Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 402536, 13 pages
http://dx.doi.org/10.1155/2015/402536

Research Article

Predicting Flavin and Nicotinamide Adenine Dinucleotide-Binding Sites in Proteins Using the Fragment Transformation Method

Chih-Hao Lu,[1,2] Chin-Sheng Yu,[3,4] Yu-Feng Lin,[5] and Jin-Yi Chen[1]

[1] *Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung 40402, Taiwan*
[2] *Graduate Institute of Basic Medical Science, China Medical University, Taichung 40402, Taiwan*
[3] *Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan*
[4] *Master's Program in Biomedical Informatics and Biomedical Engineering, Feng Chia University, Taichung 40724, Taiwan*
[5] *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 30068, Taiwan*

Correspondence should be addressed to Chih-Hao Lu; chlu@mail.cmu.edu.tw

Received 16 June 2014; Accepted 21 July 2014

Academic Editor: Hao-Teng Chang

We developed a computational method to identify NAD- and FAD-binding sites in proteins. First, we extracted from the Protein Data Bank structures of proteins that bind to at least one of these ligands. NAD-/FAD-binding residue templates were then constructed by identifying binding residues through the ligand-binding database BioLiP. The fragment transformation method was used to identify structures within query proteins that resembled the ligand-binding templates. By comparing residue types and their relative spatial positions, potential binding sites were identified and a ligand-binding potential for each residue was calculated. Setting the false positive rate at 5%, our method predicted NAD- and FAD-binding sites at true positive rates of 67.1% and 68.4%, respectively. Our method provides excellent results for identifying FAD- and NAD-binding sites in proteins, and the most important is that the requirement of conservation of residue types and local structures in the FAD- and NAD-binding sites can be verified.

1. Background

Over the past 12 years, projects involving structural genomics have generated structural data for ~12,000 proteins within the Protein Data Bank (PDB) [1]. For most of these proteins, however, biological function is unknown. It is therefore important to develop computational methodologies that can identify a protein's function from its structure. Many biochemical processes depend on interactions between proteins and cofactors, such as metal ions, vitamins, and adenine dinucleotides, for example, flavin adenine dinucleotide (FAD) and nicotinamide adenine dinucleotide (NAD). Adenine dinucleotides play important roles in many central biological processes, including DNA repair [2, 3], glycolysis, photosynthesis, and transcription [4–7]. By June 2010, 5293 proteins in PDB were annotated "nucleotide binding," and nucleotides constitute ~15% of biologically relevant ligands [8]. These statistics demonstrate how ubiquitous and essential protein-nucleotide interactions are to biological processes.

Although protein-ligand interactions are fundamental to most biochemical reactions, structural information concerning these binding sites is still inadequate. Once ligand-binding sites can be predicted from structural data, putative functions can be assigned to these proteins. More complete annotation of protein function will benefit both basic science and the pharmaceutical industry. Mutations or deletions within these ligand-binding domains often alter biochemical reactions and are the root causes of many diseases. This makes binding sites attractive targets for drug therapies, including anticancer chemotherapy. In recent years computational methods have been used to identify ligand-binding sites within proteins. These methods include empirical approaches [9], support vector machines (SVM) [8, 10, 11], random forest

[12, 13] and artificial neural networks [14], and structure comparison approaches [15–17]. These prediction methods can be divided into two broad categories: ones that use protein-sequence information, for example, amino acid composition, position-specific scoring matrix, and physicochemical properties, and ones that use protein-structure information, for example, dihedral angles, secondary structure, and 3D-structure comparison. The most effective prediction methodologies, however, tend to use a combination of sequence and structure data.

The structural genomics initiative resolves 20 new protein structures each week, and more than 60,000 structures have been deposited into PDB. The functional surfaces of proteins, which interact with cofactors, tend to be more structurally conserved than internal structures [18]. Residues that form a functional binding region are usually quite close to one another when the three-dimensional structure of a protein is examined. In addition, binding regions typically constitute only 10–30% of the entire protein [19–21]. We took advantage of previously generated structural information and used the fragment transformation method [22] to identify new binding sites for the NAD and FAD ligands.

2. Results

2.1. Residues that Bind NAD or FAD. To characterize the structural environment of NAD-/FAD-binding sites, we compared binding-site residues to whole-protein residues. The three-dimensional structure of the NAD/FAD molecule was divided into three moieties according to function. Within the spherical environment of NAD, the adenosine-binding site typically contained glycine, isoleucine, tyrosine, and aspartic acid residues; the phosphate-binding site contained glycine, isoleucine, serine, threonine, methionine, phenylalanine, tyrosine, tryptophan, arginine, and histidine residues; and the nicotinamide-binding site contained serine, threonine, cysteine, phenylalanine, asparagine, tyrosine, tryptophan, histidine, and asparagine residues. For FAD, adenosine was bound by glycine, valine, cysteine, and tryptophan; phosphate was bound by glycine, serine, and arginine; and flavin was bound by cysteine, methionine, phenylalanine, tyrosine, tryptophan, and histidine. The residue types whose ratio of binding-site residues frequency to whole-protein residues frequency was greater than 1.2 were listed above. As such, the binding residues were primarily polar residues, containing charged groups, amide groups, and nucleophilic groups (Figure 1).

We also characterized the types of atoms that were within 3.5 Å of the three moieties of each NAD/FAD ligand (Figure 2). Nicotinamide and flavin moieties were most commonly associated with nitrogen and oxygen atoms within the backbone and side-chains of the protein. Phosphate moieties were commonly bound by backbone and side-chain nitrogen or side-chain oxygen. Each ligand moiety preferentially bound certain atoms within certain residues.

2.2. Prediction Performance. We chose two criteria to evaluate the performance of our binding-site predictions: performance at less than 5% FPR and the Matthews correlation coefficient (MCC). We used a combination of features that included the number of aligned residues, RMSD, BLOSUM, and DSSP. Using a 5% FPR threshold, NAD-binding sites were predicted with an accuracy of 93.46%, a sensitivity of 67.09%, and an MCC of 0.52. Under these same conditions, FAD-binding-site predictions yielded 93.59% accuracy, 68.43% sensitivity, and an MCC of 0.54 (Table 1). When MCCs were maximized, NAD-binding proteins were identified with 95.34% accuracy, 57.88% sensitivity, 97.64% specificity, and an MCC of 0.57. Under these same conditions, FAD-binding residues were identified with an accuracy of 94.33%, a sensitivity of 64.13%, a specificity of 96.27%, and an MCC of 0.55 (Table 2). These data indicated that our method could predict binding residues for these two ligands.

Table 1: The performance of binding-site predictions at a 5% FPR threshold.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
NAD	93.46	67.09	95.08	0.52
FAD	93.59	68.43	95.22	0.54

Table 2: The performance of binding-site predictions at a maximum MCC threshold.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
NAD	95.34	57.88	97.64	0.57
FAD	94.33	64.13	96.27	0.55

2.3. Comparison with Other Methods. We next compared our results with other prediction methodologies. For these comparisons we chose two published methods that use similar criteria for analyzing these kinds of ligand-protein complexes [10, 11]. These chosen methods assign binding or nonbinding status to each residue within NAD-/FAD-binding proteins. Because these published methods use an equal number of binding and nonbinding residues, we applied our prediction method to a similar dataset to make the results comparable. Random-selection processes were performed five times for all nonbinding residues within ligand-protein complexes to generate the same scale for binding and nonbinding residues within each protein. For NAD-binding proteins, our method predicted binding residues with a sensitivity of 86.21% and an MCC of 0.75 compared with 86.13% and 0.75 for the method developed by Ansari and Raghava [10] (Table 3). For FAD-binding proteins, our method yielded 85.68% sensitivity and an MCC of 0.75. These values compared with the performance of the published method (83.36% and 0.66) developed by Mishra and Raghava [11] (Table 4). Our method, therefore, has similar performance in NAD-binding sites predicted but better in FAD-binding sites. However, in native proteins, the number of binding and nonbinding residues should not be equal. The equal number model needs to be further discussed.

