UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA
INTERNATIONAL DOCTORATE IN STRUCTURAL BIOLOGY

CICLO XXVII

COORDINATORE Prof. Claudio Luchinat

NEW COMPUTATIONAL APPROACHES TO THE STUDY OF METALS IN BIOLOGY

Settore Scientifico Disciplinare CHIM/03

Dottorando	**Tutore**
Dott.ssa Yana Valasatava	Dr. Claudia Andreini

Coordinatore
Prof. Claudio Luchinat

Anni 2012/2014

This thesis has been approved by the University of Florence, the University of Frankfurt and the Utrecht University

Acknowledgments

Doing a PhD is a challenging enterprise, and I would like to thank all the people who provided me help and support along this journey. First and foremost, thanks must go to the members of my research group. I was fortunate to work and study in a very friendly, spirited and productive team. I can no longer estimate the number of points in the process of research when their comments led to significant improvements of the project and choosing more meaningful directions. Their influence on my life was profound, and I thank them deeply.

Most importantly, I'm thankful to my supervisor Dr. Claudia Andreini for giving me the chance to work on this project, for sharing her limitless enthusiasm about the fascinating world of metalloproteins, for giving me freedom but never leaving without a piece of a good advice. I am especially grateful to Prof. Antonio Rosato for his patience, fruitful discussions, constructive critique and providing additional supervision along the way. His endless knowledge, experience, and determination were crucial ingredients that made this work possible. I would also like to acknowledge the help of Dr. Gabriele Cavallaro and express my regret at not being able to work together longer. It was an absolute pleasure working with you.

I am no less indebted to my friends and folks, with which I shared my working space over the past three years. Thanks for giving me emotional support and encouragement. Other special mentions go to my friends in the world of Ultimate Frisbee, which has been a source of distraction and counter-balance to academic work.

On a more personal note, special thanks and hugs for keeping me sane during times of intense work to Leanid Krautsevich, without whose boundless support I would have probably tried to run away. Last but not least, I want to acknowledge that, despite of being very far away, my family has always been there for me. I'm thankful to my parents and my brother for all their love, supporting my decisions and wanting the best for me and my life.

Thank you!

Table of Contents

List of Figures

1. Introduction

1.1 Metals in biological systems

The importance of many metals in biological systems is now widely appreciated (1), (2). About 12 different metals, found in living systems, are essential and support fundamental biological functions. These metals are sodium (Na), magnesium (Mg), potassium (K), calcium (Ca), molybdenum (Mo), tungsten (W), manganese (Mn), iron (Fe), cobalt (Co), nickel (Ni), copper (Cu), and zinc (Zn). The essential role of vanadium (V) has been unambiguously demonstrated only for certain organisms, such as ascidians, polychaete worms and *Amanita* mushrooms (3). There are also a number of non-essential metals that however are present in biological systems and affect important physiological processes. Some of them interact with living systems and others have beneficial or pharmacological effects. Figure 1 shows the essential (highlighted in blue and green) and non-essential metals (highlighted in yellow).

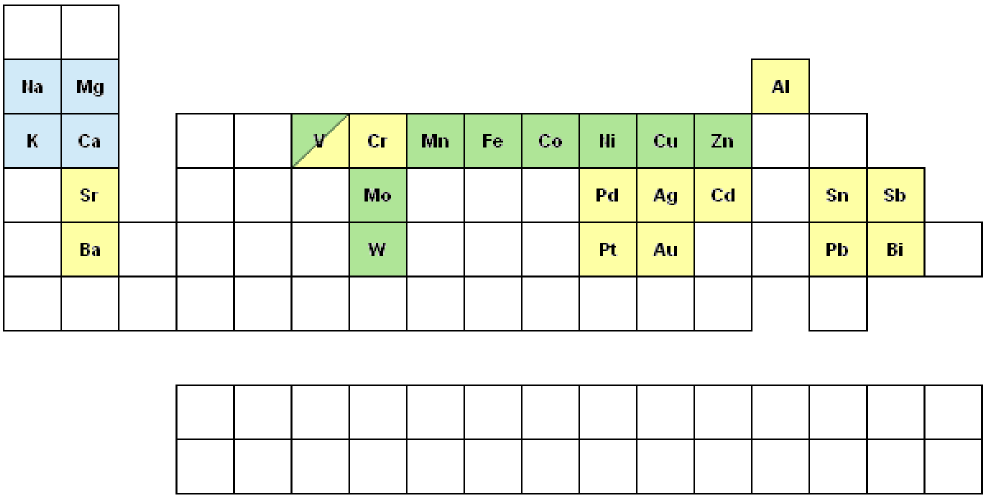


FIGURE 1. A PERIODIC TABLE OF METALS IN BIOLOGICAL SYSTEMS

In course of the evolution essential metal ions have been selected on the basis of their "biological availability" within the environment. This means that ions must be in an easily extractable form and have to be relatively abundant. The cations of the first and second group (Na^+, K^+, Mg^{2+}, Ca^{2+}) are the most abundant metal ions in the environment and in the biological systems (4), (5). They are "bulk" biological elements and are required for a great variety of biochemical functions. Other metals are needed in very small amounts, hence called "trace" metals, despite of abundance in the environment. Iron, for example, is the fourth most abundant element in the Earth's crust but as it can be potentially toxic for the cells its homeostasis is carefully supervised and its distribution is tightly controlled. Although trace metals are present in only small quantities, they have important biological effects and their availability is crucial.

For every organism there is a concentration range within which the requirement of the organism for a given essential metal is met. Within this range, organisms are able to regulate their internal essential metal concentration by means of homeostatic mechanisms without experiencing excessive stress. Below the concentration limit the organism experiences deficiency of the metal and above the limit the metal becomes toxic. In contrast, non-essential metals have a negligible effect on organisms at a below-threshold level but become increasingly toxic as the dose increases above this level (6). This is illustrated in Figure 2.

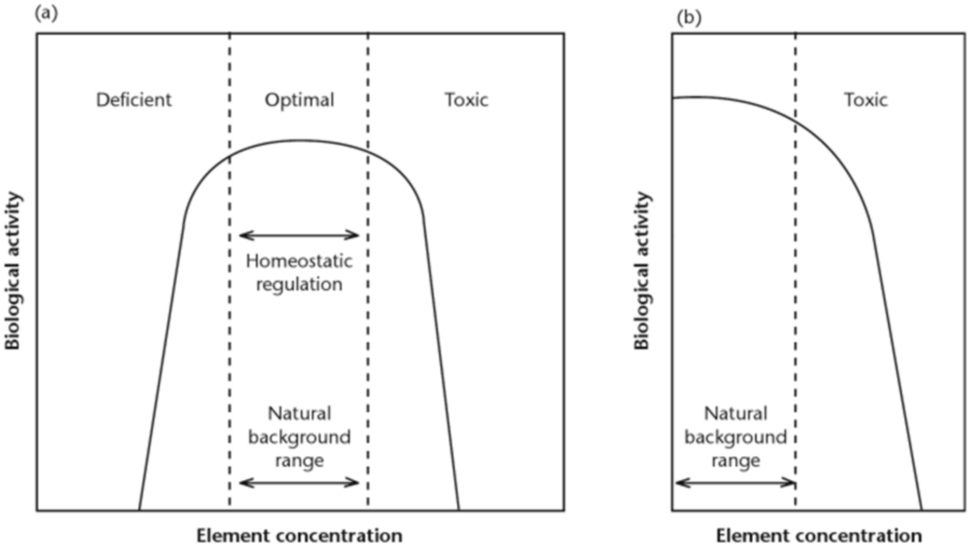


FIGURE 2. THE QUALITATIVE EFFECT OF ELEMENT CONCENTRATION IN LIVING MATTER ON BIOLOGICAL ACTIVITY: (A) ESSENTIAL ELEMENTS; (B) NON-ESSENTAL ELEMENTS

The specific role of essential metal ions within biological systems depends on their chemical and physical properties. For example, iron ions, which readily exchange electrons, may participate in reduction-oxidation reactions whereas zinc ions have a constant oxidation state (even if can it be involved in the catalysis of redox reductions, e.g. if there is an organic electron acceptor/donor). In biological systems metal ions exhibit a great variety of binding partners including biological macromolecules such as proteins, polynucleotides, and carbohydrates. These macromolecules are polymers consisted of basic building block units, some of which contain charged and polar groups that are capable of binding metals.

1.2 The importance of metalloproteins

Proteins that bind metals are called *metalloproteins*. Several facts indicate the importance of the roles covered by metal ions and by the metalloproteins associated with them. First, it is estimated that metalloproteins constitute more than half of the proteins used by living organisms and approximately 40% of enzymes with known structure bind metal ions in their active site and use them to catalyze reactions (7). Second, metalloproteins participate extensively in many essential biochemical processes, including respiration, nitrogen fixation and photosynthesis. Third, it is now generally accepted that deregulation of metal ions amount and usage in cells is associated with important diseases such as cancer and neurodegenerative disorders (8). Therefore, it is important to better understand the roles of metal ions in organisms and support the study of metalloproteins.

1.3 Roles of metals in metalloproteins

Metals contribute to biochemical and physiological properties of metalloproteins. This contribution depends on the interaction between metals and metalloproteins, the strength of which varies from very loose to very tight. Metals can be reversibly bound to metalloproteins, e.g. in case of metal transport, or can be firmly incorporated in a specific location. The classification of metals with regard to their function (9), (10) highlights the following roles that metals can play in metalloproteins:

i. *Structural* metal ions stabilize the structure of folded proteins or help to create a particular physiologically active conformation of the protein. For example, zinc functions as structural element in zinc finger domains. Structural metal ions can also serve as a cross-linking agent binding together polymer chains, different parts of the same chain, or formation of protein-protein interface.

ii. *Catalytic* metal ions are located in sites at which enzyme catalysis occurs, e.g. copper ions in superoxide dismutase or cytochrome c oxidase, iron ions in mono- or di-oxygenases, or iron and molybdenum ions in nitrogenase.

iii. Metal ions can play a *regulation* role in many various cell processes being first, second, or third messengers (11) or act as triggers for protein activity. For example, regulation of transcription is coupled with numerous intracellular signaling processes often mediated by second messengers, like calcium, which is one of the most versatile second messengers. In addition, metals may induce conformation changes in enzymes or in other proteins which may themselves enhance or inhibit enzyme activity (12).

iv. *Transport* of electrons or small molecules can occur with the help of metal ions. Transition metals that exist in multiple oxidation states serve as electron carriers – that is, iron ions in cytochromes or in iron–sulfur clusters or copper ions in blue copper proteins. Another important role is oxygen transport – that is, iron ions in hemoglobin or copper ions in hemocyanin.

1.4 How proteins bind metal ions

Metalloproteins can bind metals as individual ions or within metal-containing cofactors. Metalloproteins can bind more than one metal ion, not necessarily of the same nature. In such cases, the metals may either be distant from each other in space and can reasonably be regarded as independent or be assembled into polynuclear sites where the ions are close in space and often coordinated by bridging ligands. Metal-containing cofactors can be extremely diverse in their chemical complexity, ranging from organic ligands binding a single metal ion, such as porphyrins, to highly elaborate polymetallic clusters, such as the FeMoCo cofactors of nitrogenases. Figure 3 shows several examples of the metal binding centers.

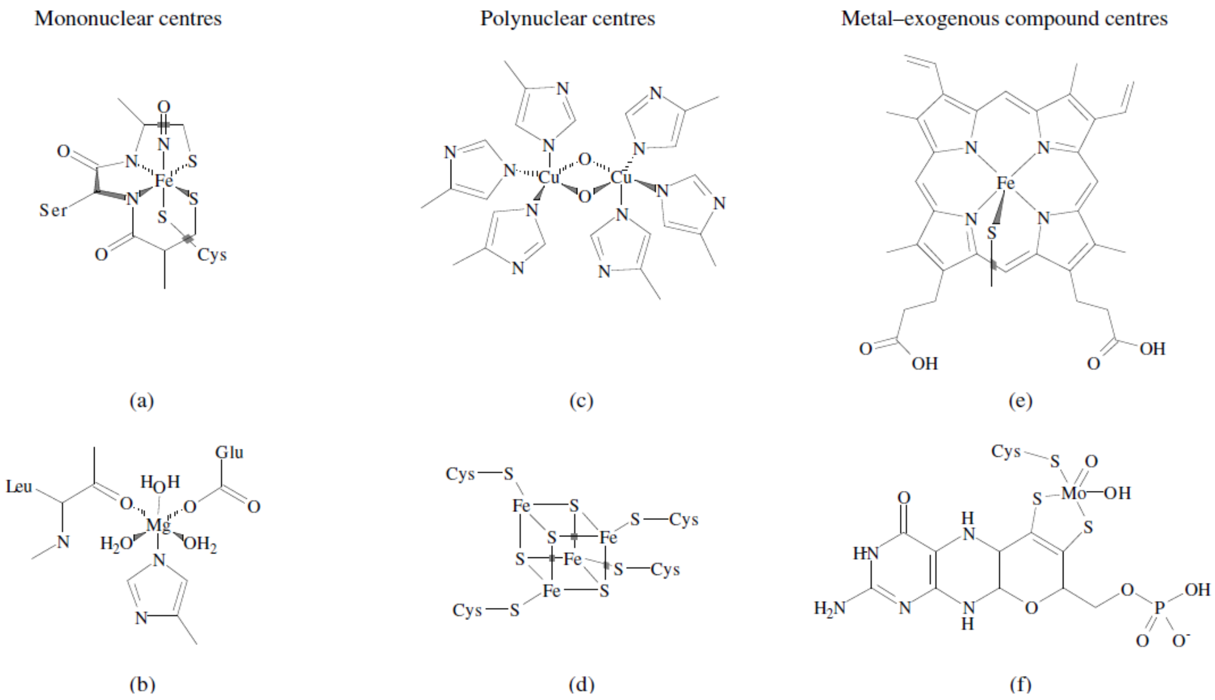


FIGURE 3. EXAMPLES OF METAL BINDING CENTRES

In Figure 3 (a) mononuclear iron center in photosensitive nitrile hydratase; (b) mononuclear magnesium center in Ni-Fe hydrogenase; (c) dinuclear copper center in oxyhaemocyanin; (d) polynuclear iron-sulfur center; (e) haem iron coordination in haem-thiolate proteins; (f) molybdenum center in sulphite oxidase.

Metals are bound to proteins via coordination bonds. Atoms directly involved in metal coordination are termed *donor atoms*. Donor atoms can be provided by the protein (*endogenous* ligands) or can be donated by *exogenous* ligands not derived from proteins, which range from organic compounds as oligopeptides, small organic molecules to small inorganic entities like anions, water molecules, and other convenient ions in the physiological environment (13). A ligand with one donor atom is termed *monodentate*. A *polydentate* ligand is attached to a central metal ion by bonds from two or more donor atoms. Proteins can be regarded as polydentate ligands, but it is often easier to think of the amino acid residues as separate ligands. The ligands surrounding the metals are also collectively called the *first coordination sphere*. The ensemble of atoms comprising the metal ion (complex or cluster of metal ions) and its ligands defines the *metal-binding site*.

Figure 4 illustrates the structure of oxalate oxidase (2ET1) in the presence of glycolate molecule bound to the active site manganese ion. The magnesium ion is shown as a red sphere. Endogenous ligands are His 88, His 90, Glu 95, and His 137 shown as blue sticks. Light blue color highlights exogenous ligands: water and glycolate molecules.

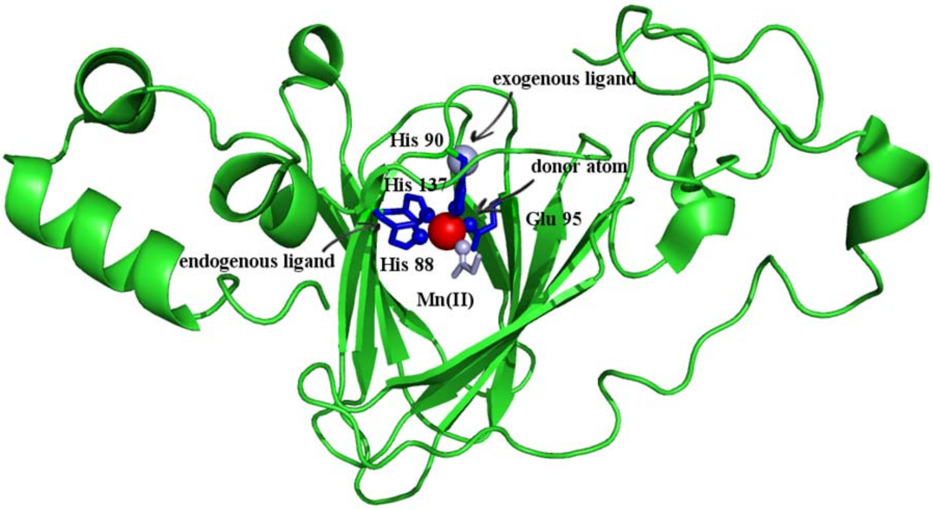


FIGURE 4. AN EXAMPLE OF A PROTEIN STRUCTURE CONTAINING METAL-BINGING SITE

Donor atoms can belong to the protein backbone or side chains/bases. Typical protein donor atoms are the oxygen, nitrogen, and sulfur. Backbone carbonyl groups can bind metal via oxygen. An analysis of the first coordination spheres of the available structures of metalloproteins has shown that about 65% of the various types of amino acid side chains are potential metal-binding groups (14). The most common metal-binding amino acid side chains in proteins are: the carboxylate groups of aspartic acid and glutamic acid, which have negatively charged oxygen atoms available for coordination; the imidazole ring of histidine, which can coordinate via its nitrogen atoms; the thiol group of cysteine, which can form a negatively-charged thiolate whose sulfur atom is available for coordination. Less frequently observed residues are tryptophan, which can bind via the indole group nitrogen, methionine, which can bind via the sulfur atom of its thioether group, and serine, threonine, and tyrosine, which can use the oxygen atom of their hydroxyl group. Asparagine and glutamine amide groups potentially coordinate metals via oxygen atom as well as lysine and arginine amino groups may interact with a metal ion via nitrogen atoms but do it relatively rarely. In rare cases, also the protein N- and C-termini can provide donor atoms to metal ions.

The utilization of protein residues at the binding sites relates to a chemical property termed as “hardness” (”softness”). The word “hard” has been introduced to indicate a low polarizability so that the electron cloud is difficult to deform. By contrast “soft” means high polarizability so that the electron cloud is readily deformed (15). The metals of the first and second group (e.g. Ca and Mg) are “hard” and interact through electrostatic forces mainly with main-chain and side-chain oxygen atoms. On the other hand, most of the transition metals relevant to biological systems (e.g., Fe, Zn, Cu, Ni, Co) are “soft” or have intermediate hard-soft properties. Soft metal ions form stronger complexes and ligate mainly with nitrogen and sulfur atoms of amino acids side chains.

1.5 How the metal environment tunes metal properties and functions

The biological function of metals is related to the chemical properties of the metal and its environment, in the first place the metal-ligand interaction. The important aspect of the metal-ligand interaction is that once a ligand bound to a metal the reactivity of the ligand is modified. This happens due to the perturbation of the ligand-centered energy by the metal proximity (16). Ligand polarization effects resulting from the Lewis acidity of the metal increase the susceptibility of the ligand to nucleophilic attack. Conversely, binding of a ligand to a metal may enhance its susceptibility to electrophilic attack or its tendency to undergo solvolysis reactions. Metal ions with partially filled *d*-orbitals may provide a convenient "bridge" for electron transfer between the ligands. Metal ions can greatly affect the conformational stability of proteins, and many proteins undergo conformational changes upon formation of the complex with metal ions. The opposite is also true as the energy levels, associated with a metal ion, are sensitive functions of its local chemical environment. The ligands often contribute to the stereochemistry required for reaction. There is an increasing number of examples which show clearly that various ligands can modulate the catalytic properties of metal ions (17). These factors contribute to a wide range in the functional properties of the metal-ligand complexes.

The metal ion environment, beyond the first coordination sphere, is also very important in tuning the function of a metal, although a full description is more difficult. For example, there may be interactions of electrostatic charge with dipolar groups in neighboring molecules (18) or of the ligands with hydrogen-bonding groups of the protein main chain and side chain groups (19). Such interactions may result in small changes in coordination geometry around the metal, and hence in coordination energy, they may be critical to the formation and stabilization of the sites or facilitate interaction with substrate or ligands, or in other ways affect the protein function. In support of the role of the local environment, one also finds a strong evidence for the involvement of specific residues, not linked directly to metals, in the mechanisms of catalysis.

Some specific cases have been investigated by structural and kinetic analysis of mutants designed to modify the metal environment. For instance, Ataie et al. (20) have studied the effect of substitution of conserved protein residues outside of the first coordination sphere on the activity of a zinc aminopeptidase. Dudev et al. (21) surveyed second shell ligands found in Mg, Mn, Ca and Zn proteins in the PDB database and carried out energy calculations. The second coordination shell can consist of ions (especially in charged complexes), molecules (especially those that hydrogen bond to ligands in the first coordination sphere) and portions of a ligand backbone. Authors found that the outer shell is apparently designed to stabilize and protect the inner-shell and complement and enhance its properties. Some studies have looked at neighboring groups as far as 3.5Å, but residues within distances up to 4–6Å may well be significant (22).

1.6 Bioinformatics studies on metalloproteins

The detailed study of the structural, physicochemical, and functional properties of metal-binding proteins is an important and actual task. When a protein is synthesized its amino acid sequence is determined by DNA sequence, but the requirement for and position of any metal ions is not a part of this information and must be determined independently. The structures of many metalloproteins are now available as a result of structural genomics efforts, while their function has to be fully characterized. Experimental methods for the identification and characterization of metalloproteins are expensive, time consuming and difficult to automate. Therefore, there is a great demand of computational methods for structural and functional studies of metalloproteins.

In recent years a great deal of bioinformatics research was carried out at CERM (the University of Florence). This research has focused on the development of methods and tools specifically targeted to the study of metalloproteins, and ultimately aimed at facilitating knowledge discovery processes to advance our understanding of the interaction among metal ions and biological macromolecules.

1.6.1 Bioinformatics models of the metal environment: the state of the art

The concept of metal-binding site was extensively used to model the metal environment in metalloproteins and evaluate their functional properties. This model was shown as useful for the bioinformatic analysis of metalloproteins and, in particular, for the prediction of metalloproteins at the whole proteome level (23), (24), (25).

Andreini et al. (25) used sequence information on metalloproteins to determine all known metal-binding signatures. These signatures, termed Metal Binding Patterns (MBP), include the binding residues and their spacing along the sequence. Each MBP is used together with the primary sequence of the corresponding metalloprotein to browse any ensemble of sequences of interest. One of the limitations of this approach is that it requires identification of conserved spacing patterns between binding residues and these spacings are not always conserved. Hence, it is not possible to search for a binding residue that is far away in sequence from other binding residues, since the exact spacing can vary greatly among sequences. A further limitation is that unprecedented sites cannot be predicted.

Many other methods are based on learning machines, such as SVM's (26), (27), (28), (29). For instance, Lin et al. (30) operated on subsequences of proteins, under the assumption that metal binding residues are influenced by the surrounding environment in nature. The amino acid at the center of the fragment is the target amino acid, whereas the others are the "neighbors". The fragment sequence is encoded to a feature vector, which contains information on the occurrence probability of the amino acid, the propensities of the secondary structure, and the metal-binding propensity of the amino acid. The feature vector is fed into a neural-network learning machine. The learning machine decides whether the target amino acid binds metal or not. This process is repeated by shifting each time one position along the protein sequence, resulting in a new fragment. The limitation of this approach is that it predicts metal binding residues rather than metal binding sites. Therefore, it analyses the probability of each putative binding residue individually, instead of taking into consideration the combined context of all residues belonging to one unified site.

Other approaches focused on a description of metal sites in proteins that would consider also the structural information of conformations of ligands in metal-binding site (31). Structural information on metal binding sites comes both from crystallographic and solution studies. It is derived from the coordinate files deposited in the PDB database (32). The metal-binding sites were described as three-dimensional (3D) templates of metal-binding ligands. Then a 3D search used to locate relative conformations of groups of residues in a given structure that closely match a specific metal-binding template.

The crucial limitation of the approaches which use models that include only the metal ligands is that such model may not be sufficiently accurate to reproduce the biochemical function of metal site. As already mentioned, 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 (33), (21), (34). The interactions involving the protein atoms beyond the ligands play a role in tuning the chemical reactivity of metal (e.g. H-bonds, salt-bridges between ligands and neighboring atoms, effects of the three-dimensional conformation upon the local environment of a site) (34). Therefore, the model of metal sites should describe a composition of the metal and the local protein environment, whose properties as a whole are optimized for function.

Consequently, there have been attempts to enhance the description of metal sites adding information about its local environment. One of the early approaches (35) is based on the finding that many metal sites in proteins share a common feature: they are cantered in a shell of hydrophilic ligands, surrounded by a shell of carbon-containing groups. Therefore, it is possible to measure the contrast between groups located at the center of the sphere (more hydrophilic), and groups located at the outer shell (more hydrophobic) within a radius of threshold distance. The contrast function is generally maximal when cantered at or near a metal binding site. However, this algorithm also identified regions of high contrast that were not associated with metal binding, such as charged surface residues and buried, positively-charged residues (36).

In this thesis we follow a novel model proposed by Andreini et al. (37). The model describes metal-binding sites in metal-binding biological macromolecules, including metalloproteins. The model is aimed at increasing the strength of the relationship with functional properties. The model takes into account the surroundings of the metal-binding site and can be thought of as the minimal environment determining metal function, hence dubbed the “Minimal Functional Site” (MFS). An MFS is defined as the ensemble of atoms containing the metal ion or metal containing cofactor, all its ligands and any other atom belonging to a chemical species within 5Å from a ligand. A distance threshold of 5Å appears to be a reasonable compromise between the need of including all residues that interact with metal ligands and the need of describing metal sites only on in terms of their local structure (without exceeding too far from the metal).

MFSs are extracted from coordinate files in the PDB format describing the three-dimensional structures of metal-binding biological macromolecules derived from the PDB database. In PDB structures macromolecules are often deposited in a complex with other biologically relevant molecules and ions such as water, metal ions, nucleic acids, ligands and so on, which can be also described in the PDB format. Availability of special coordinates of such complexes allows capturing the interaction between metal ion(s) and biological macromolecules as well as exogenous ligands at atomic level. The data deposited into PDB is validated against the unified format and so can be processed using automated computational protocols.

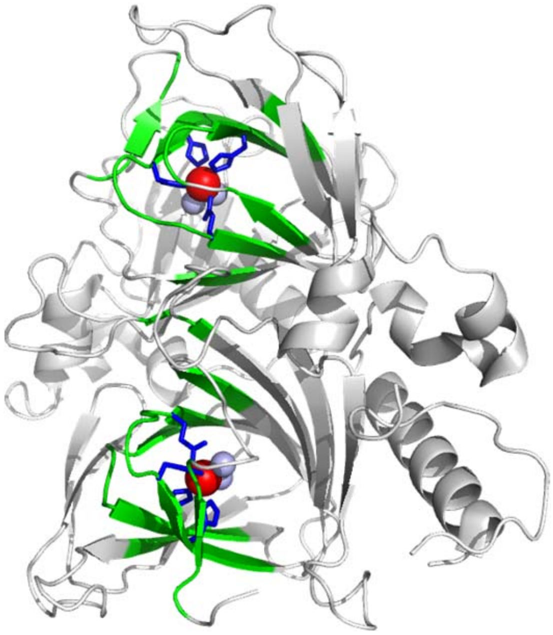


FIGURE 5. AN EXAMPLE OF A STRUCTURE WITH TWO MINIMAL FUNCTIONAL SITES

The MFS describes the local 3D environment around the cofactor containing enough information on the functional properties of the site, independently of the larger context of the protein fold in which it is embedded. The usefulness of the MFS concept outlined above has its chemico-physical foundation in the fact 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 (38) or partner recognition (39).

1.6.2 Bioinformatics resources and metalloproteins: the state of the art

Nowadays hundreds of bioinformatics resources are available, which contain and make easily accessible various types of biological information. There are databases of protein sequences, for example, the Universal Protein Resource (UniProt) (40) that provides a comprehensive and freely accessible central resource of protein sequences and functional annotation (http://www.uniprot.org/). There are databases containing the solved structures of biological macromolecules. The principal resource for many bioinformatics studies is Protein Data Bank (PDB) (http://www.rcsb.org) (32) (managed by the RCSB, Research Collaboratory for Structural Bioinformatics). It contains the structural information resulting from X-ray diffraction structure determinations of protein crystals and from the nuclear magnetic resonance (NMR) structure determinations of proteins in solution. There are databases collecting the available knowledge on specific biological systems, e.g. BRENDA (http://www.brenda-enzymes.info/) that contains enzyme-specific data manually extracted from primary scientific literature and additional data derived from automatic information retrieval methods such as text mining. Resources such as CATH (41) or SCOP (42) are able to capture distant relationships between protein domains through the analysis of their 3D structures. They provide the notion of protein superfamily, which is the ensemble of all the protein domains with the same overall structural features. Proteins that share significant sequence similarity are organized in Pfam database (http://pfam.xfam.org/), which is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (43).

Despite of the swift development of resources and importance of metals in biology, bioinformatics has paid a little attention specifically to metalloproteins. Previously created databases of metalloproteins such as PROMISE (44) and MDB (45) were short-lived and are not currently updated. PROMISE was aimed at providing ample manual annotations for several groups of metalloproteins, but it lacked well-defined classification system and terminology, therefore it could not be easily queried. Its development was abandoned in 2002. MDB had the primary goal of automatically extracting and collecting information on the geometry of metal sites from the known structures of metalloproteins. A limit was that MDB it classified the sites (and could be queried) only with respect to the type of metal and to the coordination number, and did not give any functional annotation. Also its development was abandoned, in 2001.

Our work exploits the MetalPDB database (http://metalweb.cerm.unifi.it) that has been developed by Andreini et al (46). The database contains knowledge on metal-binding sites in biological macromolecules with known structure and is automatically updated monthly. The central objects of MetalPDB are Minimal Functional Sites, introduced in Section 1.6.1. To facilitate analysis, the MFSs in MetalPDB are automatically grouped into clusters of equivalent sites, i.e. sites found in the same position within similar structures and occupied by the same metal ion(s). Developing the MetalPDB database has reduced the existing gap between bioinformatics and systematic analysis of metalloproteins, although many central questions, such as those regarding the relationship between the structure of the metal site and the function of the metal ion, still remain unanswered.

2. Research project

My doctoral project within three years was focused on the development of computational tools and resources for the structural and functional analysis of metal sites in biological macromolecules. The main idea underlying all the approaches, developed in the course of the doctorate, is to make use of the metal-centered Minimal Functional Site model in order to get biological hints on metal-containing macromolecules. MFSs have two advantages: they can be straightforwardly extracted from PDB structures and can be systematically compared via structural alignment.

In the previous studies Andreini et al have demonstrated that proteins that are structurally and evolutionary unrelated, according to CATH (41) and SCOP (47) classifications, may contain similar metal sites (48), (37). Furthermore, it has been shown that MFS support meaningful structure-based functional analysis. For example, for non-heme iron sites at least 17% of the sites found in unrelated proteins were highly similar. In another study, the systematic structural comparison of MFSs of zinc proteins allowed a classification of 77% of a non-redundant ensemble of zinc sites into 10 clusters. Each represented a zinc-binding motif conserved across different protein superfamilies. MFSs can therefore be extensively grouped into MFS folds, i.e. MFSs that share similar structure contained in protein with different shapes. Often, metal ions found in MFSs with the same fold perform the same function, so a structure-based classification is tightly connected to the functional properties of each site.

By the time the aforementioned analysis of zinc sites was performed, the authors did not have a computational tool specifically designed to compare MFSs. The analysis thus required extensive human intervention as existing tools for overall structural alignment often failed to align MFSs. This also hampers the reproducibility of the analysis as well as the application of the same approach to other metals and by other groups. The huge variability and diversity of metal sites indeed warrants the development of specific tools for their analysis.

To address the above bottleneck, we developed the MetalS^2 (Metal Sites Superposition) tool for the structural alignment of MFSs in any pair of metal-binding biological macromolecules. On the example data sets, that were used for an assessment, MetalS^2 unveiled structural similarities that other programs for protein structure comparison did not consistently point out, and overall identified a larger number of structurally similar MFSs. MetalS^2 supports the comparison of MFSs harboring different metals and/or with different nuclearity, and is available both as a web tool at http://metalweb.cerm.unifi.it/tools/metals2 and a stand-alone software tool. The paper describing the tool was published in the Journal of Chemical Information and Modeling in 2013 (Andreini C, Cavallaro G, Rosato A, and Valasatava Y. *MetalS^2: a tool for the structural alignment of Minimal Functional Sites in metal-binding proteins and nucleic acids*). The article is reported in Section 4.1.

