

# Metals<sup>2</sup>: A Tool for the Structural Alignment of Minimal Functional Sites in Metal-Binding Proteins and Nucleic Acids

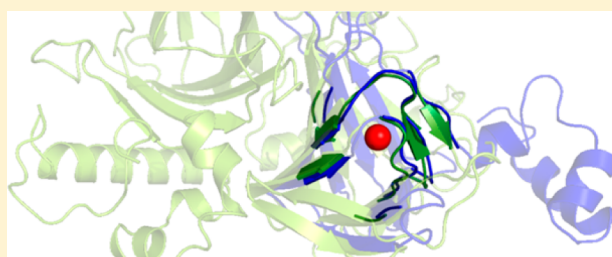
Claudia Andreini,<sup>\*,†,‡</sup> Gabriele Cavallaro,<sup>†</sup> Antonio Rosato,<sup>†,‡</sup> and Yana Valasatava<sup>†</sup>

<sup>†</sup>Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Florence, Italy

<sup>‡</sup>Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Florence, Italy

**S** Supporting Information

**ABSTRACT:** We developed a new software tool, Metals<sup>2</sup>, for the structural alignment of Minimal Functional Sites (MFSs) in metal-binding biological macromolecules. MFSs are 3D templates that describe the local environment around the metal(s) independently of the larger context of the macromolecular structure. Such local environment has a determinant role in tuning the chemical reactivity of the metal, ultimately contributing to the functional properties of the whole system. On our example data sets, Metals<sup>2</sup> unveiled structural similarities that other programs for protein structure comparison do not consistently point out and overall identified a larger number of structurally similar MFSs. Metals<sup>2</sup> supports the comparison of MFSs harboring different metals and/or with different nuclearity and is available both as a stand-alone program and a Web tool (<http://metalweb.cerm.unifi.it/tools/metals2/>).



## INTRODUCTION

Bioinorganic or biological inorganic chemistry is the discipline dealing with the interaction between inorganic substances and molecules of biological interest.<sup>1–3</sup> It is a rather wide field, because it addresses the role, uptake, and fate of elements essential for life, the response of living organisms to toxic inorganic substances, the function of metal-based drugs, the synthetic production of functional models, and so on. Within this scientific domain, the interaction between metal ions or metal-containing cofactors and biological macromolecules is often addressed at the 3D structural level, with atomic detail. These studies constitute an intersection between bioinorganic chemistry and structural biology.<sup>4</sup> The availability of the atomic coordinates of metal-macromolecule adducts allows a deeper understanding of the mechanisms by which the inorganic and protein or nucleic acid moieties influence the biochemical function of one another.<sup>5</sup>

Metal ions are bound to biological macromolecules via coordination bonds. The bonds are made by so-called donor atoms that can belong to either the polymer (protein or nucleic acid) backbone or side chains/bases. Additional donor atoms may belong to nonmacromolecular ligands, such as oligopeptides, small organic molecules, anions, water molecules. The ensemble comprising a metal ion (or cluster of metal ions) together with its donor atoms defines the metal-binding site. Metal-binding sites are occasionally extended to include all of the atoms in the amino acid or nucleotide. Such sites can be structurally characterized in high detail through X-ray crystallography and X-ray absorption spectroscopy.<sup>6–8</sup> Databases reporting on the geometric properties of metal-binding sites in proteins<sup>9</sup> or nucleic acids<sup>10</sup> are available. They are

derived from the coordinate files deposited in the Protein Data Bank<sup>11</sup> (PDB) resulting from the structural biology studies mentioned in the previous paragraph. Metal-binding sites have been shown to be useful for the bioinformatic analysis of metal-binding proteins (metalloproteins) and, in particular, for the prediction of metalloproteins from whole proteome sequences.<sup>12–14</sup> However, the functional properties associated with the occurrence of metal sites in biological macromolecules are not adequately described only on the basis of the metal coordination sphere.<sup>15–17</sup> For example, models of metal sites in proteins that include only the metal ligands may not be sufficiently accurate to reproduce biochemical functions. To increase the strength of the relationship with functional properties, the surroundings of the metal-binding site must also be taken into account. This larger ensemble can be thought of as the minimal environment determining metal function, which in previous work we dubbed the “minimal functional site” (MFS).<sup>18</sup> In practice, we defined an 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. 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 systematic structural comparison of MFSs of zinc proteins allowed a structure-based classification to be developed that is tightly connected to the functional properties of each site.<sup>18</sup> Indeed, the usefulness of the MFS concept outlined above has its chemico-physical foundation in the fact

**Received:** March 26, 2013

**Published:** October 11, 2013

that the local environment of the metal has a determinant role in tuning its properties and thus its chemical reactivity. Instead, the macromolecular matrix is instrumental to determine, e.g. substrate selection<sup>19</sup> or partner recognition.<sup>20</sup> A database of MFSs extracted from the structures deposited in the PDB is available.<sup>21</sup>

Here, we report the development of a software tool, called Metals<sup>2</sup> (Metal Sites Superposition), which allows two MFSs to be structurally aligned. Because MFSs are fragments of 3D macromolecular structures, this task is not possible with several of the available programs for structure comparison. In addition, by design Metals<sup>2</sup> starts its procedure to determine the best alignment of two MFSs with the superposition of the metal ions (or of the geometric center of polymetallic cofactors) and the comparison of the position of donor atoms. Consequently, the metal sites are always at the center of the structural alignment. This intrinsically reflects the philosophy underlying the construction of MFSs. Metals<sup>2</sup> is available both as a stand-alone program and a Web tool.

## METHODS

Determining the best global 3D alignment of two proteins is an NP-hard<sup>22</sup> problem. Even though the number of atoms in metal sites is somewhat smaller than in proteins, explicit methods are still not appropriate to tackle the determination of the best superposition of two metal sites. Consequently, we decided to rely on heuristics for this task.

In short, the basic idea underlying the Metals<sup>2</sup> program is to perform the superposition between two metal sites using a multistep approach. First, Metals<sup>2</sup> systematically computes initial poses built by superposing the geometric centers of the two metal cofactors and all the possible pairs of donor atoms from the two sites. Second, the poses are ranked on the basis of the Metals<sup>2</sup> score and the best 50% retained. Finally, the score of these structural alignments is optimized by allowing the geometric centers and the ligands to displace with respect to one another. Only the best scoring superposition is retained. The score that is optimized consists of three terms accounting respectively for the biochemical similarity of the amino acids put in correspondence, the ratio between the total length of the sequence alignment and the length of the smallest site (i.e., the fractional coverage of the smallest site), and the number and length of consecutive sequence segments in the superposition.

The whole procedure is detailed in the following paragraphs (a flow diagram is provided in Supplementary Figure S1).

In our previous work, we defined a metal-binding site as an assembly of residues around a metal ion or a cluster of metal ions.<sup>18</sup> This definition gives a pivotal role to the metal ion or the geometric center of polymetallic sites. It descends logically from this setting that when comparing the structures of the metal sites, the first step is to superpose these centers so that they coincide with the origin of coordinates (step 2 in Figure S1). Then, a number of initial poses are generated. To accomplish this task, all possible *local elementary patterns* (LEPs) are derived from the first coordination sphere for each structure (called qLEP for the query MFS and tLEP for the target MFS). For metal sites with at least two donor atoms, each LEP corresponds to three points in 3D space: one point coincides with the metal (or geometric center of metal cluster) and two other points correspond to two donor atoms. In practice, each LEP is a triangle whose vertices are the metal (or geometric center of metal cluster) and two of its donor atoms (step 3 in Figure S1). Consequently, for a site with  $N$  ligands,  $N(N-1)/2$  LEPs can be identified (for example, a site with four ligands has six LEPs). However, the comparison between a given tLEP and a given qLEP must be performed twice, as there are two possible ways to put the two pairs of donor atoms in correspondence. We do this by creating a permuted version of each tLEP in which the two donor atoms are swapped (step 4 in Figure S1). Thus, for two sites with  $N$  and  $M$  ligands respectively, a total of  $N(N-1) \times M(M-1)/2$  initial poses are created. For metal sites with a single monodentate ligand a LEP corresponds to two points in 3D space; in practice the LEP becomes a segment closed by a point coinciding with the metal (or geometric center of metal cluster) at one end and a point coinciding with the unique donor atom at the other end.

