JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. XXXX, XXX, XXX-XXX

pubs.acs.org/jcim

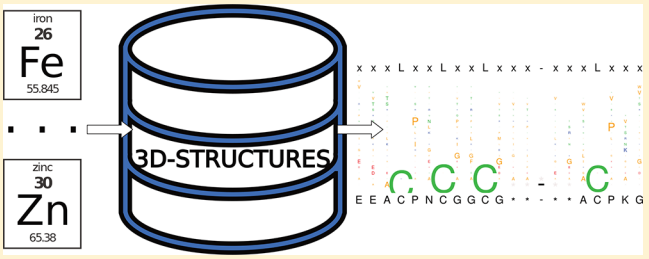
Patterns of Ligands Coordinated to Metallocofactors Extracted from the Protein Data Bank

Luca Belmonte[†] and Sheref S. Mansy*

CIBIO, University of Trento, Via Sommarive 9, 38123 Povo, Italy

Supporting Information

ABSTRACT: A new R tool is described that rapidly identifies, ranks, and clusters sequence patterns coordinated to metallocofactors. This tool, PdPDB, fills a void because, unlike currently available tools, PdPDB searches through sequences with metal coordination as the primary determinant and can identify patterns consisting of amino acids, nucleotides, and small molecule ligands at once. PdPDB was tested by analyzing structures that coordinate $Fe^{2+/3+}$, [2Fe-2S], [4Fe-4S], Zn^{2+}, and Mg^{2+} cofactors. PdPDB confirmed previously identified sequence motifs and revealed which residues are enriched (e.g., glycine) and are under-represented (e.g., glutamine) near ligands to metal centers. The data show the similarities and differences between different metal-binding sites. The patterns that coordinate metallocofactors vary, depending upon whether the metal ions play a structural or catalytic role, with catalytic metal centers exhibiting partial coordination by small molecule ligands. PdPDB 2.0.1 is freely available as a CRAN package.



INTRODUCTION

Metal ions are essential for living cells, and approximately half of the protein structures deposited in the Protein Data Bank (PDB) coordinate a metallocofactor.[1] Metal ion binding motifs can be extraordinarily conserved. For example, the metal binding motif CX_2C, where X is any amino acid, is found in all three Kingdoms of life in a manner that is not correlated with a specific tertiary fold[2] nor a specific metal ion. However, the simplified view of metal binding sites as a protein sequence motif ignores features that are critically important for activity. For example, in addition to ligation by amino acid side-chains, metal ions are often coordinated to small molecules (e.g., H_2O) or nucleic acids. Such information is frequently lost from a strictly motif-based analysis.

Patterns, rather than motifs, give a more complete picture of how metal ions are coordinated to biological molecules. Here, a pattern denotes a sequence potentially consisting of a mixture of protein and nucleic acid residues plus anything else that ligates the metal center, including small molecules. A ligand is a molecule that binds a metal center. Patterns provide a better picture of how metallocofactors coordinate to biological ligands, because patterns reveal the complete coordination sphere of the metal center. Such information is important for those interested in bioinorganic chemistry, the de novo synthesis of enzymes, and the role of metal ions in the origins of life.[3] However, there are no modern tools to identify conserved, metal ion coordinating patterns within proteins. The metalloprotein database and browser (MDB) was useful in identifying metal ligands, metal geometry, and residues that interacted with metal ligands,[4] but this tool is no longer available. Similarly, Spratt2[5] provided results based on the coordination sphere of the metal ion; however, Spratt2 is no longer accessible. There are currently available powerful tools that detect motifs, such as MOTIF (http://www.genome.jp/tools/motif/) and ASSAM,[6] but these tools are driven by the structure of the proteins and not the coordinated metal ions. PDBeMotif[7] does search for target motifs within metalloenzymes, but the complete sequence responsible for metal coordination is missing in the output, since this software is devoted to motif recognition and not pattern extraction. Also, the currently available software cannot handle mixed sequences,[8] i.e., information containing a mixture of amino acids, nucleotides, and small molecule ligands and are typically devoted to the analysis of a single structure. In summary, previously available tools could analyze the metal binding site of a single sequence or analyze a library of sequences without consideration of coordinated metal ions. What is lacking is an accessible method to easily identify the frequency of ligand-containing sequences associated with queried metallocofactors in a manner that can handle proteins, nucleic acids, and small molecule ligands simultaneously.

To rapidly identify patterns associated with the coordination of specific metal ions, we devised a new tool (PdPDB, Pattern discovery in PDB structures of metalloproteins). PdPDB extracts all the existing patterns associated with the queried metal ion from one or more PDB entries. The PDB is used, because PDB files contain all of the necessary information to decipher the complete coordination sphere of the metal center. Since PdPDB uses the metallocofactor as the primary

Received: August 4, 2017
Published: November 8, 2017

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

determinant, sequences are not required to be homologous on either a sequence or a structural level. Sequences that do not coordinate the queried metallocofactor are discarded. The ligands of the extracted patterns are then aligned. To run PdPDB, the user must input the PDB chemical name of the metallocofactor of interest, the size of the sequence window containing the ligand to be explored, and a trimming threshold.

PdPDB was used to analyze the ligand patterns associated with the binding of $Fe^{2+/3+}$, [2Fe-2S], [4Fe-4S], Zn^{2+}, and Mg^{2+}. Zn^{2+} and the Fe ions of [4Fe-4S], [2Fe-2S], and mononuclear, rubredoxin-like iron−sulfur clusters are often tetrahedrally coordinated, whereas Mg^{2+} typically assumes an octahedral geometry. Mononuclear, noncysteine coordinated $Fe^{2+/3+}$ is more heterogeneous in geometry with tetrahedral and octahedral geometries being more frequently observed. Coordinated metal ions can either play a structural or a catalytic role. The data confirm the ubiquity of CX_2C, forming part of the coordinating patterns for a significant fraction of the sequences that coordinate all of the investigated metallocofactors except for Mg^{2+}. Nonproteinaceous ligands, including substrate molecules and water, were frequently observed for metal ions within the active sites of enzymes, whereas the coordination of metal ions that served to stabilize the folding of the protein were completely coordinated by proteinaceous ligands. Differences were also observed within the surrounding residues of the ligands in a manner specific for the coordinated metallocofactor.

METHODS

Algorithm and Implementation. PdPDB 2.0.1 is freely available as a CRAN package. The workflow of PdPDB (Figure S1 in the Supporting Information) takes advantage of some of the features of the PDB standards. First, PdPDB identifies the ligands (L) to a given metal (m) by searching the PDB files for the keyword LINK. Next, PdPDB identifies the residues in the $\pm n$ positions from L, where $+n$ are the n residues that follow (f) and $-n$ are the residues that precede (p) L. This last step is not performed if L is not part of the primary structure of a protein or nucleic acid.

