國立交通大學

生物資訊研究所

碩士論文

利用蛋白質結構預測蛋白質內的重要功能位置

Prediction of functional sites of proteins from protein structures

研究生：于松桓

指導教授：黃鎮剛　教授

中華民國九十七年六月

利用蛋白質結構預測蛋白質內的重要功能位置

Prediction of functional sites of proteins from protein structures

研 究 生：于松桓　　Student：Sung-Huan Yu

指導教授：黃鎮剛　　Advisor：Jenn-Kang Hwang

國 立 交 通 大 學

生物資訊研究所

碩 士 論 文

A Thesis

Submitted to Department of Computer and Information Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Bioinformatics

June 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年六月

利用蛋白質結構預測蛋白質內的重要功能位置

學生：于松桓　　指導教授：黃鎮剛

國立交通大學生物資訊研究所碩士班

摘　　要

第一章－活性位置（active site）

由於現在結構基因組學（structural genomics）的研究以驚人的速度發展，相當多的蛋白質結構已被解出並存放於蛋白質資料銀行（Protein Data Bank – PDB）這個資料庫中。因著前面說到的情形，逐漸出現許多不知道功能的蛋白質，而發展利用蛋白質結構直接預測蛋白質內活性位置的方法也變得日漸重要。有許多特性與蛋白質的活性位置有關聯，例如：越密集的區域（higher packing density）、越靠近蛋白質幾何中心（structural centrality）、熱擾動（thermal fluctuations）越低的殘基（residues），越有可能是活性位置，根據這些特性我們發展出一個簡單的方法來預測蛋白質的活性位置。若我們給予這些方法所計算出來的結果一個合適的閾值（threshold），我們可以在 760 個非同源性酵素（nonhomologous enzyme）中預測到 76%的活性位置，並且只有 27%的假陽性（false positive）。倘若我們加入蛋白質序列（sequence）的資訊，用此資訊來加權原來的資料，可以預測到 80%活性位置，只有 20%的假陽性。我們的方法不需要序列或結構的比對（alignment），或利用結構模版庫（structural template library），此方法也避免了繁雜的溶劑表面易溶性（solvent accessible surface）和分子力學（molecular mechanical）的計算。 我們相信我們的方法會是一個預測蛋白質活性位置相當有用的方法，並且比其他的方法還要完整。

第二章－金屬離子鍵結位置（metal binding site）

金屬離子在生物體中扮演相當重要的角色，例如：幫助酵素催化、調節生物體內機能、提高結構穩定性等。由於目前蛋白質結構快速增加的現代，預測蛋白質內金屬離子的鍵結位置也就日趨重要。我們知道若是要讓金屬離子穩定的存在在蛋白質中，必須產生螯合物（chelate）。而要形成螯合物其中有一個因素非常重要，就是金屬離子周圍必須有足夠的原子與它產生配位（coordinate）。這個特性非常類似我們第一章提到的依賴距離之接觸點數（distance-dependent protein contact-number 簡稱 CN）的模型，即指明若有許多能夠與金屬離子反應的原子在一個殘基的周圍，此殘基就極有可能是金屬離子鍵結位置。一般來說，會與金屬離子產生螯合物的原子為－氮（N）、硫（S）、氧（O）。根據這個想法，我們利用 CN 模型的想法，但是將 Cα換成像氮（N）、硫（S）、氧（O）的原子，用此方法來預測金屬離子鍵結位置。此方法可以在 Sodhi 的資料組中正確預測 72.4%鈣離子、94.7%銅離子、86.5%鐵離子、77.6%鎂離子、88.5%錳離子和 91.5%鋅離子的鍵結位置。

Prediction of functional sites of proteins from protein structures

Student: Sung-Huan Yu Advisor: Jenn-Kang Hwang

Institute of Bioinformatics

National Chiao Tung University

ABSTRACT

Chapter 1 – active site

Due to the tremendous advances in structural genomics research, an incredible number of protein structures has been solved and deposited in PDB. As a result, the number of structures with unknown function also climbs up accordingly. It becomes increasingly important that one can predict functional sites directly from protein structures. Based on the distinct properties associated with the active-site residues such as higher packing density, proximity to structural centrality and smaller thermal fluctuations, we developed a simple method for detection of the active sites of enzymes to compute profiles based on the aforementioned properties. Using proper threshold values for the profiles, we are able to detect up to 76% of catalytic residues with 27% of false positives for a data set comprising 760 nonhomologous enzymes. If additional sequence information is included, the sequence-weighed profile method can be improved to detect 80% of catalytic residues with 20% of false positives. Our method does not require sequence or structural alignment, or a structural template library, and it avoids solvent accessible surface or molecular mechanical calculations. We believe that our method will be a useful tool for detection of possible active sites from protein structures to complement other existing methods.

Chapter 2 – metal binding site

Metal ions are crucial role in organisms. They participate in enzyme catalysis, play regulatory roles, and help maintain protein structure. In this era, there is incredible number of protein structures solved. So, the importance of predicting metal binding site is increased. We all know that if there are metal ions stable existed in protein, the metal ions should form chelate. One of the important factors to form chelate is there should be enough atoms to coordinate with metal ion. The characteristic is very similar as distance-dependent protein contact-number model (CN) that we introduced in chapter 1. This means that if there are more atoms that are high probability to interact with metal ion around the residue, that would be probably metal binding residue. In general, the atoms that have high probability to interact with metal are such as N, S, O. Base on the thought, we follow the aspect of CN but use the atoms, like N, S, O, to replaced Cα to predict metal binding residues. This method can detect Ca – 72.4%, Cu – 94.7%, Fe – 86.5%, Mg – 77.6%, Mn – 88.5%, and Zn – 91.5% in Sodhi's dataset.

誌　　謝

我謹將此論文獻給再這兩年一直祝福我的主耶穌，是他的祝福使我可以如此順利的完成我的碩士論文，也感謝我的家人一直支持我、幫助我。謝謝黃鎮剛教授，在我初踏進生資領域時，耐心的指導我如何做研究，告訴我如何解決問題，他對科學的熱情與執著是我學習的目標。謝謝實驗室的夥伴，少偉、建華、志鵬、志豪，在研究上給予我極大的協助，沒有你們我也無法完成論文。啟文、存操、書偉、志杰、涵堃、景盛，當我有問題時熱心且有耐心的幫我一起尋找答案。士中、彥龍，我們一同完成了我碩士班的第一個計畫，讓我對生資這個領域有了興趣與信心。瓊文、仙蕾，我們一同進入實驗室，與你們一同學習一同討論，讓我很快可以適應環境。慧雯、子琳、勇欣、人維，在生活中、研究室中，帶給我很多歡樂，你們認真的學習也讓我很受鼓勵。也謝謝新竹市召會和頭份、造橋召會的所有弟兄姊妹，你們背後的代禱，給我力量。特別是弟兄之家的弟兄們，積福、亦謙、易翰、暐弘、俊元、承彥、侯泰、新和、亦新、新城、亦俠，在我軟弱與到環境與困難時，給我極大的幫助與安慰。也特別謝謝服事者勞苦的服事，謝謝博愛小排和動工小排的成員，在週中可以與你們相聚是極其喜樂的事。謝謝所有我曾經照顧及現在照顧的小羊，你們是我最大的安慰與加力。謝謝以上所以提名的人，也願主更多加強你們祝福你們。

CONTENTS

TABLE CONTENTS

FIGURE CONTENTS

CHAPTER 1 – ACTIVE SITE

1. INTRODUCTION

Due to the enormous advances made in recent years in structural biology, the number of protein structures deposited in Protein Data Bank (PDB) has increased from 13622 in 2000 to around 49620 as of March 11, 2008 – the total number nearly quadrupled during this period. The vast number of structures provides a great opportunity to study the structure-function relationship directly from the protein structures. It becomes especially important nowadays due to an increasing number of structures with unknown function being deposited in PDB. Currently, a number of methods[1; 2; 3; 4; 5; 6; 7; 8; 9], based on the observation that most catalytic site structures are highly conserved between remotely related enzymes, predict protein function by searching protein structures for the known three-dimensional catalytic templates. For example, Thornton and co-workers[5; 6] developed a methodology, utilizing a library of three-dimensional structural templates composed of small number of residues, to detect catalytic sites and ligand binding sites of proteins. Lu *et al.*[4] developed a local fragment transformation method to detect the ligand-binding sites based on a loosely defined structural template. This method is useful for detecting DNA-binding sites, which are usually of highly variable conformations[10; 11; 12]. The effectiveness of these methods depends on whether the pre- defined templates will provide a fairly thorough coverage of the known structures[6]. These methods are unable to detect novel catalytic residue conformation that is not matched by any known structural templates in the library. There are other methods for prediction of protein function based on distinct structural or dynamical properties associated with active-site structures[13; 14; 15; 16]. For example, Amitai *et al.*[13] transformed the protein structure into residue interaction graphs with each amino acid residue represented as a graph node and the interaction between them as a graph edge, from which they compute network closeness of each residue. They were able to identify active site residues in 70% of 178 representative

structures by computing residues' closeness together with their solvent surface accessibility. Ben-Shimon and Eisenstein[14] observed that the catalytic residues are usually located in small fractions of the exposed residues closest to the protein centroid. They developed a novel algorithm called EnSite to detect the active sites of enzymes. EnSite examines only 5% of the exposed surface closest to the centroid, instead of identifying all the cavities or depressions on the enzyme surface. Ensite clusters these surface segments, which are then ranked by their area size for possible active sites. Recently, Sacquin-Mora *et al.*[16] computed the force constants of moving any given amino acid with respect to other residues in the protein. They found that the force constants associated with the catalytic residues are usually higher than those of other non-catalytic residues. Choosing an appropriate threshold value, they are able to detect potential active-site residues using Brownian dynamics simulations. The distinct property of large force constant associated with active-site residues is consistent with the recent reports[17; 18] that the catalytic residues usually have lower B-factors than other non-catalytic residues. Since the B- factor is a measure of the atomic mean-square displacement, a residue with smaller B-factors will be more rigid and, hence, be associated with a larger force constant. There are recent reports[19; 20] that the atom's B-factor is linearly proportional to its squared distance from the protein centroid. In other word, the residues in proximity to the protein centroid will have smaller thermal fluctuations or more rigid than those farther away from the protein centroid. In addition, a recent study[21] shows that the atom's thermal fluctuations is in linear inverse proportion to the number of noncovalent neighboring atoms (or protein contact number) of this atom. Here we will develop some simple methods for catalytic sites to compute the profiles based on the properties like contact number, residue centrality and B-factors.

2. METHODS

2.1. The B-factor profile

The X-ray B-factor profile of a protein is denoted as $\mathbf{b} = (b_1, b_2, \ldots b_N)$, where b_i is the B-factor of the Cα atom of the i^{th} residue taken from the PDB file and N is the number of residues of the protein. We will also normalize the B-factor profiles to the corresponding z-scores: $z_i^b = (b_i - \bar{b})/\sigma_b$, where $\bar{b}$ and σ_b are the mean and standard deviation of the B-factors. We will refer to the normalized B-factor profile as the z^b-profile or the BF profile.

2.2. The protein contact-number profile

2.2.1. The naive contact-number model

The protein contact number is conventionally defined as the number of the neighboring residues that are within a cut-off radius of the central residue, which amounts to giving an equal unitary weight to every contacting atom regardless of its distance to the central atom.

$$c_i = \sum_{j \neq i}^{N} \delta(r_0 - r_{ij}) \qquad (1)$$

where r_{ij} is the distance between Cα atoms of residue i and j, and $\delta(x) = 1$ if $x \geq 0$ and $\delta(x) = 0$ if $x < 0$. The cut-off distance r_0 is usually defined in the range 10 to 12 Å. This definition ignores that an atom at a nearer distance will have a greater effect than the atoms farther sway. For convenience, we will refer to this as the naïve contact-number (nCN) model.

2.2.2. The contact-number model

To take into account the distance factor, we define a distance-dependent contact number n_i by weighing the integral contact number with the factor $1/r_{ij}^2$, which is the distance between Cα atoms of residue i and j:

$$n_i = -\sum_{j \neq i}^{N} \frac{1}{r_{ij}^2} \qquad (2)$$

where N is the total number of the residues of the protein. As in the case of the B-factor profile, we also normalize n_i to its Z-score: $Z_i^n = (n_i - \bar{n})/\sigma_n$, where $\bar{n}$ and σ_n are the mean and the standard deviation of n. Since the contact number is defined as a negative value (Eq. 2), we can directly compare the z_b profile with the z_n profile. For convenience, we will refer to this model as the contact-number (CN) model or the z_n model.

2.2.3. The weighted contact-number model

From the CN model (i.e., Eq. 2), we can further the weighted contact-number (WCN) model.

$$\nu_i = -w_i \sum_{j \neq i}^{N} \frac{1}{r_{ij}^2} \qquad (3)$$

where, given that the residue i is of type a, w_i is calculated by

$$w_i = \log(c_a) + 1 \qquad (4)$$

where c_a is the frequency of a catalytic amino acid of type a. The addition of the constant 1 is for making w_i positive. The weighted contact number ν_i is normalized to $Z_i^{\nu} = (\nu_i - \bar{\nu})/\nu_{\nu}$, where $\bar{\nu}$ and σ_{ν} are the mean and the standard deviation of ν. This weighted contact-number model will be referred to as the z_{ν} model. For illustration, we show schematically the nCN model, the CN model and the WCN model in Figure 1.

2.3. The centroid-model profile

The previous study[20; 21] showed that there is a good correlation between the B-factor B_i and the square of the centroid distance r_i^2,

$$r_i^2 = (\mathbf{r}_i - \mathbf{r}_0) \bullet (\mathbf{r}_i - \mathbf{r}_0) \qquad (5)$$

where $\mathbf{r}_i$ is the coordinate of Cα atom of the i^{th} residue, and $\mathbf{r}_0$ is the centroid of the

protein, i.e., $\mathbf{r}_0 = \sum_i \mathbf{r}_i / N$. We will refer to this model as the centroid model (CM). For easy of comparison, we normalize r_i^2 to $z_i^{r^2} = (r_i^2 - \bar{r}^2)/\sigma_{r^2}$, where $\bar{r}^2$ and σ_{r^2} are the mean and the standard deviation of r^2. The normalized centroid-model (CM) profile will be referred to as the z_{r^2}-profile.

Similar to the WCN model, we will also define an amino-acid weighted centroid distance,

$$\rho_i^2 = w_i(s\Delta r_i^2 - R_{\min}^2) \qquad (6)$$

where $s = (R_{\max}^2 - R_{\min}^2)/(\max\{r_i^2\} - \min\{r_i^2\})$ and $\Delta r_i^2 = r_i^2 - \min\{r_i^2\}$ .The seemingly complicated form of Eq. 6 is to normalize ρ_i^2 to the range between $R_{\min}^2$ and $R_{\max}^2$. Here, $R_{\min}^2$ and $R_{\max}^2$ are set to 0.5 and 2.5. respectively. We will refer to this model as the weighted centroid model (WCM). The normalized WCN profile will be also referred to as the z_{ρ^2}-profile.

2.4. Assessment indices

To evaluate the quality of our predictions, we use the standard definitions of sensitivity and specificity[16]. Sensitivity S_n is defined as the number of correctly predicted functional residues (i.e., true positives or TP) divided by the total number of experimentally defined functional residues (i.e., T). Specificity S_p is defined as the number of correctly predicted non-functional residues (i.e., true negatives or TN) divided by the total number of experimentally defined non-functional residues (i.e., F). The false positive rate is defined as $\alpha = 1 - S_p$, and the false negative rate $\beta = 1 - S_n$.

2.5. Data sets

We selected the structures of enzymes from the Catalytic Site Atlas[22] using `blastclust` from the collection of BLAST tools[23]. The pair sequence identity is set to $\leq 30\%$. The data

set comprises 760 x-ray structures, which include 333 monomeric enzymes and 427 multimeric enzymes. We can also divide the dataset to two sub-datasets – the ligand binding cross chains or not. The ligands no cross chains are 715 in our dataset.

3. RESULTS

3.1. The frequency distribution of amino acid types in catalytic sites

Figure 2 shows the frequency distribution of the 20 amino acid types occurring in the catalytic sites, compared with that of all structures in the data set. The top 5 amino acid types, i.e., D, H, E, R and K that occur in the catalytic sites account for 65% of catalytic residues and all are charged amino acids. The polar amino acids, i.e., C, S, N, Q, T and Y provide 27% of catalytic residues. In all, the charged and the polar amino acids account for around 92% of the catalytic residues, while the rest nonpolar amino acids account for only 8%. These results are similar to the analysis results of a previous report[24] using a smaller data set. We use the information to calculate the weight and listed in Table 1. The weight is calculated through Eq. 4.

3.2. The profiles distributions

Figure 3A shows the comparison with catalytic residues and all residues of z_b, z_n and z_v. The mean of z_b of the catalytic residues is –0.48, while that of all residues is 0.00. Using the t-test, we obtain the p-value $< 2.2 \times 10^{-16}$, which indicates the difference is statistically significant. The catalytic residues tend to be near the negative side of the z_b than the other non-catalytic residues do. For example, 38% of active residues are in the region of $z_b \leq -1$, compared with only 19% of total residues. These results are consistent with the previous reports[17; 18] that the catalytic residues have smaller B-factors than the other non-catalytic residues. The mean of z_{r^2} of the catalytic residues is –1.00 and that of all residues is 0.00. There are even more catalytic residues lying toward the negative side of z_{r^2} -- 66% of active

residues in the region of $z_{r2} \le -1$, compared with 24% of total residues. The mean of z_{ρ^2} of the catalytic residues is –1.54 and that of all residues is 0.00. There are even more catalytic residues lying toward the negative side of z_{ρ^2} -- 77% of active residues in the region of $z_{\rho 2} \le -1$, compared with 18% of total residues. The difference is more obviously than z_{r^2}. Figure 3B compares the distributions of z_b, z_n and z_ν of the catalytic residues and of all residues. The z_b profile already analyses before. The mean of z_n of the catalytic residues is –1.00 and that of all residues is 0.00. There are 70% of active residues in the region of $z_n \le -1$, compared with 23% of total residues. The mean of z_ν of the catalytic residues is –1.53, about half unit of the standard deviation shifted to the left, while that of all residues is 0.00. There are 76% of active residues in the region of $z_\nu \le -1$, compared with 17% of total residues. The difference is more obviously than z_n. Taken together, we found that the catalytic residues tend to bias toward the negative z-scores in these profiles, i.e., they tend to be more rigid, near the centroid of the protein structure and in the more compact region.

3.3. The performances of different models

To discriminate the catalytic residues from the non-catalytic residues, we will determine the optimal cutoff value by minimizing the error function[16] defined as $\varepsilon = \sqrt{(1-S_n)^2 + (1-S_p)^2}$. Figure 4 shows the curves of ε against z-scores of different profile distributions. The optimal z-score cutoff values, at which the corresponding ε is minimal, for the BF profile is –0.5, the CN profile –0.7, the CM profile –0.7, the WCN profile –0.9 and the WCM profile –0.9.

3.4. Prediction of the active residues

We followed Sacquin-Mora's paper[16] to generate ROC curves. Figure 5 compares the ROC curves with different models. Though the BF profiles can distinguish active residues from

non-active residues, it performs much worse than the other 4 models. While the CM performs better than the CN model, the WCM and the WCN model perform similarly. Table 2 compares predictive performances with the five models.