To generate one pose, Metals<sup>2</sup> superimposes a given tLEP to a given qLEP by rotating the former so that the sum of squared distances between the corresponding LEP vertices is minimized. The coincident vertex that corresponds to the superimposed metal ions is the rotation center (Supplementary Figure S2 and step 5 of Figure S1). The problem of finding the rotation matrix has been solved analytically using Kearsley's method<sup>23</sup> by means of an eigenvalue determination using quaternion algebra. In practical terms, to compute the rotation matrix we, first, construct the symmetric matrix from the coordinates of vertices in the LEPs

$$\begin{pmatrix} \sum (x_m^2 + y_m^2 + z_m^2) & \sum (y_p z_m - y_m z_p) & \sum (x_m z_p - x_p z_m) & \sum (x_p y_m - x_m y_p) \\ \sum (y_p z_m - y_m z_p) & \sum (x_m^2 + y_p^2 + z_p^2) & \sum (x_m y_m - x_p y_p) & \sum (x_m z_m - x_p z_p) \\ \sum (x_m z_p - x_p z_m) & \sum (x_m z_p - x_p z_m) & \sum (x_p^2 + y_m^2 + z_p^2) & \sum (y_m z_m - y_p z_p) \\ \sum (x_p y_m - x_m y_p) & \sum (x_m z_m - x_p z_p) & \sum (y_m z_m - y_p z_p) & \sum (x_p^2 + y_p^2 + z_m^2) \end{pmatrix} \quad (1)$$

where each sum runs over the two pairs of corresponding vertices in the tLEP and the qLEP,  $x_m = (x' - x)$ ,  $x_p = (x' + x)$ ,  $(x', y', z')$  and  $(x, y, z)$  being the coordinates of the tLEP and the qLEP vertices. Analogous definitions hold for  $y_m, y_p, z_m$ , and  $z_p$ . The next step is to find eigenvalues and eigenvectors of the matrix. The eigenvector corresponding to the smallest positive eigenvalue gives a unit quaternion representing the rotation

that minimizes the sum of the distances between all corresponding points. For the unit quaternion  $(x, y, z, w)$  the corresponding rotation matrix  $M$  is defined as follows:

$$M = \begin{pmatrix} 1 - 2y^2 - 2z^2 & 2xy + 2wz & 2xz - 2wy \\ 2xy - 2wz & 1 - 2x^2 - 2z^2 & 2yz + 2wx \\ 2xz + 2wy & 2yz - 2wx & 1 - 2x^2 - 2y^2 \end{pmatrix} \quad (2)$$

The matrix computed in this way is then applied to all the atoms in the target site (step 10 in Figure S1), generating the new coordinate set that defines one pose. The procedure is repeated for each possible qLEP and tLEP pair, including permuted tLEPs.

This approach needs an extension to deal with cases where at least one of two sites has only a monodentate ligand. To ensure alignment of donor atoms we switch from the superposition of triangles to the superposition of segments (LEPs in both sites are now considered as segments). Metals<sup>2</sup> carries out a superposition of segments in all-versus-all fashion. For each qLEP and tLEP pair, we calculate a first rotation matrix (step 7 in Figure S1) that aligns the two corresponding segments. Then, Metals<sup>2</sup> performs a number of additional rotations (step 8 in Figure S1) around the axis corresponding to the superposed segment, achieving a complete sampling in 20° steps, i.e. for a total of 17 rotations. Quaternions are used to make a rotation around a single axis. The formula for quaternion  $q$  in terms of an axis angle is

$$q = \cos \frac{\alpha}{2} + i \left( x \cdot \sin \frac{\alpha}{2} \right) + j \left( y \cdot \sin \frac{\alpha}{2} \right) + k \left( z \cdot \sin \frac{\alpha}{2} \right) \quad (3)$$

where  $x$ ,  $y$ , and  $z$  represent the axis vector about which the rotation occurs, and  $\alpha$  is an angle that defines the amplitude of the rotation about the axis. Quaternions are used to compute rotation matrices as described before for the general case of two sites with multiple donor atoms. The initial rotation matrix is then multiplied by each of the 17 subsequent rotation matrices (step 9 in Figure S1) to obtain as many complete rotations that, applied to the target site, generate 17 different initial poses for each pair of qLEP and tLEP (step 10 in Figure S1).

The above initial poses are then ranked using the Metals<sup>2</sup> score. It is first necessary to assign correspondences between the atoms in the two structures being compared (atom matching). To identify the position in space of amino acidic residues, we used the coordinates of the  $C\alpha$  and  $C\beta$  atoms (for Gly we used only the  $C\alpha$  atom). Using only the  $C\alpha$  atoms in order to determine the correspondence between residue pairs is a relatively common approach that has been successfully exploited in the widely used programs for structural alignment like MAMMOTH, CE, TM-align, FATCAT, and FAST. There is indeed extensive demonstration in the scientific literature that the knowledge of the  $C\alpha$  trace is sufficient to accurately reconstruct the full coordinate set for the backbone of a protein structure. For the present application, the additional information provided by the  $C\beta$  atoms is useful as the  $C\alpha$ - $C\beta$  bond represents the direction of the side chain with respect to the main chain.  $C\beta$  interactions provide additional information on fold energetics.<sup>24</sup> By using only  $C\alpha$  and  $C\beta$  pairs, the calculation approach is essentially independent of the amino acidic sequence, except for Gly, thus facilitating the comparison of highly different sites. For nucleic acids, the pair formed by the C1 atom of the sugar and the N1 atom for pyrimidine bases or the N9 atom for purines was used. Thus, our representation of MFSs takes into account not only the

positions of residues along the main chain but also the orientation in space of amino acidic side chains and nucleic bases. Atomic coordinates are used to establish one-to-one correspondences between the residues in the two sites being superposed. Atoms are matched based on their distance. For each  $C\alpha$  atom from the first site (query site) we assign a correspondence to the  $C\alpha$  atom in the second (target) site that is closest in space. When looking for the closest atom from the target site, we restrict the search within a radius of 2 Å around the atom of the query site. If there is no atom of the target structure in this range, the atom of the query structure will remain unmatched. If both atoms in a  $C\alpha$ - $C\alpha$  (or C1-C1) pair are bound to a  $C\beta$  (or N1/N9) atom, we also compute the distance between the two  $C\beta$  atoms and use it to assign a correspondence between them with the same criterion. Ligand residues are handled separately and can only be put in correspondence to ligand residues in the other MFS. A less restrictive threshold of 5 Å is applied for ligands in order to enhance coverage. In order to perform an efficient search, the atoms from the target structure are organized in a kd-tree. After the assignment of correspondences, it is possible to calculate the score that is used to rank the poses obtained for a pair of sites. For this purpose, we evaluate three different terms:

1. A **relative coverage term**, depending on the ratio between the number of atoms put in correspondence ( $c$ ) and the maximum possible number of atom correspondences for the sites being compared ( $C_{\max}$ );  $C_{\max}$  in practice, equals the total number of  $C\alpha$  and  $C\beta$  atoms of the site with the shortest sequence. For example, if a query site containing 10 residues is to be compared with a target site of 20 residues,  $C_{\max}$  is a fixed integer value given by the number of all  $C\alpha$  and  $C\beta$  atoms of the query structure. Instead,  $c$  is the number of matched atoms in the pose being scored. By definition,  $c/C_{\max} \leq 1$ . Values close to 1 indicate that the large majority of the atoms in the smaller site have been matched to atoms in the larger site. We decided to implement this term as  $\ln(C_{\max}/c)$ . In this way, if all atoms in the smaller site have been matched, the contribution of the current term to the total score is zero.

