Biophysical Chemistry 192 (2014) 10–19

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

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

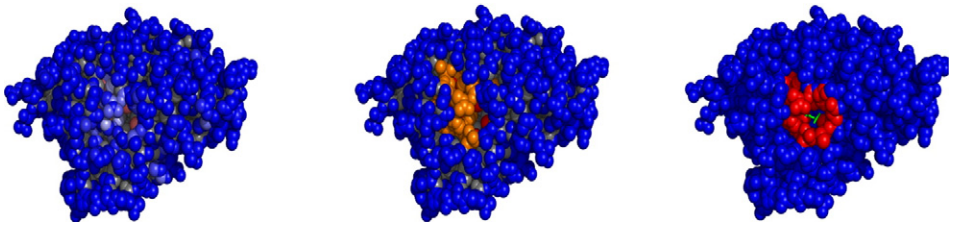
Rajasekaran Mahalingam [a,*,1], Hung-Pin Peng [a,b,c], An-Suei Yang [a,**]

[a] Genomics Research Center, Academia Sinica, 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-Fatty acid interaction
- Does not require evolutionary information for the prediction
- Useful in annotating protein structures of unknown function and computational protein models

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 April 2014
Received in revised form 22 May 2014
Accepted 22 May 2014
Available online 29 May 2014

Keywords:
Protein-fatty acid interaction
Structure-based prediction
Probability density map
Machine learning
Functional annotation

ABSTRACT

Protein-fatty acid interaction is vital for many cellular processes and understanding this interaction is important for functional annotation as well as drug discovery. In this work, we present a method for predicting the fatty acid (FA)-binding residues by using three-dimensional probability density distributions of interacting atoms of FAs on protein surfaces which are derived from the known protein-FA complex structures. A machine learning algorithm was established to learn the characteristic patterns of the probability density maps specific to the FA-binding sites. The predictor was trained with five-fold cross validation on a non-redundant training set and then evaluated with an independent test set as well as on holo–apo pair's dataset. The results showed good accuracy in predicting the FA-binding residues. Further, the predictor developed in this study is implemented as an online server which is freely accessible at the following website, http://ismblab.genomics.sinica.edu.tw/.

1. Introduction

Fatty acids (FAs) play an important role in metabolic regulation, modulation of gene expression, cell signaling, maintaining cell structure and also acting as an energy source [1–4]. Further, hundreds of bioactive lipid mediators called eicosanoids are derived from the FAs and they all are involved in pro and anti-inflammatory responses [3,5]. Essentially the FAs interact with proteins called as FA-binding proteins (FABP) to perform all these functions. These proteins are members of a superfamily of lipid-binding proteins. Some non-lipid-binding family proteins such as heat shock protein, feutin, caveolin 1, glutathione S-transferase, sterol-carrier protein-2 and fatty acid transporter also show affinity for the FAs [6–8]. Given its importance in lipid-mediated and inflammatory pathways, defects in either FAs and/or FABP protein functions lead to many metabolic diseases including obesity, diabetes and atherosclerosis [9–11]. Few therapeutic inhibitors which could be a potential therapeutic strategy to treat diabetes, insulin resistance, atherosclerosis and other fatty liver diseases have been reported [12–14]. Therefore understanding the FA-protein interaction and identifying of FA-binding sites

* Correspondence to: R. Mahalingam, Department of Physiology and Biophysics, School of Medicine, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, United States. Tel.: +1 216 368 8654.
** Correspondence to: A.-S. Yang, Genomics Research Center, Academia Sinica, 128 Academia Rd., Sec. 2, Nankang Dist., Taipei 115, Taiwan. Tel.: +8862 2 2787 1232.
E-mail addresses: rajasekaran.mahalingam@case.edu (R. Mahalingam), yangas@gate.sinica.edu.tw (A.-S. Yang).
[1] Current address: Case Western Reserve University, School of Medicine, 10900 Euclid Ave., Cleveland, OH 44106, United States.

http://dx.doi.org/10.1016/j.bpc.2014.05.002
0301-4622/

are important as it can aid in drug discovery process for developing therapeutic molecules against the metabolic diseases.