A sequence pattern is defined as a sequence of ligands interspersed with a maximum of p and f amino acids or nucleotides. The maximum number of residues in a pattern is given as $(n \times 2 + 1) \times n_L$, where n_L is the number of ligands. The pattern conserves the same order found in the protein or RNA/DNA chain. Then, if consecutive f and p residues lie beyond the user-defined window, but in the same chain of the macromolecule, a dash ("−") is inserted between the f and p. Conversely, if consecutive f and p residues in a pattern belong to different chains, plus signs ("+") are used instead of a dash. For example, for $n = 1$, the tetrahedral coordination of a given metal would appear as a sequence of four blocks, where a block is a pLf short sequence, namely, $p_{n1}L_1f_{n1}p_{n2}L_2f_{n2}p_{n3}L_3f_{n3}p_{n4}L_4f_{n4}$. This pattern would appear as $p_{n1}L_1f_{n1}-p_{n2}L_2f_{n2}+p_{n3}L_3f_{n3}p_{n4}L_4f_{n4}$ if f_{n1} and p_{n2} are more than one position apart within the same monomer, and L_1 and L_2 belong to a different monomer than L_3 and L_4. More generally, this specific pattern would appear as $xLxxLxxLxxLxx$, where x is any amino acid and L is the ligand. PdPDB highlights the ligands within patterns by brackets and by L in logo headers. Note that p and f residues may be missing in the output, because L could be at the N- or C-terminus of the protein, for example, or because p and f are less than n amino acids apart. Since a structure may contain more than a single metallocofactor with the same chemical identification (ID) (e.g., two [4Fe-4S] clusters are coordinated to *Clostridium pasteurianum* ferredoxin), the PDB peptide chain and residue number of the metal are combined to generate a unique key that identifies the pattern. If problems are encountered due to improper formatting of patterns, e.g., if preceding and following residues overlap, because of a window size that is too large, PdPDB gives a warning. In addition, PdPDB discards PDB entries in which the target metal ID is not contained as a whole in the LINK field. PdPDB handles both ".pdb" and ".cif" formats.

Additional Symbols. FASTA one letter codes are used to identify amino acids and nucleotides. Once the patterns are extracted, the three-letter codes used within the PDB files are translated to FASTA one-letter codes. Uppercase one-letter codes signify amino acids and lowercase one-letter codes indicate nucleotides. Water and small molecule ligands are labeled as O and Z, respectively. Up to nine additional symbols can be defined by the user.

Ligand Alignment. To align the patterns, the pattern matrix (P) is filled with the patterns extracted as described above. P is a $k \times m$ matrix, where k is the number of patterns and m the maximum number of residues found in the patterns. Each cell of the matrix contains a residue, either an amino acid, nucleotide, or small molecule. Extracted patterns do not always show ligated residues in the correct positions and thus adjustments must be made.

Once P has been filled by patterns, ligands are forced in the positions given by a series defined as follows:

$$l(j) = l_j = l_{j-1} + b \qquad \forall\, j \in \mathbb{N} \tag{1}$$

where l_j is the generic ligand position of the jth ligand in the pattern, and b is the block length given as $b = n \times 2 + 2$, that allows a space in the block length. Note that b refers to the single L block (e.g., pLf).

The set of possible ligand positions in P is defined as

$$\mathrm{L} = \{l_j | l_j = l_{j-1} + b \wedge l_1 \le l \le m - n\} \qquad \forall\, j \in \mathbb{N} \tag{2}$$

where l_1 is the starting value of the series given as $l(1) = l_1 = b - n - 1$ representing the position of the first L in the pattern.

The set of the generic x positions of the X residues within a pattern is defined as

$$\mathrm{X} = \{x | x \le m\} \qquad \forall\, x \in \mathbb{N} \tag{3}$$

Consequently, L ⊆ X.

We also define the shift function S, for misplaced ligands, as a function of x, as follows:

$$S(x) = l_{j+1} - x \qquad \forall\, x \notin \mathrm{L} \tag{4}$$

This applies only to L parameters in a generic x position or, in other words, in a position not given by the l-series. Once S is applied, the P elements in the range of positions $[x, l_{j+1}]$ are filled with gaps, while the misplaced L is placed at l_{j+1}. Thus, for the generic ith row, the matrix elements are

$$p_{i,y} = f(x) = \begin{cases} *, & y = x, \ldots, l_{j+1} - 1 \\ \mathrm{L}, & y = l_{j+1} \end{cases} \tag{5}$$

where y is the column index, l_{j+1} is the next L position, and the asterisk (*) denotes PdPDB-inserted gaps.

Afterward, the residues between L parameters are rearranged in order to equilibrate their distributions in the pattern. To do

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

so, a heuristic is used. Nonligated residues are assumed to be equally distributed between two ligated residues. In other words, PdPDB adjust patterns so that the same n number of residues follow/precede each L. Thus, if the generic ith row ends with * followed by a L, or, in symbols, $p_{i,l_{j+1}} - 1 = *$ and L is not a small molecule, a rearrangement of the elements occurs. The rearrangement is performed for the range (l_j, l_{j+1}) and causes half of the residues to move toward the position l_{j+1}. Gaps are put in place of moved residues. Thus,

$$p_{i,l_{j+1}-1+z} = p_{i,l_j+n-z} \quad \text{and} \quad p_{i,l_j+n-z} = * \qquad \text{with } z = 0, \ldots, \frac{(n-1)}{2} \tag{6}$$

For example, a sequence made of two XLX blocks (i.e., with n = 1) may appear as XLXX**LX after ligand alignment. PdPDB updates the sequence as XLX**XLX, using the so-far-described heuristic. Spacing between the two blocks are also allowed, such as XLX*−*XLX or XLX*+*XLX, if the two blocks are from two different monomers of the same protein. This process is iterated across the entire P. The heuristic is not applied in the range (l_j, l_{j+1}), if only * are contained. A version of the heuristic that works in reverse (i.e., for rows that contain a gap close to the first ligated residue) is used to prevent errors within the first n columns of P.

The matrix of aligned sequences is then trimmed based on the frequency of appearance of symbols. In other words, PdPDB discards columns that do not contain at least a given number of nongap characters. The frequency of appearance that is required for retention is user-defined. Here, the lowest frequency of appearance that was retained was 50%, unless specified otherwise. When PdPDB is applied to a set of structures, the results contain regular matrices where all the ligands fill well-defined positions (Figure S2 in the Supporting Information). PdPDB users must be aware that trimming also affects the spacing between chains.

Enrichment Score. For each position, the frequency of appearance of each amino acid is calculated. This information is used to calculate the frequency (%) that is relative to the number of positions in the pattern matrix P. The number of positions in the pattern matrix is relative to the dimensions of the matrix and varies, depending on the inspecting window and the trimming threshold. Gaps, which are artificially inserted by PdPDB, are not taken into account in the computation of the percentage. Aligned sequences were further compared with the average frequency of appearance of each amino acid residue in ExPASy.[9] The comparison with the average frequency in ExPASy is the default setting; however, the user may define a custom reference. The two-sided proportional test used by PdPDB is an approximation of the χ^2 test, as implemented in R.[10] The resulting p-values from the χ^2 tests are then corrected with the Bonferroni correction.[11] A further comparison on the sameness of the distribution is performed by the Mann−Whitney test.[12,13] PdPDB also compares the normality of the distribution by the Jarque−Bera test.[14] The entire comparison is iterated for the entire specimen, the ligands, the nonligands, and for residues up to n of the following and preceding positions. Furthermore, PdPDB also computes a z-score (also known as a standard score) for each proportion obtained in the specimen versus the reference. The zth element of a Z-distribution is given as

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7}$$

where p_1 is the proportion obtained from the PdPDB outcome, p_2 is the reference proportion, n_1 is the number of samples comprising the PdPDB specimen, n_2 is the number of samples composing the reference, and p is the combined proportion. For each amino acid residue, proportions are plotted for visual inspection.

Clustering. PdPDB extracts consensus sequences for each possible cluster (or family) of sequences. Consensus sequences consist of the most frequent residues per position. If a position is dominated by more than one residue with the same frequency, then this position is marked with an X. PdPDB exploits hierarchical clustering, as implemented in R.[15,16] This method exploits the Levenshtein distance matrix to compare pattern sequences on top of which the dendrogram is built.[17] The user decides where the dendrogram is cut to produce clusters of sequences. For each cluster, along with the consensus sequence, a logo is computed. A logo is drafted as follows: on the x-axis, the consensus sequence is drawn. For each position of the consensus sequence, alternative residues that can be found in that specific position are plotted. The amino acid frequency is represented by the dimension of the text; therefore, the bigger the symbol, the higher the frequency. Gaps can appear in both logos and consensus sequences. The higher the trimming threshold, the lower the chance of having gaps in the logos and vice versa. With thresholds >1%, such as those used in the present study, the meaning of gaps in consensus sequences means that the amino acids are contiguous in the P matrix and are not necessarily in the original protein sequences.

