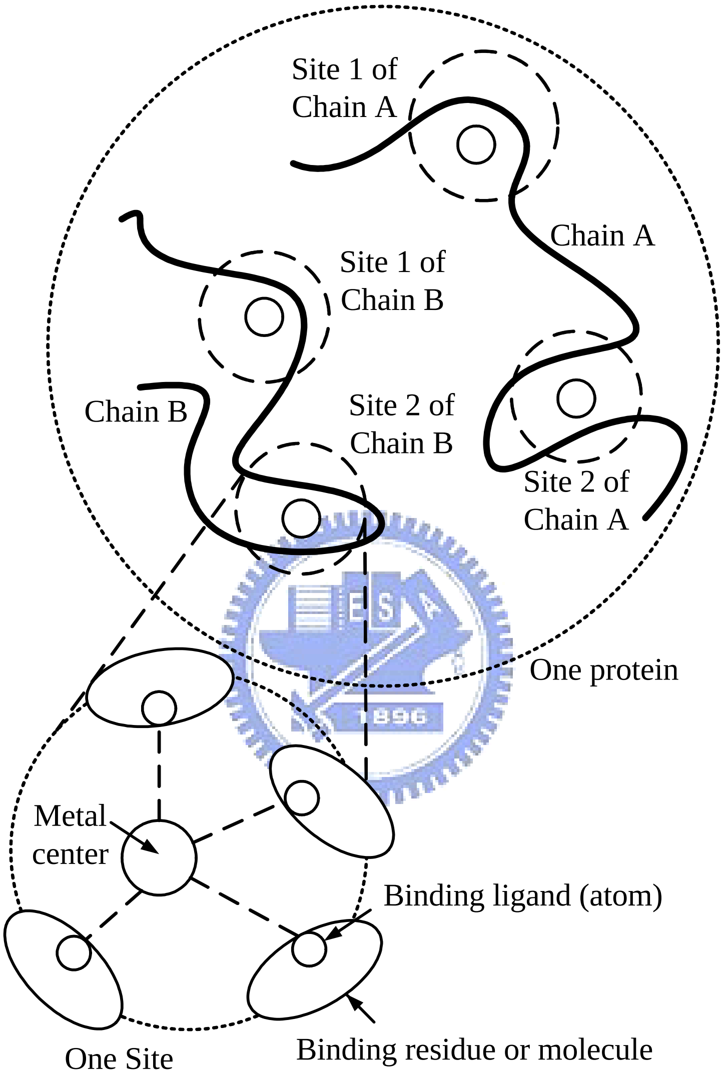


**Fig 2.2.c** Metal-binding protein data hierarchy

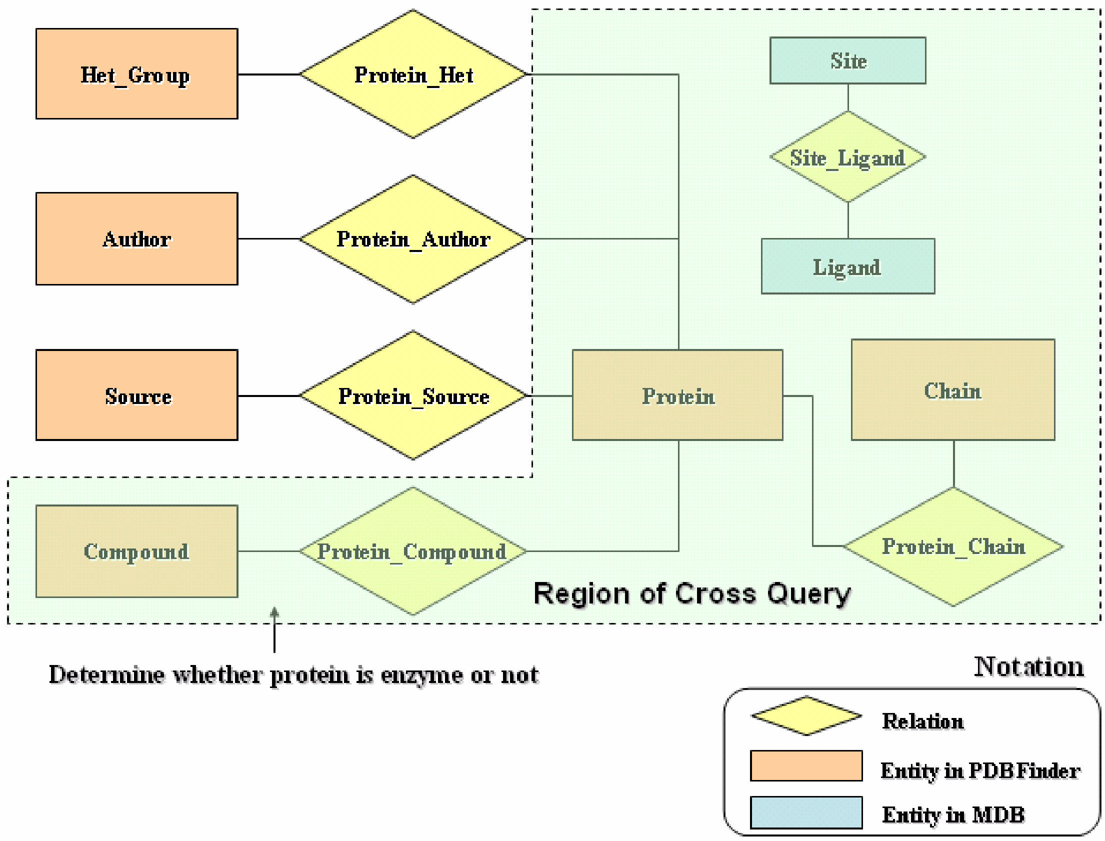


**Fig 2.2.d** Entity Relationship diagram of PDBFinder and MDB

## 2.3 Biological Data Processing and Sampling

In this thesis, there are 43 elements concerned in MDB version 17 as shown in **Table 2.3.a**. After cross querying between MDB and PDBFinder by scripts written in network programming language PHP (http://www.php.net/) on local MySQL (http://www.mysql.com) database, 41 and 35 metal types can be found in protein and enzyme respectively. **Table 2.3.b** shows the list of elements in metal binding residue prediction after cross querying. For simplicity, each instance in integrated database is treated as one chain of protein in real world; as the result, the inter-chain metal binding won’t be considered. By binding information from MDB, every position in protein chain sequence can be marked as binding or non-binding to be input for learn scheme (in chapter 3). **Figure 2.3.a** concludes all demanded data process and flow.

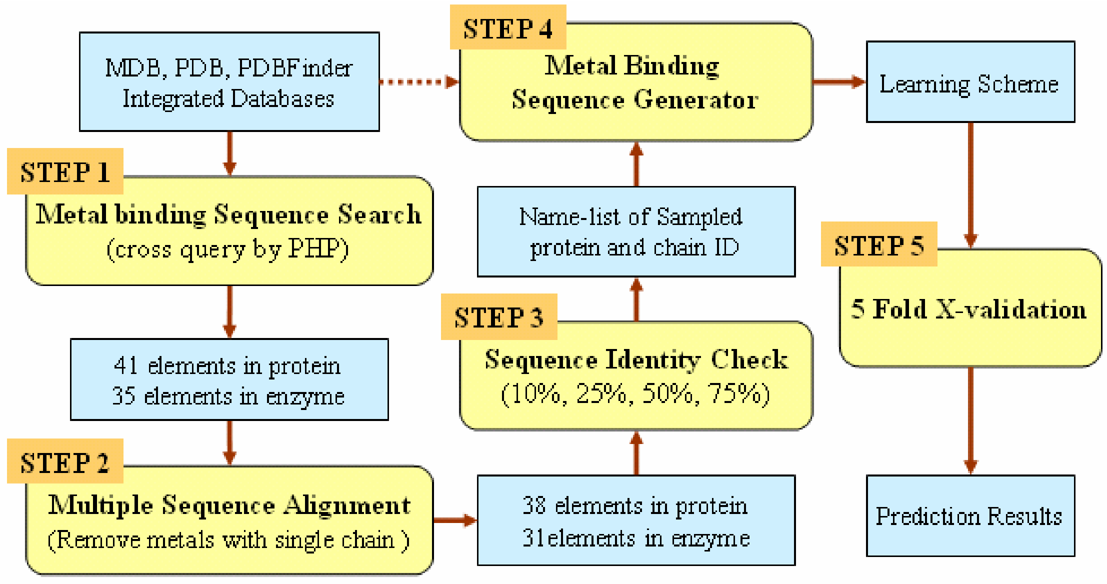


**Fig 2.3.a** Data processing pipeline

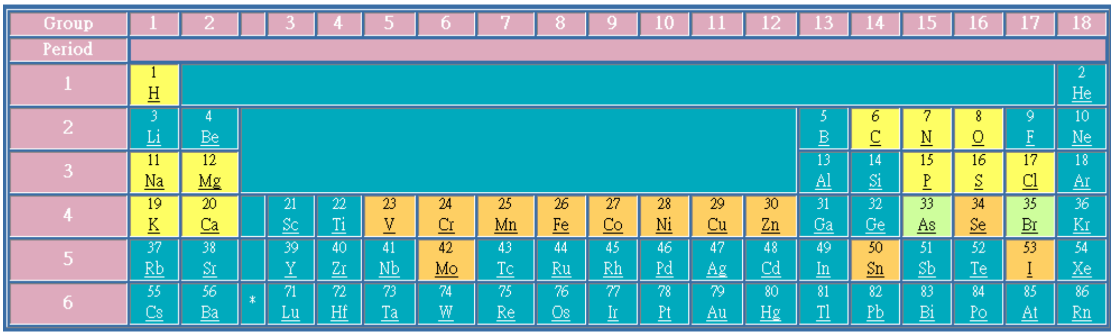
In **Table 2.3.b**, the first column indicates biological level which is the classification of life element in [3]. The third column is the element classification from periodic table. Next two columns are total number of metal binding chains in protein and enzyme. From existence of the field “EC_number” in entity “compound” of database PDBFinder, it is easy to identify whether a protein is an enzyme or not. The last column is the ratio of enzyme and all terms are ordered by this ratio.

| Element | Number of sites (Lines in ligand file) | Element | Number of sites (Lines in ligand file) |
|---|---|---|---|
| Mg | 6161 | Ho | 53 |
| Fe | 5357 | Sr | 53 |
| Se | 4861 | Sm | 52 |
| Ca | 4409 | V | 52 |
| Zn | 3326 | Pb | 46 |
| Na | 2018 | Pt | 41 |
| Mn | 1584 | Au | 35 |
| W | 1065 | Gd | 29 |
| Cu | 849 | Ba | 26 |
| Cd | 813 | Yb | 26 |
| U | 788 | Be | 25 |
| K | 784 | Cr | 23 |
| Hg | 465 | Ag | 15 |
| I | 416 | Rb | 15 |
| Co | 395 | Rh | 15 |
| Ni | 257 | La | 11 |
| As | 186 | Te | 8 |
| Mo | 102 | Eu | 7 |
| Al | 76 | Tb | 7 |
| Cs | 71 | Si | 5 |
| Tl | 59 | Li | 3 |
|  |  | In | 2 |

Ordered by number of sites w.r.t. element type

Sum : 34591

**Table 2.3.a** Number of site in MDB released files



| Group | 1 | 2 | | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period | | | | | | | | | | | | | | | | | | | |
| 1 | 1<br>H | | | | | | | | | | | | | | | | | | 2<br>He |
| 2 | 3<br>Li | 4<br>Be | | | | | | | | | | | | 5<br>B | 6<br>C | 7<br>N | 8<br>O | 9<br>F | 10<br>Ne |
| 3 | 11<br>Na | 12<br>Mg | | | | | | | | | | | | 13<br>Al | 14<br>Si | 15<br>P | 16<br>S | 17<br>Cl | 18<br>Ar |
| 4 | 19<br>K | 20<br>Ca | | 21<br>Sc | 22<br>Ti | 23<br>V | 24<br>Cr | 25<br>Mn | 26<br>Fe | 27<br>Co | 28<br>Ni | 29<br>Cu | 30<br>Zn | 31<br>Ga | 32<br>Ge | 33<br>As | 34<br>Se | 35<br>Br | 36<br>Kr |
| 5 | 37<br>Rb | 38<br>Sr | | 39<br>Y | 40<br>Zr | 41<br>Nb | 42<br>Mo | 43<br>Tc | 44<br>Ru | 45<br>Rh | 46<br>Pd | 47<br>Ag | 48<br>Cd | 49<br>In | 50<br>Sn | 51<br>Sb | 52<br>Te | 53<br>I | 54<br>Xe |
| 6 | 55<br>Cs | 56<br>Ba | * | 71<br>Lu | 72<br>Hf | 73<br>Ta | 74<br>W | 75<br>Re | 76<br>Os | 77<br>Ir | 78<br>Pt | 79<br>Au | 80<br>Hg | 81<br>Tl | 82<br>Pb | 83<br>Bi | 84<br>Po | 85<br>At | 86<br>Rn |

**Fig 2.3.b** Life elements in periodic table

**Figure 2.3.b** illustrates all life elements in periodic table in biological system., and there are 11 bulk biological elements－hydrogen (H), carbon (C), nitrogen (N), oxygen (O), sodium (Na), magnesium (Mg), phosphorus (P), sulfur (S), chlorine (Cl), potassium (K), and calcium (Ca), 12 trace elements essentials for life－vanadium (V), chromium (Cr), manganese (Mn), iron (Fe), cobalt (Co), nickel (Ni), copper (Cu), zinc (Zn), selenium (Se), molybdenum (Mo), tin (Sn), and iodine (I) and 2 possible trace elements－arsenic (As) and bromine (Br) in periodic table as indicated in [4]. After cross comparison, there are 4 of 11 (36%) bulk biological elements, 11 of 12 (91.6%) trace elements, and 1 of 2 (50%) possible trace elements in MDB as shown in **Table 2.3.b** which is classified by their biological level and order by their enzyme-protein ratio (E/P, the last column) with respect to each biological level set.

