Systematic classification of metalloproteins based on three-dimensional structural similarity of their metal sites

Yana Valasatava (✉ valasatava@cerm.unifi.it)

Magnetic Resonance Center (CERM) – University of Florence

Antonio Rosato (✉ rosato@cerm.unifi.it)

Magnetic Resonance Center (CERM) – University of Florence, Department of Chemistry – University of Florence

Claudia Andreini

Magnetic Resonance Center (CERM) – University of Florence, Department of Chemistry – University of Florence

Method Article

Keywords: Functional clustering, Iron, Zinc

Posted Date: April 18th, 2015

DOI: <https://doi.org/10.1038/protex.2015.036>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

In this work, we developed a methodology to perform a systematic classification based on three-dimensional structural similarity of the metal sites contained in metalloproteins. Our definition of metal site extended beyond the metal ion and its aminoacidic ligands by including all the chemical species \ (aminoacids, nucleotides, exogenous ligands) providing at least one donor atom as well as all any other chemical species within a radius of 5.0 Å. We previously defined this as the **Minimal Functional Site** of a metalloprotein (**MFS**), and showed that its characteristics are related to the metalloprotein function. The methodology described here leverages the MetalS² algorithm, whose total score provides a quantitative measure of structural similarity between pairs of MFS. We used this measure to build clusters of structurally similar MFSs using a two-stage hierarchical clustering algorithm. At the first stage we cluster MFSs identified in corresponding position within proteins with the same fold. For each of these clusters we then identify a representative MFS. At the second stage, all representative MFSs are clustered. The resulting groups are thus independent of the overall protein fold.

Introduction

Metal ions are bound to biological macromolecules via coordination bonds. The bonds are formed by the so-called donor atoms. Such atoms can belong to either the backbone or side chains/bases of the macromolecule \ (protein or nucleic acid) as well as to non-macromolecular ligands, such as oligopeptides, small organic molecules, anions, water molecules. A metal ion together with its donor atoms and ligands constitute the metal-binding site. However, the biochemical properties of such a site depend also on the surrounding macromolecular environment \ (5-9). Consequently, we defined the “minimal functional site” \ (MFS) in a metal-macromolecule adduct as the ensemble of atoms containing the metal ion or cofactor, all its ligands and any other atom belonging to a chemical species within 5 Å from a ligand \ (1,10). The MFS describes the local 3D environment around the cofactor, independently of the larger context of the protein fold in which it is embedded. The MetalPDB database is an updated collection of all structurally characterized MFSs \ (11). Recently, we have developed a computational approach, implemented in the MetalS² program, to quantify the structural similarity of MFSs in metalloproteins \ (1). In this work we exploited MetalS² to perform systematic, quantitative comparisons of MFS structures with the final aim of producing a classification of metal sites. This classification does not depend on the overall metalloprotein fold and describes structural variability of MFSs within a metalloprotein family. Furthermore, it indicates possible relationships between different metalloprotein families binding the same metal cofactors. The present computational protocol organizes MFSs into clusters in such a way that each cluster contains sites that are structurally similar to each other and differ from sites of the other clusters. The procedure uses a hierarchical agglomerative clustering algorithm to obtain a structure-based classification. In agglomerative clustering every individual object is initially considered as a singleton \ (i.e., a cluster containing only one member). Then the clusters are iteratively grouped by merging the two clusters at the shortest “distance”, i.e. the most similar pair. For the present work, the distance measure adopted was the global MetalS² score, which increases with increasing

structural diversity. Two merged clusters become one cluster, so after each iteration there is one less cluster. The iterations are repeated until all objects are collected into a single cluster. The result of hierarchical clustering is a nested sequence of partitions, with a single, all-inclusive cluster at the top and all singleton clusters at the bottom. Each intermediate cluster can be viewed as a combination of two clusters from the lower level or as a part of a split cluster from the higher level. Hierarchical clustering methods differ in the way they merge clusters (linkage methods). Although all methods merge the two “closest” clusters at each step, they determine differently the distance between clusters, i.e., have different metrics to compare one cluster to another. Here we used both the complete and average linkage methods (12). For complete linkage the distance between a pair of clusters corresponds to greatest distance from any member of one cluster to any member of the other cluster. In the average linkage method the distance between two clusters is the average of the distances between all the members in one cluster and all the members in the other. The final clusters are defined by cutting the nested sequence of partitions at a certain threshold. The clusters are considered to be separate if the distance between them (a value of the global MetalS² score) is bigger than this threshold. The value of the threshold also determines the extent to which the objects are similar within each cluster (the diameter of the clusters, which however is affected by the linkage type used).

