STATISTICAL CONSERVATION ANALYSIS OF ZINC-INTERACTING RESIDUES

Ioannis N. Kasampalidis[1], *Ioannis Pitas*[1] *and Kleoniki Lyroudia*[2]

[1]Department of Informatics, Aristotle University of Thessaloniki,
[2]2Department of Endodontology, School of Dentistry,
Aristotle University of Thessaloniki,
Thessaloniki, 54124, Greece
pitas@aiia.csd.auth.gr

ABSTRACT

As a result of rapid advances in genome sequencing, the pace of discovery of new protein sequences has surpassed that of structure/function determination by orders of magnitude. This is also true for metal-binding proteins, i.e. proteins that bind one or more metal atoms necessary for their biological function. With regard to metal-interacting residues, the question arising is whether these interactions apply additional evolutionary constraints and to what extent. We try to answer this question for a subset of metal-binding-proteins, namely zinc-binding proteins, which play an important role in a number of biological processes, as exemplified by the tumor-suppressor protein p53. Our results shows significantly higher evolutionary pressure on zinc-interacting residues, a result which can be used in a number of other studies, including zinc-binding site prediction.

1. INTRODUCTION

Bioinorganic chemistry is the field dealing with the crucial interactions between inorganic metals and biological molecules of interest [1]. An important subset of biological molecules, metallo-proteins, plays a fundamental role in numerous biological processes, as evidenced by the fact that about one third of determined protein structures contain metal-binding sites, as shown by a simple Protein Data Bank (PDB) search [2].

A great deal of effort has recently been devoted to the analysis and prediction of metal-binding-sites. Many researchers have focused on the analysis of their structure, either with respect to its geometry as in Harding et al. [3] and Tainer et al. [4] or its chemical properties as in Karlin et al. [5]. Other studies perform analysis based on density functional theory/continuum dielectric methods as in Dudev et al. [6]. Such analysis can have significant impact on better functional characterization of metal-binding proteins, drug design, database searching for metal-binding-proteins, as in Andreini et al. [7], or metal-binding-site (MBS) prediction.

Regarding MBS prediction, a number of recent approaches have been proposed, including an energy-based method by Laurie et al. [8], a support-vector machine predictor of cysteine binding state by Passerini and Frasconi [9], and a recursive neural network predictor of disulfide bridge connectivity by Vullo and Frasconi [10].

A key piece of information in some of these methods [9, 10] is conservation. Their results are encouraging; however, no large-scale analysis has been performed on conservation of metal-interacting residues. Such a study could provide justification for using conservation information as a feature for MBS prediction. Additionally, it could provide further insight to the functional importance of certain metal/residue combinations by comparison of the extent of residue conservation among different metal ions and protein families.

In this study, we focus on conservation analysis of zinc-interacting residues. Zinc is one of the metals playing a crucial role in a number of biologically significant proteins, including p53 [11], a tumor-suppressor protein. This work investigates whether there is higher selective pressure on zinc-interacting residues vs. non-zinc-interacting residues.

For this purpose, we derived a non-redundant set of zinc-interacting proteins from PDB. Because this set did not contain enough members for a large-scale analysis, we used homology search via PSI-BLAST [12] to obtain additional putative zinc-interacting proteins, limiting our selection to one ortholog protein from each species. The known-structure proteins were grouped according to families as defined in Structural Classification of Proteins (SCOP) [13], while their orthologs were also included in the same family. Protein grouping by family was preferred since same family membership in SCOP indicates clear evolutionary relationship. A multiple sequence alignment (MSA) was performed on all members of each family.

We applied two approaches in our analysis. In the first approach, we measured the identity ratio, in the MSA, for all zinc-interacting residues and compared it to the identity ratio for non-interacting residues. This approach was limited only to known-structure sequences, since for unknown-structure sequences the metal-interacting residues cannot be guaranteed. The mean identity ratio of all residues within a family was calculated and the means of all families were compared.

We also pursued conservation analysis based on an information theoretic approach, where we calculate separate substitution matrices for zinc-interacting and non-zinc-interacting MSA columns for each family. These matrices were compared using the relative entropy metric, as described in Altschul [14]. This metric serves as a distance measure between the actual and the theoretically expected probability distribution of residue substitutions, where higher relative entropy indicates higher selective pressure.

2. METHODS

2.1. Dataset

We created a dataset of zinc-interacting structures with the help of PDB [2] using an appropriate query, where we required structures to have resolution better than 2.5Å and no mutant residues. From the structures meeting these criteria, we chose only the ones classified in the following SCOP classes: 1) all alpha proteins, 2) all beta proteins, 3) alpha and beta proteins (a+b), 4) alpha and beta proteins (a/b) and 5) membrane and cell surface proteins and peptides. We also required sequences to have a length greater than 40 residues. A non-redundant set was derived from the proteins meeting these criteria with the help of the algorithm by Li et al. [15], at the 90% identity level, as implemented in PDB. In total, 481 PDB files containing zinc were identified, where some of these files may belong to more than one class. The alpha class contained 105 files, the beta class 165, the a+b class 177, the a/b class 174 and the class of membrane and cell surface proteins 1 PDB file.