Owing to avoiding bias phenomenon of homology sequences in sets corresponding to different metal elements, sequence identity check has been applied to eliminate redundant sequence from each set. **Table 2.3.c** and **Table 2.3.d** show the set size comparison between different sets with respect to binding metal under different sequence identity thresholds. The selection criteria is when the average sequence identity of one chain to all sequences in the set except itself is less than the sequence identity threshold, the sequence is chose under this threshold. Before computing the pairwise sequence identity, all sequences in set are aligned by multiple sequence alignment (MSA) software － Clustalw. Single chain subset is skipped and noted the number of chain as “n/a (not available).”

| biological level | element name | element type | chains in Protein | chains in Enzyme | E/P |
|---|---|---|---|---|---|
| bulk element | K | Alkali metal | 243 | 175 | 72.02% |
| | Na | | 707 | 450 | 63.65% |
| | Mg | Alkaline metal | 1738 | 785 | 45.17% |
| | Ca | | 2455 | 1018 | 41.47% |
| trace element | Cr | Transition metal | 6 | 6 | 100.00% |
| | V | | 12 | 10 | 83.33% |
| | Co | | 174 | 97 | 55.75% |
| | Mo | | 48 | 24 | 50.00% |
| | Ni | | 172 | 85 | 49.42% |
| | Zn | | 2329 | 1064 | 45.68% |
| | Mn | | 956 | 400 | 41.84% |
| | Cu | | 567 | 213 | 37.57% |
| | Fe | | 2795 | 803 | 28.73% |
| | Se | Non-metal | 16 | 12 | 75.00% |
| | I | Halogen | 15 | 6 | 40.00% |
| possible trace element | As | Semi-metal | 79 | 51 | 64.56% |
| n/a | Hg | Transition metal | 221 | 103 | 46.61% |
| | Ag | | 3 | 1 | 33.33% |
| | Cd | | 267 | 80 | 29.96% |
| | Pt | | 7 | 2 | 28.57% |
| | W | | 7 | 2 | 28.57% |
| | U | | 63 | 14 | 22.22% |
| | Au | | 14 | 2 | 14.29% |
| | Tb | | 1 | 0 | 0.00% |
| | Te | Semi-metal | 4 | 2 | 50.00% |
| | Yb | Rare Earth | 14 | 7 | 50.00% |
| | Eu | | 2 | 1 | 50.00% |
| | Ho | | 6 | 1 | 16.67% |
| | Sm | | 20 | 3 | 15.00% |
| | Gd | | 16 | 0 | 0.00% |
| | La | | 5 | 0 | 0.00% |
| | Tl | Basic metal | 18 | 18 | 100.00% |
| | Al | | 22 | 17 | 77.27% |
| | Pb | | 22 | 7 | 31.82% |
| | In | | 1 | 0 | 0.00% |
| | Ba | Alkaline metal | 3 | 2 | 66.67% |
| | Sr | | 12 | 3 | 25.00% |
| | Be | | 6 | 0 | 0.00% |
| | Cs | Alkali metal | 7 | 4 | 57.14% |
| | Li | | 2 | 1 | 50.00% |
| | Rb | | 1 | 0 | 0.00% |

**Table 2.3.b** Number of chains in protein set and enzyme set after cross querying

| biological level | element | Total chains in protein | Sequence Identity Threshold | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 75% | | 50% | | 25% | | 10% | |
| | | | R | R/T | R | R/T | R | R/T | R | R/T |
| bulk element | Ca | 2455 | 2455 | 100.00% | 2455 | 100.00% | 2455 | 100.00% | 2322 | 94.58% |
| | Mg | 1738 | 1738 | 100.00% | 1738 | 100.00% | 1738 | 100.00% | 1738 | 100.00% |
| | Na | 707 | 707 | 100.00% | 707 | 100.00% | 707 | 100.00% | 547 | 77.37% |
| | K | 243 | 243 | 100.00% | 243 | 100.00% | 243 | 100.00% | 173 | 71.19% |
| trace element | Fe | 2795 | 2795 | 100.00% | 2795 | 100.00% | 2795 | 100.00% | 2241 | 80.18% |
| | Zn | 2329 | 2329 | 100.00% | 2329 | 100.00% | 2329 | 100.00% | 2329 | 100.00% |
| | Mn | 956 | 956 | 100.00% | 956 | 100.00% | 956 | 100.00% | 706 | 73.85% |
| | Cu | 567 | 567 | 100.00% | 567 | 100.00% | 567 | 100.00% | 307 | 54.14% |
| | Co | 174 | 174 | 100.00% | 174 | 100.00% | 174 | 100.00% | 129 | 74.14% |
| | Ni | 172 | 172 | 100.00% | 172 | 100.00% | 172 | 100.00% | 99 | 57.56% |
| | Mo | 48 | 48 | 100.00% | 48 | 100.00% | 13 | 27.08% | 6 | 12.50% |
| | Se | 16 | 16 | 100.00% | 7 | 43.75% | 7 | 43.75% | 1 | 6.25% |
| | I | 15 | 15 | 100.00% | 15 | 100.00% | 15 | 100.00% | 5 | 33.33% |
| | V | 12 | 12 | 100.00% | 5 | 41.67% | 3 | 25.00% | 0 | 0.00% |
| | Cr | 6 | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| possible trace element | As | 79 | 79 | 100.00% | 21 | 26.58% | 21 | 26.58% | 9 | 11.39% |
| n/a | Cd | 267 | 267 | 100.00% | 267 | 100.00% | 267 | 100.00% | 267 | 100.00% |
| | Hg | 221 | 221 | 100.00% | 221 | 100.00% | 154 | 69.68% | 128 | 57.92% |
| | U | 63 | 63 | 100.00% | 32 | 50.79% | 32 | 50.79% | 17 | 26.98% |
| | Al | 22 | 22 | 100.00% | 22 | 100.00% | 10 | 45.45% | 1 | 4.55% |
| | Pb | 22 | 22 | 100.00% | 22 | 100.00% | 22 | 100.00% | 4 | 18.18% |
| | Sm | 20 | 20 | 100.00% | 20 | 100.00% | 12 | 60.00% | 6 | 30.00% |
| | Tl | 18 | 18 | 100.00% | 6 | 33.33% | 2 | 11.11% | 0 | 0.00% |
| | Gd | 16 | 16 | 100.00% | 16 | 100.00% | 10 | 62.50% | 1 | 6.25% |
| | Au | 14 | 14 | 100.00% | 14 | 100.00% | 10 | 71.43% | 0 | 0.00% |
| | Yb | 14 | 14 | 100.00% | 14 | 100.00% | 14 | 100.00% | 1 | 7.14% |
| | Sr | 12 | 12 | 100.00% | 12 | 100.00% | 12 | 100.00% | 1 | 8.33% |
| | Cs | 7 | 7 | 100.00% | 7 | 100.00% | 3 | 42.86% | 0 | 0.00% |
| | Pt | 7 | 7 | 100.00% | 7 | 100.00% | 7 | 100.00% | 0 | 0.00% |
| | W | 7 | 7 | 100.00% | 7 | 100.00% | 7 | 100.00% | 0 | 0.00% |
| | Be | 6 | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | Ho | 6 | 6 | 100.00% | 4 | 66.67% | 1 | 16.67% | 0 | 0.00% |
| | La | 5 | 5 | 100.00% | 5 | 100.00% | 1 | 20.00% | 0 | 0.00% |
| | Te | 4 | 4 | 100.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | Ag | 3 | 3 | 100.00% | 1 | 33.33% | 0 | 0.00% | 0 | 0.00% |
| | Ba | 3 | 3 | 100.00% | 1 | 33.33% | 0 | 0.00% | 0 | 0.00% |
| | Eu | 2 | 2 | 100.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | Li | 2 | 2 | 100.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | In | 1 | n/a | | | | | | | |
| | Rb | 1 | | | | | | | | |
| | Tb | 1 | | | | | | | | |

**Table 2.3.c** Protein set size under different sequence identity threshold

| biological level | element | Total chains in enzyme | Sequence Identity Threshold | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 75% | | 50% | | 25% | | 10% | |
| | | | R | R/T | R | R/T | R | R/T | R | R/T |
| bulk element | Ca | 1018 | 1018 | 100.00% | 1018 | 100.00% | 1018 | 100.00% | 892 | 87.62% |
| | Mg | 785 | 785 | 100.00% | 785 | 100.00% | 785 | 100.00% | 661 | 84.20% |
| | Na | 450 | 450 | 100.00% | 450 | 100.00% | 450 | 100.00% | 245 | 54.44% |
| | K | 175 | 175 | 100.00% | 175 | 100.00% | 175 | 100.00% | 100 | 57.14% |
| trace element | Zn | 1064 | 1064 | 100.00% | 1064 | 100.00% | 1064 | 100.00% | 994 | 93.42% |
| | Fe | 803 | 803 | 100.00% | 803 | 100.00% | 803 | 100.00% | 753 | 93.77% |
| | Mn | 400 | 400 | 100.00% | 400 | 100.00% | 400 | 100.00% | 222 | 55.50% |
| | Cu | 213 | 213 | 100.00% | 213 | 100.00% | 182 | 85.45% | 79 | 37.09% |
| | Co | 97 | 97 | 100.00% | 97 | 100.00% | 97 | 100.00% | 48 | 49.48% |
| | Ni | 85 | 85 | 100.00% | 85 | 100.00% | 66 | 77.65% | 23 | 27.06% |
| | Mo | 24 | 24 | 100.00% | 16 | 66.67% | 11 | 45.83% | 0 | 0.00% |
| | Se | 12 | 3 | 25.00% | 3 | 25.00% | 3 | 25.00% | 0 | 0.00% |
| | V | 10 | 10 | 100.00% | 3 | 30.00% | 1 | 10.00% | 0 | 0.00% |
| | I | 6 | 6 | 100.00% | 6 | 100.00% | 4 | 66.67% | 0 | 0.00% |
| | Cr | 6 | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| possible trace element | As | 51 | 11 | 21.57% | 11 | 21.57% | 11 | 21.57% | 5 | 9.80% |
| n/a | Hg | 103 | 103 | 100.00% | 103 | 100.00% | 63 | 61.17% | 41 | 39.81% |
| | Cd | 80 | 80 | 100.00% | 80 | 100.00% | 80 | 100.00% | 58 | 72.50% |
| | Tl | 18 | 18 | 100.00% | 6 | 33.33% | 2 | 11.11% | 0 | 0.00% |
| | Al | 17 | 17 | 100.00% | 17 | 100.00% | 7 | 41.18% | 0 | 0.00% |
| | U | 14 | 14 | 100.00% | 14 | 100.00% | 4 | 28.57% | 0 | 0.00% |
| | Pb | 7 | 7 | 100.00% | 7 | 100.00% | 2 | 28.57% | 0 | 0.00% |
| | Yb | 7 | 7 | 100.00% | 7 | 100.00% | 0 | 0.00% | 0 | 0.00% |
| | Cs | 4 | 4 | 100.00% | 2 | 50.00% | 0 | 0.00% | 0 | 0.00% |
| | Sm | 3 | 3 | 100.00% | 1 | 33.33% | 0 | 0.00% | 0 | 0.00% |
| | Sr | 3 | 3 | 100.00% | 3 | 100.00% | 0 | 0.00% | 0 | 0.00% |
| | Au | 2 | 2 | 100.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | Pt | 2 | 2 | 100.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | W | 2 | 2 | 100.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | Te | 2 | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | Ba | 2 | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | Ho | 1 | n/a | | | | | | | |
| | Ag | 1 | | | | | | | | |
| | Eu | 1 | | | | | | | | |
| | Li | 1 | | | | | | | | |
| | Gd | 0 | | | | | | | | |
| | Be | 0 | | | | | | | | |
| | La | 0 | | | | | | | | |
| | In | 0 | | | | | | | | |
| | Rb | 0 | | | | | | | | |
| | Tb | 0 | | | | | | | | |

**Table 2.3.d** Enzyme set size under different sequence identity threshold

# Chapter 3

# Machine Learning Scheme

The learning schemes used, in this thesis, are as simple as possible so that it becomes easy to observe the prediction performances according to various coding using non-biological or biological features. Besides, the relationship between the performance and size of sequence sampling window also can be found.

## 3.1 Neural Networks

Neural network consist of groups of parallel processing unit with connection between layers and each connection has one weight parameter. Neural networks use these weights between layers to "memorize" the patterns fed from input layer. The basic unit within a layer is an artificial neuron (node) shown as one circle in **Figure 3.1.a.** In this thesis, multi-layer Perceptron (MLP) neural networks with back-propagation (BP) algorithm are chosen as learning machine to complete our experiments. In the NNs, we used one hidden layer with 30 hidden nodes as shown in **Figure 3.1.a** so that there are (30 × dimension of input layer) weights between input layer and hidden layer and (30 × dimension of output layer) weights between hidden layer and output layer respectively.