Equipment

****Data**** The structures of all MFSs can be downloaded from MetalPDB database web-interface: <http://metalweb.cerm.unifi.it/download/sites/>. ****Software**** The source code exploited by this protocol implements the pipeline as a library of Python scripts. The code requires the following freely-available software/libraries: • Python 2.6.6 (<https://www.python.org/download/releases/2.6.6/>). • MetalS² (http://metalweb.cerm.unifi.it/tools/metals2_download_package/) (1). MetalS² is a software tool that allows pairwise superimposition of MFSs’ structures and assessment of their structural similarity. The measure of similarity is given by a score calculated over the residues put in correspondence in structural alignment of MFSs (sequence alignment based on structural alignment). In the context of the present work the global MetalS² similarity score is used as the distance measurement in clustering procedure (the lower the score the more similar the two MFSs are). • p3d (<http://p3d.fufezan.net/>) (2), a bioinformatic analysis tools to process and analyze three dimensional protein structure files (PDB files). With root privileges, copy the p3d directory in the Python packages directory /usr/lib/python2.6/site-packages. • SciPy 0.7.2 (<http://www.scipy.org/>) (3), a Python-based open-source software for mathematics, science, and engineering. In particular, we used the hierarchy module (scipy.cluster.hierarchy) of the hierarchical clustering package. • NumPy 1.4.1 (<http://www.numpy.org/>) (4), the package for scientific computing with Python.

Procedure

The protocol leverages the organization of MFSs in equistructural groups (EGs hereafter) that is already provided by the MetalPDB database (<http://metalweb.cerm.unifi.it/>). Such groups contain the MFSs that

are found in proteins with the same fold and occur at the same position within that fold. EG are computed in MetalPDB by superimposing the entire domain containing the MFS in the protein structures under consideration and then computing the distance between the metal centers. MFSs whose metal centers are within a threshold of 3.5 Å from one another are assigned to the same equistructural group. The workflow consists of the following steps: 1. Select all structures of MFSs that bind the metal cofactor of interest (e.g. heme-binding sites, zinc sites, copper sites, etc.). Each site has an EG identifier assigned (e.g. 3c25_4; 21566). 2. Divide the sites into two sets: the first set contains sites from EGs populated with more than one member, the second set contains the sites that are the only members of their corresponding EGs. 3. For each EG from the first set perform an all-vs-all superposition of its sites with Metals² (1). This provides a matrix containing the pairwise distances of a set of site. 4. Cluster sites within each EG into groups of highly similar structures (first or intra-group stage). For this, we used the hierarchical clustering algorithm with the complete linkage method (12). The cut-off value of the global Metals² similarity score used to determine the resulting clusters was set as 2.25. This constitutes a very stringent threshold (1), so that each first stage cluster contains sites with high structural similarity. 5. For each cluster obtained at the previous stage, define a representative site as the MFS by minimizing the sum of the Metals² score values to all other MFSs in the same cluster. 6. Prepare the data for the second level analysis (second or inter-group stage) by merging the list of all representative MFSs from clusters built at the first level (point 5) and the list of sites from single-membered EGs (point 2). 7. Perform the all-vs-all superposition of sites from point #6. 8. Build the clusters (as done in points 3-4) by applying complete or average linkage and specifying the threshold value of the global Metals² similarity score that defines the maximum distance allowed between two clusters. Reasonable values are in the range between 2.25 and 3.5. Figure 1 graphically recapitulates the protocol.

Timing

The running time of the algorithm presented here depends on the number and size of the structures to be processed. The running time of the Metals2 tool to compare a pair of MFS structures on an Intel(R) Core(TM) i5 CPU 650 @ 3.20GHz processor varies from seconds to a few minutes, depending on the number of atoms in the structures of MFS. As an example, for 8891 sites of heme-binding proteins the entire procedure required approximately 10000 CPU-hours (using AMD Opteron(tm) Processor 6366 HE CPU 1.800 GHz).

