PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Function-based assessment of structural similarity measurements using metal co-factor orientation

Stefan Senn,[1]* Vikas Nanda,[2] Paul Falkowski,[1,3,4,5] and Yana Bromberg[6]*

[1] Environmental Biophysics and Molecular Ecology Program, Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, New Jersey 08901

[2] Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, New Jersey 08854

[3] Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854

[4] Department Environmental Sciences, Rutgers University, New Brunswick, New Jersey 08901

[5] Department of Earth and Planetary Sciences, Rutgers University, Piscataway, New Jersey 08854

[6] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey 08901

ABSTRACT

Structure comparison is widely used to quantify protein relationships. Although there are several approaches to calculate structural similarity, specifying significance thresholds for similarity metrics is difficult due to the inherent likeness of common secondary structure elements. In this study, metal co-factor location is used to assess the biological relevance of structural alignments. The distance between the centroids of bound co-factors adds a chemical and function-relevant constraint to the structural superimposition of two proteins. This additional dimension can be used to define cut-off values for discriminating valid and spurious alignments in large alignment sets. The hypothesis underlying our approach is that metal coordination sites constrain structural evolution, thus revealing functional relationships between distantly related proteins. A comparison of three related nitrogenases shows the sequence and fold constraints imposed on the protein structures up to 18 Å away from the centers of their bound metal clusters.

Proteins 2013; 00:000–000.

Key words: structure comparison; structural bioinformatics; metalloproteins; computational biology; structural evolution.

INTRODUCTION

Proteins that perform similar functions often share certain sequence and structural features. Although sequence comparison is a well-established approach for inferring functional similarity of proteins, structural comparison can identify relationships between proteins that have far diverged in sequence. The limitation of structural alignments is their reliance on subjective visual inspection for analysis and verification.[1,2] Several studies have tried to quantify the significance of structural alignments,[3–11] but there is still no consensus as to what is considered relevant. Note, however, that it is possible, as is sometimes performed for sequence alignments, to exploit the particular protein characteristics, for example, the knowledge of specific amino acid residues necessary for co-factor binding or catalytic activity,[12,13] in approximating alignment relevance.

Finding the correct structural alignment of metalloprotein binding sites can potentially provide insight into early protein evolution. Metals are often important for protein structural integrity[14] and catalysis of crucial metabolic processes.[15] Redox-active metalloproteins have been postulated to serve as key electronic components in the circuitry of life.[16] The evolutionary history of metals as protein co-factors coincides with their biogeochemical

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Gordon and Betty Moore Foundation (GBMF); Grant number: GBMF2807. Grant sponsor: USDA-NIFA; Grant number: 1015:0228906 (to Y.B.).

*Correspondence to: Yana Bromberg; Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, 76 Lipman Drive, NJ 08901.
E-mail: yana@bromberglab.org or Stefan Senn. E-mail: senn@marine.rutgers.edu

Received 22 May 2013; Revised 17 September 2013; Accepted 26 September 2013
Published online 11 October 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24442

availability,[17] implying that proteins incorporated metals early in the origin of life. Proper metal binding is essential to protein molecular function and restricts possible structural configurations, especially in the ligand vicinity.[18,19] Identifying evolutionary relationships between metalloproteins requires a structural alignment method that is 'extra' sensitive to these functional relationships but avoids spurious alignments.

In this study, we examine structural alignments of the protein backbones in metal co-factor binding regions, or microenvironments, specifically around iron–sulfur clusters and heme groups. Our metallo-centric approach allows for the assessment of alignment quality by considering geometric features of the corresponding protein ligands. The Euclidian distance between the metal centroids is a simple, yet surprisingly reliable similarity measure of two super-imposed protein environments around their co-factor binding sites. In this study, we use this ligand–ligand distance to establish relationships between the binding environments of three classes of ligands in structures from the protein data bank (PDB).[20] The use of co-factor distances as an indicator of alignment quality is informed by the biology of the compared proteins. Most structure alignment methods try to optimize a combination of root mean square deviation (RMSD) and alignment length.[1,2] We argue that a measure independent of these two properties is valuable in assessing a superimposition's biological significance.

Ligands, specifically metal-containing co-factors, constrain the space of possible mutations in their close vicinity. These constraints have implications for the evolution of metal-binding proteins. The ligand-enforced constraints in this study are described by the progressive loss of sequence identity and increased mutability of residues as a function of the distance from the ligand.

MATERIALS AND METHODS

Ligand microenvironments

Three classes of ligands were examined: Fe_4-S_4 (SF4 in PDB notation), Fe_2-S_2 (FES), and heme (HEM). For each co-factor we extracted all PDB files, non-redundant at 95% sequence identity (148 entries for SF4, 121 entries for FES, and 456 entries for HEM). This sequence filter limits the search space while retaining a high number of closely related structures and a wide spectrum of possible binding sites. Note that more stringent filters remove the "not identical" but highly similar structures that are important for the calibration of co-factor distance-based signal. For each structure in our set we computed microenvironment spheres of radii 10–20 Å. A microenvironment includes all Cα atoms within a given radius from the average coordinates of the iron atoms (centroid) of the examined co-factor. In heme-containing structures, only the central iron atom of the porphyrin is used as the centroid. Fragments shorter than three continuous residues were omitted. Note that the two microenvironments of neighboring ligands in the same structure have overlapping backbone regions if the microenvironment radius is larger than half the distance between two ligand centroids. From the entire set of microenvironments, we removed identical microenvironments coming from the same PDB file to yield three sets with 231 (Fe_4-S_4), 137 (Fe_2-S_2), and 690 (heme) members, respectively (Supporting Information Table SI).