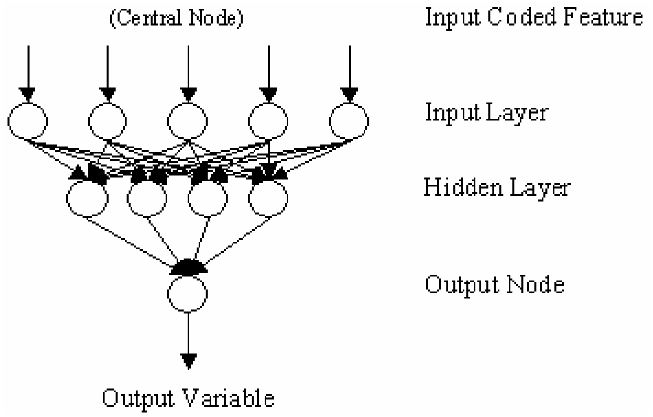
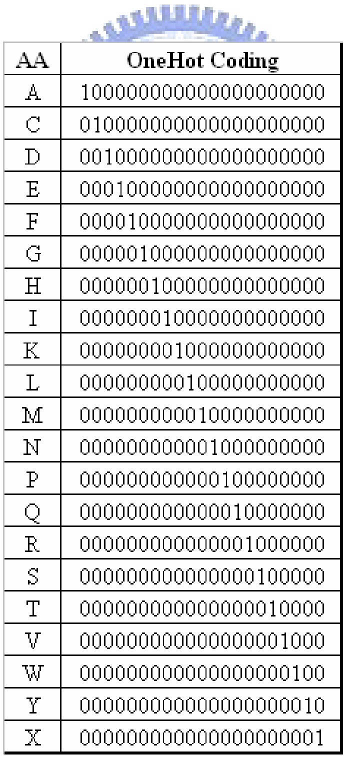


**Fig 3.1.a** simple full connection neural networks

Besides, dimension of input layer is depended on the size of sequence sample window and dimension of output layer is two. In testing phase, if first output value is larger than second one, then the prediction result is defined as positive (binding), otherwise negative (non-binding).

## 3.2 Feature Encoding

There are two input coding used in our experiments. One is direct one-hot coding which presents every amino acid as one 21-bits array. Only one bit in array is ‘1’ and other bits in array are ‘0’.  In this way, every type of natural amino acid can be indicated by the position of the only “1” bit. Owing to the unknown type (usually use the symbol ‘X’ in sequence) of amino acid in protein sequence, add one bit to record this condition. This is the non-biological coding for amino acid as illustrated in **Table 3.2.a**.



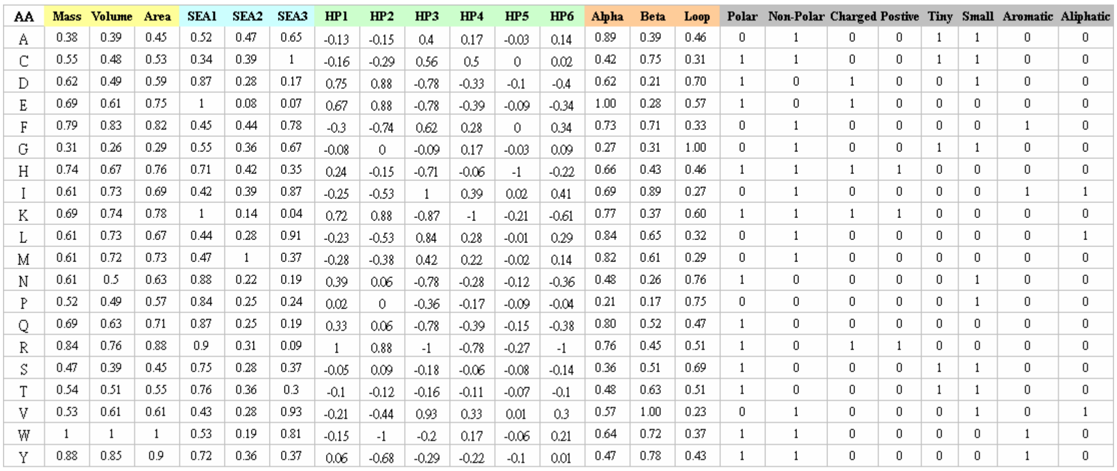
| AA | OneHot Coding |
|---|---|
| A | 100000000000000000000 |
| C | 010000000000000000000 |
| D | 001000000000000000000 |
| E | 000100000000000000000 |
| F | 000010000000000000000 |
| G | 000001000000000000000 |
| H | 000000100000000000000 |
| I | 000000010000000000000 |
| K | 000000001000000000000 |
| L | 000000000100000000000 |
| M | 000000000010000000000 |
| N | 000000000001000000000 |
| P | 000000000000100000000 |
| Q | 000000000000010000000 |
| R | 000000000000001000000 |
| S | 000000000000000100000 |
| T | 000000000000000010000 |
| V | 000000000000000001000 |
| W | 000000000000000000100 |
| Y | 000000000000000000010 |
| X | 000000000000000000001 |

**Table 3.2.a** One-hot coding table for 20 amino acids

Another coding method is done by referencing five different types of biological features about amino acid as shown in **Table 3.2.b.** and **Table 3.2.c.**

| Feature Set (size) | Definition and Content | | References |
|---|---|---|---|
| Physical (3) | mass, volume, and area | | [7]NCBI statistics |
| Solvent Exposed Area Levels (3) | three levels | SEA > 30 | [8] |
| | | 10 < SEA < 30 | |
| | | SEA < 10 | |
| Hydrophobicity Scales (6) | six scales | Engleman-Steitz | [9] |
| | | Hopp-Woods | [10] |
| | | Kyte-Doolittle | [11] |
| | | Janin | [12] |
| | | Chothia | [13] |
| | | Eisenberg Weiss | [14] |
| Secondary Structure Propensity (3) | three secondary structures | Alpha helix | [1] |
| | | Beta strand | |
| | | Turn (loop, coil) | |
| Chemical Classification (8) | eight classifications | Polar | [7] |
| | | Non-Polar | |
| | | Charged | |
| | | Positive | |
| | | Tiny | |
| | | Small | |
| | | Aromatic | |
| | | Aliphatic | |

**Table 3.2.b** Definitions of five biological feature sets



| AA | Mass | Volume | Area | SEA1 | SEA2 | SEA3 | HP1 | HP2 | HP3 | HP4 | HP5 | HP6 | Alpha | Beta | Loop | Polar | Non-Polar | Charged | Postive | Tiny | Small | Aromatic | Aliphatic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.38 | 0.39 | 0.45 | 0.52 | 0.47 | 0.65 | -0.13 | -0.15 | 0.4 | 0.17 | -0.03 | 0.14 | 0.89 | 0.39 | 0.46 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| C | 0.55 | 0.48 | 0.53 | 0.34 | 0.39 | 1 | -0.16 | -0.29 | 0.56 | 0.5 | 0 | 0.02 | 0.42 | 0.75 | 0.31 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| D | 0.62 | 0.49 | 0.59 | 0.87 | 0.28 | 0.17 | 0.75 | 0.88 | -0.78 | -0.33 | -0.1 | -0.4 | 0.62 | 0.21 | 0.70 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| E | 0.69 | 0.61 | 0.75 | 1 | 0.08 | 0.07 | 0.67 | 0.88 | -0.78 | -0.39 | -0.09 | -0.34 | 1.00 | 0.28 | 0.57 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| F | 0.79 | 0.83 | 0.82 | 0.45 | 0.44 | 0.78 | -0.3 | -0.74 | 0.62 | 0.28 | 0 | 0.34 | 0.73 | 0.71 | 0.33 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| G | 0.31 | 0.26 | 0.29 | 0.55 | 0.36 | 0.67 | -0.08 | 0 | -0.09 | 0.17 | -0.03 | 0.09 | 0.27 | 0.31 | 1.00 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| H | 0.74 | 0.67 | 0.76 | 0.71 | 0.42 | 0.35 | 0.24 | -0.15 | -0.71 | -0.06 | -1 | -0.22 | 0.66 | 0.43 | 0.46 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| I | 0.61 | 0.73 | 0.69 | 0.42 | 0.39 | 0.87 | -0.25 | -0.53 | 1 | 0.39 | 0.02 | 0.41 | 0.69 | 0.89 | 0.27 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| K | 0.69 | 0.74 | 0.78 | 1 | 0.14 | 0.04 | 0.72 | 0.88 | -0.87 | -1 | -0.21 | -0.61 | 0.77 | 0.37 | 0.60 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| L | 0.61 | 0.73 | 0.67 | 0.44 | 0.28 | 0.91 | -0.23 | -0.53 | 0.84 | 0.28 | -0.01 | 0.29 | 0.84 | 0.65 | 0.32 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| M | 0.61 | 0.72 | 0.73 | 0.47 | 1 | 0.37 | -0.28 | -0.38 | 0.42 | 0.22 | -0.02 | 0.14 | 0.82 | 0.61 | 0.29 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| N | 0.61 | 0.5 | 0.63 | 0.88 | 0.22 | 0.19 | 0.39 | 0.06 | -0.78 | -0.28 | -0.12 | -0.36 | 0.48 | 0.26 | 0.76 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| P | 0.52 | 0.49 | 0.57 | 0.84 | 0.25 | 0.24 | 0.02 | 0 | -0.36 | -0.17 | -0.09 | -0.04 | 0.21 | 0.17 | 0.75 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Q | 0.69 | 0.63 | 0.71 | 0.87 | 0.25 | 0.19 | 0.33 | 0.06 | -0.78 | -0.39 | -0.15 | -0.38 | 0.80 | 0.52 | 0.47 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R | 0.84 | 0.76 | 0.88 | 0.9 | 0.31 | 0.09 | 1 | 0.88 | -1 | -0.78 | -0.27 | -1 | 0.76 | 0.45 | 0.51 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| S | 0.47 | 0.39 | 0.45 | 0.75 | 0.28 | 0.37 | -0.05 | 0.09 | -0.18 | -0.06 | -0.08 | -0.14 | 0.36 | 0.51 | 0.69 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| T | 0.54 | 0.51 | 0.55 | 0.76 | 0.36 | 0.3 | -0.1 | -0.12 | -0.16 | -0.11 | -0.07 | -0.1 | 0.48 | 0.63 | 0.51 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| V | 0.53 | 0.61 | 0.61 | 0.43 | 0.28 | 0.93 | -0.21 | -0.44 | 0.93 | 0.33 | 0.01 | 0.3 | 0.57 | 1.00 | 0.23 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| W | 1 | 1 | 1 | 0.53 | 0.19 | 0.81 | -0.15 | -1 | -0.2 | 0.17 | -0.06 | 0.21 | 0.64 | 0.72 | 0.37 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| Y | 0.88 | 0.85 | 0.9 | 0.72 | 0.36 | 0.37 | 0.06 | -0.68 | -0.29 | -0.22 | -0.1 | 0.01 | 0.47 | 0.78 | 0.43 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

**Table 3.2.c** Values of five biological feature sets

Because the binding behavior of central metal atom is influenced by the surrounding environment in protein, it is necessary to observe in wider scope than

[7] National Biotechnology Information Center, U.S.A. http://www.ncbi.nlm.nih.gov/

single one amino acid so as to determine whether the binding happens or not. Accordingly, each input vector applied to learning machine is extracted from one segment of entire chain by the concept － continuous sliding window. Each sliding window is centered by the “target” amino acid. And the rest of the amino acids in window are the “neighbors” of the target. **Figure 3.2.c** shows the feature extraction, learning scheme and how sliding window works. For simplicity the window size illustrated is 5.