Dendrogram cuts to obtain the families presented here are shown in Figure S3 in the Supporting Information.

Ranking. PdPDB ranks patterns according to their frequency of appearance with a cumulative function as follows:

$$F(\mu) = \sum_{j}^{N} \sum_{i}^{n} H(\mu_i) \tag{8}$$

where μ is a generic pattern, n is the number of motifs in a single structure, and N is the number of entries. H is the hit function, which is defined as

$$H(\mu_i) = \begin{cases} 1 \\ 0 \qquad \text{if it is not found} \end{cases} \tag{9}$$

The patterns are ranked according to their frequency of appearance given by F. PdPDB retained identical patterns extracted from the same input structure, even though it is possible to discard duplicates.

Datasets. PdPDB was run on datasets consisting of PDB files that were deposited before February 2017. The dataset only contained structures solved by X-ray crystallography. The structures were only of wild-type sequences that did not contain sequence tags of any kind, and only sequences that had 100% identity to their UniProt reference sequence were downloaded. Datasets consisted of structures coordinated to a given metal; e.g., zinc, or were annotated as belonging to a given class of enzyme, e.g., RNA polymerase. A summary of the analyzed PDB entries is shown in Table S1 in the Supporting Information. Note that the analysis of the aconitase sequence

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

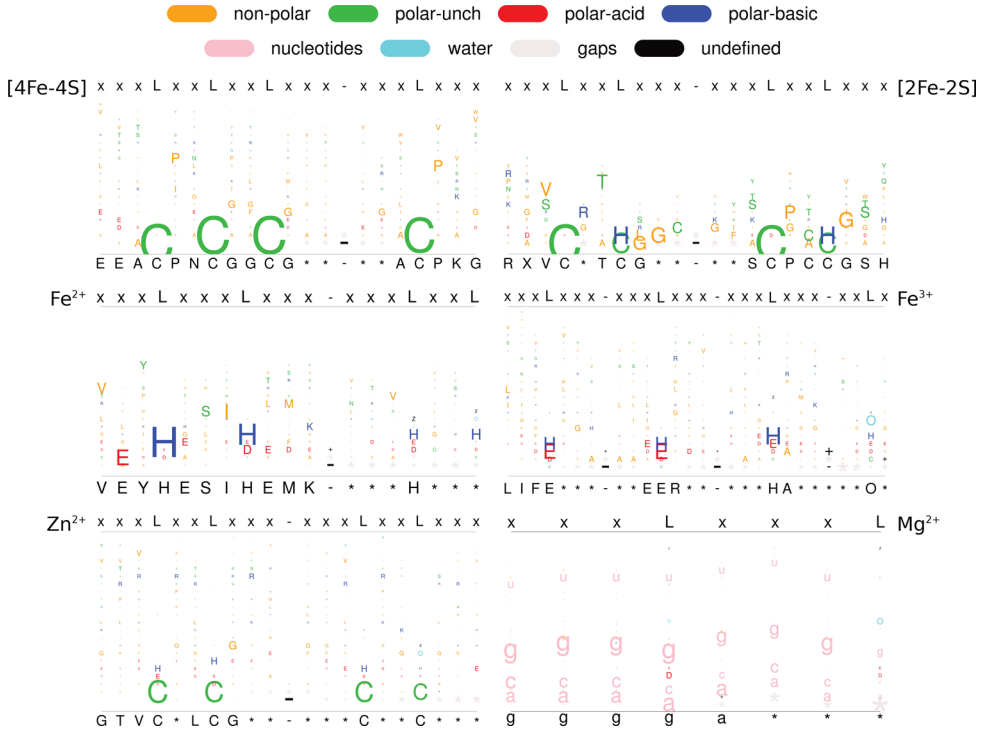


Figure 1. Root consensus sequences and associated logos extracted from metallocofactor containing structures deposited in the PDB. Data for [4Fe-4S], [2Fe-2S], Fe^{2+}, Fe^{3+}, Zn^{2+}, and Mg^{2+} protein ensembles are shown. The consensus sequences and logos were built from 131 sequences extracted from 36 PDB protein structures ([4Fe-4S]), 45 sequences extracted from 24 protein structures ([2Fe-2S]), 62 sequences extracted from 21 protein structures (Fe^{2+}), 202 sequences extracted from 49 protein structures (Fe^{3+}), 1283 sequences extracted from 315 protein sequences (Zn^{2+}), and 3140 sequences extracted from 284 protein structures (Mg^{2+}). The window size was $n = \pm 3$ from each ligand, and the trimming function was set to 50%. The consensus sequence is shown at the bottom of each panel. L signifies a ligand, *x* is any amino acid, O symbolizes water, "unch" indicates uncharged, a dash symbol (−) indicates spacing greater than the window size within the same chain, and a plus symbol (+) signifies different chains. Detailed patterns (i.e., with − and +) are provided by PdPDB as tabulated files. Amino acids are in uppercase, and nucleotides are shown in lowercase.

did not follow these criteria, because the solved structures of aconitase contained amino acid substitutions.

RESULTS AND DISCUSSION

The functionality of PdPDB was first tested by analyzing the [4Fe-4S] (PDB chemical ID: SF4) proteins deposited in the PDB. [4Fe-4S] proteins were chosen as a target, because [4Fe-4S] ferredoxins are known to frequently exploit a CX_2CX_2C motif for the coordination of the iron−sulfur cluster.[18] The fourth cysteinyl ligand needed to fully coordinate the iron−sulfur cluster either precedes ($CP\cdots CX_2CX_2C$) or follows ($CX_2CX_2C\cdots CP$) the CX_2CX_2C motif.[18] Therefore, PdPDB would be expected to identify coordinating patterns compatible with these known CX_2CX_2C motifs. PdPDB was run with a window size of three and the trimming function set to 50%, meaning that only residues ±3 away from the ligand that were conserved in at least 50% of the sequences analyzed were retained. These two parameters affect each other in that positions further away from the ligand may be less conserved and therefore not observable with a high value inserted for the trimming function. In the case of [4Fe-4S] proteins, the exploited parameters gave an uncut dendrogram with a root consensus sequence of EEA(C)PN(C)GG(C)G−A(C)PKG (ligands are given in parentheses), which fit well the known [4Fe-4S] ferredoxin motif in which CP either precedes or follows the core CX_2CX_2C motif (see Figure 1).

The sequences extracted from the PDB files could be divided into four different families (see Figure S4 in the Supporting Information). The consensus sequences of these four different families were LVS(C)SN(C)PG(C)DXP−XGL(C)AAW (39 sequences), ADT(C)IG(C)GX(C)XAA−VSV(C)PVG (42 sequences), VTA(C)PA(C)IF(C)GA(C)VNV (40 sequences), and EVX(C)XC(C)XF(C)X−(C)/(Z) (10 sequences), where Z represents quinolinic acid (Figure S4d). Taken together, 87% of the analyzed sequences contained a CX_2C ligand motif, half of which were part of a longer CX_2CX_2C motif (Table S2 in the Supporting Information). The analysis of [4Fe-4S] proteins with PdPDB shows how the PdPDB algorithm can be used to successfully identify coordinating patterns associated with the binding of metal ions, and that PdPDB can detect coordinating patterns that exploit small molecule ligands. The latter feature is not possible by exploiting BLAST-type[19,20] searches through sequences of proteins or nucleic acids.