2.4. Template Matching. Figures 3–6 show alignments of predicted NAD-/FAD-binding proteins and corresponding templates. Structures within these figures were drawn using

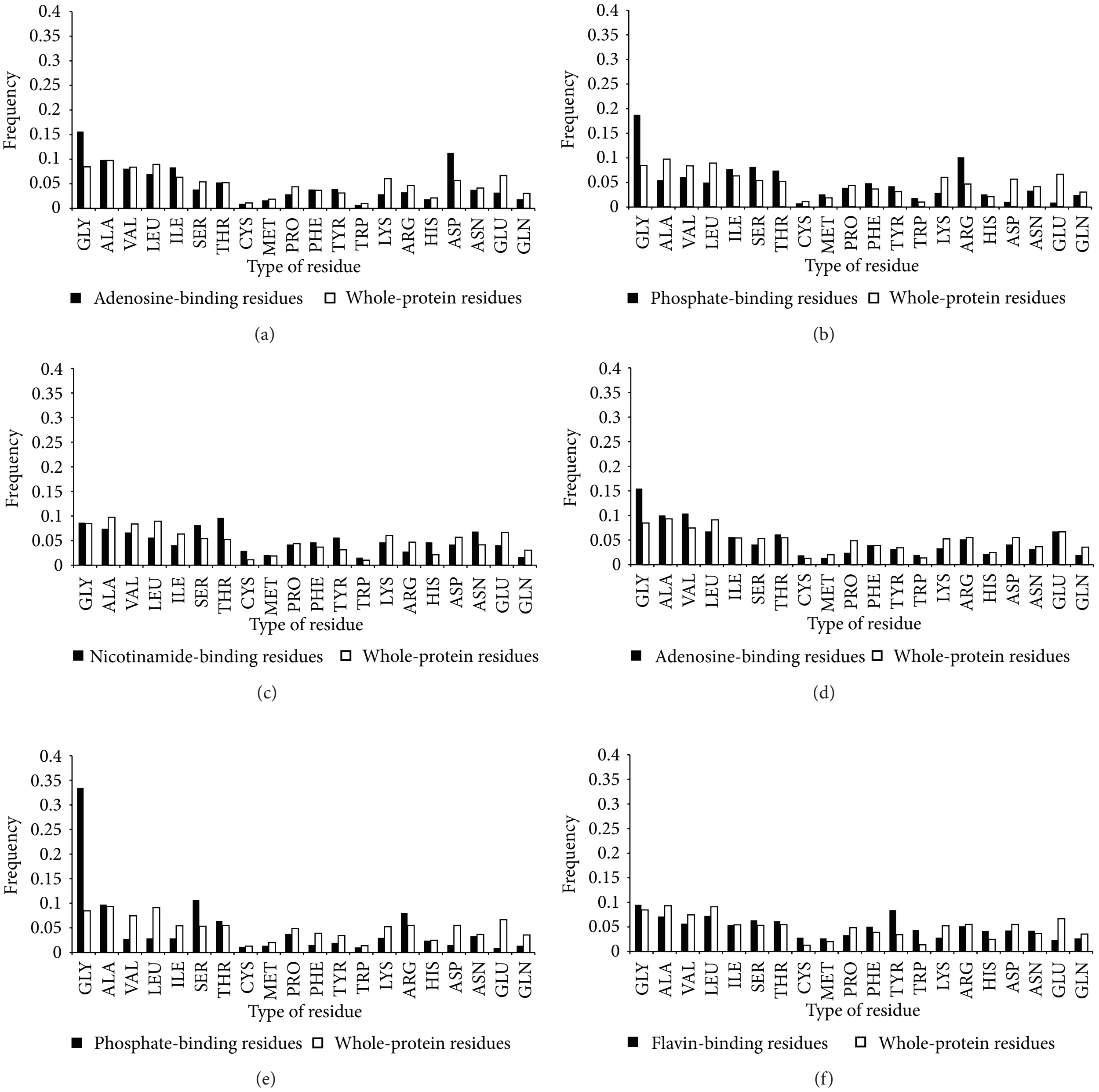


Figure 1: Amino acid frequencies within NAD-/FAD-binding sites. Frequencies within NAD-/FAD-binding sites (black) are compared with whole-protein frequencies (white). (a) Adenosine-binding of NAD. (b) Phosphate-binding of NAD. (c) Nicotinamide-binding of NAD. (d) Adenosine-binding of FAD. (e) Phosphate-binding of FAD. (f) Flavin-binding of FAD. The preferred types of amino acids surrounding the different moiety of NAD/FAD are shown.

Table 3: Comparison between the fragment transformation and SVM methods for predicting NAD-binding-site residues.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Random 1	87.46	86.45	88.48	0.75
Random 2	87.23	85.79	88.67	0.74
Random 3	87.38	85.65	89.11	0.75
Random 4	87.46	86.91	88.01	0.75
Random 5	87.38	86.25	88.51	0.75
Average	**87.38**	**86.21**	**88.56**	**0.75**
SVM [10]	87.25	86.13	88.37	0.75