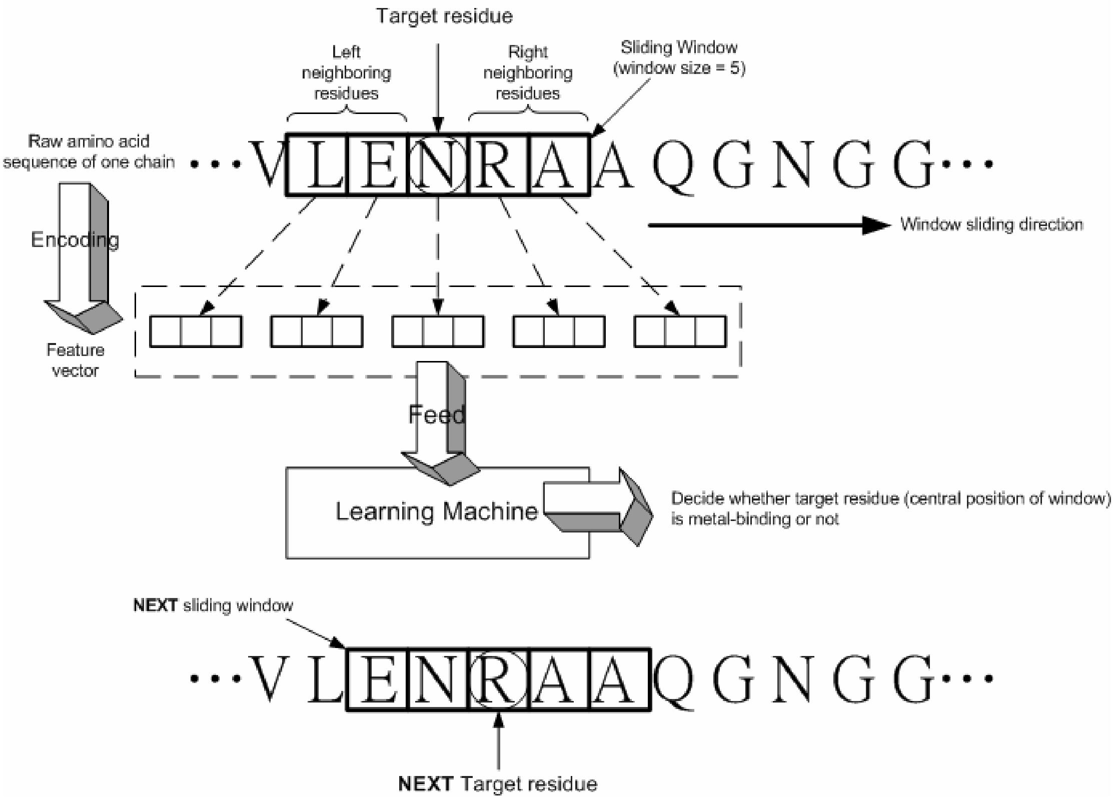

Target residue
Left neighboring residues
Right neighboring residues
Sliding Window (window size = 5)
Raw amino acid sequence of one chain
• • • V L E N R A A Q G N G G• • •
Encoding
Window sliding direction
Feature vector
Feed
Learning Machine
Decide whether target residue (central position of window) is metal-binding or not
NEXT sliding window
• • • V L E N R A A Q G N G G• • •
NEXT Target residue

**Fig 3.2.c** Feature extraction, learning scheme and sliding window

# Chapter 4

# Results and Conclusion

In out experiments, there are two major sets — protein and enzyme sets with specified sequence identity constraint. To avoid sampling bias, the sequence identity threshold is set as 25% — the threshold of homology modeling. Each set corresponding to different metal element has its own neural network which is trained for 150 epochs to observe its time-varied characteristics. Five fold cross validation is used to calculate performance, shown in **Fig 4.a**.

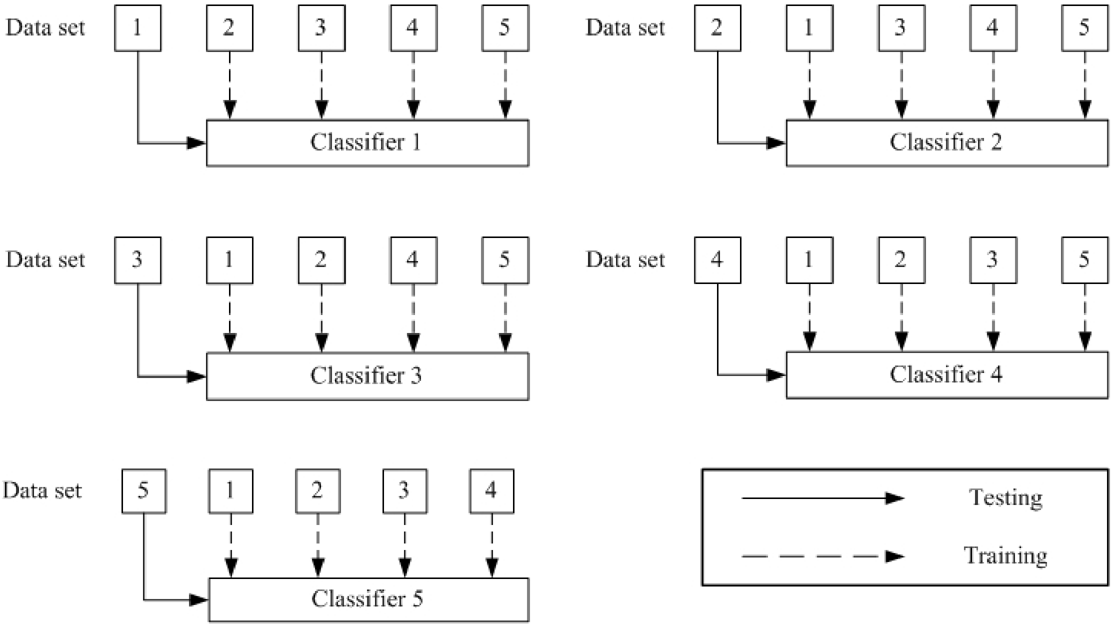


**Fig 4.a** five fold cross validation

## 4.1 Performance Measures

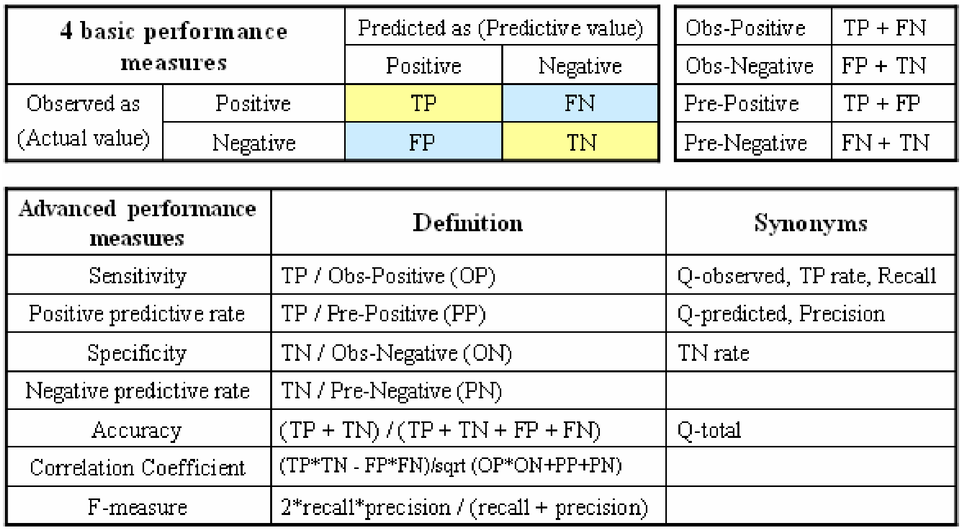
Four basic performance measures are used in the experiment — TP (true positive, when an instance (residue) is observed as positive, and predicted as positive), TN (true negative, when an instance is observed as negative, and predicted as negative), FP (false positive, when an instance is observed as negative, but predicted as positive), and FN (false negative, when an instance is observed as positive, but predicted as negative).

Besides, three performance measures, $Q_{total}$ (Equation 1), $Q_{predicted}$ (Equation 2) and $Q_{observed}$ (Equation 3), are also used in our experiments. $Q_{predicted}$ is defined as the ratio between the "true" and total (true and false) instances predicted as positive (binding) and it also shows that how likely the result of prediction would be true when an instance predicted as positive. $Q_{observed}$ is defined as the ratio between the instances truly predicted as positive and instances observed as positive and it also shows the ability to discover binding residues so that it is also called "sensitivity." More detailed performance measures and comparison are listed in **Table 4.1**.

$$Q_{total} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

$$Q_{predicted} = \frac{TP}{TP + FP} \qquad (2)$$

$$Q_{observed} = \frac{TP}{TP + FN} \qquad (3)$$



| 4 basic performance measures | | Predicted as (Predictive value) | |
|---|---|---|---|
| | | Positive | Negative |
| Observed as (Actual value) | Positive | TP | FN |
| | Negative | FP | TN |

| | |
|---|---|
| Obs-Positive | TP + FN |
| Obs-Negative | FP + TN |
| Pre-Positive | TP + FP |
| Pre-Negative | FN + TN |

| Advanced performance measures | Definition | Synonyms |
|---|---|---|
| Sensitivity | TP / Obs-Positive (OP) | Q-observed, TP rate, Recall |
| Positive predictive rate | TP / Pre-Positive (PP) | Q-predicted, Precision |
| Specificity | TN / Obs-Negative (ON) | TN rate |
| Negative predictive rate | TN / Pre-Negative (PN) | |
| Accuracy | (TP + TN) / (TP + TN + FP + FN) | Q-total |
| Correlation Coefficient | (TP*TN - FP*FN)/sqrt (OP*ON+PP+PN) | |
| F-measure | 2*recall*precision / (recall + precision) | |

**TABLE 4.1** DETAILED PERFORMANCE MEASURES AND COMPARISON

## 4.2 Experiments on One-hot Coding Method

In this section, one-hot coding method is used in all experiments varied by size of window from 5 to 17 so as to observe the change of performance according to different window size. Owing to the extremely low P/N (positive and negative instance ratio), specificity and negative prediction rate (almost approach 100%) are relatively higher than sensitivity (Q-observed). As the result, sensitivity (Q-observed) becomes only one critical term in performance measures in these absolutely unbalanced (positive and negative) training. **Table 4.2.a** shows all Q-observed in enzyme set with respect to different window size. **Figure 4.2.a** and **Figure 4.2.b** offer detailed comparison in bulk and trace elements respectively.

| **biological level** | **element** | **window size** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **5** | **7** | **9** | **11** | **13** | **15** | **17** | **P/N** |
| Bulk element | Ca | 21.01% | 17.31% | 16.13% | 16.47% | 15.97% | 18.49% | 20.50% | 0.67% |
| | K | 2.99% | 14.93% | 17.91% | 23.88% | 28.36% | 37.31% | 34.33% | 0.46% |
| | Mg | 8.50% | 10.46% | 12.09% | 10.46% | 13.40% | 14.05% | 18.63% | 0.62% |
| | Na | 9.59% | 13.01% | 13.70% | 19.18% | 19.18% | 19.18% | 24.66% | 0.62% |
| Trace element | Co | 31.43% | 34.29% | 35.71% | 45.71% | 48.57% | 50.00% | 54.29% | 1.25% |
| | Cr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 1.67% |
| | Cu | 32.73% | 33.64% | 41.82% | 42.73% | 40.91% | 42.73% | 46.36% | 0.92% |
| | Fe | 40.40% | 35.82% | 35.82% | 36.39% | 37.25% | 40.40% | 38.40% | 0.62% |
| | I | 0.00% | 25.00% | 62.50% | 75.00% | 75.00% | 75.00% | 87.50% | 0.46% |
| | Mn | 21.94% | 31.12% | 29.08% | 33.16% | 32.65% | 31.63% | 35.71% | 1.22% |
| | Mo | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% | 20.00% | 0.95% |
| | Ni | 42.42% | 42.42% | 51.52% | 54.55% | 51.52% | 54.55% | 63.64% | 1.04% |
| | Se | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.50% |
| | V | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.51% |
| | Zn | 24.22% | 15.74% | 14.19% | 24.74% | 29.58% | 27.51% | 30.10% | 0.57% |
| Possibly Trace element | As | 25.00% | 25.00% | 25.00% | 50.00% | 50.00% | 75.00% | 62.50% | 0.84% |
| N/A | Al | 0.00% | 10.00% | 80.00% | 90.00% | 90.00% | 90.00% | 100.00% | 0.19% |
| | Au | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.76% |
| | Ba | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.58% |
| | Cd | 31.08% | 30.41% | 37.16% | 38.51% | 39.86% | 43.92% | 47.30% | 0.52% |
| | Cs | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 20.00% | 60.00% | 0.46% |
| | Hg | 29.73% | 43.24% | 45.95% | 51.35% | 58.11% | 56.76% | 56.76% | 0.16% |
| | Pb | 50.00% | 50.00% | 58.33% | 58.33% | 66.67% | 75.00% | 75.00% | 0.90% |
| | Pt | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.36% |
| | Sm | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 71.43% | 85.71% | 0.26% |
| | Sr | 0.00% | 0.00% | 0.00% | 50.00% | 100.00% | 75.00% | 100.00% | 0.88% |
| | Te | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.42% |
| | Tl | 62.50% | 87.50% | 87.50% | 87.50% | 100.00% | 100.00% | 100.00% | 0.18% |
| | U | 42.86% | 42.86% | 85.71% | 85.71% | 100.00% | 71.43% | 100.00% | 0.38% |
| | W | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.17% |
| | Yb | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 1.09% |

**Table 4.2.a** Q-observed of 31 elements in enzyme set w.r.t. different window size

Increasing window size indeed improves the sensitivity in each metal set; but in some sets, it is not necessary to have better performance with longer window size, such as in metal sets calcium (Ca) and zinc (Zn). Nevertheless, the large computation cost resulted from the extension of sampling window doesn’t bring great and rapid improvement on performance.