The availability of MetalS^2 opened up several possibilities for further investigations. Our first efforts were dedicated to systematic comparison of all the content of the MetalPDB database. This analysis aimed at building a structure-based classification of all metal-binding sites with known structure. Such classification brings together structures of MFSs in distinct folds thereby revealing common structural motifs in structurally unrelated proteins. Conversely, it also highlights differences in the metal sites of proteins that have similar structure. In this regard we developed a computational protocol to systematically compare and classify metal-binding sites on the basis of the structural similarity of their MFSs. This protocol can be applied to analyze all the structures of MFSs. The protocol is based on MetalS^2 by exploiting its ability to quantitatively compare a pair of MFSs, and uses the available organization of the MetalPDB database. In the submitted paper *Hidden relationships between metalloproteins unveiled by structural comparison of their metal sites*, reported in Section 4.3, the usefulness of the analysis exploiting the protocol has been demonstrated, e.g. by showing previously undetected similarities in multi-heme cytochromes.

Another application that we pursued has been the use of MetalPDB as a dataset of MFSs with known function against which an MFS of unknown function (a query site) can be compared systematically. Detected structural similarities can indicate possible functional identity of the query site. Such analysis is useful to get functional hints for metals found in proteins of unknown function (e.g. newly determined structures of metalloproteins). In this regard we developed MetalS^3 (Metal Sites Similarity Search), a database search tool that search for structural similarities within the MetalPDB database. MetalS^3 uses a suitably adapted version of the algorithm implemented in MetalS^2, and can be accessed through a web interface at http://metalweb.cerm.unifi.it/tools/metals3/. It systematically compares the structure of the query metal site to each MFS in MetalPDB, and keeps the best superposition for each MFS. All these superpositions are then ranked according to the MetalS^3 scoring function and presented to the user in tabular form. The user can interact with the output web page to visualize the structural alignment or the sequence alignment derived from it. Options to filter the results are also available. Test calculations showed that the MetalS^3 output correlates well with expectations from protein homology considerations. Furthermore, we provide several usage scenarios that highlight the usefulness of MetalS^3 to obtain mechanistic and functional hints regardless of homology. The paper describing MetalS^3 was published in the Journal of Biological Inorganic Chemistry in 2014 (*MetalS^3, a database-mining tool for the identification of structurally similar metal sites.* Valasatava Y, Rosato A, Cavallaro G, Andreini C.). The article is reported in Section 4.2.

Further analysis involved the integration of diverse data from multiple data sources. It is now widely appreciated that proteins can be organized into superfamilies of structurally related molecules with very similar or radically diverse functions. These, of course, include metalloproteins. At present, there is no systematic investigation of the occurrence and function of metal sites across known protein superfamilies. Indeed, bringing together metal sites and databases of structurally related proteins may provide insights on the structural and functional diversification of proteins e.g. in superfamilies including members both with and without metal sites.

To this end, we planned a strategy to integrate data on protein superfamilies with MFSs. This was implemented as a new resource that currently manages biological data from three sources: CATH database, Metal-MACiE database (49), and the MetalPDB database (46). The CATH database contains structural domains derived from the Protein Data Bank, organized in superfamilies according to their Class, Architecture, Topology, and Homology. Proteins are clustered into evolutionarily-related families if they have high sequence similarity or high structural similarity and some sequence/functional similarity. Metal-MACiE is a resource that contains functional annotations for catalytic metal ions (i.e. about the role that metals or metal-containing cofactors play in the catalytic mechanism of metalloenzymes). Metalloenzymes are an important subclass of metalloproteins wherein the metallic cofactor is essential for the catalytic activity. The mechanisms of functional diversification of metalloenzymes have been extensively analyzed (50), (51), (52). Current resources either provide details on just a particular type of data or advance extensive detailed analysis on a relatively small number of enzymes (53), (54). Metalloenzymes catalyze numerous reactions of physiological importance utilizing a relatively small number of metallic cofactors. Giving an importance to catalytic metal sites in governing the function of enzymes, this work aimed at creating an overview of the differentiation of the functional properties of enzymes in connection with the differentiation in a local structure of their sites. Finally, MetalPDB contains detailed information on three-dimensional structures of metal-binding sites in proteins.

In this frame we analyzed CATH superfamilies that contain catalytic metal sites contained in Metal-MACiE. For every member of such CATH superfamilies we obtained a metal-binding annotation. The annotation takes into account the conservation of metal-binding ligands over the sequence alignment based on the structural alignment of all superfamily members. The structural alignment is based on either the superimposition of a site (if the site is similar and occupies the same position as catalytic site) or the superimposition of the entire protein domain. The procedure of obtaining site-based superimpositions inherits their principals from MetalS^3 search. In practice, each superfamily is identified by a catalytic metal site. Then for each metal-binding domain in a given superfamily the number of MFSs is known and MetalS^2 tool is used to find the most similar catalytic sites. This strategy also allows analyzing the proteins that do not contain metals but are structurally related to metal-containing molecules.

To begin to understand in detail how enzymes depend on metals in the differentiation of their function we have initially deployed the strategy outlined above to the contents of the FunTree database (55), a public resource that brings together sequence, structure, phylogenetic, chemical and mechanistic information for structurally defined enzyme superfamilies, instead of using directly CATH. The structural information in FunTree is indeed originating from the CATH database, but is already preorganized to focus on the mechanistic aspects of enzyme families by including only the catalytic domains of enzymes. Metal-binding sites identified in MetalPDB as corresponding to a Metal-MACiE entry were thus mapped on FunTree sequence alignments. FunTree sequence alignments are derived from structural alignments of protein domains featuring the same CATH classification. In practice, each FunTree alignment represents a given CATH superfamily of enzymes. This allowed us to evaluate the conservation of the ligands within each enzyme superfamily. This work is described in more details in a paper that is in preparation at the moment and its draft is included in Section 4.4.

3. Methodological aspects

This section describes in detail some of the methods and algorithms we used for our research and the interpreting the data to obtain results. Section 3.1 describes the protocol to extract MFSs from PDB files. Section 3.2 contains a mathematical framework for structural comparison of MFSs. We combined the proposed method for structural comparison of MFSs with a hierarchical clustering approach to obtain a structure-based classification of MFS in Section 3.3. The analysis of CATH superfamilies required custom development of the procedure to create a multiple sequence alignment based on the structural alignments. The procedure, described in Section 3.4, operates on both the structural alignments of MFSs and protein domains. Finally, Section 3.5 provides the details on implementation of aforementioned approaches in Python.

3.1. How Minimal Function Site is defined

In order to select atoms that comprise an MFS a simple distance-based protocol is applied to PDB files describing metal-containing macromolecules. For each MFS we first identify a metallic cofactor which can be an individual metal ion or a polymetallic complex. Metal ions are assembled in a cluster if they are separated by a distance smaller than 5Å or bridged together by a ligand. This allows identifying polynuclear sites, e.g. an iron-sulfur cluster in ferredoxins will be identified as an individual four-nuclear site. Then ligands are identified as residues having at least one non-hydrogen atom within 2.8Å from any metal ion. The ensemble of metal ions(s) and ligands identifies a metal-binding site. Finally, a metal-binding site is extended by adding all atoms of residues that have at least one non-hydrogen atom within 5Å from any ligand. The latest assemble of metals ions and atoms from ligands and surrounding residues constitute the MFS and described in the PDB format.

Figure 6 represents the process of assembling an MFS around a single metal ion (a red sphere) where ligands are shown in blue color and surrounding residues in green color.

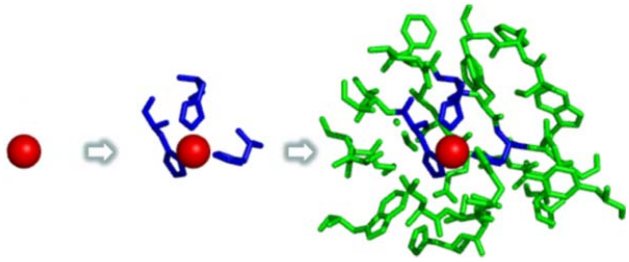


FIGURE 6. STEPS TO ASSEMBLE A MINIMAL FUNCTIONAL SITE

3.2. Structural comparison of Minimal Function Sites

We developed a new algorithm which allows comparing the structures of MFS via structural alignment based on rigid-body superimposition. Structural alignment is an alignment based on comparison of shapes. There are two main issues when dealing with structural alignment: (i) a pairing of atoms must be established. In the general structural alignment problem we have to compare different proteins having different length and potentially very different sequence. We therefore do not know in advance the correspondence between the atoms in the two structures we want to align; (ii) transformation must be found that superimposes the two structures with given pairing. The naive approach to the first problem would be to simply check all possible pairings of atoms resulting in O(n!) time complexity. Even if the number of atoms present in MFSs is smaller than in proteins an exhaustive search is undesirable. Consequently, we decided to rely on heuristics to tackle the determination of the best superimposition of two MFSs.

In our previous work, we defined a MFS as an assembly of residues around a metal ion or a cluster of metal ions. 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 MFSs, the first step is to superpose these centers so that they coincide with the origin of coordinates. In the subsequent alignment, rather than looking through each possible combination of residues, the method generates a pool of possible candidate alignments (see Section 3.2.2). To accomplish this task the structures of MFSs are structurally superimposed in a number of relative orientations. The method used for structural superimposition is described in more detail in Section 3.2.3. In making an alignment, a one-to-one correspondence is set up between the sequences of MFSs. The residues are matched on a basis of special proximity of atoms in a reduced representation described in Section 3.2.1. The atom pair coordinates are used to establish one-to-one correspondences between the residues of the two sites. The correspondence is established on the basis of the distance between atoms. The method used to match atoms is described in Section 3.2.4. The initial alignments are then ranked using the specially designed scoring function (see Section 3.2.5). Finally, the score of these structural alignments is optimized by allowing the geometric centers and the ligands to displace with respect to one another (the method for the final fitting is described in Section 3.2.6). During the final fitting, for each pose the atom matching procedure is repeated to update the atom correspondences and the similarity score is recalculated. Only the best scored alignment is retained in the end.

3.2.1. Minimal Functional Site reduced representation

A special interest concerns the modeling of an MFS as a first approximation of its structure that is used by the algorithm for structural alignment which is further described. When aligning structures, we are interested only in a subset of the atoms of the given MFS structure. The model reduces an MFS to a set of points that corresponds to a metal ion (or a geometric center metals ions), donor atoms and atoms that represent each residue, in particular its backbone core and a side chain group, except for glycine that contains a hydrogen atom as its side chain.

Donor atoms, the atoms that are directly bonded to a metal ion, are identified by a procedure described in Section 3.1. To identify the position in space of each amino acidic residue, we use the coordinates of two atoms: one backbone atom and one atom from a side chain. For amino acids residues, we use the Cα and Cβ atoms (only the Cα atom for glycine). For nucleic acids residues, the pair is formed by the C1 atom of the sugar and the N1 atom for pyrimidine bases or the N9 atom for purines was used.

Figure 7 shows relative placement of representative atoms: metal ion (in red), donor atoms (in blue), atoms from backbone (in green), and side chain atoms (in dark green).

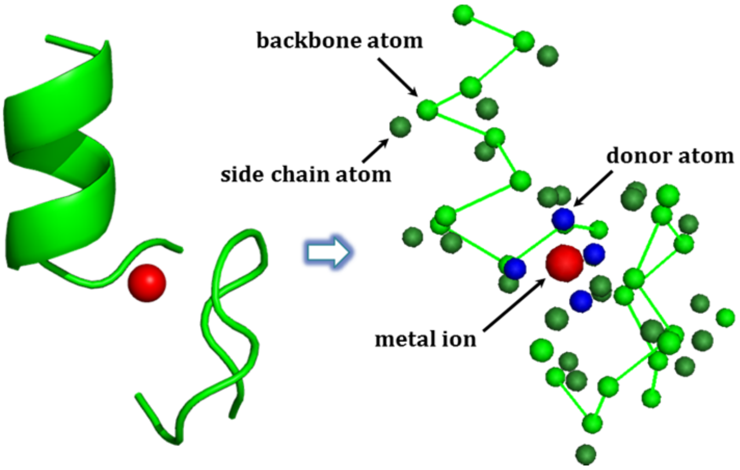


FIGURE 7. A REDUCED REPRESENTATION OF A MINIMAL FUNCTIONAL SITE

Using only the Cα atoms in order to determine the topology of a backbone trace is a relatively common approach. There is indeed extensive demonstration in the scientific literature that the knowledge of the Cα 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 Cβ atoms is useful as the Cα-Cβ bond represents the direction of the side chain with respect to the main chain. Cβ interactions provide additional information on fold energetics (56). By using only Cα and Cβ pairs, the calculation approach is essentially independent of the amino acidic sequence thus facilitating the comparison of highly different sites.

3.2.2. Generating candidate alignments

A candidate alignment is established based on proximity of residues after MFSs are relatively orientated. The position of one MFS is fixed (a query MFS) and the other MFS is respectively transformed (a target MFS). The initial relative orientations (or initial poses) are determined by the comparison of the positions of donor atoms. To accomplish this task, all possible *local elementary patterns* (LEP) are derived from the first coordination sphere for each structure of MFSs and superimposed in all-vs-all manner.

A LEP corresponds to three points in 3D space: one point coincides with a metal (or geometric center of a cluster of metals) and two other points correspond to a pair of donor atoms.

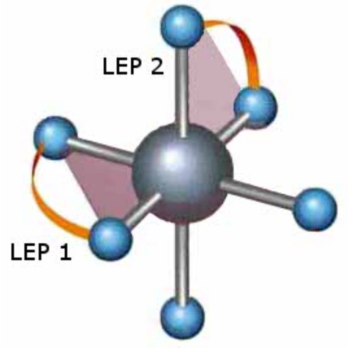


FIGURE 8. LOCAL ELEMENTARY PATTERNS

As shown in Figure 8, each LEP is a triangle unless a site has only one monodentate ligand. A monodentate ligand has only one donor atom used to bind a metal ion and in this case LEPs in both structures are linear segments. Each segment is enclosed by two points in 3D space: 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.

Triangular LEPs are superimposed in a way that donor atoms are rotated around common origin and approach each other. The rotational matrix resulting from each superposition of LEPs is then applied to the whole site to define an initial pose. The initial poses are computed in a different way if LEPs are segments. To ensure alignment of donor atoms we switch from the superposition of triangles to the superposition of segments (LEPs in both sites are considered as segments). Each superposition provides an initial rotation matrix. The number of initial rotational matrices is equal to the maximum number of donor atoms for two sites. Each initial rotation matrix implies rotation in two-dimensional space. In order to take into account all degrees of freedom, a number of additional rotations are produced. Each additional rotation is made around the axis corresponding to the superposed segments. So for each pair of superposed segments we have 17 additional matrices which represent 20° rotations.

3.2.3. Structural superposition

Optimal superposition of aligned sets of points (the correspondence between points is established) can be computed exactly and efficiently via analytical solution. To solve the rigid-body least-squares superposition problem we apply the approach proposed by Kearsley (57) using the mathematical object called quaternions (58). Quaternions are generalizations of complex numbers with direct application to transformations in the three dimensional (3D) space. A quaternion is an element of a 4 dimensional vector-space. It has an *x*, *y*, and *z* component, which represents the axis about which a rotation will occur. It also has a *w* component, which represents the amount of rotation which will occur about this axis. Quaternion is defined as $q=w + xi + yj + zk$, where *i*, *j* and *k* are imaginary numbers. Specifically, a unit quaternion (a quaternion of norm one) is a way to compactly represent 3D rotations.

The solution for the unit quaternion is shown to be the eigenvector of a symmetric 4x4 matrix *N* associated with the most positive eigenvalue. The elements of this matrix are simple combinations of sums of products of corresponding coordinates of the points.

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

Where x_m denotes the component-wise difference (x' - x) (similarly y_m and z_m) and x_p to denote the component-wise sum (x' + x) (similarly y_p and z_p). Here (x', y', z') and (x, y, z) are the coordinates of corresponding points in aligned sets.

Diagonalizing this matrix yields four eigenvalues and (corresponding) eigenvectors. The eigenvector corresponding to the smallest eigenvalue is a unit quaternion that corresponds to the rotation producing the least-squares error. For the unit quaternion *(w, x, y, z)* the corresponding rotation matrix *M* is defined as follows:

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

The computational effort that takes to solve the rigid-body superposition problem using Kearsley's quaternion approach is dominated by the computation of the *N* where each of 10 distinct terms in the matrix requires O(n) effort. The diagonalization of *N* is independent of n and shows a rapid convergence with numerical methods such as Jacobi's diagonalization algorithm (59).

Quaternions are also used to make a rotation around a single axis. The formula for quaternion q in terms of an axis angle is:

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

Where x, y, and z represent the axis vector, about which the rotation occurs, and α is an angle that defines the amplitude of the rotation about the axis.

3.2.4. Atoms matching

When finding corresponding points, there is the possibility to search the closest points using exhaustive (brute force) search. This method is very complex, because all points of one set (*i*) must be compared to all points of another (*j*). The complexity is valued as O(n_i × n_j) and so this approach is time-consuming. A high increase of the speed is achieved by using k-d trees and closest point caching. A k-d tree, or k-dimensional tree, is a data structure used for organizing some number of points in a space with k dimensions (in this work points refer to atoms and are stored in the Cartesian plane, in three-dimensional space) (60). Each level of k-d tree partitions the space into two parts, the partitioning is done along one dimension of the node at the top level of the tree, along another dimension in nodes at the next level, and so on, iterating through the dimensions. The partitioning proceeds in such a way that, at each node, approximately one half of the points stored in the subtree fall on one side, and one half fall on the other. The use of a k-d tree search permits excluding big regions in the search space. At every decision in a tree node, one side of the hyper plane can be rejected. K-d trees allow efficiently performing searches like “all points at distance lower than d from p” or “k nearest neighbors of p”. When processing such query, we find points which correspond to p. Approximate time complexity is $O(n_i \times \log(n_j))$ for building a k-d tree and $O(\log(n_{j)})$ for performing a search.

3.2.5. Calculating similarity score

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 developed the scoring function that evaluates three different terms:

1. A **relative coverage term,** which depends 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 Cα and Cβ 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 Cα and Cβ 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 Cα 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

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

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:

$$\frac{\sum_{f=1}^{F} {}^{1}/_{n_f}}{N}$$

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 score, T, the three terms above were linearly combined as follows:

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

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. 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.

3.2.6. Final fitting

At the stage of the final fitting the atom correspondence is already established. The pairs of corresponding atoms, named matched atoms, are used for fitting. Fitting minimizes the RMSD of the coordinates of matched atoms by roto-translating the target MFS. The roto-translation matrix is calculated using Singular Value Decomposition (SVD) (61) of the covariance matrix of the coordinates of the abovementioned pairs.

SVD is a method for writing an arbitrary matrix A as the product of two orthogonal matrices and a diagonal matrix:

$$A = U \cdot S \cdot V^T$$

Where the columns of U are the left singular vectors; S (the same dimensions as A) has singular values and is diagonal; and V^T has rows that are the right singular vectors. Calculating the SVD consists of finding the eigenvalues and eigenvectors of AA^T and A^TA. The eigenvectors of A^TA make up the columns of *V*; the eigenvectors of AA^T make up the columns of U. Also, the singular values in S are square roots of eigenvalues from AA^T or A^TA. The singular values are the diagonal entries of the S matrix and are arranged in descending order. Then the optimal rotation matrix is $R = V \cdot U^T$. The optimal translation T is computed as $T = \bar{d} - R \cdot \bar{m}$, where$\{d_i\}$, $\{m_i\}$ are the points sets to be mapped.

3.3. Hierarchical clustering to group Minimal Function Sites

The procedure of obtaining structure-based classification of Minimal Function Sites uses a hierarchical agglomerative clustering algorithm (62). 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 MetalS^2 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 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. 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. We used the complete and average linkage methods. 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 other words, the distance between clusters C_i and C_j is defined as:

$$d_c(C_i, C_j) = \max_{k \in C_i, l \in C_j} d(k, l)$$

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 distance for the average linkage is defined as

$$d_c\left(C_i, C_j\right) = \frac{1}{\left|C_i\right|\left|C_j\right|} \sum_{k \in C_i, l \in C_j} d(k,l)$$

Where $|C_i|$ and $|C_j|$ and are the numbers of members in the clusters C_i and C_j correspondingly.

In both formulas *k* and *l* refer to members of the clusters C_i and C_j, *d(k,l)* is the distance between the *k*-th member and *l*-th member of C_i and C_j respectively (in practice the global MetalS^2 score between the *k*-th and *l*-th MFSs). The minimum distance $d_c(C_i,C_j)$ among all the intra-cluster distances determines which pair of clusters is merged.

The clustering results are influenced by the linkage type applied. Complete linkage tends to produce clusters that are more compact (tight) with respect to clusters produced by average linkage. When a cut-off value of a similarity measure is applied in order to determine the final partition, the clusters produced by the average linkage method allows some within-cluster distances to exceed the cut-off value whereas the complete linkage method ensures that no within-cluster distance exceeds the cut-off. As a result, the complete linkage approach produces a higher number of more robust clusters while with average linkage the number of clusters is lower but within-cluster variability is higher. One of the weaknesses of the complete linkage method is its sensitivity to outliers, i.e. members that do not fit well into the global structure of the cluster. Such sensitivity may prevent the identification of even intuitive clusters, as outliers may pull similar members into different groups.

3.4. Multiple sequence alignment based on structural superimposition

Here we assume that a reasonable a multiple structural alignment is already known and provide a basis for computing a multiple sequence alignment (MSA). In a multiple sequence alignment, residues among a set of structures are aligned together in columns. A column of aligned residues occupy similar three-dimensional structural positions. Constructing MSA requires computing the pairwise alignments between all sequences and constructing an all-to-all matrix describing the similarity between each pair of sequences (the distance matrix). We compute pairwise alignments on a basis of predefined superimposition of structures. Then a multiple sequence alignment is built by “merging” these pairwise alignments. The algorithm iteratively proceeds through the distance matrix selecting a pairwise alignment having the best similarity score at each step.

The selected pairwise alignment can be (i) assigned to a new chunk, if none of the sequences from pairwise alignment has been assigned to any existing chunk; (ii) added to an existing chunk, if one of the sequences from pairwise alignment has been assigned to an existing chunk; (iii) used to merge two chunks together, if both of the sequences have been assigned to different chunks.

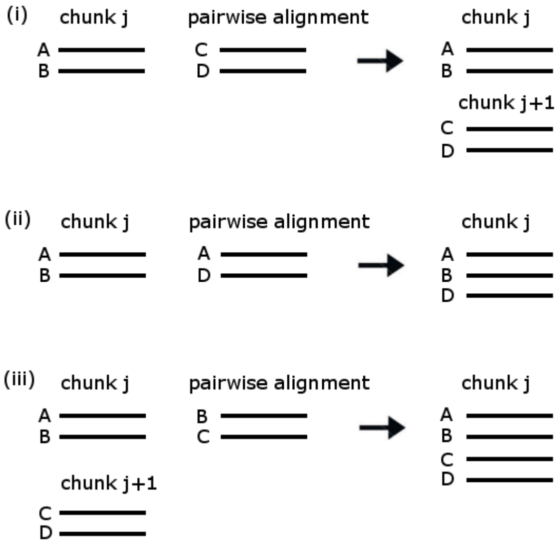


FIGURE 9. SCHEME OF MERGING SEQUENCE ALIGNMENTS

The iterations are performed until all the chunks are merged to encompass all sequences. Lastly, the pairwise comparisons, which were not used at the previous stage, are used to refine the MSA.

3.5. Implementation

3.5.1. Programming language

All the back-end scripts are implemented in Python 2.6 (http://www.python.org/) on a Linux platform. The reasons for choosing this language were:

- The availability of p3d (63), an object oriented Python module for structural bioinformatics. In particular, the Protein class with a set of methods greatly simplifies handling PDB structures.
- Multi-platform: 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.

By using the Python language, we could also exploit the following resources: SciPy 0.7.2, a library of scientific and numerical routines; NumPy 1.4.1, a language extension that adds support for large and fast, multi-dimensional arrays and matrices.

The front-end was implemented using mako, a template library written in Python included by default with the Pylons web application framework, JavaScript, and CSS.

3.5.2. Running environment

MetalS^2 application is currently hosted on an 8-CPU (Intel(R) Xeon(R) CPU E5506 @ 2.13GHz) server.

MetalS^3 application is currently hosted on a 24-CPU (AMD Opteron^{TM} 6234) server.

The running time of the program comparing 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 size of the two structures.

3.5.3. Input specification

MetalS^2 and MetalS^3 support input in PDB format. PDB format consists of lines of information in a text file. A PDB file generally contains several different types of records but only ATOM and HETATM records should mandatory be present.

Example of PDB format is given below:

```
ATOM   1058  N   ARG A 141      -6.466  12.036 -10.348  7.00 19.11           N
ATOM   1059  CA  ARG A 141      -7.922  12.248 -10.253  6.00 26.80           C
ATOM   1060  C   ARG A 141      -8.119  13.499  -9.393  6.00 28.93           C
ATOM   1061  O   ARG A 141      -7.112  13.967  -8.853  8.00 28.68           O
ATOM   1062  CB  ARG A 141      -8.639  11.005  -9.687  6.00 24.11           C
ATOM   1063  CG  ARG A 141      -8.153  10.551  -8.308  6.00 19.20           C
ATOM   1064  CD  ARG A 141      -8.914   9.319  -7.796  6.00 21.53           C
ATOM   1065  NE  ARG A 141      -8.517   9.076  -6.403  7.00 20.93           N
ATOM   1066  CZ  ARG A 141      -9.142   8.234  -5.593  6.00 23.56           C
ATOM   1067  NH1 ARG A 141     -10.150   7.487  -6.019  7.00 19.04           N
ATOM   1068  NH2 ARG A 141      -8.725   8.129  -4.343  7.00 25.11           N
ATOM   1069  OXT ARG A 141      -9.233  14.024  -9.296  8.00 40.35           O
TER
HETATM 1071 FE   HEM A   1       8.128   7.371 -15.022 24.00 16.74          FE
HETATM 1072  CHA HEM A   1       8.617   7.879 -18.361  6.00 17.74           C
HETATM 1073  CHB HEM A   1      10.356  10.005 -14.319  6.00 18.92           C
HETATM 1074  CHC HEM A   1       8.307   6.456 -11.669  6.00 11.00           C
HETATM 1075  CHD HEM A   1       6.928   4.145 -15.725  6.00 13.25           C
```

The detailed description of a PDB format can be found in Protein Data Bank online documentation: http://www.wwpdb.org/docs.html.

3.5.4. User interface

MetalS^2 standalone application has a simple command line interface working as following:

```
$python metalS2.py [--qp/--qs file] [--tp/--ts file] [--qm number] [--tm number] [-d
distance] [output directory]
```

It is mandatory to specify the input structures to start the alignment process. The following options allow specifying the types of input files:

```
--qp    specifies a PDB structure containing a query metal site
--tp    specifies a PDB structure containing a target metal site
--qs    specifies a file containing a query metal-binding site alone
--ts    specifies a file containing a target metal-binding site alone
```

If the input is a PDB file with a number of sites and a metal atom of interest is known, it can explicitly specified by passing a residue sequence number of metal atom followed by the following options:

```
--qm    specifies a metal atom of interest in a query structure
--tm    specifies a metal atom of interest in a target structure
```

All metals are considered by default.

The user can adjust a distance cutoff value for atoms alignment procedure by setting the following option:

```
-d      sets a distance cutoff value, in Angstroms, that allows controlling the upper
bound of the area where two atoms may be considered as aligned
```

The cutoff value can be any non-negative floating point number. The default value is 2.0 Å. A value of 0 prevents the program from running at all.

The user can set the location where she wants to store the results of calculations by adding a relative output path:

```
/relative/path/to/the/output/directory
```

If the output directory is omitted the results will be stored in a directory where the script is current running.

The following options give the user a summary of the usage and available options.

```
-h   --help      prints a brief reminder of command line usage and all available options
-u   --usage     prints a usage summary
```

Graphical User Interfaces of MetalS^2 and MetalS^3 are presented in Sections 0 and 4.1 respectively.

4. Results

The published, submitted papers and papers that are in preparation are listed in the following sections in a chronological order.

4.1. MetalS^2: a tool for the structural alignment of Minimal Functional Sites in metal-binding proteins and nucleic acids

Claudia Andreini[1,2,], Gabriele Cavallaro[1], Antonio Rosato[1,2] and Yana Valasatava[1]*

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

J Chem Inf Model. 2013 Nov 25;53(11):3064-75. Epub 2013 Oct 28.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

MetalS^2: A Tool for the Structural Alignment of Minimal Functional Sites in Metal-Binding Proteins and Nucleic Acids

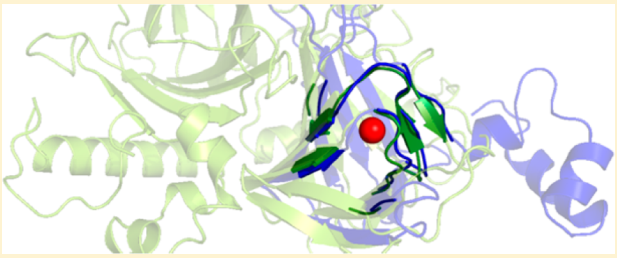
Claudia Andreini,*[†,‡] Gabriele Cavallaro,[†] Antonio Rosato,[†,‡] and Yana Valasatava[†]

[†]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Florence, Italy

[‡]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^2, 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^2 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^2 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.[1−3] 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.[4] 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.[5]

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.[6−8] Databases reporting on the geometric properties of metal-binding sites in proteins[9] or nucleic acids[10] are available. They are derived from the coordinate files deposited in the Protein Data Bank[11] (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.[12−14] 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.[15−17] 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).[18] 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.[18] 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[19] or partner recognition.[20] A database of MFSs extracted from the structures deposited in the PDB is available.[21]

Here, we report the development of a software tool, called MetalS^2 (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^2 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^2 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[22] 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^2 program is to perform the superposition between two metal sites using a multistep approach. First, MetalS^2 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^2 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.[18] 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 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) × 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^2 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[23] 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

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

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:

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

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^2 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^2 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

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

where x, y, and z represent the axis vector about which the rotation occurs, and α 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^2 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 Cα and Cβ atoms (for Gly we used only the Cα atom). Using only the Cα 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 Cα 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 Cβ atoms is useful as the Cα-Cβ bond represents the direction of the side chain with respect to the main chain. Cβ interactions provide additional information on fold energetics.[24] By using only Cα and Cβ 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 Cα atom from the first site (query site) we assign a correspondence to the Cα 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 Cα-Cα (or C1−C1) pair are bound to a Cβ (or N1/N9) atom, we also compute the distance between the two Cβ 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 Cα and Cβ 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 Cα and Cβ 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 Cα 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

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

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

$$\frac{\sum_{f=1}^{F} \frac{1}{n_f}}{N} \tag{5}$$

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^2 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

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

where $x_i^A - x_i^B$ is the distance between the i-th atom pair, and $C^*_{\max}$ is the number of matched Cα, Cβ 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^2 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,[25] 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^2, on a statistical basis: FAST,[26] MAMMOTH,[27] and TM-align.[28] These tools were chosen among the relevant programs included in a recent review,[29] because they are able to handle protein fragments despite being designed for the alignment of entire structures. The only exception was the program MUSTANG,[30] 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[29] but was successfully used by some of us in the past for similar applications.[18,31] 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^2 we used two data sets previously analyzed by some of us. The first one (Fe-data set) consists of 86 MFSs containing nonheme iron,[31] whereas the second one (Zn-data set) consists of 367 MFSs containing zinc.[18] 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^2 in the analysis of the above test sets was evaluated using the following parameters

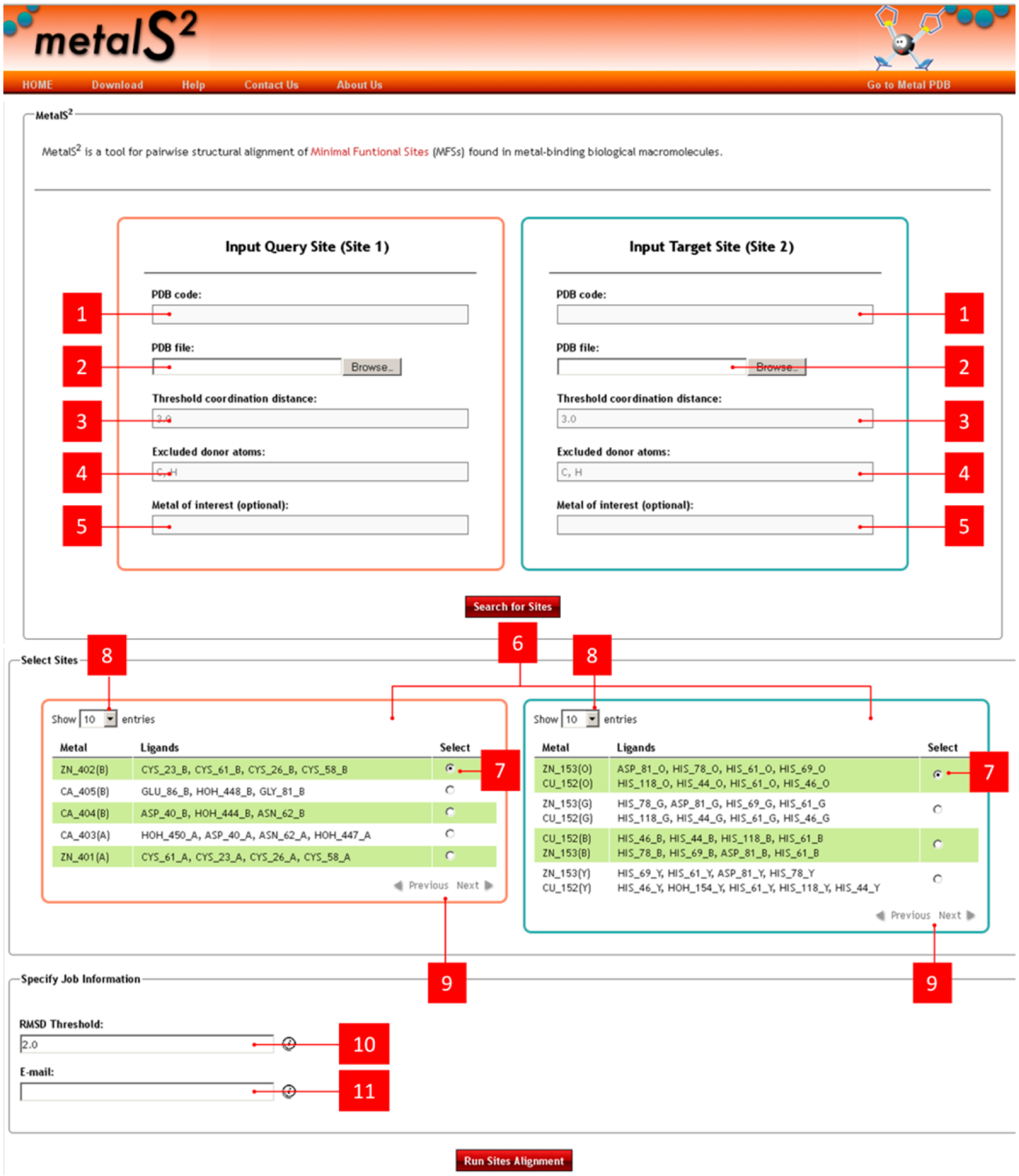
$$\text{Matthews correlation coefficient (MCC)} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

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

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

MetalS^2 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^2 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.[18,31] Zinc MFSs were assigned one function among catalytic, structural, regulatory, substrate, and unknown; nonheme iron MFSs were



MetalS²

MetalS² is a tool for pairwise structural alignment of Minimal Funtional Sites (MFSs) found in metal-binding biological macromolecules.