2. A **sequence similarity term**, depending on the ratio between the similarity score ( $S$ ) computed using the BLOSUM62 matrix for the sequence alignment derived from the  $C\alpha$  correspondences and the similarity score that would be obtained if the two sites being aligned had identical sequences ( $S_{\max}$ ). To compute  $S_{\max}$  we consider the sequence giving the lowest similarity score to itself. For nucleic acids we used a simple scoring system that consists of a "reward" for a match (+5) and a "penalty" for a mismatch (-4). The term is formulated as

$$\left( 1 - \frac{S}{S_{\max}} \right) \quad (4)$$

3. A **fragmentation term**, which takes into account how many fragments the alignment is broken into and how long each segment is. This term is formulated as follows

$$\frac{\sum_{f=1}^F \frac{1}{n_f}}{N} \quad (5)$$

where  $F$  is the total number of fragments,  $n_f$  is the length of (i.e., number of residues in) the  $f$ -th fragment, and  $N$  is the total alignment length.  $N$  is used as a kind of normalization factor, as larger sites are less likely to overlap completely. Because MFSs



are often discontinuous fragments of protein structure, this term is generally not null even for self-alignments.

Each term describes quantitatively an essential property of the structural alignment. We believe it is preferable to rank MFS structural alignments on the basis of a small number of terms that are interpretable by the user. We therefore place emphasis on the physical and chemical interpretation of the terms in the scoring function. The implicit assumption is that by comparing two very similar sites one will obtain “good” scores for all terms.

To give the total Metals<sup>2</sup> score,  $T$ , the three terms above were linearly combined as follows

$$T = w_1 \frac{\sum_{f=1}^F \frac{1}{n_f}}{N} + w_2 \ln\left(\frac{C_{\max}}{c}\right) + w_3 \left(1 - \frac{S}{S_{\max}}\right) \quad (6)$$

where  $w_1$ ,  $w_2$ , and  $w_3$  are the relative weight factors of the three terms, which were set equal to 1.5, 1.0, and 2.5, respectively. Note that with the current formulation the better solutions are those with the *lower* scores. The present scoring scheme allows metal sites in proteins to be compared with other metal sites in proteins as well as metal sites bound to nucleic acids to be compared with other metal sites in nucleic acids. “Cross-category” alignments are not possible.

After ranking all the poses generated for a pair of MFSs, those having a score in the best 50% of the observed score range are retained for optimization. In this stage, the atom correspondences already established are used to minimize the RMSD of the coordinates of the two sites

$$\text{RMSD} = \sqrt{\sum_{i=1}^{C_{\max}^*} \frac{(x_i^A - x_i^B)^2}{C_{\max}^*}} \quad (7)$$

where  $x_i^A - x_i^B$  is the distance between the  $i$ -th atom pair, and  $C_{\max}^*$  is the number of matched  $C\alpha$ ,  $C\beta$  atom pairs to which we added the pair of the metal ions (or of the geometric centers of polymetallic sites). The RMSD is minimized by roto-translating the target site; the roto-translation matrix is calculated using Singular Value Decomposition of the covariance matrix of the coordinates of the above-mentioned pairs. After roto-translation, for each pose the atom matching procedure is repeated to update the atom correspondences and the Metals<sup>2</sup> score is recalculated. Poses are then reranked. If the new best scoring pose has a total score worse than the best scoring pose before RMSD minimization, then the change is rejected. Otherwise the new best scoring pose is retained as the final solution.

For the final solutions, the correlation between the various terms was examined on the basis of the simple Pearson correlation coefficients. Pearson coefficients were used to discriminate different sets of weights, with the aim of finding the set balancing the different terms with respect to one another and also with respect to their contribution to the total score.

**Implementation.** All scripts are implemented in Python (<http://www.python.org/>) on a Linux platform. The reasons for choosing this language were as follows:

- The availability of p3d,<sup>25</sup> a Python module for structural bioinformatics. In particular, the Protein class with a set of functions greatly simplifies handling structures.
- Multiplatform: runs on Windows, Linux/Unix, Mac OS X and has been ported to the Java and .NET virtual machines.
- Free to use, even for commercial products, because of its OSI-approved open source license.

The running time of the program comparing a pair of metal site structures on an Intel(R) Core(TM) i5 CPU 650 @ 3.20 GHz processor varies from seconds to a few minutes, depending on the size of the two structures.

**Calculations with Other Programs for Structural Alignment.** We used the following structure alignment programs to compare their results with Metals<sup>2</sup>, on a statistical basis: FAST,<sup>26</sup> MAMMOTH,<sup>27</sup> and TM-align.<sup>28</sup> These tools were chosen among the relevant programs included in a recent review,<sup>29</sup> because they are able to handle protein fragments despite being designed for the alignment of entire structures. The only exception was the program MUSTANG,<sup>30</sup> which can align protein fragments. However, we were not able to exploit it, because its output score, which includes the RMSD of the superposition and the number of atoms superimposed, was not readily applicable to discriminate positive and negative alignments; in addition, no indications of thresholds were available from the authors. FAST was not included in the aforementioned review<sup>29</sup> but was successfully used by some of us in the past for similar applications.<sup>18,31</sup> All the programs were run with default parameters. The thresholds used to identify reliable alignments were as follows: > 1.5 for FAST; > 4.0 for MAMMOTH; > 0.5 for TM-align.

**Data Sets Used.** To test the results of Metals<sup>2</sup> we used two data sets previously analyzed by some of us. The first one (Fe-data set) consists of 86 MFSs containing nonheme iron,<sup>31</sup> whereas the second one (Zn-data set) consists of 367 MFSs containing zinc.<sup>18</sup> The small size of the Fe-data set allowed us to inspect results manually. For the sake of performance characterization, we classified MFS pairs that all the programs for structural alignment used in this work aligned with a poor score (i.e., lower than one-third of the recommended threshold for meaningful alignments given by each program's authors) as negative examples. For positive cases, we adopted pairs of MFSs that at least one program could align with a score better than the program's recommended threshold. For the Fe-data set, all positive examples were manually checked to remove instances where the metal ions were not superimposed in the structural alignment.

The performance of Metals<sup>2</sup> in the analysis of the above test sets was evaluated using the following parameters

$$\begin{aligned} \text{Matthews correlation coefficient (MCC)} \\ = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{aligned} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

Metals<sup>2</sup> as well as all other programs were further run to structurally align all possible MFS pairs from both the full Fe- and full Zn-data sets. To check whether there was a statistically significant difference between the results of Metals<sup>2</sup> and those of each other program, we applied the Wilcoxon rank sum test using its Matlab implementation.

For functional analysis, we took the functional assignments available from the articles describing the two data sets.<sup>18,31</sup> Zinc MFSs were assigned one function among catalytic, structural, regulatory, substrate, and unknown; nonheme iron MFSs were

**metalS<sup>2</sup>**

HOME Download Help Contact Us About Us Go to Metal PDB

MetalS<sup>2</sup> is a tool for pairwise structural alignment of *Minimal Functional Sites* (MFSs) found in metal-binding biological macromolecules.