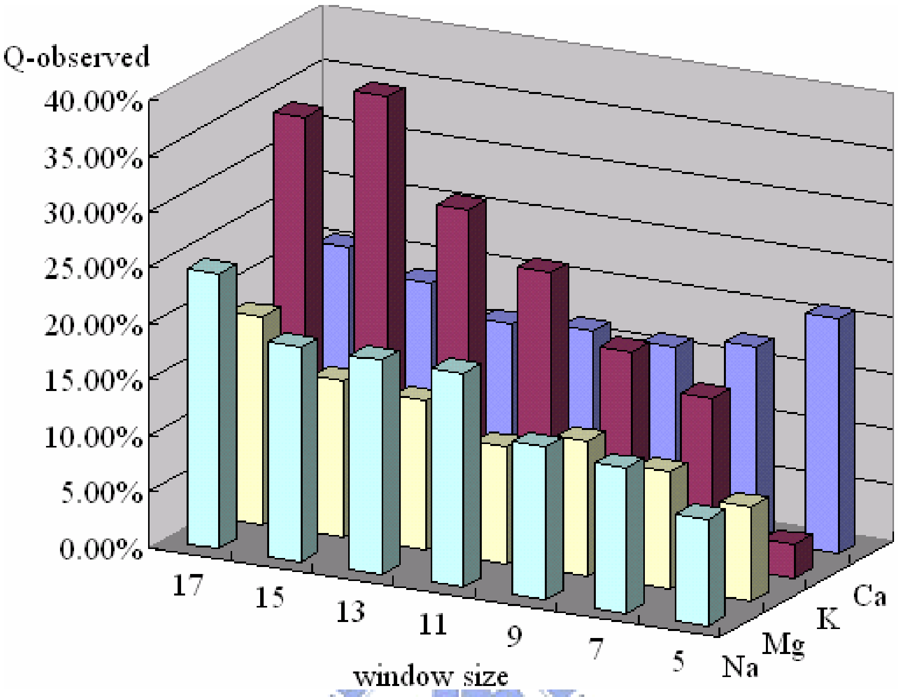


**Fig 4.2.a** Q-observed of 4 bulk elements by one-hot coding

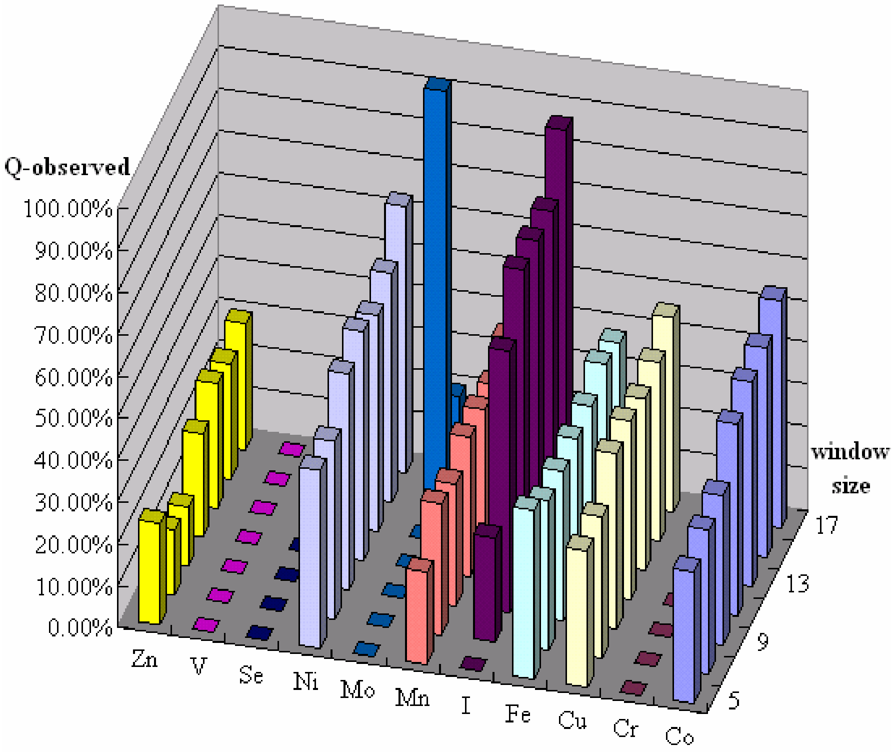


**Fig 4.2.b** Q-observed of 11 trace elements by one-hot coding

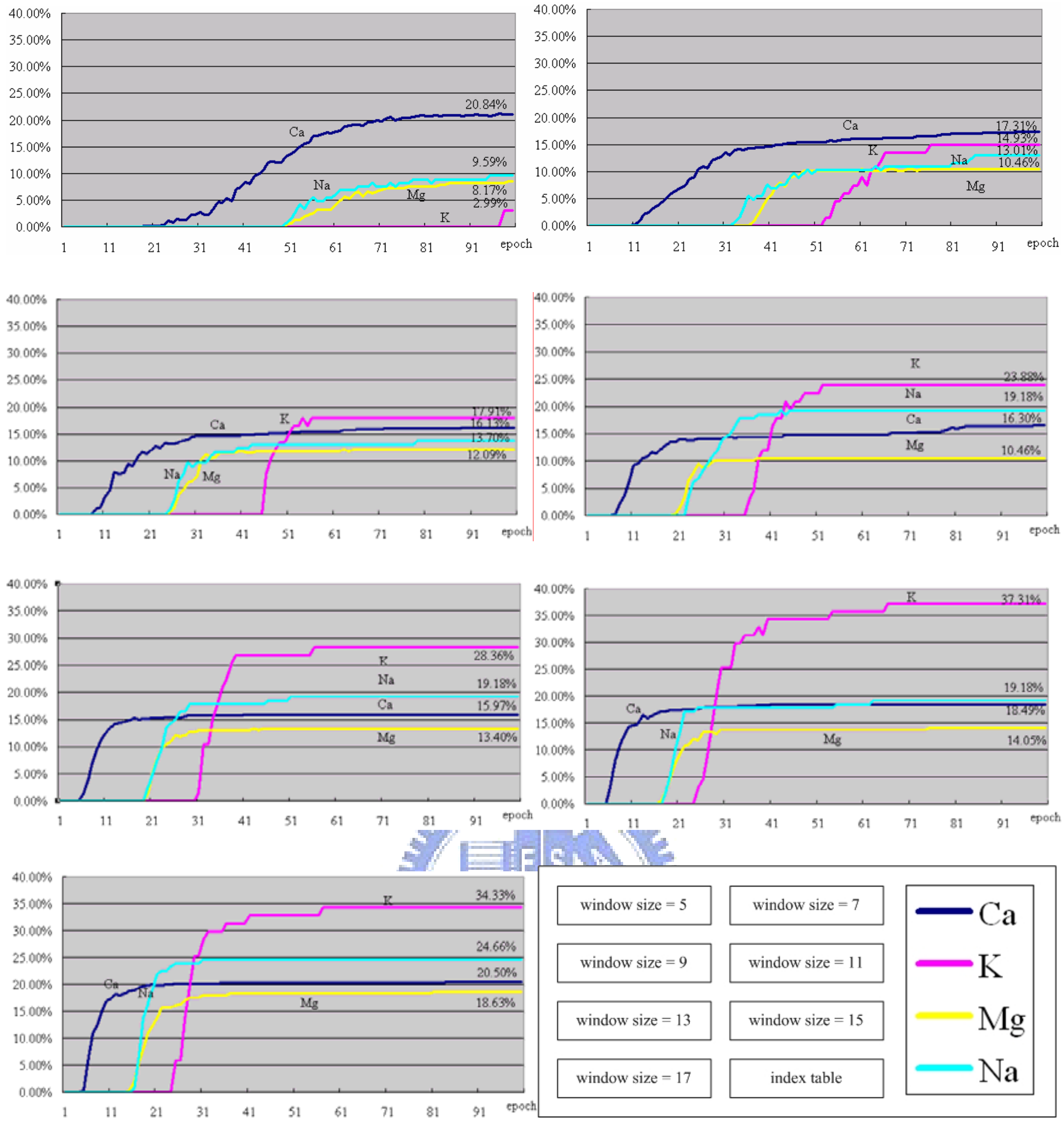


**Fig 4.2.c** Q-observed versus training epoch of bulk elements

Besides, every binding metal specified subset is trained for 100 epochs in experiments and Q-observed training curves are shown as **Figure 4.2.c** and right-button corner of figure is index table for these subfigures in it. In these subfigures, there are two labels (element name and Q-observed value at 100 epochs) on each training curve. By comparing these curves, one can observe how Q-observed values grow under window extension：

(1) All Q-observed values are not greater than 40% under one-hot coding method in bulk elements. It might be the limitation of one-hot coding method to this problem.

(2) While size of window increases, in general, every training curve rises earlier, and achieves higher Q-observed value at end of training. In addition, the rising edge of curve becomes sharper (curve converges earlier).

(3) Following the last observation in (2) and comparing the curve of four bulk elements, potassium (K) is the most sensitive element than other three elements while window extends.

## 4.3 Comparison between Different Feature sets

In last section, the one-hot coding method does not give contented results and computation cost (time and space) after extension of window size is not proportional to the improvement of performance; hence in this section, one-hot coding method is replaced by biological feature sets as shown in **Table 3.2.b** and **Table 3.2.c**. Data set focus on four bulk element (calcium, potassium, magnesium and sodium) subsets with less than 25% sequence identity and sliding window size is 15. The comparison between different feature sets is listed in **Table 4.3.a**. For simplicity, only Q-observed and Q-predicted values are listed in the table.

| Feature set | Element | TP | TN | FP | FN | Q-observed | Q-predicted |
|---|---|---|---|---|---|---|---|
| 2nd | Ca | 160 | 47471 | 25 | 435 | 26.89% | 86.49% |
| | K | 61 | 13054 | 0 | 6 | 91.04% | 100.00% |
| | Mg | 100 | 53897 | 21 | 206 | 32.68% | 82.64% |
| | Na | 99 | 19311 | 4 | 47 | 67.81% | 96.12% |
| Phy | Ca | 0 | 47496 | 0 | 595 | 0.00% | n/a |
| | K | 0 | 13054 | 0 | 67 | 0.00% | n/a |
| | Mg | 0 | 53918 | 0 | 306 | 0.00% | n/a |
| | Na | 0 | 19315 | 9 | 146 | 0.00% | 0.00% |
| SEA | Ca | 2 | 47496 | 0 | 593 | 0.34% | 100.00% |
| | K | 0 | 13054 | 0 | 67 | 0.00% | n/a |
| | Mg | 0 | 53918 | 0 | 306 | 0.00% | n/a |
| | Na | 1 | 19315 | 9 | 145 | 0.68% | 10.00% |
| HP | Ca | 120 | 47491 | 5 | 475 | 20.17% | 96.00% |
| | K | 67 | 13054 | 0 | 0 | 100.00% | 100.00% |
| | Mg | 67 | 53895 | 23 | 239 | 21.90% | 74.44% |
| | Na | 84 | 19314 | 1 | 62 | 57.53% | 98.82% |
| CC | Ca | 594 | 47496 | 0 | 1 | 99.83% | 100.00% |
| | K | 67 | 13054 | 0 | 0 | 100.00% | 100.00% |
| | Mg | 306 | 53918 | 0 | 0 | 100.00% | 100.00% |
| | Na | 146 | 19315 | 0 | 0 | 100.00% | 100.00% |
| OneHot | Ca | 110 | 47495 | 1 | 485 | 18.49% | 99.10% |
| | K | 25 | 13054 | 0 | 42 | 37.31% | 100.00% |
| | Mg | 43 | 53918 | 0 | 263 | 14.05% | 100.00% |
| | Na | 28 | 19315 | 0 | 118 | 19.18% | 100.00% |

**Table 4.3.a** Comparison of one-hot coding and 5 biological sets in bulk elements

By comparing the Q-observed, physical and solvent exposed area feature sets do not work well in discrimination of metal-binding and non-metal-binding residues, even worst than direct one-hot coding method. Other three biological feature sets (secondary structure propensity, hydrophobicity scales and chemical classification) get better performance than one-hot coding.

These results reflect and correspond to the characteristics of metal-binding chelates, a three dimension cave for metal ion to “reside” in protein and it also can be interpreted as that the formation of metal-binding chelate is highly related to the secondary structure tendency, degree of hydrophobicity and chemical classification of neighboring amino acids of which the entire protein molecule is composed. It is also apparent that metal-binding phenomena don’t be dominated by the physical features of surrounding amino acids only before these experiments began. However, the results in this section have proved this idea true and show that solvent exposed area is not quite highly related to the formation of metal-binding chelates in protein.

**Figure 4.3.b** and **Figure 4.3.c** show the comparison between different feature sets in Q-observed and Q-predicted. The major and significant difference of different feature sets is Q-observed as mentioned before. “Chemical Classifications” of amino acids indeed performs better than other feature sets in metal-binding residue prediction when compare their Q-observed together. **Figure 4.3.d** shows the growth and trend of Q-observed curve with training time for 6 different feature sets (5 biological feature sets and one-hot coding).

From section 4.2 and 4.3, it is clear that biological insight indeed play an important role in prediction the biochemical phenomena in nature. Although one-hot coding is straight-forward idea in feature encoding of 20 amino, it can not completely represent the behavior and characteristics of metal-binding in protein. After these verbose experiments in this thesis, eventually a direct metal-binding prediction method is proposed and proven to be useful and absolutely accurate in proteins binding four bulk elements under 5 fold cross validation.