Structure alignments

We used the precise setting of TopMatch[21] to calculate structure alignments between protein backbones of all extracted ligand microenvironments. Microenvironments can be described in fragment notation used as TopMatch input (Supporting Information Table SI). The TopMatch algorithm combines two crucial features of structural alignments, spatial deviation and number of aligned residues, into one similarity score, *S*. Structural distance is calculated [Eq. (1)] as described in,[22] where Q_L and T_L denote the number of residues in the query and target structures, respectively.

$$D = Q_L + T_L - 2S \quad (1)$$

Structural distance, *D*, then corresponds to the total number of residues in the target and query that are not aligned and can, for the ease of use, be considered as having the unit '*residues*'.

Defining clustering thresholds

We established structural and co-factor centroid distance thresholds to create a set of high confidence structural relationship clusters from the total set of alignments (structures as nodes and alignment structural distances as edges). The thresholds for co-factor distances were defined from manual inspection of Figure 3(A–C) (2.8 for Fe_4-S_4, 2.1 for Fe_2-S_2, and 3.0 for heme). These cut-off values were selected to cover the highly similar alignments 'tail' of the plots in Figure 3(A–C), but not the alignments of unknown quality [cloud of alignments in Fig. 3(A–C)].

To find cut-off values for *D* usable for clustering, we performed precision/recall analysis. We considered the empirically defined co-factor distance cut-off our 'gold-standard' that is, all alignments with a co-factor distance lower than the cut-off is considered "true". We then compiled the confusion matrix for each *D* cut-off value in steps of five residues. Precision and recall are defined as

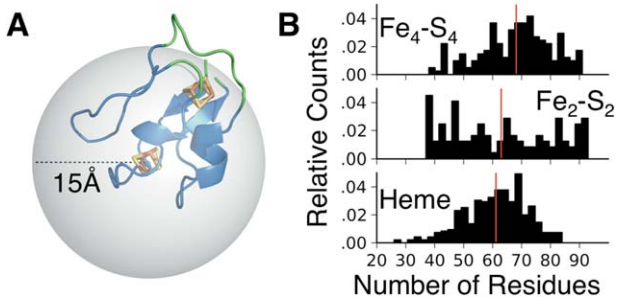

Figure 1

Microenvironments and population sizes. Microenvironments are defined as all Cα atoms of the protein backbone that are closer to the center of the examined ligand than a defined cut-off (radius). In (**A**) the ferredoxin (2fdn) microenvironment is shown. The sphere marks 15 Å from the Fe_4-S_4 center. Blue colored residues in the structure are part of the microenvironment. Green colored residues are outside the sphere and are thus not part of the microenvironment. (**B**) shows the distributions of the number of residues in microenvironments of 15 Å radius around Fe_4-S_4 (top), Fe_2-S_2 (middle), and heme (bottom) ligands. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$Precision = TP/(FP+TP); Recall = TP/(TP+FN) \quad (2)$$

The threshold for D was selected as the value where 90% precision in identifying relevant alignments was achieved (for Fe_4-S_4 $D = 65$, for Fe_2-S_2 and heme $D = 50$). For each cluster of structures, we removed the alignments scoring below these thresholds. The structure with the most alignments in each cluster was selected as its representative (Supporting Information Table SII).

Mutation analysis

To establish the potential functional disruptiveness of single amino acid substitutions we used the program SNAP.[23] SNAP uses solely sequence-based information to produce a score, ranged from −100 to 100, where positive scores indicate mutations disruptive of protein function. Higher scores, indicate more severe effects. The SubMat score is the average SNAP score over all evolutionarily 'allowed' substitutions[24] (i.e., amino acid exchanges scoring >0 in the BLOSUM62[25] matrix).