**Input Query Site (Site 1)**

1 PDB code:

2 PDB file:

3 Threshold coordination distance:

4 Excluded donor atoms:

5 Metal of interest (optional):

**Input Target Site (Site 2)**

1 PDB code:

2 PDB file:

3 Threshold coordination distance:

4 Excluded donor atoms:

5 Metal of interest (optional):

**Select Sites**

Show 10 entries

Metal	Ligands	Select
ZN_402(B)	CYS_23_B, CYS_61_B, CYS_26_B, CYS_58_B	<input checked="" type="radio"/>
CA_405(B)	GLU_86_B, HOH_448_B, GLY_81_B	<input type="radio"/>
CA_404(B)	ASP_40_B, HOH_444_B, ASN_62_B	<input type="radio"/>
CA_403(A)	HOH_450_A, ASP_40_A, ASN_62_A, HOH_447_A	<input type="radio"/>
ZN_401(A)	CYS_61_A, CYS_23_A, CYS_26_A, CYS_58_A	<input type="radio"/>

Previous Next

**Select Sites**

Show 10 entries

Metal	Ligands	Select
ZN_153(O)	ASP_81_O, HIS_78_O, HIS_61_O, HIS_69_O	<input checked="" type="radio"/>
CU_152(O)	HIS_118_O, HIS_44_O, HIS_61_O, HIS_46_O	<input type="radio"/>
ZN_153(G)	HIS_78_G, ASP_81_G, HIS_69_G, HIS_61_G	<input type="radio"/>
CU_152(G)	HIS_118_G, HIS_44_G, HIS_61_G, HIS_46_G	<input type="radio"/>
CU_152(B)	HIS_46_B, HIS_44_B, HIS_118_B, HIS_61_B	<input type="radio"/>
ZN_153(B)	HIS_78_B, HIS_69_B, ASP_81_B, HIS_61_B	<input type="radio"/>
ZN_153(Y)	HIS_69_Y, HIS_61_Y, ASP_81_Y, HIS_78_Y	<input type="radio"/>
CU_152(Y)	HIS_46_Y, HOH_154_Y, HIS_61_Y, HIS_118_Y, HIS_44_Y	<input type="radio"/>

Previous Next

**Specify Job Information**

RMSD Threshold:

E-mail:

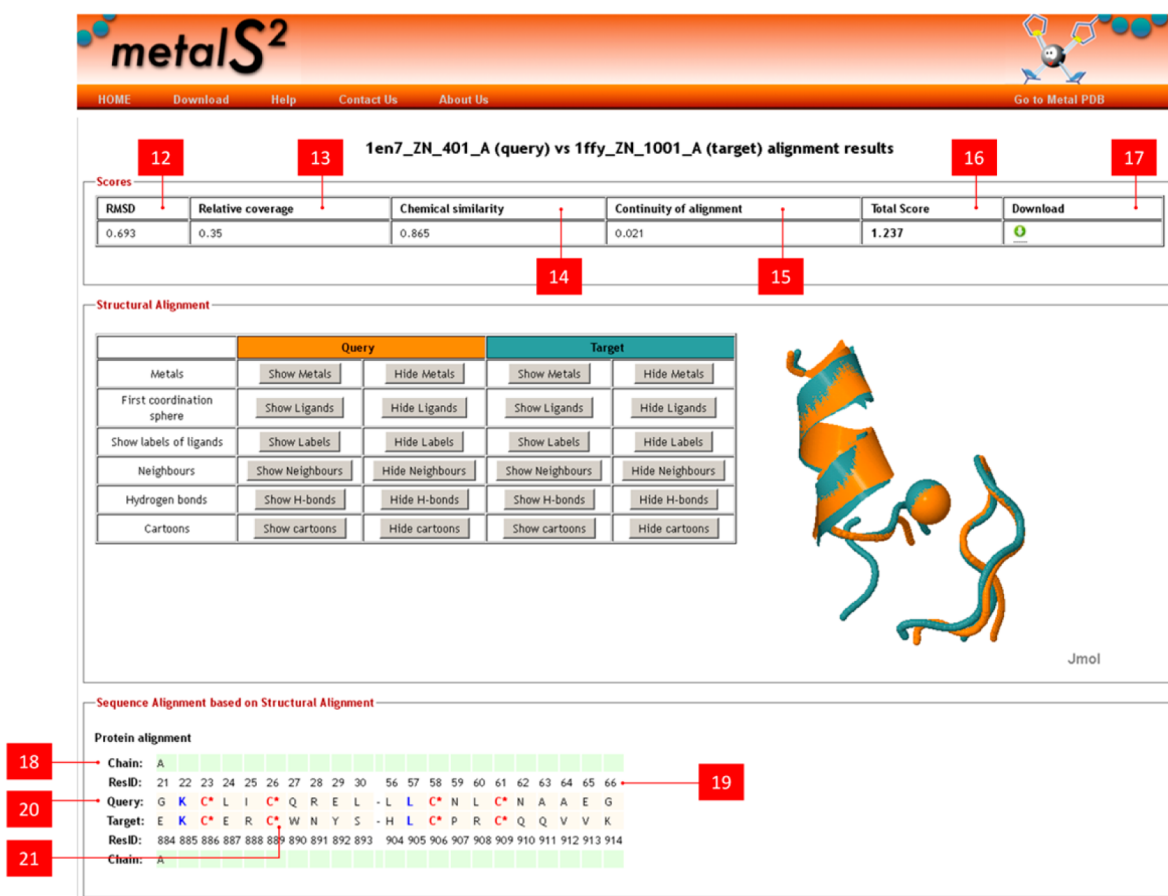
**Figure 1.** Input form on the MetalS<sup>2</sup> Web page. Top: Selection of PDB entry/upload of PDB file. (1) PDB code fields; (2) Button to upload a PDB file from the local disk (alternative to 1); (3) Distance from the metal used to identify donor atoms; (4) Comma-separated list of chemical elements not allowed to be donor atoms; (5) Selection of specific metal elements for MFS identification (optional). Bottom: selection of an individual MFS within each structure. Each record in the Tables (6) represents an MFS contained in one of the input PDB files. The two MFSs to be aligned are selected by checking the corresponding radio buttons in the “Select” columns (7). The number of MFSs shown per page can be adjusted from the default value of 10 (8), while the different pages can be navigated using the Next/Previous links (9). The threshold for the assignment of correspondences between the atoms of the two MFSs can be adjusted (10). The field (11) can be used to provide an e-mail address to which the link to the results will be sent.

assigned one function among catalytic, structural, electron transfer, sensing, and unknown. Unknown-unknown matches were not taken into account.

## RESULTS

MetalS<sup>2</sup> has been implemented and made available both as a stand-alone program and via a Web portal within our MetalPDB platform (Figure 1). The metal sites to be compared by MetalS<sup>2</sup> are identified in the input protein structures using a previously described approach.<sup>21</sup> In practice, the ligands to each metal atom in each structure are first identified, as having at least one non-hydrogen atom at a distance smaller than 2.8 Å

(this threshold can be adjusted by the user) from the metal. They can be residues in a polypeptide or a polynucleotide chain (endogenous ligands) as well as different ions or molecules such as water, sulfide, acetate (exogenous ligands). Organic cofactors such as heme are considered exogenous ligands. Each pair of metal atoms that have at least one common ligand, such as a bridging amino acidic side chain or exogenous anion, or whose distance is lower than 5 Å is included into a single polynuclear site. This procedure is iterated such that if metal A and metal B are to be included into a single site and then metal B and metal C are also to be included in a single site, eventually a three-nuclear site is formed that contains all three metal ions.



**Figure 2.** Results of the MFS comparison shown on the MetalS<sup>2</sup> output page. Top: table of scores and download button; (12) shows the RMSD calculated over the atoms paired in the two MFSs. Cells (13)–(15) display the MetalS<sup>2</sup> score components; the total score appears in (16). An archive containing all output files can be downloaded by clicking on the arrow in the “Download” column (17). Middle: Interactive display of the MFS superposition. Bottom: sequence alignment derived for the superposition of the MFSs. The alignment is visualized as follows: the first line of the description gives a reference to the chain (18) containing the aligned residues; the second line displays the number of each aligned residue within its chain (19); the third line shows the aligned residues using the one-letter code (20). Ligands (21) are highlighted by an asterisk.