Input Query Site (Site 1)

PDB code:

PDB file:

Threshold coordination distance: 3.0

Excluded donor atoms: C, H

Metal of interest (optional):

Input Target Site (Site 2)

PDB code:

PDB file:

Threshold coordination distance: 3.0

Excluded donor atoms: C, H

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	●
CA_405(B)	GLU_86_B, HOH_448_B, GLY_81_B	○
CA_404(B)	ASP_40_B, HOH_444_B, ASN_62_B	○
CA_403(A)	HOH_450_A, ASP_40_A, ASN_62_A, HOH_447_A	○
ZN_401(A)	CYS_61_A, CYS_23_A, CYS_26_A, CYS_58_A	○

◀ Previous Next ▶

Show 10 entries

Metal	Ligands	Select

ZN_153(O)
CU_152(O)
	ASP_81_O, HIS_78_O, HIS_61_O, HIS_69_O
	HIS_118_O, HIS_44_O, HIS_61_O, HIS_46_O
		●

ZN_153(G)
CU_152(G)
	HIS_78_G, ASP_81_G, HIS_69_G, HIS_61_G
	HIS_118_G, HIS_44_G, HIS_61_G, HIS_46_G
		○

CU_152(B)
ZN_153(B)
	HIS_46_B, HIS_44_B, HIS_118_B, HIS_61_B
	HIS_78_B, HIS_69_B, ASP_81_B, HIS_61_B
		○

ZN_153(Y)
CU_152(Y)
	HIS_69_Y, HIS_61_Y, ASP_81_Y, HIS_78_Y
	HIS_46_Y, HOH_154_Y, HIS_61_Y, HIS_118_Y, HIS_44_Y
		○

◀ Previous Next ▶

Specify Job Information

RMSD Threshold: 2.0

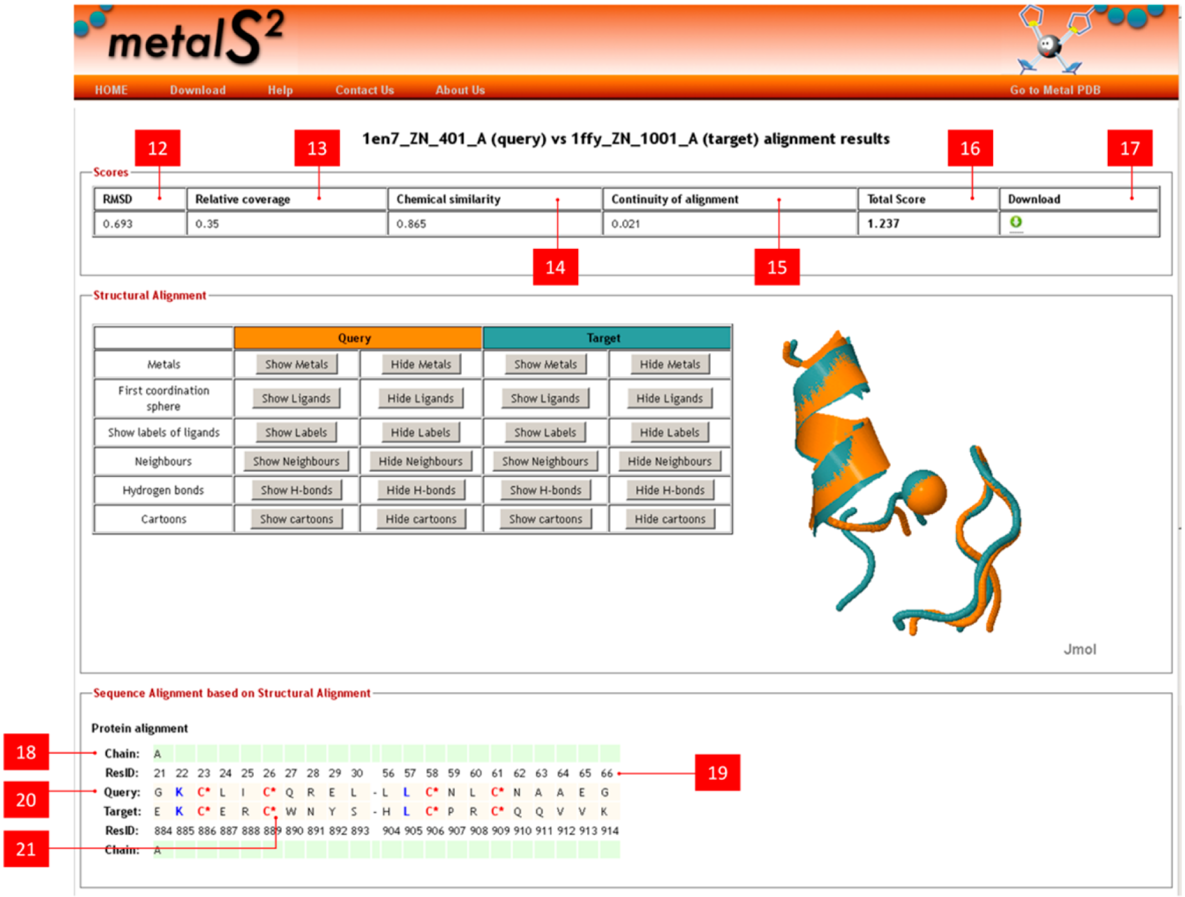
E-mail:

Figure 1. Input form on the MetalS² 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

MetalS² 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 MetalS² are identified in the input protein structures using a previously described approach.[21] 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.



1en7_ZN_401_A (query) vs 1ffy_ZN_1001_A (target) alignment results

Scores

RMSD	Relative coverage	Chemical similarity	Continuity of alignment	Total Score	Download
0.693	0.35	0.865	0.021	**1.237**	⬇

Structural Alignment

	Query		Target	
Metals	Show Metals	Hide Metals	Show Metals	Hide Metals
First coordination sphere	Show Ligands	Hide Ligands	Show Ligands	Hide Ligands
Show labels of ligands	Show Labels	Hide Labels	Show Labels	Hide Labels
Neighbours	Show Neighbours	Hide Neighbours	Show Neighbours	Hide Neighbours
Hydrogen bonds	Show H-bonds	Hide H-bonds	Show H-bonds	Hide H-bonds
Cartoons	Show cartoons	Hide cartoons	Show cartoons	Hide cartoons

Sequence Alignment based on Structural Alignment

Protein alignment

```
Chain:  A
ResID:  21  22  23  24  25  26  27  28  29  30    56  57  58  59  60  61  62  63  64  65  66
Query:  G   K   C*  L   I   C*  Q   R   E   L   - L   L   C*  N   L   C*  N   A   A   E   G
Target: E   K   C*  E   R   C*  W   N   Y   S   - H   L   C*  P   R   C*  Q   Q   V   V   K
ResID:  884 885 886 887 888 889 890 891 892 893   904 905 906 907 908 909 910 911 912 913 914
Chain:  A
```

Figure 2. Results of the MFS comparison shown on the MetalS^2 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^2 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_4S_4 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.[18,21] The MetalS^2 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.[18,31] 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[32] or CATH[33] 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[34] 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,[26] MAMMOTH,[27] and TM-align.[28] Among these, the FAST program was already shown by some of us to have an acceptable performance when applied to similar analyses.[31] 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^2 on the Test Sets Derived from the Fe- and Zn-Data Sets[a]

	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%

[a] TP: True positives (number of MFS pairs aligned by MetalS^2 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^2 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^2 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^2 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^2 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^2 total score indicates a good structural alignment. With this threshold, the precision of MetalS^2 on the combined test set is 91.6% and its accuracy is 92.7%; at a threshold of 2.25 the precision of MetalS^2 is 99%.

Over the entire Fe-data set, 27 MFS pairs could be superimposed by MetalS^2 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^2 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^2 and by the other programs is 4.16 (Table 2). The ratio increases with increasing threshold for the MetalS^2 score. The MetalS^2 score of the top 1,000 structural alignments of MFS pairs from the Zn-data set (excluding self-alignments) ranges from 0.271 to 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^2 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^2 and by the other methods over the entire data sets. The test demonstrated that this was actually the case (Supplementary Table S3).

Table 2. Number of Structural Alignments for Which the MetalS^2 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^2 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

We then checked whether the MFS pairs that MetalS^2 could align with a score below a given threshold were functionally related (Table 3). To this end, we exploited the functional assignments already published.[18,31] 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^2 As a Function of the Total Score

MetalS^2 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.[18] 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 metallloproteins and metalloprotein families.[18] 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[35] or shape descriptors[36]) 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,[29] CE,[37] DALI,[38] TOPMATCH,[39] and SALIGN[40] do not yield any result when MFSs are used as input.

As its first step, MetalS^2 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^2 to the results of other programs. A notable implication of these considerations is that MetalS^2 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^2 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^2 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 β-sheets in a crossed manner.

The benchmark used to assess the performance of MetalS^2 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^2 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^2 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^2 provides a better way to recognize them. Still, by inspecting MetalS^2 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^2 was the only program that could produce a good quality alignment; some examples are given in Figure 3. The

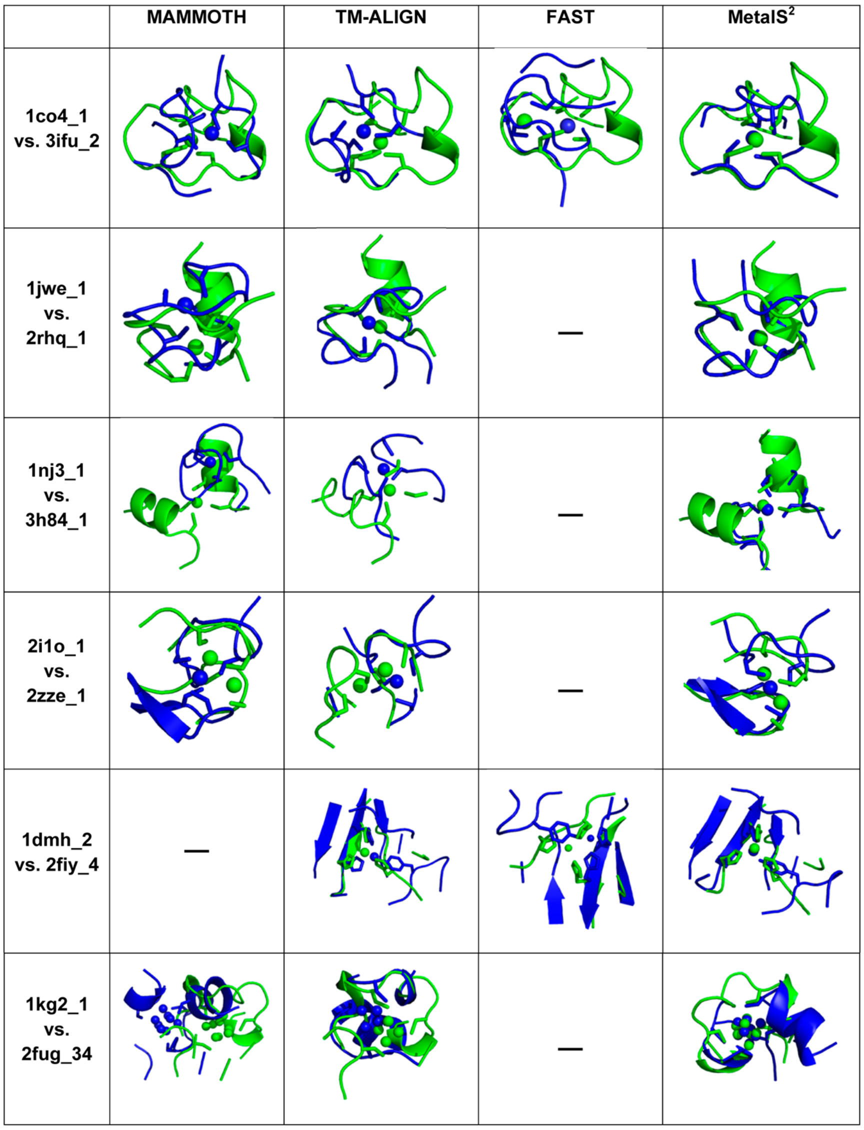


Figure 3. Comparison of selected structural alignments by MetalS^2 and the other programs tested in this work. The absence of the image indicates that the program did not produce any output. Site names in the first column correspond to those adopted in the MetalPDB database.[21]

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^2 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^2 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.[41−43] Hence, the need for an approach to 3D structure comparison that incorporates the underlying philosophy of MFSs such as MetalS^2. 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.[29,44]

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^2 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^{2}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

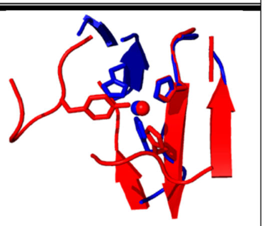
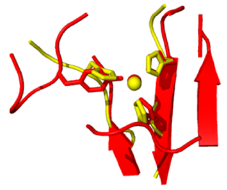
	CATH	SCOP	Pfam	EC	$MetalS^2$ Superposition
1dmh_2 **(Catalytic)**	2.60.130.10	b.3.6.1	Dioxygenase_C	1.13.11.1	
2b5h_1 **(Catalytic)**	n/a	b.82.1.19	CDO_I	1.13.11.20	
1dmh_2 **(Catalytic)**	2.60.130.10	b.3.6.1	Dioxygenase_C	1.13.11.1	
2fiy_4 **(Unknown)**	3.90.1670.10	e.59.1.1	FdhE	n/a	

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.[21]

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, $MetalS^2$ 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^{2+} ion in an octahedral coordination environment that includes three protein ligands, one DNA ligand and two water molecules, and one Zn^{2+} 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 $MetalS^2$, 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, $MetalS^2$ is available both as a stand-alone program and a Web tool (http://metalweb.cerm.unifi.it/tools/metals2/) within our MetalPDB platform.[21] $MetalS^2$, 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.[18] 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 $MetalS^2$ 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, $MetalS^2$ 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 $MetalS^2$. 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 $MetalS^2$ 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 $MetalS^2$ 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 $MetalS^2$ generated high quality alignments, whereas other programs did not perform satisfactorily. Conversely, in a few cases $MetalS^2$ 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 $MetalS^2$, which could identify a much larger number of structurally related MFS pairs in our example data sets (Table 2). Regarding possible usage scenarios, $MetalS^2$ 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, $MetalS^2$ can constitute the basis for innovative MFS classification essentially by conveniently replacing FAST within approaches similar to those we previously applied.[31]

■ ASSOCIATED CONTENT

Supporting Information

Tables S1 (Scores of all-versus-all alignments calculated by MAMMOTH, FAST, TM-align, and $MetalS^2$ on the Fe-data set), **S2** (Scores of all-versus-all alignments calculated by MAMMOTH, FAST, TM-align, and MetalS^2 on the Zn-data set), and **S3** (Wilcoxon rank sum test for aligned MFS pairs). **Figures S1** (Flowchart of MetalS^2), **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.; Prlic, 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.; Hovmoller, 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) Konc, J.; Janezic, 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.

4.2. MetalS^3, a database-mining tool for the identification of structurally similar metal sites

Valasatava Y[1], Rosato A[1, 2], Cavallaro G[1], and Andreini C[1, 2].

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

J Biol Inorg Chem. 2014 Apr 4.

J Biol Inorg Chem (2014) 19:937–945
DOI 10.1007/s00775-014-1128-3

ORIGINAL PAPER

MetalS^3, a database-mining tool for the identification of structurally similar metal sites

Yana Valasatava · Antonio Rosato · Gabriele Cavallaro · Claudia Andreini

Received: 23 January 2014 / Accepted: 19 March 2014 / Published online: 4 April 2014

Abstract We have developed a database search tool to identify metal sites having structural similarity to a query metal site structure within the MetalPDB database of minimal functional sites (MFSs) contained in metal-binding biological macromolecules. MFSs describe the local environment around the metal(s) independently of the larger context of the macromolecular structure. Such a local environment has a determinant role in tuning the chemical reactivity of the metal, ultimately contributing to the functional properties of the whole system. The database search tool, which we called MetalS^3 (Metal Sites Similarity Search), can be accessed through a Web interface at http://metalweb.cerm.unifi.it/tools/metals3/. MetalS^3 uses a suitably adapted version of an algorithm that we previously developed to systematically compare the structure of the query metal site with each MFS in MetalPDB. For each MFS, the best superposition is kept. All these superpositions are then ranked according to the MetalS^3 scoring function and are presented to the user in tabular form. The user can interact with the output Web page to visualize the structural alignment or the sequence alignment derived from it. Options to filter the results are available. Test calculations show that the MetalS^3 output correlates well with expectations from protein homology considerations. Furthermore, we describe some usage scenarios that highlight the usefulness of MetalS^3 to obtain mechanistic and functional hints regardless of homology.

Keywords Metalloprotein · Metalloenzyme · Bioinorganic chemistry · Structural biology · Zinc · Iron

Introduction

Bioinorganic or biological inorganic chemistry is the discipline dealing with the interaction between inorganic substances and molecules of biological interest [1–3]. It is a wide scientific field that 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. The interaction between metal ions or metal-containing cofactors and biological macromolecules can be studied in atomic detail through 3D structural studies, thus providing a connection between bioinorganic chemistry and structural biology [4].

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, and 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 donor amino acid or nucleotide. Databases reporting on the geometric properties of metal-binding sites in proteins [5] or nucleic acids [6] are available. They are derived from the

Electronic supplementary material The online version of this article (doi:10.1007/s00775-014-1128-3) contains supplementary material, which is available to authorized users.

Y. Valasatava · A. Rosato · G. Cavallaro · C. Andreini (✉)
Magnetic Resonance Center (CERM), University of Florence,
Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
e-mail: andreini@cerm.unifi.it

A. Rosato · C. Andreini
Department of Chemistry, University of Florence,
Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

coordinate files deposited in the Protein Data Bank (PDB) [7]. 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 genome sequences [8–10]. We have described how the inclusion of the surroundings of the metal-binding site in structure-based analyses strengthens the relationship of the sites with functional properties [11, 12]. 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). 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 [11, 13] (Fig. S1). 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 usefulness of the MFS concept outlined above has its chemicophysical foundation in the fact that the local environment of the metal has a determinant role in tuning its properties and thus its chemical reactivity [14, 15]. Instead, the macromolecular matrix is instrumental in determining, e.g., substrate selection [16] or partner recognition [17].

To make MFS analyses available to the scientific community, we developed two different resources: (1) MetalPDB [18], a database of all MFSs contained in the PDB, which is automatically updated, providing access to structural and functional information, including atomic coordinates, for each MFS in any metal-binding macromolecule of known 3D structure; (2) MetalS^2 (Metal Sites Superposition) [12], a tool for the metal-centered superposition of MFS pairs, applicable to structures already in the PDB or to structural files belonging to the user. In the present work, we present MetalS^3 (Metal Sites Similarity Search), a new tool that bridges the two aforementioned resources by allowing researchers to input the coordinates of one MFS and perform a systematic search of the entire MetalPDB database to identify structurally similar sites, regardless of overall fold similarity or protein homology. MetalS^3 is based on the same conceptual approach of MetalS^2, with some minor modifications. However, its implementation as a tool for a database search makes possible a completely different usage scenario, with a main focus on knowledge discovery through the unbiased exploration of the structural space of metal sites.

Methods

The MetalS^3 algorithm

MetalS^2 performs the superposition of two MFSs by performing the following steps [12]: (1) computing and overlapping the geometric centers of the metal atoms contained in each MFS; (2) systematically computing a set of initial configurations (poses), in each of which the geometric centers of the metals and two different pairs of donor atoms from the two sites are used to superimpose the MFSs (Fig. S2); (3) ranking all the poses on the basis of a specifically designed scoring function; (4) optimizing a subgroup of the poses (by default, those in the best 40 % of the entire score range) by allowing the geometric centers and the ligands to be displaced with respect to one another. The MetalS^2 score consists of three terms that account, respectively, for the biochemical similarity of the amino acids put in correspondence (sequence similarity term), the ratio between the total length of the sequence alignment and the length of the smallest site (i.e., the fractional coverage of the smaller site) (fractional coverage term), and the number and length of consecutive sequence segments in the superposition (fragmentation term). Amino acid correspondences are established on the basis of Cα–Cα and Cβ–Cβ distances. In step 4 of the procedure, the root mean square deviation (RMSD) of the coordinates in the superposition is optimized and amino acid correspondences are reevaluated. Note that atoms from exogenous (i.e., nonprotein, non-nucleic acid) ligands are not included in the computation neither of the RMSD nor of the score. The reason for this is that, especially in the context of MetalS^3, we want to identify and quantify similarities among the macromolecular components of the MFSs. Exogenous ligands contribute to the definition of each MFS geometry as well as to the calculation of the set of initial poses, which is based purely on geometrical considerations. Thereafter, and especially for the purpose of scoring the solutions, such ligands are no longer taken into account. This makes the final ranking dependent only on the similarities between the macromolecular structures, as desired, and avoids possible biases due to common arrangements of the ligands around the metal ion, e.g., as for chelators such as hydroxamic acid derivatives in zinc enzymes, which maintain a fixed geometry in most or all structures.

For the present work, we implemented a new Web interface, MetalS^3, that allows a user to upload a metal-containing macromolecular structure (or select it from the MetalPDB database) in PDB format, select any MFS (automatically detected) contained in it, and systematically compare it against all MFSs in MetalPDB using the MetalS^2 algorithm. A list of hits is returned by MetalS^3, sorted by the corresponding score. We introduced some minor modifications to the MetalS^2 procedure and scoring function described in the previous paragraph. In MetalS^3, the fractional coverage term always refers to the input (query) MFS rather than to the smallest site of the pair being superposed. In addition, the optimization step is iterated as long as the superposition score keeps decreasing.

To reduce the computational effort, we imposed some limitations on the difference in the number of donor atoms between the query MFS and any MFS from MetalPDB, which are recapitulated by the following formula:

$$\begin{cases} a = \frac{N}{4}, \text{ if}, \frac{N}{4} > 2, \text{ else } 2 \\ b = 4N, \text{ if } 4N < 20, \text{ else } 20 \end{cases} \tag{1}$$

where a and b are, respectively, the smallest and largest number of donor atoms that an MFS from the database can have for it to be included in the search set and N is the number of donor atoms in the query. In practice, any MFS in MetalPDB with a number of donor atoms outside the [a; b] range is excluded from the search. For example, a query MFS with four donor atoms will be compared only with MFSs from MetalPDB having between two and 16 donors. We believe that the application of the above-mentioned restriction does not reduce the usefulness of the results, as it seems reasonable to assume that any structural similarity between MFSs with a disparity in the number of donor atoms beyond the limits imposed by Eq. 1 does not have functional relevance.

Implementation of MetalS^3

All back-end scripts are implemented in Python 2.6.6 (http://www.python.org/) on a Linux platform. The front end was implemented using Mako, a template library written in Python included by default with the Pylons Web application framework, JavaScript, and Cascading Style Sheets. By using the Python language, we could also exploit the following resources: SciPy 0.7.2, a library of scientific and numerical routines; NumPy 1.4.1, a language extension that adds support for large and fast, multidimensional arrays and matrices; and p3d [19], a Python module for structural bioinformatics. The MetalS^3 server is currently hosted on a 24-CPU (AMD Opteron$^{\text{TM}}$ 6234) server.

The MetalS^3 Web interface

The Web interface of MetalS^3 allows the user to run queries against all representative MFSs of the equistructural MFS clusters defined in MetalPDB. Each of these MFSs represents a group of sites that are found in proteins with the same fold, as judged from sequence similarity and Pfam [20] domain assignments, and occur at the same spatial location within that fold. For example, a single representative MFS represents all the sites of rubredoxins from various organisms and with different metalation. MetalPDB currently contains 17,936 clusters of equistructural MFSs. As mentioned previously, the dataset of representative MFSs against which the query is actually compared is the subgroup of all 17,936 sites that satisfies Eq. 1. Thus, the size and the characteristics of the subgroup depend on the input query MFS, and particularly on the number of donor atoms it contains. In turn, this influences the overall calculation time.

After a calculation is finished, the user is presented with a list of hits having structural similarity to the query, ordered by the total MetalS^3 score (the list can be resorted according to different parameters, such as individual score components). It is then possible to select a specific hit, i.e., a specific representative MFS, and run a refinement calculation in which the query is compared with each individual site in the corresponding equistructural MFS cluster. A link to the results of the search is e-mailed to the user at the end of each of these two stages.

Results

A brief description of the input and output interfaces of MetalS^3 is available as electronic supplementary material (text and Figs. S3 and S4). We conducted various experiments to assess our implementation of MetalS^3 with respect to its capability to identify relevant hits within the MetalPDB database as well as with respect to the typical times required to obtain the results of a calculation.

Because MetalS^3 searches are initially performed only against representative MFSs and not the entire content of the MetalPDB database, it is important to assess whether this approach consistently returns relevant functional information. To do this, we used an example dataset of 100 different MFSs randomly picked from deposited PDB structures (Table S1). These examples, which differed in metal content as well as coordination number and geometry, were used as input queries to MetalS^3. Crucially, the examples were selected in order to avoid including any representative MFS as defined in MetalPDB. In this way, we could straightforwardly classify the output of MetalS^3 depending on whether the best-scoring hit corresponded to the representative of the cluster to which each query MFS was known to belong. In fact, even though the clustering procedure implemented in MetalPDB does not directly compare the structure of the different MFSs assigned to a cluster, in the large majority of cases the MFSs within a cluster should be similar to each other because the proteins in the cluster can be assumed to be homologous. In 75 % of cases, this was indeed observed. Notably, if we optimize all the poses, instead of a well-scoring subgroup, the above-mentioned result increases only to 76 %. We then analyzed manually the 25 cases for which the best hit identified by MetalS^3 was not the representative MFS of the cluster to which the query belongs in the MetalPDB database. For 20 of them we observed that the result obtained depended on the clustering within MetalPDB being incomplete, i.e., failing to group together MFSs that indeed are bound to homologous proteins. In turn, this is due to missing Pfam assignments or, often, to a given protein superfamily being mapped to multiple Pfam domains [21]. Instead, in five cases MetalS^3 identified a structural similarity between a pair of MFSs (the query and the returned hit) that was higher than that between the query and the representative MFS of its equistructural cluster in MetalPDB. These are cases where either highly similar MFSs are embedded in different folds (three) or the MFS representative does not adequately represent the cluster (two). The representative MFS of a cluster is chosen solely on the basis of the resolution (i.e., quality) of the corresponding 3D structure [18]. Consequently, the representative MFS cannot be regarded as a sort of "average" MFS, and there is no specific property regarding its structural similarity to the other MFSs in the cluster. A third option is that the assignment of the query MFS to the MetalPDB cluster, which was performed automatically, did not reflect the large structural variability of the MFSs within the cluster. This was not observed here. An additional consideration is that, because of the way the score is constructed, smaller query MFSs tend to be less discriminative and therefore may more easily provide high-scoring hits also to MFSs not closely related (but still structurally similar).

If one looks at the five best scoring hits, then in only ten cases from the 100 examples run was the MFS representative of the cluster of the query site not included. As already mentioned, in two instances we observed that the specific representative MFS did not reflect the "consensus" coordination geometry of its cluster. However, in most cases, the reason for the observed behavior was an incomplete clustering of the structures, in turn typically resulting from problems in the mapping of Pfam domains. This caused structures highly similar to the query not to be included in the same equistructural cluster.

The calculation times are dependent on the number of donor atoms (N) in the query MFS, as the number of poses that need to computed and compared scales with $N(N-1)$ [12] (Fig. 1). For a given number of atoms, calculations are faster the higher the number of donor atoms from exogenous ligands (such as small metal-binding molecules or ions) because these are not considered in amino acid matching and RMSD computations (see "Methods"). The calculation times are less than 2 h for sites with up to four protein donor atoms, whereas, owing to the parabolic increase of calculation times, they are as long as 10 h for sites with nine donor atoms (if all are from protein ligands) and within 24 h for multinuclear sites with 12 donor atoms from the protein moiety. Of all representative MFS sites collected in MetalPDB, 95.1 % have nine donor atoms or fewer. Under the assumption that MetalPDB adequately describes the diversity of MFSs occurring in nature, the data given above may suggest that users will most often submit queries that can be dealt with in 10 h or less. In any case, results are always sent to the users via e-mail, as even the simplest calculations require at least a few minutes.