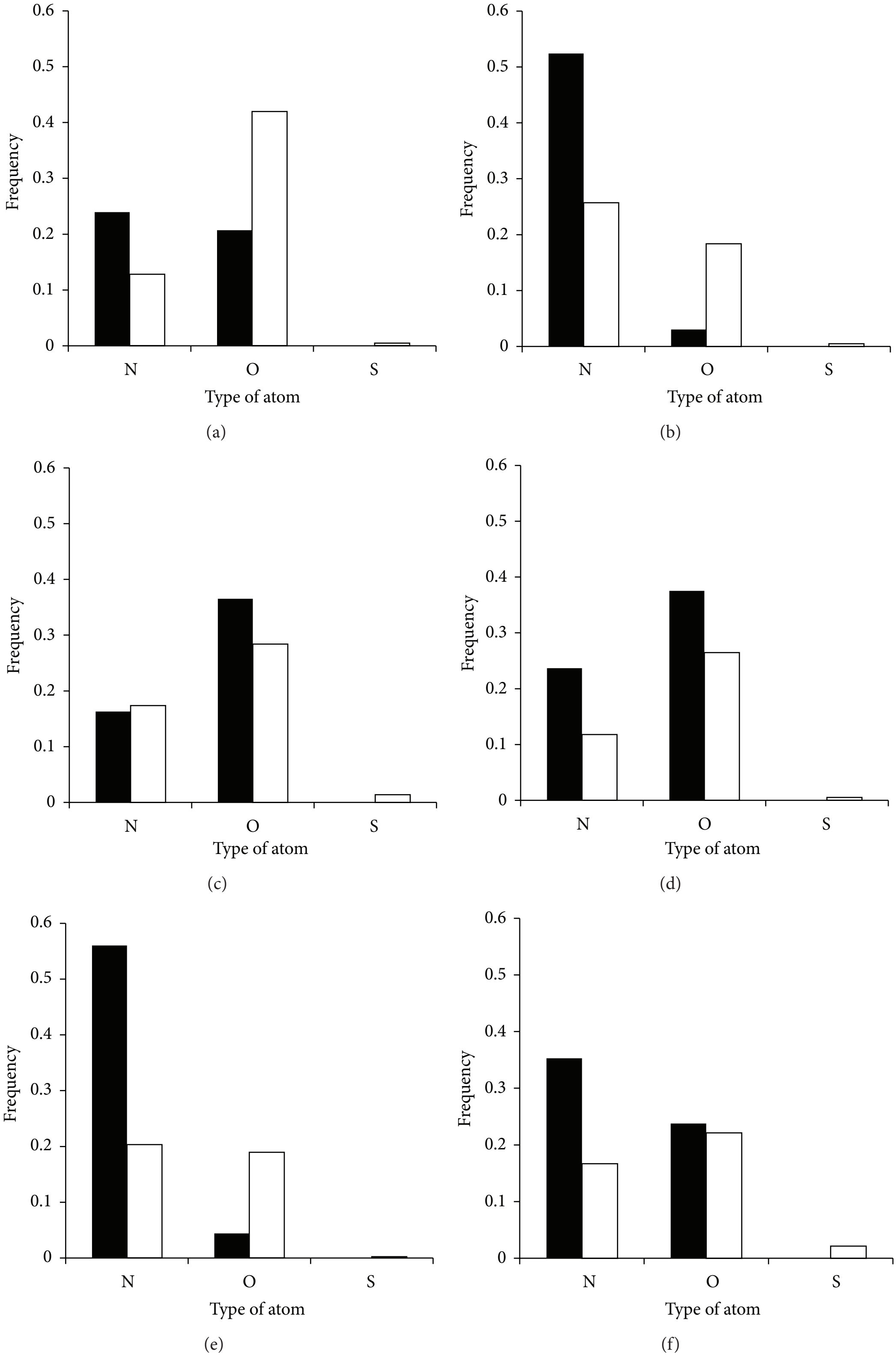
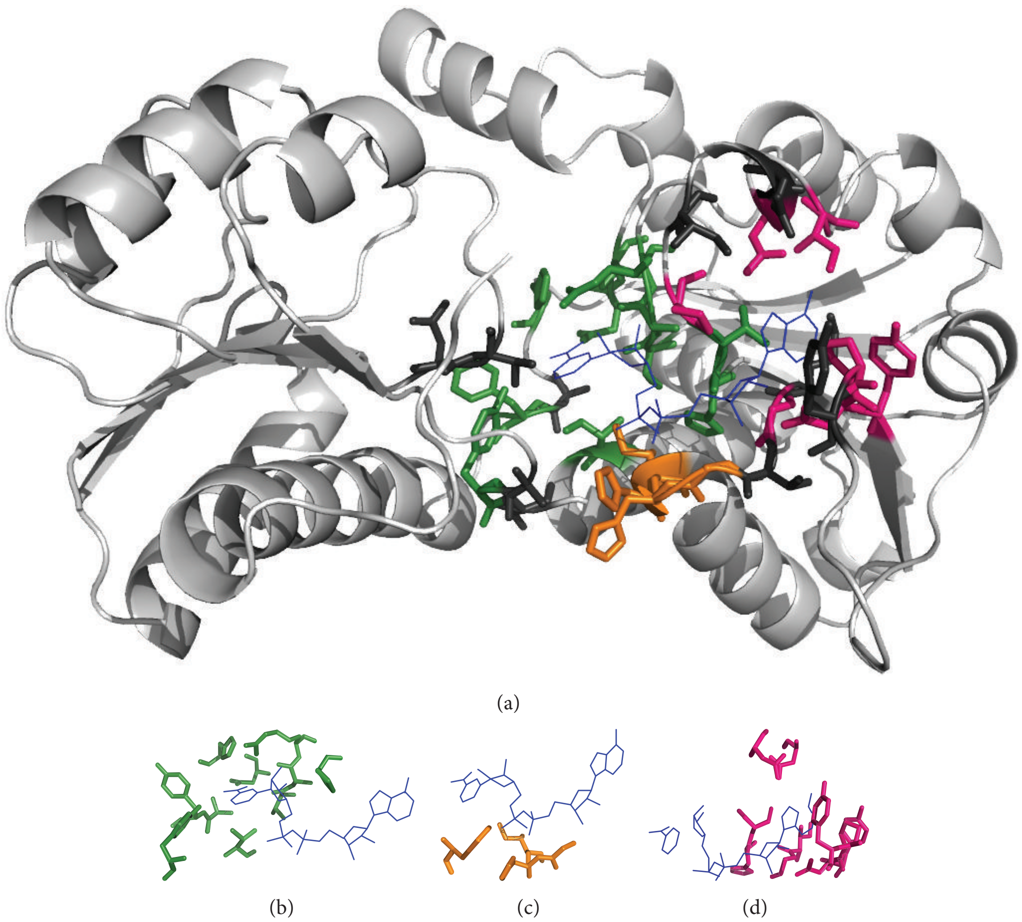


Figure 2: Atom-type frequencies within NAD-/FAD-binding sites. Frequencies for both backbone (black) and side-chain (white) atoms are shown. (a) Adenosine-binding of NAD. (b) Phosphate-binding of NAD. (c) Nicotinamide-binding of NAD. (d) Adenosine-binding of FAD. (e) Phosphate-binding of FAD. (f) Flavin-binding of FAD. The preferred types of atoms surrounding the different moiety of NAD/FAD are shown.



(a)

(b) (c) (d)

Figure 3: Identification of NAD-binding sites. (a) Chain A of D-2-hydroxyisocaproate dehydrogenase (PDB ID:1DXY) was the query protein. Templates were constructed from (b) D-Lactate dehydrogenase (chain A; PDB ID:3KB6), (c) phosphoglycerate dehydrogenase (chain A; PDB ID:1YBA), and (d) C-terminal-binding protein/brefeldin A-ADP ribosylated substrate (chain A; PDB ID:1HKU).

Table 4: Comparison between the fragment transformation and SVM methods for predicting FAD-binding-site residues.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Random 1	87.38	85.68	89.08	0.75
Random 2	87.48	85.73	89.23	0.75
Random 3	87.35	85.55	89.15	0.75
Random 4	87.58	85.73	89.43	0.75
Random 5	87.44	85.73	89.15	0.75
Average	**87.45**	**85.68**	**89.21**	**0.75**
SVM [11]	82.86	83.36	82.36	0.66

PyMOL [23] and color coded: light gray for the query protein; blue lines for the ligand; hot pink, orange, and forest sticks for adenosine-, phosphate-, and nicotinamide-/flavin-binding residues that are predicted correctly; and dark gray sticks for nonbinding residues that are predicted to be binding residues. Our method accurately identified 21 NAD-binding residues within chain A of D-2-hydroxyisocaproate dehydrogenase (PDB ID:1DXY) [24, 25], with ten false positives (Figure 3). Nine nicotinamide-binding residues were identified based on D-Lactate dehydrogenase (chain A; PDB ID:3KB6) [26, 27], three phosphate-binding residues were identified based on phosphoglycerate dehydrogenase (chain A; PDB ID:1YBA) [28], five adenosine-binding residues were identified based on C-terminal-binding protein/brefeldin A-ADP ribosylated substrate (chain A; PDB ID:1HKU) [29], and four were identified based on other protein templates. Our method also accurately predicted 23 NAD-binding residues within chain C of 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (PDB ID:2D4E), with only eight false positives (Figure 4). Nine nicotinamide-binding residues were

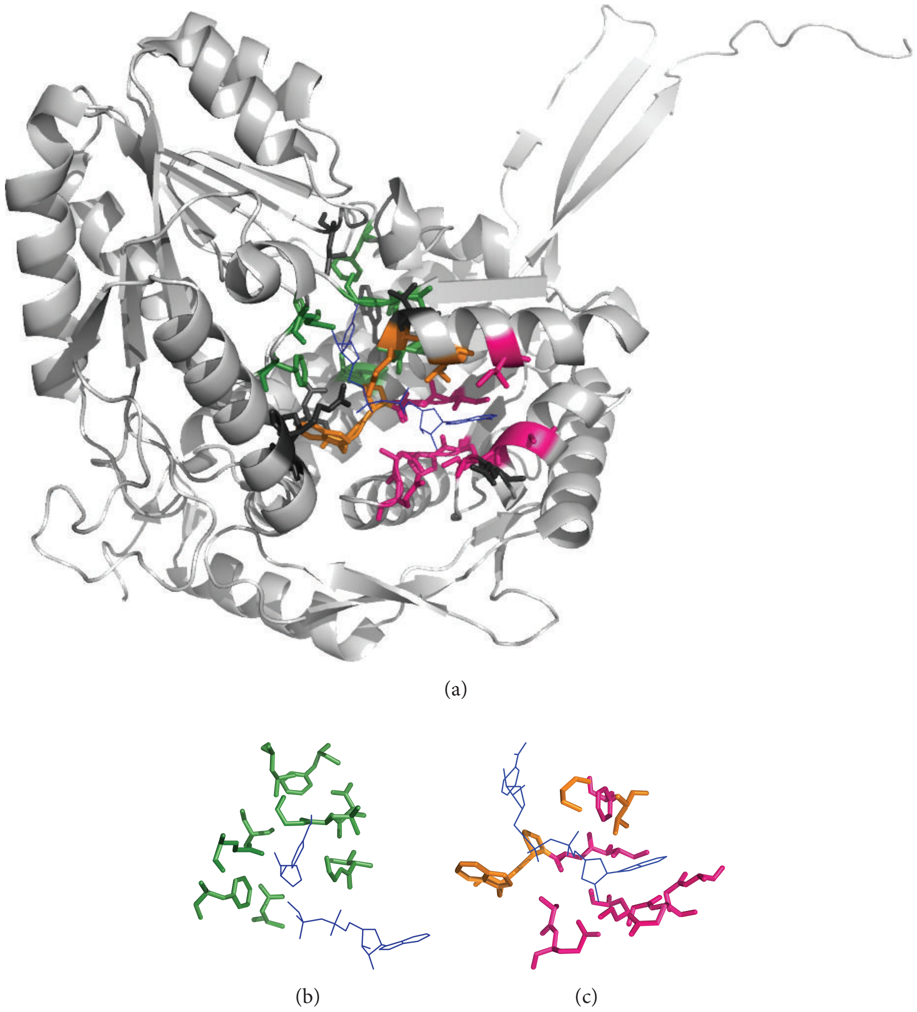


Figure 4: Identification of NAD-binding sites. (a) Chain C of 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (PDB ID:2D4E) was the query protein. Templates were constructed from (b) aldehyde dehydrogenase (chain A; PDB ID:3B4W) and (c) 1-pyrroline-5-carboxylate dehydrogenase (chain A; PDB ID:2EHU).