This procedure allows e.g. Fe<sub>4</sub>S<sub>4</sub> clusters found in ferredoxins to be defined as an individual four-nuclear site. The neighbors of all the ligands are then identified as containing at least one non-hydrogen atom at a distance smaller than 5.0 Å from any ligand. The ensemble of the neighbors, the ligands, and the metal atom(s) constitute the MFS.<sup>18,21</sup> The MetalS<sup>2</sup> portal can automatically search structures deposited in the PDB for MFSs, taking as input the corresponding PDB code. The MFSs are presented to the user in a table, from which it is possible to select one of the MFSs for superposition (Figure 1). Thus, there is no need for the user to download/upload metal-containing structures that are available from the PDB, whereas it is mandatory for structures not publicly available. For each superposition, the user is presented with information on the values of the different components of the score, the RMSD value of the best solution, and the superposition-derived sequence alignment (Figure 2). In addition, the tool allows the superposition to be visualized and manipulated, using Jmol. The MFSs coordinates rotated in the same Cartesian reference frame can be downloaded in PDB format and visualized e.g. with Pymol, using a script output by the program. A link to the results is optionally sent by e-mail (Figure 1).

The program has been tested using two data sets containing respectively proteins binding nonheme iron ions (Fe-data set) and zinc ions (Zn-data set), which were described in previous

publications by some of us.<sup>18,31</sup> The Fe-data set contains 86 proteins; its relatively small size allowed us to manually analyze the results. The Zn-data set contains 367 proteins, resulting in 67161 pairwise comparisons, which constitute a large enough basis for statistical analysis. Both data sets are nonredundant, i.e. for all proteins belonging to the same SCOP<sup>32</sup> or CATH<sup>33</sup> superfamily only one representative was kept. In this way, we minimized the number of homologous proteins in the data set, whose structures are expected to be very similar<sup>34</sup> and thus would result, if included, in a less stringent testing of the program.

We systematically aligned all the MFSs in the two data sets with different programs: FAST,<sup>26</sup> MAMMOTH,<sup>27</sup> and TM-align.<sup>28</sup> Among these, the FAST program was already shown by some of us to have an acceptable performance when applied to similar analyses.<sup>31</sup> When analyzing the Fe-data set with default parameters, the programs produced a number of reliable (i.e., having a score for the structural alignment better than the threshold indicated by the program's authors) superpositions between 3 (MAMMOTH) and 23 (FAST). However, we observed that in some cases (e.g., five FAST superpositions) despite the good score, the metal ions and the ligands were not structurally aligned. After removing these instances, we obtained a total of 21 superpositions classified as reliable by at least one of the programs, which we took as our test set of

Table 1. Analysis of the Output Produced by Metals<sup>2</sup> on the Test Sets Derived from the Fe- and Zn-Data Sets<sup>a</sup>

	threshold									
	1.75	2	2.25	2.5	2.75	3	3.25	3.5	3.75	4
Fe-Data Set										
TP	2	7	9	12	16	18	18	20	21	21
TN	16	16	16	16	16	15	10	9	5	3
FP	0	0	0	0	0	1	6	7	11	13
FN	19	14	12	9	5	3	3	1	0	0
MCC	0.209	0.422	0.495	0.605	0.762	0.788	0.500	0.574	0.453	0.340
Zn-Data Set										
TP	418	546	696	795	857	888	920	944	951	961
TN	1637	1637	1629	1609	1557	1436	1211	880	520	223
FP	0	0	8	28	80	201	426	757	1117	1414
FN	546	418	268	169	107	76	44	20	13	3
MCC	0.570	0.672	0.780	0.839	0.845	0.782	0.671	0.525	0.364	0.228
Cumulative										
TP	420	553	705	807	873	906	938	964	972	982
TN	1653	1653	1645	1625	1573	1451	1221	889	525	226
FP	0	0	8	28	80	202	432	764	1128	1427
FN	565	432	280	178	112	79	47	21	13	3
MCC	0.564	0.667	0.774	0.834	0.844	0.782	0.669	0.526	0.365	0.230
Performance Metrics										
precision	100.0%	100.0%	98.9%	96.6%	91.6%	81.8%	68.5%	55.8%	46.3%	40.8%
accuracy	78.6%	83.6%	89.1%	92.2%	92.7%	89.3%	81.8%	70.2%	56.7%	45.8%

<sup>a</sup>TP: True positives (number of MFS pairs aligned by Metals<sup>2</sup> with a total score below the selected threshold, and aligned by at least one of the other programs tested with a satisfactory score), TN: True negatives (number of MFS pairs aligned by Metals<sup>2</sup> with a total score above the selected threshold, and aligned by all the other programs tested with a poor score), FP: False positives (number of MFS pairs aligned by Metals<sup>2</sup> with a total score below the selected threshold, and aligned by all the other programs tested with a poor score), FN: False negatives (number of MFS pairs aligned by Metals<sup>2</sup> with a total score above the selected threshold, and aligned by at least one of the other programs tested with a satisfactory score), MCC: Matthews correlation coefficient.

positive examples. For negative examples, we used MFS pairs whose superpositions were classified very poorly (i.e., lower than one-third of the indicated threshold) by all programs (16 instances). With these assumptions, we could test the performance of Metals<sup>2</sup> as a function of the selected threshold for its total score (Table 1). Based on the Matthews correlation coefficient, the optimal threshold lies between 2.75 and 3.0. A similar reasoning could be applied to the Zn-data set, resulting in a somewhat larger test set of 964 positive examples and 1637 negative examples. For these, the Matthews correlation coefficient is maximum between 2.5 and 2.75. Combining the two test sets derived from the Fe- and Zn-data sets results in a broad maximum at 2.75 (Table 1), which can therefore be taken as the threshold below which the Metals<sup>2</sup> total score indicates a good structural alignment. With this threshold, the precision of Metals<sup>2</sup> on the combined test set is 91.6% and its accuracy is 92.7%; at a threshold of 2.25 the precision of Metals<sup>2</sup> is 99%.

Over the entire Fe-data set, 27 MFS pairs could be superimposed by Metals<sup>2</sup> with a score lower than 2.25 (as compared with 21 for the three other programs tested altogether). Over the entire Zn-data set, 4072 MFS pairs could be superimposed by Metals<sup>2</sup> with a score lower than 2.25 (as compared with 964 for the three other programs tested altogether). The complete output is given in Supplementary Tables S1 and S2. Altogether, at the 2.25 threshold the ratio between the number of alignments produced by Metals<sup>2</sup> and by the other programs is 4.16 (Table 2). The ratio increases with increasing threshold for the Metals<sup>2</sup> score. The Metals<sup>2</sup> score of the top 1,000 structural alignments of MFS pairs from the Zn-data set (excluding self-alignments) ranges from 0.271 to

Table 2. Number of Structural Alignments for Which the Metals<sup>2</sup> Score Was below the Threshold and Its Ratio to the Number of Positive Cases Identified by the Other Programs

threshold	no. of structural alignments below the threshold	ratio of Metals <sup>2</sup> vs all other programs combined
1.75	1268	1.29
2.0	2362	2.40
2.25	4099	4.16
2.5	6590	6.69
2.75	9901	10.1
3.0	14,945	15.2

1.675. These include 362 reliable alignments from FAST as well as 240 instances where instead FAST was unable to produce an output. TM-align, instead, produced only 44 reliable superpositions with no failures and MAMMOTH featured no reliable superpositions as well as two failures. According to the authors' criteria, 35 of these MFS pairs would be dubbed as having no similarity by TM-align, with a Metals<sup>2</sup> score ranging between 0.997 and 1.673. We used the Wilcoxon rank sum test to check whether there was a statistically significant difference between the results provided by Metals<sup>2</sup> and by the other methods over the entire data sets. The test demonstrated that this was actually the case (Supplementary Table S3).