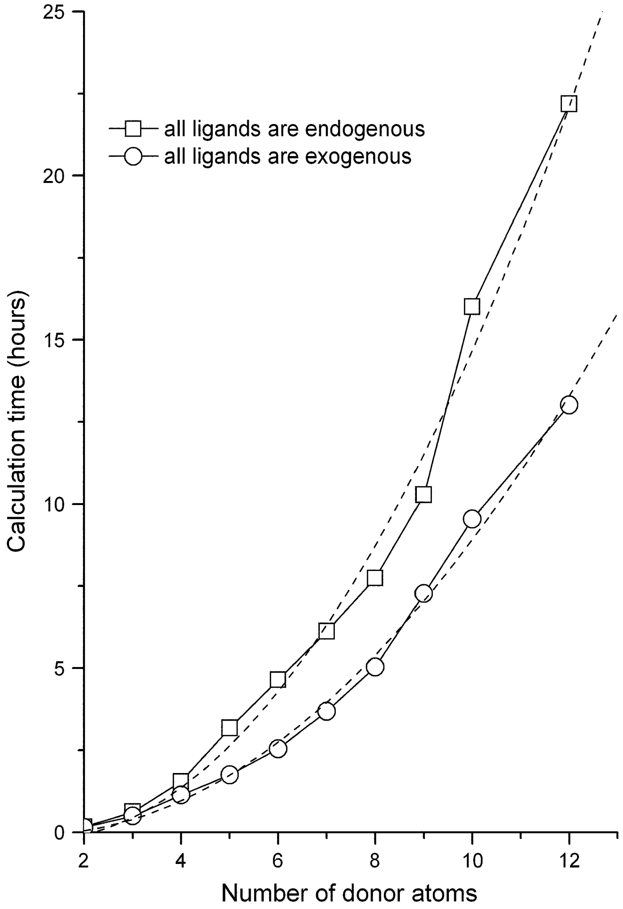


Fig. 1 Calculation times for MetalS^3 queries as a function of the number and type of donor atoms. *Dashed lines* are the best fit to a second-order polynomial

Discussion

MetalS^3 is a Web interface that allows the user to systematically compare an MFS of interest (query) with the contents of the MetalPDB database [18], i.e., with an ensemble representing the diversity of known MFSs. This is achieved through a suitably modified implementation of the MetalS^2 algorithm [12]. Typically, the hits returned for a query will comprise sites that are contained within a protein homologous to the protein containing the query MFS as well as sites from unrelated proteins. The presence in the output page of one or the other type of hit, as well as their relative abundance, will depend on the cutoffs defined to exclude hits from the visualization (Fig. S3). The cutoffs can be adjusted also after the calculation has finished, through the “Filter Results” button on the output page (Fig. S4). Increasing the cutoff values will result in a longer list of hits being displayed.

Our test calculations show that the top position in the list of the hits is highly likely to be occupied by an MFS contained within a homolog of the query protein; when the top five hits are considered, this is verified for as many as 90 % of the examples that we run. According to the definition of the MetalPDB database, on which MetalS^3 builds, this situation corresponds to the query and hit MFSs belonging to the same equistructural cluster. For MFSs in MetalPDB to be clustered, it is actually requested that the sites occupy the same position within the fold after the entire protein structure has been superimposed, and the structures of the MFSs belonging to a given cluster are not compared with one another. The approach of MetalS^3 is entirely different, as it operates only on the MFSs, disregarding the rest of the protein structure. The very good correlation between the fold-based clustering results and the MetalS^3 output points to the high similarity of the local 3D structure around the metal site being a possible indicator of metalloprotein homology. This is supported also by the fact that in 20 of the examples, MetalS^3 indicated that the clustering within MetalPDB was incomplete. Incomplete clustering typically results from the homology relationship between metalloproteins bearing structurally similar MFSs being hidden by the fact that the Pfam domain assignments we use in the definition of equistructural clusters are fine-grained and may occasionally separate a single superfamily into multiple domain definitions. To address this issue, the user can verify if the Pfam domains of interest belong to the same Pfam clan [21]. One can possibly further speculate that if the MFS properties must be defined tightly to make possible the correct protein function [i.e., to correctly define the reactivity of the metal ion(s) in the MFS], then conservation of the 3D structure of the MFS will be particularly strict among homologous proteins. Consequently, the intracluster variability of the MFS structure may be informative on the requirements imposed by the catalysis on the MFS features or, in other words, on how the functional and mechanistic properties of the system are encoded in the structure.

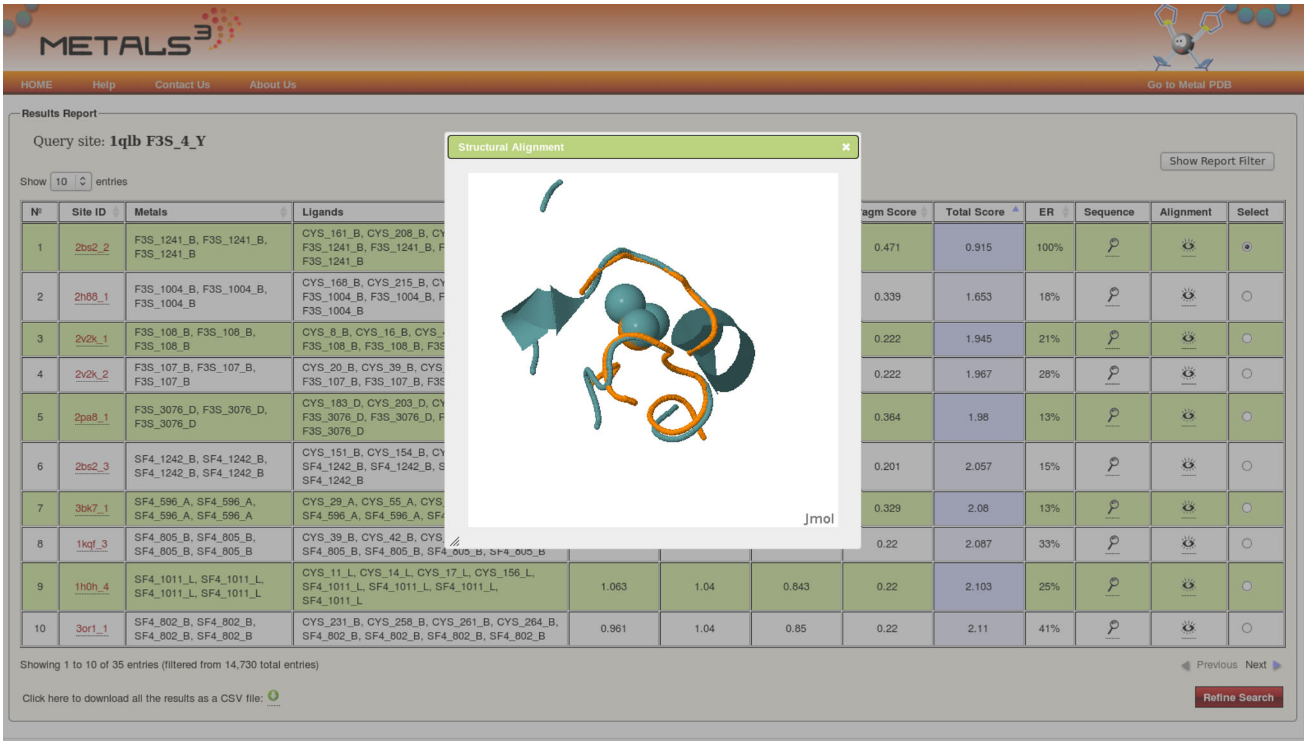
A practical application of MetalS^3 is to detect MFS structural similarities that are not associated with a homology relationship among the proteins harboring the MFSs (indicated by the MFS mapping to a shared Pfam domain or domain clan). These situations may be indicative of the occurrence of common functional properties that are endowed by the MFS itself. Such observations can provide useful hints for experimental work. In this usage scenario, the best hit returned by MetalS^3 is often uninteresting (i.e., when it is bound to a protein with the same domain composition as the protein containing the query MFS), and one should focus on worse-scoring hits. Operatively, the domain composition of a hit MFS can be immediately obtained by looking up that MFS in the MetalPDB database [18]. Below, we briefly discuss some examples not included in the 100 test dataset.

As a first example, we took one of the two equivalent Fe_3S_4 clusters in the PDB structure of fumarate reductase from *Wolinella succinogenes* (PDB ID 1QLB [22]), which is identified as site 1qlb_4 in MetalPDB (hereafter, we will use the PDB code in lowercase letters followed by an underscore and a number to indicate a specific MFS within the MetalPDB database, whereas we will use the PDB code in uppercase letters to indicate the PDB entry). This site is located with a ferredoxin-type domain, and it is likely to be part of the electron transfer pathway. MetalS^3 returns as the fifth hit, with a total score of 1.98, a site harboring an Fe_3S_4 cluster in the D subunit of the structure of the DNA-directed RNA polymerase from *Sulfolobus solfataricus* P2 (PDB ID 2PA8 [23]). Despite a sequence identity between these two MFSs of only 13 % over 15 amino acids, the superposition is good (RMSD 0.799 Å) (Fig. 2).

The latter cluster, which is possibly an Fe_4S_4 cluster in vivo, is found in the corresponding subunits of the polymerases from various species of Archaea and Eukarya, but not of Bacteria [24]. The domain containing the MFS within subunit D is not present in all archaeal RNA polymerases, but it is actually characteristics of a specific evolutionary lineage of Archaea. Here we observed that the binding mode of the Fe_3S_4 cluster within subunit D of *S. solfataricus* P2 polymerase actually bears some similarity to an unrelated episilonproteobacterial system.

A second example is provided by the MFS containing the magnesium(II) ion identified as residue 9,018 (MetalPDB entry 1g0u_1) within the structure of the core particle of the yeast proteasome (PDB ID 1G0U [25]). This MFS is interfacial, as it contains protein ligands from subunits I and Y. MetalS^3 returns hits also to sites containing metal ions other than magnesium. One of these is the MFS defined around the calcium(II) ion identified as residue 501 in the structure of human calcium and integrin binding protein 1 (PDB ID 1Y1A [26]), with a total score of 2.427 and, in particular, a sequence identity of 0 % (Fig. 3). This MFS is located within an EF-hand motif. Such a structural similarity would be extremely hard to identify by any other method, especially a sequence-based method. Magnesium(II) and calcium(II) are known to compete for binding in EF-hand sites [27]. The similarity between the two MFSs may thus underlie commonalities in the atomic mechanism by which the metal affinity is tuned.

3ZFJ is a recently solved NMR structure of a PhtD domain from *Streptococcus pneumoniae* that binds a single zinc(II) ion [28]. At the time of writing, it is not yet



Results Report

Query site: **1qlb F3S_4_Y**

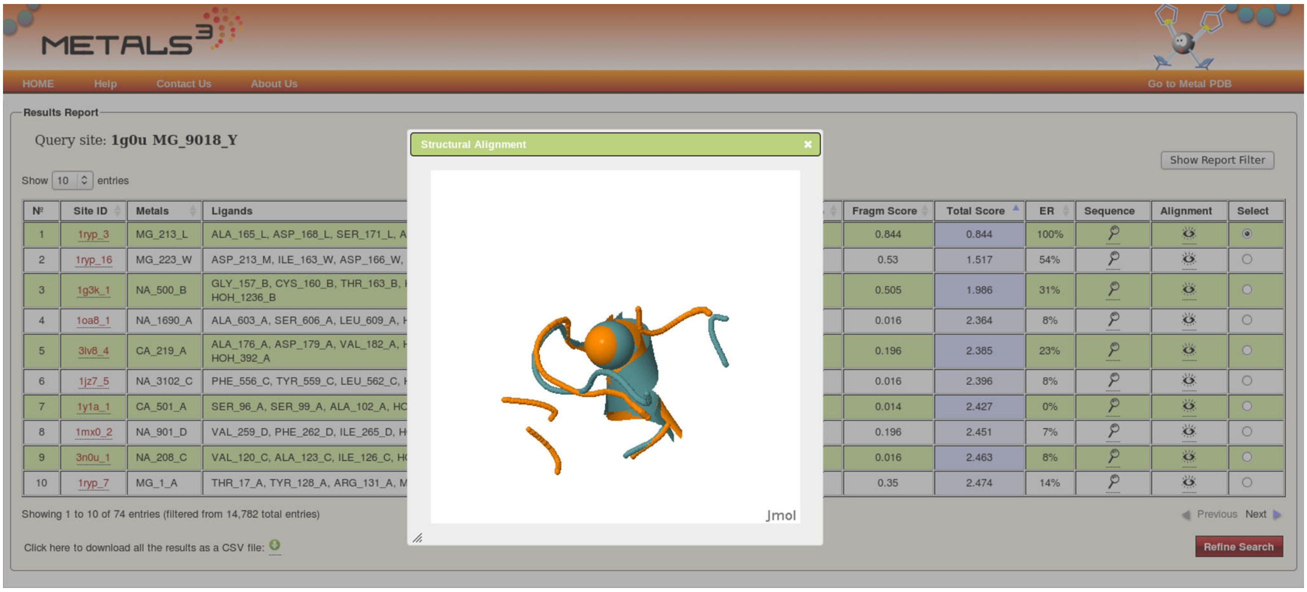
Show 10 entries

Nº	Site ID	Metals	Ligands				agm Score	Total Score	ER	Sequence	Alignment	Select
1	2bs2_2	F3S_1241_B, F3S_1241_B, F3S_1241_B	CYS_161_B, CYS_208_B, CY F3S_1241_B, F3S_1241_B, F F3S_1241_B				0.471	0.915	100%	🔍	👁	●
2	2h88_1	F3S_1004_B, F3S_1004_B, F3S_1004_B	CYS_168_B, CYS_215_B, CY F3S_1004_B, F3S_1004_B, F F3S_1004_B				0.339	1.653	18%	🔍	👁	○
3	2v2k_1	F3S_108_B, F3S_108_B, F3S_108_B	CYS_8_B, CYS_16_B, CYS_ F3S_108_B, F3S_108_B, F3S				0.222	1.945	21%	🔍	👁	○
4	2v2k_2	F3S_107_B, F3S_107_B, F3S_107_B	CYS_20_B, CYS_39_B, CYS F3S_107_B, F3S_107_B, F3S				0.222	1.967	28%	🔍	👁	○
5	2pa8_1	F3S_3076_D, F3S_3076_D, F3S_3076_D	CYS_183_D, CYS_203_D, CY F3S_3076_D, F3S_3076_D, F F3S_3076_D				0.364	1.98	13%	🔍	👁	○
6	2bs2_3	SF4_1242_B, SF4_1242_B, SF4_1242_B, SF4_1242_B	CYS_151_B, CYS_154_B, CY SF4_1242_B, SF4_1242_B, S SF4_1242_B				0.201	2.057	15%	🔍	👁	○
7	3bk7_1	SF4_596_A, SF4_596_A, SF4_596_A, SF4_596_A	CYS_29_A, CYS_55_A, CYS SF4_596_A, SF4_596_A, SF4				0.329	2.08	13%	🔍	👁	○
8	1kqf_3	SF4_805_B, SF4_805_B, SF4_805_B, SF4_805_B	CYS_39_B, CYS_42_B, CYS SF4_805_B, SF4_805_B, SF4_805_B, SF4_805_B				0.22	2.087	33%	🔍	👁	○
9	1h0h_4	SF4_1011_L, SF4_1011_L, SF4_1011_L, SF4_1011_L	CYS_11_L, CYS_14_L, CYS_17_L, CYS_156_L, SF4_1011_L, SF4_1011_L, SF4_1011_L, SF4_1011_L	1.063	1.04	0.843	0.22	2.103	25%	🔍	👁	○
10	3or1_1	SF4_802_B, SF4_802_B, SF4_802_B, SF4_802_B	CYS_231_B, CYS_258_B, CYS_261_B, CYS_264_B, SF4_802_B, SF4_802_B, SF4_802_B, SF4_802_B	0.961	1.04	0.85	0.22	2.11	41%	🔍	👁	○

Showing 1 to 10 of 35 entries (filtered from 14,730 total entries)

Click here to download all the results as a CSV file:

Fig. 2 Output result page for a calculation performed using the 1qlb_4 site as the query. The *inset* shows the structural alignment to the fifth hit, 2pa8_1



Results Report

Query site: **1g0u MG_9018_Y**

Show 10 entries

Nº	Site ID	Metals	Ligands	Fragm Score	Total Score	ER	Sequence	Alignment	Select
1	1ryp_3	MG_213_L	ALA_165_L, ASP_168_L, SER_171_L, A	0.844	0.844	100%	🔍	👁	●
2	1ryp_16	MG_223_W	ASP_213_M, ILE_163_W, ASP_166_W,	0.53	1.517	54%	🔍	👁	○
3	1g3k_1	NA_500_B	GLY_157_B, CYS_160_B, THR_163_B, HOH_1236_B	0.505	1.986	31%	🔍	👁	○
4	1oa8_1	NA_1690_A	ALA_603_A, SER_606_A, LEU_609_A,	0.016	2.364	8%	🔍	👁	○
5	3lv8_4	CA_219_A	ALA_176_A, ASP_179_A, VAL_182_A, HOH_392_A	0.196	2.385	23%	🔍	👁	○
6	1jz7_5	NA_3102_C	PHE_556_C, TYR_559_C, LEU_562_C,	0.016	2.396	8%	🔍	👁	○
7	1y1a_1	CA_501_A	SER_96_A, SER_99_A, ALA_102_A, HC	0.014	2.427	0%	🔍	👁	○
8	1mx0_2	NA_901_D	VAL_259_D, PHE_262_D, ILE_265_D, H	0.196	2.451	7%	🔍	👁	○
9	3n0u_1	NA_208_C	VAL_120_C, ALA_123_C, ILE_126_C, H	0.016	2.463	8%	🔍	👁	○
10	1ryp_7	MG_1_A	THR_17_A, TYR_128_A, ARG_131_A, M	0.35	2.474	14%	🔍	👁	○

Showing 1 to 10 of 74 entries (filtered from 14,782 total entries)

Click here to download all the results as a CSV file:

Fig. 3 Output result page for a calculation performed using the 1g0u_1 site as the query. The *inset* shows the structural alignment to the seventh hit, 1y1a_1

included in the MetalPDB database and therefore simulates well the situation of a real user. MetalS^3 identifies the 2CS7 structure [29] as the second best hit. In fact, both proteins contain the Pfam domain "Strep_his_triad," and have 23 % sequence identity. This is a case where the next update of MetalPDB would put the two in the same equistructural cluster. The above-mentioned proteins have a role in the uptake of zinc(II), by scavenging zinc(II) ions and then providing them to the extracellular membrane-anchored AdcAII transporter at the surface of *S.*

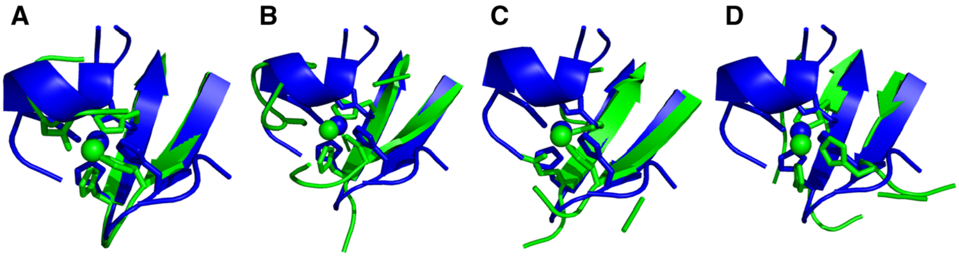


Fig. 4 Selected high-scoring zinc sites among the search results for a zinc-containing minimal functional site (MFS) from 3ZFJ. The 3ZFJ query structure is always in *blue* and in the same orientation. The superpositions to the sites **a** 2cs7_1, **b** 4hhj_1, **c** 2e26_5, and **d** 1txl_1 are displayed. Only protein ligands are shown; ZN(603) in 2E26 is additionally coordinated by two water molecules; ZN(216) in 1TXL is additionally coordinated by a water molecule

pneumoniae. The first hit is an iron-binding MFS from *Escherichia coli* galactose 1-phosphate uridylyltransferase (structure 1GUP [30]). This iron ion plays a structural role and is not essential to the enzyme activity [31]. It is useful to compare the hits returned by using either 3ZFJ or 2CS7 as queries. Among the shared top-scoring zinc proteins, one finds an MFS from structure 4HHJ [32], identified by the zinc ion with residue number 1,001. This ion has been proposed to have a structural and/or regulatory role for the activity of this RNA-dependent RNA polymerase [33]. Another common hit is from PDB entry 2E26 [34], identified by the zinc ion with residue number 603, which describes the structure of mouse reelin, a secreted glycoprotein. This ion is observed in the structures of both reelin alone and reelin in complex with apolipoprotein E receptor 2 [35], where it has fractional occupancy. Finally, MetalS^3 identifies the zinc-containing MFS of the ZinT protein (PDB ID 1TXL; S. Eswaramoorthy and S. Swaminathan, unpublished) as a further hit to the MFS in 2CS7; the MFSs of 3ZFJ and 1TXL also display good structural similarity (Fig. 4). ZinT is a periplasmic zinc transporter that facilitates metal recruitment during zinc shortage by binding zinc(II) with high affinity and subsequently transferring it to the ZnuA component of the ZnuABC membrane transporter [36, 37]. Intriguingly, in the zinc(II)-specific ABC uptake system AdcABC of *S. pneumoniae*, the AdcA protein, which does not interact with PhtD domains (see above), is a fusion between a ZnuA-like protein and a ZinT-like protein [38]. In summary, the present MetalS^3 analysis identified a minimal zinc-binding structure as being associated with reversible metal ion binding in zinc(II) transport, where different protein systems for zinc(II) uptake contain structurally similar MFSs, and in (hypothesized) zinc(II)-dependent regulation of intermolecular interactions.

An additional example is provided by the 4NAO structure, a homodimer that contains a single iron(II) ion per subunit [39], which was released in the PDB on January 15, 2014, and is not yet included in MetalPDB. This enzyme is an iron(II)/2-ketoglutarate-dependent dioxygenase that hydroxylates an *N*-(D-lysergyl-aminoacyl) lactam in the ergot fungus *Claviceps purpurea*. MetalS^3 identifies similarities to various other dioxygenases that are active against different substrates. In particular, the best hit is the iron(II) site of the 2CSG structure, an uncharacterized protein addressed by the Midwest Center for Structural Genomics, with 17 % sequence identity between the sites. Both structures feature organic ligands (2-ketoglutarate for 4NAO; succinate, which is a reaction product, and isocitrate for 2CSG) bound to the metal ion in corresponding positions (Fig. 5a). The second hit is a isopenicillin N synthase from *Emericella nidulans* (PDB ID 1ODM) [40]. This site has lower RMSD and higher sequence similarity to the query, and also features an organic ligand chelating the iron(II) ion in a manner relatively similar to that of 2-ketoglutarate of 4NAO (Fig. 5b). Notably, isopenicillin N synthase is not dependent on 2-ketoglutarate, whose functional role is performed by the tripeptide substrate [41]. The third hit contains a group of dioxygenases more closely related to 4NAO, which includes human phytanoyl-CoA dioxygenase (PhyH; PDB ID 2A1X). The article describing 4NAO provides a detailed comparison with PhyH and its homolog PhyHD1, which are actually the best results returned by a Dali [42] search based on the entire structure [39]. The 2-ketoglutarate molecules present in the 4NAO and PhyH structures chelate the metal ion in a closely similar manner (Fig. 5c). Finally, the fourth hit is a manganese(II) site in the 2-ketoglutarate-dependent dioxygenase AlkB (PDB ID 4JHT) [43] (Fig. 5d). AlkB is an iron(II)/2-ketoglutarate-dependent dioxygenase that catalyzes the oxidative demethylation of nucleic acids and histones [44]. It can bind manganese(II) in its catalytic site, yielding an inactive enzyme. Indeed, the aforementioned 4jht_1 site is the representative of a relatively large equistructural cluster in MetalPDB that contains the other structurally characterized AlkB MFSs. The cluster contains, for example, also the 3O1T structure [45], where the iron(II) ion is chelated by succinate, again in a position close to that of 2-ketoglutarate in 4NAO. The systems described in this paragraph map to three different, but

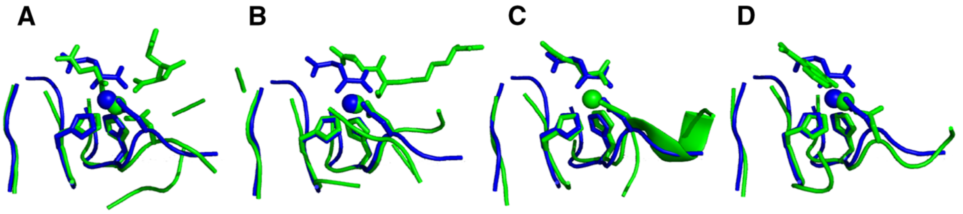


Fig. 5 The four top-scoring sites among the search results for the iron(II)-containing MFS in the A chain of 4NAO. The 4NAO query structure is always in *blue* and in the same orientation. The superpositions to the sites **a** 2csg_1, **b** 1odm_1, **c** 2a1x_1, and **d** 4jht_1 are displayed. The organic iron(II) ligands present in the various MFSs are shown as *sticks*. Water molecules are not shown

related to the same superfamily, Pfam domains: DUF1479 (2CSG), 2OG-FeII_Oxy (1ODB, 4JHT), and PhyH (4NAO, 2A1X). The results include also a case of a system where the physiological iron(II) ion was substituted in vitro. Thus, even for a large and widely studied protein superfamily such as that of iron(II)/2-ketoglutarate-dependent dioxygenases, MetalS^3 proves useful in the analysis of a newly solved structure to identify relationships across different subgroups in a manner that is independent of overall fold similarity.

Concluding remarks

MFSs in metal-binding biological macromolecules constitute a novel viewpoint for the elucidation of the mechanisms of function in these systems [11]. In this frame, we have developed the MetalPDB database [18]. MetalPDB contains a systematic analysis of all known MFSs. In particular, within the database all MFSs were grouped into so-called equistructural clusters. Each cluster contains all MFSs located at corresponding positions within the fold of homologous proteins. Recently, we developed the MetalS^2 program and Web server to perform pairwise structural superpositions of MFSs, providing a ground for the quantitative evaluation of MFS similarity [12]. MetalS^3, which is described in this work, is a Web-based tool (http://metalweb.cerm.unifi.it/tools/metals3/) that adopts the MetalS^2 algorithm to perform searches in the MetalPDB database. This is implemented as a first coarse-grained search against the ensemble of the MFSs representing MetalPDB equistructural clusters, followed by a refinement step in which the query MFS is compared with all the MFSs in a user-selected cluster. Although algorithmically very similar, MetalS^2 and MetalS^3 have somewhat different usage scenarios and make possible access to distinct information. MetalS^2 requires the user to have prior knowledge of the structures to be compared, either a pair or a group of related metalloproteins. In contrast, MetalS^3 constitutes an unbiased approach to seeking structural similarities between metal sites, independently of the user's prior knowledge. The hits returned by MetalS^3 can be a combination of relatively obvious ones (e.g., homologs of the query metalloprotein) and unexpected ones. The latter can be identified only through the present approach, whereas MetalS^2 is a tool to quantify structural similarities within groups of sites already familiar to the user.

The MetalS^3 approach may help researchers in the field of bioinorganic chemistry to assess the relationships or evaluate possible evolutionary links between different groups of metalloproteins and may help guide experimentalists' work in understanding the function of uncharacterized metalloproteins. Overall, this contributes to achieving a better comprehension of the role of metal ions in living systems.

Acknowledgments This work was supported by the Ministero Italiano dell'Università e della Ricerca through the FIRB project RBFR08WGXT 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 JJR, Williams RJP (2001) The biological chemistry of the elements: the inorganic chemistry of life. Oxford University Press, New York
2. Bertini I, Sigel A, Sigel H (2001) Handbook on metalloproteins. Dekker, New York
3. Bertini I, Gray HB, Stiefel EI, Valentine JS (2006) Biological inorganic chemistry. University Science Books, Sausalito
4. Bertini I, Rosato A (2003) Proc Natl Acad Sci USA 100:3601–3604
5. Hsin K, Sheng Y, Harding MM, Taylor P, Walkinshaw MD (2008) J Appl Crystallogr 41:963–968
6. Schnabl J, Suter P, Sigel RKO (2012) Nucleic Acids Res 40:D434–D438
7. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) Nucleic Acids Res 39:D392–D401
8. Andreini C, Bertini I, Rosato A (2009) Acc Chem Res 42:1471–1479
9. Andreini C, Bertini I, Rosato A (2004) Bioinformatics 20:1373–1380
10. Shu N, Zhou T, Hovmoller S (2008) Bioinformatics 24:775–782
11. Andreini C, Bertini I, Cavallaro G (2011) PLoS ONE 10:e26325

12. Andreini C, Cavallaro G, Rosato A, Valasatava Y (2013) J Chem Inf Model 53:3064–3075
13. Andreini C, Bertini I, Cavallaro G, Najmanovich RJ, Thornton JM (2009) J Mol Biol 388:356–380
14. Maret W, Li Y (2009) Chem Rev 109:4682–4707
15. Choi M, Davidson VL (2011) Metallomics 3:140–151
16. Banci L, Bertini I, Calderone V, Della Malva N, Felli IC, Neri S, Pavelkova A, Rosato A (2009) Biochem J 422:37–42
17. Bertini I, Fragai M, Luchinat C, Melikian M, Venturi C (2009) Chem Eur J 15:7842–7845
18. Andreini C, Cavallaro G, Lorenzini S, Rosato A (2013) Nucleic Acids Res 41:D312–D319
19. Fufezan C, Specht M (2009) BMC Bioinform 10:258
20. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) Nucleic Acids Res 40:D290–D301
21. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Nucleic Acids Res 34:D247–D251
22. Lancaster CR, Kroger A, Auer M, Michel H (1999) Nature 402:377–385
23. Hirata A, Klein BJ, Murakami KS (2008) Nature 451:851–854
24. Hirata A, Murakami KS (2009) Curr Opin Struct Biol 19:724–731
25. Groll M, Bajorek M, Kohler A, Moroder L, Rubin DM, Huber R, Glickman MH, Finley D (2000) Nat Struct Biol 7:1062–1067
26. Blamey CJ, Ceccarelli C, Naik UP, Bahnson BJ (2005) Protein Sci 14:1214–1221
27. Malmendal A, Linse S, Evenas J, Forsen S, Drakenberg T (1999) Biochemistry 38:11844–11850
28. Hastie KM, Kimberlin CR, Zandonatti MA, MacRae IJ, Saphire EO (2011) Proc Natl Acad Sci USA 108:2396–2401
29. Riboldi-Tunnicliffe A, Isaacs NW, Mitchell TJ (2005) FEBS Lett 579:5353–5360
30. Thoden JB, Ruzicka FJ, Frey PA, Rayment I, Holden HM (1997) Biochemistry 36:1212–1222
31. Geeganage S, Frey PA (1999) Biochemistry 38:13398–13406
32. Noble CG, Lim SP, Chen YL, Liew CW, Yap L, Lescar J, Shi PY (2013) J Virol 87:5291–5295
33. Yap TL, Xu T, Chen YL, Malet H, Egloff MP, Canard B, Vasudevan SG, Lescar J (2007) J Virol 81:4753–4765
34. Yasui N, Nogi T, Kitao T, Nakano Y, Hattori M, Takagi J (2007) Proc Natl Acad Sci USA 104:9988–9993
35. Yasui N, Nogi T, Takagi J (2010) Structure 18:320–331
36. Petrarca P, Ammendola S, Pasquali P, Battistoni A (2010) J Bacteriol 192:1553–1564
37. Ilari A, Alaleona F, Tria G, Petrarca P, Battistoni A, Zamparelli C, Verzili D, Falconi M, Chiancone E (2014) Biochim Biophys Acta 1840:535–544
38. David G, Blondeau K, Schiltz M, Penel S, Lewit-Bentley A (2003) J Biol Chem 278:43728–43735
39. Havemann J, Vogel D, Loll B, Keller U (2014) Chem Biol 21:146–155
40. Elkins JM, Rutledge PJ, Burzlaff NI, Clifton IJ, Adlington RM, Roach PL, Baldwin JE (2003) Org Biomol Chem 1:1455–1460
41. Roach PL, Clifton IJ, Fulop V, Harlos K, Barton GJ, Hajdu J, Andersson I, Schofield CJ, Baldwin JE (1995) Nature 375:700–704
42. Holm L, Sander C (1995) Trends Biochem Sci 20:478–480
43. Hopkinson RJ, Tumber A, Yapp C, Chowdhury R, Aik W, Che KH, Li XS, Kristensen JBL, King ONF, Chan MC, Yeoh KK, Choi H, Walport LJ, Thinnes CC, Bush JT, Lejeune C, Rydzik AM, Rose NR, Bagg EA, McDonough MA, Krojer TJ, Yue WW, Ng SS, Olsen L, Brennan PE, Oppermann U, Muller S, Klose RJ, Ratcliffe PJ, Schofield CJ, Kawamura A (2013) Chem Sci 4:3110–3117
44. Yu B, Edstrom WC, Benach J, Hamuro Y, Weber PC, Gibney BR, Hunt JF (2006) Nature 439:879–884
45. Yi C, Jia G, Hou G, Dai Q, Zhang W, Zheng G, Jian X, Yang CG, Cui Q, He C (2010) Nature 468:330–333

4.3. Hidden relationships between metalloproteins unveiled by structural comparison of their metal sites

Yana Valasatava[1], *Claudia Andreini* [1, 2], *and Antonio Rosato* [1, 2,*]

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Submitted

Abstract

Metalloproteins account for a substantial fraction of all proteins. They incorporate metal atoms, which are required for their structure and/or function. Here we describe a new computational protocol to systematically compare and classify metal-binding sites on the basis of their structural similarity. These sites are extracted from the MetalPDB database of minimal functional sites (MFSs) in metal-binding biological macromolecules. Structural similarity is measured by the scoring function of the available MetalS^2 program. Hierarchical clustering was used to organize MFSs into clusters, for each of which a representative MFS was identified. The comparison of all representative MFSs provided a thorough structure-based classification of the sites analyzed. As examples, the application of the proposed computational protocol to all heme-binding proteins and zinc-binding proteins of known structure highlighted the existence of structural subtypes, validated known evolutionary links and shed new light on the occurrence of similar sites in systems at different evolutionary distances. The present approach thus makes available an innovative viewpoint on metalloproteins, where the functionally crucial metal sites effectively lead the discovery of structural and functional relationships in a largely protein-independent manner.