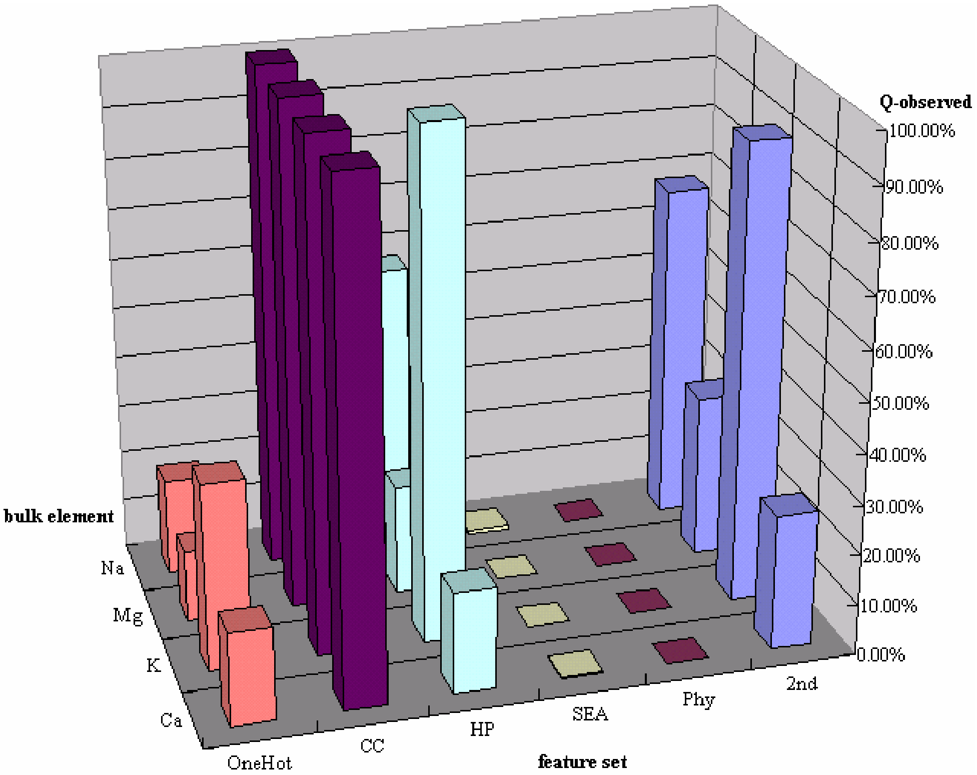


**Fig 4.3.b** Q-observed comparison between different feature sets and bulk elements

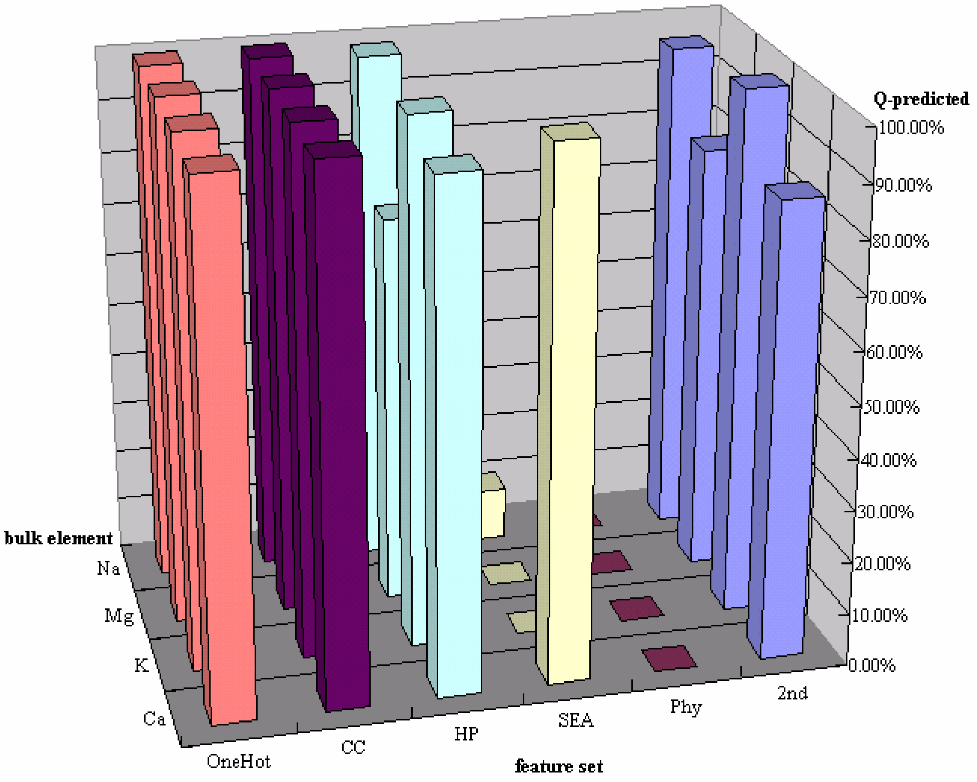


**Fig 4.3.c** Q-predicted comparison between different feature sets and bulk elements

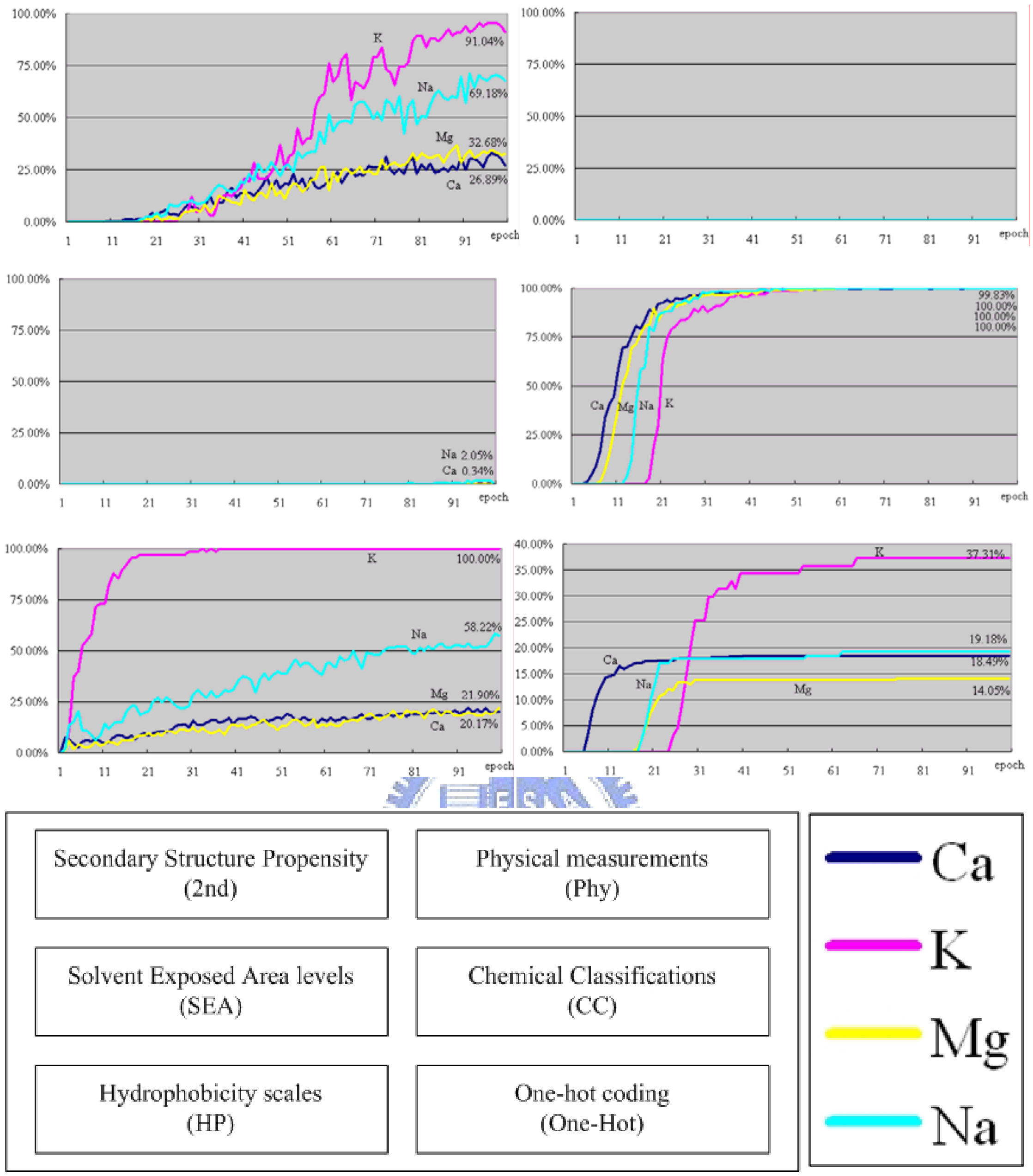


**Fig 4.3.d** Q-predicted versus training epoch between different feature sets

# References

[1] C. H. Wu and J. W. McLarty, *Neural Networks and Genome Informatics*, Elsevier Science Ltd, UK, pp. 67-86, 2000.

[2] C. T. Lin and C. S. George Lee, *Neural Fuzzy Systems,* Prentice-Hall, Inc. N.J., U.S.A. 1996.

[3] C. Branden and J. Tooze, *Introduction to Protein Structure*, $2^{nd}$ edition, Garland Publishing, Inc., New York, pp. 205-220, 1999.

[4] M. J. Kendrick, M. T. May, M. J. Plishka, and K. D. Robinson, *Metals in Biological System*, Ellis Horwood Limited, England, pp. 11-48, 1992.

[5] R. A. Copeland, *Enzymes A Practical Introduction to Structure, Mechanism and Data Analysis*, $2^{nd}$ edition, *Wiley-VHC, Inc*, Canada, pp. 42-74, 2000.

[6] J. M. Castagnetto, S. W. Hennessy, V. A. Roberts, E. D. Getzoff, J. A. Tainer and M.E. Pique, "MDB: the Metalloprotein Database and Browser at The Scripps Research Institute", *Nucleic Acids Res.* ,Vol. 30, No.1 , pp.379-382, 2002.

[7] W. R. Taylor, "The Classification of Amino Acid Conservation", *J. Theor. Biol.*, Vol.119, pp. 205-218, 1986.

[8] D. Bordo and P. Argos, "Suggestions for **Safe** Residue Substitutions in Site-Directed Mutagensis", *J. Mol. Biol.* Vol.217, pp. 721-729, 1991.

[9] D. M. Engelman, T. A. Steitz, and A. Goldman, "Identifying nonpolar transbilayer helices inamino acid sequences of membrane proteins", *Annu. Rev. Biophys. Biophys. Chem.* Vol.15, pp. 321-353, 1986.

[10] T. P. Hoop and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences". *Proc Natl Acad Sci*, Vol.78, pp.3824, 1981.

[11] J. Kyte and R. Doolit, "A Simple Method for Displaying the Hydropathic Character of a Protein", *J. Mol Biol.* Vol.157, pp.105-132, 1982.

[12] J. Janin, "Surface and Inside Volumes in Globular Proteins", *Nature*, Vol. 277, pp.491-492, 1979.

[13] C. Chothia, "Hydrophobic bonding and accessible surface area in proteins", *Nature*, Vol.248, pp.338-339, 1974.

[14] Eisenberg D., Weiss R.M., Terwilliger C.T., Wilcox W., 1982. Hydrophobic moments and protein structure, Faraday Symp. Chem. Soc. 17:109-120.

[15] S.Dietmann and C. Frommel, "Prediction of 3D neighbours of molecular surface patches in proteins by artificial neural networks", *Bioinformatics*, Vol. 18, No.1, pp. 167-174, 2002.

[16] E. Roulet, P. Bucher, R. Schneider, E. Wingender, Y. Dusserre, T. Werner and N. Mermod, "Experimental Analysis and Computer Prediction of CTF/NFI Transcription Factor DNA Binding Sites", *J. Mol. Biol.*, Vol. 297, pp. 833-848, 2000.

[17] I. Jonassen, I. Eidhammer, D. Conklin and W. R. Taylor, "Structure motif discovery and mining the PDB", *Bioinformatics*, Vol. 18, No. 2, pp. 362-367, 2001.

[18] M. Shah, S. Passovets, D. Kim, K. Ellrott, L. Wang, I. Vokler, P. L. Cascio, D. Xu and Y. Xu, "A Computational pipeline for protein structure prediction and analysis at genome scale", *Bioinformatics*, Vol. 19, No. 15, pp. 1985-1996, 2003.

[19] M. Cline, R. Hughey and K. Karplus, "Predicting reliable regions in protein sequence alignments", *Bioinformatics*, Vol. 18, No. 2, pp. 306-314, 2002.

# Appendix

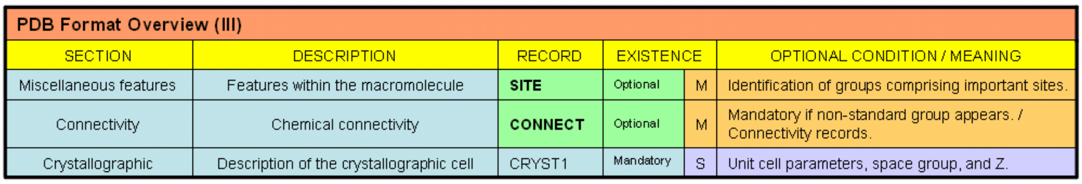
## [A] MySQL, Apache and PHP

The environment of experiments is set up on X86 computer with Microsoft Windows XP OS. User can install all these components ([8]MySQL database, Apache web server, and [9]PHP webpage preprocessor) individually or use integrated tool kit — *Foxserv* (http://www.foxserv.net/portal.php) to easily set them ready at once on X86 machine with Microsoft windows or Linux.

## [B] PDB File Format

The full document is available on PDB website and current version is 2.2 (20 December, 1996). Here the document is condensed as tabular representation as shown as follows. There are totally 10 sections shown in **Table B.4** in current version, but there are 12 sections in Table B.1 ~ 3 owing to the mergence of sections. Title and Remark sections are combined into Title section. Crystallographic and Coordinate Transformation sections are joined into one section.

