Journal of Theoretical Biology 343 (2014) 154–161

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Prediction of FMN-binding residues with three-dimensional probability distributions of interacting atoms on protein surfaces

Rajasekaran Mahalingam [a,*], Hung-Pin Peng [a,b,c], An-Suei Yang [a,**]

[a] Genomics Research Center, Academia Sinica, 128 Academia Rd., Sec. 2, Nankang Dist., Taipei 115, Taiwan
[b] Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan
[c] Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan

HIGHLIGHTS

- First structure-based approach for prediction of protein–FMN interaction.
- Does not require evolutionary information for the prediction.
- Useful in annotating proteins structures of unknown function and computational protein models.

ARTICLE INFO

Article history:
Received 26 July 2013
Received in revised form
29 October 2013
Accepted 30 October 2013
Available online 7 November 2013

Keywords:
Structure-based
Computational method
Machine learning
Drug discovery
Functional annotation

ABSTRACT

Flavin mono-nucleotide (FMN) is a cofactor which is involved in many biological reactions. The insights on protein–FMN interactions aid the protein functional annotation and also facilitate in drug design. In this study, we have established a new method, making use of an encoding scheme of the three-dimensional probability density maps that describe the distributions of 40 non-covalent interacting atom types around protein surfaces, to predict FMN-binding sites on protein surfaces. One machine learning model was trained for each of the 30 protein atom types to predict tentative FMN-binding sites on protein structures. The method's capability was evaluated by five-fold cross-validation on a dataset containing 81 non-redundant FMN-binding protein structures and further tested on independent datasets of 30 and 15 non-redundant protein structures respectively. These predictions achieved an accuracy of 0.94, 0.94 and 0.96 with the Matthews correlation coefficient (MCC) of 0.53, 0.53 and 0.65 respectively for the three protein structure sets. The prediction capability is superior to the existing method. This is the first structure-based approach that does not rely on evolutionary information for predicting FMN-interacting residues. The webserver for the prediction is available at http://ismblab.genomics.sinica.edu.tw/.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

FMN is an essential cofactor in flavoproteins which are involved in (i) redox reactions in the energy producing metabolic pathways and (ii) non-redox reactions in which FMN acts as acid or base in the covalent-intermediate formation (Mansoorabadi et al., 2007; Serrano et al., 2012). Flavodoxin, which is one of the flavoproteins, is considered as one of the potential drug targets against microbial infections because it plays a critical role in the electron transfer system of pathogenic bacteria but not in mammals. *Helicobacter pylori* flavodoxin acts as an electron acceptor in pyruvate metabolic pathway, and thus inhibition of this protein can affect the bacterial survival (Cremades et al., 2005). Chorismate synthase is another FMN-binding protein involved in shikimate pathway (Macheroux et al., 1999) and is considered as a primary target in developing antibacterial therapeutics against tuberculosis (Fernandes et al., 2007). Hence, identification of FMN-binding proteins and binding site residues can aid in the drug discovery processes for antimicrobial therapeutics.

A computational method for predicting the FMN-binding residues on proteins would greatly facilitate defining FMN-binding sites on protein structures. Computational methods have been developed to predict FMN (Wang et al., 2012), flavin adenine dinucleotide (FAD) (Mishra and Raghava, 2010) and nicotinamide adenine dinucleotide (NAD) (Ansari and Raghava, 2010) binding

* Corresponding author. Current address: Department of Physiology and Biophysics, School of Medicine, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, USA. Tel.: +1 216 368 8654.
** Corresponding author. Tel.: +886 2 2787 1232.
E-mail addresses: rajasekaran.mahalingam@case.edu (R. Mahalingam), yangas@gate.sinica.edu.tw (A.-S. Yang).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.10.020

sites. These computational methods are reasonably successful in their respective predictions. Nevertheless, they are all sequence-based predictors relying on evolutionary information. Consequently, these methods may have difficulty in predicting binding site in orphan proteins, which shares very low sequence similarity with existing proteins. A structure-based method which does not rely on evolutionary information has yet to be developed. In this study, we have developed such a structure-based method for the prediction of FMN-interacting residues on protein surfaces.

This method uses machine learning approach to predict FMN-binding sites on protein surfaces by recognizing characteristic interacting atom distribution patterns associated with the FMN binding. The basic principle has been already applied to predict the protein–protein (Chen et al., 2012) and protein–carbohydrate (Tsai et al., 2012) interactions successfully. Here we extend this approach to predicting FMN-interacting residues. In the prediction, protein surface atoms were first categorized into 30 atom types and one machine learning model was trained for each of the atom types. The input attributes for the machine learning algorithm were normalized distance–weighted sum of three-dimensional probability density maps (PDMs) of 40 interacting atom types (30 atom types from protein, one from water and nine from FMN) on the protein surfaces. The PDMs around the query protein atoms for the protein interacting atom types and water have been described in previous publications (Chen et al., 2012; Tsai et al., 2012); the PDMs for the nine FMN interacting atom types were constructed with the protein–FMN interacting atom pairs from the dataset of 192 FMN–protein complex structures. The machine learning algorithm learned the patterns of the attributes to distinguish the binding atoms from the non-binding atoms on the protein surfaces. We evaluated our predictor performance on the training dataset P81 as well independent test sets P30 and P15 (Wang et al., 2012). The results indicate that our approach is the best method for predicting the FMN-binding sites on protein structures.