Keywords: metalloprotein; metalloenzyme; bioinorganic chemistry; structural biology; cytochrome; heme; multiheme

Metal ions are bound to biological macromolecules via coordination bonds. The bonds are formed by the so-called donor atoms, which can belong to either the backbone or side chains/bases of the macromolecule (protein or nucleic acid). Additional donor atoms may belong to non-macromolecular ligands, such as oligopeptides, small organic molecules, anions, water molecules. The metal ion (or cluster of metal ions) together with its donor atoms constitute the metal-binding site. To achieve a satisfactory understanding of the biochemical properties of metal sites through the analysis of 3D structural features it is important to go beyond metal-binding sites by taking into account the surrounding macromolecular environment [1-6]. Altogether, this larger ensemble of atoms defines the minimal environment determining metal function, i.e. the "minimal functional site" (MFS). 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 (Supplementary Figure S1)[7]. 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, which is derived in an automated manner from the Protein Data Bank (PDB)[8], collects all known MFSs[9]. Recently, we have developed a computational approach, implemented in the MetalS^2 program, to quantify the structural similarity of MFSs in metalloproteins[10].

Structure-based as well as domain-based classifications of protein structures are well established. Resources such as CATH[11] or SCOP[12] are able to capture distant relationships between protein domains through the analysis of their 3D structures. They provide the notion of protein superfamily, which is the ensemble of all the protein domains with the same overall structural features. In MetalPDB we exploited such classifications to assign MFSs to so-called equistructural groups[9]. Such groups contain the MFSs that are found in proteins with the same fold and occur at the same position within that fold. This is evaluated 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. This approach is simple and intuitive, but can potentially overlook structural variations occurring within each metalloprotein family.

In this work we implemented and evaluated an approach based on the MetalS^2 program to perform systematic, quantitative comparisons of MFS structures with the final aim of producing a classification of metal sites. This is achieved by organizing 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 resulting classification is independent of the overall metalloprotein fold and can capture the fine structural variability of sites even within the same metalloprotein family. In addition, it provides unbiased indications on relationships between different metalloprotein families harboring the same metal cofactors. This contribution provides an unprecedented approach in bioinorganic structural biology that puts metal sites, the true center of research in bioinorganic chemistry, at the center of structural analysis. In fact, our new protocol innovatively recombines available algorithms to support out-of-the-box thinking about relationships among metalloproteins. The box we are referring to here is that constituted by the conventional tools based on global sequence or structural domain similarity. The present protocol is not meant to replace this kind of analysis, which has been successfully applied to metalloproteins[13-18], but to provide an additional, new tool to the portfolio of the structural biologist with an interest in bioinorganic chemistry that has been specifically designed for the specific challenges of the latter field of research.

We demonstrate the protocol using two test cases, namely heme-binding and zinc-binding proteins. Heme is one of the most abundant and widely used biological metalloporphyrins. As a protein cofactor, heme shuttles electrons between different redox centers in aerobic and anaerobic respiration as well as photosynthesis, or transports and stores O_2 as with the globins. Furthermore, numerous heme-dependent enzymes are known, which can catalyze both reductive and oxidative chemistry. MetalPDB shows that the iron coordination geometry in heme-containing MFSs is quite constant, being either square pyramidal or octahedral in the vast majority of cases, with four donor atoms out of a maximum of six provided by the porphyrin moiety. This makes it difficult to exploit the features of the iron coordination for functional or structural classification. Zinc proteins are one of the largest groups of metalloproteins within MetalPDB. Estimates of zinc proteomes in various organisms indicated that the amount of genes encoding zinc proteins varies from 4% to 10% of the genome [19,20]. Zinc enzymes in which zinc plays a catalytic role are present across all living organisms and constitute the largest share of prokaryotic zinc proteins. The main reason for the selection of zinc as a catalytic cofactor lies in its distinctive chemical properties, which combine Lewis acid strength, lack of redox reactivity, and fast ligand exchange [21]. The coordination geometry of the zinc(II) ion and the number of cysteine ligands can be quite informative on function, both for enzymes [7,22] and non-catalytic systems such as zinc fingers [23]. The application of our newly developed protocol can provide means to verify structure similarities beyond the first coordination sphere, and their relationship to functional properties.

Results

Analysis of equistructural groups of MFSs (first stage)

The present new computational protocol highlights local structure features that may distinguish members within a given metalloprotein family or reveal similarities across different families. To do so, the protocol leverages the organization of sites in equistructural groups (EGs hereafter) that is already provided by the MetalPDB database. These are groups of corresponding sites in the structures of metalloproteins belonging to the same family. Comparisons are first done within EGs, i.e. within metalloprotein families. Then representative MFSs are defined for the various structural subtypes occurring within a family. Finally, representative MFSs are exploited to systematically compare sites across different subtypes and, most importantly, across different metalloprotein families.

For heme-containing MFSs (hMFSs hereafter), we started from 187 EGs that had more than one member. The procedure yielded 344 clusters of hMFSs, of which 17 clusters did not contain hMFSs and thus were discarded. Our approach readily separated sites that bind individual metal ions from hMFSs, such as in the case of the EG containing the sites corresponding to the interfacial heme of bacterioferritins. This EG additionally includes various, possibly adventitious, sites from Dps-like proteins binding cations such as iron(II), copper(II), nickel(II). The complete separation achieved upon structural comparison of these two kinds of sites is not surprising given the difference in size and interactions with the protein of the cofactor. On the other hand, the Fe-coproporphyrin III site of *Desulfovibrio desulfuricans* bacterioferritin was clustered together with all other bacterioferritin hMFSs, in keeping with its structural and functional similarity to the typical heme site[24]. Another example is that of the separation of the interfacial hMFS of *Haemophilus ducreyi* superoxide dismutase[25] from adventitious metal sites in other superoxide dismutase structures. The 327 clusters that contained hMFSs (or other MFS binding heme analogs such as metal-substituted protoporphyrin IX, Supplementary Table S1) included 21 clusters with at least 100 sites, whereas 23 clusters contained a single hMFS. We manually inspected how the larger EGs were split into clusters. Typically, the clustering reflected defined structural features of the hMFSs. For example, in the EG corresponding to animal heme-dependent peroxidases, the two major clusters, which cumulatively accounted for 96% of the EG sites, contained myeloperoxidases together with lactoperoxidases (92 hMFSs), and prostaglandin synthases (113 hMFSs) (Figure 1). Another example is given by tryptophan 2,3 dioxygenases, which formed two clusters (18 and 23 members respectively) differing for the presence or absence of the substrate bound in the cavity (with one exception, Figure 2). In a few instances the clustering procedure generated an apparently too fine-grained separation of hMFSs. For example, the EG of cytochrome P450s, which contains 992 members, was split in as many as 22 clusters, containing 8 to 151 hMFSs. Here it is difficult to rationalize the outcome of the procedure as well as to correlate it to specific structural features. Notably, EGs including even more than 100 hMFSs constituted a single cluster when the structural similarity of the sites was sufficiently high; this was, for example, the case of the 531 hMFSs of mammalian nitric oxide synthases.

For zinc-binding MFSs (zMFSs hereafter), we started from 1752 EGs with more than one member (for a total of about 19,500 zMFss) and obtained 2263 clusters. In addition, 1640 zMFSs did not belong to any EG, and were carried on directly to the second stage of the procedure. 19 first-stage clusters included 100 sites or more. The largest cluster comprised all 335 zMFSs of the EG of alcohol dehydrogenases. As described above for hMFSs, in several cases EGs were split into two or more clusters. An interesting example is that of an EG containing 61 zMFSs from various aminoacyl-tRNA synthetases and closely related enzymes, which is gave rise to four distinct clusters. Among these, the two larger clusters contained respectively 28 and 29 sites, differing for the size and binding mode of the substrate analogues present in the structure (Figure 3).

We quantified the structural deviation within clusters by computing the root-mean-square-deviation (RMSD) of the Cα and Cβ atoms of the sites. We observed that the largest average RMSD within a cluster was of only 1.5 Å. Nearly 95% of the clusters had an average RMSD smaller than 1.0 Å and the median value for the average RMSD was 0.75 Å. The very high degree of structural similarity within clusters supports the usefulness of defining a single representative hMFS for each of them.

Comparison of representative MFSs (second stage)

In the second stage of our procedure we compared representative MFSs to one another, independently of EG assignments, thus avoiding possible biases due to domain assignments. We tried different clustering approaches (complete vs. average linkage) and different thresholds (T) to evaluate the stability of the outcome (note that a higher threshold indicates lower similarity).

Depending on the above factors, representatives hMFSs were grouped in a number of clusters ranging from 51 (average linkage clustering, T=3.5) to 199 (complete linkage clustering, T=2.25), whereas zMFSs were grouped in a number of clusters ranging from 840 (average linkage clustering, T=2.75) to 1661 (complete linkage clustering, T=2.25). Hereafter, we will use the following notation: CC or AC to indicate complete vs. average linkage, respectively, followed by the value of the threshold used (e.g. CC2.75 is the result of the clustering of representative hMFSs using complete linkage clustering and T= 2.75).

At the second stage of the computational procedure, there are three possible causes for representative MFSs to get clustered. (i) The first reason is that sites with very high structural similarity and found in different metalloprotein families are identified. (ii) The second cause becomes relevant when MetalPDB did not group metalloproteins of the same family, typically because of missing domain information, and consequently assigned them to different EGs. In this case, our second stage analysis puts together sites that should have been clustered already at the first stage, but actually were not compared because of the inconsistent EG assignments. (iii) The MFSs representing two clusters originating from the same EG may be regrouped because the distance between a pair of representative MFSs is shorter than the distance assigned by the CC algorithm to the corresponding clusters, as the latter equals the *largest* distance between any possible pair of cluster members. The representative MFS approximates a "central" position within the cluster it represents. This effectively reduces the distance between first stage clusters. It is possible to draw an analogy here to the use of consensus sequences to represent multiple sequence alignments, which hides some of the existing diversity. The aforementioned three causes may simultaneously concur to the formation of a second stage cluster of representative MFSs. The first and third causes should become more and more effective with reduced stringency of the clustering approach applied, whereas the relevance of the second cause is limited by the number of incomplete EG assignments and presumably declines, in relative terms, with increasing threshold.

The most stringent CC2.25 approach, which is the same approach implemented for the first stage clustering, yielded a total of 199 clusters out of 389 input hMFSs (327 representative hMFSs plus 62 singletons), each containing between 1 and 9 hMFSs. The largest clusters were formed by representative hMFSs belonging to the same EG that were re-grouped (reason iii), e.g. for some, but not all, representatives of cytochrome P450s. The representative hMFSs of tryptophan 2,3 dioxygenases (Figure 2) were also clustered together; in addition, the same cluster included the representative hMFS of the related indoleamine 2,3-dioxygenase. Example of clusters formed only at the second stage because of missing domain assignments (reason ii) in MetalPDB were that of the sirohemes in the catalytic sites of sulfite reductases, or of dye peroxidases (DyP). For the latter case, the appropriate domain is not identified within the sequence of DyP2 from *Amycolatopsis* sp. ATCC 39116 (PDB entry 4G2C [26]) but our approach correctly identified the similarity between DyP hMFSs. Finally, the cluster containing heme 4 of the cytochrome *c* subunit of *Rhodopseudomonas viridis* photosyntethic reaction center and the cysteine-coordinated heme of SoxA (heme 1263 in the 1H32 structure[27]) is an example of a cluster formed with CC2.25 for reason (i), i.e. because highly similar hMFS occurred in proteins with unrelated fold. With increasing threshold or passing from the CC to the AC approach, the number of clusters diminished as the reduced stringency allowed more dissimilar sites to be clustered together than for CC2.25. In particular, 110 clusters were formed with AC2.75 (Supplementary Table S2). We previously showed that 2.75 is a reasonable threshold for the MetalS^2 score[10] to identify meaningful structural similarities. At this level all cytochrome P450s were clustered together but one (PDB entry 3R9C [28]), due to the presence of a sodium(I) ion within the latter hMFS. Other metalloprotein families remained split even at this level, such as the family of ABM monooxygenases, which include various heme-degrading enzymes, reflecting their different modes or stoichiometries of heme binding[29]. When applying the AC2.75 approach, clusters formed with CC2.25 can merge. This occurred, for example, for the aforementioned cluster of tryptophan and indoleamine 2,3-dioxygenases, which additionally included the heme site of proteins related to PnrB, the second enzyme in the pyrrolnitrin biosynthesis pathway. Thus, our approach recomposed the full group of related dioxygenase folds, which eventually comprised proteins from three different EGs of MetalPDB.

For zMFSs, we analyzed in detail the output of the AC2.5 clustering, which provided 1083 clusters (of which 763 with more than one member; Supplementary Table S3). Our analysis focused instead on the ten largest clusters, which ranged in size between 25 and 382 members. The superpositions corresponding to two of these clusters are shown in Figure 4. In Figure 4A, cluster 820 encompasses 66 representative zMFSs of different types of related peptidases, largely from the metallopeptidase MA clan [30]. The cluster further includes the active sites of the anthrax toxin lethal factor[31] and, curiously, the zinc-substituted catalytic site of iron-dependent tyrosine 3-monooxygenase (PDB ID 2XSN, unpublished). The superposition clearly reveals that the local structural similarity extends to the region of substrate binding. Figure 4B instead refers to cluster 193 (31 members), which mainly includes zinc-finger-type sites from a variety of systems. These zMFSs are identified in proteins from prokaryotic as well as eukaryotic organisms and their functional role has not always been ascertained. Whereas interaction with DNA seems the most obvious role[32], also because the majority of these systems are involved in DNA recognition and/or modification and repair, there are other possibilities, such as ubiquitin-binding[33]. In previous articles, the zMFS of Figure 4B has been described as unique to a specific system[34] or not relevant to function[35]. Instead, the present data show that it is relatively widespread and thus likely to have functional relevance. This highlights the usefulness of the present approach as a knowledge discovery tool in bioinorganic chemistry. Finally, cluster 877 (Figure 5) contains 25 zMFSs from enzymes, mostly di-nuclear metal sites formed by zinc(II) and another divalent cation. The zinc ion is the catalytic center of these enzymes, whereas the second metal ion might be bound to the substrate (e.g. Mg- cytidine diphosphate for 2C-methyl-d-erythritol-2,4-cyclodiphosphate synthase[36]) or can be bound to the protein independently of the presence of substrate/cofactors (e.g. Mn(II) in yeast Pop2p[37]). The site is found either in 2C-methyl-d-erythritol-2,4-cyclodiphosphate synthases or in DNA polymerases with exonuclease activity as well as other nucleases (Figure 5). These groups of enzymes share a similar architecture but different topologies, according to the CATH[38] classification. Intriguingly, despite the different fold, the substrate binding site is closely located in these two groups. The same zMFS is exploited to perform a phosphorus-oxygen lyase reaction by the synthases, with respect to the hydrolysis of a phosphodiester bond in the nucleases.

A detailed analysis of multiheme c-type cytochromes

Multiheme *c*-type cytochromes (MHCs), which are proteins that bind several heme groups to a single polypeptide chain via a pair of thioether bonds, are of particular interest in the context of the present work. For these systems fold assignments tend to be less informative, also because their 3D structure is largely determined by cofactor-protein hydrophobic interactions rather than by protein-protein interactions in the hydrophobic core[39]. Our protocol provided a complete picture of structural similarities among the various hMFSs contained in MHCs, from di-heme to sixteen-heme proteins (Figure 4). It is possible to immediately identify two major blocks of related MHCs, namely those linked to (or, in evolutionary terms, presumably derived from) the four hMFSs of the tetra-heme cytochrome c_3 and those linked to the sites of NrfA. The first block includes cytochrome c_3, cytochrome c_7, nona-, dodeca- and exadeca-heme cytocromes. In the first block, all hMFSs can be related to one of the hMFS of cytochrome c_3, with two exceptions. One is a unique site present in nonaheme cytochromes that acts as a connector between two cytochrome *c_3* domains[40]. A search of the MetalPDB database using this site as input to the MetalS^3 search tool [6] revealed a weak similarity to one hMFS of NrfB (not shown). The other exception was within the structure of dodecaheme cytochromes, which have been described as a combination of four cytochrome *c_7* domains[41]. Our analysis indicated that this is true for two out of three hMFSs, whereas the other hMFS is structurally diverse and gave rise to a separate cluster (Figure 6 and Supplementary Figure S2). The second block includes NrfA (a five-heme nitrite reductase), NrfB (a five-heme electron donor to NrfA), eight-heme nitrite reductase, hydroxylamine oxidase (a eight-heme enzyme), tetrathionate reductase (a eight-heme enzyme) and tetra-heme cytochrome *c_{554}*. Here all hMFSs can be related to one of the sites of NrfA, with one or two specific exceptions for NrfB as well as the various eight-heme enzymes. Furthermore, we indentified a tight relationship between the hMFSs of two of the simplest MHCs, namely the di-heme proteins NapB, a subunit of periplasmic nitrate reductase, and *Geobacter sulfurreducens* DHC2 cytochrome *c*. The analysis summarized by Figure 4 provides an objective guidance to comparison at the whole structure level for pairs of MHCs with different folds. Indeed, after superposition of the hMFSs of the two proteins contained in the same clusters MetalS^2 provides roto-traslational matrices that can be applied to the entire structure. Cluster assignments indicate how to combine various hMFSs to obtain a single overall matrix that yields a best fit for all of them simultaneously. The global structural superposition obtained in this way can indicate relationships also between sites not clustered together, based on the spatial proximity of the heme groups (Supplementary Figure S3). As an example, Figure 7 provides an overview of the hMFS correspondences obtained by superposing various MHCs to the structure of eight-heme nitrite reductase (PDB entry 3GM6 [42]) as indicated above. The known [43] relationships between the sites of these proteins are independently re-discovered. Notably, the catalytic sites of nitrite reductase, hydroxylamine oxidase, NrfA and cytochrome *c_{554}* are related by spatial proximity after superposition in addition to their belonging to the same cluster. For cytochrome *c_{554}*, a NO reducing activity has been reported [44]; its structural correspondence to hydroxylamine oxidase, including the then unknown catalytic site, had already been highlighted [45]. A less obvious relationship is that between three sites of fumarate reductase and three sites of the small tetraheme cytochrome *c* from S*hewanella*. (Figure 7)

Discussion

In this work, we developed a methodology to perform a systematic comparison based on structural similarity of metal sites extracted from metalloproteins. Our definition of metal site extended beyond the metal ion and its aminoacidic ligands by involving all the chemical species (aminoacids, nucleotides, exogenous ligands) containing at least one donor atom (shown in blue in Supplementary Figure S1) as well as all any other chemical species within a radius of 5.0 Å (shown in green in Supplementary Figure S1). We previously defined this as the minimal functional site of a metalloprotein (MFS), and showed that its characteristics are related to the metalloprotein function[3,7]. The present methodology leverages the MetalS^2 algorithm, whose total score provides a quantitative measure of structural similarity between pairs of MFSs[10]. We used this measure to build clusters of structurally similar MFSs using a hierarchical clustering algorithm. The proposed computational strategy is a two-stage procedure, mainly for the sake of simplicity and calculation speed. In the first stage, predefined groups of MFSs contained in corresponding regions of metalloproteins having the same fold (equistructural groups, EGs) are retrieved from the MetalPDB database[9]. Then, all MFSs in each EG are systematically compared to one another. After the application of a complete linkage clustering algorithm with a very restrictive threshold (2.25) each EG gave rise to one or more clusters characterized by a low degree of internal structural variability (less than 1 Å backbone RMSD in more than 90% of the cases). The different clusters resulting from a given EG provide a thorough view of homogeneous structural features across the members of the group. Because each EG corresponds to a specific metalloprotein family, the first stage clusters recapitulate systematically the known structural variants of the metal-binding site of that family. These variants can be associated to biochemical events such as ligand binding (Figure 2 and Figure 3) or reflect the structural features of different subfamilies (Figure 1). The low structural variability within clusters enabled us to meaningfully define a single representative MFS for each cluster.

Representative MFSs allow the comparison of the sites of different metalloprotein families (second stage clustering), at the level of their structural subtype, in an innovative manner that is independent of the global sequence or structural similarity of the metalloproteins containing the MFSs. Indeed, the clusters obtained after the second stage often grouped MFSs from metalloproteins with different but related folds (e.g. as defined by so-called clans in the Pfam database of domains[46]). This supports the idea that the 3D structures of the whole metalloprotein and of its metal site differentiate at comparable rates. The detection of structural similarity between MFSs can thus be taken as good an indication of homology as overall structural similarity is for proteins not binding metal cofactors. This result provides also a means to assign potential biological functions to the so-called domains of unknown function, when they contain MFSs structurally similar to sites of functionally characterized metalloproteins. Finally, discovering structural similarities among representative MFSs also allows establishing relationships involving completely unrelated protein domains.

We demonstrated a practical implementation of the proposed procedure for heme-binding proteins as well as for zinc-binding proteins. The unique usefulness of the present tool resides in its capability to address comprehensively relationships among different metalloprotein families, i.e. in systems with different folds. As observed for MHCs, such relationships can be related to evolutionary patterns (Figure 6) but can also correct or shed a different light on previously proposed such patterns (Supplementary Figure S2). Furthermore, our approach identified common occurrences of zinc-binding sites across different protein folds, showing how the same local structure is harnessed by different systems to perform different metal-based catalysis (Figure 5).

In conclusion, we showed here for the first time that the structures of MFS, i.e. of small portions of the larger 3D structures of metalloproteins and metalloenzymes centered around the metal cofactor, can be systematically compared and clustered to obtain useful insight into the structural, functional and evolutionary features of metalloproteins. This kind of analysis complements the information that can be gained through more conventional approaches, such as sequence or fold comparison[13-18]. The present protocol constitutes a unique, innovative tool in the portfolio of computational tools of bioinorganic chemists. Its unicity stems from the concept of centering structural comparisons at the metal center itself, which is crucial to define the cellular role of metal-binding proteins. By performing comparisons at the level of the whole MetalPDB database, users can achieve a systematic view of metalloproteins based on the structural properties of the metal-sites rather than on the structural properties of the protein fold in which the site is embedded, as afforded by currently available approaches. This is a dramatically different viewpoint on metalloproteins, which only now becomes available.

Materials and Methods

Background

In our previous work[9] we organized MFSs into groups of equistructural sites. Such sites are extracted from metal-binding polypeptide chains that have similar fold, using the approach summarized below. After superimposing all the chains with the same fold, the distance between the metal ions (or the geometric center of all metal ions for polymetallic cofactors) is measured. MFSs whose metal ions are separated by a distance shorter than a predefined threshold (3.5 Å) are put in the same group, regardless of the chemical identity of the ions. This leads to e.g. all sites of the same metalloprotein after different metal replacement experiments belonging to the same equistructural group. Broadly speaking, the condition described above identifies sites that occupy the same location within a given protein fold. At the computational level, a single linkage clustering approach has been implemented to build the groups. A practical implication of this is that for any given MFS in a group the aforementioned condition will be fulfilled by at least another group member, but not necessarily by all. By construction, the structural similarity that is described by equistructural groups is mainly the result of overall fold similarity. Conversely, structurally similar MFSs that are bound to proteins with different fold were associated with different equistructural groups. Here, we combine the use of our MetalS^2 algorithm, which provides a quantitative approach to the structural comparison of pairs of MFSs[10], with a hierarchical clustering method to cluster MFS structures independently of the overall metalloprotein fold.

Datasets used

The datasets used for this study consist of the three-dimensional structures of all MFSs present in the MetalPDB database (http://metalweb.cerm.unifi.it/) as of April 2014 that were members of an equistructural group containing at least one heme-binding site or at least one zinc ion.

The number of heme sites in the dataset was 8891, separated into 249 EGs. Of these, 14 contain at least 100 members, with the largest one having more than 2000, whereas 62 are singletons, i.e. contain only one site. To achieve the greatest coverage and potentially gain more information, the above included also sites that harbor chemically or biosynthetically modified heme cofactors as well as inorganic complexes mimicking the heme moiety (Supplementary Table S1).

For zinc-binding sites, we firstly removed all sites with less than 10 amino acids as well as all sites where the zinc ion had only aminoacidic ligand will all other ligands being water molecules. The number of zinc sites was 21483, of which we kept 20478. After the first stage, we obtained 2263 clusters, plus 1640 singletons.

Clustering procedure

Our procedure was based on a hierarchical agglomerative clustering algorithm[47]. 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 operative distance measure adopted was the global MetalS^2 score, which increases with increasing structural diversity[10]. 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 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. 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. We used the complete and average linkage methods. 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 other words, the distance between clusters C_i and C_j is defined as

$$d_c(C_i, C_j) = \max_{k \in C_i, l \in C_j} d(k,l)$$

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 distance for the average linkage is defined as

$$d_c(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{k \in C_i, l \in C_j} d(k,l)$$

where $|C_i|$ and $|C_j|$ and are the numbers of members in the clusters C_i and C_j correspondingly.

In both formulas *k* and *l* refer to members of the clusters C_i and C_j, *d(k,l)* is the distance between the *k*-th member and *l*-th member of, respectively, C_i and C_j (in practice the global MetalS^2 score between the *k*-th and *l*-th MFSs). The minimum distance $d_c(C_i,C_j)$ among all the intra-cluster distances determines which pair of clusters is merged.

The clustering results are influenced by the linkage type applied. Complete linkage tends to produce clusters that are more compact (tight) with respect to clusters produced by average linkage. When a cut-off value of a similarity measure is applied in order to determine the final partition, the clusters produced by the average linkage method allows some within-cluster distances to exceed the cut-off value whereas the complete linkage method ensures that no within-cluster distance exceeds the cut-off. As a result, the complete linkage approach produces a higher number of more robust clusters while with average linkage the number of clusters is lower but within-cluster variability is higher. One of the weaknesses of the complete linkage method is its sensitivity to outliers, i.e. members that do not fit well into the global structure of the cluster. Such sensitivity may prevent the identification of even intuitive clusters, as outliers may pull similar members into different groups.

For the analysis of our dataset, we used the algorithm described above within a multi-step procedure, which included: (i) dividing existing equistructural groups into smaller clusters (first, intra-group stage); (ii) defining a representative MFS for each cluster; (iii) building broader clusters by comparing the representative MFSs from clusters built at the first level (second, inter-group stage).

First stage

This stage of analysis is designed to capture the structural variations possibly occurring among the MFSs in each group of equistructural sites of MetalPDB. For each group we systematically compared all possible pairs of MFSs, using the MetalS^2 algorithm[10]. The result was a matrix of all-versus-all comparison scores for each group. The matrix was then used as the input to perform hierarchical clustering within the equistructural group, applying a cut-off value of 2.25 for the MetalS^2 similarity score to build the clusters. At this stage we applied a complete linkage clustering approach, so that two MFSs whose structural superposition results in a MetalS^2 score greater than 2.25 are always associated to different clusters. As we described previously[10], the 2.25 threshold is quite stringent, i.e. corresponds to a high level of structural similarity. The first-stage procedure thus resulted in a fine-grained clustering of each EG, which highlighted intra-group structural variations.

For each cluster obtained that contained more than one MFS, we defined a single representative MFS as the most similar on average to all other members of the cluster. In practice the representative MFS was defined as the MFS minimizing the sum of the MetalS^2 global scores resulting from its pairwise comparisons to all other MFSs in the same cluster. All the MFSs that did not cluster at the first stage or that formed an equistructural group (singleton) by themselves were taken as a representative. Clusters that did not contain heme-binding sites were removed, together with their corresponding representatives.

Second stage

The second stage of comparison aims to obtain a set of clusters, each representing a distinct MFS shape, independently of overall protein fold. The dataset used for this analysis included all representative structures of MFSs from the first level clustering. Similarly to the first stage procedure, we generated a single all-versus-all similarity matrix. Both complete and average linkage clustering algorithms were then applied to generate clusters at this stage. Different cut-off values, from 2.25 to 3.5, were tested.

Multi-heme c-type cytochromes

To investigate the full network of structural relationships across multi-heme *c*-type cytochromes (MHCs) we compiled a list of all clusters that included sites from different MHCs (we excluded MHCs containing multiple single-heme mitochondrial-type cytochrome c domains [14]). For each protein we then added the clusters containing only hMFSs specific to it in order to cover its entire set of sites (either common to other MHCs or unique to that MHC). This was done for all proteins in the MHC list, so that the set of selected clusters eventually contained all MHC sites present in the MetalPDB database.

References

Reference List

1. Lin,I.J., Gebel,E.B., Machonkin,T.E., Westler,W.M. & Markley,J.L. Changes in hydrogen-bond strenght explain reduction potentials in 10 ruvredoxin variants. *Proc. Natl Acad. Sci. , USA* **102**, 14581-14586 (2005).

2. Dey,A. *et al.* Solvent tuning of electrochemical potentials in the active sites of HiPIP versus ferredoxin. *Science* **318**, 1464-1468 (2007).

3. 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.* **388**, 356-380 (2009).

4. Lee,Y.M. & Lim,C. Factors Controlling the Reactivity of Zinc Finger Cores. *J. Am. Chem. Soc.* **133**, 8691-8703 (2011).

5. Dudev,T. & Lim,C. Competition among metal ions for protein binding sites: determinants of metal ion selectivity in proteins. *Chem. Rev.* **114**, 538-556 (2014).

6. Valasatava,Y., Rosato,A., Cavallaro,G. & Andreini,C. MetalS^3, a database-mining tool for the identification of structurally similar metal sites. *J. Biol. Inorg. Chem.* **19**, 937-945 (2014).

7. Andreini,C., Bertini,I. & Cavallaro,G. Minimal functional sites allow a classification of zinc sites in proteins. *Plos ONE* **10**, e26325 (2011).

8. Berman,H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).

9. Andreini,C., Cavallaro,G., Lorenzini,S. & Rosato,A. MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **41**, D312-D319 (2013).

10. Andreini,C., Cavallaro,G., Rosato,A. & Valasatava,Y. MetalS^2: a tool for the structural alignment of minimal functional sites in metal-binding proteins and nucleic acids. *J. Chem. Inf. Model.* **53**, 3064-3075 (2013).

11. Sillitoe,I. *et al.* New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* **41**, D490-D498 (2013).

12. Andreeva,A., Howorth,D., Chothia,C., Kulesha,E. & Murzin,A.G. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* **42**, D310-D314 (2014).

13. Bertini,I., Luchinat,C., Provenzani,A., Rosato,A. & Vasos,P.R. Browsing gene banks for Fe_2S_2 ferredoxins and structural modeling of 87 plant-type sequences: an analysis of fold and function. *Proteins Struct. Funct. Genet.* **46**, 110-127 (2002).

14. Bertini,I., Cavallaro,G. & Rosato,A. Cytochrome c: occurrence and functions. *Chem. Rev.* **106**, 90-115 (2006).

15. Zambelli,B., Musiani,F., Savini,M., Tucker,P. & Ciurli,S. Biochemical studies on Mycobacterium tuberculosis UreG and comparative modeling reveal structural and functional conservation among the bacterial UreG family. *Biochemistry* **46**, 3171-3182 (2007).