In Table B.1 ~ 3, each section contains types several records and the field "EXISTENCE" indicates that record exists mandatorily or optionally and record type. There are 6 record types (**S**ingle, **S**ingle **C**ontinued, **M**ultiple, **M**ultiple **C**ontinued, **G**rouping, and **O**ther). Their differences are shown in **Table B.5**.



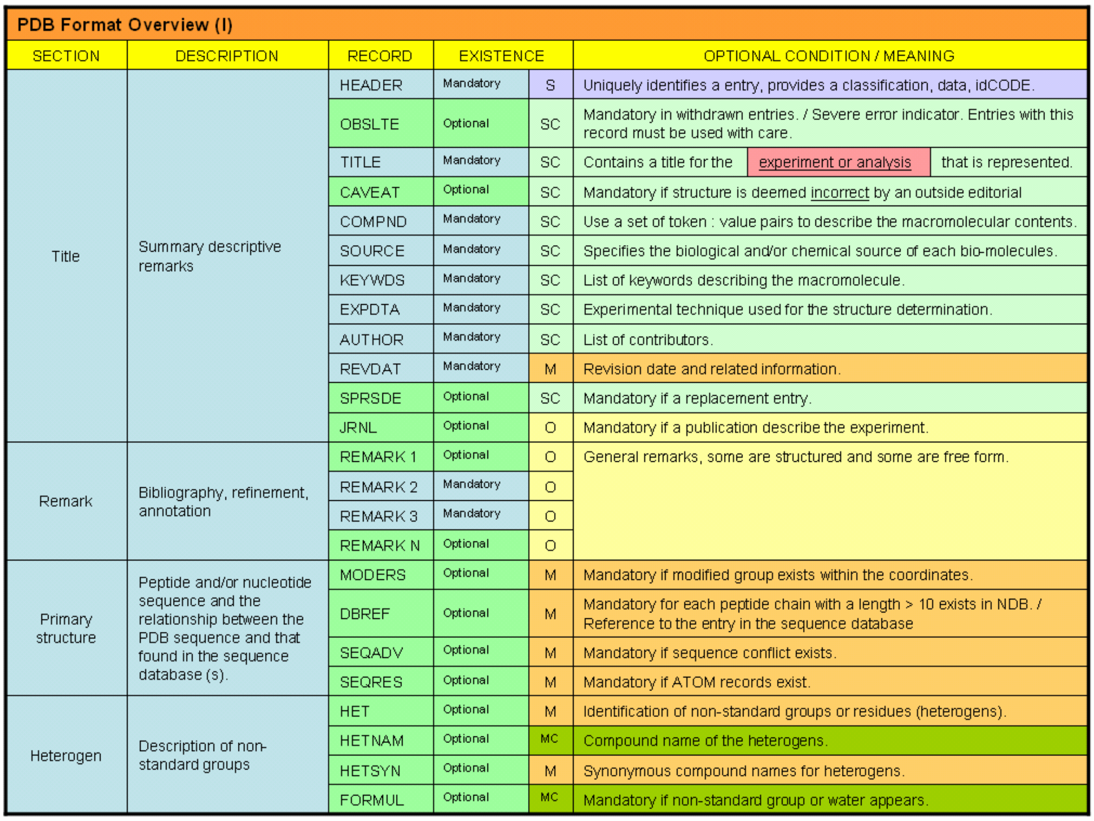
PDB Format Overview (III)

| SECTION | DESCRIPTION | RECORD | EXISTENCE | | OPTIONAL CONDITION / MEANING |
|---|---|---|---|---|---|
| Miscellaneous features | Features within the macromolecule | **SITE** | Optional | M | Identification of groups comprising important sites. |
| Connectivity | Chemical connectivity | **CONNECT** | Optional | M | Mandatory if non-standard group appears. / Connectivity records. |
| Crystallographic | Description of the crystallographic cell | CRYST1 | Mandatory | S | Unit cell parameters, space group, and Z. |

**Table B.3** PDB file format overview part 3

[8] an world-wide open source database system, http://www.mysql.com/

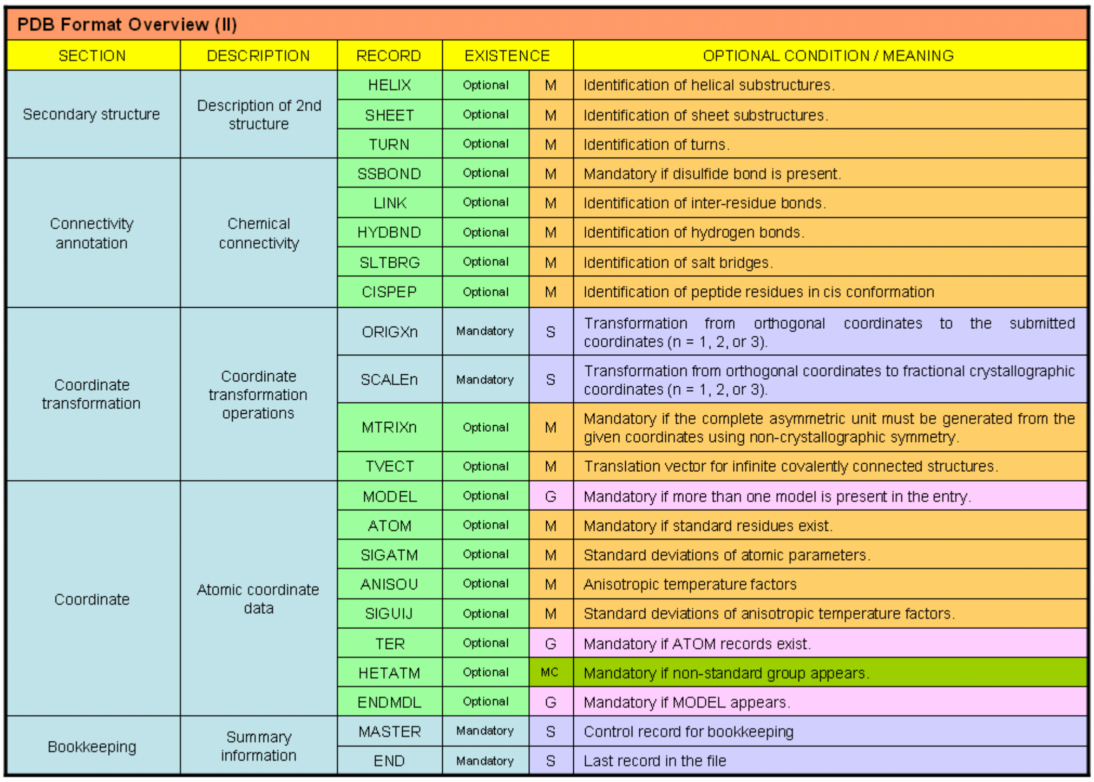
[9] cross-platform server-side scripting language used to create dynamic web pages, http://www.php.net



PDB Format Overview (I)

| SECTION | DESCRIPTION | RECORD | EXISTENCE | | OPTIONAL CONDITION / MEANING |
|---|---|---|---|---|---|
| Title | Summary descriptive remarks | HEADER | Mandatory | S | Uniquely identifies a entry, provides a classification, data, idCODE. |
| | | OBSLTE | Optional | SC | Mandatory in withdrawn entries. / Severe error indicator. Entries with this record must be used with care. |
| | | TITLE | Mandatory | SC | Contains a title for the experiment or analysis that is represented. |
| | | CAVEAT | Optional | SC | Mandatory if structure is deemed incorrect by an outside editorial |
| | | COMPND | Mandatory | SC | Use a set of token : value pairs to describe the macromolecular contents. |
| | | SOURCE | Mandatory | SC | Specifies the biological and/or chemical source of each bio-molecules. |
| | | KEYWDS | Mandatory | SC | List of keywords describing the macromolecule. |
| | | EXPDTA | Mandatory | SC | Experimental technique used for the structure determination. |
| | | AUTHOR | Mandatory | SC | List of contributors. |
| | | REVDAT | Mandatory | M | Revision date and related information. |
| | | SPRSDE | Optional | SC | Mandatory if a replacement entry. |
| | | JRNL | Optional | O | Mandatory if a publication describe the experiment. |
| Remark | Bibliography, refinement, annotation | REMARK 1 | Optional | O | General remarks, some are structured and some are free form. |
| | | REMARK 2 | Mandatory | O | |
| | | REMARK 3 | Mandatory | O | |
| | | REMARK N | Optional | O | |
| Primary structure | Peptide and/or nucleotide sequence and the relationship between the PDB sequence and that found in the sequence database (s). | MODERS | Optional | M | Mandatory if modified group exists within the coordinates. |
| | | DBREF | Optional | M | Mandatory for each peptide chain with a length > 10 exists in NDB. / Reference to the entry in the sequence database |
| | | SEQADV | Optional | M | Mandatory if sequence conflict exists. |
| | | SEQRES | Optional | M | Mandatory if ATOM records exist. |
| Heterogen | Description of non-standard groups | HET | Optional | M | Identification of non-standard groups or residues (heterogens). |
| | | HETNAM | Optional | MC | Compound name of the heterogens. |
| | | HETSYN | Optional | M | Synonymous compound names for heterogens. |
| | | FORMUL | Optional | MC | Mandatory if non-standard group or water appears. |

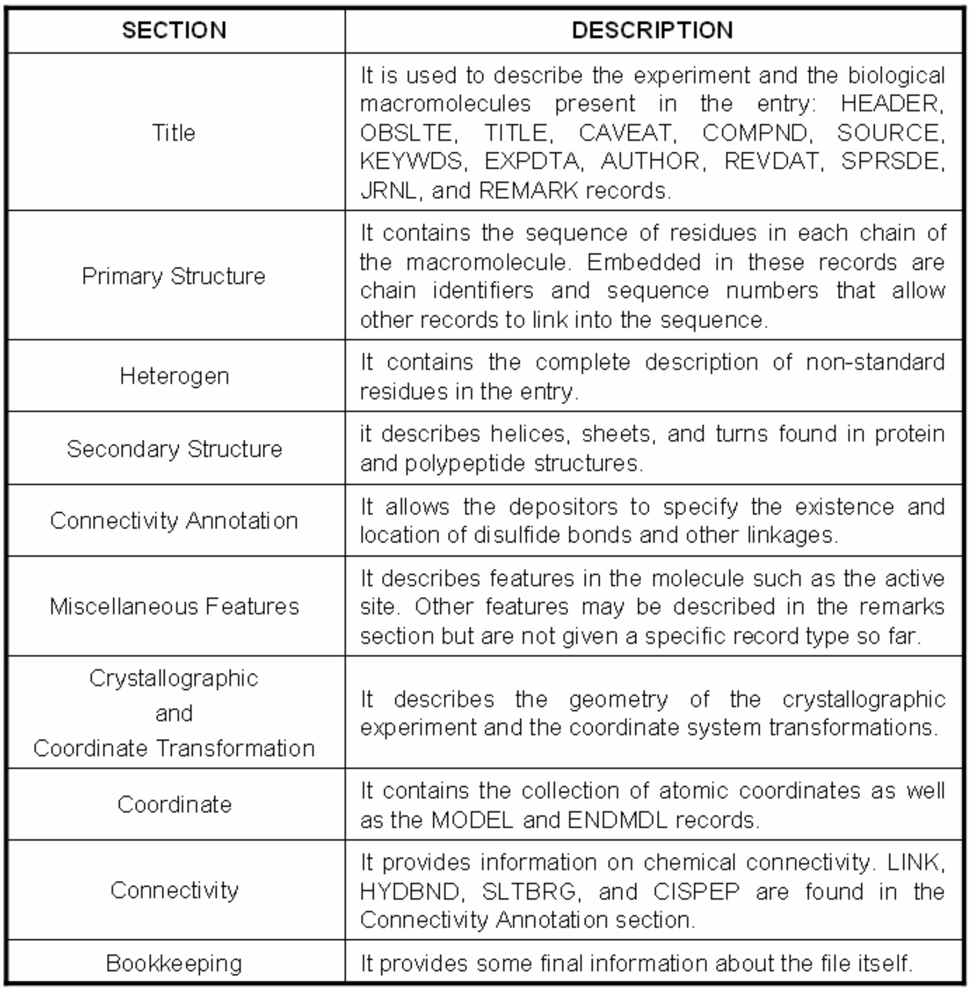
**Table B.1** PDB file format overview part 1



PDB Format Overview (II)