RESULTS AND DISCUSSION

Influence of the co-factor microenvironment size

The number of residues in the query and the target proteins affects the alignment structural distance. Aligning identical structures should result in a distance of $D = 0$ [Eq. (1)], which is only possible if the microenvironments around metal clusters are of equal size. A microenvironment's size is defined by a cut-off distance from the co-factor centroid [Fig. 1(A)] – the radius of the microenvironment sphere. The larger the radius, the more structural information is involved in assessing D. However, smaller metalloproteins in the dataset limit the maximal value of this radius. For the three classes of ligands in our study (Fe_4-S_4, Fe_2-S_2, and heme), the population (number of residues assigned to the microenvironment) distributions for 15 Å microenvironments are shown in Figure 1(B). The distributions show average populations of 68 residues per Fe_4-S_4 microenvironment, 63 for Fe_2-S_2, and 61 for hemes. The shapes of Fe_4-S_4 and heme microenvironment population size distributions are similar, with the Fe_4-S_4 environments, on average, more populated [distribution shifted to the right in Fig. 1(B)]. This is expected since heme is significantly larger than Fe_4-S_4; that is, in the presence of a larger ligand fewer residues fit into the same radius sphere. The distribution of population sizes around Fe_2-S_2 is different from both the heme and Fe_4-S_4 and is maximal at 40 residues per microenvironment. We speculate that this might be due to the fact that Fe_2-S_2 is often observed at protein structural "edges", for example, in Rieske-like domains. Note again that our microenvironments are spheres and may not be optimal for all ligand shapes, for example, more oblong hemes. The microenvironment radius was selected to include most of the amino acids relevant to co-factor binding and function. Hence we believe that the shape of the ligand on this scale of comparison does not influence the structure alignment results dramatically.

With an increasing radius, all proteins reach their limit for the microenvironment population size, that is, the complete protein is fully contained in the microenvironment (Supporting Information Table SII). For the analysis reported here, solely the 15 Å radius microenvironments were examined as they provided the most consistent residue count between all proteins of different co-factor types.

Limits of the distance *D* in evaluating microenvironment alignments

Absent a well-defined functional constraint it is difficult to infer structural relatedness of dissimilar folds. For example, aligning the 15 Å microenvironment from the *Desulfovibrio vulgaris* NiFe-selenium-hydrogenase[26] (PDB: 2wpn) with that from *Pseudomonas aeruginosa* ferredoxin[27] (2fgo) results in 13 aligned residues and a structural distance of 97 residues at an RMSD of 1.91 Å [alignment A, Fig. 2(A)]. The alignment of the *Geobacillus stearothermophilus* MutY adenine glycosylase[28] (1rrq) microenvironment with one from the *Clostridium difficile* (R)-2-Hydroxyisocarpoyl-CoA dehydratase[29] (3o3m) results in a similar structural distance of 91 residues, 13

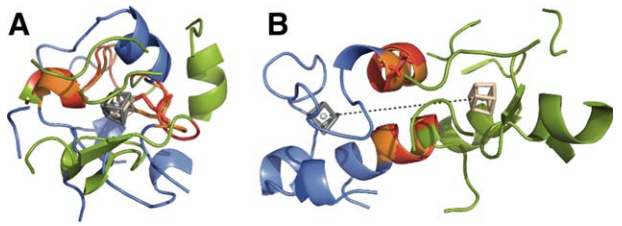

Figure 2

Similarly scored alignments have different co-factor distances. Cartoon representation of alignments between the 15 Å microenvironments of (**A**) an iron–sulfur cluster in the *Desulfovibrio vulgaris* NiFeSe-hydrogenase (residue A1284 in 2wpn; blue if unaligned, orange if aligned) and an iron–sulfur cluster in the *Pseudomonas aeruginosa* ferredoxin (residue A 202 in 2fgo; green if unaligned, red if aligned), and (**B**) an iron–sulfur in the *Geobacillus stearothermophilus* MutY adenine glycosylase (residue A 400 in 1rrq; blue if unaligned, orange if aligned), and the *Clostridium difficile* (R)−2-hydroxyisocarpoyl-CoA dehydratase (residue C 409 in 3o3m; green if unaligned, red if aligned). Both comparisons result in 13 aligned residues (one identical, stick representation) and comparable structural distances (97 and 91 residues, respectively), and RMSD values (1.91 and 1.04 Å, respectively). The co-factors in (A), however, are closer than in (B) with the Euclidian distances between the co-factor centroids of 0.99 and 19.18 Å (black dashed line), respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

residues aligned at an RMSD of 1.04 Å [alignment B, Fig. 2(B)]. Both alignments exhibit 8% sequence identity, which corresponds to a single amino acid. The comparisons have similar results and alignment B can even be considered slightly better due to a lower RMSD (1.04 vs. 1.91 Å) and a lower structural distance (91 vs. 97 residues). However, the functional relevance of these alignments cannot be verified without subjective, visual inspection (Fig. 2).

In alignment A, the microenvironments come from functionally related proteins; hydrogenases and ferredoxins are both are involved in electron transport. Moreover, the single residue identical between the structures in alignment A is an iron-binding cysteine. In contrast, alignment B is a comparison of a hydrolase with a lyase—functionally distinct enzymes. It is not surprising, therefore, that the corresponding residue in this alignment is an alanine, irrelevant in the context of iron–sulfur binding. Considering the ligand–ligand distance in the alignment also gives insight into alignment quality. Alignment B exhibits a high (19.2 Å) distance between the centroids of the two iron–sulfur co-factors, whereas that distance in alignment A is much lower (1.0 Å).

The findings that, unlike in alignment B, in alignment A (1) the aligned residues of are close to the bound co-factors and (2) the structural superimposition places the iron–sulfur cluster, and the one iron-binding cysteine, at equivalent positions, lead us to conclude that the latter has more functional relevance than the former.