identified based on aldehyde dehydrogenase (chain A; PDB ID:3B4W), three phosphate-binding and eight adenosine-binding residues were identified based on 1-pyrroline-5-carboxylate dehydrogenase (chain A; PDB ID:2EHU), and three were identified based on other protein templates.

For the FAD-binding proteins, our method accurately predicted chain A of deoxyribodipyrimidine photolyase (PDB ID:1OWL) [30] which contains 24 residues that bind FAD (Figure 5) and only six false positives occurred. Three adenosine-binding residues were identified based on human cryptochrome DASH (chain X; PDB ID:2IJG) [31, 32], six phosphate-binding residues were identified based on photolyase-like domain of cryptochrome 1 (chain A; PDB ID:1U3C) [33], eleven flavin-binding residues were identified based on photolyase (chain A; PDB ID:1IQR) [34], and four were identified based on other protein templates. In addition, 30 FAD-binding residues were accurately predicted within chain H of D-amino acid oxidase (PDB ID:1DDO) [35] with 14 false positives. Five adenosine-binding residues were predicted based on putidaredoxin reductase (chain B; PDB ID:1Q1R) [36, 37], three adenosine-binding and nine flavin-binding residues based on D-amino acid oxidase (chain A; PDB ID:1C0I) [38], three phosphate-binding and five flavin-binding residues based on glycine oxidase (chain B; PDB ID:1NG3) [39], and five based on other protein templates (Figure 6).

3. Discussion

Small molecular cofactors (ligands) are essential for cells to perform numerous biological functions. NAD and FAD, for example, bind to proteins that play critical roles in energy transfer, energy storage, and signal transduction, to name just a few. To understand the mechanism by which these ligands affect protein function, it is important to identify ligand-binding residues within relevant proteins. The experimental identification of these interacting residues is so difficult; however, that computational methods to accomplish this task are in high demand.

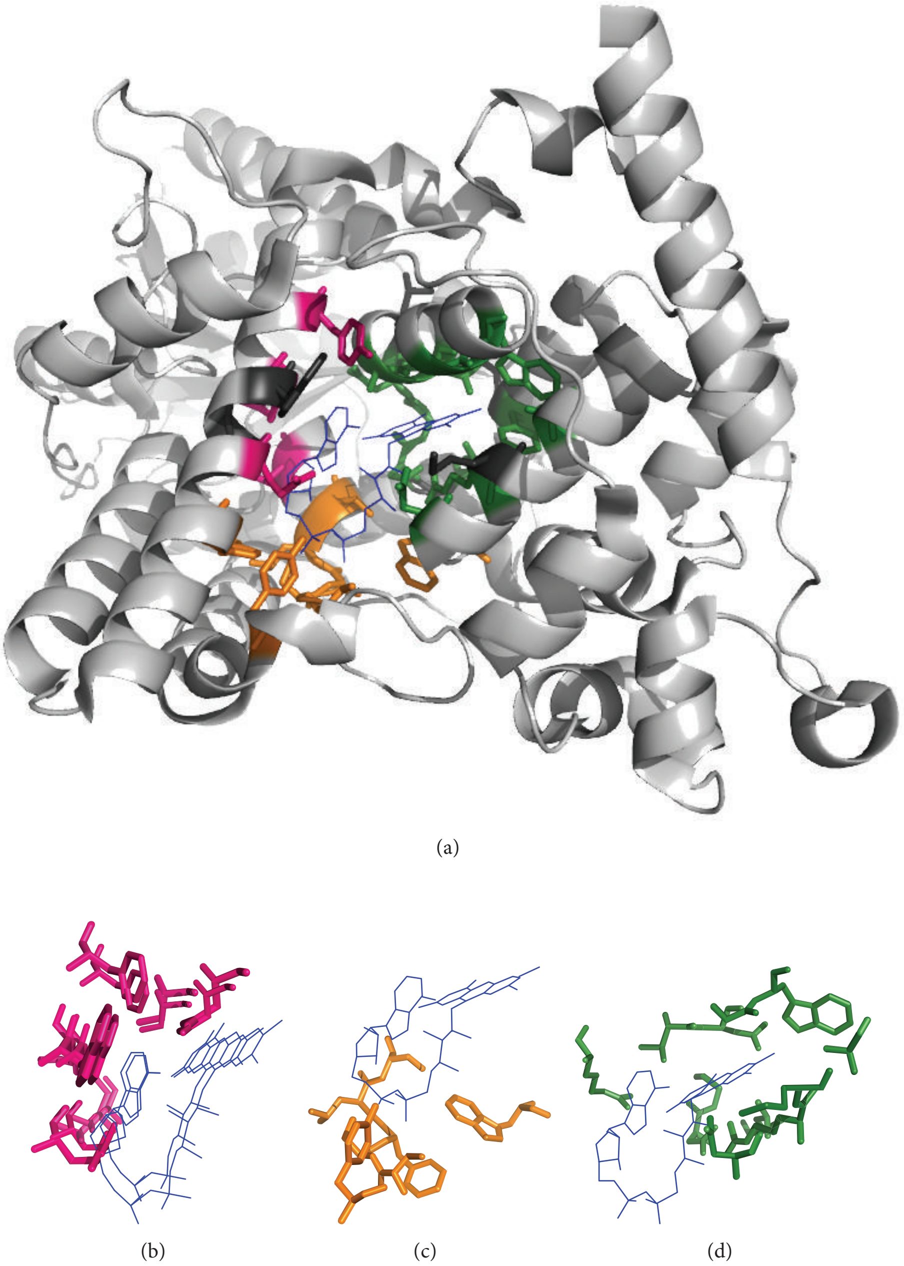
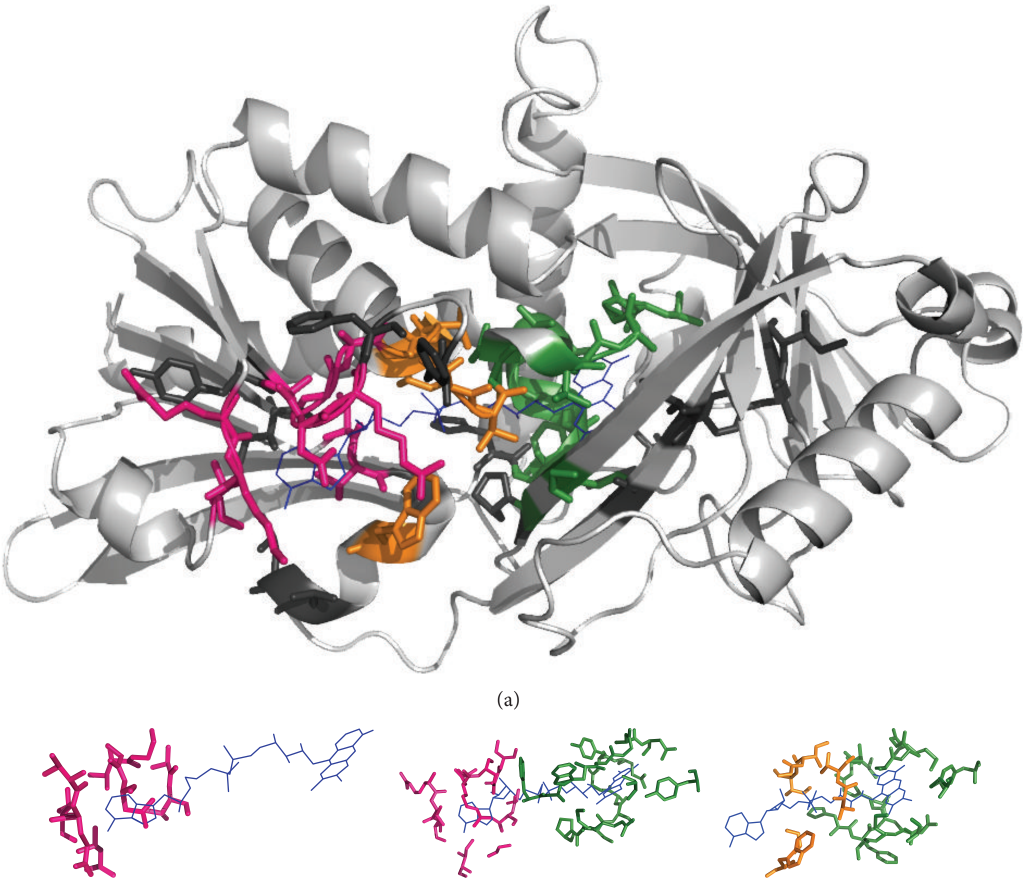