We then checked whether the MFS pairs that Metals<sup>2</sup> could align with a score below a given threshold were functionally related (Table 3). To this end, we exploited the functional assignments already published.<sup>18,31</sup> At the threshold of 2.25, which defines high-quality structural alignments, the percentage of functional matches was as high as 96.1%. The percentage of



**Table 3. Number of Matching Functional Assignments for MFS Pairs Aligned by Metals<sup>2</sup> As a Function of the Total Score**

Metals <sup>2</sup> score	matches	mismatches	% of matches
<2.0	1637	44	97.4%
<2.25	2838	115	96.1%
<2.5	4562	245	94.9%
<2.75	6574	549	92.3%

matches is threshold-dependent and decreases with increasing threshold, reaching 92.5% at a threshold of 2.75.

## DISCUSSION

In the present work, we present a tool that has been developed specifically for the structural comparison of pairs of MFSs. MFSs extend beyond metal-binding sites as the latter include only the metal ion (or polymetallic cluster) and the ligands to it, whereas MFSs additionally include ligand neighbors, i.e. other residues or chemical species in contact with the ligands.<sup>18</sup> Focusing on MFSs allows functional linkages between different proteins of known structure to be made with greater confidence than with metal-binding sites. This is because the ligand neighbors play a crucial role in tuning the properties of the metal-binding site and, in particular, the reactivity of the metal ion. The systematic comparison of MFSs thus is quite informative on the functional features of metalloproteins and metalloprotein families.<sup>18</sup> Therefore, the availability of a dedicated tool for the structural comparison of MFSs is of interest to bioinorganic chemists. A crucial feature of such a tool must be that it takes explicitly into account the fact that MFSs are built around metal sites. Consequently the structural comparison should start from there. Even approaches aimed at the structural comparison of protein binding sites, a task conceptually similar to ours, often include various other features in addition to “simple” 3D structure (e.g., surface structural patches<sup>35</sup> or shape descriptors<sup>36</sup>) but without taking into account the metals explicitly. Programs designed to compare protein structures at the entire chain level also do not exploit the presence of the metal sites and sometimes are actually unable to manage MFSs altogether. Out of a list of seven widely used tools whose performance was recently analyzed,<sup>29</sup> CE,<sup>37</sup> DALI,<sup>38</sup> TOPMATCH,<sup>39</sup> and SALIGN<sup>40</sup> do not yield any result when MFSs are used as input.

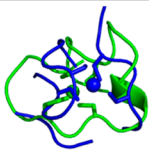
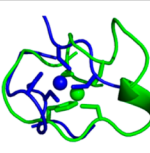
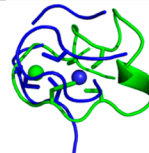
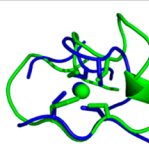
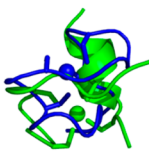
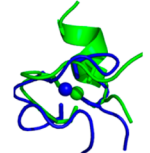
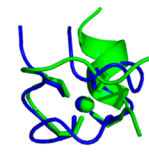


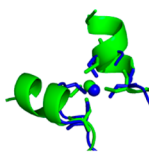
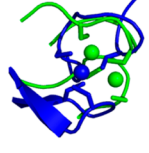
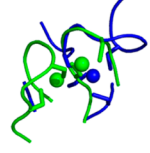
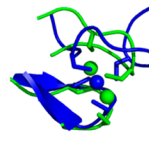
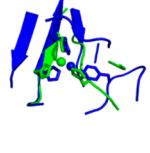
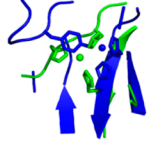
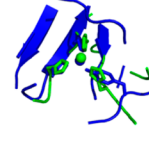
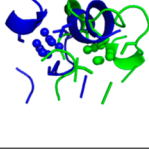
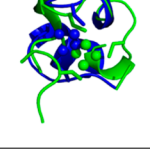
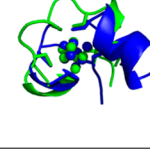
As its first step, Metals<sup>2</sup> identifies and extracts the portion of the metal-bound structure of interest (i.e., the MFSs), through a relatively simple distance-based protocol. Then, differently than any other program for global or local structure comparison we apply a metallo-centric view by immediately superposing the centers of the metal ions or polymetallic cofactors contained in the two MFSs. This and the subsequent alignment of metal ligands drive the rest of the structure comparison, thus effectively pruning configurations in which the two metal sites are not well superposed. All possible superpositions that do not fulfill this precondition are in practice never explored. This philosophy is unique among programs for either local or global macromolecular structure comparison and has to be taken into account when comparing the results of Metals<sup>2</sup> to the results of other programs. A notable implication of these considerations is that Metals<sup>2</sup> is unable to identify configurations in which traditional programs would obtain a satisfactory superposition of e.g. the backbone of the

polypeptide chain at the expense of putting the metal sites far apart.

Metals<sup>2</sup> was tested by systematically performing pairwise superpositions of all MFSs in two data sets of respectively 86 nonheme iron-binding proteins and 367 zinc-binding proteins that did not contain homologues. The three contributions to the total score of Metals<sup>2</sup> have different relative importance in determining its output: the size term spans the largest range (from 0 to 3.42), the biochemical similarity term spans the smallest range (from 0 to 1.28), and the fragmentation term spans an intermediate range (from 0.01 to 2.35). The range spanned by the total score is from 0.271 to 6.64. The three terms do not have a statistically significant correlation (the Pearson coefficients between them being all lower than 0.4). The size part term is relevant to penalize superpositions where only a minor portion of one of the two MFSs can be matched to the other. This is important as MFSs are a shell of relatively small thickness around the metal center and thus it is unlikely that superpositions in which only a minority of the atoms is overlapped can reveal meaningful relationships. This is at variance with the case of protein structures, where, for example, the superposition of a relatively small motif or domain to a full structure can provide insightful indications. As a reference, 80% coverage corresponds to a value of the size term of 0.33 whereas 50% coverage corresponds to 1.04. The chemical similarity term is presumably limited by the fact that the present data set does not contain sequences with particularly unbalanced aminoacidic composition. The fragmentation term, finally, penalizes cases where extensive coverage of the MFSs being aligned could be obtained by combining many small, nonconsecutive regions of the sites, e.g. by overlapping two  $\beta$ -sheets in a crossed manner.

The benchmark used to assess the performance of Metals<sup>2</sup> recruited only about 2,500 out of the 70,816 MFS pairs (3.5%) resulting from the complete Zn- and Fe-data sets. The full set of MFS pairs could thus be meaningfully used as a basis to compare the outputs of three different programs for structural alignment. This analysis indicates that for each MFS pair there is typically no consistency between the programs (Supplementary Tables S1 and S2). In particular, FAST is the program that provides, according to its own scoring measures, the largest number of potentially meaningful structural alignments, even though it is also the program that fails to provide an output in the largest number of cases. This lack of consistency may partly be due to the fact that the scoring functions of the programs and their corresponding confidence thresholds have been calibrated for the alignment of full protein structures. Indeed, this may prevent the user from discriminating good and bad alignments, especially when analyzing large structural data sets. Metals<sup>2</sup> on the other hand is consistently capable of aligning MFSs and its total score can be used as an indicator of the quality of alignment. Specifically, we expect that nearly all of the alignments having a score lower than 2.25 are meaningful. With this threshold, Metals<sup>2</sup> identifies a number of MFS pairs that can be superposed well more than four times larger than the other programs (Table 2). This does not imply that some of these MFS pairs cannot be well aligned also by another tool, rather that Metals<sup>2</sup> provides a better way to recognize them. Still, by inspecting Metals<sup>2</sup> alignments close to the 2.25 threshold also in comparison with the output of the other programs, it was possible to identify various cases where Metals<sup>2</sup> was the only program that could produce a good quality alignment; some examples are given in Figure 3. The