2.2. Zinc-interacting residues

For each structure, the metal-interacting residues were identified using a distance cut-off of 4Å from the metal atom. Although this criterion does not take into account the biological significance of this interaction, it is probably the best criterion currently available for automated metal-interacting residue identification. The distance cut-off of 4Å was chosen as an upper empirical bound, as described in Harding et al. [3]. Only the domains, as defined in SCOP, containing zinc-interacting residues were selected for multiple sequence alignment and these domains were afterwards grouped by SCOP family. A small number of domains from the original set were discarded because their species could not be identified based on the NCBI taxonomy database [16, 17].

2.3. Multiple sequence alignments(MSA)

Initial multiple sequence alignments for the domains containing metal-interacting residues were performed using PSI-BLAST (2) against the NCBI NR database [16, 17], with an e-value cutoff of 10^{-5}. In order to identify orthologs, for each protein, we selected only the reciprocal best hit from each species in the PSI-BLAST reports. In the reports, sequences corresponding to the same species as the query sequence were also discarded. Sequences were further filtered by discarding entries with the following keywords: synthetic, putative, probable, predicted, hypothetical, unnamed, unknown, unidentified, designed, vector. The resulting sequences were grouped with known-structure sequences into SCOP families and the multiple sequence alignments were further refined using MUSCLE (multiple sequence comparison by log-expectation) [18].

2.4. Identity ratio

The identity ratio for a single residue was calculated as the ratio of identical residues and the length of the MSA column. This ratio was calculated only for residues of known-structure sequences, while sequence gaps in the MSA columns were not included in the calculation. The mean zinc-interacting and non-interacting identity ratios of each family were calculated by simple averaging over all zinc-interacting and non-interacting residues of each family respectively.

2.5. Substitution matrices

Within each family MSA, a MSA column was defined as zinc-interacting if it contained at least one zinc interacting residue. Substitution matrices were created separately for non-zinc-interacting and zinc-interacting MSA columns, using all the sequences in each MSA, as described in Henikoff & Henikoff [19]. More specifically, each element $s_{i,j}$ of the substitution matrix is calculated as in (1)

$$s_{i,j} = log_2 \left(\frac{c_{i,j}}{e_{i,j}} \right), \tag{1}$$

where $c_{i,j}$ and $e_{i,j}$ are the observed and expected frequencies respectively. The observed frequencies are calculated separately for the zinc-interacting and non-interacting MSA columns. The expected frequencies are calculated from all MSA columns using the formula described in Henikoff and Henikoff [19]. The matrices were then compared based on the relative entropy metric, shown in (2), as described in Altschul [14].

$$H = \sum_{i,j} c_{i,j} \times s_{i,j}, \tag{2}$$

where $c_{i,j}$ and $s_{i,j}$ are the observed frequency and the elements of the substitution matrix respectively.

3. RESULTS

3.1. Identity ratio

The mean identity ratio for zinc interacting residues is 0.7, while for non-interacting residues, it is 0.51. The t-test on the two means resulted in a p-value of 7.25×10^{-23}, which is highly significant. The histogram for the mean identity ratio per family is shown in Figure 1. This result clearly shows the higher evolutionary constraints for zinc-interacting residues of known-structure sequences.

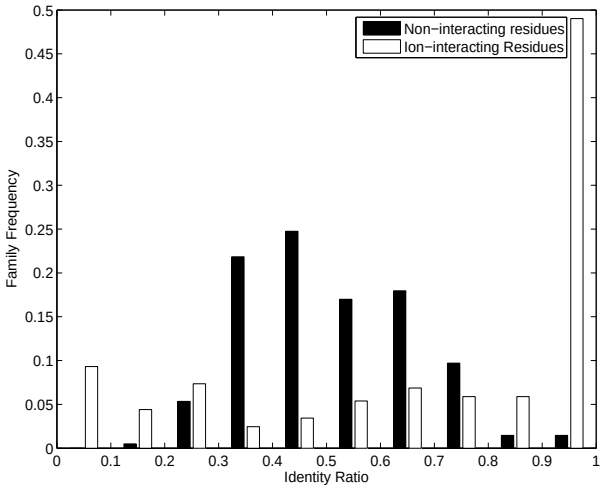

Figure 1. Histogram of the identity ratio of non-zinc-interacting vs. zinc-interacting residues. Zinc-interacting residues exhibit higher selection constraints.