Figure 5: Identification of FAD-binding sites. (a) Chain A of deoxyribodipyrimidine photolyase (PDB ID:1OWL) was the query protein. Templates were constructed from (b) human cryptochrome DASH (chain X; PDB ID:2IJG), (c) photolyase-like domain of cryptochrome 1 (chain A; PDB ID:1U3C), and (d) photolyase (chain A; PDB ID:1IQR).

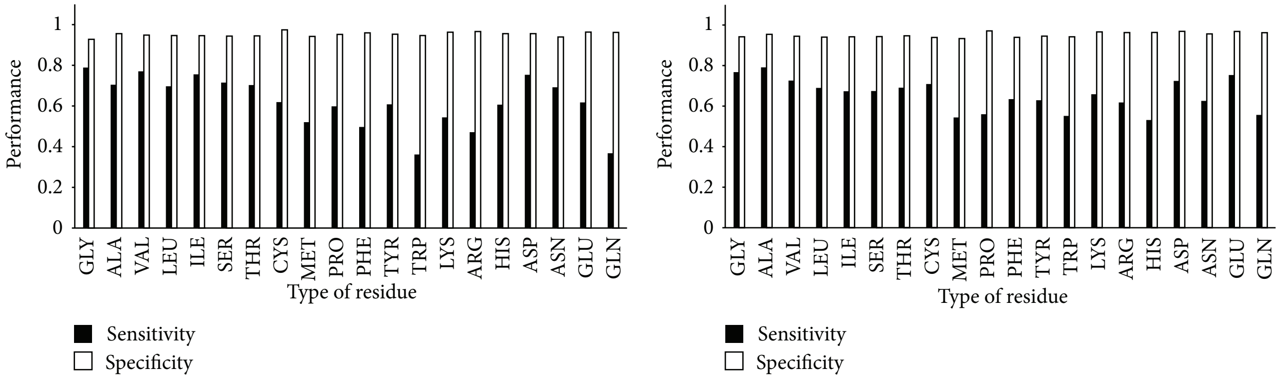
Here we developed a structure comparison method that uses both sequence and structure information to predict NAD-/FAD-binding residues within proteins. This approach also provides valuable information concerning the microenvironment of the protein-ligand interaction. The composition of NAD-/FAD-binding residues that we identified here is generally similar to previous studies [10, 11]. Interestingly, glycine was the most frequent binding residue, binding to NAD through phosphate or adenosine moieties more often than through the nicotinamide moiety. In contrast, arginine preferentially interacted with phosphate moieties and aspartic acid preferentially interacted with adenosine moieties of NAD, whereas threonine, cysteine, and histidine bound to nicotinamide. The most common residue within FAD-binding sites was also glycine, which preferentially bound phosphate and adenosine moieties. Serine interacted with phosphate moieties, whereas cysteine, tyrosine, and tryptophan primarily bound to nicotinamide. By taking advantage of this kind of structural information, details concerning these critical binding sites may be revealed. To investigate the influence of amino acids on prediction performance, the sensitivity and specificity associated with each



(a)

(b) (c) (d)

Figure 6: Identification of FAD-binding sites. (a) Chain H of D-amino acid oxidase (PDB ID:1DDO) was the query protein. Templates were constructed from (b) putidaredoxin reductase (chain B; PDB ID:1Q1R), (c) D-amino acid oxidase (chain A; PDB ID:1C0I), and (d) glycine oxidase (chain B; PDB ID:1NG3).



(a) (b)

Figure 7: Sensitivity and specificity associated with each amino acid in NAD-/FAD-binding-site predictions. (a) NAD. (b) FAD.

residue were calculated (Figure 7). For NAD-binding-site predictions, specificity for each residue was excellent (0.927–0.966), but sensitivity was relatively low for phenylalanine, tryptophan, arginine, and glutamine which were less than 0.5. For FAD-binding sites, all residues achieved high specificity (0.933–0.971) and sensitivity (0.532–0.791). It should be noted that the ratio of NAD-/FAD-binding residues to nonbinding residues is about 1 to 16 in our dataset. This large difference might cause lots of false positives when predicted. That is the reason for high specificity and accuracy but low sensitivity in our prediction results. Hence, the positions of false positive residues in sequence were also investigated; 20% and 25% of false positive residues of NAD- and FAD-binding prediction occurred next to the true positive residues in sequence. It was shown that these residues are also located near the ligand in the coordinate space. If these residues were treated as true positive residues, our prediction results of NAD-binding yielded 71.55% sensitivity and 0.61 MCC at a 5% FPR threshold. Under the same conditions, FAD-binding-site predictions yielded 73.34% sensitivity and an MCC of 0.64. Compared with other prediction methods, ours did not use protein evolutionary information but only used protein structure and did not need to use equal number dataset for training but predicted whole-proteins through comparing structures of template database. Our results yielded excellent prediction performance when analyzing NAD-/FAD-binding residues and thus provide important details concerning the binding-site microenvironment. This approach, therefore, may be used to predict putative NAD-/FAD-binding proteins and the specific residues involved in the interaction.

4. Methods

4.1. Overview. We extracted structures of proteins bound to NAD or FAD from PDB and constructed a database of NAD-/FAD-binding residue templates. Residues that were defined as binding residues by the ligand-binding database BioLiP [40] were included in the template. Query protein structures were then compared with each template in the database using a "leave-one-out" comparison method. The fragment transformation method [22] was used to align query and template structures. After comparing the local protein structure, each residue was assigned a score based on both protein sequence and structure. Sequence similarity was calculated using the BLOSUM62 substitution matrix [41], whereas structural similarity was calculated by measuring the root mean square deviation (RMSD) of the Cα carbons from local structure alignments and using a secondary structure substitution matrix [22] according to the Dictionary of Secondary Structure of Proteins' (DSSP) definition of secondary structure [42]. Residues with an alignment score that exceeded a predetermined threshold were predicted to bind NAD/FAD. This method is illustrated in Figure 8.

4.2. NAD-/FAD-Binding Proteins and Binding Residue Templates. We adopted the same datasets with previous research [10, 11]. All protein complexes were collected from PDB and had pairwise sequence identity <40% by using CD-HIT. Proteins chains that are not involved in NAD/FAD binding were excluded. Residues that were defined as binding or nonbinding residues by using the ligand-binding database BioLiP. The main dataset included 184 and 165 polypeptide chains for NAD and FAD, respectively. Because NAD is composed of a nicotinamide moiety, an adenosine moiety, and a phosphate moiety, binding residues were divided into three groups: nicotinamide binding, adenosine binding, and phosphate binding. FAD-binding sitessimilarly contain flavin-binding residues, adenosine-binding residues, and phosphate-binding residues. Groups of residues that contained more than or equal to two binding residues were considered a binding residue template (see Figures 9 and 10).

