Biochimica et Biophysica Acta 1844 (2014) 171–180

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins ☆

Zexian Liu [a], Yongbo Wang [b], Changhai Zhou [a], Yu Xue [b], Wei Zhao [a,*], Haiyan Liu [a,*]

[a] Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei, Anhui 230027, China
[b] Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

ARTICLE INFO

Article history:
Received 3 December 2012
Received in revised form 2 March 2013
Accepted 4 March 2013
Available online 14 March 2013

Keywords:
Zinc-binding
Geometric restriction
Training-independent
Cancer gene
Drug target

ABSTRACT

Zinc is one of the most essential metals utilized by organisms, and zinc-binding proteins play an important role in a variety of biological processes such as transcription regulation, cell metabolism and apoptosis. Thus, characterizing the precise zinc-binding sites is fundamental to an elucidation of the biological functions and molecular mechanisms of zinc-binding proteins. Using systematic analyses of structural characteristics, we observed that 4-residue and 3-residue zinc-binding sites have distinctly specific geometric features. Based on the results, we developed the novel computational program *G*eometric *RE*striction for *Zn*c-binding (GRE4Zn) to characterize the zinc-binding sites in protein structures, by restricting the distances between zinc and its coordinating atoms. The comparison between GRE4Zn and analogous tools revealed that it achieved a superior performance. A large-scale prediction for structurally characterized proteins was performed with this powerful predictor, and statistical analyses for the results indicated zinc-binding proteins have come to be significantly involved in more complicated biological processes in higher species than simpler species during the course of evolution. Further analyses suggested that zinc-binding proteins are preferentially implicated in a variety of diseases and highly enriched in known drug targets, and the prediction of zinc-binding sites can be helpful for the investigation of molecular mechanisms. In this regard, these prediction and analysis results should prove to be highly useful be helpful for further biomedical study and drug design. The online service of GRE4Zn is freely available at: http://biocomp.ustc.edu.cn/gre4zn/. This article is part of a Special Issue entitled: Computational Proteomics, Systems Biology & Clinical Implications. Guest Editor: Yudong Cai.

1. Introduction

Although the total net content of zinc in organisms is very low, it is still essential for survival. Free or loosely bound zinc ions function as an intracellular signal [1], and were shown to act as a second messenger recently [2]. However, the major role of the zinc ion is tightly coordinated with protein residues [3], and it is estimated that as mach as 10% of the human proteome is made up of potential zinc-binding proteins [4]. A large number of studies have been conducted to describe the molecular mechanisms of zinc-binding [5,6]. Although zinc ions can be penta- or even hexa- coordinated, tetrahedral coordination is the predominant form for most of the zinc-binding sites [6]. The majority of the identified zinc-binding residues are made up of cysteine (C) and histidine (H) [3], while various other residues, such as glutamic acid (E), aspartic acid (D), serine, threonine, lysine and methionine can also coordinate with zinc ions [5]. In fact, cysteine, histidine, aspartic acid and glutamic acid constitute almost all of the zinc-coordinating protein residues, while oxygen, nitrogen and sulfur donors in water molecules or other free ligands can also serve as coordinating moieties for zinc ions [5,6]. As a structural component that binds with amino acid (AA) residues, the zinc ion is critical for the functions of proteins, such as helping to stabilize the structure of "zinc-finger" transcription factors [3,7] and acting as the catalytic site in enzymes [3]. Thus, it is critical to identify zinc-binding sites in order to dissect the molecular functions and mechanisms in the proteins that contain them.

To date, a variety of experimental approaches, including X-ray diffraction, Nuclear Magnetic Resonance (NMR) and X-ray absorption fine structure (XAFS) techniques have been employed to identify zinc-binding residues in proteins [5]. However, since these experimental studies are both time- and labor-intensive, only a small proportion of potential zinc-binding proteins have characterized, even though genome-scale analysis have suggested that there are thousands of zinc-binding proteins [4]. Recently, a number of computational approaches have made contributions to this area to promote the discovery of zinc-binding proteins together with their binding sites. Since zinc ions only coordinate with restricted types of residues

☆ This article is part of a Special Issue entitled: Computational Proteomics, Systems Biology & Clinical Implications. Guest Editor: Yudong Cai.

* Corresponding authors. Tel./fax: +86 551 63607450.
E-mail addresses: zhaowei@ustc.edu.cn (W. Zhao), hyliu@ustc.edu.cn (H. Liu).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.03.001

and the binding seems to follow certain specific patterns, a series of prediction studies were carried out based on sequence analysis [8–17]. A variety of algorithms were employed alone or in combination in these studies, including support vector machines (SVMs) [10,11,13,14,17], neural networks (NNs) [9,11,13,17], machine learning (ML) [12] and the homology-based method of PHI-Blast [8,15]. Since zinc-binding is heavily dependent on the three-dimensional conformations of protein residues, the structure-based predictions might be expected to achieve a better performance [5,6].

Previously, a handful of computational studies have contributed to the effort to predict zinc-binding sites based on protein structures [18–27]. Structural features such as secondary structure states (SS), solvent-accessible surface areas (SASAs), inter-residue distance matrices, geometrical features and residue properties were combined with various algorithms including NNs [18], SVMs [20,24], machine learning [24], random forest algorithm [21], and Bayesian classifier [22] in an effort to provide accurate predictions. Furthermore, the empirical Fold-X force field, Rosetta software and the fragment transformation method (FTM) were also employed to characterize zinc-binding structures [19,25,27] (Table 1). Recently, Zheng et al. presented a powerful computational framework which integrates various features including sequence, structure and network properties with the random forest algorithm to predict zinc-binding sites [28]. In addition, we previously developed a structure-based method (TEMSP or 3D TEmplate-based Metal Site Prediction) for predicting zinc-binding sites [26]. In these studies, complex classifiers, force field-based modeling and template-based calculation among these approaches afforded excellent results.