3.5. Examples

3.5.1. Examples of CN and CM

As a typical example, figure 6A shows the profiles of 8-amino-7-oxononanoate synthase (PDB ID: 1bs0)[25]. It has 4 catalytical residues, i.e., H133, E175, D204 and K236. As shown in the z_b profile, the z-scores D204 and K236 are close to the minima, indicating that they are quite rigid. But the z-scores H133 and E175 are higher, indicating that they are relatively flexible. On the other hand, as shown in the z_{r^2} profile, the z-scores of the catalytic residues all coincide with the minima of the profile, indicating that these residues are all close to the centroid position of the protein structure. The z_n profile shows that the catalytic residues are located in the compact regions of the structure. This is consistent with our recent finding[21] that protein centroid region is usually the protein's most compact region. Note that the shape of the z_{r^2} profile of this particular example is very similar to the old z_n profile. The relationship between these 2 types of profiles is in fact a general one, as shown in our previous study[21]. Figure 6B shows the three-dimensional structure of 8-amino-7-oxononanoate synthase with colors ramped according to the z_b, z_{r^2} and z_n profiles, respectively. Another example is given by S-adenosylmethionine decarboxylase (PDB ID: 1jen)[26], which has 5 catalytic residues C82, S229 and H243 (located in chain A), and E11 and E67 (located in chain B). Its profiles are shown in Figure 6C. All of the catalytic residues except E67 on chain B are quite rigid. All of them are close to the centroid position and are buried in the very compact regions. Notice again the similarity between z_{r^2} and z_n profiles in Figure 6C. The three-dimensional structure with colors mapped according to the

z_b , z_{r^2} and z_n profiles, respectively, are shown in Figure 6D. More examples phosphoenolpyruvate carboxylase (PDB ID: 1jqn)[27], Acid beta-glucosidase (PDB ID: 2f61)[28] are shown in Figure 6E-H respectively.

3.5.2. Examples of WCN and WCM

As typical examples, Figure 7 shows the three-dimensional structure of (A) prokaryotic phospholipase A2(1IT4)[29], hydrolyzing the 2-acyl ester bonds of 1,2-diacylalycero-3-phospholipids, has two catalytic residues, His-64 and Asp-85; (B) deoxyribose-5-phosphate aldolase (1P1X) [30] has three catalytic residues: Asp-102, Lys-167 and Lys-201. It is the only known aldolase that uses aldehydes as both aldol donor and accepter molecules in the aldol reaction; (C) ASV integrase (1A5V)[31] has three catalytic residues: Asp-64, Asp-121 and Lys-164. Avian sarcoma virus (ASV) is a retrovirus with many similarities to HIV. Integrase would help the cDNA inserted into the cellular DNA of host to form integrated proviral DNA; (D) rhinovirus 3C protease (1CQQ)[32]. The catalytic residues of the enzyme are His-40, Glu-71 and Cys-147. The 3C proteinase is a cysteine protease with a serine protease-like fold that are responsible for the bulk of polyprotein processing in the Picornaviridae. Most cleavages occur between Gln-Gly peptide bonds.

The colors of these structures are ramped from red (negative z-score) to white (positive z-score) in accord with the z_v and z_{ρ^2} profile, i.e., the residues of the most negative Z-score values are colored on the red end of the red-white spectrum, while the most positive Z-score values are colored on the other end of the red-white spectrum. As shown in the figures, most catalytic residues of these enzymes have more negative z_v and z_{ρ^2} values.

3.6. Comparison with other methods

Sacquin-Mora, Laforet and Lavery (SLL)[16] have recently developed a method for detection of active-site residues to calculate force constants to move any given amino acid residue with respect to the other residues in the protein. Their results indicate that the catalytic residues are usually associated with higher force constants or, equivalently, they are more rigid than other non-catalytic residues. Using Brownian dynamics simulation, they detected 78% of catalytic residues with 26% of false positives for a dataset[18] of 98 nonhomologous enzymes, which covers 6 EC classes: 93 monomeric enzymes and 5 multimeric enzymes. In Table 3, we compare the results of our methods with those of SLL. The results of z_ν and z_{ρ^2} models are significantly better than those of SLL. It is interesting to note that, though both SLL and the z_b model are based on the rigidity of catalytic residues, SLL performs significantly better. In fact, all the profile models based on contact number and centroid distance outperform the z_b, despite that these properties are closely related to the B-factors. This may suggest that the X-ray B-factors are probably not a good measure of atomic rigidity as others.

It is worth noting that, in the cases of ASV integrase (Fig. 4C) and rhinovirus 3C protease (Fig. 4D), their catalytic residues are located on the protein surface. These two proteins are difficult to predict because that the catalytic residues of both these proteins locate on the surface, not within a cleft. SLL can identify one of three (1/3) in the former and two of four (2/4) in the latter, while we can detect all the active sites in these two protein.

Ben-Shimon, and Eisenstein (BE)[14] have recently shown that the active sites residues tend to lie near the protein centroid. They were able to detect around 74% active-site residues for 177 hand annotated enzymes from CSA version 1.0[22]. At the present, the enzymes in CSA have been expanded to 880 hand annotated enzymes and the original dataset is obsolete. We tested z_ν-profile and z_{ρ^2}-profile method on this expanded CSA dataset and we able to detect about 80% active-site residues.

3.7. Ligand cross chains

We all know that not every ligand is interacted with the residues in one chain. There are some ligands which interacted cross two or more chains. So we selected the ligands that cross two or more chains to predict by WCN and WCM models. Table 4 is the result. We found results of using biological unit are better than one chain. Because the ligands that cross two or more chains are almost on the interface. So the functional unit is not one chain. That is why using biological unit is better. We also compared the datasets that include and exclude ligand cross multimer. Table 5 and 6 showed the results are almost the same.

4. DISCUSSION

Based on the distinct properties associated with catalytic residues, we developed a simple profiles based on these properties to discriminate between the catalytic residues and non-catalytic residues. This method is easy to implement and computationally fast -- it needs only a single structure; it does not require sequence alignment or structural template search; and it does not compute solvent accessible surface or perform molecular mechanical calculation. Our method will be useful for prediction of active sites from protein structures. However, it is not clear why these properties (i.e., residue centrality, thermal fluctuations or protein packing density) are related to catalytic sites. Warshel and co-workers[33; 34; 35; 36] have long argued that enzyme catalysis mainly arises from smaller reorganization of the active site residues, i.e., catalytical residues usually maintain similar conformations in both the reactant and the transition states. To lower activation barrier, the enzyme structures are optimized through evolution to partially pre-organize the catalytical residues, thus reducing the reorganization energy required for reaching the transition state. As a result, the catalytic residues tend to be more rigid than other non-catalytic residues. Properties such as B-factors, packing density or residue centrality are all related to residue's rigidity. Interestingly, the profile based on B-factors does perform as well as those based on other properties. It is known

that various experimental factors such such as temperature, crystallization or structural refinement may affect the final B-factor values. Consequently, the B-factor profiles of similar structures may be quite different from each other. For example, Figure 8(A) compares the z_b profiles of 3 X-ray structures of lysozyme with a root-mean-square-deviation of their structures in the range 0.6-0.8 Å. It is clear that their z_b profiles are all indeed very different. For comparison, (B) and (C) shows their z_v and z_{r^2} profiles. The z_v and z_{r^2} profiles overlap each other almost perfectly.

CHAPTER 2 – METAL BINDING SITE

1. INTRODUCTION

Metal ions are crucial for protein function. They participate in enzyme catalysis, play regulatory roles, and help maintain protein structure. Due to the enormous advances made in recent years in structural biology, the number of protein structures deposited in Protein Data Bank (PDB) has increased from 13622 in 2000 to around 49620 as of March 11, 2008 – the total number nearly quadrupled during this period. The vast number of structures provides a great opportunity to study the structure-function relationship directly from the protein structures. The importance is increased more and more for using structure to predict metal binding sites. Several computational methods have been explored for identifying and detecting metal-binding proteins. Some base on sequence searching[37; 38]; some are base on graph or structural information[39; 40; 41], and some combine sequence and structural information[42; 43]. And there are also many methods that predicting metal binding residues by using machine learning[43; 44]. Jaspreet Singh Sodhi *et al.*[43] developed a method called MetSite, represents a fully automatic approach for the detection of metal-binding residue clusters applicable to protein models of moderate quality. MetSite involves using sequence profile information in combination with approximate structural data. Several neural network classifiers are shown to be able to distinguish metal sites from non-sites with a mean accuracy of 94.5%. Kshama Goyal and Shekhar C. Mande[41] use 3D-structural motifs to predict more than 1000 novel metal-binding sites in proteins using three-residue templates, and more than 150 novel metal-binding sites using four-residue templates.

In 2006, Hai Deng, *et al*. based on an aspect related to contact number to detect calcium binding sites[45]. He said if there are four O formed a clique, this region have high percentage be a calcium binding site. This aspect is very similar with CN model[21] which constructed by computing how the residues crowed comparing all residues in the protein.

In 2007, Kasampalidis, I. N., *et al*. used statistics method to analysis the metal binding residues[37]. They proved that certain residues are preferred to bind to certain metals, such as Glu, Asp and His. They also established a statistically significant difference in conservation between metal-coordinating and non-coordinating residues. they mentioned that metal would form chelate with N, O, and S in protein[37].

In our research, we use CN model and replaced Cα by using the atoms that have high probability interact with metal, such as N, O, S. We also consider that different metals would interact with different specific atoms and different residues to form chelate. So predict different kinds of metals should select specific atoms from some specific residues to do CN model. We thought that if there are more atoms that are high probability to interact with metal ion in the region where would be probably metal binding sites, because there is more opportunity to form chelate. We call this model mCN. Base on this thought, we use mCN to predict metal binding residues for Jaspreet Singh Sodhi's dataset. He used NN to predict metal binding site in his non redundant dataset. We also use the same coding scheme as Sodhi's, but training by SVM. No matter NN or SVM, mCN's results are better than them.

2. METHODS

2.1. The contact-number model

The conventional contact number is usually defined as the number of the neighboring residues that are within a cut-off radius of the central residue, which amounts to giving an equal unitary weight to every contacting atom regardless of its distance to the central atom. This definition ignores that an atom at a nearer distance makes a greater contribution than the atoms farther sway. To take the distance factor into account, we define the distance-dependent contact number n_i by weighing the integral contact number by the factor $1/r_{ij}^2$, which is the distance between Cα atoms of residue i and j. is defined as

$$n_i = -\sum_{j \neq i}^{N} \frac{1}{r_{ij}^2} \qquad (1)$$

where N is the number of residues of the protein. Note that this contact number is defined as a negative value. This is for easy for comparing the contact number with the B-factor. We will refer to this model as the contact-number (CN) model. We will normalize n_i to its z-score defined as $z_i^n = (n_i - \bar{n})/\sigma_n$, where $\bar{n}$ and σ_n are the mean and the standard deviation of n. B-factor and CN model are quite similar.

2.2. Using CN to compute metal binding residues – mCN model

Like we mentioned before, metal would form chelate with N, O, and S. So, if we want to use CN to predict metal binding residues, we need to replace Cα to N, O and S. We also add the information that every kind of metals would prefer to interact with specific residues[37; 41; 46], such as, Fe would interact with S in Cys, N in His, and O in hydrophilic residues' side chain more probably. We use statistic method to analysis which residues should be selected, the results show in Figure 9. The conclusion listed in Table 7. Base on these changes we create the method – mCN.

$$m_i = -\sum_{j \neq i}^{M} \frac{1}{r_{ij}^2} \qquad (4)$$

where M is how many atoms may interact with this metal of the protein. Then we choose the atom that the value is lowest of each residue to represent the residue. And get the Z-score. The low m_i means that the neighboring residues around the ith position prefer to interact with the metal. If the Z-score is lower than the cut-off, the residue is high percentage be metal binding residue.

2.3. Assessment indices

To evaluate the quality of our predictions, we use the standard definitions of sensitivity and specificity[16]. Sensitivity S_n is defined as the number of correctly predicted functional

residues (i.e., true positives or TP) divided by the total number of experimentally defined functional residues (i.e., T). Specificity S_p is defined as the number of correctly predicted non-functional residues (i.e., true negatives or TN) divided by the total number of experimentally defined non-functional residues (i.e., F). The false positive rate is defined as $\alpha = 1 - S_p$, and the false negative rate $\beta = 1 - S_n$.

2.4. Data sets

We used Sodhi's dataset[43] which include six kinds of metal binding protein - Ca, Cu, Fe, Mg, Mn and Zn, and the sequence identity is $\leq 25\%$. There are total 982 proteins in this dataset which can divide into six subset according six different kinds of metal ions. There are 261 proteins in Ca dataset, 45 proteins in Cu dataset, 49 proteins in Fe dataset, 216 proteins in Mg dataset, 104 proteins in Mn dataset, and 361 proteins in Zn dataset. The total number is not the same as the sum of the proteins in the six dataset, because there are some proteins interact not only one kind of metal ion. Our dataset has a little different with Sodhi's dataset. 2stv, and 1e53 are replaced by other proteins in PDB (2buk and 1z60). And 1iw7 have 485 Mg ion in the protein, 1f83 is obsolete Structure. So, we deleted the two structures. The metal binding residues are defined by PDBsum[47]. We use CN model, and mCN model to predict Sodhi's dataset[43], and compare his results.

3. RESULTS

3.1. Comparison with metal binding residues and other residues

In Figure 10 we compare the distribution with metal binding residues and other residues by using mCN and CN models. We use calcium for an example. The upper plot of Figure 10A shows the mCN profile of the calcium binding residues, compared with CN profile of the calcium binding residues. The mean of the Z-score (z_m) of the calcium binding residues of

mCN profile is –1.40, while the Z-score (z_n) of CN profile is -0.15. The difference is statistically significant (the p-value $< 2.2 \times 10^{-16}$). These results are showed that the calcium binding residues computed by mCN profile have smaller Z-score than that of CN profile, which means that using specific atoms compute contact number to predict calcium binding residues is better than using Cα. There are around 67% of calcium binding residues of mCN profile with $z_m \le -1$, compared with 23% of CN profile with $z_n \le -1$. The middle plot of Figure 10A shows the distribution of the calcium binding residues and all residues that compute by mCN profile. The mean of z_m of the calcium binding residues is –1.40, and the mean of z_m of all residues is 0.00. The difference is statistically significant (the p-value $< 2.2 \times 10^{-16}$). These results are showed that the calcium binding residues have smaller Z-score than all residues by using mCN model, which means that they are more crowed. There are 67% of calcium binding residues in the region of $z_m \le -1$, compared with 17% of total residues. The bottom plot of Figure 10A shows the distribution of the calcium binding residues and all residues that compute by CN profile. The mean of z_n of the calcium binding residues is –0.15. The difference is statistically significant (the p-value $< 2.51 \times 10^{-8}$). There are 23% of calcium binding residues in the region of $z_n \le -1$, compared with 18% of total residues. In summary, the calcium binding residues tend to bias toward more negative Z-scores in mCN and CN profiles, but mCN is more obvious than CN. In other word, they tend to be more crowed, especially prefer to be located in the region of having the atoms that they are high probability to interact with metal. The other cases are similar and showed in Figure 10B-G.

3.2. Optimize cutoff value

To discriminate the metal binding residues from the other residues, we will determine the optimal cutoff value by minimizing the error function[16] defined as $\varepsilon = \sqrt{(1-S_n)^2 + (1-S_p)^2}$,

which is equivalent to minimizing essentially both the false positives and false negatives. Figure 11 shows the curves of ε against Z-scores of different kinds of metals for CN and mCN models. From this, we can determine the optimal cutoff values for different metals. We can see that although using the optimal cutoff value for CN, the ε still larger than mCN. The optimal z-score cutoff values, at which the corresponding ε is minimal, for the mCN profile: Ca –0.8, Cu –0.7, Fe –1.3, Mg –0.9, Mn –1.2, and Zn -0.1.

3.3. The performances of mCN

According to the optimal cutoff, we can get the best S_n and S_p. The performance of CN and mCN listed in Table 8 and Table 9 which showed that the results are not bad by using CN model. This reveals that contact number really have some relationship with metal binding residues. The metal binding residues tend to locate on the high packing region. But some metal ions like Ca, Mg and Zn are not good enough by computing CN model. Because of that the metal can't interact with Cα. If we compute by mCN model, the results are much better than CN model. This proves that the type of atoms and residues are also important factors should be considered.

3.4. Examples

Figure 12 is some examples that we computed by mCN model. The values below the cut-off are painted by green; these are the metal binding residues that we predicted. The experimental metal binding residues are represented by sticks. If mCN detect the metal binding residues, we painted red.

(A) shows Psoriasin (PDB ID: 2PSR)[48]. The metal binding residues are D-62, N-64, D-66, K-68, and E-73. Psoriasin is a small calcium-binding protein first found in psoriatic lesions and also up-regulated in other inflammatory skin diseases and cancer tissues. The protein responds to transient changes in the cellular calcium concentration by binding yet unidentified

receptor molecules. (B) shows pseudoazurin (PDB ID: 1BQK)[49]. The metal binding residues are H-40, C-78, H-81 and M-86. pseudoazurin is type-I Blue copper-containing proteins. The role of type-I copper-containing redox proteins are to shuttle electrons from an electron donor to an electron acceptor in bacteria and plants. The contribution of the copper ion in pseudoazurin is the stability and the unfolding pathway[50]. (C) shows ferritin (PDB ID: 1FHA)[51]. The metal binding residues are E-27, E-62, and H-65. Ferritin is important in iron homeostasis. Ferritin evolved as the only protein able to solve the problem of iron/oxygen chemistry and metabolism. Its twenty-four chains of two types, H and L, assemble as a hollow shell providing an iron-storage cavity. Ferritin molecules in cells containing high levels of iron tend to be rich in L chains, and may have a long-term storage function, whereas H-rich ferritins are more active in iron metabolism. (D) shows myosin (PDB ID: 1KQM chain B)[52]. The metal binding residues are D-28, D-30, D-32 and F-34. Myosins are a large family of motor proteins found in eukaryotic tissues. Myosins are almost composed of two domains – head domain and tail domain. The role of magnesium ion in myosin is critical for activating ATP hydrolysis. (E) shows Ribonuclease III (PDB ID: 1JFZ chain B)[53]. The metal binding residues are E-240, D-307 and E-310. Ribonuclease III (RNase III) belongs to the family of endoribonucleases that show specificity for double-stranded RNA (dsRNA). Manganese ion has significant impact on crystal packing, intermolecular interactions, thermal stability, and the formation of two RNA-cutting sites within each compound active center. (F) shows p300 protein (PDB ID: 1L3E, chain B)[54]. The form of p300 protein like a triangle composed of four α-helices with three zinc binding sites. The metal binding residues are H-125, C-129, C-142 and C-147 for first zinc; H-156, C-160, C-166 and C-171 for second zinc; H-180, C-184, C-189 and C-192 for third zinc. p300 can form a complex with CBP. They can interact with numerous transcription factors to increase the expression of target genes. The role of the zinc ions is to organize and stabilize the structural conformation.

3.5. Comparison with ROC curves and other methods

Table 9 is the performance of mCN. Table 10 shows we use the standard – FPR 5% to compare with mCN model, Sodhi's NN results[43] and SVM results. Figure 13 compares with the ROC curves of the results of mCN model, Sodhi's NN results and SVM results. In Table 10, when FPR below 5%, the Zn's result is not better than SVM. But Figure 13F reveal that when FPR higher than 7.5%, the Zn's TPR grow rapidly. So, if the users want to get low FPR results, they may return to use SVM. If the users want to get high TPR results, they may return to use mCN.