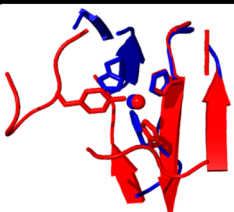
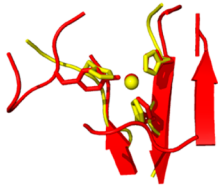
	MAMMOTH	TM-ALIGN	FAST	MetalS <sup>2</sup>
1co4_1 vs. 3ifu_2				
1jwe_1 vs. 2rhq_1			—	
1nj3_1 vs. 3h84_1			—	
2i1o_1 vs. 2zze_1			—	
1dmh_2 vs. 2fiy_4	—			
1kg2_1 vs. 2fug_34			—	

**Figure 3.** Comparison of selected structural alignments by MetalS<sup>2</sup> and the other programs tested in this work. The absence of the image indicates that the program did not produce any output. Site names in the first column correspond to those adopted in the MetalPDB database.<sup>21</sup>

analysis of MCC's as a function of the threshold suggests that 2.75 is the best value in terms of trade-off between the number of additional true and false positives introduced by raising the threshold with respect to 2.25. We thereby identify scores between 2.25 and 2.75 as a "shadow zone" where alignments are often meaningful, but some care in interpreting the results is needed and the alignments should be closely inspected. MFS pairs that are aligned by MetalS<sup>2</sup> with scores even higher than 2.75 are in the majority of cases not structurally similar. However, it is possible that potentially informative alignments fall in this range of scores (see the false negative rows in Table 1). These are commonly cases where the superposition requires the metal ions (or the geometric centers of polymetallic cofactors) not to be exactly coincident. MetalS<sup>2</sup> is in fact unable to identify elements of structural similarity in sites where the relative position of the metal cofactors with respect to the protein frame is different. There is merit in both metal-driven structural alignments and traditional protein/nucleic acid-driven alignments and thus both should be examined. Nevertheless, the concept of MFS, in its various but related forms, is of primary and central concern to the bioinorganic

chemist.<sup>41–43</sup> Hence, the need for an approach to 3D structure comparison that incorporates the underlying philosophy of MFSs such as MetalS<sup>2</sup>. On the other hand, structural similarities that do not take into account or do not highlight metal site similarity can be retrieved by a wide portfolio of software tools.<sup>29,44</sup>

As a general procedure, one would presumably rely on a combination of traditional protein-centered and metal-centered structural alignments to obtain functional hints from 3D structures. The correlation between the quality of the MFS alignments produced by MetalS<sup>2</sup> and the percentage of functional matches (Table 3) suggests that MFS alignments alone are already useful indicators of the functional properties of the metal site. Thus, they can be exploited in cases where the sites are found within different protein folds. For example, the iron MFS in 1dmh, a catechol dioxygenase, was identified by MetalS<sup>2</sup> as being structurally similar to that of 2b5h, a cysteine dioxygenase, even though the protein folds are different (Figure 4). The EC numbers of these two enzymes differ only at the fourth level. The same MFS was also identified as being similar to one in 2fiy, a structure solved within a structural genomics

	CATH	SCOP	Pfam	EC	MetalS <sup>2</sup> Superposition
<b>1dmh_2</b> (Catalytic)	2.60.130.10	b.3.6.1	Dioxygenase_C	1.13.11.1	
<b>2b5h_1</b> (Catalytic)	n/a	b.82.1.19	CDO_I	1.13.11.20	
<b>1dmh_2</b> (Catalytic)	2.60.130.10	b.3.6.1	Dioxygenase_C	1.13.11.1	
<b>2fiy_4</b> (Unknown)	3.90.1670.10	e.59.1.1	FdhE	n/a	

**Figure 4.** An example of functionally relevant MFS alignments. 1dmh is a catechol dioxygenase; 2b5h is a cysteine dioxygenase; 2fiy is a protein of unknown function. Fold classification according to three different databases is reported in the CATH, SCOP, and Pfam columns. The EC column specifies the Enzyme Commission classification, where known. Site names in the first column correspond to those adopted in the MetalPDB database.<sup>21</sup>

initiative for which there is no experimentally validated functional assignment. One can thus hypothesize that the iron site of 2fiy is similarly involved in redox catalysis. Also 1dmh and 2fiy have unrelated folds, preventing functional predictions on the basis of structural domain analysis.

It is also relevant to mention here that even though in this contribution we focused on examples of MFSs derived from metalloproteins, MetalS<sup>2</sup> can handle also sites where some or all of the ligands are provided by nucleic acids. As an example, Supplementary Figure S3 shows the structural alignment of two sites containing respectively one Mn<sup>2+</sup> ion in an octahedral coordination environment that includes three protein ligands, one DNA ligand and two water molecules, and one Zn<sup>2+</sup> ion in a tetrahedral coordination environment that includes three protein ligands and one DNA ligand.

## CONCLUDING REMARKS

In this paper we developed a new software tool, which we called MetalS<sup>2</sup>, for the comparison of two metal-binding biological macromolecules of known 3D structure. To facilitate its use and make it readily available to the scientific community, MetalS<sup>2</sup> is available both as a stand-alone program and a Web tool (<http://metalweb.cerm.unifi.it/tools/metals2/>) within our MetalPDB platform.<sup>21</sup> MetalS<sup>2</sup>, by design, does not take into consideration the entire structure. Instead, it focuses on the MFSs contained in the structures. Each MFS is an ensemble of atoms built around and incorporating a metal site in a metal-binding macromolecule. As such, it contains all the structural information on the metal site itself and its surroundings while discarding all the information related to higher-level structural features, such as the overall protein fold. In this way, each MFS embeds the major part of the structural determinants of the functional properties of the metal site.<sup>18</sup> At the same time, this approach prevents possible biases in the structural comparison due to the larger (in terms of number of atoms) macromolecular chains. We believe that the MetalS<sup>2</sup> strategy supports well one of the intellectual attitudes of bioinorganic chemists dealing with 3D structural data, i.e. understanding how the macromolecular environment tunes the metal site properties and, conversely, how the presence of the metal site defines the

functional properties of the system. Of course, MetalS<sup>2</sup> is meant to complement and not replace the large variety of available tools for the comparison of whole 3D structures, as the latter kind of comparison will provide insight that is exquisitely complementary to that of MetalS<sup>2</sup>. For systems having high structural similarity, such as pairs of homologous proteins, the two approaches will likely provide essentially the same information.

To provide an indication of a possible threshold to identify high-quality structural alignments, we relied on a benchmark generated in a semiautomated manner, which contained proteins binding nonheme iron ions and zinc ions. We could identify a safe threshold of 2.25 for the MetalS<sup>2</sup> total score, below which alignments are essentially always of high quality and a range between 2.25 and 2.75 where the superpositions are good in the majority of cases. The score of MetalS<sup>2</sup> allows users to more easily identify good alignments than with other programs, whose thresholds and scoring functions have been optimized for application to entire protein structures. In addition, there were several MFS pairs for which MetalS<sup>2</sup> generated high quality alignments, whereas other programs did not perform satisfactorily. Conversely, in a few cases MetalS<sup>2</sup> could not reproduce the good alignments provided by other tools. This happened typically when some displacement of the metal cofactors was needed. Globally, at the safe threshold of 2.25, the balance was in favor of MetalS<sup>2</sup>, which could identify a much larger number of structurally related MFS pairs in our example data sets (Table 2). Regarding possible usage scenarios, MetalS<sup>2</sup> can be exploited to define the variability of MFS structure within a superfamily of metalloproteins or to analyze structural changes upon ligand/inhibitor binding. Additionally, MetalS<sup>2</sup> can constitute the basis for innovative MFS classification essentially by conveniently replacing FAST within approaches similar to those we previously applied.<sup>31</sup>

## ASSOCIATED CONTENT

### Supporting Information

Tables S1 (Scores of all-versus-all alignments calculated by MAMMOTH, FAST, TM-align, and MetalS<sup>2</sup> on the Fe-data

set), S2 (Scores of all-versus-all alignments calculated by MAMMOTH, FAST, TM-align, and Metals<sup>2</sup> on the Zn-data set), and S3 (Wilcoxon rank sum test for aligned MFS pairs). **Figures S1** (Flowchart of Metals<sup>2</sup>), **S2** (Superposition of different LEPs for a given MFS pair), and **S3** (Alignment of the Zn1138 site of 2xqc and of the Mn1133 site of 2vju). This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*Phone: +39 055 4574267. Fax: +39 055 4574253. E-mail: andreini@cerm.unifi.it. Corresponding author address: Magnetic Resonance Center, University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was supported by MIUR (Ministero Italiano dell'Università e della Ricerca), through the FIRB projects RBFR08WGXT and RBRN07BMCT and the PRIN project 2009FAKHZT, and by the European Commission through the BioMedBridges project (grant no 284209). We gratefully acknowledge the technical help of Enrico Morelli.