We next sought to analyze a related class of iron−sulfur proteins that coordinates a [2Fe-2S] cluster (PDB chemical ID: FES). In this case, PdPDB gave a root, dendrogram consensus sequence of RXV(C)T(C)G−S(C)PC(C)GSH (Figure 1). Although there were not proteins that contained a CX_2CX_2C motif within the 45 sequences used to generate the root consensus sequence, sequences could be described as containing a shorter, persistent CX_2C motif (47% of the analyzed [2Fe-2S] sequences). The [2Fe-2S] protein dataset could be divided into four families of sequences. Two of these families of sequences completely ligated the [2Fe-2S] cluster with four cysteine residues, with one family exploiting a $CX_2C\cdots CX_2C$ motif and the other family only containing a

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

single CX_2C motif (see Figure S5 in the Supporting Information). The two remaining families of sequences exploited two histidine and two cysteine ligands to coordinate the [2Fe-2S] cluster, both containing a single CX_2H motif; 46% of the analyzed [2Fe-2S] coordinating sequences contained this CX_2H ligand motif (see Table S2 in the Supporting Information). Rieske-type of proteins are known to exploit CXH···CX_2H sequences for the coordination of a [2Fe-2S] cluster.[21] No structures were identified that exploited non-proteinaceous ligands.

To better characterize the coordination of Fe ions, PdPDB was used to investigate nonheme, mononuclear iron-binding proteins (PDB chemical IDs: FE2 for the Fe^{2+} ion and FE for the Fe^{3+} ion). $Fe^{2+/3+}$ showed a greater degree of sequence diversity than that of [4Fe-4S] or [2Fe-2S] proteins. Nevertheless, the root dendrogram sequence of both ensembles showed a propensity for coordination by histidine, glutamate, and water ligands (Figure 1). The data were consistent with what is known about the coordination of mononuclear, nonheme Fe ions. That is, a variety of ligands and geometries are associated with the coordination of $Fe^{2+/3+}$ centers, including coordination by different combinations of one or two histidines or with one or two aspartate or glutamate residues.[22] In addition, the coordination number can vary between four and six with small molecules, including water, functioning as the remaining ligands. When the Fe^{2+} and Fe^{3+} datasets were divided into families of sequences, additional small molecule ligands were detected, including ATP, lysine NZ-carboxylic acid, and acetone (see Figures S6 and S7, as well as Table S3 in the Supporting Information). Coordinating patterns with more than four ligands were only identified if the trimming function was reduced to 5% (see Figures S8 and S9 in the Supporting Information). Note that one family of sequences from the Fe^{3+} dataset was significantly different in that this family could exploit four cysteine ligands to coordinate the Fe ion (see Figure S9). The data were consistent with a class of iron−sulfur proteins known as rubredoxins. When PdPDB was used to analyze only structures annotated as rubredoxins, a CX_2C···CX_2C motif emerged (see Figure S10 in the Supporting Information).

It has been hypothesized that modern-day Zn^{2+} and iron−sulfur proteins emerged from short CX_2CG and CX_2CX_2C peptide ancestors, respectively.[23] Therefore, we wondered if PdPDB would detect similar coordinating patterns for Zn^{2+}-binding proteins (PDB chemical ID: ZN), as was observed for [2Fe-2S] and [4Fe-4S] proteins. The root dendrogram consensus sequences was composed of 1283 sequences. The Zn^{2+} consensus sequence showed predominantly ligation by four cysteine residues, although ligation by histidine, glutamate, aspartate, and water was also detected (Figure 1). The classical Zn^{2+} finger binding motif coordinates Zn^{2+} through either two cysteines and two histidines,[24] three cysteines and one histidine,[25−27] or by four cysteines.[28] If the Zn^{2+}-containing sequences were split into three families, 718 of the 1283 total sequences clustered into one family with a consensus sequence of GRR(C)RR(C)GR−FR(C)RG(C)SRE (see Figure S11 in the Supporting Information). That is, the majority of the Zn^{2+}-binding proteins analyzed contained a sequence pattern that consisted, in part, of the putative ancestral CX_2CG peptide sequence and was compatible with the CX_2C···CX_2C motif associated with the binding of iron−sulfur clusters. In fact, 40% of the analyzed Zn^{2+}-binding sequences contained the same CX_2C···CX_2C ligand motif as found in rubredoxin (see Table S2 in the Supporting Information). This sequence pattern was also consistent with the well-characterized GATA-type zinc finger motif.[29] One of the remaining families of sequences was dominated by three histidine ligands and one ligand position populated by a water molecule or another small molecule. One out of the three families of Zn^{2+}-binding sequences had a less well-defined coordination sphere (see Figure S11b in the Supporting Information).

To compare the ligand patterns associated with a completely different group of metal ions, PdPDB was used to characterize structures containing the alkaline-earth-metal Mg^{2+} (PDB chemical ID: MG). The Mg^{2+} coordinating sequences were much more diverse than the transition-metal complexes analyzed above and, therefore, did not yield a clear consensus pattern (Figure 1). Coordination by guanosine was predominant with some examples of ligation by aspartate, glutamate, and water (see Figure S12 in the Supporting Information). Ligation by aspartate was consistent with a known DXDXD motif associated with the binding of Mg^{2+}.[30] Nevertheless, none of the consensus sequences accounted for all of the ligands of hexacoordinate Mg^{2+}. If the window of ±3 from the ligand was reduced to ±1, then the diversity present in the sequences surrounding the ligand interfered less (see Figure S13). However, many of the coordinating patterns continued to lack all of the ligands necessary to fully coordinate Mg^{2+}. The lack of well-defined sequence motifs associated with Mg^{2+} not only reflects the sequence heterogeneity of Mg^{2+}-binding proteins and nucleic acids but also likely arises as a consequence of a predominant coordination by nonproteinaceous ligands, including water, and the known difficulties of assigning the electron densities of Mg^{2+} and water.[31]

Since PdPDB provides the complete list of coordinating patterns from which the motifs are generated, it was possible to search for coordinating patterns that provided a complete coordination sphere from the output, even if such patterns were not frequent enough to generate a motif. The most common ligand pattern associated with hexacoordinate Mg^{2+} was (O)...(O)...(O)...(O)...(O)...(O), where O indicates a water molecule. Complete ligation by water molecules was consistent with the fact that Mg^{2+} often coordinates to proteins through a shell of water molecules, i.e. through outer-sphere coordination.[32] The second-most common, fully coordinated pattern was I(D)G−F(E)E−G(E)Q···(Z)...(O), where Z represents (2R,3R,4R)-*N*,2,3,4,5-pentakis(oxidanyl)pentanamide, which binds the Mg^{2+} through two oxygens in a bidentate fashion. ADP, ATP, and the nucleotide analogue phosphomethylphosphonic acid adenylate ester were also frequently found to ligate Mg^{2+} with a frequency of 49, 15, and 5, respectively (see Figure S13 in the Supporting Information). However, these ligands were too infrequent to appear in the consensus sequences.

Since catalytically active metal ions are typically not fully coordinated by proteinaceous ligands, we next sought to evaluate the ligand patterns of metalloenzymes to see if such a pattern emerged. When searching for [4Fe-4S] proteins, as described above, *Pyrococcus horikoshii* NadA was identified as having three cysteine ligands with the substrate of this enzyme, quinolinic acid, providing the final ligand to the [4Fe-4S] cluster. NadA (quinolinate synthase) catalyzes the formation of a precursor to NAD^+ (quinolinic acid) from aspartate-enamine and dihydroxyacetone phosphate[33] and is a member of the same hydro-lyase enzyme family as aconitase. Both NadA and aconitase bind carboxylate-containing substrates through the coordination of hydroxyl moieties to the [4Fe-4S] cluster of the

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

enzyme.[33] Therefore, we examined the sequence pattern associated with the binding of the [4Fe-4S] cluster of aconitase. Because WT sequences were not available for aconitase, the wild-type criterion was removed for this analysis. Only 10 structures were identified in the PDB, giving a CX_2C sequence pattern associated with the ligation of the [4Fe-4S] cluster of *Escherichia coli*, *Bos taurus*, *Sus scrofa*, and *Homo sapiens* aconitase as IGS(C)TNS−ANA(C)GP(C)IGQ−(Z)+(O) (see Figure S14 in the Supporting Information). As indicated above, O represents water, and, here, Z represents either hydrogen peroxide, acetate, citrate, isocitrate, or nitroisocitrate.