Note that a global alignment of structures 2wpn (NiFe-selenium-hydrogenase chain A, 277 residues) and 2fgo (ferredoxin, chain A, 81 residues) has a maximum length 33 aligned residues and an RMSD of 3.23 Å. Of the 33 residues, three are identical (9%). The closest iron–sulfur cluster pair from the different structures in the superimposition is ~10 Å apart. A BLAST[30] of the corresponding protein sequences is also inconclusive (*e*-value 0.067). Thus, the local relationship described in Figure 2(A) cannot be detected by either a global structure alignment or sequence alignment.

Ligand–ligand distance as measure of alignment quality

To automate the discrimination between the 'good' and 'bad' alignments (e.g., A *vs.* B in Fig. 2), we used the distance between the centroids of the ligands in the microenvironment superimposition as a metric of alignment quality. In meaningful alignments, the co-factors are expected be located at equivalent structural positions. We calculated all possible alignments between the 15 Å microenvironments of each co-factor. The distribution of alignment ligand–ligand distances versus structural distances, *D*, exhibits a distinct shape with a clearly isolated 'tail' of highly similar structures; that is structural alignments with a low co-factor distance across a range of *D* [Fig. 3(A–C)]. The non-linear relationship between *D* and the co-factor distance suggests a threshold for discriminating alignment quality. The ligand–ligand distance remains low up to $D = 60$ residues for Fe_4-S_4 and $D = 40$ residues for heme and Fe_2-S_2 environments (Fig. 3). Beyond these cut-offs, random (non-meaningful) alignments cloud the structural alignment landscape; that is, ligand–ligand distances exhibit significantly higher averages [structural distance bin stepsize = 1; black line in Fig. 3(A–C)] and increased fluctuations, indicating little functional relevance and likely poor quality of the alignments.

We examined the distribution of co-factor distances across all alignments [Fig. 3(D)] and across only the alignments of very high structural distance ($D > 120$), which we assume to be random [Fig. 3(E)]. The redundancy of our dataset guarantees a number of very good alignments and provides a guideline for acceptable co-factor distances. The former can be observed as a high fraction of alignments with short ligand–ligand distances, while the latter is represented by a dip between that maximum and the remaining random-like fraction of alignments [Fig. 3(D)].

The expected random distribution of aligned co-factor distances is approximated by sampling centers of arbitrarily superimposed 15 Å spheres and plotting their pairwise Euclidian distances [dashed lines in Fig. 3(D,E)]. This random distribution differs from both the distribution of all alignments and the distribution of alignments at high structural distances; that is, the baseline distribution is normal and right-shifted as opposed to the heavy left skew of both real distributions. [Fig. 3(D,E)]. This