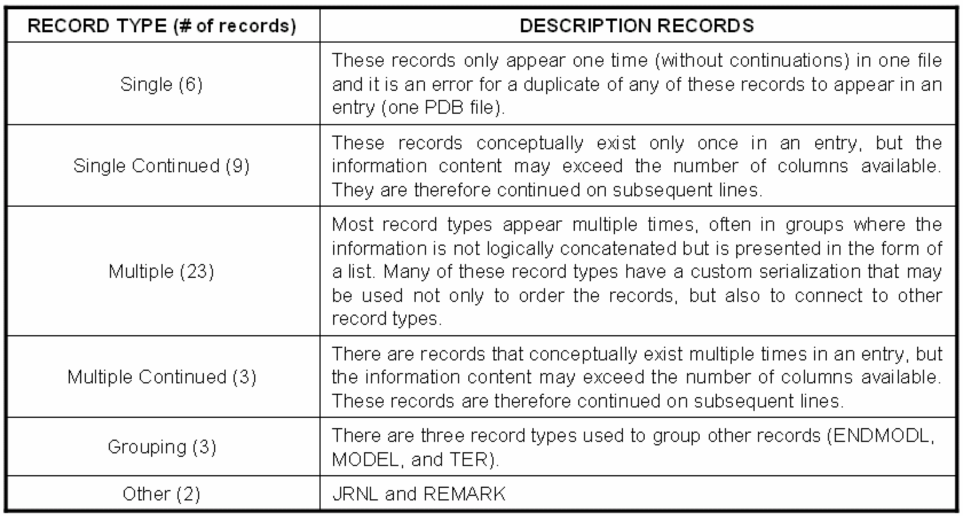
| SECTION | DESCRIPTION | RECORD | EXISTENCE | | OPTIONAL CONDITION / MEANING |
|---|---|---|---|---|---|
| Secondary structure | Description of 2nd structure | HELIX | Optional | M | Identification of helical substructures. |
| | | SHEET | Optional | M | Identification of sheet substructures. |
| | | TURN | Optional | M | Identification of turns. |
| Connectivity annotation | Chemical connectivity | SSBOND | Optional | M | Mandatory if disulfide bond is present. |
| | | LINK | Optional | M | Identification of inter-residue bonds. |
| | | HYDBND | Optional | M | Identification of hydrogen bonds. |
| | | SLTBRG | Optional | M | Identification of salt bridges. |
| | | CISPEP | Optional | M | Identification of peptide residues in cis conformation |
| Coordinate transformation | Coordinate transformation operations | ORIGXn | Mandatory | S | Transformation from orthogonal coordinates to the submitted coordinates (n = 1, 2, or 3). |
| | | SCALEn | Mandatory | S | Transformation from orthogonal coordinates to fractional crystallographic coordinates (n = 1, 2, or 3). |
| | | MTRIXn | Optional | M | Mandatory if the complete asymmetric unit must be generated from the given coordinates using non-crystallographic symmetry. |
| | | TVECT | Optional | M | Translation vector for infinite covalently connected structures. |
| Coordinate | Atomic coordinate data | MODEL | Optional | G | Mandatory if more than one model is present in the entry. |
| | | ATOM | Optional | M | Mandatory if standard residues exist. |
| | | SIGATM | Optional | M | Standard deviations of atomic parameters. |
| | | ANISOU | Optional | M | Anisotropic temperature factors |
| | | SIGUIJ | Optional | M | Standard deviations of anisotropic temperature factors. |
| | | TER | Optional | G | Mandatory if ATOM records exist. |
| | | HETATM | Optional | MC | Mandatory if non-standard group appears. |
| | | ENDMDL | Optional | G | Mandatory if MODEL appears. |
| Bookkeeping | Summary information | MASTER | Mandatory | S | Control record for bookkeeping |
| | | END | Mandatory | S | Last record in the file |

**Table B.2** PDB file format overview part 2



| SECTION | DESCRIPTION |
|---|---|
| Title | It is used to describe the experiment and the biological macromolecules present in the entry: HEADER, OBSLTE, TITLE, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, AUTHOR, REVDAT, SPRSDE, JRNL, and REMARK records. |
| Primary Structure | It contains the sequence of residues in each chain of the macromolecule. Embedded in these records are chain identifiers and sequence numbers that allow other records to link into the sequence. |
| Heterogen | It contains the complete description of non-standard residues in the entry. |
| Secondary Structure | it describes helices, sheets, and turns found in protein and polypeptide structures. |
| Connectivity Annotation | It allows the depositors to specify the existence and location of disulfide bonds and other linkages. |
| Miscellaneous Features | It describes features in the molecule such as the active site. Other features may be described in the remarks section but are not given a specific record type so far. |
| Crystallographic<br>and<br>Coordinate Transformation | It describes the geometry of the crystallographic experiment and the coordinate system transformations. |
| Coordinate | It contains the collection of atomic coordinates as well as the MODEL and ENDMDL records. |
| Connectivity | It provides information on chemical connectivity. LINK, HYDBND, SLTBRG, and CISPEP are found in the Connectivity Annotation section. |
| Bookkeeping | It provides some final information about the file itself. |

**Table B.4** Sections in PDB file



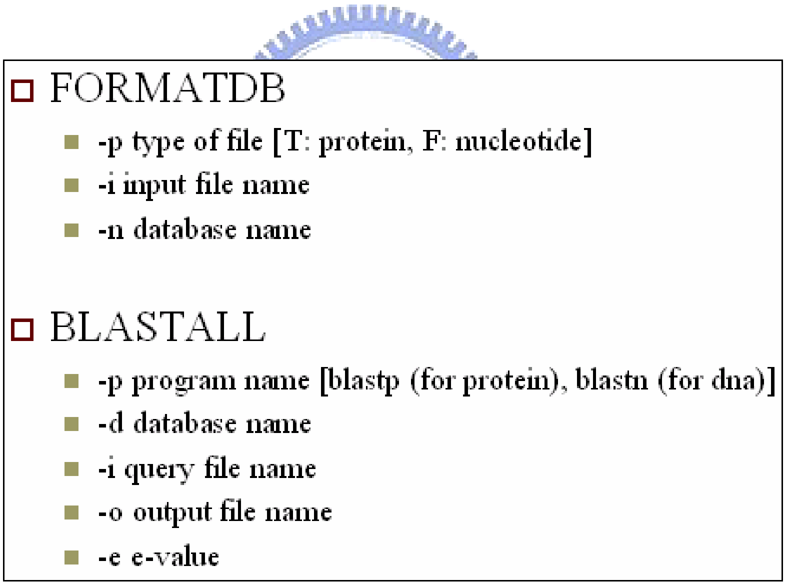
| RECORD TYPE (# of records) | DESCRIPTION RECORDS |
|---|---|
| Single (6) | These records only appear one time (without continuations) in one file and it is an error for a duplicate of any of these records to appear in an entry (one PDB file). |
| Single Continued (9) | These records conceptually exist only once in an entry, but the information content may exceed the number of columns available. They are therefore continued on subsequent lines. |
| Multiple (23) | Most record types appear multiple times, often in groups where the information is not logically concatenated but is presented in the form of a list. Many of these record types have a custom serialization that may be used not only to order the records, but also to connect to other record types. |
| Multiple Continued (3) | There are records that conceptually exist multiple times in an entry, but the information content may exceed the number of columns available. These records are therefore continued on subsequent lines. |
| Grouping (3) | There are three record types used to group other records (ENDMODL, MODEL, and TER). |
| Other (2) | JRNL and REMARK |

**Table B.5** Record types in PDB file

# [C] Clustalw and Blast

Here illustrate several important commands about these multiple sequence alignment tools in terminal mode when sequence sampling. Usually you can download “GUI” version from internet but it needs step by step to click buttons on it so as to complete your task. As the result, this section tends to give a practical guide about how to work on batch mode when you use these tools.

When you download the “terminal” version of these tools (always with their source code), you can set it up on various of machines or OS which has C language complier, such as free gcc, g++ from GNU or other commercial compliers. If you use PC with window OS, you can compile it on window command mode environment. If you are work station user, you don’t need to worry about the purchase of complier and environment.



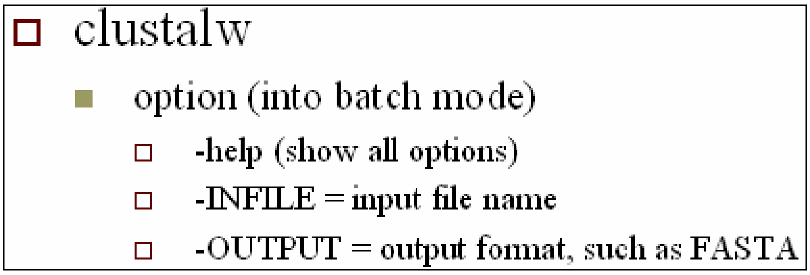
□ FORMATDB

- ■ -p type of file [T: protein, F: nucleotide]
- ■ -i input file name
- ■ -n database name

□ BLASTALL

- ■ -p program name [blastp (for protein), blastn (for dna)]
- ■ -d database name
- ■ -i query file name
- ■ -o output file name
- ■ -e e-value

**Table C.1** Important commands in BLAST package



□ clustalw

- ■ option (into batch mode)
  - □ -help (show all options)
  - □ -INFILE = input file name
  - □ -OUTPUT = output format, such as FASTA

**Table C.2** Important commands in clustalw package

**Command Line Parameters**

| Classification | | Usage | Meaning |
|---|---|---|---|
| DATA (sequence) | | -INFILE=input.txt | input sequence file |
| | | -PROFILE1=prof1.txt | profile 1 |
| | | -PROFILE2=prof2.txt | profile 2 |
| VERBS (do things) | | -OPTIONS | list the command line parameters |
| | | -HELP or -CHECK | outline the command line params. |
| | | -ALIGN | do full multiple alignment |
| | | -TREE | calculate NJ tree. |
| | | -BOOTSTRAP(=n) | bootstrap a NJ tree (n= number of bootstraps; def. = 1000). |
| | | -CONVERT | output the input sequences in a different file format. |
| PARAMETERS (set things) | General Setting | -INTERACTIVE | read command line, then enter normal interactive menus |
| | | -QUICKTREE | use FAST algorithm for the alignment guide tree |
| | | -TYPE= | PROTEIN or DNA sequences |
| | | -NEGATIVE | protein alignment with negative values in matrix |
| | | -OUTFILE= | sequence alignment file name |
| | | -OUTPUT= | GCG, GDE, PHYLIP, PIR or NEXUS |
| | | -OUTORDER= | INPUT or ALIGNED |
| | | -CASE | LOWER or UPPER (for GDE output only) |
| | | -SEQNOS= | OFF or ON (for Clustal output only) |
| | | -SEQNO_RANGE= | OFF or ON (NEW: for all output formats) |
| | | -RANGE=m,n | sequence range to write starting m to m+n. |
| | Fast Pariwise Alignment | -KTUPLE=n | word size |
| | | -TOPDIAGS=n | number of best diags. |
| | | -WINDOW=n | window around best diags. |
| | | -PAIRGAP=n | gap penalty |
| | | -SCORE | PERCENT or ABSOLUTE |
| | Slow Pariwise Alignment | -PWMATRIX= | Protein weight matrix=BLOSUM, PAM, GONNET, ID or filename |
| | | -PWDNAMATRIX= | DNA weight matrix=IUB, CLUSTALW or filename |
| | | -PWGAPOPEN=f | gap opening penalty |
| | | -PWGAPEXT=f | gap opening penalty |
| | Multiple Alignment | -NEWTREE= | file for new guide tree |
| | | -USETREE= | file for old guide tree |
| | | -MATRIX= | Protein weight matrix=BLOSUM, PAM, GONNET, ID or filename |
| | | -DNAMATRIX= | DNA weight matrix=IUB, CLUSTALW or filename |
| | | -GAPOPEN=f | gap opening penalty |
| | | -GAPEXT=f | gap extension penalty |
| | | -ENDGAPS | no end gap separation pen. |
| | | -GAPDIST=n | gap separation pen. range |
| | | -NOPGAP | residue-specific gaps off |
| | | -NOHGAP | hydrophilic gaps off |
| | | -HGAPRESIDUES= | list hydrophilic res. |
| | | -MAXDIV=n | % ident. for delay |
| | | -TYPE= | PROTEIN or DNA |
| | | -TRANSWEIGHT=f | transitions weighting |
| | Profile Alignment | -PROFILE | Merge two alignments by profile alignment |
| | | -NEWTREE1= | file for new guide tree for profile1 |
| | | -NEWTREE2= | file for new guide tree for profile2 |
| | | -USETREE1= | file for old guide tree for profile1 |
| | | -USETREE2= | file for old guide tree for profile2 |
| | Seq-Profile Alignment | -SEQUENCES | Sequentially add profile2 sequences to profile1 alignment |
| | | -NEWTREE= | file for new guide tree |
| | | -USETREE= | file for old guide tree |
| | Structure Alignment | -NOSECSTR1 | do not use secondary structure-gap penalty mask for profile 1 |
| | | -NOSECSTR2 | do not use secondary structure-gap penalty mask for profile 2 |
| | | -SECSTROUT={} | {STRUCTURE or MASK or BOTH or NONE} output in alignment file |
| | | -HELIXGAP=n | gap penalty for helix core residues |
| | | -STRANDGAP=n | gap penalty for strand core residues |
| | | -LOOPGAP=n | gap penalty for loop regions |
| | | -TERMINALGAP=n | gap penalty for structure termini |
| | | -HELIXENDIN=n | number of residues inside helix to be treated as terminal |
| | | -HELIXENDOUT=n | number of residues outside helix to be treated as terminal |
| | | -STRANDENDIN=n | number of residues inside strand to be treated as terminal |
| | | -STRANDENDOUT=n | number of residues outside strand to be treated as terminal |
| | Trees | -OUTPUTTREE={} | {nj OR phylip OR dist OR nexus} |
| | | -SEED=n | seed number for bootstraps. |
| | | -KIMURA | use Kimura's correction. |
| | | -TOSSGAPS | ignore positions with gaps. |
| | | -BOOTLABELS={} | {node OR branch} position of bootstrap values in tree display |

**Table C.3** Full commands of clustalw package