Next, Zn^{2+}-binding hydrolases, Zn^{2+}-binding RNA polymerases, and Mg^{2+}-binding kinases were evaluated with PdPDB. The root consensus sequence of the hydrolases showed predominantly carboxylate ligands (aspartate and glutamate), a histidine, and a water molecule (see Figure S15 in the Supporting Information). This is in contrast to the root, consensus sequence from a dataset of all Zn^{2+}-proteins, where complete cysteinyl ligation dominated (Figure 1). Furthermore, the sequence pattern of the hydrolases clearly showed the importance of the ligated water molecule. This is consistent with the catalytic mechanism of hydrolases, where the Zn^{2+}-bound water molecule acts as the nucleophile.[34] Conversely, the Zn^{2+}-binding site of RNA polymerases is not thought to play a catalytic role but rather is involved in assembling the complex of protomers.[35] The root, consensus sequence of the RNA polymerases was KRI(C)GT(C)GXM−FE(C)KG(C)S, which exploits a CX_2C motif (see Figure S16 in the Supporting Information) to fully coordinate the metal ion. The relationship between the Zn^{2+}-binding site of RNA polymerase and zinc finger proteins has previously been noted.[36] Attempts at defining a Mg^{2+} binding site for kinases revealed a high frequency of water molecule ligands (see Figure S17 in the Supporting Information) but no well-defined protein sequence motifs. The metalloproteins with catalytic metal centers showed the presence of a water ligand, whereas the enzyme that contained a noncatalytic metal center showed complete ligation by protein side-chains.

To gain a better sense of the significance of specific residues for the coordination of metal ions, an enrichment score (i.e., proportions from χ^2) was calculated for each analyzed position with PdPDB. This was done, in part, by comparing the frequency of each amino acid at each specified position against the average frequency of occurrence in ExPASy[9] of that amino acid for natural proteins. Enrichment scores were not calculated for nonproteinaceous ligands or nucleic acids. As would be expected, the residues with the highest enrichment score were the ligands to the metal centers. Consistent with previous analyses,[3] the enrichment score was the largest for histidine, aspartate, and cysteine for $Fe^{2+/3+}$, Mg^{2+}, and Zn^{2+}, respectively (see Figure 2, as well as Table S4 in the Supporting Information). The enrichment score for cysteine at the ligand position was greatest for [2Fe-2S] clusters, [4Fe-4S] clusters, and Zn^{2+} (see Figure S18 in the Supporting Information). For example, cysteine is found a total of 64 times across the entire pattern for the root of the [4Fe-4S]-cluster proteins, while the ExPASy reference has a frequency of 10 808 717. The Bonferroni corrected p-value for the χ^2 is <0.01, indicating that there is a statistically significant difference between the cysteine in [4Fe-4S] coordinating protein sequences and protein sequences that do not coordinate a [4Fe-4S] cluster. Noncysteine ligation was frequently observed as well (see Table S2). For example, in addition to cysteine, $Fe^{2+/3+}$ was also found to be coordinated by glutamate and aspartate. After cysteine ligation, the ligand with the largest enrichment score for Zn^{2+} was histidine, followed by glutamate/aspartate. Aspartate had the largest enrichment score for Mg^{2+}, followed by glutamate.

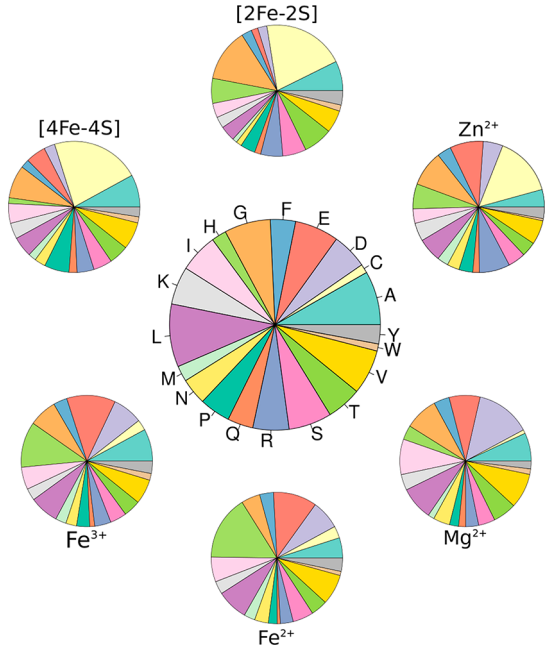


Figure 2. Amino acid enrichment for patterns that coordinate metallocofactors. Sequences include the ligand position and residues ±3 away from the ligand position with the trimming function set to 50%. The frequency of amino acids found in ExPASy was used as a reference. The center, larger wheel shows how the results would look if the distributions perfectly matched that found in ExPASy. The distributions of proportions were different between metalloproteins and the ExPASy reference, as determined by the Mann−Whitney test. Since most of the distributions were not normal, according to the Jarque-Bera test, a χ^2 test was used to compare the ligand containing patterns of different classes of metalloproteins.

The analysis of the enrichment scores helped to reveal the lesser-used residues for the ligation of metal ions, thus highlighting the possibilities exploited by biology. For example, while aspartate was not the most enriched residue occupying the ligand position for [2Fe-2S] clusters, aspartate was more enriched than most of the other amino acids, except for cysteine and histidine (see Figure S18). In fact, *E. coli* succinate dehydrogenase (complex II) (PDB ID: 1NEK) and *Thermotoga maritima* ferredoxin:NADP oxidoreductase (PDB ID: 4YLF) both have a [2Fe-2S] cluster with three cysteine ligands and one aspartate ligand. Interestingly, inspection of the crystal structures shows that *E. coli* succinate dehydrogenase uses a $CX_4CX_2DX_{11}C$ motif to coordinate the [2Fe-2S] cluster, whereas *T. maritima* ferredoxin:NADP oxidoreductase uses ligands with the same spacing but different order, i.e., $DX_4CX_2CX_{11}C$. *T. maritima* ferredoxin:NADP oxidoreductase also contains a [4Fe-4S] cluster with three cysteine ligands and one glutamate ligand, which is similar to a recently solved structure of *Thermosipho melanesiensis* HydF, which exploits three cysteine ligands and one exchangeable glutamate ligand.[37] Ligation of a [4Fe-4S] cluster with a histidine side-chain is quite rare (see Figure S19 in the Supporting Information). However, ethylbenzene dehydrogenase from *Aromatoleum aromaticum* (PDB ID: 2IVF), *Clostridium scatologenes* 4-hydroxyphenylace-