Troubleshooting

We suggest that the users remove from the datasets all sites with less than 10 amino acids as well as all sites where the metal ions has only one aminoacidic ligand with all other ligands being water molecules. These sites tend to produce low Metals² scores even in the absence of significant structural similarity.

Anticipated Results

Clusters of MFS structures obtained independently of the overall metalloprotein fold.

References

1. MetalS⁽²⁾: a tool for the structural alignment of minimal functional sites in metal-binding proteins and nucleic acids. Andreini C, Cavallaro G, Rosato A, Valasatava Y. 11, Nov 25, 2013, Vol. 53, pp. 3064-3075.
2. p3d–Python module for structural bioinformatics. Fufezan, C. and Specht, M. s.l. : BioMed Central, 2009, BMC Bioinformatics, Vol. 10, p. 258.
3. Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python. 2001–.
4. The NumPy Array: A Structure for Efficient Numerical Computation. Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. 2, Computing in Science & Engineering, Vol. 13, pp. 22 - 30.
5. Changes in hydrogen-bond strengths explain reduction potentials in 10 rubredoxin variants. I-Jin Lin, Erika B. Gebel, Timothy E. Machonkin, William M. Westler, and John L. Markley. 41, 2005, Proceedings of the National Academy of Sciences, Vol. 102, pp. 14581-14586.
6. Solvent Tuning of Electrochemical Potentials in the Active Sites of HiPIP Versus Ferredoxin. Abhishek Dey, Francis E. Jenney, Michael W. W. Adams, Elena Babini, Yasuhiro Takahashi, Keiichi Fukuyama, Keith O. Hodgson, Britt Hedman, Edward I. Solomon. 2007, Science, Vol. 318, pp. 1464-1468. 5855.
7. Structural analysis of metal sites in proteins: non-heme iron sites as a case study. Andreini C., Bertini I., Cavallaro G., Najmanovich R.J., Thornton J.M. s.l. : Journal of Molecular Biology, 2009, Vol. 388, pp. 356–380.
8. Factors controlling the reactivity of zinc finger cores. Lee YM, Lim C. 22, 2011, Journal of the American Chemical Society, Vol. 133, pp. 8691–8703.
9. Competition among Metal Ions for Protein Binding Sites: Determinants of Metal Ion Selectivity in Proteins. Dudev T., Lim C. 1, 2014, Chemical Reviews, Vol. 114, pp. 538–556.
10. Minimal functional sites allow a classification of zinc sites in proteins. Andreini C., Bertini I., Cavallaro G. 10, s.l. : PLoS One, 2011, Vol. 6, p. e26325.
11. MetalPDB: a database of metal sites in biological macromolecular structures. C. Andreini, G. Cavallaro, S. Lorenzini, A. Rosato. s.l. : Oxford Journals, 2013, Nucleic Acids Research, Vol. 41, pp. D312-D319.
12. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. s.l. : Springer Series in Statistics, 2009.
13. MetalS⁽³⁾, a database-mining tool for the identification of structurally similar metal sites. Valasatava Y, Rosato A, Cavallaro G, Andreini C. 6, 2014, JBIC Journal of Biological Inorganic Chemistry, Vol. 19, pp. 937-945.

Acknowledgements

This work was supported by MIUR (Ministero Italiano dell'Università e della Ricerca) through the FIRB projects RBF08WGXT and by the European Commission through the BioMedBridges and EGI-Engage project (grant no 284209 and 654142).

Figures

Figure 1

Graphical representation of the computational protocol to systematically compare and classify metal-binding sites on the basis of their structural similarity. The workflow includes the following steps: (1) select MFSs organized in equistructural groups (EGs, contain MFSs that are found in proteins with the same fold and occur at the same position within that fold); (2) cluster MFSs within each EG into groups of highly similar structures (first or intra-group stage); (3) for each cluster built at the first stage select a

representative MFS; (4) build the broader clusters comparing representative MFSs and organizing them in groups of similar structures (second or inter-group stage).