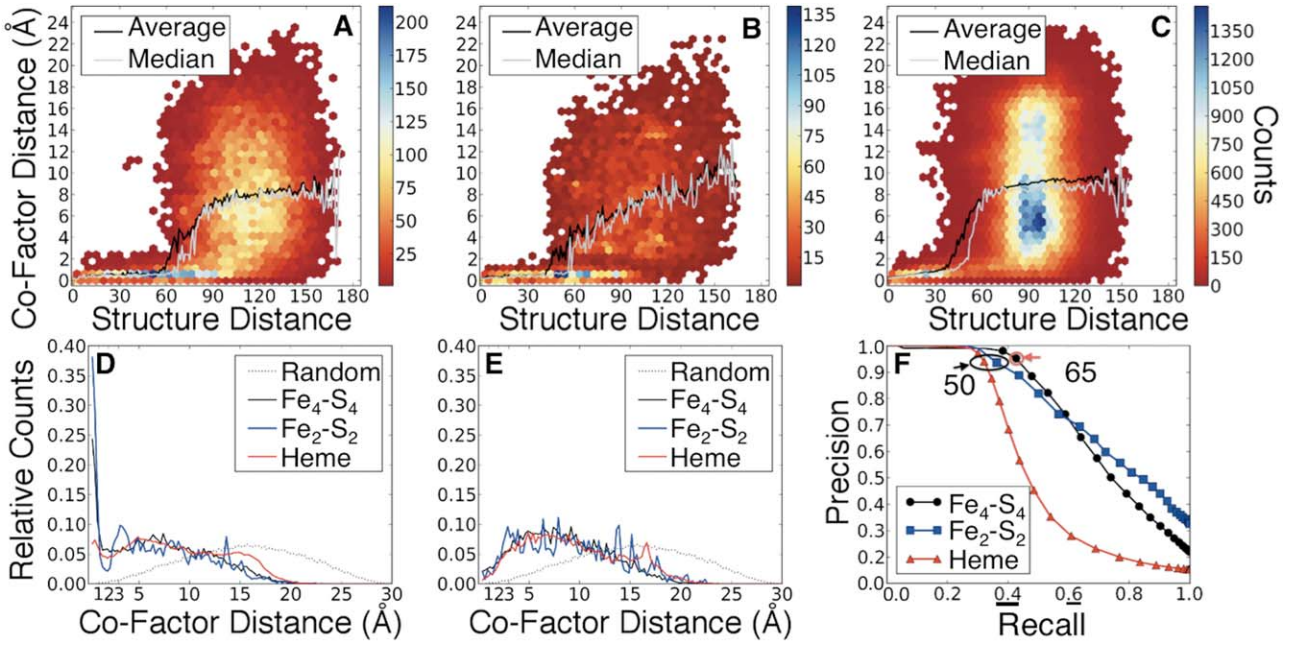

Figure 3

Ligand–ligand distance indicates structural relationships of ligand-binding protein structure environments. In (**A**) 26,565 microenvironment alignments between Fe_4-S_4 containing protein structures are plotted, where the *x*-axis represents the structural distance and the *y*-axis represents the ligand–ligand distance in the calculated superimposition. The colors correspond to the counts of alignments in each bin (structural distance interval). For each structural distance value, the average (black line) and median (gray line) ligand–ligand distance are also plotted. (**B**) and (**C**) are plotted as in (A), but with data from Fe_2-S_2 (9,316 alignments) and heme (237,698 alignments) microenvironments, respectively. (**D**) shows the frequencies of ligand–ligand distances (in 0.25 Å bins) for the complete dataset of alignments for Fe_4-S_4 (black line), Fe_2-S_2 (blue line), and heme-containing environments (red line), compared with the pairwise-distance frequencies from the idealized random-point model (dotted line). (**E**) shows the same data as in (D), but restricted to high structural distances (>120 residues), where the alignments represent typical TopMatch random results. In (**F**) a precision/recall [Eq. (2)] curve is shown for the three alignment datasets. Alignments with ligand–ligand distances below the (>90% precision) distance thresholds of 2.8 Å (Fe_4-S_4, black line), 2.1 Å (Fe_2-S_2, blue line), and 3.0 Å (heme, red line) were deemed true (correct) and false (random/noise) otherwise.

difference is expected as, for most protein pairs, at least one alignment can be constructed that fulfills the criteria of the structure comparison method. Such an alignment places the co-factors closer together than the theoretical maximum distance. Thus, the random model has the average and the median ligand–ligand distances of 15 Å – the microenvironment radius. The average real co-factor distance for the dataset of alignments with high structural distances (D >120 residues) is 9.38 Å for hemes, 8.97 Å for Fe_2-S_2, and 8.37 Å for Fe_4-S_4 clusters; the medians are 8.75, 8.52, and 7.95 Å, respectively. These numbers also mean that low random co-factor distances are highly unlikely, since <5% of baseline (random) alignments have co-factor distances <2.0 Å. The threshold of reliably related alignments can thus be chosen as the structural distance where the average co-factor distance starts to drastically increase, for example, $D = 40$ for heme and Fe_2-S_2 and $D = 60$ for Fe_4-S_4 (Fig. 3). At these distances, more than 98% of alignments have ligand–ligand distances lower than 2 Å.

To find the threshold for D that defines high confidence alignments we generated a precision-recall plot [Fig. 3(F); Eq. (2)] at various threshold values of D. We calculated the precision and recall using alignments defined as correct at ligand–ligand distance cut-offs of at or below 2.8 Å for Fe_4-S_4, 2.1 Å for Fe_2-S_2, and 3.0 Å for hemes, and incorrect otherwise. These co-factor distances were chosen manually after close inspection of the plots in Figure 3(A–C) (materials and methods). Setting the D threshold to 65 for Fe_4-S_4 and to 50 for Fe_2-S_2, and heme yielded precision over 90%.

Applications of ligand–ligand distance in structural comparison

The applications of our approach range from facilitating large-scale comparisons of various ligand-binding proteins to the assessment of protein relationships in highly specific cases. In this study, we focused on a small subset of metalloproteins, but our method could be generalized to other protein centers such as active or binding sites. For example, the many available serine protease structures can be compared and their relationships quantified using the distances between their catalytic triads.

One advantage of using protein co-factors for such an assessment is their independence from most alignment generation processes. Protein co-factors can thus serve as functional beacons for navigating the space of structural relationships to find biological relevance.

Clustering of microenvironments

We used the thresholds of D ($D = 65$ for Fe_4-S_4 and $D = 50$ for Fe_2-S_2, and heme) described above to cluster the microenvironments of proteins binding the same co-factors. For Fe_4-S_4, the 231 structures grouped into 39 distinct clusters, 18 of which were single structures (Supporting Information Table SIII). The largest of the remaining 21 clusters contained 101 structures (43.72% of all Fe_4-S_4 structures) and represented the ferredoxin fold family. This cluster showed a completeness of 41.8% (defined as the fraction of all possible edges/alignments between all nodes/microenvironments in the cluster), while the representative structure directly connected to 83.00% of all other structures in the group. The ferredoxin group included the N- and C-terminal halves of the classic ferredoxin fold and those of the α-helical ferredoxin.