16. Andreini,C., Bertini,I. & Rosato,A. Metalloproteomes: a bioinformatic approach. *Acc. Chem. Res.* **42**, 1471-1479 (2009).

17. Zhang,Y. & Gladyshev,V.N. Comparative genomics of trace element dependence in biology. *J. Biol. Chem.* **286**, 23623-23629 (2011).

18. Gladyshev,V.N. & Zhang,Y. Comparative genomics analysis of the metallomes. *Met. Ions. Life Sci.* **12:529-80. doi: 10.1007/978-94-007-5561-1_16.**, 529-580 (2013).

19. Andreini,C., Banci,L., Bertini,I. & Rosato,A. Zinc through the three domains of life. *J. Proteome Res.* **5**, 3173-3178 (2006).

20. Andreini,C., Banci,L., Bertini,I. & Rosato,A. Counting the zinc proteins encoded in the human genome. *J. Proteome Res.* **5**, 196-201 (2006).

21. Andreini,C., Bertini,I., Cavallaro,G., Holliday,G.L. & Thornton,J.M. Metal ions in biological catalysis: from enzyme databases to general principles. *J. Biol. Inorg. Chem.* **13**, 1205-1218 (2008).

22. Lee,Y.M. & Lim,C. Physical basis of structural and catalytic Zn-binding sites in proteins. *J. Mol. Biol.* **379**, 545-553 (2008).

23. Lee,S.J. & Michel,S.L. Structural metal sites in nonclassical zinc finger proteins involved in transcriptional and translational regulation. *Acc. Chem. Res.* **47**, 2643-2650 (2014).

24. Macedo,S. *et al.* The nature of the di-iron site in the bacterioferritin from Desulfovibrio desulfuricans. *Nat. Struct. Biol.* **10**, 285-290 (2003).

25. Toro,I. *et al.* Structural basis of heme binding in the Cu,Zn superoxide dismutase from Haemophilus ducreyi. *J. Mol. Biol.* **386**, 406-418 (2009).

26. Brown,M.E., Barros,T. & Chang,M.C. Identification and characterization of a multifunctional dye peroxidase from a lignin-reactive bacterium. *ACS Chem. Biol.* **7**, 2074-2081 (2012).

27. Bamford,V.A. *et al.* Structural basis for the oxidation of thiosulfate by a sulfur cycle enzyme. *EMBO J* **21**, 5599-5610 (2002).

28. Agnew,C.R. *et al.* An enlarged, adaptable active site in CYP164 family P450 enzymes, the sole P450 in Mycobacterium leprae. *Antimicrob. Agents Chemother.* **56**, 391-402 (2012).

29. Contreras,H., Chim,N., Credali,A. & Goulding,C.W. Heme uptake in bacterial pathogens. *Curr. Opin. Chem. Biol.* **19**, 34-41 (2014).

30. Rawlings,N.D., Morton,F.R., Kok,C.Y., Kong,J. & Barrett,A.J. MEROPS: the peptidase database. *Nucleic Acids Res.* **36**, D320-D325 (2008).

31. Panchal,R.G. *et al.* Identification of small molecule inhibitors of anthrax lethal factor. *Nat. Struct. Mol. Biol.* **11**, 67-72 (2004).

32. Zhu,L. *et al.* Sexual dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. *Genes Dev.* **14**, 1750-1764 (2000).

33. Page,A.N., George,N.P., Marceau,A.H., Cox,M.M. & Keck,J.L. Structure and biochemical activities of Escherichia coli MgsA. *J. Biol. Chem.* **286**, 12075-12085 (2011).

34. Bernstein,D.A., Zittel,M.C. & Keck,J.L. High-resolution structure of the E.coli RecQ helicase catalytic core. *EMBO J.* **22**, 4910-4921 (2003).

35. Grant,R.P., Buttery,S.M., Ekman,G.C., Roberts,T.M. & Stewart,M. Structure of MFP2 and its function in enhancing MSP polymerization in Ascaris sperm amoeboid motility. *J. Mol. Biol.* **347**, 583-595 (2005).

36. Steinbacher,S. *et al.* Structure of 2C-methyl-d-erythritol-2,4-cyclodiphosphate synthase involved in mevalonate-independent biosynthesis of isoprenoids. *J. Mol. Biol.* **316**, 79-88 (2002).

37. Andersen,K.R., Jonstrup,A.T., Van,L.B. & Brodersen,D.E. The activity and selectivity of fission yeast Pop2p are affected by a high affinity for Zn2+ and Mn2+ in the active site. *RNA.* **15**, 850-861 (2009).

38. Sillitoe,I. *et al.* New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* **41**, D490-D498 (2013).

39. Sharma,S., Cavallaro,G. & Rosato,A. A systematic investigation of multi-heme c-type cytochromes in prokaryotes. *J. Biol. Inorg. Chem.* **15**, 559-571 (2010).

40. Umhau,S. *et al.* Three-dimensional structure of the nonaheme cytochrome c from Desulfovibrio desulfuricans Essex in the Fe(III) state at 1.89 A resolution. *Biochemistry* **40**, 1308-1316 (2001).

41. Pokkuluri,P.R. *et al.* Structure of a novel dodecaheme cytochrome c from Geobacter sulfurreducens reveals an extended 12 nm protein with interacting hemes. *J. Struct. Biol.* **174**, 223-233 (2011).

42. Polyakov,K.M. *et al.* High-resolution structural analysis of a novel octaheme cytochrome c nitrite reductase from the haloalkaliphilic bacterium Thioalkalivibrio nitratireducens. *J. Mol. Biol.* **389**, 846-862 (2009).

43. Tikhonova,T.V., Trofimov,A.A. & Popov,V.O. Octaheme nitrite reductases: structure and properties. *Biochemistry (Mosc.)* **77**, 1129-1138 (2012).

44. Upadhyay,A.K., Hooper,A.B. & Hendrich,M.P. NO reductase activity of the tetraheme cytochrome C554 of Nitrosomonas europaea. *J. Am. Chem. Soc.* **128**, 4330-4337 (2006).

45. Iverson,T.M. *et al.* Heme packing motifs revealed by the crystal structure of the tetra-heme cytochrome c554 from Nitrosomonas europaea. *Nat. Struct. Biol.* **5**, 1005-1012 (1998).

46. Finn,R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222-D230 (2014).

47. Hastie,T., Tibshirani,R. & Friedman,J. The Elements of Statistical Learning. Springer, New York (2009).

Acknowledgements

This work was supported by MIUR (Ministero Italiano dell'Università e della Ricerca) through the FIRB projects RBFR08WGXT and by the European Commission through the BioMedBridges project (grant no 284209).

Author contributions

AR and CA designed research. YV and CA carried out research. AR wrote the main manuscript text. All authors prepared the figures and reviewed the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figure Legends

Figure 1. Comparison of the structures of the hMFSs contained in the two major clusters originating from the equistructural group of animal heme-dependent peroxidases. Left: myeloperoxidases and lactoperoxidases; right: prostaglandin synthases.

Figure 2. Comparison of the structures of the hMFSs contained in the two clusters originating from the equistructural group of tryptophan 2,3 dioxygenases. In the most populated cluster a molecule of substrate (L-tryptophan) is contained in the enzyme cavity (left).

Figure 3. Comparison of the structures of the zMFSs contained in the two largest clusters originating from the equistructural group of aminoacyl-tRNA synthetases and closely related enzymes. The zMFSs in the two clusters differ because of the size and binding mode of their organic ligands.

Figure 4. Example clusters of representative zMFSs. A) superposition of 66 representative zMFSs of different related metallopeptidases; the common position for substrate binding, as indicated by the binding of ligands (hidden for clarity) in the 3D structures of the cluster, faces the reader; B) superposition of 31 representative zMFSs of non-standard zinc fingers.

Figure 5. A cluster formed by zMFS from DNA polymerases with nuclease activity and 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthases. The ligands or substrates present in the structures are also shown. The right panel depicts a selection of two synthases (blue structures), two exonuclease sites of polymerases (yellow structures) and of the fission yeast Pop2p exonuclease (orange).

Figure 6. Structural relationships between the hMFSs of multi-heme cytochromes. The number of hMFSs for a given MHC (rows) included in a given cluster (columns) is reported. Each column corresponds to a cluster of Supplementary Table S2. Each row corresponds to a different MHC family. The first column reports the PDB entry corresponding to the structure of a typical member of the family (not necessarily the one from which representative hMFSs are derived). The last column reports the number of hemes in the MHC. The last row reports the number of hemes in each cluster.

Figure 7. hMFS relationships in eight-heme, NrfA and related MHCs derived from cluster-guided structural superpositions. Top: Superpositions of the heme groups of selected MHCs resulting from the simultaneous overlay of the protein part of the hMFSs of each MHC to the sites of eight-heme nitrite reductase (PDB entry 3GM6) belonging to the same AC2.75 clusters (2K3V, cyan; 2P0B, magenta; 2CZS, gray; 3ML1, dark green; 3GM6, blue; 1QDB, light green; 1FGJ, yellow; 1BVB, red; 1SP3, orange; 1Q9I, brown). Residue numbering for the heme groups is shown for structure 3GM6. Bottom: summary of the relationships, color coded according to the cluster assignments of Figure 4 (green: cluster 65; blue: cluster 68; magenta: cluster 64; yellow, cluster 55; pink, cluster 61). Heme sites are labeled by their residue numbers in the PDB structure. Relationships are derived from spatial proximity after superposition and all refer to the sites of nitrite reductase. Only clusters containing hMFSs from different MHCs have been highlighted. A star indicates sites that fulfill the requirement of spatial proximity but are not satisfactorily superimposed (e.g. iron ligands do not overlay or the heme orientation is somewhat different). The heme groups 802 of structure 1SP3 and 804 of structure 1Q9I have been omitted for clarity. The figure independently re-discovers the known [43] relationships between between hemes I-VIII of nitrite reductase and hemes I-VIII of hydroxylamine oxidase, between hemes IV-VIII of nitrite reductase and hemes I-V of NrfA, or between seven out of the eight groups of nitrite reductase and tetrathionate reductase.

Figure 1. Comparison of the structures of the hMFSs contained in the two major clusters originating from the equistructural group of animal heme-dependent peroxidases. Left: myeloperoxidases and lactoperoxidases; right: prostaglandin synthases.

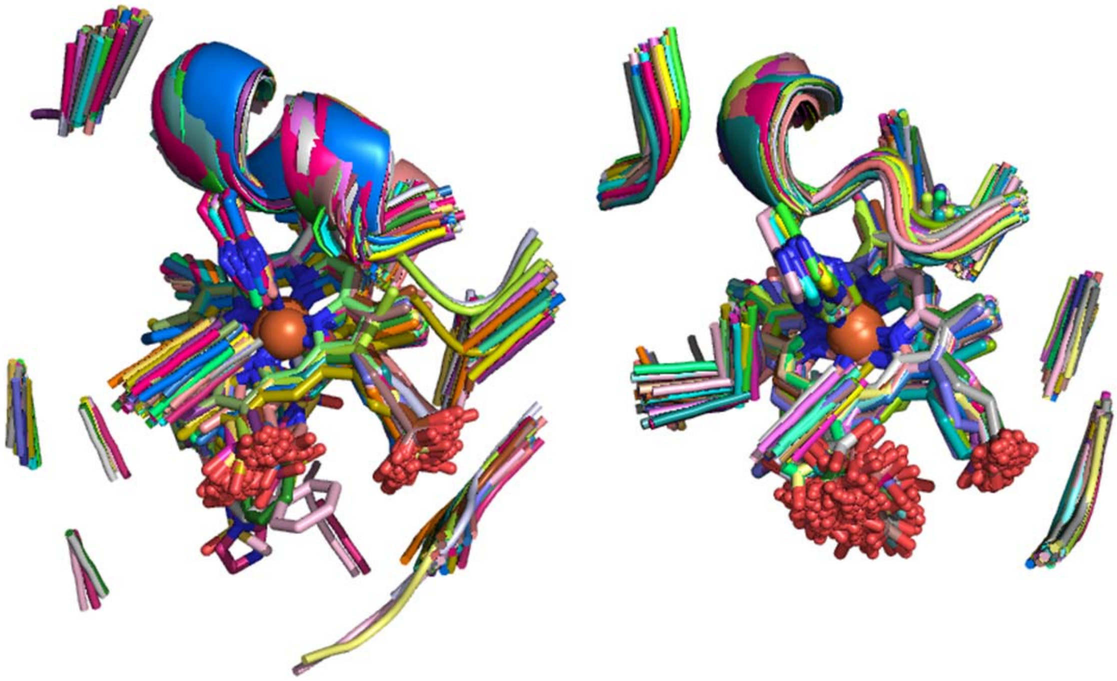


Figure 2. Comparison of the structures of the hMFSs contained in the two clusters originating from the equistructural group of tryptophan 2,3 dioxygenases. In the most populated cluster a molecule of substrate (L-tryptophan) is contained in the enzyme cavity (left).

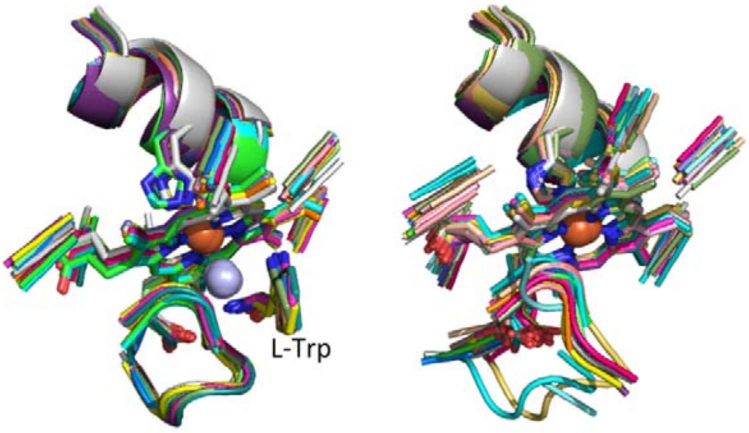


Figure 3. Comparison of the structures of the zMFSs contained in the two largest clusters originating from the equistructural group of aminoacyl-tRNA synthetases and closely related enzymes. The zMFSs in the two clusters differ because of the size and binding mode of their organic ligands.

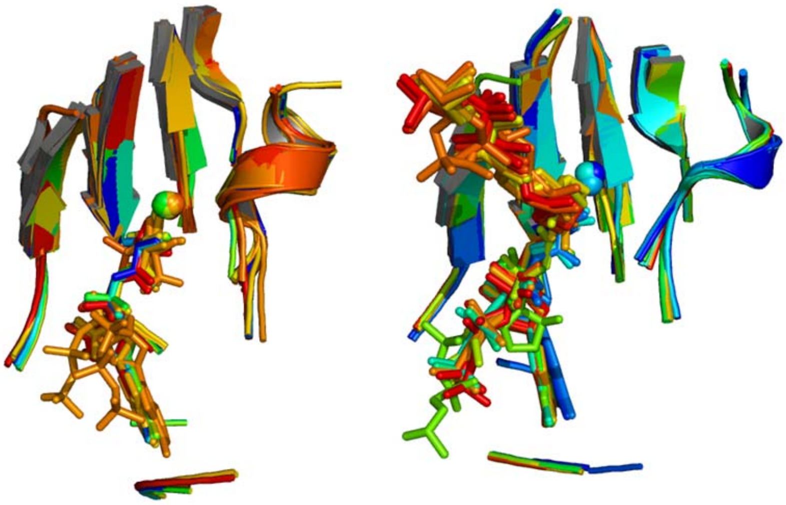


Figure 4. Example clusters of representative zMFSs. A) superposition of 66 representative zMFSs of different related metallopeptidases; the common position for substrate binding, as indicated by the binding of ligands (hidden for clarity) in the 3D structures of the cluster, faces the reader; B) superposition of 31 representative zMFSs of non-standard zinc fingers.

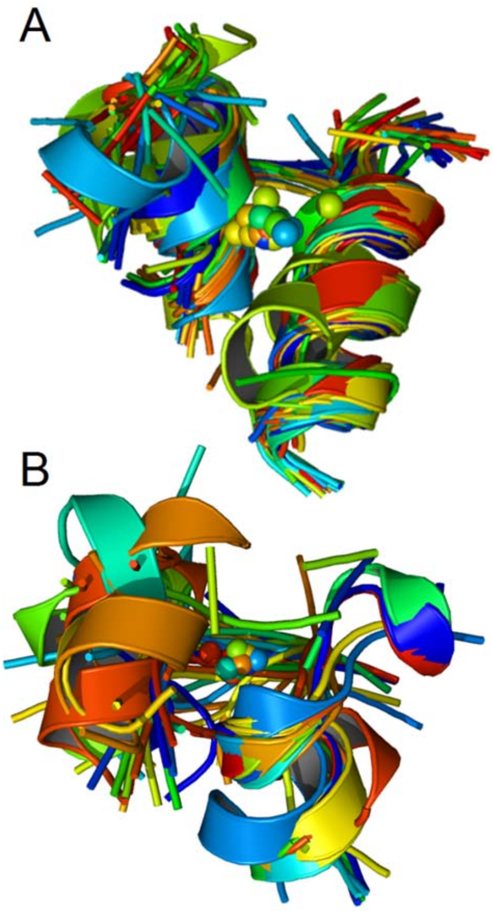


Figure 5. A cluster formed by zMFS from DNA polymerases with nuclease activity and 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthases. The ligands or substrates present in the structures are also shown. The right panel depicts a selection of two synthases (blue structures), two exonuclease sites of polymerases (yellow structures) and of the fission yeast Pop2p exonuclease (orange).

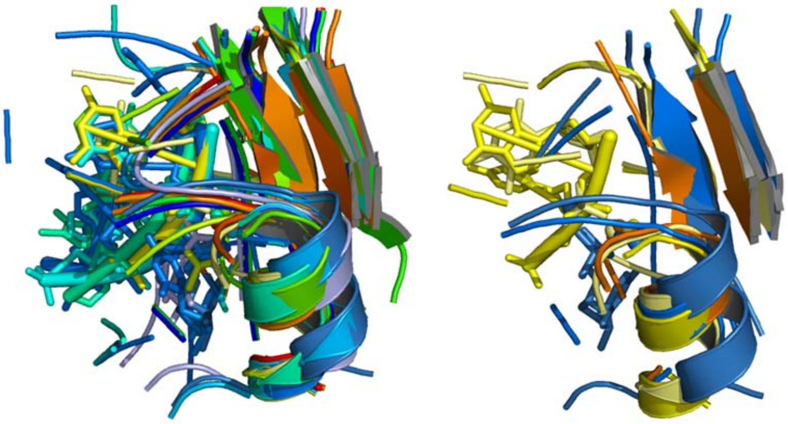
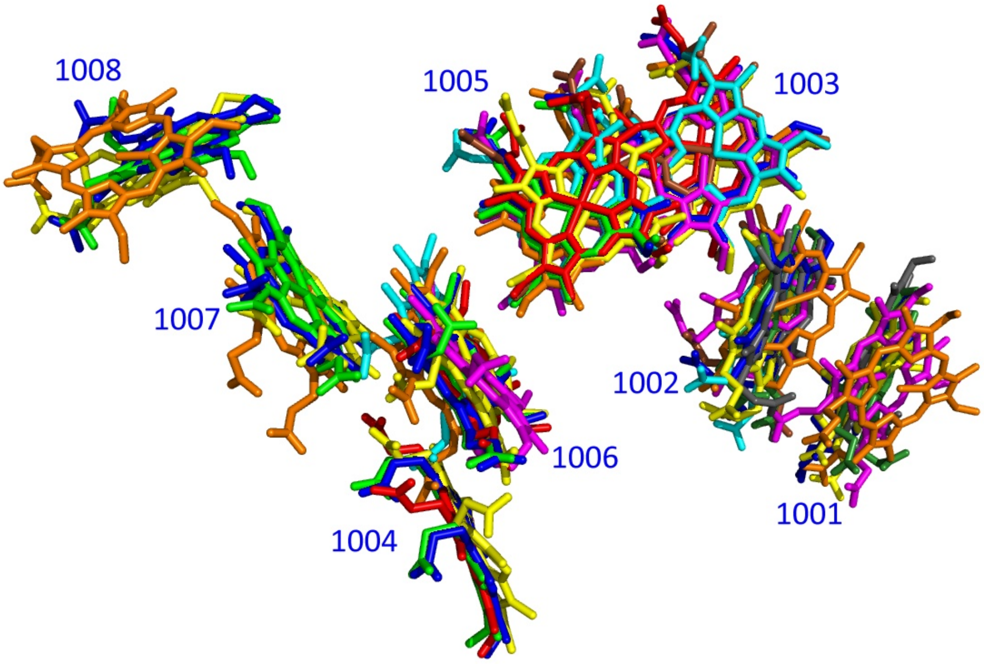


Figure 6. Structural relationships between the hMFSs of multi-heme cytochromes. The number of hMFSs for a given MHC (rows) included in a given cluster (columns) is reported. Each column corresponds to a cluster of Supplementary Table S2. Each row corresponds to a different MHC family. The first column reports the PDB entry corresponding to the structure of a typical member of the family (not necessarily the one from which representative hMFSs are derived). The last column reports the number of hemes in the MHC. The last row reports the number of hemes in each cluster.

	60	63	66	70	69	84	53	54	59	73	65	68	64	55	74	61	58	67	83	71	57	75	105	108	
1GWS	4	4	4	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**Exadecaheme cytochrome *c***
1DUW	2	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**Nonaheme cytochrome c**
1UP9	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**Cytochrome *c_3***
1EHJ	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**Cytochrome *c_7***
3OV0	0	4	4	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**Dodecaheme cytochrome *c***
3U99	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	**Diheme cytochrome DHC**
3PMQ	0	0	4	0	0	0	0	0	2	1	0	0	0	0	2	0	1	0	0	0	0	0	0	0	**Decaheme cytochrome MtrF**
3UFK	0	0	4	0	0	0	2	1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**Undecaheme cytochrome UndA**
2K3V	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	**Small tetraheme cytochrome *c***
2P0B	0	0	1	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	**NrfB**
2CZS	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	**Dhc2**
3ML1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	**NapB**
3GM6	0	0	0	0	0	0	0	0	0	0	2	3	1	1	1	0	0	0	0	0	0	0	0	0	**Eight-heme nitrite reductase**
1QDB	0	0	0	0	0	0	0	0	0	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0	**NrfA**
1FGJ	0	0	0	0	0	0	0	0	0	0	0	3	3	1	0	0	1	0	0	0	0	0	0	0	**Hydroxylamine oxidoreductase**
1BVB	0	0	0	0	0	0	0	0	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0	**Tetraheme cytochrome c_{554}**
1SP3	0	0	0	0	0	0	0	0	0	0	2	3	1	0	0	0	0	1	0	0	0	0	0	1	**Octaheme tetrathionate reductase**
1Q9I	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	0	**Fumarate reductase**
1H21	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	**Split-Soret cytochrome *c***
2VR0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	**NrfH**
Total	*8*	*12*	*23*	*6*	*2*	*4*	*2*	*1*	*5*	*2*	*15*	*17*	*7*	*4*	*3*	*2*	*2*	*2*	*1*	*1*	*1*	*1*	*1*	*1*	

Figure 7. hMFS relationships in eight-heme, NrfA and related MHCs derived from cluster-guided structural superpositions. Top: Superpositions of the heme groups of selected MHCs resulting from the simultaneous overlay of the protein part of the hMFSs of each MHC to the sites of eight-heme nitrite reductase (PDB entry 3GM6) belonging to the same AC2.75 clusters (2K3V, cyan; 2P0B, magenta; 2CZS, gray; 3ML1, dark green; 3GM6, blue; 1QDB, light green; 1FGJ, yellow; 1BVB, red; 1SP3, orange; 1Q9I, brown). Residue numbering for the heme groups is shown for structure 3GM6. Bottom: summary of the relationships, color coded according to the cluster assignments of Figure 4 (green: cluster 65; blue: cluster 68; magenta: cluster 64; yellow, cluster 55; pink, cluster 61). Heme sites are labeled by their residue numbers in the PDB structure. Relationships are derived from spatial proximity after superposition and all refer to the sites of nitrite reductase. Only clusters containing hMFSs from different MHCs have been highlighted. A star indicates sites that fulfill the requirement of spatial proximity but are not satisfactorily superimposed (e.g. iron ligands do not overlay or the heme orientation is somewhat different). The heme groups 802 of structure 1SP3 and 804 of structure 1Q9I have been omitted for clarity. The figure independently re-discovers the known 29 relationships between between hemes I-VIII of nitrite reductase and hemes I-VIII of hydroxylamine oxidase, between hemes IV-VIII of nitrite reductase and hemes I-V of NrfA, or between seven out of the eight groups of nitrite reductase and tetrathionate reductase.



Small tetraheme cytochrome c	NrfB	Dhc2	NapB	Eight-heme nitrite reductase	NrfA	Hydroxylamine oxidoreductase	Tetraheme cytochrome c_{554}	Octaheme tetrathionate reductase	Fumarate reductase	
2K3V	2P0B	2CZS	3ML1	3GM6	1QDB	1FGJ	1BVB	1SP3	1Q9I	*Notes*
-	201	501	1129	1001	-	547	-	801	-	
218	202	500	1128	1002	-	548	-	803	801	
238	203	-	-	1003	-	549	213	804	802	
-	-	-	-	1004	515	550	214	-	-	*Catalytic site*
261	204	-	-	1005	516	551	215	805	803	
278*	205	-	-	1006	517	552	216*	806	-	
-	-	-	-	1007	518	553	-	807	-	
-	-	-	-	1008	519	552	-	808*	-	
-	-	-	-	-	-	-	-	802	-	
-	-	-	-	-	-	-	-	-	804	

4.4. Analysis of the relationship of functional changes in metalloenzymes and changes in metal sites

Claudia Andreini^{1,2,*}, Antonio Rosato^{1,2,} Yana Valasatava^{1}, Janet Thornton^{3}, and Nicholas Furnham

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

[3]EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

In preparation

Introduction

A number of metal ions are essential to life [1;2]. A major determinant of their functional relevance in living systems is that a substantial fraction of enzymes require metals for their catalytic activity and are thus called metal-dependent enzymes or metalloenzymes. If only enzymes of known structure are taken into account, then metal-dependent enzymes constitute about 40% of all enzymes [3]. Metals can bind to enzymes as individual ions or within metal-containing cofactors. The latter can be extremely diverse in their chemical complexity, ranging from organic ligands binding a single metal ion, such as porphyrins, to highly elaborate polymetallic clusters, such as the FeMoCo cofactors of nitrogenases. Metal ions as well as metal-containing cofactors can play either a catalytic or a structural role. When the metal is catalytic, its interaction with the protein and the substrate are crucial to determine the details of the mechanism of catalysis. Importantly, the relevant interaction with the protein extends beyond the first coordination sphere of the metal ion (i.e. the nature and distribution along the sequence of protein ligands) [4;5]: so-called second sphere interactions play a role in defining features such as substrate recognition and selectivity, or the structure of the transition state during the reaction [6]. In addition, the protein part contributes to the catalytic mechanism also through residues that are not interacting with the metal-containing cofactor.

The diversity of enzymatic reactions is absolutely remarkable, both in terms of the variety of substrates processed and of the chemical reactivity deployed. Such diversity is much more extensive than the variety of the folds of enzymes. Resources such as CATH [7] or SCOP [8;9] are able to capture distant relationships between protein domains through the analysis of their three-dimensional atomic structures and thus recapitulate their structural variety. A very useful outcome of such classifications is the notion of protein superfamily, which is the ensemble of all the protein domains with the same overall structural features. By construction, enzymes with the same structural features belong to the same protein superfamily. The next question is how the chemistry performed by the various enzymes is differentiated within a given superfamily and what are, if any, the aspects that are common to the various superfamily members. Indeed, many studies have shown that there is often conservation of some aspects of chemistry between relatives in enzyme superfamilies; at the same time, there are examples of relatives which have diversified to perform very different functions (as defined by the overall reaction they perform), and/or to use different chemical mechanisms (the method by which the substrates transform), and/or act with different specificity [10]. In many cases evolution has managed to significantly modify the chemistry performed by an enzyme through small-scale local changes, rather than by remodeling the entire structure [11;12]. In multi-domain enzymes, functional changes may also arise from the combination of different domains within a single protein chain [11;13].

In the present work we aimed at gathering a complete overview of the differentiation of the functional properties of metal-dependent enzymes and of the underlying mechanisms. Metal-dependent enzymes require a specific focus because the presence of the metal-containing cofactor enables mechanisms for functional differentiation that are not operative in the other enzymes. The present analysis is also useful to investigate the interplay of the local structure of the metal-binding site and of the remainder of the protein active site in the determination and evolution of the enzyme reactivity [4]**.** In particular, individual metals can feature broad, overarching mechanisms of catalysis [3], which can be exploited by different enzymes to achieve different chemistry. We observed that the functional landscape of metal-dependent enzymes is significantly more complex and diversified than simple gain/loss of a metal-binding site or replacement of metal ligands.

Materials and methods:

Our starting point was Metal-MACiE, a database of catalytic mechanisms for metalloenzymes of known 3D structure [14]. Hereafter, Metal-MACiE entries will be labeled as MM# followed by the corresponding database identifier (e.g. MM#0137). All catalytic metals in Metal-MACiE were selected and mapped to the MetalPDB database [15]. The latter is a database of metal-binding sites automatically generated from the Protein Data Bank (PDB, [16]). For each metalloenzyme in Metal-MACiE, the EC number is known. The latter is used to describe the reaction catalyzed by an enzyme and consists of a hierarchical classification system developed and maintained by the Enzyme Commission (EC). It is of a four-level descriptor, with the first three levels broadly categorising the overall chemistry and the fourth level being a serial number that is assigned to differentiate the substrate specificity. There is no correlation between the differences between the reactions catalysed and the numerical identifiers in the EC classification; so for example EC number 1.1.1.1 is no more similar to 1.1.1.2 than it is to 1.1.1.25.

Each metal-binding site identified in MetalPDB as corresponding to a Metal-MACiE entry belongs to a group of equivalent sites, i.e. sites bound at the same location within a common fold. The latter were mapped to the sequences in Funtree [17] alignments in order to identify the positions of the protein residues acting as metal ligands. FunTree alignments are sequence alignments derived from structural alignments of protein domains featuring the same CATH classification. In practice, each FunTree alignment represents a given CATH superfamily [17]. This allowed us to evaluate the conservation of the ligands within each superfamily.

When it was not possible to map a metal-binding site within the FunTree alignment of a given superfamily, e.g. because the corresponding PDB structure of the holo-protein was not included in the FunTree selection, the PDB structure of the original Metal-MACiE was added by aligning it to the underlying FunTree structural alignment. For this we used the program TM-ALIGN [18]. The newly obtained structural superposition allowed us to derive a sequence alignment containing the metal-binding site of interest.

After mapping a metal-binding site within a FunTree sequence alignment, each sequence in the alignment was inspected for the occurrence of the metal ligands and then labeled as “metal-binding” if it conserved at least 50% of the ligands. For sites with only two ligands, the label was assigned to sequences conserving the entire site.

For each CATH superfamily, we performed an analysis based on the EC numbers associated with its members. In particular, we separated the superfamily into groups with the same EC number. This grouping was used first to derive statistics about the metal-binding capability within each group, and then to compare the metal-binding capabilities across different groups belonging to the same superfamily (i.e. across enzymes with different EC number but the same fold). For each family of enzymes, functional and structural details were manually extracted from various sources, which typically included: the publications describing the structures deposited in the PDB; MetalPDB [15]; Metal-MACiE [14]; PDBSprotEC [19]; BRENDA [20]. The metal-binding sites were structurally superimposed with MetalS^2 [21].

Results

The Metal-MACiE resource contains an annotation for 174 catalytic metal ions (i.e. metals or metal-containing cofactors that play a role in the catalytic mechanism), which map to 88 CATH superfamilies. Of these, we were able to analyze the conservation of the first sphere residues (or ligands) within CATH domains for just 106 catalytic metals. The causes of this are: for 21 catalytic metals there is no structural information and thus, the metal coordinating residues are unknown; 26 catalytic metals bind to a region of the structure that is not associated to any superfamily; 21 catalytic metals are bound to superfamilies that are not in FunTree data. The remaining 106 catalytic metal sites fall within 65 CATH superfamilies. Of these, just nine superfamilies contain enzymes all associated with the same EC code. In these superfamilies, the metal site is present on average in 96% of the enzymes. These metal sites, according to the Metal-MACiE annotation, bind always the same metal. The remaining 56 superfamilies include metal-dependent enzymes with different EC numbers, and thus correspond to instances where the same protein fold is adopted to perform different catalytic functions. To assess the extent to which the function is dependent on the occurrence and features of metal sites, we undertook a systematic analysis of all of them. Any pair of functionally diverse enzymes within each superfamily can (i) harbor the same metal-binding site, (ii) harbor metal-binding sites with differences in their first coordination spheres, (different number or identity of metal ligands), (iii) harbor sites with varying nuclearity (i.e. number of metal ions bound); (iv) use different catalytic metal ions and finally (v) some members can miss metal-binding capability altogether. For each of the 56 superfamilies, we identified one or more of these five types of behavior, resulting in a total of 82 instances.