4. DISCUSSION

Base on the contact-number and the residues frequency, we create a useful method to predict metal binding residues – mCN. The metal binding residues tend to locate on the high packing region or crowed part, and each metal ion has their preference to interact with specific atoms and residues. The more atoms that have high probability to interact with metal, the region are more high percentage to be metal binding site. The lowest sensitivity can reach to 72.4% (Ca), the highest is 94.7% (Cu). The lowest specificity is about 78.1% (Ca), the highest is 97.1% (Fe). And the highest error rate is 35.2% (Ca), the lowest is 10.1% (Cu).

In this method, there are some proteins that mCN model can't predict. First, there are some metals are just select side chain, such as Cu, Fe and Zn. So, if there are some metal binding residues that interact with metal by their backbone, mCN can't find it. Similar situation that if we just selected S from Cys and N from His, mCN can't find the metal interact with N and S from other residues, such as Arg, Lys, Trp and Met. But these situations are very few. Then we analysis mCN's results, we can see that if the metal would often interact with backbone, the results are not as well as the metal almost interact with side chain. Because of that the sum of the atoms of backbone interacts with metal are not many and they

dispersing to 20 amino acid types. So, let the FPR increase.

There are other cases that can't predict very well. Some of the metal would use O of water to form chelate. For example, Peroxisome targeting signal 1 receptor pex5 (PDBID:1hxi)[55]. Figure 14 showed the three dimensional structure of this protein, and we can see that Calcium is stabilize by interact with E397 and five O of water. Because of that we didn't select these five O from water, so we can't detect the metal binding residues very well. There are also some other examples in dataset, especially Ca and Mg, but if we select O from water, these results would create too many false positive. There are many of this kind of proteins in Mg, Ca, some in Zn, Mn, very few in Cu, Fe. We analysis the mCN results and see that the performance of Ca and Mg is also not better than others. So, how to fix O from water is an important future work.

REFERENCES

1. Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W. & Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol* **243**, 327-44.
2. Oldfield, T. J. (2002). Data mining the protein data bank: residue interactions. *Proteins* **49**, 510-28.
3. Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N. & Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol* **326**, 955-78.
4. Lu, C. H., Lin, Y. S., Chen, Y. C., Yu, C. S., Chang, S. Y. & Hwang, J. K. (2006). The fragment transformation method to detect the protein structural motifs. *Proteins* **63**, 636-43.
5. Torrance, J. W., Bartlett, G. J., Porter, C. T. & Thornton, J. M. (2005). Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol* **347**, 565-81.
6. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). Protein function prediction using local 3D templates. *J Mol Biol* **351**, 614-26.
7. Kristensen, D. M., Ward, R. M., Lisewski, A. M., Erdin, S., Chen, B. Y., Fofanov, V. Y., Kimmel, M., Kavraki, L. E. & Lichtarge, O. (2008). Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* **9**, 17.
8. Watson, J. D., Sanderson, S., Ezersky, A., Savchenko, A., Edwards, A., Orengo, C., Joachimiak, A., Laskowski, R. A. & Thornton, J. M. (2007). Towards fully automated structure-based function prediction in structural genomics: a case study. *J Mol Biol* **367**, 1511-22.
9. Tseng, Y. Y. & Liang, J. (2007). Predicting enzyme functional surfaces and locating key residues automatically from structures. *Ann Biomed Eng* **35**, 1037-42.
10. Li, C. L., Hor, L. I., Chang, Z. F., Tsai, L. C., Yang, W. Z. & Yuan, H. S. (2003). DNA binding and cleavage by the periplasmic nuclease Vvn: a novel structure with a known active site. *Embo J* **22**, 4014-25.
11. Hsia, K. C., Chak, K. F., Liang, P. H., Cheng, Y. S., Ku, W. Y. & Yuan, H. S. (2004). DNA binding and degradation by the HNH protein ColE7. *Structure (Camb)* **12**, 205-14.
12. Hsia, K. C., Li, C. L. & Yuan, H. S. (2005). Structural and functional insight into sugar-nonspecific nucleases in host defense. *Curr Opin Struct Biol* **15**, 126-34.
13. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. & Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *J Mol Biol* **344**, 1135-46.

14. Ben-Shimon, A. & Eisenstein, M. (2005). Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J Mol Biol* **351**, 309-26.
15. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. (2006). Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci* **15**, 2120-8.
16. Sacquin-Mora, S., Laforet, E. & Lavery, R. (2007). Locating the active sites of enzymes using mechanical properties. *Proteins* **67**, 350-9.
17. Yuan, Z., Zhao, J. & Wang, Z. X. (2003). Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng* **16**, 109-14.
18. Yang, L. W. & Bahar, I. (2005). Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* **13**, 893-904.
19. Shih, C. H., Huang, S. W., Yen, S. C., Lai, Y. L., Yu, S. H. & Hwang, J. K. (2007). A simple way to compute protein dynamics without a mechanical model. *Proteins* **68**, 34-38.
20. Lu, C. H., Huang, S. W., Lai, Y. L., Lin, C. P., Shih, C. H., Huang, C. C., Hsu, W. L. & Hwang, J. K. (2008). On the relationship between the protein structure and protein dynamics. *Proteins*.
21. Lin, C. P., Huang, S. W., Lai, Y. L., Yen, S. C., Shih, C. H., Lu, C. H., Huang, C. C. & Hwang, J. K. (2008). Deriving protein dynamical properties from weighted protein contact number. *Proteins*.
22. Porter, C. T., Bartlett, G. J. & Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* **32**, D129-33.
23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.
24. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **324**, 105-21.
25. Alexeev, D., Alexeeva, M., Baxter, R. L., Campopiano, D. J., Webster, S. P. & Sawyer, L. (1998). The crystal structure of 8-amino-7-oxononanoate synthase: a bacterial PLP-dependent, acyl-CoA-condensing enzyme. *J Mol Biol* **284**, 401-19.
26. Ekstrom, J. L., Mathews, II, Stanley, B. A., Pegg, A. E. & Ealick, S. E. (1999). The crystal structure of human S-adenosylmethionine decarboxylase at 2.25 A resolution reveals a novel fold. *Structure* **7**, 583-95.
27. Matsumura, H., Xie, Y., Shirakata, S., Inoue, T., Yoshinaga, T., Ueno, Y., Izui, K. & Kai, Y. (2002). Crystal structures of C4 form maize and quaternary complex of E. coli phosphoenolpyruvate carboxylases. *Structure* **10**, 1721-30.

28. Liou, B., Kazimierczuk, A., Zhang, M., Scott, C. R., Hegde, R. S. & Grabowski, G. A. (2006). Analyses of variant acid beta-glucosidases: effects of Gaucher disease mutations. *J Biol Chem* **281**, 4242-53.
29. Sugiyama, M., Ohtani, K., Izuhara, M., Koike, T., Suzuki, K., Imamura, S. & Misaki, H. (2002). A novel prokaryotic phospholipase A2. Characterization, gene cloning, and solution structure. *J Biol Chem* **277**, 20051-8.
30. Heine, A., Luz, J. G., Wong, C. H. & Wilson, I. A. (2004). Analysis of the class I aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase at 0.99A resolution. *J Mol Biol* **343**, 1019-34.
31. Lubkowski, J., Yang, F., Alexandratos, J., Wlodawer, A., Zhao, H., Burke, T. R., Jr., Neamati, N., Pommier, Y., Merkel, G. & Skalka, A. M. (1998). Structure of the catalytic domain of avian sarcoma virus integrase with a bound HIV-1 integrase-targeted inhibitor. *Proc Natl Acad Sci U S A* **95**, 4831-6.
32. Matthews, D. A., Dragovich, P. S., Webber, S. E., Fuhrman, S. A., Patick, A. K., Zalman, L. S., Hendrickson, T. F., Love, R. A., Prins, T. J., Marakovits, J. T., Zhou, R., Tikhe, J., Ford, C. E., Meador, J. W., Ferre, R. A., Brown, E. L., Binford, S. L., Brothers, M. A., DeLisle, D. M. & Worland, S. T. (1999). Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. *Proc Natl Acad Sci U S A* **96**, 11000-7.
33. Warshel, A. (1978). Energetics of enzyme catalysis. *Proc Natl Acad Sci U S A* **75**, 5250-4.
34. Warshel, A., Naray-Szabo, G., Sussman, F. & Hwang, J. K. (1989). How do serine proteases really work? *Biochemistry* **28**, 3629-37.
35. Warshel, A. (1981). Calculations of enzymatic reactions: calculations of pKa, proton transfer reactions, and general acid catalysis reactions in enzymes. *Biochemistry* **20**, 3167-77.
36. Shurki, A., Strajbl, M., Villa, J. & Warshel, A. (2002). How much do enzymes really gain by restraining their reacting fragments? *J Am Chem Soc* **124**, 4097-107.
37. Kasampalidis, I. N., Pitas, I. & Lyroudia, K. (2007). Conservation of metal-coordinating residues. *Proteins* **68**, 123-30.
38. Rigden, D. J. & Galperin, M. Y. (2004). The DxDxDG motif for calcium binding: multiple structural contexts and implications for evolution. *J Mol Biol* **343**, 971-84.
39. Gregory, D. S., Martin, A. C., Cheetham, J. C. & Rees, A. R. (1993). The prediction and characterization of metal binding sites in proteins. *Protein Eng* **6**, 29-35.
40. Chakrabarti, P. (1990). Geometry of interaction of metal ions with histidine residues in protein structures. *Protein Eng* **4**, 57-63.
41. Goyal, K. & Mande, S. C. (2008). Exploiting 3D structural templates for detection of metal-binding sites in protein structures. *Proteins* **70**, 1206-18.

42. Karlin, S., Zhu, Z. Y. & Karlin, K. D. (1997). The extended environment of mononuclear metal centers in protein structures. *Proc Natl Acad Sci U S A* **94**, 14225-30.
43. Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L. & Jones, D. T. (2004). Predicting metal-binding site residues in low-resolution structural models. *J Mol Biol* **342**, 307-20.
44. Lin, H. H., Han, L. Y., Zhang, H. L., Zheng, C. J., Xie, B., Cao, Z. W. & Chen, Y. Z. (2006). Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. *BMC Bioinformatics* **7 Suppl 5**, S13.
45. Deng, H., Chen, G., Yang, W. & Yang, J. J. (2006). Predicting calcium-binding sites in proteins - a graph theory and geometry approach. *Proteins* **64**, 34-42.
46. Dudev, T. & Lim, C. (2003). Principles governing Mg, Ca, and Zn binding and selectivity in proteins. *Chem Rev* **103**, 773-88.
47. Laskowski, R. A., Chistyakov, V. V. & Thornton, J. M. (2005). PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res* **33**, D266-8.
48. Brodersen, D. E., Nyborg, J. & Kjeldgaard, M. (1999). Zinc-binding site of an S100 protein revealed. Two crystal structures of Ca2+-bound human psoriasin (S100A7) in the Zn2+-loaded and Zn2+-free states. *Biochemistry* **38**, 1695-704.
49. Inoue, T., Nishio, N., Suzuki, S., Kataoka, K., Kohzuma, T. & Kai, Y. (1999). Crystal structure determinations of oxidized and reduced pseudoazurins from Achromobacter cycloclastes. Concerted movement of copper site in redox forms with the rearrangement of hydrogen bond at a remote histidine. *J Biol Chem* **274**, 17845-52.
50. Stirpe, A., Sportelli, L. & Guzzi, R. (2006). A comparative investigation of the thermal unfolding of pseudoazurin in the Cu(II)-holo and apo form. *Biopolymers* **83**, 487-97.
51. Lawson, D. M., Artymiuk, P. J., Yewdall, S. J., Smith, J. M., Livingstone, J. C., Treffry, A., Luzzago, A., Levi, S., Arosio, P., Cesareni, G. & et al. (1991). Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. *Nature* **349**, 541-4.
52. Himmel, D. M., Gourinath, S., Reshetnikova, L., Shen, Y., Szent-Gyorgyi, A. G. & Cohen, C. (2002). Crystallographic findings on the internally uncoupled and near-rigor states of myosin: further insights into the mechanics of the motor. *Proc Natl Acad Sci U S A* **99**, 12645-50.
53. Blaszczyk, J., Tropea, J. E., Bubunenko, M., Routzahn, K. M., Waugh, D. S., Court, D. L. & Ji, X. (2001). Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage. *Structure* **9**, 1225-36.
54. Freedman, S. J., Sun, Z. Y., Poy, F., Kung, A. L., Livingston, D. M., Wagner, G. & Eck, M. J. (2002). Structural basis for recruitment of CBP/p300 by hypoxia-inducible

factor-1 alpha. *Proc Natl Acad Sci U S A* **99**, 5367-72.

55. Kumar, A., Roach, C., Hirsh, I. S., Turley, S., deWalque, S., Michels, P. A. & Hol, W. G. (2001). An unexpected extended conformation for the third TPR motif of the peroxin PEX5 from Trypanosoma brucei. *J Mol Biol* **307**, 271-82.

APPENDIX – A simple method predict active site by structure.

According chapter 2, we thought we can use the same aspect to predict active site. Catalytic residues also interact with some specific residues – polar or charged residues. So, we go to check what kind of atoms that the ligands prefer to interact. Then we select O from backbone and side chain, N form side chain, and S from side chain. And using the method like mCN model to compute the contact-number of these selected atoms and set the lowest value of one residue to represent the residue's contact number. Get Z-score to normalize the value. Finally follow the chapter 1 to set cutoff to get the predicted catalytic residues. We called this method aCN model.

Dataset is the same as chapter 1. Assessment indices follow the chapter 1 and 2. To use specificity, sensitivity and error function to evaluate the accuracy of the method.

The performance is showed in Table A1. We can see aCN results almost the same as the WCN model. The weight and selected atoms maybe have some relationship, because the weight and selected atoms both base on the probability that the residues can be catalytic residues. For WCN and WCM model, the probability is higher; the weight is bigger. For aCN model, the probability is higher; the more atoms are selected from the residue.

We also use the seleced atoms to compute CM model, but the results are not better than WCM model. We thought that is because of that the position of centroid is not change very much.

For example, the centroid shifts 1.1Å of 9pap, 0.68 Å of 1a0i, and just 0.15 Å of 4kbp.

Then we compute ROC curve to compare with many models. The results are showed in Figure A1. Figure A2 also show some examples that aCN model can predict very well. We can found aCN model is a good and simple method to predict catalytic residues.

TABLE CAPTIONS

Table 1: The left column is the amino acid type, the middle column is the fraction of each amino acid type (%) for our dataset, and the right column is the weight for each amino acid type (w).

Table 2: The table shows sensitivity and specificity for our dataset. S_n means sensitivity, S_p means specificity, ε means the value of error function, α means false positive rate, and β means false negative rate. In WCN and WCM, the sensitivity is 78% ~ 80%, the specificity is about 80%, and the error rate is ~ 30%. The results are better than others.

Table 3: In WCN and WCM, S_n is 80% ~ 82%, S_p is about 80% ~ 81%, and ε is 27%. We compare with the results that published by Sophie Sacquin-Mora et al in 2007.

Table 4: We use 5 models to predict the catalytic residues that ligand cross two or more chains. Compare with the results by using one chain and biological unit to compute 5 models. We can see use biological unit is better than one chain.

Table 5: The results that include the proteins that ligand cross two or more chains.

Table 6: The results that exclude the proteins that ligand cross two or more chains. The results compare to Table 5 are almost the same.

Table 7: The specific atoms selected for mCN. The atoms have high probability to interact with the six metals for Sodhi's dataset.

Table 8: The performances of using CN (selected Cα) predict metal binding residues.

Table 9: The performances of using mCN (selected specific atoms see Table 7) predict metal binding residues. The results are much better than CN.

Table 10: Using another standard – FPR = 5% to compare with CN, mCN, and Sodhi's results (we also use his features to run SVM). And mCN is the much better than others expect

for Zn.

Table A1: The results of aCN and aCM model and compare with the other 5 models mentioned in chapter 1. We found aCN almost the same as WCN and WCM model.

FIGURE CAPTION

Figure 1: (A) The schematic illustration of the nCN model. The spheres represent the residues, and r_0 is the cut-off value for the nCN model. In this particular example, the contact-number of the central residue (the black sphere) is 2 (i.e., the 2 spheres in gray). The same sizes of the spheres indicate that they contribute equally. The contribution of residues outside the cut-off radius is ignored (spheres in white). (B) The CN model. The size of each sphere (gray) indicates its relative contribution to the central residue. No cut-off radius is used in the CN model. The contribution of each residue is scaled down by its reverse squared distance from the central residue. (C) The WCN model. The central residue is weighted by the statistical probability of its amino acid type occurring in the active site. In the left, the central residue is a Pro, which rarely occurs in an active site, and hence, is weighted by a smaller probability. In the right, the central residue is a Glu, which occurs more frequently than Pro, and is weighted by a larger probability

Figure 2: the frequency distribution of the 20 amino acid types occurring in the catalytic sites, compared with that of all structures in the data set. We can see their distribution is very different.

Figure 3: (A) The figure shows the histogram of active site and all residues of the z_b (upper), the z_{r^2} (middle) and the z_{ρ^2} (bottom) models. The black bars are the residues of active site and the gray bars are all residues. **(B)** The figure shows the histogram of active site and all residues of the z_b (upper), the z_n (middle) and the z_v (bottom) models. The black bars are the residues of active site and the gray bars are all residues.

Figure 4: In the figure, five curve represented five models. When the threshold is -0.5, the error rate of BF would be lowest. When the threshold is -0.7, the error rate of CN and CM would be lowest. And we can see when the threshold is -0.9, the error rate of WCN and WCM

would be lowest.

Figure 5: This figure shows the ROC curves of different models. WCN and WCM are almost the same, and they both better than CN, CM and BF. The result of BF is worst.

Figure 6: (A) The profiles of 8-amino-7-oxononanoate synthase (PDB ID: 1bs0). The z_b profile is shown on the top, the z_{r^2} profile on the middle and the z_n profile on the bottom. The catalytical residues, i.e., H133, E175, D204 and K236, are marked in empty circles. (B) The three-dimensional structures of 8-amino-7-oxononanoate synthase. The colors of the structures are ramped from blue (negative z-score) to red (positive z-score) according to the z_b (top), z_{r^2} (middle) and z_n (bottom) profiles. (C) (D) The profiles and three-dimensional structures of S-adenosylmethionine decarboxylase (PDB ID: 1jen). The catalytic residues are C82, S229 and H243 (located in chain A), and E11 and E67 (located in chain B). (E) (F) The profiles and three-dimensional structures of phosphoenolpyruvate carboxylase (PDB ID: 1jqn). The catalytic residues are H138, R196, R581, R699 and R713. (G) (H) The profiles and three-dimensional of Acid beta-glucosidase The catalytic residues are E235, E340 and C342 (PDB ID: 2f61).

Figure 7: The three-dimensional structure of (A) phospholipase A2 (PDB ID: 1IT4), (B) deoxyribose-5-phosphate aldolase (PDB ID: 1P1X), and (C) ASV integrase (1A5V) and (D) rhinovirus protease (1CQQ). For each figure, the upper one is WCN model, the bottom one is WCM model. The catalytic residues are represented by the CPK model. The colors of the structures are ramped from red (negative Z-score) to white (positive Z-score) according to z_ν and z_{ρ^2} profile.

Figure 8: (A) The z_b profiles and (B) the z_{r^2} profiles of 3 lysozymes: 6lyt (thick solid), 2bqo (dotted) and 2lzt (thin solid).

Figure 9: The statistics of the metal binding residues for every amino acid type. (A) is

calcium, (B) is Copper, (C) is iron, (D) is magnesium, (E) is manganese, and (F) is zinc.