For Fe_2-S_2, the 137 structures grouped into 19 clusters, 10 of which were singletons. The largest cluster contained 57 Fe_2-S_2 ferredoxins (ubiquitin-like fold). The second largest cluster contained 42 structures of the rieske-like Fe_2-S_2 binding domain. The Rieske-like structures were more densely connected (53.08% of all possible alignments) than the Fe_2-S_2 ferredoxin cluster (31.29% of all alignments), implying lower structural diversity and possibly later emergence – a claim supported by the relatively high redox potential of Rieske-like proteins.[31]

The 690 heme-binding structures clustered into 85 distinct groups (45 singletons). The largest group (223 structures) contained various loosely connected (5.44% of all alignments) heme-binding microenvironments from cytochromes. The second largest group (133 structures) formed a denser cluster (71.20% of all alignments) of globin binding sites. The low connectedness of the cytochrome group might be due to the wide range of redox potentials covered by the heme folds of this group. These results support the hypothesis that cytochromes are further evolutionary diverged than globins.[32,33]

Ligand environments in the immediate co-factor proximity are functionally significant

We assume that selection pressure is higher for the amino acids closest to the metal co-factor, causing high levels of conservation of the immediate metal-binding fold. This might explain low co-factor distances at $D >> 0$, and the presence of similar ligand binding motifs in overall different structural contexts. We expect that this trend is reflected in sequence and can be shown with related proteins from phylogenetically distant organisms.

We examined three related proteins: two nitrogenase iron proteins, from *Azotobacter vinelandii* (PDB:1de0[34]) and *Clostridium pasteuranium* (PDB:1cp2[35]), and the dark-operative protochlorophyllide reductase from *Rhodobacter sphaeroides* (PDB:3fwy[36]). The latter is structurally similar to the nitrogenase iron proteins.[37] All three proteins are homo-dimers with one Fe_4-S_4 cluster at the chain interface. The two nitrogenases have 70% identical residues (of the 528 in the structural alignment). The protochlorophyllide reductase is less similar with 33% (of 471 aligned residues), and 35% (of 459 aligned residues) sequence identity with 1de0 and 1cp2, respectively. Nitrogenase iron protein is crucial for the function of the nitrogenase complex. It hydrolyzes ATP and transfers electrons via the bound Fe_4-S_4 cluster to the nitrogen-fixating molybdenum–iron protein. Crystallography studies of the *C. pasteuranium* nitrogenase identified 76 residues (of 273 total) as being functionally important either for the MoFe-protein interaction, ATP binding, or dimer interface generation.[35] A third of these functional residues is located within an 18 Å sphere around the iron–sulfur cluster. The fraction of functionally important residues within the sphere is roughly two-fold higher (47%, 25 of 53) than elsewhere in the protein (23%, 51 residues of 220). This enrichment of functionally significant residues closer to the iron–sulfur cluster suggests more rigid structural constraints in close proximity of the metal co-factor.

To compensate for the effects of structure-driven bias of experimentally defined functionally important residues, we additionally made sequence-based computational predictions of residue functional importance. Functional importance was defined as the level of amino acid mutability of the amino acids – ability to replace the wild-type amino acid by another without effect on function. We calculated the mutability of each residue in the three proteins (SubMat score; materials and methods). A positive SubMat score means that changing the residue is likely disruptive to protein function, but unlikely to have an effect on structure. Substituting residues closer to the ligand is more disruptive of protein function than mutating amino acids further away [Fig. 4(A,B) and Supporting Information Fig. S1]. Note that the patterns of residue mutability are not well explained by protein sequence [Fig. 4(A) and Supporting Information S1(A,C)] but are clearly visible in structure [Fig. 4(B) and Supporting Information S1(B,D)]. These results are indicative of the function-constrained evolutionary pressure acting to maintain the structure of metal-binding motifs.

We further calculated the microenvironment structural alignments for all three possible pairs, increasing sphere radii by 1 Å starting at 10 Å. The sequence identities of