In superfamilies that contain multiple EC numbers, the metal sites are most commonly conserved (45% of all instances) (Figure 1). The second most frequent case is that of variations of the first coordination sphere while leaving the same metal ion(s) (27% of the instances). Changes in nuclearity or changes in the identity of the bound metal ion(s) are comparatively less common (4% and 8% respectively). 16% of the instances, finally, correspond to cases of loss/gain of metal-binding capabilities. However, in some of these (6% of all instances) the metal site is only apparently lost/gained, as it is actually shifted within the protein fold. In the rest of this section we present a more detailed analysis of the characteristics of the 56 superfamilies, by separating them according to their specific behavior in terms of metal-binding properties in order to identify common trends.

Superfamilies containing enzymes with different EC numbers and conserved metal sites

In these instances (45% of the total, Figure 1), the first coordination sphere provided by the polypeptide chain to the metal cofactor is conserved across members of each superfamily having different EC numbers. This behavior was observed in 37 superfamilies, and it is the dominant one when the functional diversification is little (Figure 2). We then focused on the instances with the largest functional diversity, as evaluated from the level of change in the EC numbers. Typically, the functional changes in these superfamilies are the result of modifications in the local structure of the active site, as the overall fold is maintained, e.g. leading to different or a different number of cofactors or co-substrates being bound in the enzymes with different EC numbers. Within this context, the specific contribution of the metal cofactor to the overall catalytic mechanism is essentially unaltered.

As an example, superfamily 2.60.120.330 contains a variety of oxidoreductases, including as many as 14 different EC numbers. Two of these are described in Metal-MACiE i.e. deacetoxycephalosporin-C synthase (EC number 1.14.20.1, MM#M0137) and isopenicillin-N synthase (EC number 1.21.3.1, MM#M0145). The former enzyme binds three different substrates: penicilin N, 2-oxoglutarate and O_2 [22]; instead, isopenicillin-N synthase operates on two substrates: delta(L-2-aminoadipyl)-L-cysteinyl-D-valine and O_2 [23]. In both enzymes, O_2 binds directly to the catalytic iron(II) ion, producing a peroxo intermediate which then converts to oxo-iron(IV). The latter is the oxidizing species that acts on the penicillin or delta(L-2-aminoadipyl)-L-cysteinyl-D-valine substrate. This mechanism is maintained for all the oxidoreductases of the superfamily [24;25], thus the iron(II) function is the same across the various EC numbers.

The only exception to the generally observed trend of the function of the metal site being conserved despite the changes in EC number within each superfamily is that of superfamily 3.40.50.280, which contains enzymes harboring different cobalamin-based cofactors. In these enzymes, the contribution of the protein to the coordination sphere of the cobalt ion is limited to a single His. In addition, the coordination sphere of the catalitically relevant form of the metal-containing cofactor is completed, besides the macrocycle ring, by either a methyl (methylcobalamin) or an adenosyl (adenosylcobalamin) group. Methylcobalamin is produced in the active site of methionine synthase (EC number 2.1.1.13, MM#0268) from the reaction of cobalt(I) with methyl-tetrahydrofolate; the cobalt-carbon bond undergoes a heterolytic cleavage to transfer the methyl group to homocysteine, releasing methionine. Adenosylcobalamin is found in various mutases (EC number 5.4.99.2, MM#0062; EC number 5.4.99.1, MM#0063). Here the cofactor serves as a source of reactive free radicals that are generated by homolytic scission of the coenzyme's cobalt–carbon bond. Bond homolysis is promoted by the electrostatic interaction between the ribose and the protein [26]. Thus in superfamily 3.40.50.280, the cobalt ion performs different roles in the catalytic mechanism mainly as a function of its ligation within the cofactor and the interactions between the organic part of the cofactor itself and the protein chain.

Superfamilies containing enzymes with different EC numbers and variations in the coordination sphere of the catalytic metal ions

In these instances (27% of the total, Figure 1), the commonly located active site of different metal-dependent enzymes in a given functionally diverse superfamily provides different ligands to the same metal ion. This behavior occurred in 22 superfamilies. The variations that are relevant to the present discussion are different number and/or different identity of the protein ligands in the first coordination sphere of the metal ion, whose identity is nevertheless maintained. Such variations generally affect one or two ligands, and never more than 50% of the protein ligands, i.e. involve a minor portion of the entire site. We observed this behavior in superfamilies with any degree of functional diversity (Figure 2), similarly to what observed for superfamilies with conserved sites.

Superfamily 3.90.245.10 contains enzymes whose EC numbers differed only at the fourth level, namely 3.2.2.1, 3.2.2.3, 3.2.2.8. These enzymes are all nucleosidases and differ only in the substrate they recognize [27]. Their metal binding sites differ only by the identity of one protein ligand, which binds to the catalytic calcium(II) ion through its oxygen atom of the main chain. Therefore, the replacement of the amino acid did not alter the metal coordination environment directly and, similarly, did not alter the enzyme mechanism appreciably. Nevertheless, the different side chain may impact on the second coordination sphere, possibly contributing to the variation of EC number.

A more complex example is that of superfamily 3.40.50.970, which contains enzymes spanning as many as 12 different EC codes, including oxidoreductases, transferases and lyases. They all use a magnesium(II) ion to bind the thiamine diphosphate cofactor and properly orient it within the active site. Similarly to the previous example, the residues whose identity is different in the various magnesium(II)-binding sites interact with the metal ion through their backbone oxygen atoms. The reaction mechanism always starts with the formation of a carboanion upon deprotonation of the thiamine cofactor. The carboanion then performs a nuclophilic attack on the substrate. The rest of the reaction, which is different in the different enzymes, does not depend on the metal, and is the result of other structural properties of the active site and/or of the presence of additional cofactors.

The present type of variation of the metal-binding site is somewhat less common in sites that contain only donor atoms from protein side chains. Such coordination environments are typical of metals softer than magnesium(II) or calcium(II), such as transition metal ions. Despite this tendency to be quite strictly conserved, which in the past we showed to be useful for the prediction of metal-binding properties [28;29], we could identify in our dataset some relevant instances. The protein ligands that are replaced within the superfamily are often conservatively substituted, so that the changes in the structure of the metal-binding site still leave the mechanism largely unaffected. As an example, superfamily 3.40.225.10 contains L-ribulose-5-phosphate 4-epimerase (EC number 5.1.3.4, MM#0273) and L-fuculose-phosphate aldolase (EC number 4.1.2.17, MM#0072), as well as various other related enzymes (4.1.2.19, 4.2.1.109). In 4.1.12.17 and 4.1.12.19 the catalytic zinc(II) ion is bound by the enzyme in its resting state through the side chains of three His and one Glu. In 5.1.3.4, the latter is replaced by Asp, which however is not involved in zinc(II)-binding. Notably, the catalytic mechanism of 4.1.12.17 involves breaking the Glu-zinc(II) coordination bond upon substrate binding, leaving the free Glu side chain to play a role in proton shuttling [30]. A corresponding role is played by an unrelated Asp residue in aldolases [31]. The role of the metal ion in catalysis is maintained**,** as it promotes the deprotonation of the substrate.

Superfamilies containing enzymes with different EC numbers and different nuclearity of the metal site

These instances (4% of the total, Figure 1) feature a change of the number of metal ions forming the polymetallic unit that constitutes the catalytic center. Such a change is associated to a variation in the number of amino acidic ligands recruited in the formation of the site, in order to provide the adequate number of ligands given the number of metal ions involved. Variations in nuclearity are relatively uncommon, as they were observed in only three of the functionally diverse superfamilies analyzed.

As an example, superfamily 3.20.20.150 contains enzymes that bind one, two or even three divalent cations to perform their function. Three divalent, either zinc(II) or manganese(II) ions, are present in the active site of deoxyribonuclease IV [32] (EC number 3.1.21.2, MM#M0011) (Figure 3). The various isomerases that are present in this superamily bind two (xylose isomerase, EC number 5.3.1.5; L-rhamnose isomerase EC number 5.3.1.14) or one divalent metal ion (hydroxypyruvate isomerase, EC number 5.3.1.22). Xylose isomerase is the best characterized enzyme of this group (MM#0308) and binds two manganese(II) ions [33]. 3NGF is the only available structure for EC number 5.3.1.22 (unpublished); the structure binds one manganese(II) ion but its structural alignment to xylose isomerase reveals that it is endowed also with the putative ligands to the second ion. Finally, among lyases mannonate dehydratase (EC number 4.2.1.8), binds one manganese(II) ion (e.g. PDB ID 4EAC [34]). Instead, the structure of a putative myo-inosose-2 dehydratase (EC number 4.2.1.44), PDB code 3CNY (unpublished), is devoid of metals but actually contains all the ligands of the dinuclear site. The single metal ion of mannonate dehydratase is structurally equivalent to the second one of deoxyribonuclease IV, whereas the two ions of isomerases are equivalent to the second and third ions of deoxyribonuclease IV (Figure 3). In all systems, at least one of the metal ions directly binds the substrate, and activates it by increasing the electrophilicity of the atom bound to the donor oxygen atom. Additionally, the metals in the sites with higher nuclearity also bind and activate water, by increasing its acidity and/or enhancing its nucleophilicity. The single metal ion of mannonate dehydratase corresponds to the one ion of the two of xylose isomerase that is involved in substrate binding, thus highlighting a functional correspondence beyond the structural one. A similar correspondence is not obvious to draw for deoxyribonuclease IV as here all the zinc ions bind to the substrate.

Superfamilies containing enzymes with different EC numbers and different catalytic metal ions

In these instances (8% of the total, Figure 1), different metal-dependent enzymes that share the same fold bind in their commonly located active site different catalytic metal ions. The coordination sphere of the metal ion is typically modified to only a minor extent. These seven superfamilies featured the more profound levels of variation of the EC number (Figure 2). We observed two different effects on the reaction mechanism. One possibility is that the different metals perform a similar or even essentially identical function, such as substrate activation through an increase in acidity. However, the different structures of the enzyme cause the reaction to proceed to different degrees, hence the variation of EC code. Alternatively, the different metal ions play a different role in the catalytic mechanism, e.g. because of the different redox properties, thereby leading to widely different reactions.

Superfamily 1.20.1090.10 includes family III metal-dependent polyol dehydrogenases, such as glycerole dehydrogenase (EC number 1.1.1.6) or 1,3-propanediol dehydrogenase (EC number 1.1.1.202), as well as dehydroquinate synthase (EC number 4.2.3.4). The latter is a zinc(II)-dependent enzyme, whereas the polyol dehydrogenases can depend on either zinc(II) or iron(II). All these enzymes share the same catalytic mechanism, regardless of the bound metal. In fact, the metal ion binds to the substrate, often in a bidentate manner, and increases the acidity of one of the hydroxyl groups, favoring proton dissociation followed by oxidation of the alcoholate to a carbonyl via the transfer of a hydride to NAD^+. Thus, the different redox properties of zinc(II) and iron(II) do not matter: both metals are acting only as Lewis acids. Dehydroquinate synthase builds upon the same mechanism, which actually constitutes the first step of the complex reaction catalyzed. The oxidation of the alcohol is followed by beta-elimination of the phosphate group of the substrate, which is promoted by the presence of a phosphate-binding pocket in the enzyme. The third step is a reversal of the first step, as the ketone initially formed is reduced by NADH. Notably, dehydroquinate synthase does not effectively uses the zinc(II) ion in the reaction mechanism after the step, other than to keep the substrate in the binding pocket.

The superfamily of metallo beta lactamases (CATH code: 3.60.15.10) contains enzymes belonging to two distinct EC classes: hydrolases (glyoxalase II, EC number 3.1.2.6; beta-lactamases, EC number 3.5.2.6 and tRNase Z, EC number 3.1.26.11) or oxidoreductases involved in the response to nitrosative and/or oxidative stress. While hydrolases are most commonly zinc(II)-dependent enzymes (only glioxalase II is active also in the presence of other metals than zinc, such as iron(II) and manganese(II)), oxidoreductases strictly require iron to perform the catalytic reaction. The metal-binding sites are located in corresponding positions and are structurally similar in the two groups of enzymes (Figure 4), and the metal cofactor is generally dinuclear (with the exception of type B2 metallo beta lactamases). The metal ions bind directly to the substrate, properly orienting it within the active site. However, during the catalytic cycle the function of the metals is completely different in the hydrolases vs. the oxidoreductases. In the latter enzymes, each iron(II) ion transfers an electron to the substrate, thus providing two electrons in total upon formation a di-iron(III) site that is subsequently reduced by a $FMNH_2$ molecule. On the other hand, the zinc(II) site of hydrolases is responsible for the activation of a water molecule for the nucleophylic attack on the substrate**.** This type of mechanism is commonly observed in zinc-dependent hydrolases, as zinc(II) is a strong Lewis acid. Interestingly, the only metal ligand that appears to change between the two classes of enzymes is a Glu residue in the di-iron(II) sites replacing a conserved His in the hydrolytic di-zinc(II) sites (Figure 4). It has been hypothesized that this Glu residue is able to suppress any possible hydrolytic cross-reactivity in the oxidoreductases.

Superfamilies containing enzymes with different EC numbers and gaining/losing a catalytic metal-binding site

In these instances (16% of the total, Figure 1), the structure-based sequence alignment of FunTree indicates that the metal ligands observed in the starting Metal-MACiE entry are not conserved in all superfamily members. However, a closer manual inspection of these enzymes revealed that in a subset (6% of the total of instances) a metal-binding site with a catalytic role is still present, but differently located within the fold. In practice, each of these alignments contained two groups of metal-dependent enzymes with the same fold but different ligands. Thus, within the present instances we identified eight superfamilies containing both metal-dependent and non-metal-dependent enzymes, and five superfamilies containing only metal-dependent enzymes whose site is in different positions within the fold. For the first group of superfamilies, the presence or absence of a catalytic metal cofactor has a deep impact on the catalytic mechanism and thus it is not surprising that these superfamilies are characterized by the largest functional diversity. On the other hand, the impact on the mechanism is smaller for the enzymes whose metal site is simply shifted within the structure.

As an example of a superfamily containing both metal-dependent and non-metal-dependent enzymes, we describe superfamily 3.30.1130.10, which includes the enzyme GTP cyclohydrolase IA (MM#0038). The latter is a zinc-dependent hydrolase (EC number 3.5.4.16). In addition, the same superfamily includes two non-metal dependent enzymes: PreQ$_0$ reductase (EC number 1.7.1.13) and dihydroneopterin aldolase (EC number 4.1.2.25). For all these enzymes, a 3D structure in complex with the substrate or a substrate analog is available (Figure 5). The substrates are dicyclic compounds, either functionalized purines or pteridines, both containing a 2-amino-pyrimidine ring. The different functionalizations occur on the other ring, which is either a imidazole or a pyrazine. The latter ring is also the region which the enzyme acts upon. The three enzymes in object indeed interact very similarly with the common 2-amino-pyrimidine ring, through the formation of H-bonds with the side chain of a conserved Glu (Figure 5), whereas the interaction is substantially different on the other side of the substrate. In GTP cyclohydrolase IA, the zinc(II) ion faces the imidazole ring of the substrate and activates a water molecule that acts as a nucleophile. The intermediate generated after the nucleophilic attack is proposed to remain bound to the zinc(II) ion, also on the basis of a structure in presence of a substrate analogue, eventually leading to formation of zinc-bound formate as one of the reaction products. Instead, $PreQ_0$ reductase catalyzes the reduction of a nitrile group to a primary amine, whereas dihydroneopterin aldolase is a lyase that catalyzes the release of glycoaldehyde from the substrate, 7,8- dihydroneopterin. Intriguingly, $PreQ_0$ reductase forms a covalent thioimide, a putative intermediate in the reaction, using the side chain of a Cys residue that is structurally equivalent to one of the zinc(II) ligands in GTP cyclohydrolase IA (Figure 5).

Instead, superfamily 3.40.630.10 is an example for the superfamilies whose metal-binding site shifted to a different position within the fold thus resulting in an *apparent* gain/loss of the site, as judged from ligand conservation in sequence alignments. This superfamily contains two different enzymes: bacterial leucyl aminopeptidase, EC number 3.4.11.10 (MM#0167), and carboxypeptidase A, EC number 3.4.17.1 (MM#0171). The EC number thus changes only at the third level. The rationale for the shift in the position of the site is due to the substrates having to penetrate within the protein structure to different depths. Indeed, the substrates are cleaved in a more or less exposed position in exo- vs. endo-peptidases. This, in turn, determines the need for the active site to accommodate, respectively, a smaller or larger portion of the polypeptide chain. Aminopeptidases contain a di-zinc(II) site, whereas carboxypeptidase A contains a mononuclear zinc(II) site. The two enzymes adopt the same mechanism, i.e the metal site binds the substrate and activates a water molecule to act as the nucleophile eventually leading to the cleavage of the peptide bond. After structural alignment, it appears that the two binding sites are located in different positions within the protein fold, with the site of carboxypeptidase A having shifted toward the protein core with respect to aminopeptidase. In fact, the sequence alignment derived from the structural alignment indicates that only one ligand is shared between the two sites, namely D293 of aminopeptidase A corresponds to E72 of carboxypeptidase.

Discussion

We have identified five different patterns of conservation/change of the catalytic metal-binding site within CATH superfamilies containing metal-dependent enzymes. Such patterns affect the structural properties of the site to an increasing extent, from complete conservation to complete loss (or gain) of the site, through variations in the number and/or chemical identity of the priotein ligands, variations in site nuclearity and variations in metal identity. On the basis of the analysis presented in the Results section and the data summarized in Figure 2, we can postulate the existence of a relatively clear-cut correlation between structural change in the site and functional diversity. Indeed, the wider-scale changes in the structure of the metal site occur in superfamilies where the observed diversity affects the first or the second level of the EC number of its member enzymes (Figure 6). In particular, changes in catalytic metal and gain/loss of the metal-binding site always affected, in our dataset, the first level of the EC number, i.e. were associated to changes in the enzyme class. The aforementioned relationship cannot be reversed: superfamilies with large functional diversity can maintain their catalytic metal-binding sites completely or largely unchanged (Figure 2). In these cases thus enzymes with the same fold catalyze significantly different reactions, as gauged by the different EC numbers, using essentially the same metal-binding site. At the mechanistic level, this functional diversity is seldom caused by an appreciable variation of the role of the metal cofactor within the catalytic cycle. In fact, the role of the metal in catalysis tends to be somewhat conserved. Typically, the metal is involved in substrate binding and thus determines substrate orientation and activation within the active site. These can be the first steps of the overall catalytic mechanism, common to various enzymes of a given superfamily regardless of their specific EC numbers. So the specific reactivity of a metal-dependent enzyme can actually be tuned by the protein moiety not via rearrangements of the metal coordination sphere but through the interaction of the protein with the (metal-bound) substrate, i.e. second sphere interactions, or by mutating other residues involved in the catalytic mechanism, or, in systems with larger modifications of the protein fold, by binding additional cofactors and/or co-substrates. In other words, while the metal-containing cofactor is responsible for getting the reaction started, it is the last part of the catalytic cycle that decides the fate of the reaction, under the main influence of the protein contribution to the structure of the substrate-binding pocket. Of course, this is not a strict rule. As an example, cobalamin-dependent enzymes can exploit different catalytic mechanisms as a consequence of the exact coordination of the cobalt ion.

When the first coordination sphere of the metal cofactor changes while the identity of the metal ion bound is retained, the mechanism of reaction typically does not change much. This is more commonly observed for the harder magnesium(II) or calcium(II) ions, whose first coordination sphere contains a number of oxygen atoms from the protein main chain, thus making some positions relatively insensitive to the replacement of the corresponding side chains. Consequently, this kind of variations is compatible also with the absence of change in EC numbers. In other words, enzymes in a given superfamily can catalyze the same reaction, i.e. have the same EC number, despite the occurrence of changes in the first coordination sphere of their common catalytic metal. On the other hand, changing the identity of the metal ion bound may have a larger impact on the functional properties (Figure 2). This is likely to happen when the metals differ in having or lacking redox activity under biologically relevant conditions (i.e. zinc(II) vs. iron(II) as shown in Figure 4). However, even under these conditions, it is still possible that the catalytic mechanism is maintained to some extent, e.g. as for alcohol dehydrogenases. Although not explicitly addressed in this work, when different enzymes sharing the same EC number and belonging to the same superfamily bind different metal ions, the catalytic mechanism is typically conserved. An extreme example of this behavior is that of cambialistic superoxide dismutases, which, depending on bioavailability, can incorporate different metal ions in their active site and retain catalyitic activity.

Changes in site nuclearity and loss/gain of entire metal-binding sites require larger changes in the structure of the active site, by recruiting (or dismissing) various metal ligands. Such structural changes are accommodated while maintaining the overall protein fold, indicating that metalloprotein design by nature sometimes exploits an existing fold to add or modify metal-binding properties similarly to the artificial metalloprotein engineering concept**.** As already mentioned these types of variations are likely to be associated to larger functional variations (Figure 2). Yet, the impact at mechanistic level can be quite different. Within a given superfamily, sites with different nuclearity can share similar roles during the catalytic cycle. A possible rationale for the presence of additional metal ions in one enzyme with respect to another of the same superfamily is that the former needs to activate additional co-substrates or that its substrate is more inert. On the other hand, gain/loss of the metal-binding site leads necessarily to a drastic change in the catalytic mechanism.

A final consideration regards the “forces” that drive the evolution of catalytic metal-binding sites in enzymes. Successful sites, i.e. sites that are catalytically efficient at least for some specific steps, are maintained and re-used in different contexts to expand the enzymatic portfolio available in nature.

The modulation of second-sphere or protein-substrate and protein-cofactor interactions constitutes a crucial factor to support this expansion, which can be as important as the composition of the coordination sphere of the metal. On the other hand, some metal-binding sites are uniquely associated to a given enzymatic activity. This might happen because of their comparatively recent appearance, or because mutations around the metal site would be too detrimental for metal affinity, or because the substrate (and related compounds) is relatively uncommon. There are two possible reasons why the evolution of the metal site, especially in the cases with the largest structural changes and the highest impact on functional properties (Figure 6), is not accompanied by a significant variation of the protein fold. One possibility is that the protein moiety provides a suitable environment for binding metal ions with good affinity. This can be particularly relevant e.g. when different divalent cations replace one another in the active site, where the exact identity of the metal ion incorporated in the enzyme may be linked to the biosynthetic process also through the action of specific metallochaperones. The kinetics of metal release from the active site and the stability of the protein structure represent other factors that may warrant reuse of the same fold. A second argument is that the protein fold allows the substrate to be recognized with appropriate specificity. In this case, changing the identity of the metal or the nuclearity of a metal-binding site or introducing a novel site allows an organism to carry out different chemistry on the same substrate or, upon modifications of the protein moiety, on chemically related substrates (e.g. Figure 3). Beyond the need for evolving and differentiating the enzymatic repertoire of an organism, the above mechanisms may also contribute to coping with changes in the bioavailability of specific metal ions.

Figures

Figure 1: Pie chart showing the changes of metal sites observed in the 56 superfamilies that contain multiple EC numbers. One superfamily can be associated with more than one type of change, resulting in a total of 82 cases.

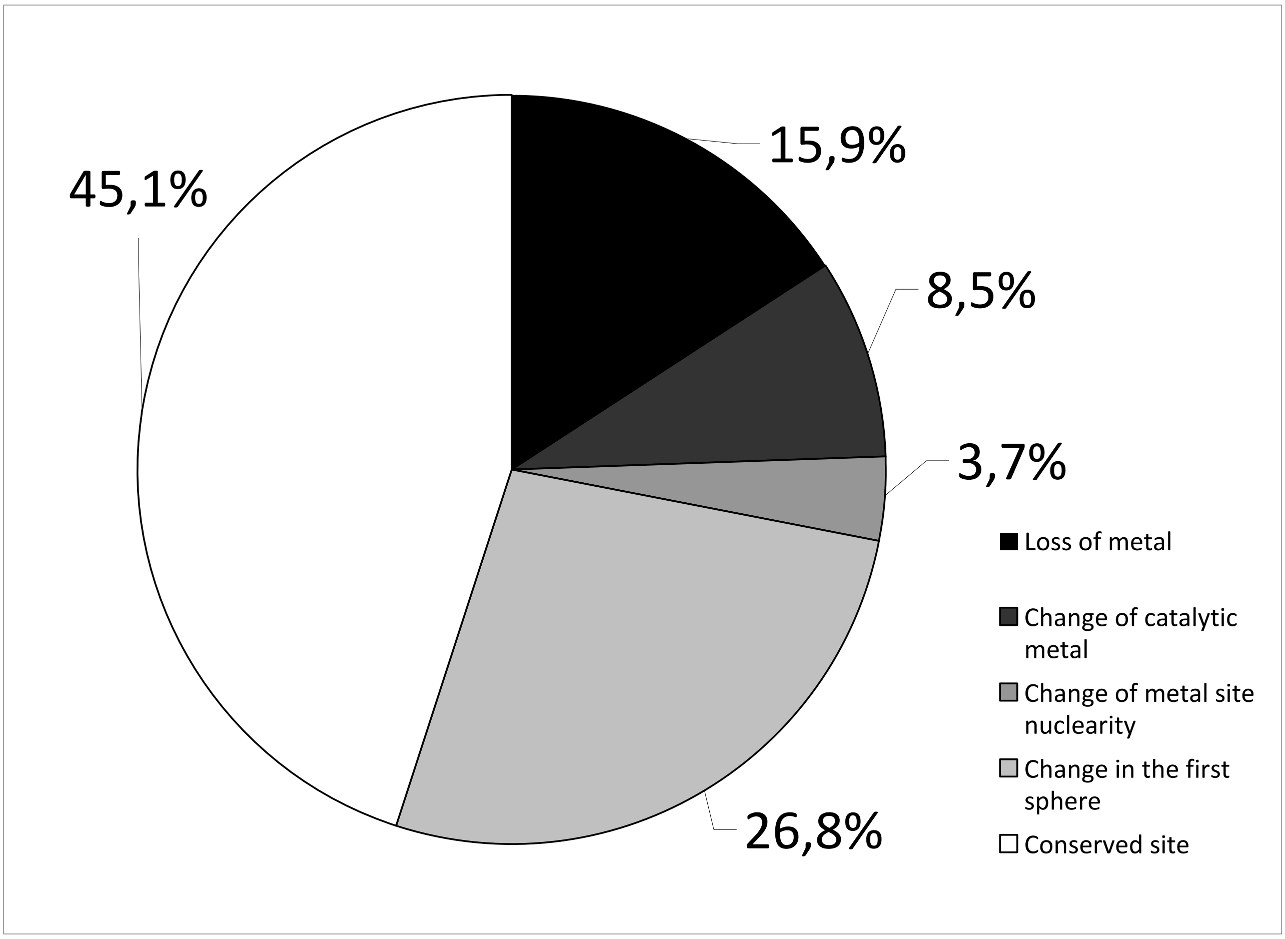


Figure 2: Histogram showing the changes of metal sites in the 56 superfamilies that contain multiple EC numbers. The changes were grouped according to the level of EC classification at which EC numbers in the superfamily differ. One superfamily can be associated with more than one type of change.

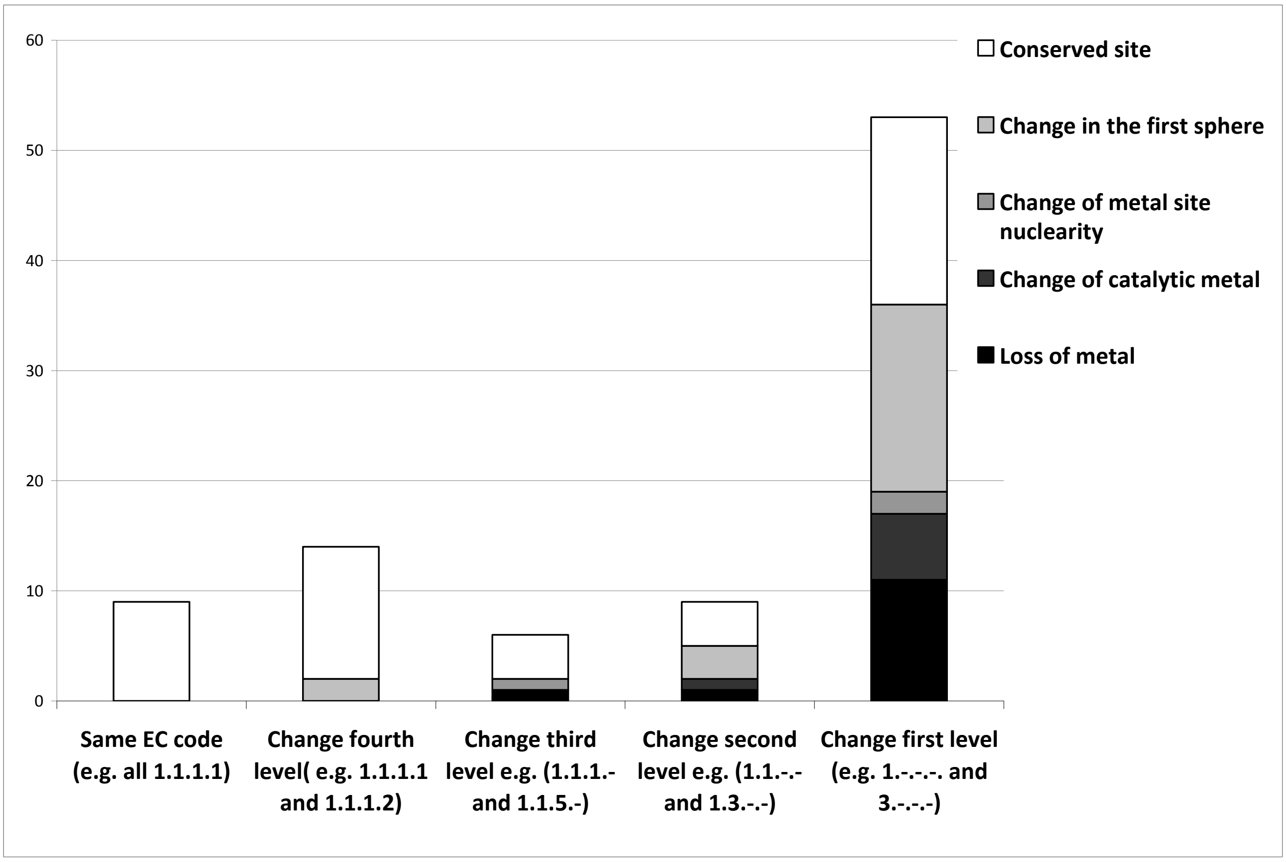
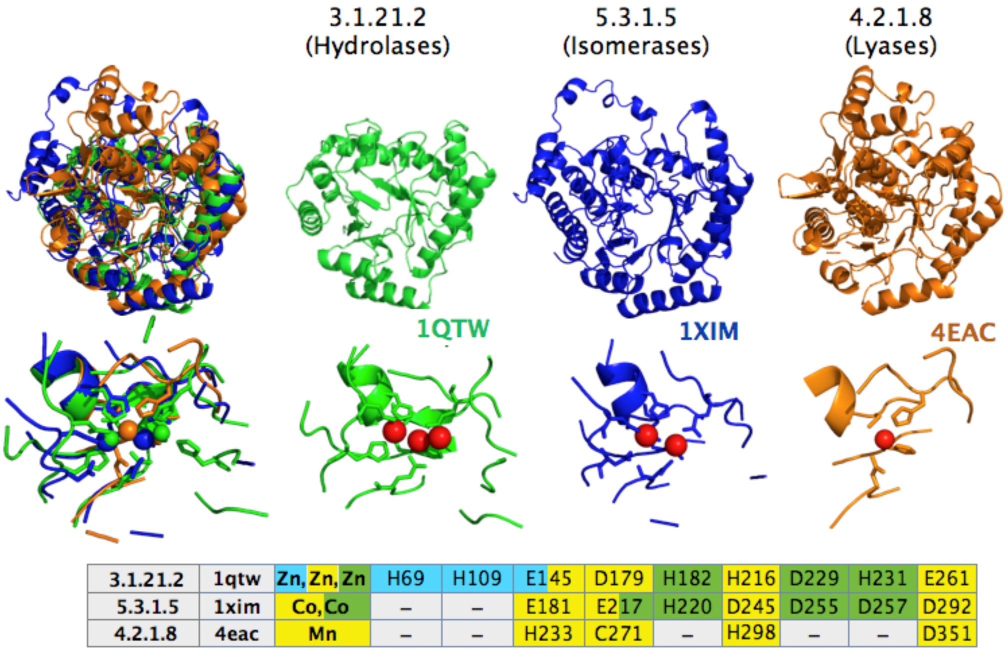
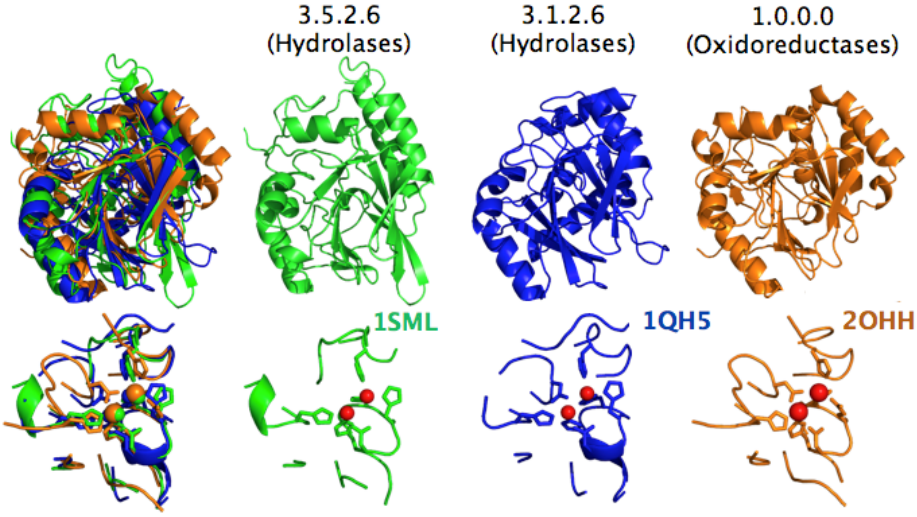


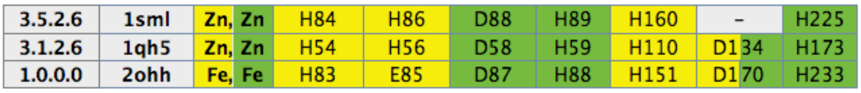
Figure 3: A superfamily (3.20.20.150) containing enzymes with different EC numbers and different nuclearity of the metal site. The aligned protein structures (top), the aligned metal site structures (middle, metal ions are depicted as red spheres), and the structure-based alignment of the metal ligands (bottom, different colors indicate the ligands of individual metal ions) are shown.