Figure 10: The histograms of the proportion distributions of metal binding residues and all residues that computed by mCN and CN profile. The upper plot is compared with the metal binding residues that computed by mCN model and CN profile. The middle plot is compared with the metal binding residues and total residues by using mCN model. The bottom plot is compared with the metal binding residues and total residues by using CN model. (A) is the results of Ca, (B) is Cu, (C) is Fe, (D) is Mg, (E) is Mn, (F) is ZN and (G) is total proteins in the dataset.

Figure 11: The error function ε curves vs. Z-scores of different profile models. (A) is the optimal cutoff for CN and (B) is the optimal cutoff for mCN. Although using the best cutoff for CN, the Err. (ε) is still larger than mCN.

Figure 12: (A) The upper plot is the three-dimensional structures of Psoriasin (PDB ID: 2PSR). The green ball is calcium. The green part is the metal binding residues that predicted by mCN. The sticks are experimental metal binding residues (real metal binding residues). The red part is the metal binding residues and mCN detected. So, if the sticks are red, it means the metal binding residues are predicted by mCN (true positive). If the sticks are white, it means the metal binding residues are missed. The bottom plot is the line chart of z_m. The circle is metal binding residues, almost lie on the wave trough. (B) The result of pseudoazurin, copper binding protein (PDB ID: 1BQK). (C) The result of Ferritin, iron binding protein (PDB ID: 1FHA). (D) The result of Myosin, magnesium binding protein (PDB ID: 1KQM, chain B). (E) The result of Ribonuclease III, manganese binding protein (PDB ID: 1JFZ, chain B). (F) The result of p300 protein, zinc binding protein (PDB ID: 1L3E, chain B).

Figure 13: This figure shows the ROC curves of different metals that predict by 4 different models. (A) is the result of Ca, (B) is the result of Cu, (C) is the result of Fe, (D) is the result

of Mg, (E) is the result of Mn, (F) is the result of Zn.

Figure 14: The three-dimensional structures of Peroxisome targeting signal 1 receptor pex5. The blue part is the metal binding residue of this protein. And we zoom in the metal binding site. The blue stick is metal binding residue – E397, and the small blue balls are water that would interact with magnesium ion, too.

Figure A1: The ROC curve of many different models. aCN is almost the same as WCN and WCM model, but aCM is not better than WCM model. This reason is the centriod of WCM and aCM are not change very much.

Figure A2: The three-dimensional structure of (A) Actinidin (PDB ID: 1AEC), (B) Endo-1,4-beta-xylanase (PDB ID: 1BVV), and (C) Ricin (PDB ID:1BR6) and (D) DNase I (PDB ID:1DNK). The catalytic residues are represented by the CPK model. The colors of the structures are ramped from red (negative Z-score) to white (positive Z-score) according to aCN model.

TABLES

Table 1. The fraction and weight of each amino acid type

Amino acid type	The fraction of amino acid type (%)	*w*
ALA	1.43	1.16
ARG	9.66	1.98
ASN	3.58	1.55
ASP	16.37	2.21
CYS	5.11	1.71
GLN	2.07	1.32
GLU	12.95	2.11
GLY	3.33	1.52
HIS	16.33	2.21
ILE	0.42	0.63
LEU	0.71	0.85
LYS	9.2	1.96
MET	0.46	0.66
PHE	1.56	1.19
PRO	0.21	0.32
SER	5.61	1.75
THR	2.66	1.42
TRP	1.73	1.24
TYR	6.08	1.78
VAL	0.21	0.32

Table 2. The performance of WCN & WCM models

model	S_n	S_p	ε	α	β	cutoff z-score
WCN	0.78	0.81	0.29	0.19	0.22	-0.9
CN	0.69	0.74	0.40	0.26	0.31	-0.7
WCM	0.80	0.80	0.28	0.20	0.20	-0.9
CM	0.76	0.73	0.36	0.27	0.24	-0.7
B-factor	0.62	0.65	0.52	0.35	0.38	-0.5

Table 3. Comparison with our methods and SLL

Methods	S_n	S_p	ε	α	β
WCN	0.8	0.81	0.27	0.19	0.2
CN	0.71	0.74	0.39	0.26	0.29
WCM	0.82	0.8	0.27	0.2	0.18
CM	0.77	0.74	0.35	0.26	0.23
B-factor	0.66	0.65	0.49	0.35	0.34
SLL	0.78	0.74	0.35	0.26	0.22

Table 4. Ligand interact on the interface of multimer

methods	S_n	S_p	ε	α	β	S_n
WCN (biological unit)	0.81	0.81	0.27	0.19	0.19	-0.9
WCN (one chain)	0.74	0.71	0.39	0.29	0.26	-0.6
CN(biological unit)	0.76	0.75	0.35	0.25	0.24	-0.8
CN(one chain)	0.58	0.57	0.6	0.43	0.42	-0.2
WCM (biological unit)	0.82	0.77	0.29	0.23	0.18	-0.8
WCM (one chain)	0.77	0.76	0.34	0.24	0.23	-0.8
CM(biological unit)	0.71	0.68	0.43	0.32	0.29	-0.6
CM(one chain)	0.78	0.55	0.5	0.45	0.22	-0.3
B-factor	0.64	0.64	0.5	0.34	0.34	-0.5

Table 5. The dataset include ligand that interact on the interface of multimer

methods	S_n	S_p	ε	α	β	S_n
WCN (biological unit)	0.8	0.81	0.28	0.19	0.2	-0.9
WCN (one chain)	0.78	0.81	0.29	0.19	0.22	-0.9
CN(biological unit)	0.74	0.72	0.38	0.28	0.26	-0.7
CN(one chain)	0.69	0.74	0.4	0.26	0.31	-0.7
WCM (biological unit)	0.78	0.8	0.3	0.2	0.22	-0.9
WCM (one chain)	0.8	0.8	0.28	0.2	0.2	-0.9
CM(biological unit)	0.68	0.72	0.42	0.28	0.32	-0.7
CM(one chain)	0.76	0.73	0.36	0.27	0.24	-0.7
B-factor	0.62	0.65	0.52	0.35	0.38	-0.5

Table 6. The dataset exclude ligand that interact on the interface of multimer

methods	S_n	S_p	ε	α	β	S_n
WCN (biological unit)	0.81	0.81	0.28	0.19	0.19	-0.9
WCN (one chain)	0.79	0.81	0.28	0.19	0.21	-0.9
CN(biological unit)	0.71	0.72	0.4	0.28	0.29	-0.7
CN(one chain)	0.72	0.74	0.39	0.26	0.28	-0.7
WCM (biological unit)	0.79	0.8	0.29	0.2	0.21	-0.9
WCM (one chain)	0.81	0.8	0.28	0.2	0.19	-0.9
CM(biological unit)	0.64	0.72	0.45	0.28	0.36	-0.7
CM(one chain)	0.78	0.74	0.34	0.26	0.22	-0.7
B-factor	0.62	0.65	0.52	0.35	0.38	-0.5

Table 7. The specific atoms for different metals.

Metal	The atoms have high probability to interact with metal
Ca	O(backbone) O(side chain of ASP GLU ASN)
Cu	S(side chain),N(His),O(side chain)
Fe	S(Cys),N(His),O(side chain)
Mg	N(His),O(backbone) O(side chain of ASP GLU ASN THR SER)
Mn	N(His),O(backbone) O(side chain of ASP GLU ASN)
Zn	S(Cys),N(His),O(side chain)

Table 8. The performance of CN model

metal	selected atoms	S_n	S_p	ε	α	β	cutoff z-score
Ca	Cα	0.53	0.54	0.66	0.46	0.47	-0.1
Cu	Cα	0.73	0.79	0.34	0.21	0.27	-0.9
Fe	Cα	0.77	0.74	0.35	0.26	0.23	-0.7
Mg	Cα	0.68	0.67	0.46	0.33	0.32	-0.5
Mn	Cα	0.76	0.71	0.38	0.30	0.24	-0.6
Zn	Cα	0.57	0.64	0.56	0.36	0.43	-0.4

Table 9. The performance of mCN model

metal	Selected atoms	S_n	S_p	ε	α	β	cutoff z-score
Ca	O(backbone), O(side chain in Asp Glu Asn)	0.73	0.78	0.35	0.22	0.28	-0.8
Cu	S(side chain), N(His), O(side chain)	0.95	0.92	0.10	0.09	0.05	-0.7
Fe	S(Cys), N(His), O(side chain)	0.87	0.97	0.14	0.03	0.14	-1.3
Mg	N(His), O(backbone), O(side chain of Asp Glu Asn Thr Ser)	0.78	0.81	0.29	0.19	0.22	-0.9
Mn	N(His),O(backbone), O(side chain of Asp Glu Asn)	0.89	0.89	0.16	0.11	0.12	-1.2
Zn	S(Cys),N(His), O(side chain)	0.92	0.84	0.18	0.16	0.09	-0.1

Table 10. When FPR = 5%, TPR for four methods

methods \ metals	Ca	Cu	Fe	Mg	Mn	Zn
Sodhi (using NN)	0.30	0.36	0.49	0.32	0.39	0.48
Sodhi (using SVM)	0.30	0.40	0.49	0.35	0.56	0.67
CN	0.07	0.31	0.31	0.13	0.23	0.16
mCN	0.51	0.89	0.87	0.57	0.83	0.60

Table A1. The performance of aCN and aCM model

methods	S_n	S_p	ε	α	β	cutoff z-score
WCN	0.78	0.81	0.29	0.19	0.22	-0.9
CN	0.69	0.74	0.4	0.26	0.31	-0.7
WCM	0.8	0.8	0.28	0.2	0.2	-0.9
CN	0.76	0.73	0.36	0.27	0.24	-0.7
B-factor	0.62	0.65	0.52	0.35	0.38	-0.5
aCN	0.78	0.81	0.29	0.19	0.22	-0.9
aCM	0.74	0.81	0.32	0.19	0.26	-0.9

FIGURES

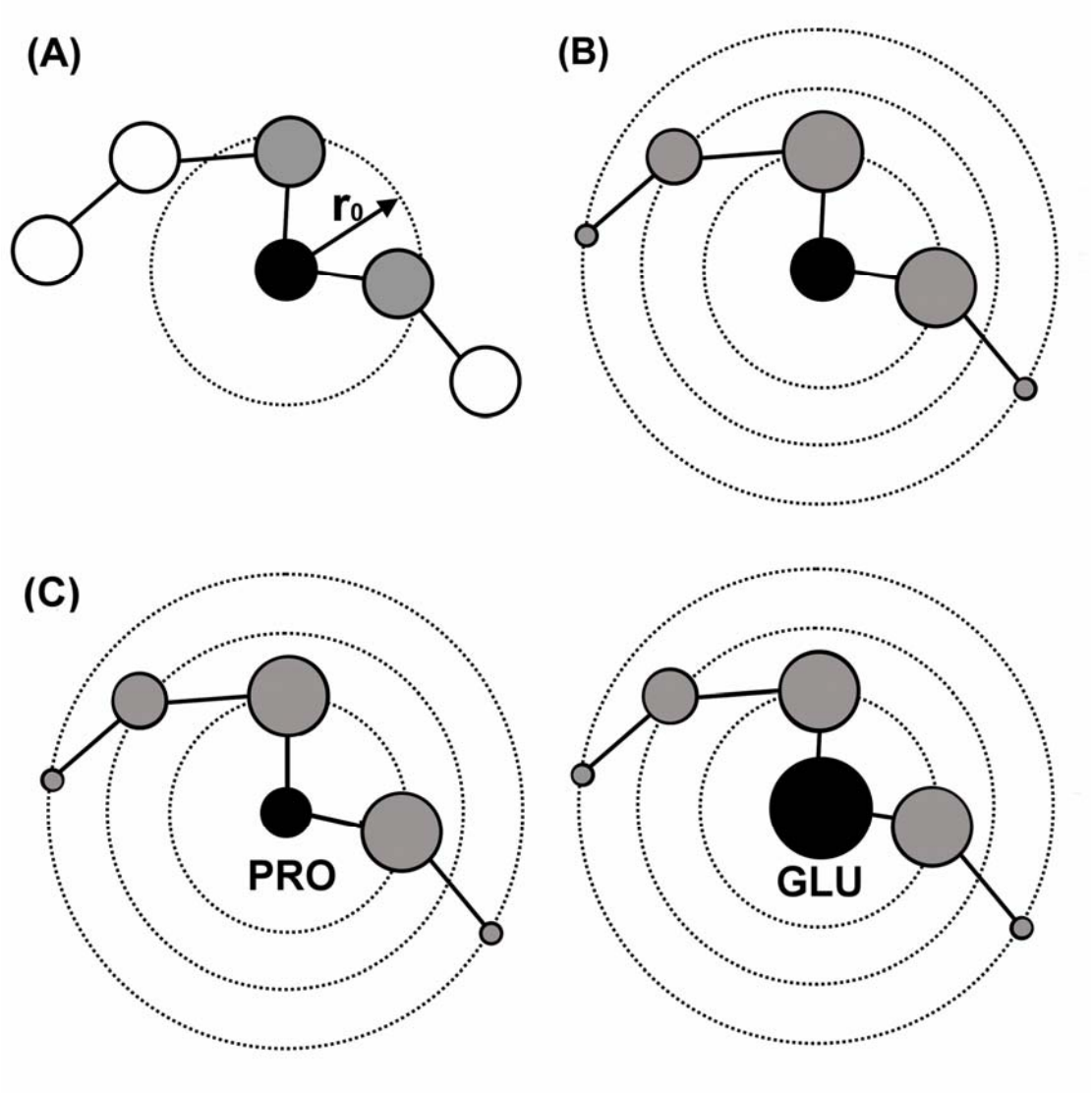


Figure 1 : The diagrams of three methods – naive CN, CN and WCN model.

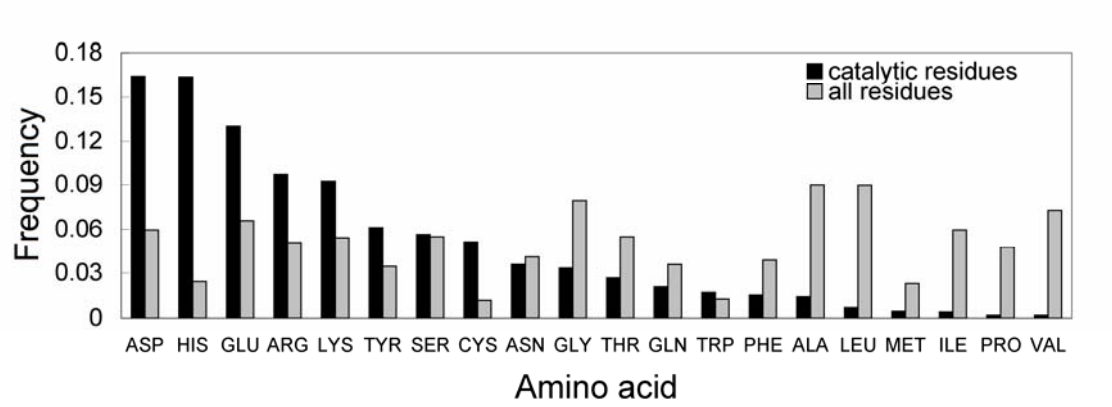


Figure 2 : The proportion of each amino acid type in active sites.

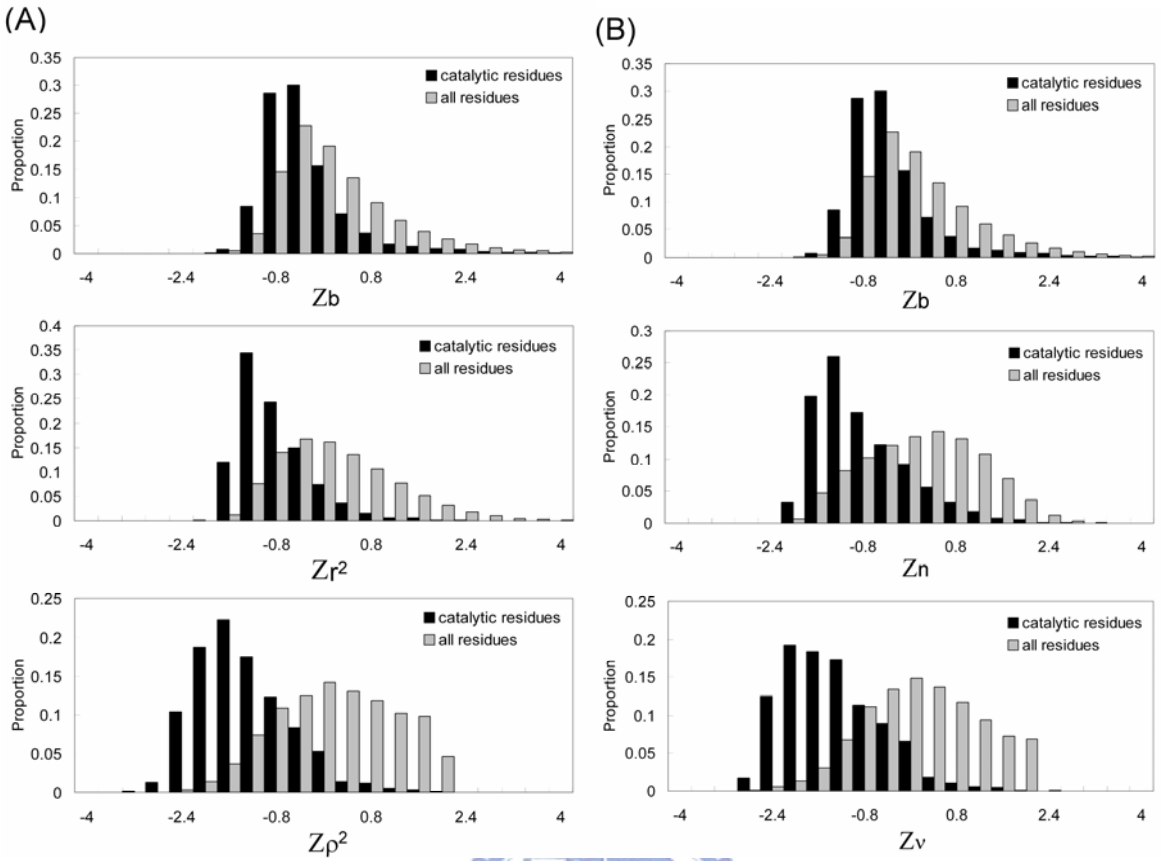


Figure 3 : The histograms of the comparison with BF, CM, CN, WCM and WCN models.

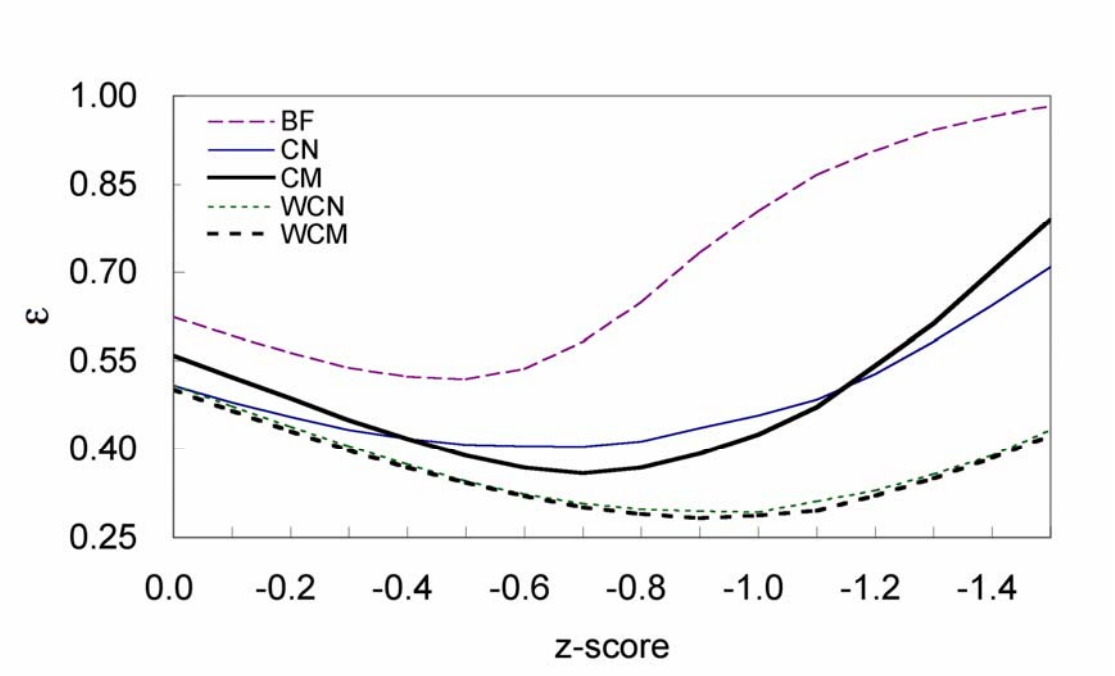


Figure 4 : The error function ε curves vs. Z-scores of different profile models.