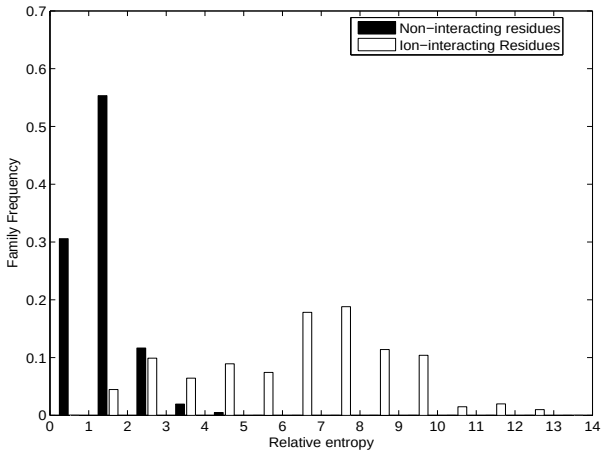

Figure 2. Histogram of the relative entropy of substitution matrices of zinc-interacting vs. non-zinc-interacting MSA columns. Relative entropy is higher for zinc-interacting MSA columns.

3.2. Relative Entropy

Identity ratio analysis on known-structure sequences does not provide a full picture of each family MSA, since it focuses only on sequences of known-structure. For this reason, substitution matrices were created for each family, for zinc-interacting and non-zinc interacting columns. Relative entropy was calculated for each of these matrices and the histogram for all 212 families is shown in Figure 2. The mean relative entropy for zinc interacting MSAs is 6.35, while for non-interacting residues it is 1.42. The difference of the two means is highly significant, as indicated by the t-test p-value of 1.84×10^{-93}.

4. CONCLUSION

In this study, we pursued analysis of zinc-interacting proteins' conservation. Zinc-interacting proteins take part in a number of important biological process as exemplified by the tumor-suppressor protein p53. Our statistical methodology showed significantly higher conservation of zinc-interacting residues compared to non-zinc-interacting residues. This conclusion is drawn from two types of metrics, identity ratio, which is based only on known-structure sequences, and relative entropy, which is based on all orthologous sequences.

However, a great deal of work remains to be done. More specifically, analysis needs to be extended to other biologically significant metal ions. Moreover, the conservation levels between different families need to be compared, in order to extract useful biological hindsight into metal-binding site structure and function. The completion of these studies can have significant implications for metal-binding site prediction, protein functional characterization and drug design.

5. ACKNOWLEDGMENTS

This work was supported by the EU project Biopattern: Computational Intelligence for biopattern analysis in Support of eHealthcare, Network of Excellence Project No. 508803.

6. REFERENCES

[1] I. Bertini and A. Rosato, "Bioinorganic chemistry in the postgenomic era," *Nucleic Acids Research*, vol. 100, pp. 3601–3604, 2003.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[3] M. Harding, "The architecture of metal coordination groups in proteins," *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, pp. 849–859, 2004.

[4] J. A. Tainer, R. V. A., and E. D. Getzoff, "Protein metal-binding sites," *Current Opinions in Biotechnology*, vol. 3, pp. 378–387, 1992.

[5] S. Karlin, Z. Zhu, and K. D. Karlin, "The extended environment of mononuclear metal centers in protein structures," *Proceedings of the National Academy of Sciences, USA*, vol. 94, pp. 14225–14230, 1997.

[6] T. Dudev, Y. Lin, M. Dudev, and C. Lim, "First-second shell interactions in metal binding sites in proteins: a pdb survey and dft/cdm calculations," *Journal of The American Chemical Society*, vol. 125, pp. 3168–3180, 2003.

[7] C. Andreini, B. I., and R. A., "A hint to search for metalloproteins in gene banks," *Bioinformatics*, vol. 20, pp. 1373–1380, 2004.

[8] A. T. Laurie and R. M. Jackson, "Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites," *Bioinformatics*, vol. 21, pp. 1908–1916, 2005.

[9] A. Passerini and P. Frasconi, "Learning to discriminate between ligand bound and disulfide bound cysteines," *Protein Engineering Design and Selection*, vol. 7, pp. 367–373, 1995.

[10] A. Vullo and P. Frasconi, "Disulfide connectivity prediction using recursive neural networks and multiple alignments," *Bioinformatics*, vol. 20, pp. 653–659, 2004.

[11] Y. Cho, S. Gorina, P. D. Jeffrey, and N. P. Pavletich, "Crystal structure of a p53 tumor suppressor-dna complex: understanding tumorigenic mutations," *Science*, vol. 265, pp. 346–355, 1994.

[12] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 1997.

[13] A. G. Murzin, B. S. E., H. T., and C. C., "Scop: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, pp. 536–540, 1995.

[14] S. F. Altschul, "Amino acid substitution matrices from an information theoretic perspective," *Journal of Molecular Biology*, vol. 219, pp. 555 – 565, 1991.

[15] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics*, vol. 17, pp. 282–283, 2001.

[16] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, R. B. A., and W. D.L., "Genbank," *Nucleic Acids Research*, vol. 28, pp. 15–18, 2000.

[17] D. L. Wheeler, A. E. Chappey, C. Lash, D. D. Leipe, M. T. L., G. D. Schuler, T. A. Tatusova, and B. A. Rapp, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 28, pp. 10–14, 2000.

[18] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, pp. 1792–1797, 2004.

[19] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences, USA*, vol. 89, pp. 10915–10919, 1992.