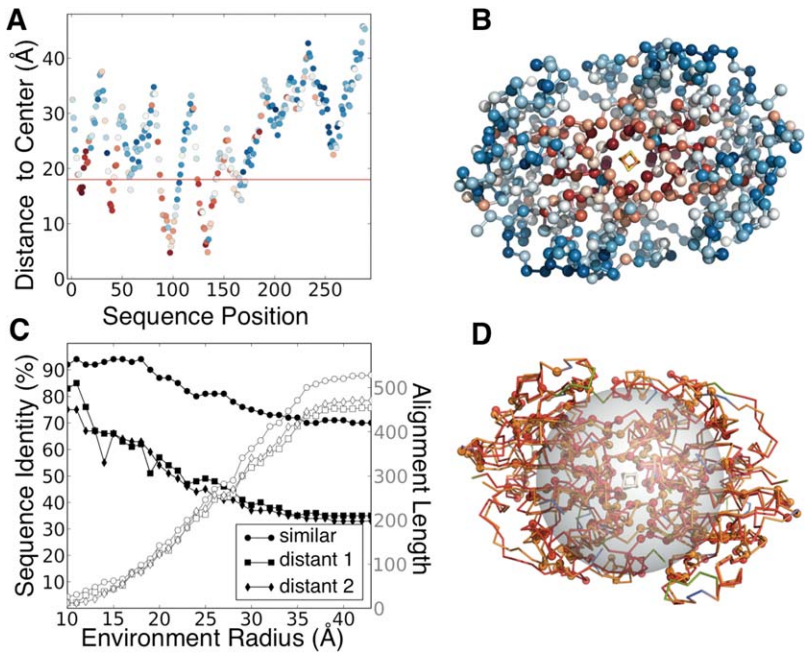

Figure 4

Sequence conservation of the metal binding motif is highest in the immediate vicinity of the ligand. (**A**) For the *Azotobacter vinelandii* nitrogenase iron protein (1de0) the sequence position versus the distance from the Fe_4-S_4 cluster is plotted. Every dot represents a residue and is colored according to its SNAP SubMat score (red - highest/not neutral, blue - lowest/neutral). The red horizontal line marks 18 Å distance from the iron–sulfur cluster. The corresponding structure is shown in (**B**). The Cα-atoms are shown as spheres and are colored the same as in (A). The highlighting demonstrates that the residues closest to the ligand-binding site are more evolutionarily constrained than residues further away. (**C**) shows the loss of sequence identity between structure pairs as a function of the distance from the iron–sulfur cluster center (black line). The alignment length is also shown (grey line). The similar pair (circles) is 1de0 and 1cp2, the distant1 pair (squares) is 1cp2 and 3fwy, and the distant2 pair (triangles) is 1de0 and 3fwy. (**D**) shows the structural alignment between the *Clostridium pasteuranium* nitrogenase (1cp2, orange if aligned, blue otherwise) and the *Rhodobacter sphaeroides* protochlorophyllide reductase (3fwy red if aligned, green otherwise). Identical amino acids are shown as spheres at the Cα position of the residues. The translucent sphere marks an 18 Å distance from the center of the iron–sulfur cluster of 1cp2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the structure alignments decrease as the microenvironment radii increase [Fig. 4(C)].

For the pair of nitrogenases, we observe a plateau of >90% sequence identity up to a radius of 18 Å [alignment length 50–100 residues, Fig. 4(C)]. Between 18 and 24 Å, the sequence identity declines, reaching a second plateau between 24 and 28 Å at around 80% sequence identity. At radii higher than 28 Å, the sequence identity falls to its global value (70%), resulting in a 1.3-fold decrease in percentage of identical amino acids from local (10 Å) to global alignment. For the two more distant pairs we observe similar plateaus between 12 and 18 Å with 65–70% identity and between 24 and 28 Å with ∼50% sequence identity for both pairs. At radii >28 Å the sequence identity drops to the global limits of 33 and 35%, respectively. Sequence identity decreases ∼2.3-fold from local to global alignments for both distant pairs.

We interpret the first plateau of sequence identity in all alignments as the influence of iron–sulfur cluster binding. We speculate that any mutations in that shell of up to 100 amino acids has the potential to severely disrupt either the binding, the orientation, or the redox potential of the cluster and therefore render the protein less functional; that is, there is strong selection pressure in this region of the protein. The residues within that 18 Å sphere around the ligand of all three proteins exhibit significantly higher (*P*-value $\leq 2.53 \times 10^{-6}$, rank-sum test[38]) SubMat scores than the other residues. Within an 18 Å sphere the median SubMat scores were positive, that is, changing the residue is indeed functionally disruptive, for all three proteins (1cp2 at 11.33, 1de0 at 21.33, and 3fwy at 18.00) and negative, that is, functionally non-disruptive, outside the sphere (−36.92, −28.75, and −20.20, respectively). The second plateau between 24 and 28 Å represents the radii where parallel beta-sheets are located to the left and right of the central iron–sulfur cluster [Fig. 4(D)]. The beta sheet contains a high number of identical residues [spheres in Fig. 4(D)] and may be more sensitive to mutations than the surrounding helices or loops. Alignment lengths increase uniformly and steadily until a radius of about 28 Å,

where the high-homology pair shows increasingly higher values compared with the distant pairs, which is due to better alignments on the 'outskirts' of the orthologous structures.

CONCLUSION

Structure comparison significantly contributes to protein analysis but suffers from subjective definition of structural similarity.[1] Structural alignments of similar length and/or RMSD can describe either noise/random or functionally relevant relationships (Fig. 2). In this work, we show that the Euclidian distance between centers of bound co-factors of aligned proteins can be used to substantiate the alignment relevance and thereby helps finding homologous structures. In addition, clusters of structurally aligned proteins provide a good estimate of the evolutionary stability of certain binding motifs.

Our examination of three closely related protein structures shows that the selection pressure influence of a bound iron–sulfur cluster may reach as far 18 Å from the co-factor centroid. We propose that the usefulness of the ligand–ligand distance in judging alignment quality is motivated by the strong conservation of structural motifs in the immediate vicinity of the functionally important bound co-factors.

ACKNOWLEDGMENTS

We thank Markus Wiederstein (University of Salzburg) who helped us with the installation of TopMatch. Fei Xu, John Kim (both Rutgers University), and Markus Gruber (University of Salzburg) for valuable discussion.