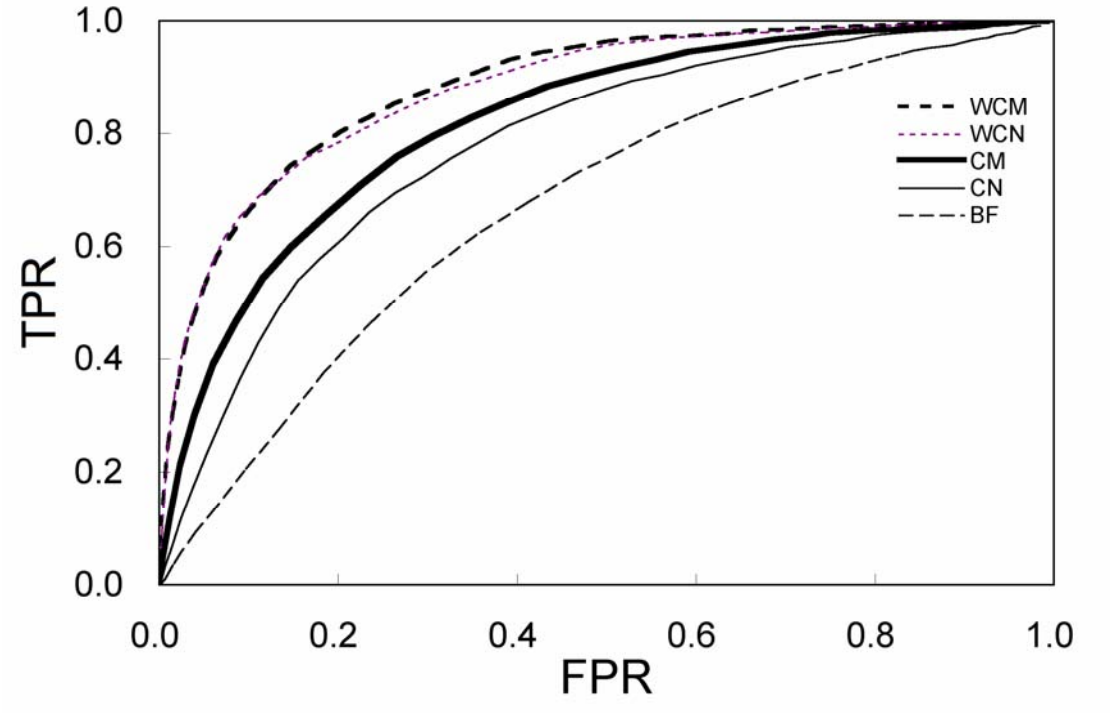


Figure 5 : The ROC curves of the BF, CM, CN, WCM and WCN models.

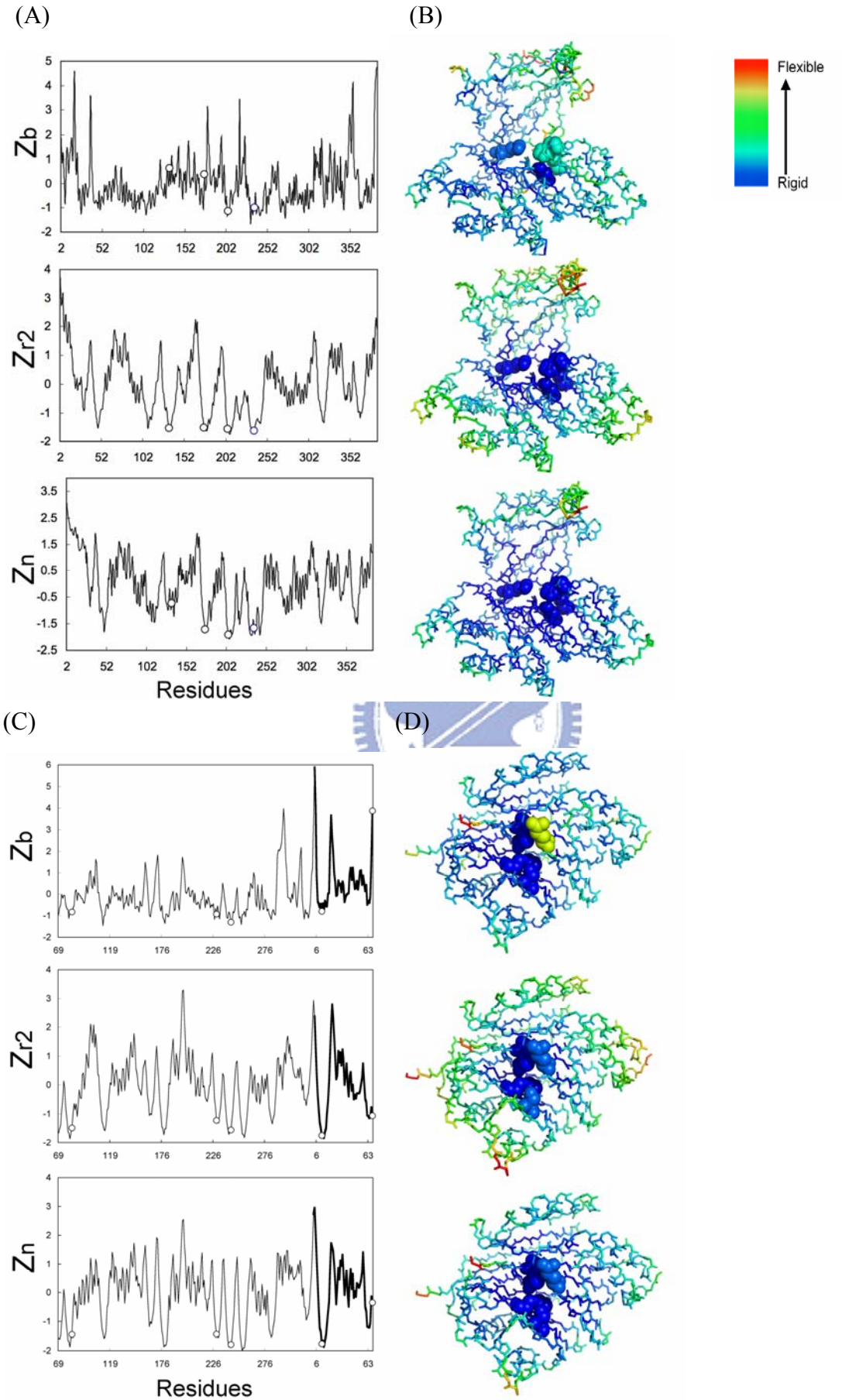


Figure 6 : The examples of comparison with BF, CN and CM models.

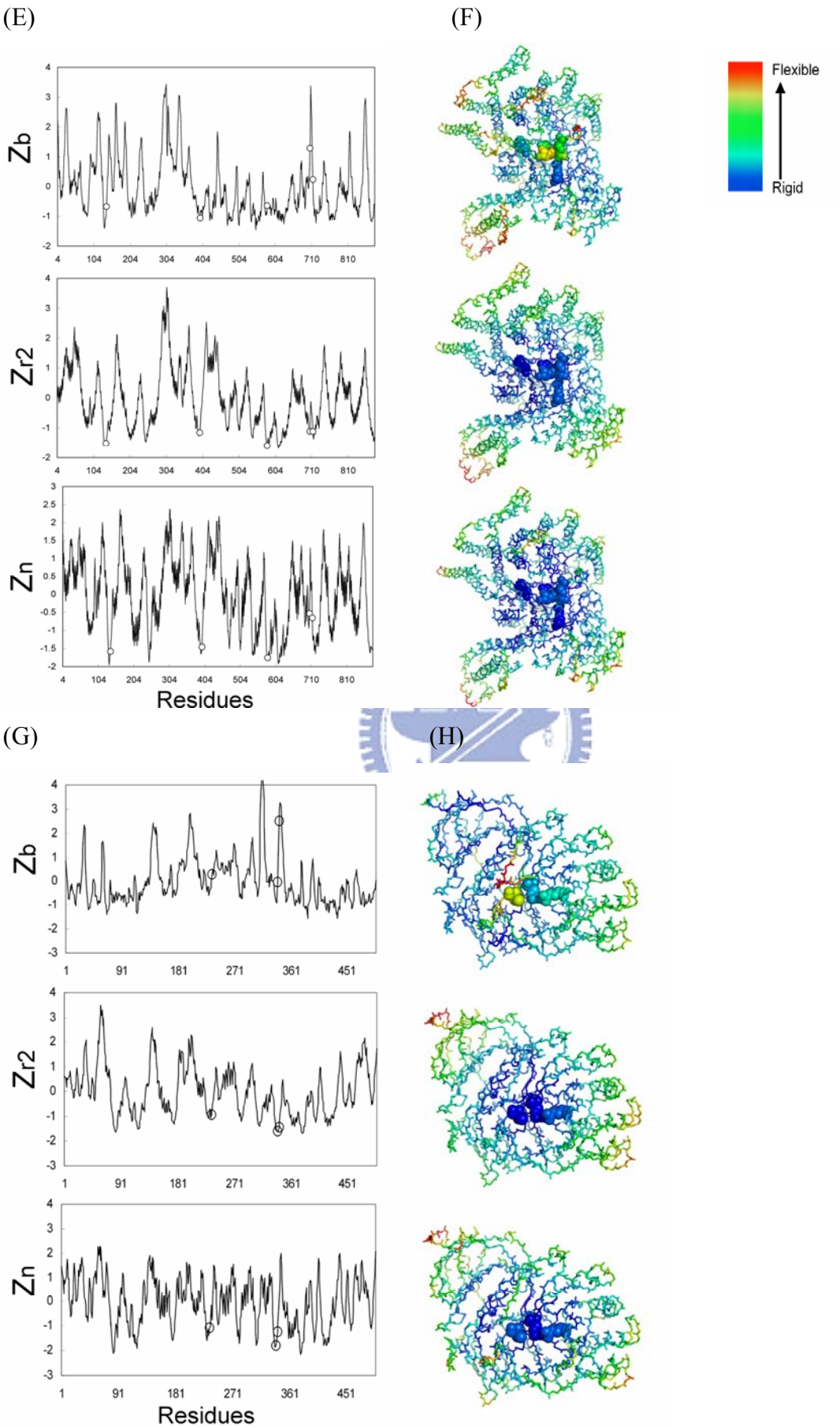


Figure 6 : The examples of comparison with BF, CN and CM models

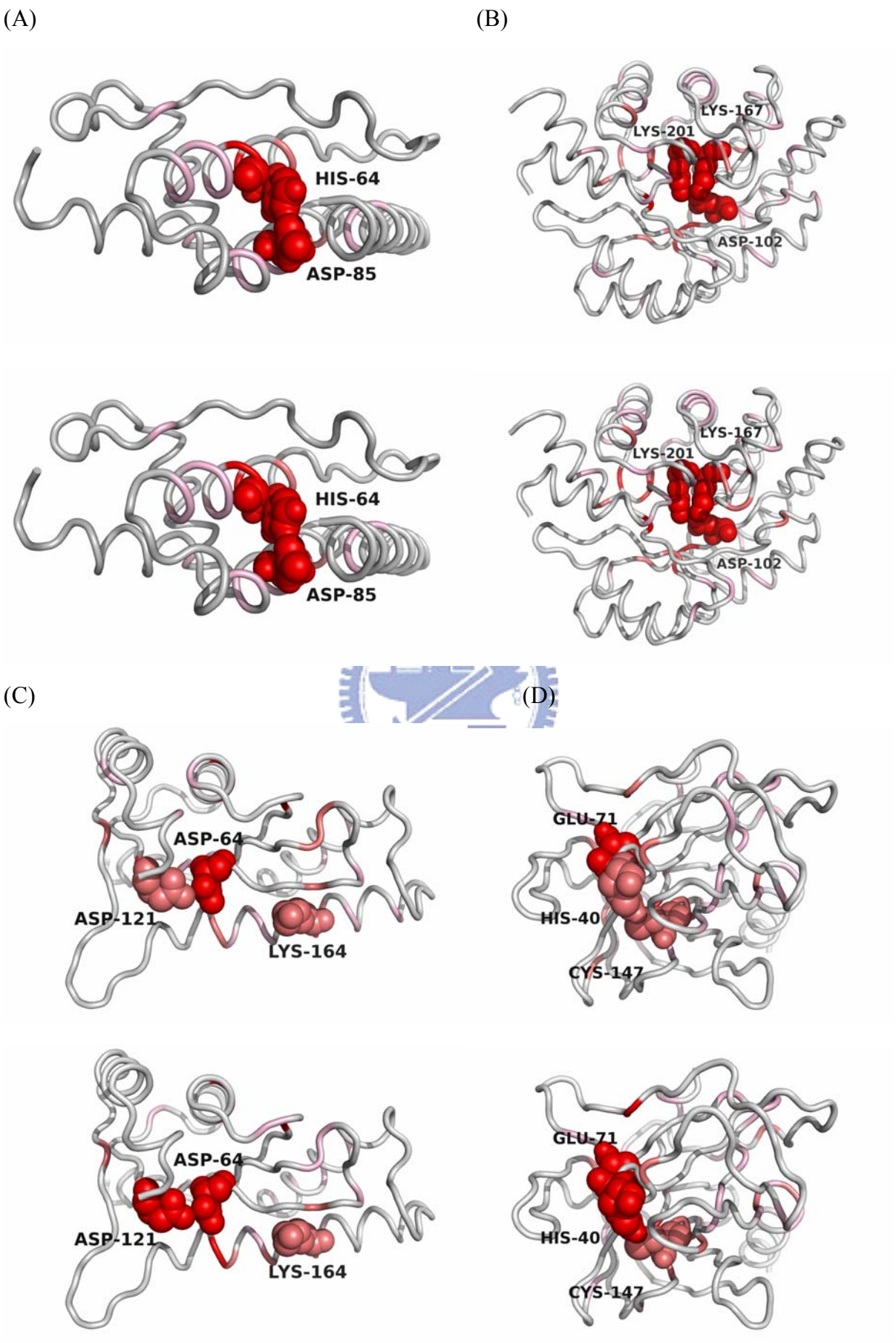


Figure 7 : The examples of WCN and WCM models. (upper is WCN, bottom is WCM)

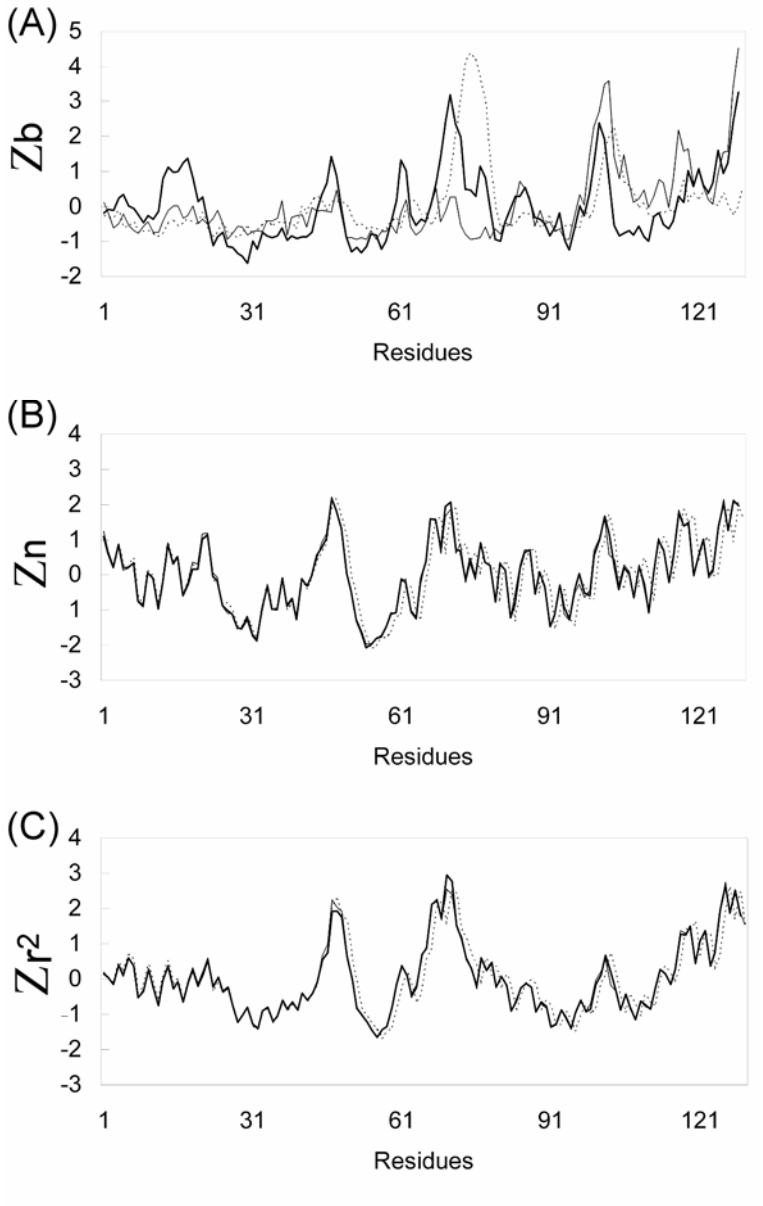
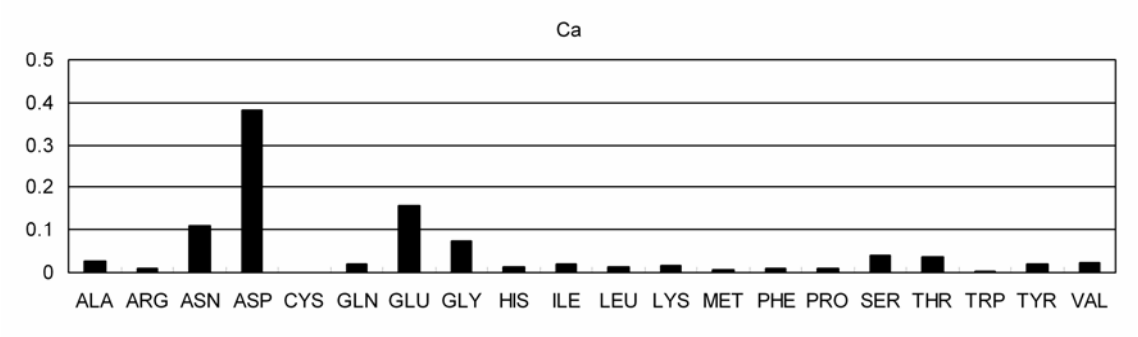
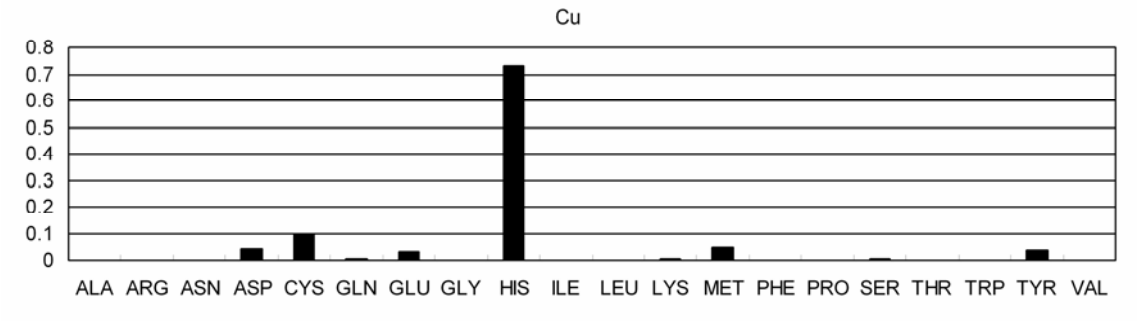


Figure 8 : The Z-score fluctuation of three lysozyme computed by BF, CN and CM models.