## REFERENCES

- (1) Frausto da Silva, J. J. R.; Williams, R. J. P. *The biological chemistry of the elements: the inorganic chemistry of life*; Oxford University Press: New York, 2001.
- (2) Bertini, I.; Sigel, A.; Sigel, H. *Handbook on Metalloproteins*; Marcel Dekker: New York, 2001; pp 1–1800.
- (3) Bertini, I.; Gray, H. B.; Stiefel, E. I.; Valentine, J. S. *Biological Inorganic Chemistry*; University Science Books: Sausalito, CA, 2006.
- (4) Bertini, I.; Rosato, A. Bioinorganic chemistry in the post-genomic era. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3601–3604.
- (5) Miller, A. F. Redox tuning over almost 1 V in a structurally conserved active site: Lessons from Fe-containing superoxide dismutase. *Acc. Chem. Res.* **2008**, *41*, 501–510.
- (6) Hasnain, S. S.; Hodgson, K. O. Structure of metal centres in proteins at subatomic resolution. *J. Synchrotron Radiat.* **1999**, *6*, 852–864.
- (7) Cotelesage, J. J. H.; Pushie, M. J.; Grochulski, P.; Pickering, I. J.; George, G. N. Metalloprotein active site structure determination: Synergy between X-ray absorption spectroscopy and X-ray crystallography. *J. Inorg. Biochem.* **2012**, *115*, 127–137.
- (8) Sarangi, R. X-ray absorption near-edge spectroscopy in bioinorganic chemistry: Application to M-O-2 systems. *Coord. Chem. Rev.* **2013**, *257*, 459–472.
- (9) Hsin, K.; Sheng, Y.; Harding, M. M.; Taylor, P.; Walkinshaw, M. D. MESPEUS: a database of the geometry of metal sites in proteins. *J. Appl. Crystallogr.* **2008**, *41*, 963–968.
- (10) Schnabl, J.; Suter, P.; Sigel, R. K. O. MINAS—a database of Metal Ions in Nucleic AcidS. *Nucleic Acids Res.* **2012**, *40*, D434–D438.
- (11) Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlc, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukich, B.; Zardecki, C.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.* **2011**, *39*, D392–D401.
- (12) Andreini, C.; Bertini, I.; Rosato, A. Metalloproteomes: a bioinformatic approach. *Acc. Chem. Res.* **2009**, *42*, 1471–1479.
- (13) Andreini, C.; Bertini, I.; Rosato, A. A hint to search for metalloproteins in gene banks. *Bioinformatics* **2004**, *20*, 1373–1380.
- (14) Shu, N.; Zhou, T.; Hovmöller, S. Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* **2008**, *24*, 775–782.
- (15) Karlin, S.; Zhu, Z. Y.; Karlin, K. D. The extended environment of mononuclear metal centers in protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 14225–14230.
- (16) Dudev, T.; Lin, Y. L.; Dudev, M.; Lim, C. First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. *J. Am. Chem. Soc.* **2003**, *125*, 3168–3180.
- (17) Dudev, T.; Lim, C. Metal binding affinity and selectivity in metalloproteins: insights from computational studies. *Annu. Rev. Biophys.* **2008**, *37*, 97–116.
- (18) Andreini, C.; Bertini, I.; Cavallaro, G. Minimal functional sites allow a classification of zinc sites in proteins. *PLoS One* **2011**, *10*, e26325.
- (19) Banci, L.; Bertini, I.; Calderone, V.; Della Malva, N.; Felli, I. C.; Neri, S.; Pavelkova, A.; Rosato, A. Copper(I)-mediated protein-protein interactions result from suboptimal interaction surfaces. *Biochem. J.* **2009**, *422*, 37–42.
- (20) Bertini, I.; Fragai, M.; Luchinat, C.; Melikian, M.; Venturi, C. Characterization of the MMP-12-elastin adduct. *Chem.—Eur. J.* **2009**, *15*, 7842–7845.
- (21) Andreini, C.; Cavallaro, G.; Lorenzini, S.; Rosato, A. MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **2013**, *41*, D312–D319.
- (22) Lathrop, R. H. The protein threading problem with sequence amino-acid interaction preferences is Np-complete. *Protein Eng.* **1994**, *7*, 1059–1068.
- (23) Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr.* **1989**, *A45*, 208–210.
- (24) Sippl, M. J. Recognition of errors in the three-dimensional structures. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 355–362.
- (25) Fufezan, C.; Specht, M. p3d—Python module for structural bioinformatics. *BMC Bioinf.* **2009**, *10*, 258.
- (26) Zhu, J.; Weng, Z. FAST: a novel protein structure alignment algorithm. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 618–627.
- (27) Ortiz, A. R.; Strauss, C. E.; Olmea, O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* **2002**, *11*, 2606–2621.
- (28) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309.
- (29) Slater, A. W.; Castellanos, J. I.; Sippl, M. J.; Melo, F. Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. *Bioinformatics* **2013**, *29*, 47–53.
- (30) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: a multiple structural alignment algorithm. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 559–574.
- (31) Andreini, C.; Bertini, I.; Cavallaro, G.; Najmanovich, R. J.; Thornton, J. M. Structural analysis of metal sites in proteins: non-heme iron sites as a case study. *J. Mol. Biol.* **2009**, *388*, 356–380.
- (32) Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **2008**, *36*, D419–D425.
- (33) Sillitoe, I.; Cuff, A. L.; Dessailly, B. H.; Dawson, N. L.; Furnham, N.; Lee, D.; Lees, J. G.; Lewis, T. E.; Studer, R. A.; Rentzsch, R.; Yeats, C.; Thornton, J. M.; Orengo, C. A. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* **2013**, *41*, D490–D498.
- (34) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291–325.
- (35) Koc, J.; Janežic, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160–1168.
- (36) Yeturu, K.; Chandra, N. PocketAlign a novel algorithm for aligning binding sites in protein structures. *J. Chem. Inf. Model.* **2011**, *51*, 1725–1736.
- (37) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739–747.

- (38) Holm, L.; Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **1995**, *20*, 478–480.
- (39) Sippl, M. J.; Wiederstein, M. A note on difficult structure alignment problems. *Bioinformatics* **2008**, *24*, 426–427.
- (40) Madhusudhan, M. S.; Webb, B. M.; Marti-Renom, M. A.; Eswar, N.; Sali, A. Alignment of multiple protein structures based on sequence and structure features. *Protein Eng., Des. Sel.* **2009**, *22*, 569–574.
- (41) Degtyarenko, K. N. Bioinorganic motifs: towards functional classification of metalloproteins. *Bioinformatics* **2000**, *16*, 851–864.
- (42) Harding, M. M.; Nowicki, M. W.; Walkinshaw, M. D. Metals in protein structures: a review of their principal features. *Crystallogr. Rev.* **2010**, *16*, 247–302.
- (43) Kasampalidis, I. N.; Pitas, I.; Lyroudia, K. Conservation of metal-coordinating residues. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 123–130.
- (44) Hasegawa, H.; Holm, L. Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.* **2009**, *19*, 341–348.