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

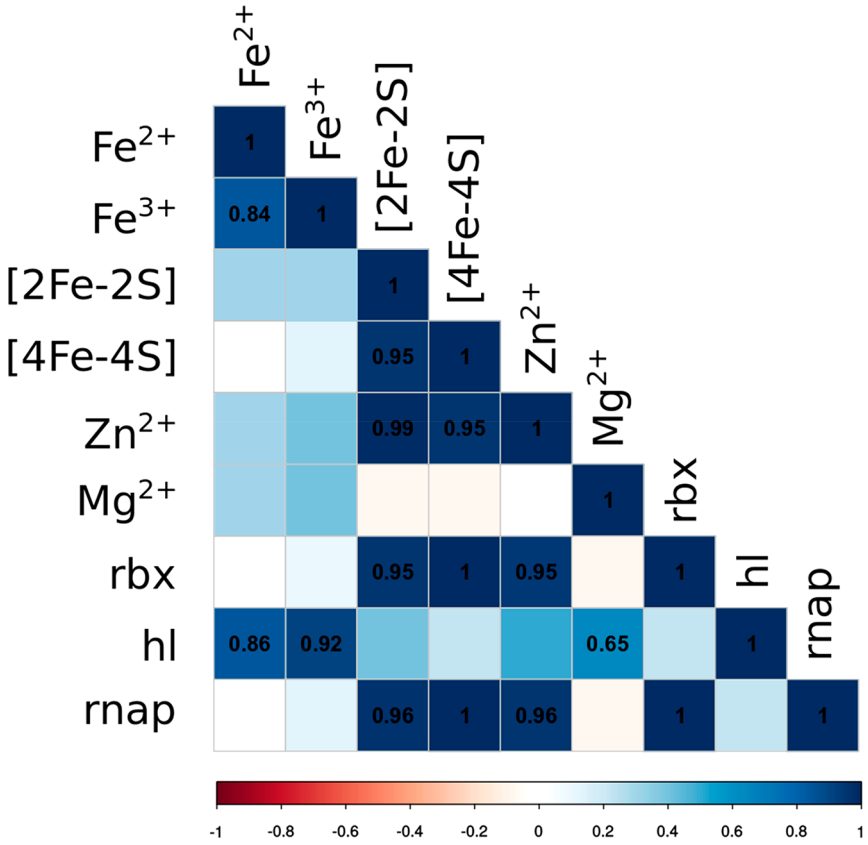


Figure 3. Correlations between the frequency of appearance of ligand patterns for each metal ion. The Pearson correlation coefficient was calculated from the enrichment score of the ligand distribution. For each column, the comparison of a given frequency distribution was compared with all of the other ligand frequency distributions. Larger values (blue) indicate a stronger correlation. Zero (white) indicates no correlation, and negative (red) values represent negative correlations. Rubredoxin (rbx) contains a mononuclear $Fe^{2+/3+}$ coordinated by four cysteines, hydrolase (hl) contains a catalytic Zn^{2+}, and RNA polymerase (rnap) contains a Zn^{2+} that plays a structural role. Data are based on the ligand position and residues ±3 away from the ligand position with the trimming function set to 50%. Only correlations with a Bonferroni corrected *p*-value of <0.01 are given.

tate decarboxylase (PDB ID: 2Y8N), and *Escherichia coli* nitrate reductase A (PDB ID: 3IR7) all coordinate a [4Fe-4S] cluster with three cysteine ligands and one histidine ligand. Surprisingly, PdPDB identified one example of a [4Fe-4S] cluster with one backbone carbonyl ligand provided by a proline residue; however, this metallocofactor was from a low-resolution structure of plant photosystem I from *Pisum sativum* (PDB ID: 3LW5) with poorly annotated metallocofactors. A more recent, higher-resolution structure (PDB ID: 4XK8) showed four cysteine ligands, with this specific proline residue clearly not coordinating. Although not detected by the analysis described herein with PdPDB, because the crystallized protein was not a wild-type sequence, *Aquifex aeolicus* IscU (PDB ID: 2Z7E) contains a [2Fe-2S] cluster ligated by three cysteines and one histidine.

The ligation patterns of Zn^{2+}, [2Fe-2S], and [4Fe-4S] proteins were more similar to each other than to mononuclear, nonheme iron or Mg^{2+} proteins. Pearson coefficients were used to identify correlations across different categories of metalloproteins[38,39] by comparing the Bonferroni corrected proportions χ^2. Such analyses revealed that the ligand preferences of mononuclear, nonheme iron proteins did not correlate well with [2Fe-2S] or [4Fe-4S] proteins (see Figure 3). However, if only cysteine ligated mononuclear iron proteins (i.e., rubredoxins) were analyzed, then a correlation with polynuclear iron−sulfur clusters was apparent (Figure 3). Zn^{2+}-binding proteins correlated well with [2Fe-2S], [4Fe-4S], and rubredoxins, but not mononuclear, nonheme iron proteins as a whole. However, the ligand pattern associated with the enzyme hydrolase, which contained a catalytic Zn^{2+} ion, differed substantially from the average computed from the entire Zn^{2+} dataset. For example, hydrolase correlated best with $Fe^{2+/3+}$, and not with [2Fe-2S], [4Fe-4S], nor rubredoxin proteins. Conversely, the structural Zn^{2+} site of RNA polymerase showed similar correlation trends to the entire Zn^{2+} dataset. Mg^{2+} did not correlate well with any of the other analyzed metal-binding datasets (Figure 3). If the analysis was expanded to include residues ±3 positions away from the ligand, the observed correlations were largely the same, because the enrichment at the ligand positions was much greater than for the surrounding positions (see Figures S20−S26 in the Supporting Information). If the ligand position was removed from the analysis of the ±3 window, then some degree of similarity began to emerge that previously was not apparent. For example, the composition of residues for Mg^{2+} and Fe^{3+} showed a correlation of 0.82 (see Figure S27 in the Supporting Information).

To better understand if there are sequence preferences for residues surrounding the ligand positions, the enrichment scores for each amino acid were evaluated for positions ±3 away from, but not including, the ligand. From this analysis, it was immediately apparent that glutamines were rarely present within the ±3 region for any metal binding site analyzed (Figure 2). Methionine was similarly not enriched. Tryptophan was also not frequently present near the ligand position;

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

however, tryptophan was enriched within the $Fe^{2+/3+}$ ±3 region of rubredoxin. Glutamates and aspartates were frequently found near the ligand position of [2Fe-2S] binding proteins (see Figures S23 and S24), whereas histidine was not commonly present within ±3 window surrounding the ligands of a [4Fe-4S] cluster. Conversely, glycine was commonly found near metal ligands (Figure S24). Proline was enriched for [4Fe-4S] proteins (Figure S24), and Ile was enriched near the ligand positions of Mg^{2+}-binding proteins (see Figure S23 and the Supporting Information). It is unclear if the observed enrichments reflect criteria imposed by the metal ion or result from the structural scaffold provided by the overall tertiary and quaternary fold of the protein. Such ambiguity is particularly pronounced for datasets with low structural diversity. For the datasets used herein, the [2Fe-2S] dataset was the most structurally diverse, whereas the rubredoxin dataset had the lowest Shannon and Simpson diversities[40−42] (Table S5 in the Supporting Information). For the case of [2Fe-2S] proteins, the large degree of structural diversity coupled with the relative lack of diversity of ligand motifs (Table S5) suggests the presence of significant constraints on the coordination of the Fe ions of polynuclear iron−sulfur clusters that require multiple proteinaceous ligands for each metal center.

CONCLUSIONS

PdPDB is an easy-to-use tool that can reliably identify metallocofactor coordinating patterns. PdPDB can also identify patterns pertaining to specific classes of enzymes by aligning the ligand positions, followed by dropping the information pertaining to nonconserved residues. Therefore, PdPDB can be used in a variety of ways to interrogate metal-binding sequence patterns. The patterns associated with the binding of iron−sulfur clusters, Mg^{2+}, and Zn^{2+} are consistent with previously identified motifs. Importantly, PdPDB also provides additional information pertaining to nonproteinaceous ligands and sequences. Such information is highly valuable for the design of metal coordinating peptides, proteins, and nucleic acids, potentially with catalytic activity. In fact, for the case of Zn^{2+}-binding hydrolases, catalytically active metal centers exploit different ligand patterns than for structural Zn^{2+}-binding sites. This is not surprising, because metal ions that play a structural role would be expected to be more tightly bound and thus fully coordinated by the protein scaffold. Conversely, catalytically active metal ions often need to coordinate to both the protein scaffold and the substrate. Therefore, catalytic metal ions would be expected to exploit, in part, exchangeable ligands, such as water molecules. However, it is important to note that the tool described herein (PdPDB) only analyzes sequences of biological molecules with known three-dimensional structures. The advantage of this is that a more complete and accurate picture of metal-binding coordinating patterns can be extracted. The limitation is that the protein and nucleic acid structures deposited in the Protein Data Bank are not necessarily representative of the breadth or frequency of the sequences used by biology.