(A)



(B)



(C)

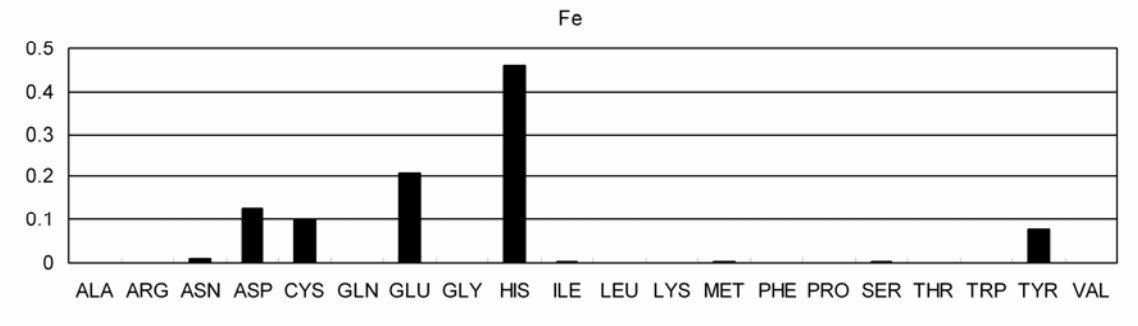
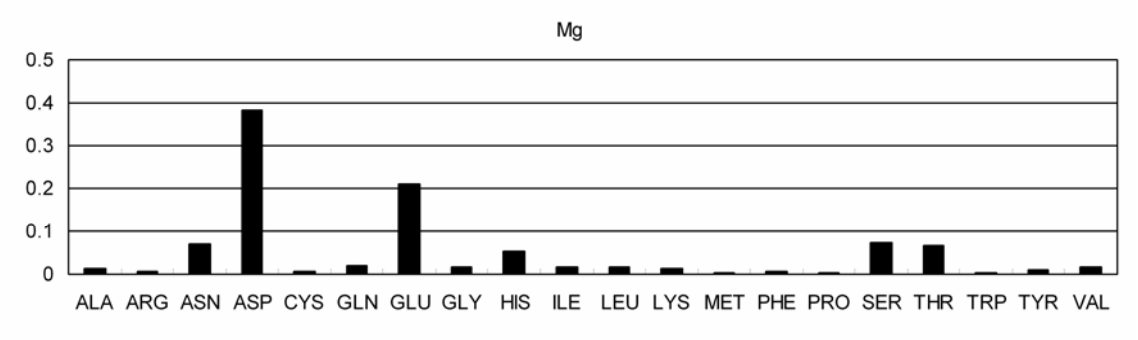
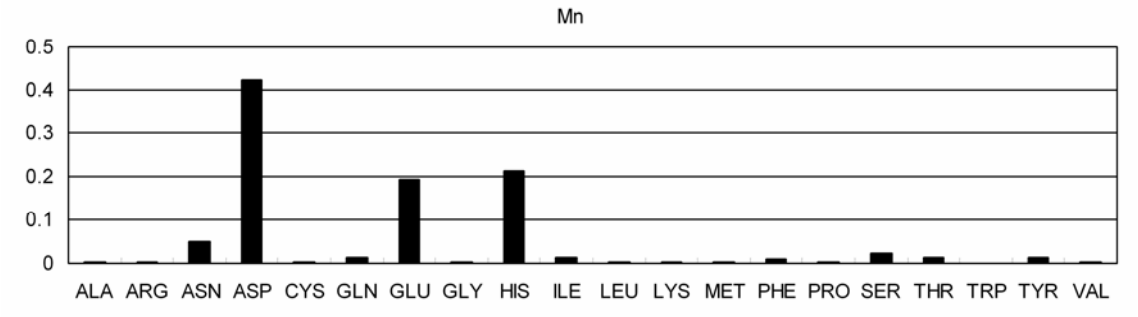


Figure 9 : The proportion of each amino acid type be metal binding residues.

(D)



(E)



(F)

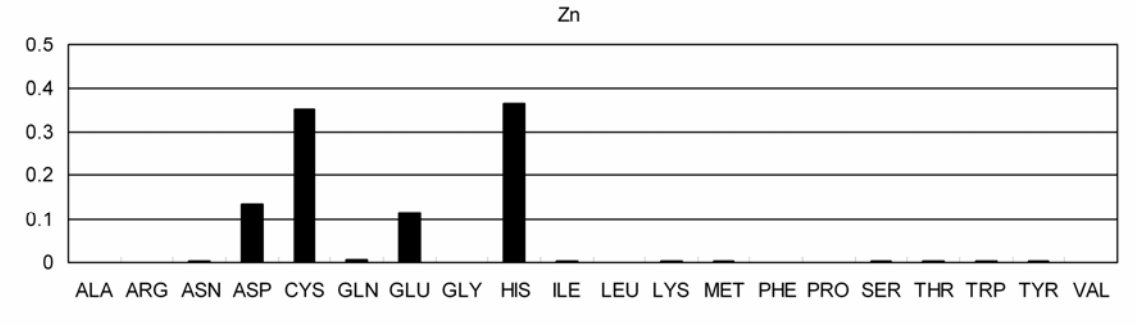


Figure 9 : The proportion of each amino acid type be metal binding residues.

(A) (B)

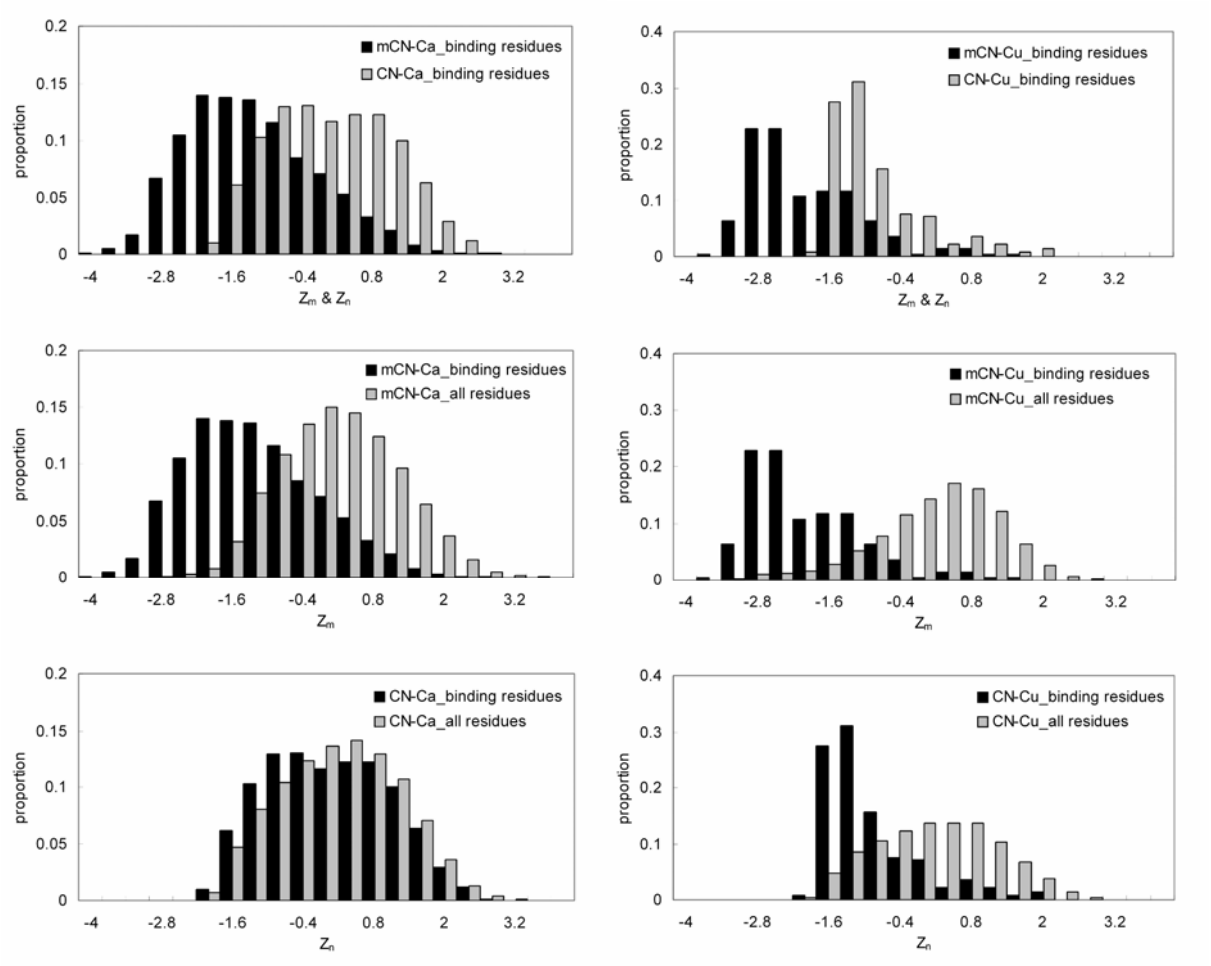


Figure 10 : The histograms of the comparison with CN and mCN models for each metal.

(C) (D)

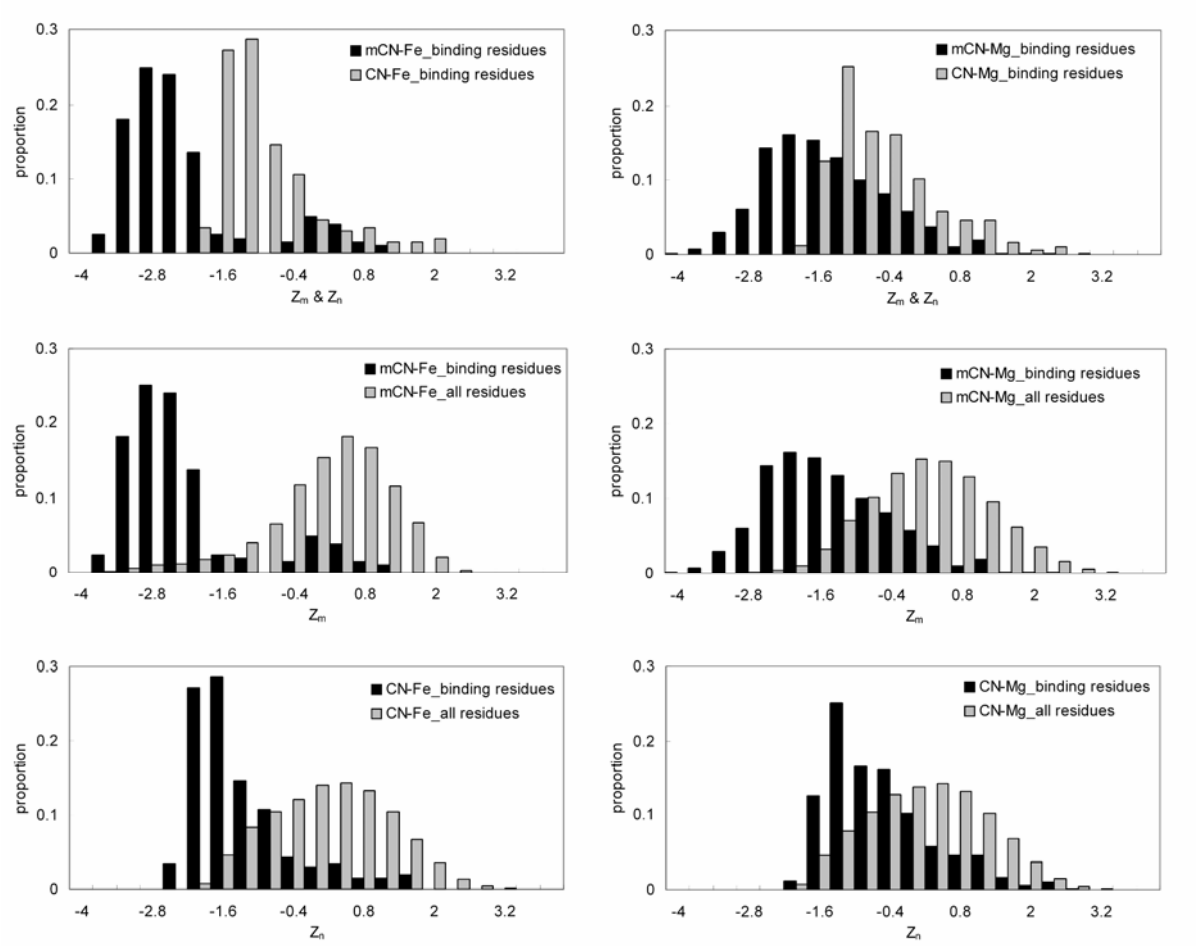


Figure 10 : The histograms of the comparison with CN and mCN models for each metal.

(E) (F)

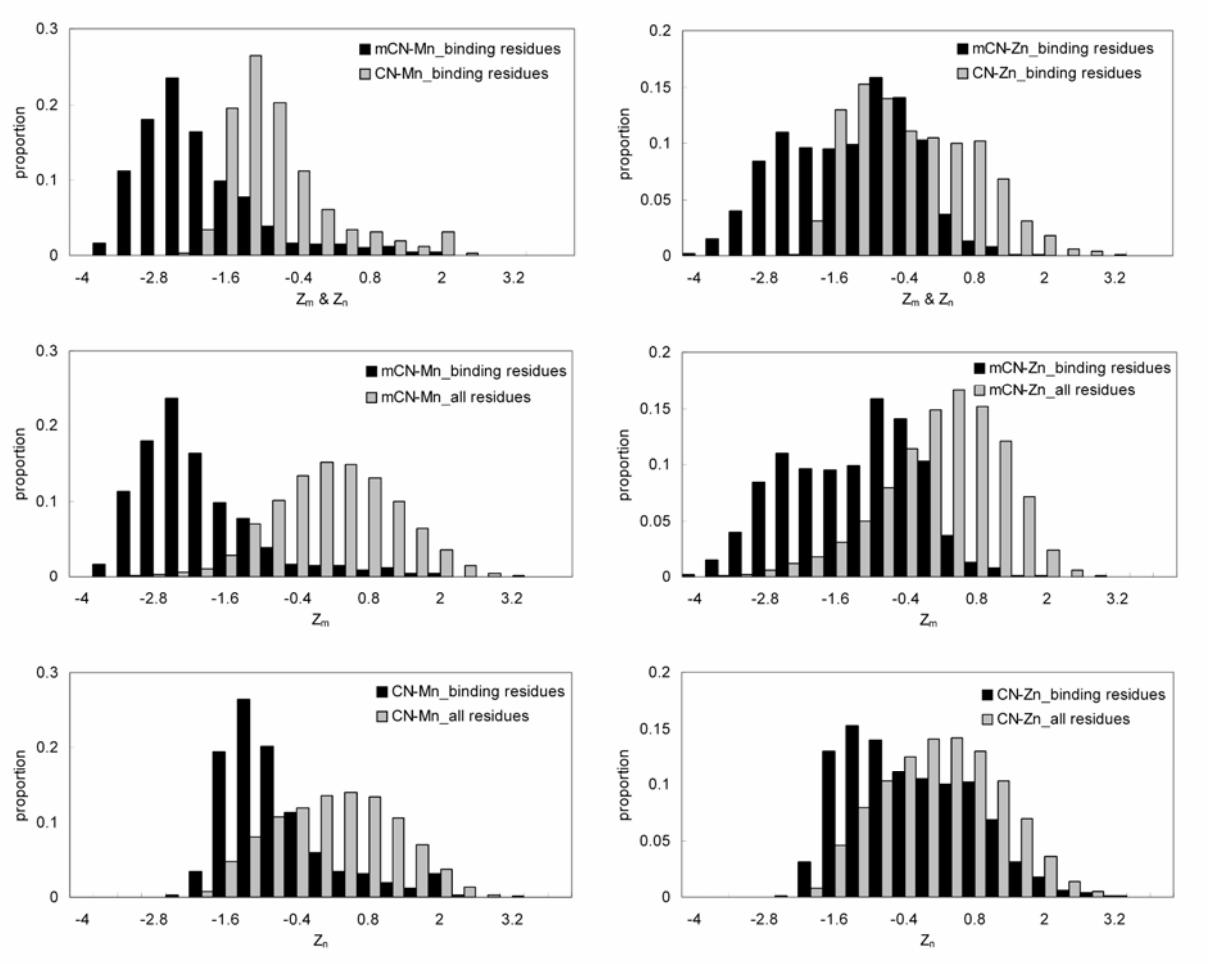


Figure 10 : The histograms of the comparison with CN and mCN models for each metal.