4.3. The Fragment Transformation Method. We used the fragment transformation method to align NAD-/FAD-binding residues. Each residue was treated as an individual unit and was used to align the query protein S with the binding template T. The structural unit consists of a triplet formed by the N–C$_\alpha$–C atoms within a given residue. S denotes the query protein of length m, and T denotes the template of n residues. The query protein S of length m and the template T of n residues can therefore be expressed in terms of triplets as $S = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ and $T = \{\tau_1, \tau_2, \ldots, \tau_n\}$, where $\sigma_i = (p_N, p_{C\alpha}, p_C)$, $\tau_j = (q_N, q_{C\alpha}, q_C)$, and p and q are PDB coordinates for each atom.

A matrix of dimensions $m \times n$ was then constructed for the residues of S and T as

$$M = \begin{vmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,n} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ M_{m,1} & M_{m,2} & \cdots & M_{m,n} \end{vmatrix}, \tag{1}$$

where the element M_{ij} is a rigid-body transformation matrix that transforms the triplet σ_i to τ_j (i.e., $M_{ij}\sigma_i = \tau_j$).

4.4. Performing Triplet Clustering. D_{kl}^{ij} is the Cartesian distance between the target τ_l and the transformed triplet $M_{ij}\sigma_k$, providing a measure of how similarly the triplet pairs (σ_i, τ_j) and (σ_k, τ_l) are oriented. This allows clustering of triplet fragments using the single-linkage algorithm [43] as follows. If for two triplet pairs, (σ_i, τ_j) and (σ_k, τ_l), $D_{kl}^{ij} < D_0$, $i \neq k$ and $j \neq l$, then the triplets are clustered. Let G_1 and G_2 be two clusters, with the first containing (σ_i, τ_j) and (σ_k, τ_l) and the second containing $(\sigma_{i'}, \tau_{j'})$ and $(\sigma_{k'}, \tau_{l'})$. If $D_{k'l'}^{ij} < D_0$, then G_1 and G_2 are merged to form a new cluster G_3, where $G_3 = G_1 \cup G_2$. These procedures are performed iteratively until no new clusters can be formed. For each final cluster G_μ, we can obtain the transformation matrix $M_{k,l}^{\mu}$ and aligned substructure pair $S_\mu = \bigcup_{\sigma_k \in G_\mu} \sigma_k$ and $T_\mu = \bigcup_{\tau_l \in G_\mu} \tau_l$, where G_μ has the minimum Cartesian distance when using $M_{k,l}^{\mu}$.

4.5. Scoring Function. For each residue i, the binding score C_i is defined as

$$C_i = \underset{\sigma_i \in G_\mu}{\mathrm{MAX}} \left(\varepsilon_\mu \times C_\mu^R \times C_\mu^B \times C_\mu^D \right), \tag{2}$$

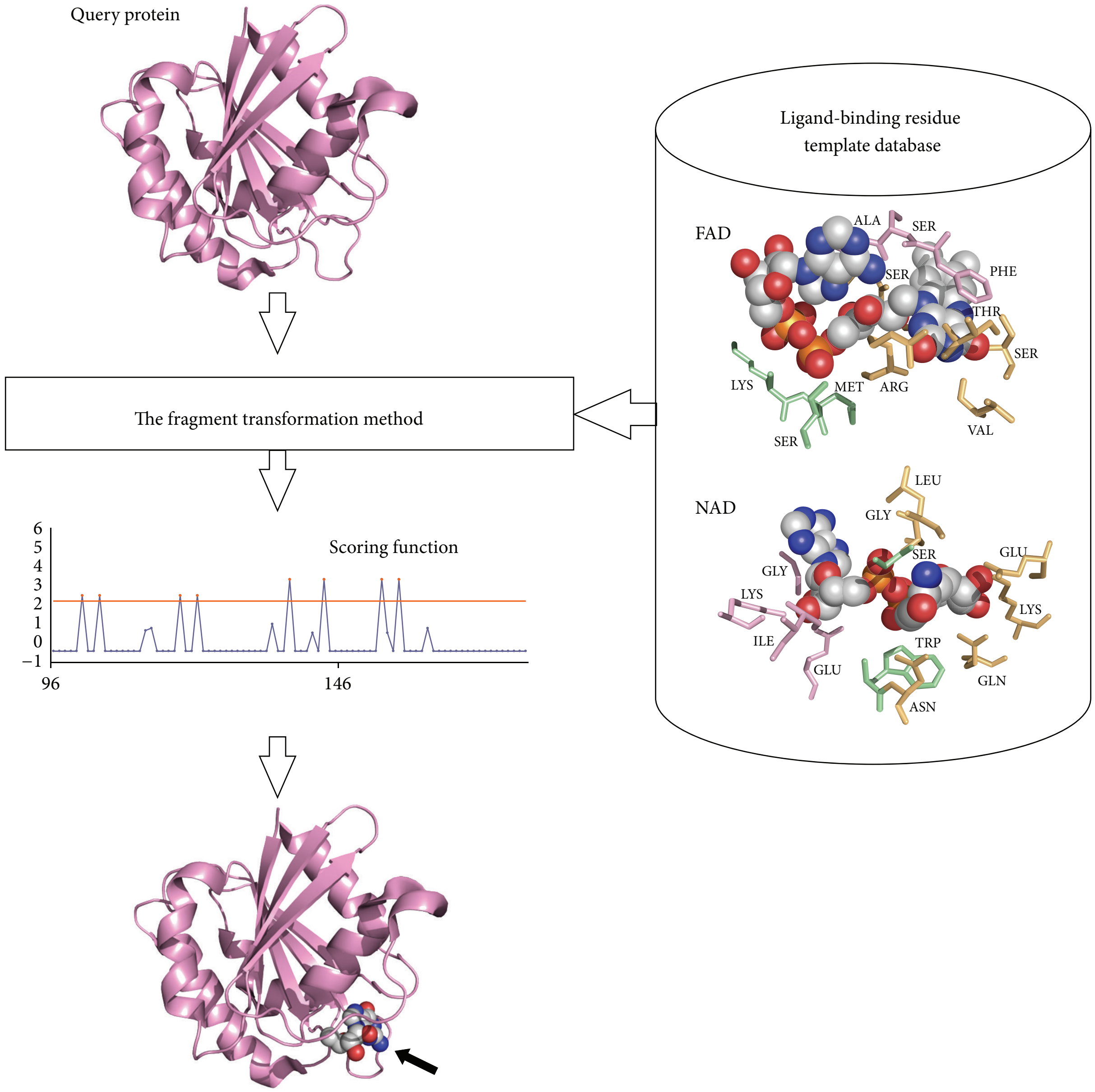


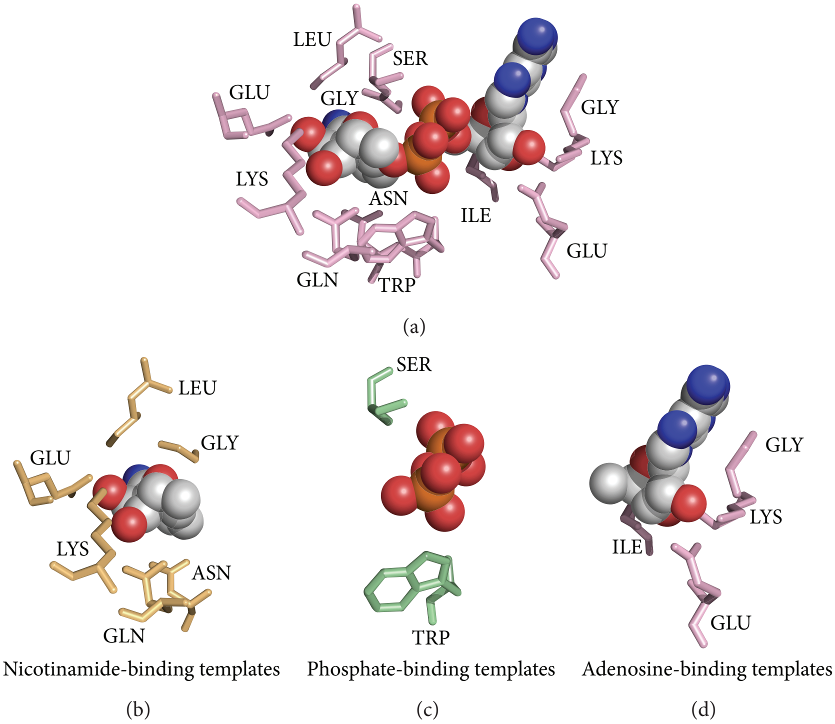
Figure 8: Schematic of the method for predicting NAD-/FAD-binding sites.