2. Materials and methods

2.1. Dataset

The training and test datasets except P15 were obtained from Wang et al. (2012). The autor obtained 111 protein chains from PDB (Berman et al., 2002). Then they randomly selected 30 proteins (P30) for the independent test and the remaining 81 protein chains (P81) were used as a training set. For the P15, we extracted protein–FMN complexes from PDB and then used PISCES program (Wang and Dunbrack, 2003) to remove structures which has the sequence identity more than 10% with P81 and P30 datasets that finally yielded 15 protein chains.

2.2. Construction of three-dimensional probability density maps on protein surfaces

The detailed method for the PDMs construction has been discussed previously (Chen et al., 2012; Tsai et al., 2012; Yu et al., 2012). In brief, the interacting atom types from protein, water, and FMN are shown in Table 1. The PDMs for these interacting atom types were constructed with interacting atoms retrieved from the interacting atom database described previously (Chen et al., 2012; Tsai et al., 2012; Yu et al., 2012). The interacting atom database for protein–FMN interacting atom pairs was constructed with the dataset of 192 protein–FMN complexes.

Table 1
Protein and FMN atom types.

ID #	Atom type	Radius (Å)	Description
1	NH1	1.65	Backbone NH
2	C	1.76	Backbone C
3	CH1E	1.87	Backbone CA (exc. Gly)
4	O	1.40	Backbone O
5	CH0	1.76	Arg CZ, Asn CG, Asp CG, Gln CD, Glu CD
6	CH1S	1.87	Sidechain CH1: Ile CB, Leu CG, Thr CB, Val CB
7	CH2E	1.87	Tetrahedral CH2 (except CH2P,CH2G) All CB
8	CH3E	1.87	Tetrahedral CH3
9	CR1E	1.76	Aromatic CH (except CR1W, CRHH, CR1H)
10	OH1	1.40	Alcohol OH (Ser OG, Thr OG1, Tyr OH)
11	OC	1.40	Carboxyl O (Asp OD1, OD2, Glu OE1, OE2)
12	OS	1.40	Sidechain O: Asn OD1, Gln OE1
13	CH2G	1.87	Gly CA
14	CH2P	1.87	Pro CB, CG, CD
15	NH1S	1.65	Sidechain NH: Arg NE, His ND1, NE1, Trp NE1
16	NC2	1.65	Arg NH1, NH2
17	NH2	1.65	Asn ND2, Gln NE2
18	CR1W	1.76	Trp CZ2, CH2
19	CY2	1.76	Tyr CZ
20	SC	1.85	Cys S
21	CF	1.76	Phe CG
22	SM	1.85	Met S
23	CY	1.76	Tyr CG
24	CW	1.76	Trp CD2, CE2
25	CRHH	1.76	His CE1
26	NH3	1.50	Lys NZ
27	CR1H	1.76	His CD2
28	C5	1.76	His CG
29	N	1.65	Pro N
30	C5W	1.76	Trp CG
31	HOH	1.40	Water
32	P	2.10	Phosphate
33	O1O2	1.68	Phosphate oxygen
34	N1N9	1.82	Base
35	N	1.82	ring
36	C	1.90	ring
37	N	1.82	link
38	O	1.68	link
39	C.3	1.90	Sp3 carbon
40	O.3	1.68	Sp3 oxygen

The protein atom types 1–31 have been previously defined by Laskowski et al. (1996) with minor modifications. The atom types 32–40 were defined in this work for FMN molecule.

2.3. PDM-based attributes as inputs for machine learning algorithms

Protein surface atoms were categorized into 30 protein atom types (Table 1, 1–30), and one machine learning model was trained for each of the atom types. The input attributes for each of the protein atom i ($a_{i,j}$ (j=1,41): 40 attributes from the 40 interacting atom type PDMs plus one attribute from geometry) for each of the machine learning models were calculated from the PDMs on the protein surface and from the geometry of the protein surface as the following: for each atom i on the surface of the query protein (solvent accessible surface area of atom $i>0$), the PDM values associated with the grids within 5 Å radius centered at the atom are summed in the following equation:

$$s_{i,j} = \sum_k^{r_{i,k} \le 5\ \text{Å}} g_{k,j} \quad (1)$$

where $S_{i,j}$ is the PDM sum for interacting atom type j at atom i; $r_{i,k}$ is the distance between atom i to a grid point k; and $g_{k,j}$ is the PDM value of interacting atom type j at grid point k. $A_{i,j}$ (j=1,40) associated with each atom i was calculated with the following equation:

$$A_{i,j} = S_{i,j} + \sum_k^{d_{i,k} \le 10\ \text{Å}} S_{k,j} \times d_{i,k}^{-2} / \sum_n^{d_{i,n} \le 10\ \text{Å}} d_{i,n}^{-2} \quad (2)$$

where $S_{i,j}$ is defined in Eq. (1); $d_{i,k}$ is the distance between atom i and atom k. The attribute set ($a_{i,j}$ (j=1,40)) for the machine learning

Download English Version:

https://daneshyari.com/en/article/4496281

Download Persian Version:

https://daneshyari.com/article/4496281

Daneshyari.com