A computational method for predicting the FA-binding site on the proteins would greatly facilitate the identification of FA-binding sites on the protein structure. Few computational methods have been developed to predict the lipid-binding residues from the protein sequences. Tempel et al. [15] and Scott et al. [16] developed methods to predict lipid-binding residues for cytoskeleton and cytoskeleton-associated proteins respectively. Wang et al. [17] and Xiong et al. [18] used support vector machine approach to predict the lipid-binding residues from the protein sequences. Lin et al. [19] developed a method to identify the functional class of lipid-binding proteins from protein sequences. Although these methods are reasonably successful in their respective prediction, they all are not specific for protein-FA interaction prediction and moreover most of them are sequence based as well as use evolutionary information for their prediction. These methods may have difficulty in predicting the binding sites from the orphan proteins. Therefore a reliable structure-based method for predicting FA-binding residues without using the evolutionary information is necessary.

In this study, we have developed a structure-based method which uses machine learning approach to predict the FA-binding sites on the protein surfaces. This method mainly recognizes characteristics interacting atom distribution patterns associated with the FA-binding. The basic principle has been already applied successfully to predict protein–protein [20], protein–carbohydrate [21] and protein–FMN interactions [22]. Here we have extended this method to predict the FA-binding residues. In the prediction, protein surface atoms (it refers to all the protein atoms including interior 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 35 interacting atom types (30 atom types from protein, 1 from water and 4 from FA) 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 [20,21]; the PDMs for the 4 FA interacting atom types were constructed with the protein–FA interacting atom pairs from the dataset of 440 protein–FA 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 by five-fold cross validation on the training dataset P75 and then the trained model used to predict the independent test set P25 and holo–apo pairs. The results indicate that our approach can predict the FA-binding sites with very good accuracy.

2. Materials and methods

2.1. Datasets

All the structures were extracted from PDB [23]. The training set P75 contains 75 chains that released before the 31st of December 2010 and that binds different FAs. The test set P25 contains structures which released after the 31st of December 2010 and retained 25 structures which shares less than 5% sequence similarity with training set [24]. Holo and apo datasets consist of 10 proteins in each set. The given residue is annotated as a FA-binding, if any of the FA atoms within 5 Å distance with any protein atoms. The negative dataset of S108 and S142 were collected from protein–carbohydrate [21] and protein–protein interaction [20] predictions respectively.

2.2. Construction of three-dimensional probability density maps of non-covalent interacting atoms on protein surfaces

The methodology for the PDM construction for protein-non covalent interacting atom pair (Table 1, atom types 1–31) has been described previously [20,21]. The PDMs for FA atoms (Table 1, atom types 32–35) were constructed with protein–FA interacting atom pair database derived from 440 protein–FA complexes. In order to keep the PDMs high in information content and low in noise from irrelevant interactions, non-interacting pairs were eliminated with the filter system based on the work by McConkey et al. [25].

Table 1
Protein and fatty acid 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	ZC3	1.90	Sp3 carbon
33	ZO3	1.68	Sp3 oxygen
34	ZO2	1.66	Sp2 oxygen
35	ZC3	1.90	Sp2 carbon

The protein atom types 1–31 have been previously defined by Laskowski et al. [39] with minor modifications. The atom types 32–35 were defined in this work for fatty acid molecule.

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

The input attributes were derived from PDMs on the protein surfaces. Atoms from the protein surface and interior were categorized into 30 protein atom types and for each atom type one machine learning model was trained. 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 were summed in Eq. (1).

$$S_{i,j} = \sum_{k}^{r_{i,k} \le 5A} g_{k,j} \tag{1}$$

where S_{ij} 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; $g_{k,j}$ 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 Eq. (2).

$$A_{i,j} = S_{i,j} + \frac{\sum_{k}^{d_{i,k} \le 10A} S_{k,j} \times d_{i,k}^{-2}}{\sum_{n}^{d_{i,n} \le 10A} d_{i,n}^{-2}} \tag{2}$$