3.1.21.2	1qtw	Zn, Zn, Zn	H69	H109	E145	D179	H182	H216	D229	H231	E261
5.3.1.5	1xim	Co, Co	–	–	E181	E217	H220	D245	D255	D257	D292
4.2.1.8	4eac	Mn	–	–	H233	C271	–	H298	–	–	D351

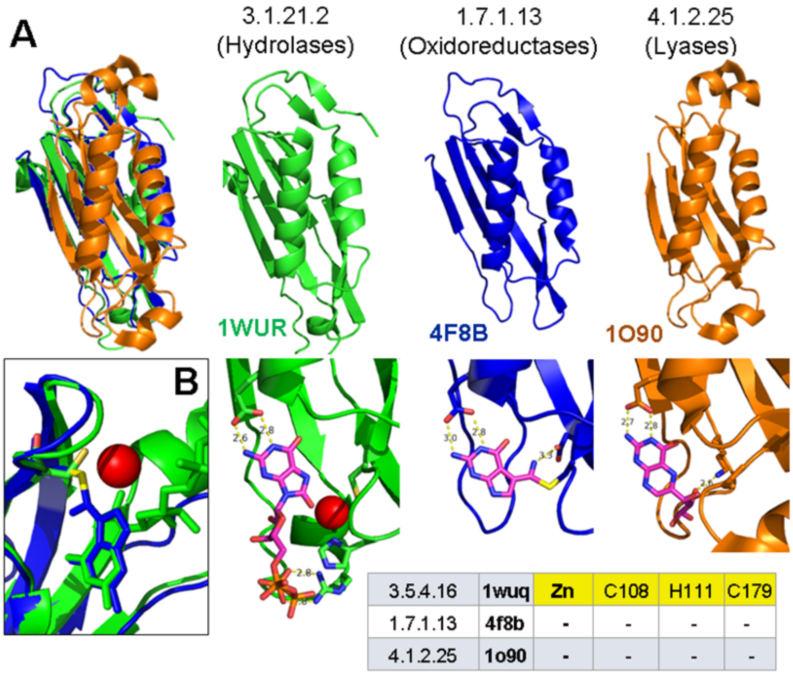
Figure 4: A superfamily (3.60.15.10) containing enzymes with different EC numbers and different catalytic metal ions. The aligned protein structures (top), the aligned metal site structures (middle, metal ions are depicted as red spheres), and the structure-based alignment of the metal ligands (bottom, different colors indicate the ligands of individual metal ions) are shown.





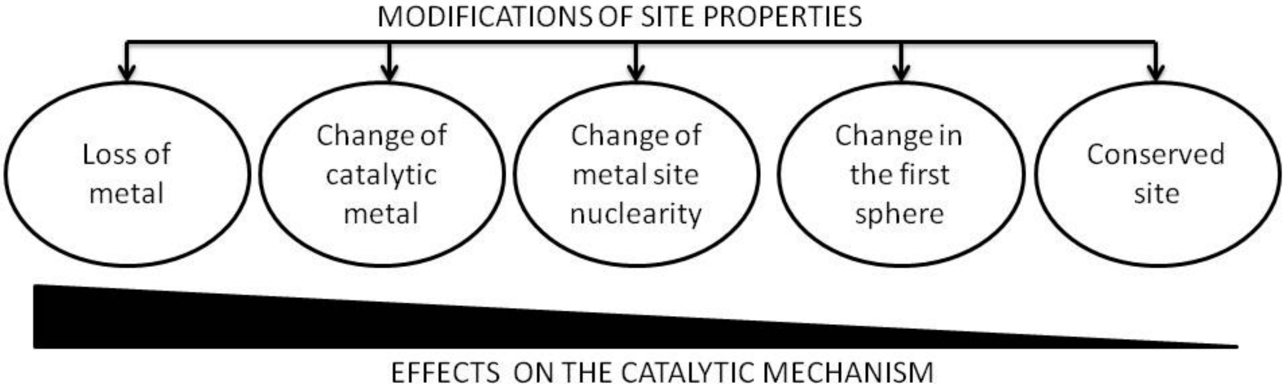
3.5.2.6	1sml	Zn, Zn	H84	H86	D88	H89	H160	–	H225
3.1.2.6	1qh5	Zn, Zn	H54	H56	D58	H59	H110	D134	H173
1.0.0.0	2ohh	Fe, Fe	H83	E85	D87	H88	H151	D170	H233

Figure 5: A superfamily (3.30.1130.10) containing enzymes with different EC numbers and gaining/losing a catalytic metal-binding site. The aligned protein structures (top), the aligned active sites with substrate-analogues bound (middle, the metal ion is depicted as a red sphere), and the structure-based alignment of the metal ligands (bottom) are shown.



3.5.4.16	**1wuq**	**Zn**	C108	H111	C179
1.7.1.13	**4f8b**	-	-	-	-
4.1.2.25	**1o90**	-	-	-	-

Figure 6: The relationship between the change in the structural properties of the catalytic metal site and the variation of the catalytic mechanism.


MODIFICATIONS OF SITE PROPERTIES
Loss of metal
Change of catalytic metal
Change of metal site nuclearity
Change in the first sphere
Conserved site
EFFECTS ON THE CATALYTIC MECHANISM

References

[1] Bertini,I., Sigel,A., & Sigel,H. (2001) *Handbook on Metalloproteins*, 1 edn. Marcel Dekker, New York.

[2] Frausto da Silva,J.J.R. & Williams,R.J.P. (2001) *The biological chemistry of the elements: the inorganic chemistry of life*. Oxford University Press, New York.

[3] Andreini,C., Bertini,I., Cavallaro,G., Holliday,G.L., & Thornton,J.M. (2008) Metal ions in biological catalysis: from enzyme databases to general principles. *J. Biol. Inorg. Chem.*, **13,** 1205-1218.

[4] Andreini,C., Bertini,I., & Cavallaro,G. (2011) Minimal functional sites allow a classification of zinc sites in proteins. *Plos ONE*, **10,** e26325.

[5] Andreini,C., Bertini,I., Cavallaro,G., Najmanovich,R.J., & Thornton,J.M. (2009) Structural analysis of metal sites in proteins: non-heme iron sites as a case study. *J. Mol. Biol.*, **388,** 356-380.

[6] Zhao,M., Wang,H.B., Ji,L.N., & Mao,Z.W. (2013) Insights into metalloenzyme microenvironments: biomimetic metal complexes with a functional second coordination sphere. *Chem. Soc. Rev.*, **42,** 8360-8375.

[7] 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. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, **41,** D490-D498.

[8] Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C., & Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36,** D419-D425.

[9] Andreeva,A., Howorth,D., Chothia,C., Kulesha,E., & Murzin,A.G. (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, **42,** D310-D314.

[10] Furnham,N., Sillitoe,I., Holliday,G.L., Cuff,A.L., Laskowski,R.A., Orengo,C.A., & Thornton,J.M. (2012) Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PloS Comput. Biol.*, **8,** e1002403.

[11] Todd,A.E., Orengo,C.A., & Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307,** 1113-1143.

[12] Glasner,M.E., Gerlt,J.A., & Babbitt,P.C. (2006) Evolution of enzyme superfamilies. *Current Opinion in Chemical Biology*, **10,** 492-497.

[13] Levitt,M. (2009) Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, **106,** 11079-11084.

[14] Andreini,C., Bertini,I., Cavallaro,G., Holliday,G.L., & Thornton,J.M. (2009) Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics*, **25,** 2088-2089.

[15] Andreini,C., Cavallaro,G., Lorenzini,S., & Rosato,A. (2013) MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.*, **41,** D312-D319.

[16] Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D., Young,J., Yukich,B., Zardecki,C., Berman,H.M., & Bourne,P.E. (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39,** D392-D401.

[17] Furnham,N., Sillitoe,I., Holliday,G.L., Cuff,A.L., Rahman,S.A., Laskowski,R.A., Orengo,C.A., & Thornton,J.M. (2012) FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Research*, **40,** D776-D782.

[18] Zhang,Y. & Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33,** 2302-2309.

[19] Martin,A.C. (2004) PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, **20,** 986-988.

[20] Barthelmes,J., Ebeling,C., Chang,A., Schomburg,I., & Schomburg,D. (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.*, **35,** D511-D514.

[21] Andreini,C., Cavallaro,G., Rosato,A., & Valasatava,Y. (2013) MetalS2: a tool for the structural alignment of minimal functional sites in metal-binding proteins and nucleic acids. *J. Chem. Inf. Model.*, **53,** 3064-3075.

[22] Roach,P.L., Clifton,I.J., Fulop,V., Harlos,K., Barton,G.J., Hajdu,J., Andersson,I., Schofield,C.J., & Baldwin,J.E. (1995) Crystal structure of isopenicillin N synthase is the first from a new structural family of enzymes. *Nature*, **375,** 700-704.

[23] Roach,P.L., Clifton,I.J., Hensgens,C.M., Shibata,N., Schofield,C.J., Hajdu,J., & Baldwin,J.E. (1997) Structure of isopenicillin N synthase complexed with substrate and the mechanism of penicillin formation. *Nature*, **387,** 827-830.

[24] Zhang,Z., Ren,J.S., Clifton,I.J., & Schofield,C.J. (2004) Crystal structure and mechanistic implications of 1-aminocyclopropane-1-carboxylic acid oxidase--the ethylene-forming enzyme. *Chem. Biol.*, **11,** 1383-1394.

[25] Welford,R.W., Clifton,I.J., Turnbull,J.J., Wilson,S.C., & Schofield,C.J. (2005) Structural and mechanistic studies on anthocyanidin synthase catalysed oxidation of flavanone substrates: the effect of C-2 stereochemistry on product selectivity and mechanism. *Org. Biomol. Chem.*, **3,** 3117-3126.

[26] Sharma,P.K., Chu,Z.T., Olsson,M.H., & Warshel,A. (2007) A new paradigm for electrostatic catalysis of radical reactions in vitamin B12 enzymes. *Proc. Natl. Acad. Sci. U. S. A*, **104,** 9661-9666.

[27] Versees,W. & Steyaert,J. (2003) Catalysis by nucleoside hydrolases. *Curr. Opin. Struct. Biol.*, **13,** 731-738.

[28] Andreini,C., Bertini,I., & Rosato,A. (2009) Metalloproteomes: a bioinformatic approach. *Acc. Chem. Res.*, **42,** 1471-1479.

[29] Andreini,C., Bertini,I., Cavallaro,G., Decaria,L., & Rosato,A. (2011) A simple protocol for the comparative analysis of the structure and occurence of biochemical pathways across superkingdoms. *J. Chem. Inf. Model.*, **51,** 730-738.

[30] Joerger,A.C., Mueller-Dieckmann,C., & Schulz,G.E. (2000) Structures of l-fuculose-1-phosphate aldolase mutants outlining motions during catalysis. *J. Mol. Biol.*, **303,** 531-543.

[31] Samuel,J., Luo,Y., Morgan,P.M., Strynadka,N.C., & Tanner,M.E. (2001) Catalysis and binding in L-ribulose-5-phosphate 4-epimerase: a comparison with L-fuculose-1-phosphate aldolase. *Biochemistry*, **40,** 14772-14780.

[32] Tsutakawa,S.E., Shin,D.S., Mol,C.D., Izumi,T., Arvai,A.S., Mantha,A.K., Szczesny,B., Ivanov,I.N., Hosfield,D.J., Maiti,B., Pique,M.E., Frankel,K.A., Hitomi,K., Cunningham,R.P., Mitra,S., & Tainer,J.A. (2013) Conserved structural chemistry for incision activity in structurally non-homologous apurinic/apyrimidinic endonuclease APE1 and endonuclease IV DNA repair enzymes. *J. Biol. Chem.*, **288,** 8445-8455.

[33] Kovalevsky,A.Y., Hanson,L., Fisher,S.Z., Mustyakimov,M., Mason,S.A., Forsyth,V.T., Blakeley,M.P., Keen,D.A., Wagner,T., Carrell,H.L., Katz,A.K., Glusker,J.P., & Langan,P. (2010) Metal ion roles and the movement of hydrogen during reaction catalyzed by D-xylose isomerase: a joint x-ray and neutron diffraction study. *Structure*, **18,** 688-699.

[34] Qiu,X., Tao,Y., Zhu,Y., Yuan,Y., Zhang,Y., Liu,H., Gao,Y., Teng,M., & Niu,L. (2012) Structural insights into decreased enzymatic activity induced by an insert sequence in mannonate dehydratase from Gram negative bacterium. *J. Struct. Biol.*, **180,** 327-334.

5. Conclusions

In summary, the main contribution of my Ph.D. project is the development and the application of a novel approach to the study of metals in biology. The approach is based on the alignment of three dimensional structures of Minimal Functional Sites (MFSs) in metal-binding biological macromolecules. MFSs constitute a novel viewpoint of metal sites that facilitates elucidating their mechanisms of function. The first part of my Ph.D. project was dedicated to development of a tool for systematic comparison of MFSs. The second part of the project was devoted to application of the method to address several scientific problems and scenarios with the aim at exemplifying use-cases and demonstrating the potentiality of the approach.

The tool for the pairwise comparison of MFSs was called MetalS^2 (http://metalweb.cerm.unifi.it/tools/metals2/). We specifically designed a scoring function that quantitatively assesses the similarity of two MFSs (the better the alignment, the less is the value of scoring function). The tool performs rotational and translational transformation of an MFS with respect to another MFS to minimize a scoring function composed by three terms that relate to basic physical and biochemical concepts. As an extension of MetalS^2, we developed MetalS^3 (http://metalweb.cerm.unifi.it/tools/metals3/), a tool that allows searching for similar sites among the ensemble of MFSs in MetalPDB. The core of the tool is the algorithm of MetalS^2 with some minor modifications to adjust the scoring function for the database search. MetalS^3 is a client-server application that may help researchers in the field of bioinorganic chemistry to assess the relationships or evaluate possible evolutionary links between different groups of metalloproteins as well as help experimentalists' work in understanding the function of uncharacterized metalloproteins. In fact, we previously demonstrated that if two metalloproteins (also having different folds) have similar MFSs, then the metal ions perform the same general function within these sites. Overall, this contributes to achieve a better comprehension of the role of metal ions in living systems. Although algorithmically very similar, MetalS^2 and MetalS^3 have somewhat different usage scenarios and enable access to distinct information. MetalS^2 requires the user to have prior knowledge of which structures are to be compared, either a pair or a group of related metalloproteins. It allows highlighting subtle similarities and/or differences in the local structure of the metal site. On the contrary, MetalS^3 constitutes an unbiased approach to seek structural similarities between metal sites, independently of the user's prior knowledge. The hits returned by

MetalS^3 can be a combination of relatively obvious (e.g. homologs of the query metalloprotein) and unexpected ones, making it a true knowledge discovery tool.

To further assess the usefulness of the developed approaches we applied them to get interesting biological hints on metalloproteins. In this frame, we exploited MetalS^2 as a part of a newly developed computational protocol to obtain a completely MFS-based classification of metalloproteins. By applying the protocol to all heme-binding proteins in the MetalPDB database we obtained a thorough view of structural variation across these systems. In addition, we unveiled structural relationships across different families in a manner that is unbiased by homology considerations. Indeed, the approach was able to highlight undetected similarities in multi-heme cytochromes. The obtained classification was shown to be complementary to and more inclusive than the existing structure-based and domain-based classifications of metalloproteins. In fact, the protocol can be used in data reduction with respect to existing fold-based classifications (e.g. obtained with a use of CATH, SCOP), since metal sites in proteins of different superfamilies and with different folds can share MFSs. In perspective, a new fold-independent classification of MFSs will provide an organized source of information to be embedded into the MetalPDB database.

Finally, we applied the MFSs comparison method to obtain hints on the evolution of metalloenzymes. This analysis was performed by exploiting a strategy planned for a new resource that integrates the information from three sources: CATH database, Metal-MACiE database, and the MetalPDB database. This resource was designed to functionally characterize metal sites in structurally related protein superfamilies. We have initially deployed the strategy outlined above to the contents of the FunTree database, a public resource containing the information on structurally defined enzyme superfamilies, in order to understand how enzymes depend on metals in the differentiation of their function. We have identified five different patterns of conservation/change of the catalytic metal-binding site within superfamilies containing metal-dependent enzymes. Such patterns affect the structural properties of the site to an increasing extent, from complete conservation to complete loss (or gain) of the site, through variations in the number and/or chemical identity of the protein ligands, variations in site nuclearity and variations in metal identity. When the metal site is conserved or the first coordination sphere of the metal cofactor changes while the identity of the metal ion is retained, the mechanism of reaction typically does not change much. On the other hand, changing the identity of the metal ion may have a larger impact on the functional properties, and the gain/loss of the metal-binding site leads necessarily to a drastic change in the catalytic mechanism. By mapping such changes onto the tree of life, evolutionary hints are obtained. This strategy is generally applicable beyond metalloenzymes, to obtain hints on the evolution of systems such as electron transfer chains or metal homeostasis and transport machineries.

Bibliography

1. **Bertini I., Sigel A., Sigel H.** *Handbook on Metalloproteins.* 1st. New York : Marcel Dekker, 2001.

2. **Williams R.J.P., Frausto da Silva J.J.R.** *The biological chemistry of the elements: the inorganic chemistry of life.* New York : Oxford University Press, 2001.

3. **Wolfgang Maret, Anthony Wedd, [ed.].** *Binding, Transport and Storage of Metal Ions in Biological Cells.* s.l. : Royal Society of Chemistry, 2014.

4. *Biological coordination chemistry of magnesium, sodium, and potassium ions. Protein and nucleotide binding sites.* **C.B. Black, H.-W. Huang, J.A. Cowan.** s.l. : Elsevier, 1994, Coordination Chemistry Reviews, Vol. 135/136, pp. 165–202.

5. *Magnesium chemistry and biochemistry.* **Michael E. Maguire, James A. Cowan.** s.l. : Springer, 2002, BioMetals, Vol. 15, pp. 203–210.

6. **Rankin, WJ.** *Minerals, Metals and Sustainability. Meeting Future Material Needs.* s.l. : CSIRO PUBLISHING, 2011.

7. *Metal ions in biological catalysis: from enzyme databases to general principles.* **Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM.** s.l. : Society of Biological Inorganic Chemistry, 2008, Journal of Biological Inorganic Chemistry, Vol. 13, pp. 1205–1218.

8. **Brown, David R.** *Brain Diseases and Metalloproteins.* Singapore : Pan Stanford, 2012.

9. **Pomogailo, A. and Wöhrle, D.** *Metal Complexes and Metals in Macromolecules: Synthesis, Structure and Properties.* Weinheim : Wiley-VCH Verlag GmbH & Co., 2003.

10. *Structural and functional aspects of metal sites in biology.* **Holm R.H., Kennepohl P., Solomon E.I.** s.l. : ACS Publications, 1996, Chemical Reviews, Vol. 96, pp. 2239–2314.

11. **Eugene A. Permyakov, Vladimir N. Uversky.** *Metalloproteomics.* Hoboken, New Jersey : John Wiley & Sons, 2009.

12. **Harrison P.M., Hoare R.J.** *Metals in biochemistry.* New York : Chapman and Hall, 1980.

13. **M.Roat-Malone, Rosette.** *Bioinorganic Chemistry: A Short Course.* 2nd. Hoboken, New Jersey : A John Wiley & Sons, Inc., 2007.

14. **Anfinsen, Christian B.** *Metalloproteins : Structural Aspects.* San Diego, California : Academic Press, 1991.

15. **Mildvan, A.S.** *Metals in enzyme catalysis. The Enzymes.* New York : Academic Press, 1970. Vol. 2.

16. **Constable, Edwin C.** *Metals and Ligard Reactivity: An Introduction to the Organic Chemistry of Metal Complexes.* s.l. : Wiley-VCH, 1996.

17. **Hanzlik, Robert P.** *Inorganic Aspects of Biological and Organic Chemistry.* s.l. : Academic Press Inc., 1976.

18. **Paul C. Painter, Michael M. Coleman.** *Essentials of Polymer Science and Engineering.* [ed.] Inc DEStech Publications. 2008.

19. *Hydrogen Bonds in Proteins: Role and Strength.* **Roderick E Hubbard, Muhammad Kamran Haider.** s.l. : Wiley Online Library, 2010, eLS.

20. *Zinc Coordination Geometry and Ligand Binding Affinity: The Structural and Kinetic Analysis of the Second-shell Serine 228 Residue and the Methionine 180 Residue of the Aminopeptidase from Vibrio proteolyticus.* **Ataie, N.J., et al., et al.** 2008, Biochemistry, Vol. 47, pp. 7673–7683.

21. *First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations.* **Dudev T., Lin Y. L., Dudev M., Lim C.** s.l. : Journal of the American Chemical Society, 2003, Vol. 125, pp. 3168-3180.

22. *Metals in protein structures: a review of their principal features.* **Marjorie M. Harding, Matthew W. Nowicki, Malcolm D. Walkinshaw.** 4, 2010, Crystallography Reviews, Vol. 16, pp. 247-302.

23. *Metalloproteomes: A Bioinformatic Approach.* **Andrein C., Bertini I., and Rosato A.** 10, s.l. : American Chemical Society, 2009, Vol. 42.

24. *Prediction of zinc-binding sites in proteins from sequence.* **Shu N., Zhou T., Hovmoller S.** 24, s.l. : Bioinformatics, 2008, pp. 775-782.

25. *A hint to search for metalloproteins in gene banks.* **Andreini C., Bertini I., Rosato A.** 2004, Bioinformatics, Vol. 20, pp. 1373–1380.

26. *Prediction of zinc-binding sites in proteins from sequence.* **Nanjiang Shu, Tuping Zhou and Sven Hovmöller.** 6, 2008, Bioinformatics, Vol. 24, pp. 775-782.

27. *Identifying Cysteines and Histidines in Transition-Metal-Binding Sites Using Support Vector Machines and Neural Networks.* **A. Passerini, M. Punta, A. Ceroni, B. Rost, and P. Frasconi.** 2, 2006, Proteins: Structure, Function, and Bioinformatics, Vol. 65, pp. 305-316.

28. *Improving Prediction of Zinc Binding Sites by Modeling the Linkage between Residues Close in Sequence.* **Sauro Menchetti, Andrea Passerini, Paolo Frasconi, Claudia Andreini, Antonio Rosato.** 2006. Research in Computational Molecular Biology. Vol. 3909, pp. 309-320.

29. *Predicting Metal Binding Sites from Protein Sequence.* **A. Passerini, M. Lippi, P. Frasconi.** 1, 2012, IEEE Transactions on Computational Biology and Bioinformatics, Vol. 9, pp. 203-213.

30. *Protein metal binding residue prediction based on neural networks.* **Lin, C.T., et al., et al.** 71, s.l. : International Journal of Neural Systems, 2005, Vol. 15.

31. *ProFunc: a server for predicting protein function from 3D structure.* **Laskowski, R.A., Watson,J.D., Thornton, J.M.** s.l. : Oxford Journals, 2005, Nucleic Acids Research, Vol. 33, pp. 89–93.

32. *The RCSB Protein Data Bank: redesigned web site and web services.* **Rose, P. W., et al., et al.** s.l. : Oxford Journals, 2011, Nucleic Acids Research, Vol. 39, pp. D392-D401.

33. *The extended environment of mononuclear metal centers in protein structures.* **Karlin S., Zhu Z.Y., Karlin K.D.** U.S.A : Proceedings of the National Academy of Sciences, 1997, Vol. 94.

34. *Metal binding affinity and selectivity in metalloproteins: insights from computational studies.* **Dudev T., Lim C.** s.l. : Annual Review of Biophysics, 2008, Vol. 37, pp. 97-116.

35. *Where metal-ions bind in proteins.* **Yamashita, M.M., Wesson, L., Eisenman, G., Eisenberg, D.** 87, s.l. : Kenneth R. Fulton, 1990, Proceedings of the National Academy of Sciences, pp. 5648–5652.

36. *The prediction and characterization of metal binding sites in proteins.* **Gregory, D.S., Martin, A.C., Cheetham, J.C., Rees, A.R.** s.l. : Oxford Journals, 1993, Protein Engineering Design and Selection, Vol. 6, pp. 29–35.

37. *Minimal functional sites allow a classification of zinc sites in proteins.* **Andreini C., Bertini I., Cavallaro G.** s.l. : PLoS One, 2011, Vol. 10.

38. *Copper(I)-mediated protein-protein interactions result from suboptimal interaction surfaces.* **Banci, L., et al., et al.** s.l. : Portland Press, 2009, Biochemical Journal, Vol. 422, pp. 37-42.

39. *Characterization of the MMP-12-elastin adduct.* **Bertini, I., et al., et al.** s.l. : Wiley-VCH, 2009, Chemistry: A European Journal, Vol. 15, pp. 7842-7845.

40. *Activities at the Universal Protein Resource (UniProt).* **Consortium, The UniProt.** 2014, Nucleic Acids Research.

41. *New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.* **Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA.** s.l. : Oxford Journals, 2013, Nucleic Acids Research.

42. *SCOP2 prototype: a new approach to protein structure mining.* **Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, Alexey G. Murzin.** 2014, Nucleic Acids Research, Vol. 42, pp. D310-D314.

43. *The Pfam protein families database.* **R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta.** 2014 : s.n., Nucleic Acids Research , Vol. 42, pp. D222-D230.

44. *PROMISE: a database of bioinorganic motifs.* **Degtyarenko KN, North AC, Findlay JB.** 1, 1999, Nucleic Acids Research, Vol. 27, pp. 233–236.

45. *MDB: the Metalloprotein Database and Browser at The Scripps Research Institute.* **Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME.** 1, 2002, Nucleic Acids Research, Vol. 30, pp. 379-382.

46. *MetalPDB: a database of metal sites in biological macromolecular structures.* **Andreini, C., et al., et al.** s.l. : Oxford Journals, 2013, Nucleic Acids Research, Vol. 41, pp. D312-D319.

47. *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* **Murzin A. G., Brenner S. E., Hubbard T., Chothia C.** s.l. : Elsevier, 1995, Journal of Molecular Biology, Vol. 247, pp. 536-540.

48. *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.

49. *Metal-MACiE: a database of metals involved in biological catalysis.* **Andreini, C., Bertini,I., Cavallaro,G., Holliday,G.L., & Thornton,J.M.** s.l. : Oxford Journals, 2009, Bioinformatics, Vol. 25, pp. 2088-2089.

50. *Evolution of protein function, from a structural perspective.* **Todd, A.E., Orengo,C.A. and Thornton,J.M.** s.l. : Elsevier, 1999, Current Opinion in Chemical Biology, Vol. 3, pp. 548–556.

51. *Mechanistic Diversity in a Metalloenzyme Superfamily.* **Armstrong, Richard N.** 45, s.l. : American Chemical Society, 2000, Biochemistry, Vol. 39, pp. 13625–13632.

52. *Divergence and Convergence in Enzyme Evolution.* **Koonin, Michael Y. Galperin and Eugene V.** 1, s.l. : American Society for Biochemistry and Molecular Biology, 2012, The Journal of Biological Chemistry, Vol. 287, pp. 21-28.

53. *Catalysing new reactions during evolution: economy of residues and mechanism.* **Bartlett, G.J., Borkakoti,N. and Thornton,J.M.** s.l. : Elsevier, 2003, Journal of Molecular Biology, Vol. 331, pp. 829–860.

54. *A gold standard set of mechanistically diverse enzyme superfamilies.* **Brown, S.D., Gerlt,J.A., Seffernick,J.L. and Babbitt,P.C.** s.l. : BioMed Central, 2006, Genome Biology, Vol. 7.

55. *FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies.* **Nicholas Furnham, Ian Sillitoe, Gemma L. Holliday, Alison L. Cuff, Syed A.**

Rahman, Roman A. Laskowski, Christine A. Orengo and Janet M. Thornton. 42, s.l. : Oxford Journals, 2011, Nucleic Acids Research, pp. D776-D782.

56. *Recognition of Errors in the Three-Dimensional Structures.* **M.J., Sippl.** s.l. : Proteins: Structure, Function, and Bioinformatics, 1993, Vol. 17, pp. 355-362.

57. *On the orthogonal transformation used for structural comparisons.* **S.K., Kearsley.** 2, s.l. : John Wiley & Sons Ltd, 1989, Acta Crystallographica Section A, Vol. 45, pp. 208-210.

58. **Hamilton W.R., Hamilton W.E.** *Elements of quaternions.* London, New York & Bombay : Longmans, Green, & Company, 1866.

59. *Uber ein leichtes Verfahren, die in der Theorie der Sakularstorungen vorkommenden Gleichungen numerisch aufzulosen.* **C.G.J., Jacobi.** s.l. : Walter de Gruyter, 1846, Journal fur die Reine und Angewandte Mathematik, Vol. 30, pp. 51-95.

60. *Multidimensional binary search trees used for associative searching.* **L., Bentley J.** 9, s.l. : Association for Computing Machinery, 1975, Communications of the ACM, Vol. 18, p. 509.

61. **Golub G.H., Van Loan C.F.** *Matrix Computation.* 3rd. Baltimore : Johns Hopkins Univ. Press, 1996.

62. **Hastie, T., Tibshirani,R. & Friedman,J.** *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* s.l. : Springer Series in Statistics, 2009.

63. *p3d--Python module for structural bioinformatics.* **Fufezan, C. and Specht, M.** s.l. : BioMed Central, 2009, BMC Bioinformatics, Vol. 10.

64. *Data mining of metal ion environments present in protein structures.* **Heping Zheng, Maksymilian Chruszcz, Piotr Lasota, Lukasz Lebioda,and Wladek Minor.** 9, 2008, Journal of Inorganic Biochemistry, Vol. 102, pp. 1765–1776.