where ε_μ is the number of triplets of S_μ (i.e., the aligned residues of the query structure). The alignment scores C_μ^R, C_μ^B, and C_μ^D are defined as

$$C_\mu^R = \frac{1}{1 + \mathrm{RMSD}\left(S_\mu, T_\mu\right)},$$
$$C_\mu^B = \frac{\mathrm{BLOSUM}\left(S_\mu, T_\mu\right)}{\mathrm{BLOSUM}\left(T_\mu, T_\mu\right)} + 1 \qquad (3)$$
$$C_\mu^D = \frac{\mathrm{DSSP}\left(S_\mu, T_\mu\right)}{\mathrm{DSSP}\left(T_\mu, T_\mu\right)} + 1,$$

where RMSD (S_μ, T_μ) is the RMSD of all C_α atoms between S_μ and T_μ, BLOSUM (S_μ, T_μ) is the sequence alignment score between S_μ and T_μ calculated using the BLOSUM62 [41] substitution matrix, BLOSUM (T_μ, T_μ) is the maximum sequence alignment score of T_μ, DSSP (S_μ, T_μ) represents the secondary structure alignment score based on a construction substitution matrix [22] using the definition of DSSP [42] between S_μ and T_μ, and DSSP (T_μ, T_μ) is the maximum secondary structure alignment score of T_μ. The value of RMSD (S_μ, T_μ) should be <3 Å.

For each residue i, we predict a geometric center Θ_i^ω of the ligand by $\Theta_i^\omega = M_{k,l}^{\mu\ -1} L_\omega$, where L_ω is the geometric center of the binding template type ω in template T. ω represents the three moieties of NAD/FAD: nicotinamide, adenosine,

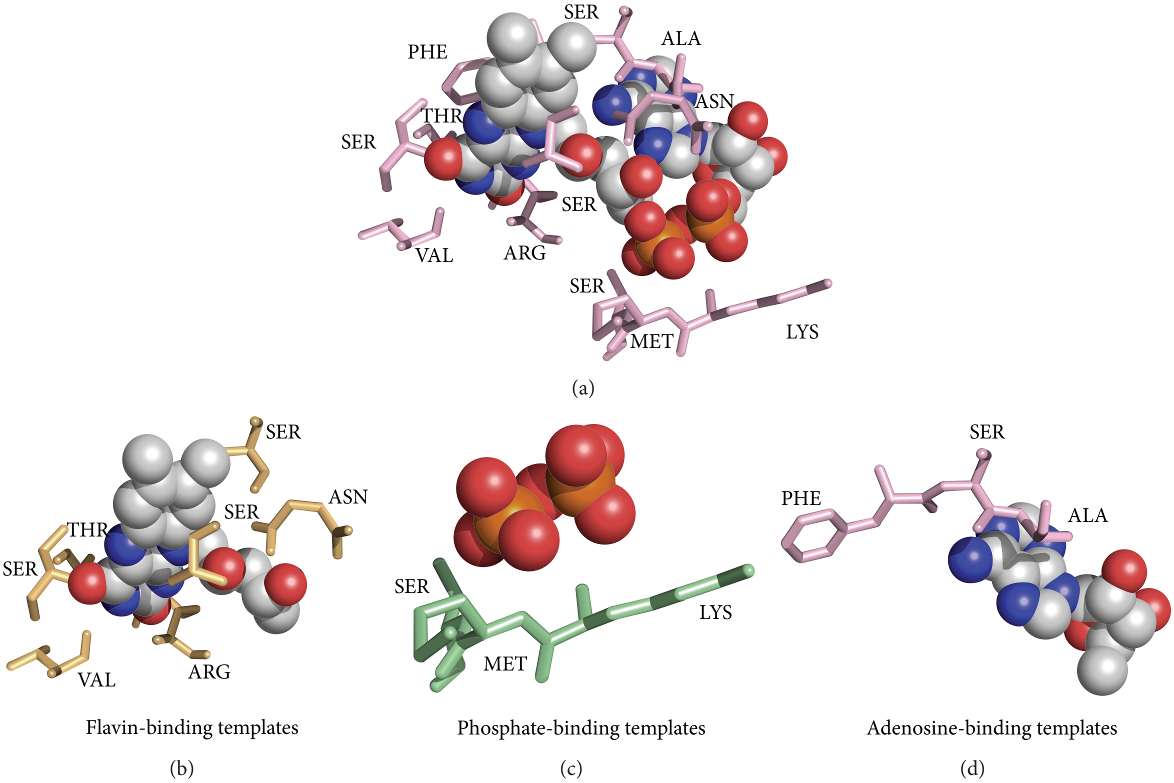


Nicotinamide-binding templates
(b)

Phosphate-binding templates
(c)

Adenosine-binding templates
(d)

Figure 9: NAD-binding residue templates. (a) The entire NAD-binding template. (b) Nicotinamide-binding templates. (c) Phosphate-binding templates. (d) Adenosine-binding templates.



Flavin-binding templates
(b)

Phosphate-binding templates
(c)

Adenosine-binding templates
(d)

Figure 10: FAD-binding residue templates. (a) The entire FAD-binding template. (b) Flavin-binding templates. (c) Phosphate-binding templates. (d) Adenosine-binding templates.

and phosphate for NAD; flavin, adenosine, and phosphate for FAD. The binding score C_k is added to C_i if the distance between Θ_i^{ω} and $\Theta_k^{\omega'}$ is between 3 and 9 Å, and $\omega \neq \omega'$. Finally, the normalized binding score Z_i^C is calculated as

$$Z_i^C = \frac{C_i - \overline{C}}{SD_C}, \quad (4)$$

where $\overline{C}$ and SD_C denote the mean and standard deviation, respectively, of the binding score C_i.

4.6. Performance Assessment. The accuracy of predicting NAD-/FAD-binding sites wasdefined as the number of true positives and true negatives and was evaluated using a leave-one-out approach. Accuracy (ACC), the true positive rate (TPR), and the false positive rate (FPR) were calculated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values as follows:

$$\begin{aligned} \mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{TPR} = \text{Sensitivity} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad (5) \\ \mathrm{FPR} = 1 - \text{Specificity} &= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}. \end{aligned}$$

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Chih-Hao Lu and Chin-Sheng Yu developed and implemented the methods; Chih-Hao Lu, Chin-Sheng Yu, Yu-Feng Lin, and Jin-Yi Chen carried out the analysis; Chih-Hao Lu and Yu-Feng Lin drafted the paper; Chih-Hao Lu supervised the work. All the authors have read and approved the content of the final paper. Chih-Hao Lu and Chin-Sheng Yu contributed equally to this work.

Acknowledgment

This work was supported by Grants from the National Science Council (NSC 99-2113-M-039-002-MY2) and China Medical University (CMU99-N2-02), Taiwan, to Chih-Hao Lu. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper. The authors are grateful to Yeong-Shin Lin (National Chiao Tung University, Taiwan) for invaluable comments.

References

[1] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[2] A. Wilkinson, J. Day, and R. Bowater, "Bacterial DNA ligases," *Molecular Microbiology*, vol. 40, no. 6, pp. 1241–1248, 2001.

[3] A. Bürkle, "Physiology and pathophysiology of poly(ADP-ribosyl)ation," *BioEssays*, vol. 23, no. 9, pp. 795–806, 2001.

[4] Q. Zhang, D. W. Piston, and R. H. Goodman, "Regulation of corepressor function by nuclear NADH," *Science*, vol. 295, no. 5561, pp. 1895–1897, 2002.

[5] J. S. Smith and J. D. Boeke, "An unusual form of transcriptional silencing in yeast ribosomal DNA," *Genes and Development*, vol. 11, no. 2, pp. 241–254, 1997.