(G)

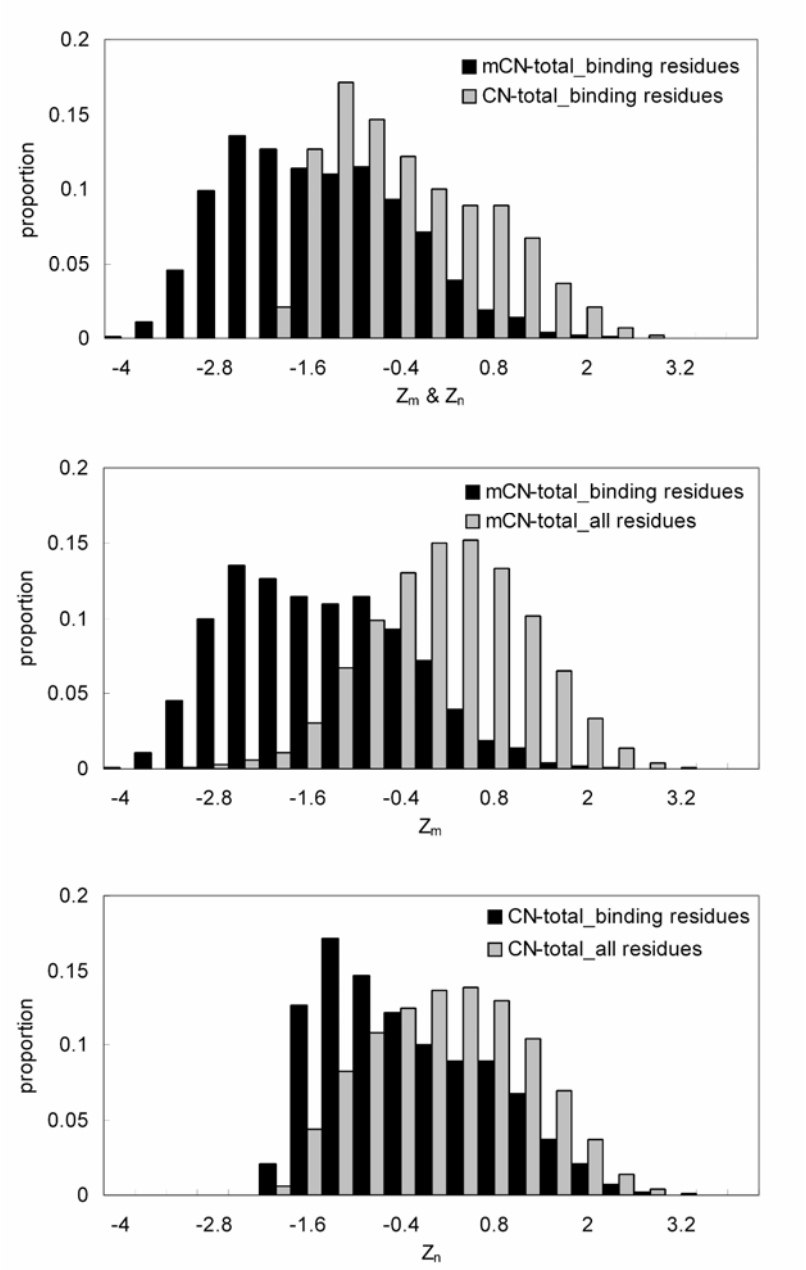
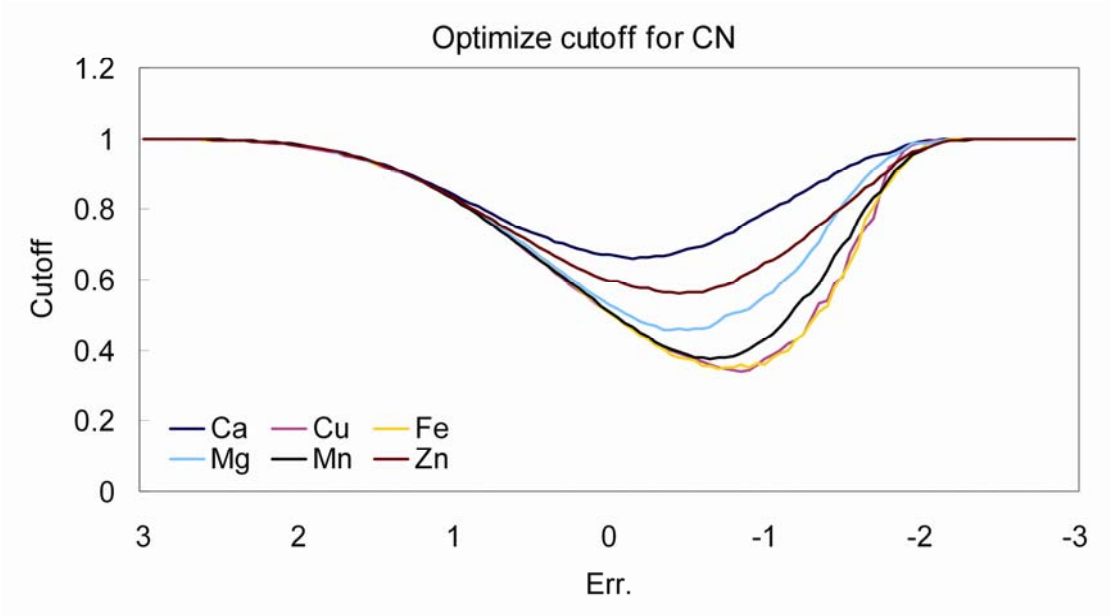


Figure 10 : The histograms of the comparison with CN and mCN models for each metal.

(A)



(B)

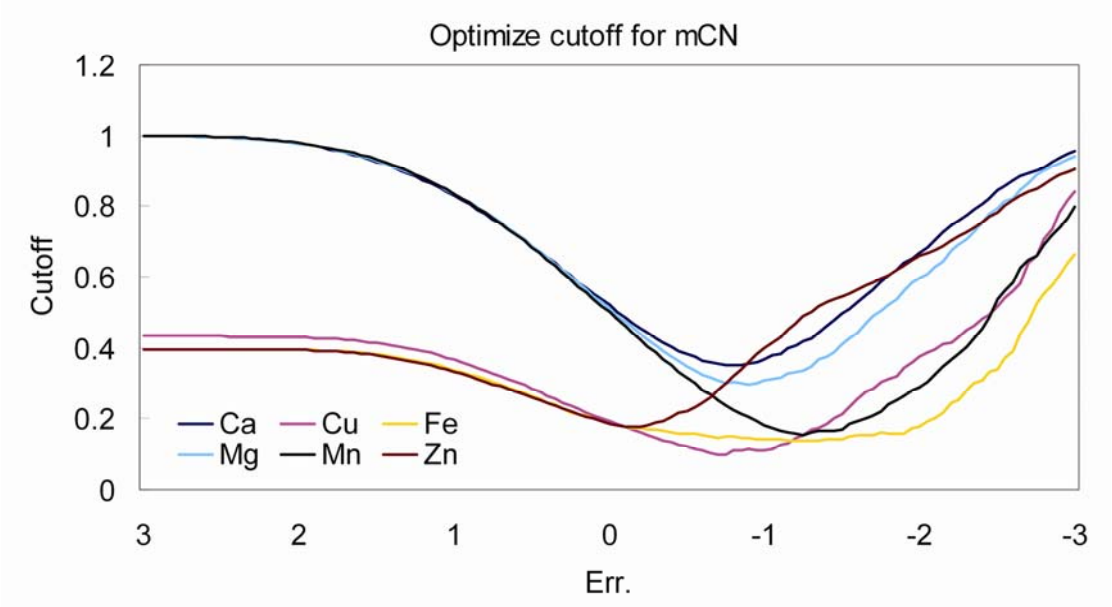


Figure 11 : The error function ε curves vs. Z-scores of two profile models for each metal.

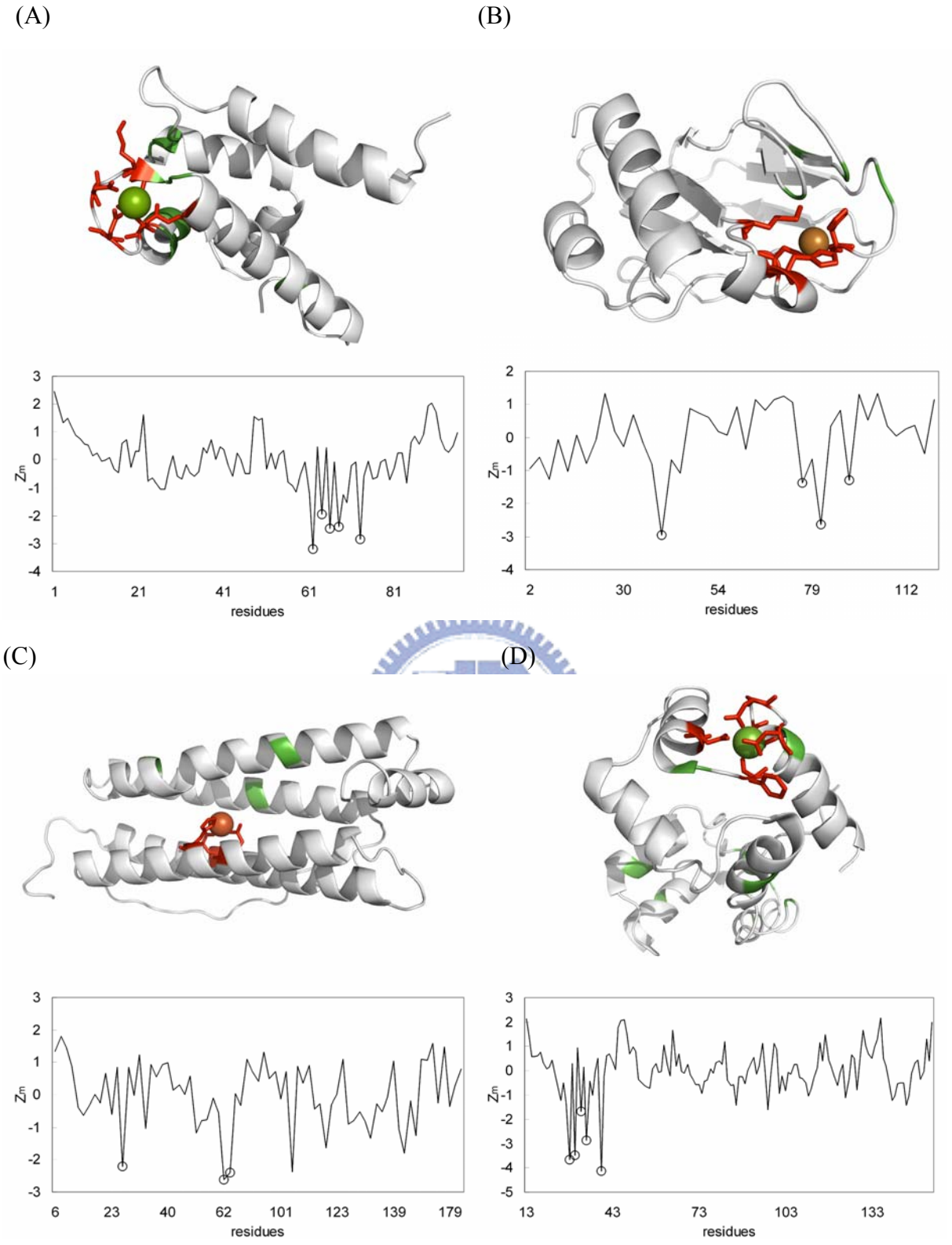


Figure 12 : The examples of mCN models for each metal.

(E) (F)

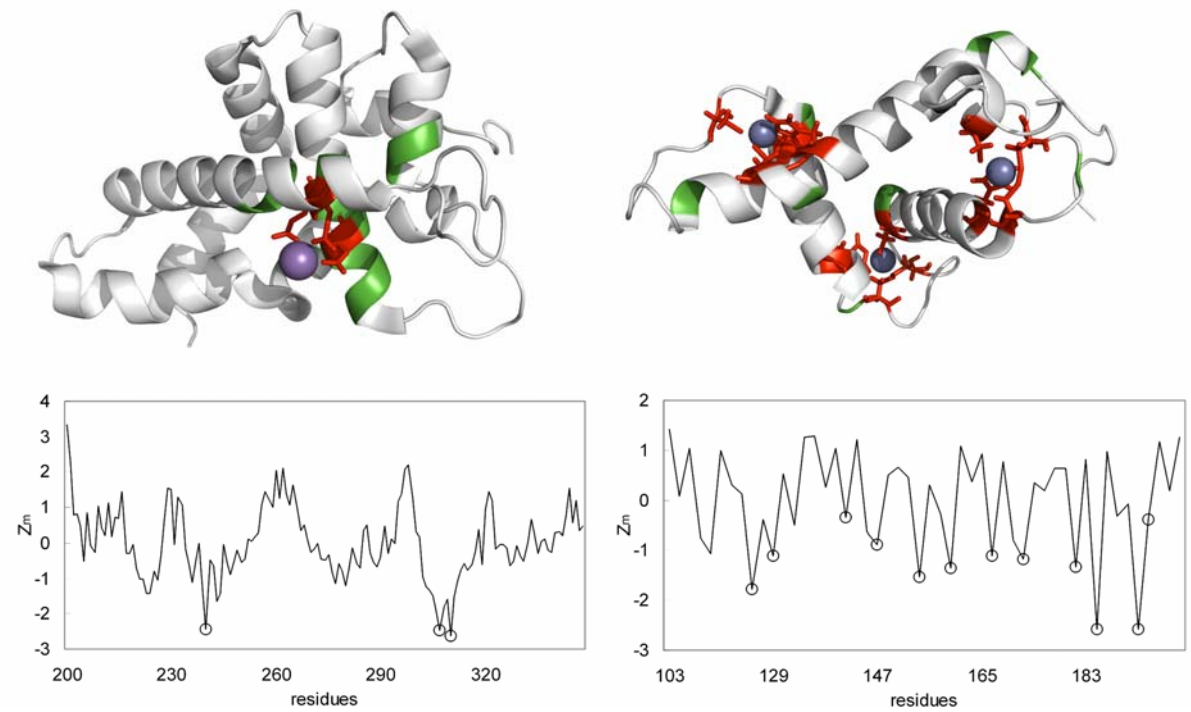


Figure 12 : The examples of mCN models for each metal.

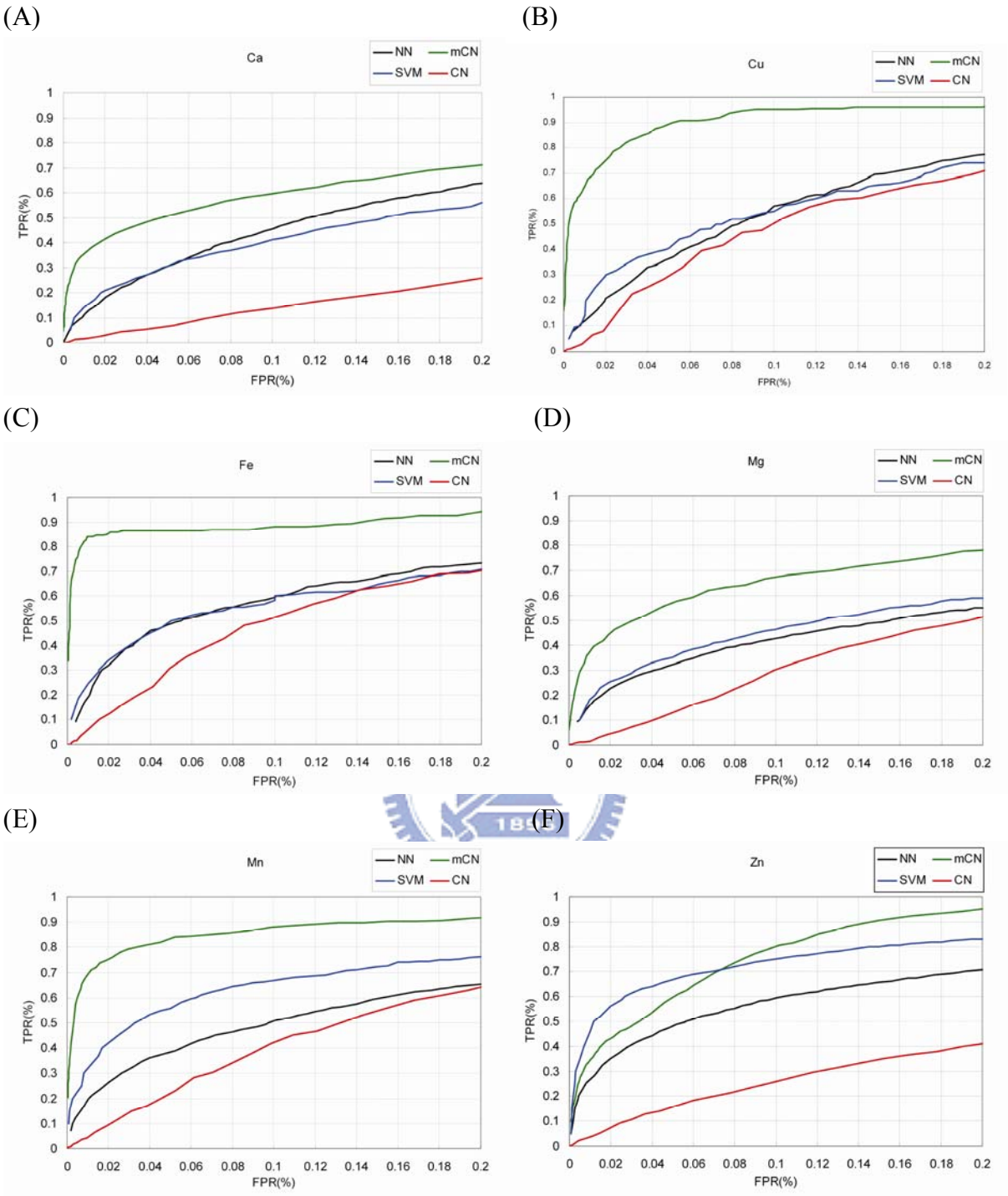


Figure 13 : Using ROC curve to Compare with mCN model and other methods.

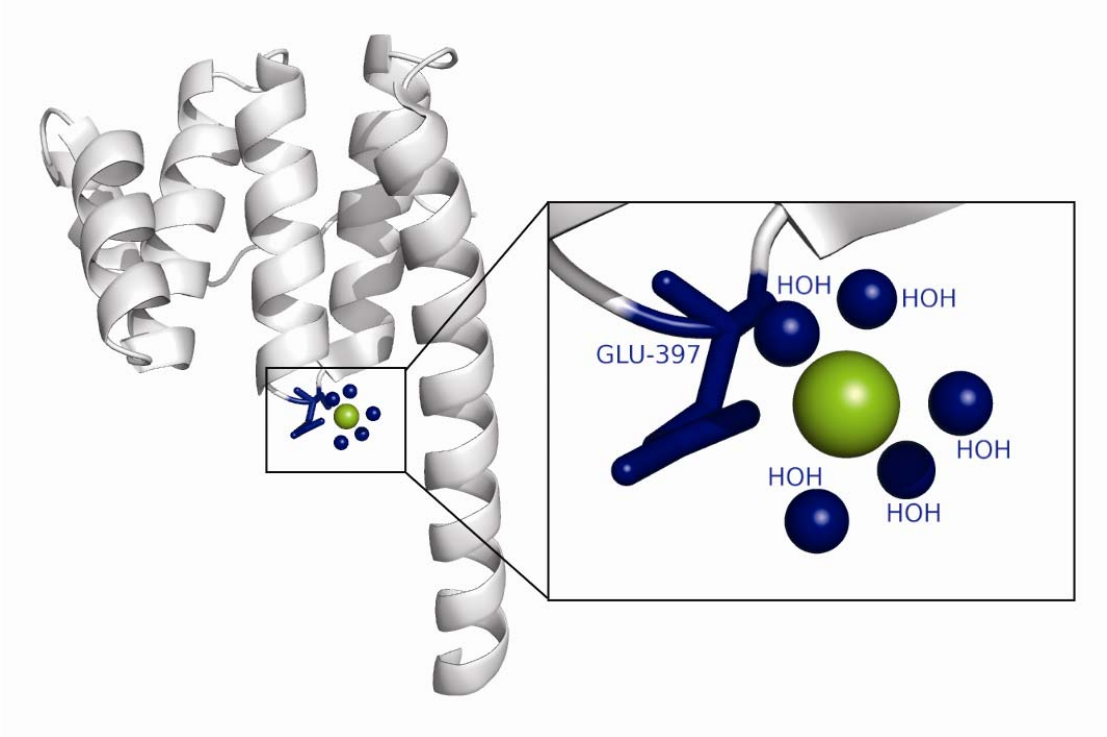
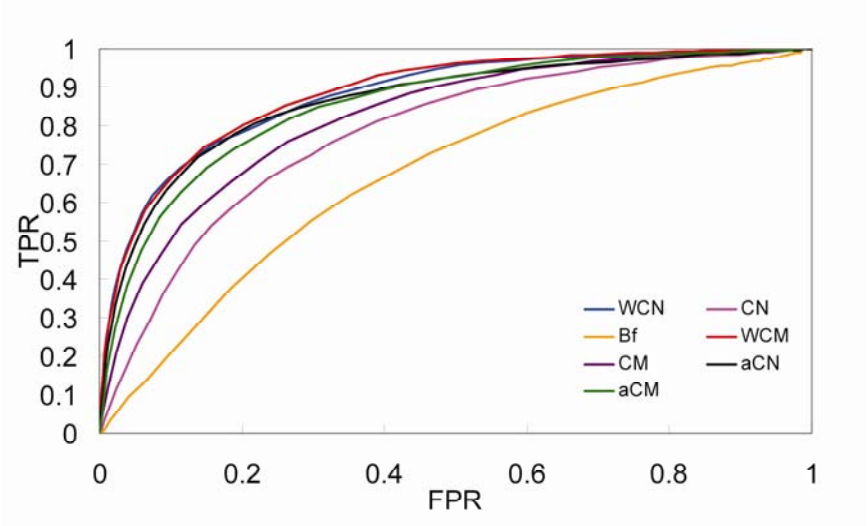
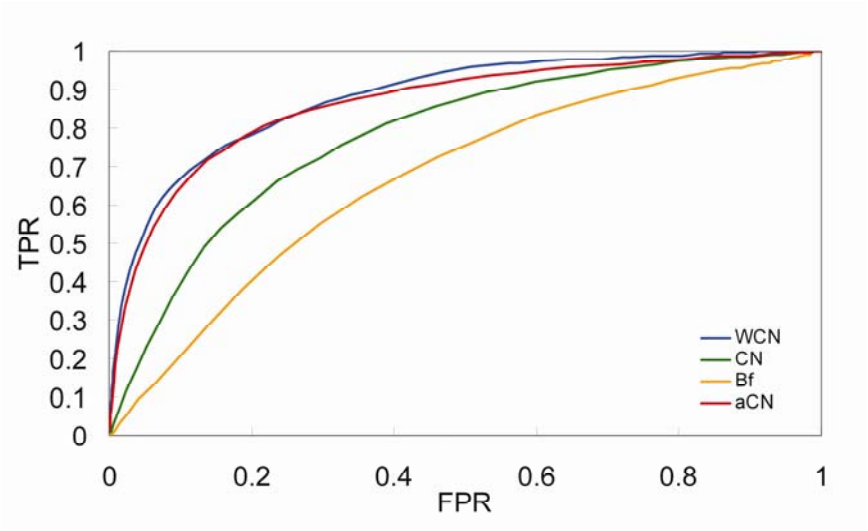


Figure 14 : Excepted case that water interact with metal.