It is striking how frequently the CX_2C sequence emerged when analyzing the ligand binding patterns of different metal ions. The only metal ion for which a CX_2C motif was not detected was Mg^{2+}, which is consistent with the incompatibility of a hard metal ion with softer ligands. Although no attempt was made to infer evolutionarily ancient sequences with PdPDB, CX_2C has been proposed to represent an ancient metal-binding motif for both Zn^{2+} and iron−sulfur proteins[23] and is even exploited by c-type cytochromes to covalently bind heme.[48] Computational and laboratory measurements of thiolate ligands and cysteine containing peptides suggest that such CX_2C sequences would bind Zn^{2+} more strongly than Fe^{2+}.[3,22] However, most of the sequences associated with the binding of mononuclear, nonheme iron ions do not exploit thiolate ligands. Instead, both Mg^{2+} and $Fe^{2+/3+}$ share a preference for oxygen ligands. In fact, it has been proposed that early life may have used Fe^{2+} instead of Mg^{2+}, because of the increased availability of Fe^{2+} on the anoxygenic, prebiotic Earth.[43] However, analyses with PdPDB show that mononuclear Fe^{2+}-binding sites also make extensive use of nitrogen ligands, as provided by histidine, which leads to a strong selection for the binding of Fe ions over Mg ions.[22]

CX_2C is also a common motif that is used to make, break, and isomerize disulfide bonds.[22] Such enzymes with CX_2C containing active sites (e.g., thioredoxins and glutaredoxins) do not exploit the CX_2C sequence to bind metal ions. That is, the same short sequence motif can play two fundamentally different and important roles in biology. Furthermore, CX_2C containing hexapeptides have been shown to coordinate redox active [2Fe-2S] clusters,[44] and CX_2C octapeptides are capable of reducing disulfide bonds in the absence of metal ions.[45,46] Since the residues immediately adjacent to the cysteine positions influence the metal-binding properties of the peptides[47] and the redox potential of the cysteines,[45,46] ancient CX_2C peptide sequences may have been predisposed to different, perhaps complementary, activities.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00468.

List of PDB IDs (Table S1); ligand patterns (Table S2); frequency of small molecules ligated to $Fe^{2+/3+}$ (Table S3); amino acid enrichments (Table S4); secondary structures and diversity indices (Table S6); PdPDB workflow (Figure S1); PdPDB and Clustal X alignments (Figure S2); dendrograms (Figure S3);[4Fe-4S] protein families (Figure S4); [2Fe-2S] protein families (Figure S5); Fe^{2+} protein families (Figure S6); Fe^{3+} protein families (Figure S7); Fe^{2+} protein root logo with 5% threshold (Figure S8); Fe^{3+} protein root logo with 5% threshold (Figure S9); Rubredoxin logo (Figure S10); Zn^{2+} protein families (Figure S11); Mg^{2+} protein families (Figure S12); Mg^{2+} protein families with n = 1 (Figure S13); aconitase logo (Figure S14); hydrolase logo (Figure S15); RNA-polymerase logo (Figure S16); kinase logo (Figure S17); *A. aromaticum* ethylbenzene dehydrogenase cluster (Figure S19); enrichment scores (Figures S18, S20−S26); correlations of nonligated positions (Figure S27) (PDF)

Enrichment scores, frequencies, proportions, p-values and statistical tests (XLS)

Complete list of patterns extracted with PdPDB (XLS)

AUTHOR INFORMATION

Corresponding Author

*E-mail: mansy@science.unitn.it.

ORCID

Luca Belmonte: 0000-0002-7977-9170

Sheref S. Mansy: 0000-0003-2382-198X

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Present Address

[†]Dipartimento di Scienze della Vita, Università di Modena e Reggio Emilia, Via Campi 287, 41125, Modena, MO, Italy.

Funding

This work was supported by the Simons Foundation (No. 290358), Armenise-Harvard Foundation, and COST action CM1304.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank M. Benelli, C. Bonfio, M. Forlin, F. Gallo, D. Prandi, and S. Scintilla for helpful discussions on this work.

■ REFERENCES

(1) Berman, H.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.

(2) Shuber, A. P.; Orr, E. C.; Recny, M. A.; Schendel, P. F.; May, H. D.; Schauer, N. L.; Ferry, J. G. Cloning, Expression, and Nucleotide Sequence of the Formate Dehydrogenase Genes from *Methanobacterium Formicicum*. *J. Biol. Chem.* **1986**, *261*, 12942−12947.

(3) Belmonte, L.; Rossetto, D.; Forlin, M.; Scintilla, S.; Bonfio, C.; Mansy, S. S. Cysteine Containing Dipeptides Show a Metal Specificity That Matches the Composition of Seawater. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20104−20108.

(4) Castagnetto, J. M.; Hennessy, S. W.; Roberts, V. A.; Getzoff, E. D.; Tainer, J. A.; Pique, M. E. MDB: The Metalloprotein Database and Browser at The Scripps Research Institute. *Nucleic Acids Res.* **2002**, *30*, 379−382.

(5) Jonassen, I.; Eidhammer, I.; Conklin, D.; Taylor, W. R. Structure Motif Discovery and Mining the PDB. *Bioinformatics* **2002**, *18*, 362−367.

(6) Nadzirin, N.; Gardiner, E. J.; Willett, P.; Artymiuk, P. J.; Firdaus-Raih, M. SPRITE and ASSAM: Web Servers for Side Chain 3D-Motif Searching in Protein Structures. *Nucleic Acids Res.* **2012**, *40*, W380−W386.

(7) Golovin, A.; Henrick, K. MSDmotif: Exploring Protein Sites and Motifs. *BMC Bioinf.* **2008**, *9*, 312.

(8) Brylinski, M.; Skolnick, J. FINDSITE-Metal: Integrating Evolutionary Information and Machine Learning for Structure-Based Metal-Binding Site Prediction at the Proteome Level. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 735−751.

(9) Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R. D.; Bairoch, A. ExPASy: The Proteomics Server for In-Depth Protein Knowledge and Analysis. *Nucleic Acids Res.* **2003**, *31*, 3784−3788.

(10) Newcombe, R. G. Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Stat. Med.* **1998**, *17*, 857−872.10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

(11) Wright, S. P. Adjusted P-Values for Simultaneous Inference. *Biometrics* **1992**, *48*, 1005−1013.

(12) Bauer, D. F. Constructing Confidence Sets Using Rank Statistics. *J. Am. Stat. Assoc.* **1972**, *67*, 687−690.

(13) Hollander, M.; Wolfe, D. A.; Chicken, E. *Nonparametric Statistical Methods*, 3rd Edition; John Wiley & Sons: New York, 2013.

(14) Cromwell, J. B.; Labys, W. C.; Terraza, M. *Univariate Tests for Time Series Models*; Quantitative Applications in the Social Sciences, No. *99*; Sage Publications: Thousand Oaks, CA, 1994.

(15) Murtagh, F.; Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **2014**, *31*, 274−295.10.1007/s00357-014-9161-z

(16) Becker, R. A.; Chambers, J. M.; Wilks, A. R. *The New S Language*; Wadsworth & Brooks: Pacific Grove, CA, 1988.