[6] R. M. Anderson, K. J. Bitterman, J. G. Wood et al., "Manipulation of a nuclear NAD$^+$ salvage pathway delays aging without altering steady-state NAD$^+$ levels," *The Journal of Biological Chemistry*, vol. 277, no. 21, pp. 18881–18890, 2002.

[7] J. Rutter, M. Reick, L. C. Wu, and S. L. McKnight, "Regulation of crock and NPAS2 DNA binding by the redox state of NAD cofactors," *Science*, vol. 293, no. 5529, pp. 510–514, 2001.

[8] K. Chen, M. J. Mizianty, and L. Kurgan, "Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors," *Bioinformatics*, vol. 28, no. 3, pp. 331–341, 2012.

[9] M. Saito, M. Go, and T. Shirai, "An empirical approach for detecting nucleotide-binding sites on proteins," *Protein Engineering, Design and Selection*, vol. 19, no. 2, pp. 67–75, 2006.

[10] H. R. Ansari and G. P. S. Raghava, "Identification of NAD interacting residues in proteins," *BMC Bioinformatics*, vol. 11, article 160, 2010.

[11] N. K. Mishra and G. P. S. Raghava, "Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information," *BMC Bioinformatics*, vol. 11, article S48, no. 1, 2010.

[12] Z.-P. Liu, L.-Y. Wu, Y. Wang, X.-S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics*, vol. 26, no. 13, pp. 1616–1622, 2010.

[13] L. Wang, Z. P. Liu, X. S. Zhang, and L. Chen, "Prediction of hot spots in protein interfaces using a random forest model with hybrid features," *Protein Engineering, Design and Selection*, vol. 25, no. 3, pp. 119–126, 2012.

[14] J. S. Chauhan, N. K. Mishra, and G. P. S. Raghava, "Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information," *BMC Bioinformatics*, vol. 11, article 301, 2010.

[15] A. Roy and Y. Zhang, "Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement," *Structure*, vol. 20, no. 6, pp. 987–997, 2012.

[16] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 14, pp. 5441–5446, 2008.

[17] J. Yang, A. Roy, and Y. Zhang, "Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment," *Bioinformatics*, vol. 29, no. 20, pp. 2588–2595, 2013.

[18] Y. T. Yan and W.-H. Li, "Identification of protein functional surfaces by the concept of a split pocket," *Proteins: Structure, Function and Bioinformatics*, vol. 76, no. 4, pp. 959–976, 2009.

[19] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[20] S. Govindarajan and R. A. Goldstein, "Evolution of model proteins on a foldability landscape," *Proteins*, vol. 29, no. 4, pp. 461–466, 1997.

[21] G. Parisi and J. Echave, “Structural constraints and emergence of sequence patterns in protein evolution,” *Molecular Biology and Evolution*, vol. 18, no. 5, pp. 750–756, 2001.

[22] C. H. Lu, Y. S. Lin, Y. C. Chen, C. S. Yu, S. Y. Chang, and J. K. Hwang, “The fragment transformation method to detect the protein structural motifs,” *Proteins: Structure, Function and Genetics*, vol. 63, no. 3, pp. 636–643, 2006.

[23] L. Schrodinger, *The PyMOL Molecular Graphics System, Version 1.3r1*, 2010.

[24] U. Dengler, K. Niefind, M. Kieß, and D. Schomburg, “Crystal structure of a ternary complex of D-2-hydroxy-isocaproate dehydrogenase from Lactobacillus casei, NAD^+ and 2-oxoisocaproate at 1.9 Å resolution,” *Journal of Molecular Biology*, vol. 267, no. 3, pp. 640–660, 1997.

[25] E. Gross, C. S. Sevier, A. Vala, C. A. Kaiser, and D. Fass, “A new FAD-binding fold and intersubunit disulfide shuttle in the thiol oxidase Erv2p,” *Nature Structural Biology*, vol. 9, no. 1, pp. 61–67, 2002.

[26] S. V. Antonyuk, R. W. Strange, M. J. Ellis et al., “Structure of d-lactate dehydrogenase from *Aquifex aeolicus* complexed with NAD^+ and lactic acid (or pyruvate),” *Acta Crystallographica F: Structural Biology and Crystallization Communications*, vol. 65, part 12, pp. 1209–1213, 2009.

[27] C. K. Wu, T. A. Dailey, H. A. Dailey, B. C. Wang, and J. P. Rose, “The crystal structure of augmenter of liver regeneration: a mammalian FAD-dependent sulfhydryl oxidase,” *Protein Science*, vol. 12, no. 5, pp. 1109–1118, 2003.

[28] J. R. Thompson, J. K. Bell, J. Bratt, G. A. Grant, and L. J. Banaszak, “Vmax regulation through domain and subunit changes. The active form of phosphoglycerate dehydrogenase,” *Biochemistry*, vol. 44, no. 15, pp. 5763–5773, 2005.

[29] M. Nardini, S. Spanò, C. Cericola et al., “CtBP/BARS: a dual-function protein involved in transcription co-repression and Golgi membrane fission,” *EMBO Journal*, vol. 22, no. 12, pp. 3122–3130, 2003.

[30] R. Kort, H. Komori, S. I. Adachi, K. Miki, and A. Eker, “DNA apophotolyase from Anacystis nidulans: 1.8 Å structure, 8-HDF reconstitution and X-ray-induced FAD reduction,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, no. 7, pp. 1205–1213, 2004.

[31] Y. Huang, R. Baxter, B. S. Smith, C. L. Partch, C. L. Colbert, and J. Deisenhofer, “Crystal structure of cryptochrome 3 from Arabidopsis thaliana and its implications for photolyase activity,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 47, pp. 17701–17706, 2006.

[32] J. B. Thoden, T. M. Wohlers, J. L. Fridovich-Keil, and H. M. Holden, “Molecular basis for severe epimerase deficiency galactosemia. X-ray structure of the human V94M-substituted UDP-galactose 4-epimerase,” *The Journal of Biological Chemistry*, vol. 276, no. 23, pp. 20617–20623, 2001.

[33] C. A. Brautigam, B. S. Smith, Z. Ma et al., “Structure of the photolyase-like domain of cryptochrome 1 from Arabidopsis thaliana,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 33, pp. 12142–12147, 2004.

[34] H. Komori, R. Masui, S. Kuramitsu et al., “Crystal structure of thermostable DNA photolyase: pyrimidine-dimer recognition mechanism,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13560–13565, 2001.

[35] F. Todone, M. A. Vanoni, A. Mozzarelli et al., “Active site plasticity in D-amino acid oxidase: a crystallographic analysis,” *Biochemistry*, vol. 36, no. 19, pp. 5853–5860, 1997.

[36] I. F. Sevrioukova, H. Li, and T. L. Poulos, “Crystal structure of putidaredoxin reductase from Pseudomonas putida, the final structural component of the cytochrome P450cam monooxygenase,” *Journal of Molecular Biology*, vol. 336, no. 4, pp. 889–902, 2004.

[37] S. Y. Song, Y. B. Xu, Z. J. Lin, and C. L. Tsou, “Structure of active site carboxymethylated D-glyceraldehyde-3-phosphate dehydrogenase from Palinurus versicolor,” *Journal of Molecular Biology*, vol. 287, no. 4, pp. 719–725, 1999.

[38] L. Pollegioni, K. Diederichs, G. Molla et al., “Yeast D-amino acid oxidase: structural basis of its catalytic properties,” *Journal of Molecular Biology*, vol. 324, no. 3, pp. 535–546, 2002.

[39] E. C. Settembre, P. C. Dorrestein, J. H. Park, A. M. Augustine, T. P. Begley, and S. E. Ealick, “Structural and mechanistic studies on thiO, a glycine oxidase essential for thiamin biosynthesis in Bacillus subtilis,” *Biochemistry*, vol. 42, no. 10, pp. 2971–2981, 2003.

[40] J. Yang, A. Roy, and Y. Zhang, “BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions,” *Nucleic Acids Research*, vol. 41, no. 1, pp. D1096–D1103, 2013.

[41] S. Henikoff and J. G. Henikoff, “Amino acid substitution matrices from protein blocks,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915–10919, 1992.

[42] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[43] J. C. Gower and G. J. S. Ross, “Minimum spanning trees and single-linkage cluster analysis,” *Journal of the Royal Statistical Society*, vol. 18, no. 1, article 11, 1969.