REFERENCES

1. Hasegawa H, Holm L. Advances and pitfalls of protein structural alignment. Curr Opin Struc Biol 2009;19:341–348.
2. Sierk ML, Kleywegt GJ. Deja vu all over again: finding and analyzing protein structure similarities. Structure 2004;12:2103–2111.
3. Slater AW, Castellanos JI, Sippl MJ, Melo F. Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. Bioinformatics 2013;29:47–53.
4. Mayr G, Domingues FS, Lackner P. Comparative analysis of protein structure alignments. BMC Struct Biol 2007;7:50.
5. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
6. Hollup SM, Sadowski MI, Jonassen I, Taylor WR. Exploring the limits of fold discrimination by structural alignment: a large scale benchmark using decoys of known fold. Comput Biol Chem 2011; 35:174–188.
7. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 2010;26:889–895.
8. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005;346:1173–1188.
9. Taylor WR. Decoy models for protein structure comparison score normalisation. J Mol Biol 2006;357:676–699.
10. Jia Y, Dewey TG. A random polymer model of the statistical significance of structure alignment. J Comput Biol 2005;12:298–313.
11. Wrabl JO, Grishin NV. Statistics of random protein superpositions: p-values for pairwise structure alignment. J Comput Biol 2008;15: 317–355.
12. Harel A, Falkowski P, Bromberg Y. TrAnsFuSE refines the search for protein function: oxidoreductases. Integr Biol 2012;4:765–777.
13. Torrance JW, Bartlett GJ, Porter CT, Thornton JM. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 2005; 347:565–581.
14. Zheng H, Chruszcz M, Lasota P, Lebioda L, Minor W. Data mining of metal ion environments present in protein structures. J Inorg Biochem 2008;102:1765–1776.
15. Fontecilla-Camps JC, Amara P, Cavazza C, Nicolet Y, Volbeda A. Structure–function relationships of anaerobic gas-processing metalloenzymes. Nature 2009;460:814–822.
16. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science 2008;320:1034–1039.
17. Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anolles G. History of biological metal utilization inferred through phylogenomic analysis of protein structures. Proc Natl Acad Sci USA 2010; 107:10567–10572.
18. Kim JD, Rodriguez-Granillo A, Case DA, Nanda V, Falkowski PG. Energetic selection of topology in ferredoxins. PLoS Comput Biol 2012;8:e1002463.
19. Krishna SS, Sadreyev RI, Grishin NV. A tale of two ferredoxins: sequence similarity and structural differences. BMC Struct Biol 2006;6:8.
20. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
21. Sippl MJ, Wiederstein M. Detection of spatial correlations in protein structures and molecular complexes. Structure 2012;20:718–728.
22. Sippl MJ. On distance and similarity in fold space. Bioinformatics 2008;24:872–873.
23. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 2007;35:3823–3835.
24. Hecht M, Bromberg Y, Rost B. News from the protein mutability landscape. J Mol Biol 2013;425:3937–3948.
25. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
26. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
27. Giastas P, Pinotsis N, Efthymiou G, Wilmanns M, Kyritsis P, Moulis JM, Mavridis IM. The structure of the 2[4Fe-4S] ferredoxin from Pseudomonas aeruginosa at 1.32-A resolution: comparison with other high-resolution structures of ferredoxins and contributing structural features to reduction potential values. J Biol Inorg Chem 2006;11:445–458.
28. Fromme JC, Banerjee A, Huang SJ, Verdine GL. Structural basis for removal of adenine mispaired with 8-oxoguanine by MutY adenine DNA glycosylase. Nature 2004;427:652–656.
29. Knauer SH, Buckel W, Dobbek H. Structural basis for reductive radical formation and electron recycling in (R)-2-hydroxyisocaproyl-CoA dehydratase. J Am Chem Soc 2011;133:4342–4347.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
31. Snyder CH, Merbitz-Zahradnik T, Link TA, Trumpower BL. Role of the Rieske iron-sulfur protein midpoint potential in the protonmotive Q-cycle mechanism of the cytochrome bc1 complex. J Bioenerg Biomembr 1999;31:235–242.
32. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol 1965;8:357–366.

33. Margoliash E. Primary structure and evolution of cytochrome C. Proc Natl Acad Sci USA 1963;50:672–679.
34. Jang SB, Seefeldt LC, Peters JW. Modulating the midpoint potential of the [4Fe-4S] cluster of the nitrogenase Fe protein. Biochemistry 2000;39:641–648.
35. Schlessman JL, Woo D, Joshua-Tor L, Howard JB, Rees DC. Conformational variability in structures of the nitrogenase iron proteins from Azotobacter vinelandii and Clostridium pasteurianum. J Mol Biol 1998;280:669–685.
36. Sarma R, Barney BM, Hamilton TL, Jones A, Seefeldt LC, Peters JW. Crystal structure of the L protein of Rhodobacter sphaeroides light-independent protochlorophyllide reductase with MgADP bound: a homologue of the nitrogenase Fe protein. Biochemistry 2008;47:13004–13015.
37. Burke DH, Hearst JE, Sidow A. Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins. Proc Natl Acad Sci USA 1993;90:7134–7138.
38. Wilcoxon F. Individual comparisons of grouped data by ranking methods. J Econ Entomol 1946;39:269.