(A)



(B)



(C)

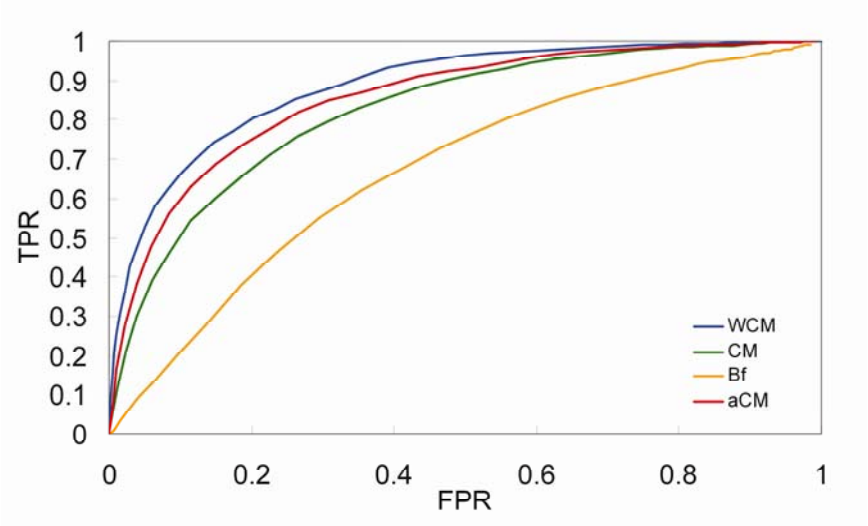
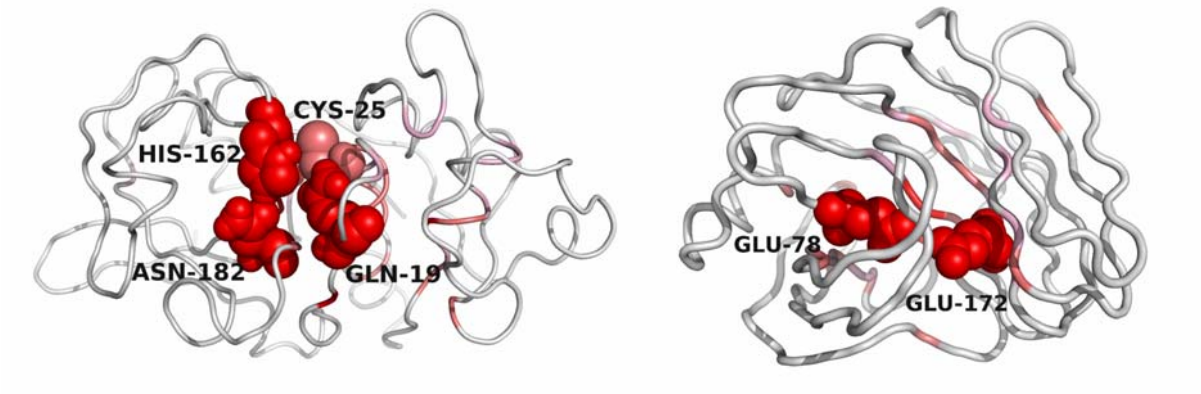


Figure A1 : Comparison with aCN and other models by using ROC curve.

(A) (B)



(C) (D)

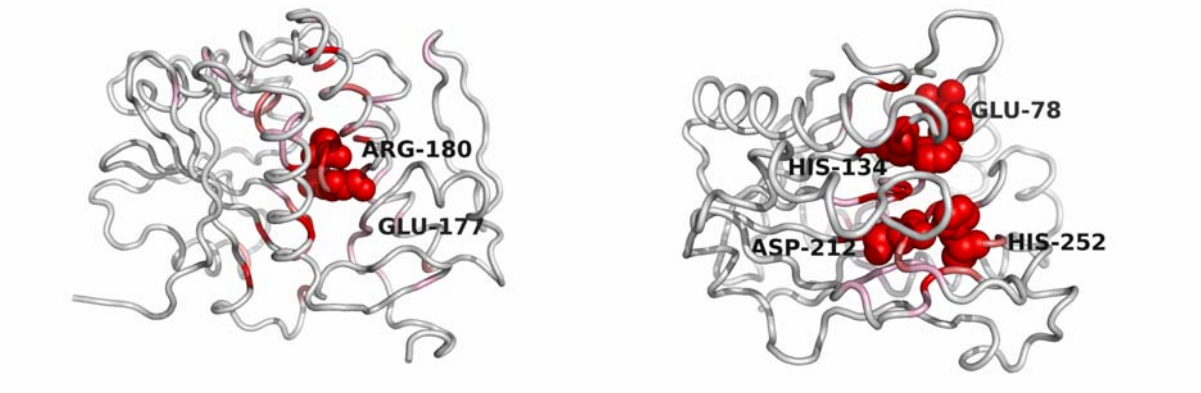


Figure A2 : The examples of aCN models.

Table S1: Dataset for sequence identity <= 30% from CSA								
12as:A	1apt:E	1bgl:C	1cf2:Q	1dae:_	1e2t:F	1f7l:A	1ge7:A	1i19:A
135l:_	1apx:A	1bh2:_	1cfr:_	1db3:A	1e3v:A	1f80:FCA	1geq:B	1i1e:A
13pk:A	1apy:AB	1bhg:B	1cg2:C	1dbf:C	1e5q:E	1f8m:C	1get:BA	1i1i:P
1a05:BA	1aq0:A	1bix:_	1cg6:A	1dbt:B	1e6e:A	1f8r:B	1gim:_	1i29:A
1a0i:_	1aq2:_	1bjo:A	1cgk:A	1dci:A	1e7l:AB	1f8x:B	1gns:A	1i6p:A
1a0j:C	1ar1:A	1bjp:C	1chd:_	1dco:C	1e7q:A	1fa0:A	1gog:_	1i78:B
1a16:_	1arz:B	1bmt:A	1chk:B	1dd8:B	1eb6:A	1fc4:B	1goj:A	1i7q:AB
1a26:_	1ast:_	1bo1:A	1chm:B	1de6:A	1ebf:A	1fcb:A	1gox:_	1i8d:C
1a2t:_	1asy:A	1bol:A	1ci8:B	1dek:A	1ec9:C	1fcq:A	1gp1:B	1i9a:A
1a30:AB	1at1:A	1boo:A	1cjy:A	1df9:B	1ecf:B	1fdy:C	1gp5:A	1idj:B
1a4g:B	1aug:D	1bou:B	1ck7:A	1dfo:B	1ecl:_	1fgh:_	1gpa:A	1ig8:A
1a4i:A	1aui:A	1bp2:_	1cl1:A	1dgs:B	1ecm:AB	1fgj:A	1gpj:A	1im5:A
1a4l:A	1auk:_	1bqc:A	1cns:A	1dhf:B	1ecx:B	1fiq:C	1gpm:C	1ima:B
1a50:B	1auo:A	1brm:B	1coy:_	1dhp:B	1eej:A	1fnb:_	1gpr:_	1inp:_
1a65:A	1auw:AC	1brw:B	1cqj:AB	1dhr:_	1ef0:A	1foa:A	1gq8:A	1iph:A
1a69:A	1avq:C	1bs0:A	1cqq:A	1di1:B	1ef8:A	1fob:A	1gqg:B	1ir3:A
1a79:B	1aw8:AE	1bs4:A	1cs1:CA	1din:_	1eg7:B	1foh:D	1gsa:_	1it4:A
1a8h:_	1ax4:B	1bs9:_	1ct9:D	1dio:L	1eh5:A	1fps:_	1gt7:A	1itq:B
1a8q:_	1ay4:A	1bt1:A	1ctn:_	1diz:A	1eh6:A	1fq0:A	1gtp:L	1itx:A
1a95:C	1azw:A	1btl:_	1ctt:_	1dj0:A	1ehk:AB	1fr2:B	1guf:B	1iu4:A
1ab4:_	1b02:A	1bvv:_	1cv2:A	1djl:A	1ehy:A	1fr8:A	1gxs:C	1iyd:B
1ab8:B	1b04:A	1bvz:A	1cvr:A	1djo:BA	1ei5:A	1fro:C	1gz6:A	1j00:A
1abr:A	1b3r:B	1bwp:_	1cw0:A	1dl2:A	1elq:B	1fua:_	1h19:A	1j09:A
1af7:_	1b57:A	1bwz:A	1cwy:A	1dli:A	1emd:_	1fug:BA	1h2r:L	1j49:B
1afr:D	1b5q:B	1bxr:B	1cz0:A	1dmu:A	1eq2:A	1fui:E	1h3i:B	1j53:A
1afw:B	1b5t:A	1bya:_	1cz1:A	1dnk:A	1esc:_	1fuq:BA	1h54:B	1j79:B
1agm:_	1b65:D	1bzc:A	1czf:B	1dnp:A	1eso:_	1fva:B	1h7o:A	1j7g:A
1agy:_	1b66:A	1bzy:B	1d0s:A	1do6:B	1et0:A	1fwk:D	1h7x:C	1jag:D
1ah7:_	1b6b:B	1c0k:A	1d1q:B	1do8:A	1eug:A	1fy2:A	1hdh:B	1jch:A
1ahj:CD	1b6g:_	1c2t:A	1d2r:E	1dod:_	1euu:_	1g0d:A	1hfe:M	1jdw:_
1aj0:_	1b73:A	1c3c:BA	1d2t:A	1dpg:A	1euy:A	1g24:C	1hfs:_	1jen:AB
1aj8:A	1b7y:A	1c3j:A	1d3g:A	1dqa:AB	1evy:A	1g64:B	1hiv:AB	1jfl:A
1ak0:_	1b8f:A	1c4z:B	1d4a:B	1dqr:AB	1exn:A	1g6t:A	1hja:BC	1jh6:A
1akd:_	1b8g:B	1c82:A	1d4c:A	1dqs:B	1ey2:A	1g72:A	1hka:_	1jhf:A
1akm:A	1b93:A	1c9u:B	1d5r:A	1dtw:BA	1eyi:A	1g79:A	1hpl:A	1jkm:B
1ako:_	1b9h:A	1ca0:CG	1d6m:A	1dup:A	1eyp:A	1g8f:A	1hqc:A	1jm6:A
1al6:_	1bcr:AB	1ca2:_	1d6o:A	1dw9:FA	1ez1:A	1g8p:A	1hr6:BE	1jms:A
1ald:_	1bd0:AB	1ca3:_	1d7r:A	1dxe:A	1ez2:A	1g99:A	1hrd:B	1jnr:AB
1alk:A	1bd3:B	1cb7:D	1d8c:A	1dzr:A	1f2d:A	1ga8:A	1hrk:A	1jof:E
1amo:A	1bf2:_	1cb8:A	1d8d:AB	1e0c:A	1f2v:A	1gal:_	1hto:B	1jqn:A
1amp:_	1bfd:_	1cd5:A	1d8h:A	1e19:B	1f48:A	1gcb:_	1hv9:A	1js4:B
1amy:_	1bg0:_	1cel:A	1d8t:A	1e1a:A	1f6d:B	1gcu:A	1hxq:B	1jxa:BC
1aop:_	1bg6:_	1cev:A	1daa:A	1e2a:B	1f75:B	1gdh:B	1hzf:A	1jxh:A

1k0w:B	1luc:A	1nsf:_	1pja:A	1qmh:B	1slm:_	1v0y:A	2ace:_	2toh:A
1k30:A	1lvh:A	1nsj:_	1pjb:A	1qpr:AB	1sml:A	1v25:B	2acy:_	2tps:A
1k32:A	1lxa:_	1nsp:_	1pjh:A	1qq5:A	1smn:B	1vao:B	2adm:A	2ts1:_
1k4l:A	1lya:AD	1nvm:G	1pkn:_	1qrg:A	1snn:A	1vas:A	2ahj:CD	2xis:_
1k4t:A	1m21:B	1nvt:B	1pma:QA	1qrr:A	1sox:A	1vid:_	2amg:_	2ypn:A
1k82:D	1m6k:A	1nw9:B	1pmi:_	1qsg:G	1ssx:A	1vie:_	2ayh:_	3cla:_
1kae:A	1mas:A	1nww:A	1pnl:B	1qtn:A	1stc:E	1vlb:A	2bbk:L	3csm:A
1kaz:_	1mbb:_	1nzy:C	1pow:A	1qum:A	1std:_	1vnc:_	2bhg:A	3mdd:A
1kc7:A	1mdr:_	1o04:E	1ps1:B	1qv0:A	1szj:R	1vom:_	2bif:B	3nos:A
1kcz:A	1mhl:CA	1o98:A	1ps9:A	1qx3:A	1t0u:B	1vzx:B	2bkr:A	3pca:N
1kdg:A	1mht:A	1o9i:A	1psd:A	1qz9:A	1t4c:AB	1w0h:A	2bmi:A	3r1r:A
1kez:C	1mhy:D	1oac:B	1pud:_	1r16:A	1t7d:A	1w1o:A	2bx4:A	4kbp:A
1kfu:L	1mj9:A	1oas:A	1pvd:A	1r1j:A	1tde:_	1w2n:A	2cpo:_	5cox:D
1kfx:L	1mka:BA	1oba:A	1pvi:B	1r30:A	1tdj:_	1wd8:A	2cpu:A	5cpa:_
1kim:B	1mla:_	1odt:C	1pwh:C	1r4f:B	1thg:_	1wgi:A	2dhn:_	5enl:_
1knp:A	1mlv:B	1oe8:B	1pwv:B	1r4z:A	1ti6:C	1wnw:C	2dln:_	5fit:_
1kny:B	1mok:D	1ofd:A	1pxv:B	1r6w:A	1tml:_	1x7d:A	2dor:B	5rsa:_
1kp2:A	1moq:_	1ofg:F	1pya:AFE	1r76:A	1tmo:_	1x9h:A	2ebn:_	7atj:A
1kra:C	1mpx:C	1og1:A	1pyl:B	1ra2:_	1tox:A	1x9y:B	2eng:_	7odc:A
1ksj:A	1mpy:B	1oh9:A	1pym:B	1rbl:A	1tph:1	1xgm:B	2f61:A	8tln:E
1kws:A	1mqw:A	1oj4:B	1pz3:B	1rdd:_	1trk:A	1xik:A	2fok:B	9pap:_
1kyq:B	1mro:AB	1ok4:H	1q18:B	1req:C	1tyf:I	1xny:AB	2gsa:A	
1kyw:F	1mrq:A	1okg:A	1q3n:A	1rgq:A	1tys:_	1xqw:A	2hdh:A	
1kzh:A	1muc:A	1onr:A	1q3q:C	1rhc:A	1tz3:A	1xrs:B	2hgs:A	
1l0o:B	1mud:A	1opm:A	1q91:A	1rhs:_	1u5u:B	1xtc:A	2isd:B	
1l1d:A	1mug:A	1or8:A	1qam:A	1rk2:C	1u7u:A	1xva:B	2jcw:_	
1l1l:D	1mvn:A	1ord:B	1qaz:A	1ro7:A	1u8v:C	1xvt:A	2lip:_	
1l1r:A	1myr:_	1oro:A	1qba:_	1roz:A	1uae:_	1xyz:A	2nac:A	
1l6p:A	1n20:A	1os7:B	1qcn:A	1rpt:_	1uag:_	1y9m:A	2nlr:A	
1l7n:B	1n8o:BC	1otg:C	1qd1:A	1rpx:C	1uam:A	1ybq:A	2nmt:A	
1l7q:A	1nba:B	1oya:_	1qd6:CD	1rql:B	1uaq:B	1ybv:A	2npx:_	
1l8t:A	1ndh:_	1oyg:A	1qdl:AB	1rtf:B	1uas:A	1ycf:A	2oat:C	
1lam:_	1ndi:A	1p1x:A	1qf6:A	1rtu:_	1uch:_	1ygh:B	2pda:A	
1lba:_	1ndo:AC	1p3d:A	1qfe:B	1ru4:A	1uf7:B	1ysc:_	2pec:_	
1lbu:_	1nf9:A	1p4n:A	1qfm:A	1rvv:A	1uk7:A	1yve:I	2pfl:A	
1lcb:_	1nhx:A	1p4r:B	1qgx:A	1s20:E	1ula:_	1z9h:B	2pgd:_	
1lci:_	1ni4:BC	1p5d:X	1qh9:A	1s2k:A	1un1:B	1ze1:A	2phk:A	
1ldm:_	1nid:_	1pa9:A	1qhf:B	1s3i:A	1uok:_	1zio:_	2pia:_	
1lij:A	1nir:A	1pfk:A	1qhg:A	1s76:D	1uqr:A	1zm2:F	2plc:_	
1lio:A	1nkk:C	1pfq:B	1qho:A	1s95:B	1uqt:B	1zrz:A	2pth:_	
1ljl:A	1nln:A	1pgs:_	1qj4:A	1s9c:B	1uro:A	1zym:B	2sqc:A	
1lml:_	1nlu:A	1pii:_	1qje:A	1sca:_	1ush:_	206l:_	2tdt:_	
1lnh:_	1nml:A	1pix:B	1qk2:B	1ses:B	1v04:A	2a0n:A	2thi:A	
1ltq:A	1nn4:B	1pj5:A	1qlh:A	1sll:_	1v0e:B	2abk:_	2tmd:B	