In this work, we systematically analyzed the structural features of zinc-binding sites from a well characterized and non-redundant dataset of 601 zinc-binding sites in 431 proteins. We observed that 4-residue (4-res) and 3-residue sites (3-res) have different sequential and structural features, such as sequence length distribution, AA preferences, SSs and geometrical distance. In particular, we found that the geometrical distance between zinc and the binding residue was specifically restricted in the 4-res and 3-res binding sites, respectively. Based on these observations, we developed a geometric restriction approach to characterize zinc-binding sites from protein structures. The geometrical distance ranges for the 4-res and 3-res binding were respectively calculated from the zinc-binding data and then were employed to predict potential zinc-binding residues. The software program *G*eometric *RE*striction for *Zin*c-binding, http://biocomp.ustc.edu.cn/gre4zn/ (GRE4Zn) was implemented, with a sensitivity of 97.68% and a precision of 98.93% under an Intersection over Union Ratio (IoUR) ≧0.5, respectively, while the values were 95.56% and 96.58% under IoUR = 1.0. With this superior predictor, we performed a large-scale study on the structurally characterized proteins in PDB database [29]. Furthermore, we systematically analyzed the potential zinc-binding sites for *Escherichia coli*, *Saccharomyces cerevisiae* and *Homo sapiens*. The statistical analyses of gene ontology (GO) annotations for the results suggested that the enriched biological processes were different for different species. It was observed that zinc-binding proteins were involved in more complicated biological processes in higher organisms than lower organisms, which implies an ongoing evolution of the functions of zinc-binding proteins. Further analyses revealed that zinc-binding proteins are also significantly enriched in cancer genes and drug targets, and could serve as a useful resource for further biomedical investigation and drug design.

Table 1
Summary of a number of previous studies on the prediction of zinc-binding.

	Software	PMID	Algorithm
Sequence-based predictions			
Andreini et al. [8]		14962940	PHI-BLAST
Lin et al. [9]		15912584	NN
Passerini et al. [11]	MLP	16927295	SVM, NN
Lin et al. [10]	SVMProt	17254297	SVM
Passerini et al. [12]	Zinc Finder	17280606	ML
Shu et al. [14]	PredZinc	18245129	SVM
Lippi et al. [13]	MetalDetector	18635571	SVM, NN
Andreini et al. [15]		19697929	PHI-BLAST
Bertini et al. [16]		20443034	HMM
Passerini et al. [17]	MetalDetector v2.0	21576237	SVM, NN
Structure-based predictions			
Sodhi et al. [18]	MetSite	15313626	NN
Schymkowitz et al. [19]		16006526	Fold-X
Babor et al. [20]	CHED	17657805	SVM
Goyal and Mande [23]		17847089	Template-based
Ebert and Altman [22]	FEATURE	18042678	Bayesian
Bordner [21]		18940825	Random forest
Levy et al. [24]	SeqCHED	19173310	SVM, ML
Wang et al. [25]	Rosetta	20054832	Rosetta
Zhao et al. [26]	TEMSP	21414989	Template-based
Lu et al. [27]		22723976	FTM
Zheng et al. [28]	Zincidentifier	23166753	Random forest

2. Materials and methods

2.1. Data preparation and analysis

The dataset of experimentally identified zinc-binding sites have been retrieved according to the methods described in our recent study [26]. The zinc-binding protein structures of less than 95% sequence identity were retrieved while redundant chains were removed. The "abnormal" zinc sites with multiple conformations, less than three coordinating atoms, metal atom occupancy lower than 0.5, or a B factor higher than 90, were discarded. Since the excess redundancy of a large number of homologous sites could lead to the overestimation of prediction accuracy, the zinc-binding proteins were clustered at a sequence identity cutoff of 30% with the CD-HIT program [30], and only one representative protein chain was retained from each cluster. Furthermore, 99 protein structures from the resulting 431 proteins were randomly selected as the testing dataset, while the others were considered as the training dataset.

Since almost all zinc-binding residues are cysteines, histidines, glutamic acids and aspartic acids (CHEDs), in this study, only the sites with CHED residues were reserved for consideration. The dataset of 601 zinc-binding sites in 431 proteins was divided into two subsets according to the size of their ligands. The larger subset of "4 residues (4-res)" contains 473 sites in 317 proteins, in which four of the coordinating ligands are CHED residues, while the smaller subset of "3 residues (3-res)" has 128 sites in 123 proteins, in which the tetrahedral ligands were constituted with three CHED residues and another donor of oxygen, nitrogen or sulfur from either a water molecule or other free ligands. The distances between atoms were computed for the 4-res and 3-res datasets, respectively. For large-scale prediction, the protein structures determined by X-ray with a resolution of less than 3.0 Å were retrieved from the PDB database [29]. The protein structures were integrated with a sequence similarity threshold of 30% in organism-specific statistical analyses of *E. coli*, *S. cerevisiae* and *H. sapiens*. From these large-scale analyses, 377, 411 and 1027 structure chains were predicted to contain zinc-binding sites among 1263, 726 and 3087 protein structures, respectively, from the PDB database.

The SS and SASA analyses were carried out by STRIDE [31]. The enrichment analyses of the annotations of SCOP structural classifications, GO, cancer genes and drug targets were performed with a hypergeometric distribution [31]. The comparison of the SCOP structural classifications for 4-res and 3-res zinc-binding was performed with Yates' Chi-square (χ^2) test with the 2×2 contingency table [31]. The GO annotation file for PDB was downloaded from the GOA database at the EBI (http://www.ebi.ac.uk/goa) [32], while the SCOP structural classification of proteins data was downloaded from the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop/) [33]. In