(17) Galili, T. Dendextend: An R Package for Visualizing, Adjusting and Comparing Trees of Hierarchical Clustering. *Bioinformatics* **2015**, *31*, 3718−3720.

(18) Kim, J. D.; Rodriguez-Granillo, A.; Case, D. a; Nanda, V.; Falkowski, P. G. Energetic Selection of Topology in Ferredoxins. *PLoS Comput. Biol.* **2012**, *8*, e1002463.

(19) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(20) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(21) Schmidt, C. L.; Shaw, L. A Comprehensive Phylogenetic Analysis of Rieske and Rieske-Type Iron-Sulfur Proteins. *J. Bioenerg. Biomembr.* **2001**, *33*, 9−26.

(22) Dudev, T.; Nikolova, V. Determinants of Fe^{2+} over M^{2+} (M = Mg, Mn, Zn) Selectivity in Non-Heme Iron Proteins. *Inorg. Chem.* **2016**, *55*, 12644−12650.

(23) Lupas, A. N.; Ponting, C. P.; Russell, R. B. On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *J. Struct. Biol.* **2001**, *134*, 191−203.

(24) Pabo, C. O.; Peisach, E.; Grant, R. A. Design and Selection of Novel Cys_2His_2 Zinc Finger Proteins. *Annu. Rev. Biochem.* **2001**, *70*, 313−340.

(25) Chan, K. L.; Bakman, I.; Marts, A. R.; Batir, Y.; Dowd, T. L.; Tierney, D. L.; Gibney, B. R. Characterization of the Zn(II) Binding Properties of the Human Wilms' Tumor Suppressor Protein C-terminal Zinc Finger Peptide. *Inorg. Chem.* **2014**, *53*, 6309−6320.

(26) Magyar, J. S.; Godwin, H. A. Spectropotentiometric Analysis of Metal Binding to Structural Zinc-Binding Sites: Accounting Quantitatively for pH and Metal Ion Buffering Effects. *Anal. Biochem.* **2003**, *320*, 39−54.

(27) Berg, J. M. Zinc-Finger Proteins. *Curr. Opin. Struct. Biol.* **1993**, *3*, 11−16.

(28) Lovering, R.; Hanson, I. M.; Borden, K. L. B.; Martin, S.; O'Reilly, N. J.; Evan, G. I.; Rahman, D.; Pappin, D. J. C.; Trowsdale, J.; Freemont, P. S. Identification and Preliminary Characterization of a Protein Motif Related to the Zinc Finger. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 2112−2116.

(29) Patient, R. K.; McGhee, J. D. The GATA Family (Vertebrates and Invertebrates). *Curr. Opin. Genet. Dev.* **2002**, *12*, 416−422.

(30) van der Gulik, P.; Massar, S.; Gilis, D.; Buhrman, H.; Rooman, M. The First Peptides: The Evolutionary Transition between Prebiotic Amino Acids and Early Proteins. *J. Theor. Biol.* **2009**, *261*, 531−539.

(31) Zheng, H.; Chordia, M. D.; Cooper, D. R.; Chruszcz, M.; Müller, P.; Sheldrick, G. M.; Minor, W. Validating Metal Binding Sites in Macromolecule Structures Using the CheckMyMetal Web Server. *Nat. Protoc.* **2013**, *9*, 156−170.

(32) Black, C. B.; Huang, H.-W.; Cowan, J. A. Biological Coordination Chemistry of Magnesium, Sodium, and Potassium Ions. Protein and Nucleotide Binding Sites. *Coord. Chem. Rev.* **1994**, *135-136*, 165−202.

(33) Esakova, O. A.; Silakov, A.; Grove, T. L.; Saunders, A. H.; McLaughlin, M. I.; Yennawar, N. H.; Booker, S. J. Structure of Quinolinate Synthase from Pyrococcus Horikoshii in the Presence of Its Product, Quinolinic Acid. *J. Am. Chem. Soc.* **2016**, *138*, 7224−7227.

(34) Hernick, M.; Fierke, C. A. Zinc Hydrolases: The Mechanisms of Zinc-Dependent Deacetylases. *Arch. Biochem. Biophys.* **2005**, *433*, 71−84.

(35) Markov, D.; Naryshkina, T.; Mustaev, A.; Severinov, K. A Zinc-Binding Site in the Largest Subunit of DNA-Dependent RNA Polymerase Involved in Enzyme Assembly. *Genes Dev.* **1999**, *13*, 2439−2448.

(36) Klug, A.; Rhodes, D. Zinc Fingers: A Novel Protein Fold for Nucleic Acid Recognition. *Cold Spring Harbor Symp. Quant. Biol.* **1987**, *52*, 473−482.

(37) Caserta, G.; Pecqueur, L.; Adamska-Venkatesh, A.; Papini, C.; Roy, S.; Artero, V.; Atta, M.; Reijerse, E.; Lubitz, W.; Fontecave, M. Structural and Functional Characterization of the Hydrogenase-Maturation HydF Protein. *Nat. Chem. Biol.* **2017**, *13*, 779−784.

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(38) Murdoch, D. J.; Chow, E. D. A Graphical Display of Large Correlation Matrices. *Am. Stat.* **1996**, *50*, 178−180.
(39) Friendly, M. Corrgrams: Exploratory Displays for Correlation Matrices. *Am. Stat.* **2002**, *34*, 1447−1449.
(40) Hurlbert, S. H. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology* **1971**, *52*, 577−586.
(41) Fisher, R. A.; Corbet, A. S.; Williams, C. B. The Relation between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *J. Anim. Ecol.* **1943**, *12*, 42−58.
(42) Heck, K. L.; van Belle, G.; Simberloff, D. Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size. *Ecology* **1975**, *56*, 1459−1461.
(43) Athavale, S. S.; Petrov, A. S.; Hsiao, C.; Watkins, D.; Prickett, C. D.; Gossett, J. J.; Lie, L.; Bowman, J. C.; O'Neill, E.; Bernier, C. R.; Hud, N. V.; Wartell, R. M.; Harvey, S. C.; Williams, L. D. RNA Folding and Catalysis Mediated by Iron(II). *PLoS One* **2012**, *7*, e38024.
(44) Scintilla, S.; Bonfio, C.; Belmonte, L.; Forlin, M.; Rossetto, D.; Li, J.; Cowan, J. A.; Galliani, A.; Arnesano, F.; Assfalg, M.; Mansy, S. S. Duplications of an Iron−sulphur Tripeptide Leads to the Formation of a Protoferredoxin. *Chem. Commun.* **2016**, *52*, 13456−13459.
(45) Siedler, F.; Rudolph-Boehner, S.; Doi, M.; Musiol, H. J.; Moroder, L. Redox Potentials of Active-Site Bis (Cysteinyl) Fragments of Thiol-Protein Oxidoreductases. *Biochemistry* **1993**, *32*, 7488−7495.
(46) Ookura, T.; Kainuma, K.; Kim, H.-J.; Otaka, A.; Fujii, N.; Kawamura, Y. Active Site Peptides with CXXC Motif on MAP-Resin Can Mimic Protein Disulfide Isomerase Activity. *Biochem. Biophys. Res. Commun.* **1995**, *213*, 746−751.
(47) Bonfio, C.; Valer, L.; Scintilla, S.; Shah, S.; Evans, D. J.; Jin, L.; Szostak, J. W.; Sasselov, D. D.; Sutherland, J. D.; Mansy, S. S. UV-Light-Driven Prebiotic Synthesis of Iron−Sulfur Clusters. *Nat. Chem.* **2017**, DOI: 10.1038/nchem.2817.
(48) Sanders, C.; Turkarslan, S.; Lee, D.-W.; Daldal, F. Cytochrome c Biogenesis: The Ccm System. *Trends Microbiol.* **2010**, *18*, 266−274.

DOI: 10.1021/acs.jcim.7b00468
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX