Modular origins of biological electron transfer chains

Hagai Raanan[a,b], Douglas H. Pike[b], Eli K. Moore[a], Paul G. Falkowski[a,c,1], and Vikas Nanda[b,d,1]

[a]Environmental Biophysics and Molecular Ecology Program, Department of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901; [b]Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854; [c]Department of Earth and Planetary Sciences, Rutgers University, New Brunswick, NJ 08901; and [d]Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers University, Piscataway, NJ 08854

Contributed by Paul G. Falkowski, December 20, 2017 (sent for review August 28, 2017; reviewed by Marilyn R. Gunner and Ivan V. Korendovych)

Oxidoreductases catalyze electron transfer reactions that ultimately provide the energy for life. A limited set of ancestral protein-metal modules are presumably the building blocks that evolved into this diverse protein family. However, the identity of these modules and their path to modern oxidoreductases is unknown. Using a comparative structural analysis approach, we identify a set of fundamental electron transfer modules that have evolved to form the extant oxidoreductases. Using transition metal-containing cofactors as fiducial markers, it is possible to cluster cofactor microenvironments into as few as four major modules: bacterial ferredoxin, cytochrome c, symerythrin, and plastocyanin-type folds. From structural alignments, it is challenging to ascertain whether modules evolved from a single common ancestor (homology) or arose by independent convergence on a limited set of structural forms (analogy). Additional insight into common origins is contained in the spatial adjacency network (SPAN), which is based on proximity of modules in oxidoreductases containing multiple cofactor electron transfer chains. Electron transfer chains within complex modern oxidoreductases likely evolved through repeated duplication and diversification of ancient modular units that arose in the Archean eon.

oxidoreductase | electron transfer | metalloprotein | evolution | network

Global electron transfer reactions maintain chemical disequilibrium across the major geophysical fluids: the atmosphere, ocean, and mantle. These disequilibria are largely dependent on life (1–5). It is thought these essential electron transfer processes are driven by a limited set of functionally homologous gene classes with critical roles in chemotropic and phototrophic metabolic reactions (3). These genes primarily belong to the Enzyme Commission 1 (EC1) proteins, the oxidoreductases (2, 6–8). The oxidoreductases are old, and most are proposed to have evolved during a period of intense genetic, microbial innovation in the Archean (9). Many extant oxidoreductases are massive protein nanomachines, consisting of multiple protein subunits and dozens of cofactors. Such complex proteins would not emerge spontaneously and must have evolved through intermediate, simpler forms.

The evolutionary origins of redox modules and how they assemble into electron transfer chains (ETCs) are obscured by lateral gene transfer and extensive selection. It is not clear whether oxidoreductases evolved from one universal common ancestor or from several independent origins (6, 7, 10). Given that oxidoreductases evolved in microbes over the first *ca.* 2.5 billion years of Earth history, the application of traditional sequence-based molecular clock methods for estimating the age of these proteins is fundamentally limited (11). To circumvent this, permissive metal-binding sequence profile alignments (6), or estimating age by extent of gene duplication (1), have been applied. Here, we use a structure-based approach to identify the fundamental modules that comprise the functional units of oxidoreductases.

Modern oxidoreductases arose from a basic set of metalloprotein modules (10), in which transition metals are often responsible for function (12, 13). The protein cofactor microenvironment tunes the electrochemical properties of the metal cofactor (14, 15), as has been demonstrated by protein engineering (16, 17). Hence, the metal cofactor and its protein microenvironment are interdependent. We examine networks of structural similarity for the identification of these modules and rules for assembly into ETCs.

Metal cofactors serve two purposes in our study that improve the power of comparative structural analysis (Fig. 1). The first is as fiducial markers of the quality of protein structure alignments (18). In addition to standard metrics of alignment quality (e.g., rmsd, sequence identity, alignment length), the metal-metal alignment was included as a second assessment of structural similarity. This allows us to potentially identify distantly related folds with weak structural similarity, which is critical, given the antiquity of oxidoreductases. Second, metal cofactors can be used to investigate the spatial arrangement of modules in ETCs. The Moser–Dutton ruler (19) specifies that for efficient electron transfer, cofactors should be located within 14 Å of each other in the protein matrix. Pairs of microenvironments co-occurring in an oxidoreductase structure with cofactor separations within this limit have the potential to engage in electron transfer. Using this criterion, we produced a spatial adjacency network (SPAN) to identify consistent patterns of how modules are wired together in ETCs.

Results and Discussion

Identifying Cofactor Modules. From 9,500 high-resolution metalloprotein structures in the Protein Data Bank (PDB), ∼32,000 microenvironments were defined by a 15-Å radius from the metal center, including surrounding amino acids (18). The term microenvironment is used instead of fold or domain as these units often contain discontinuous elements of sequence and structure, sometimes from multiple chains (20). Comparative structural alignments of microenvironments were scored according to a weighted combination of rmsd and alignment length, and subsequently filtered by distance between the centroids of the cofactors defining the microenvironments. The metal centers serve as fiducial markers to evaluate the quality of the alignment (18). The distribution of alignment scores versus metal center distances of the alignments show two clearly distinguished populations of structures (Fig. 2*A*). We chose a threshold similarity score of 1.4 as a conservative cutoff

Significance

There are no physical fossils of the original proteins at the beginning of life on Earth and phylogenetic approaches that infer the nature of the ancestral proteins from sequences and/or structures of extant molecules are of limited use over long time scales (e.g., billions of years). We analyzed the structures of proteins containing transition-metal cofactors, and identified four structural modules that comprise the diverse family of oxidoreductases, molecular nanomachines that are critical for electron transfer reactions that form the energetic basis of life. These structural modules are, in effect, relict "building blocks" of life that have descended through time with only minor modifications.

Author contributions: H.R., D.H.P., P.G.F., and V.N. designed research; H.R., D.H.P., E.K.M., and V.N. performed research; H.R., D.H.P., E.K.M., P.G.F., and V.N. analyzed data; and H.R., D.H.P., E.K.M., P.G.F., and V.N. wrote the paper.

Reviewers: M.R.G., City College of New York; and I.V.K., Syracuse University.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: falko@marine.rutgers.edu or nandavi@cabm.rutgers.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1714225115/-/DCSupplemental.

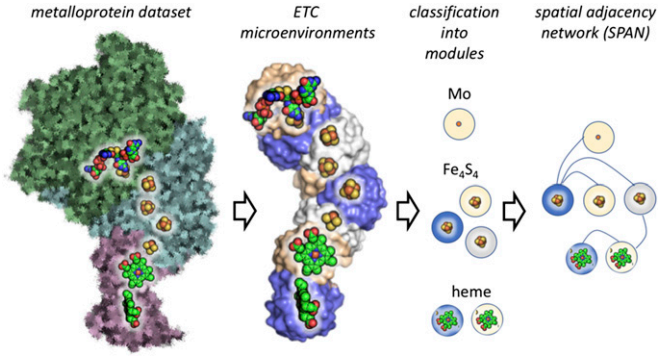

Fig. 1. General strategy for probing the modular structure of ETCs. Metalloprotein structures are decomposed into microenvironments and clustered into a set of modules. Modules with microenvironments that are close enough to allow electron transfer are connected in the SPAN. The SPAN provides unique insight into ETC topology and pathways for ETC emergence and evolution. This structural example comes from nitrate reductase (PDB ID code 1q16) (48).

to minimize potential false-positive results. After clustering, a minimal set of 1,017 unique modules was identified. Modules may include microenvironments with different types of cofactors or even different metals. The distribution of cluster sizes followed a power-law dependence (Fig. 2*B*), with half of all microenvironments contained within the 10 most populated modules. Such a power-law distribution is generally observed for protein domain family frequencies across genomes (21–23), supporting the premise that a 15-Å cofactor microenvironment captures an evolutionarily relevant functional domain. Among the most populous modules are those associated with oxidoreductase functions. These include bacterial ferredoxin, cytochrome c, symerythrin, and plastocyanin-like folds (Fig. 2*C*). The centrality of these four modules to the evolution of ETCs is described below.

Each module is a component within the network of microenvironments connected by structural similarity. This does not imply that every pair of microenvironments within a given module has similarities below the selection threshold; instead, some are connected via a series of intermediate alignments. For example, the module containing the 4Fe4S-binding bacterial ferredoxin motif contains both mixed α/β and all-helical folds. The standard $(\beta–\alpha–\beta)_2$ topology (24) consists of two microenvironments, one for each iron-sulfur cluster, that form two major subgroups related to their proximity to the N and C termini (Fig. 3*A*). A direct alignment of these two topologies shows only a minimal two-helix overlapping region that contains six of the eight cysteines required to coordinate clusters (Fig. 3*B*). However, we identified intermediate forms between the all-helical and $(\beta–\alpha–\beta)_2$ topologies, suggesting these may be evolutionarily related (Fig. 3*C*). Although intermediate forms are identified, this analysis alone does not prove a common ancestor for the ferredoxin module; evolutionary relationships based on local structural alignments can be misleading. The all-helical ferredoxin also shows strong similarity to a heme-binding globin fold (25). Additional information beyond structural similarity would be required to determine whether structures are related by homology or analogy.

The challenge of identifying evolutionary relationships is evident in the most populated module of the cytochrome c-like fold. That module includes structures from small, single-heme proteins to large, multiple cofactor complexes. The network of microenvironments within this component contains three major subgroups interconnected by a small number of edges (Fig. 3*D*). While a CXXCH motif-bound heme is common to all microenvironments,

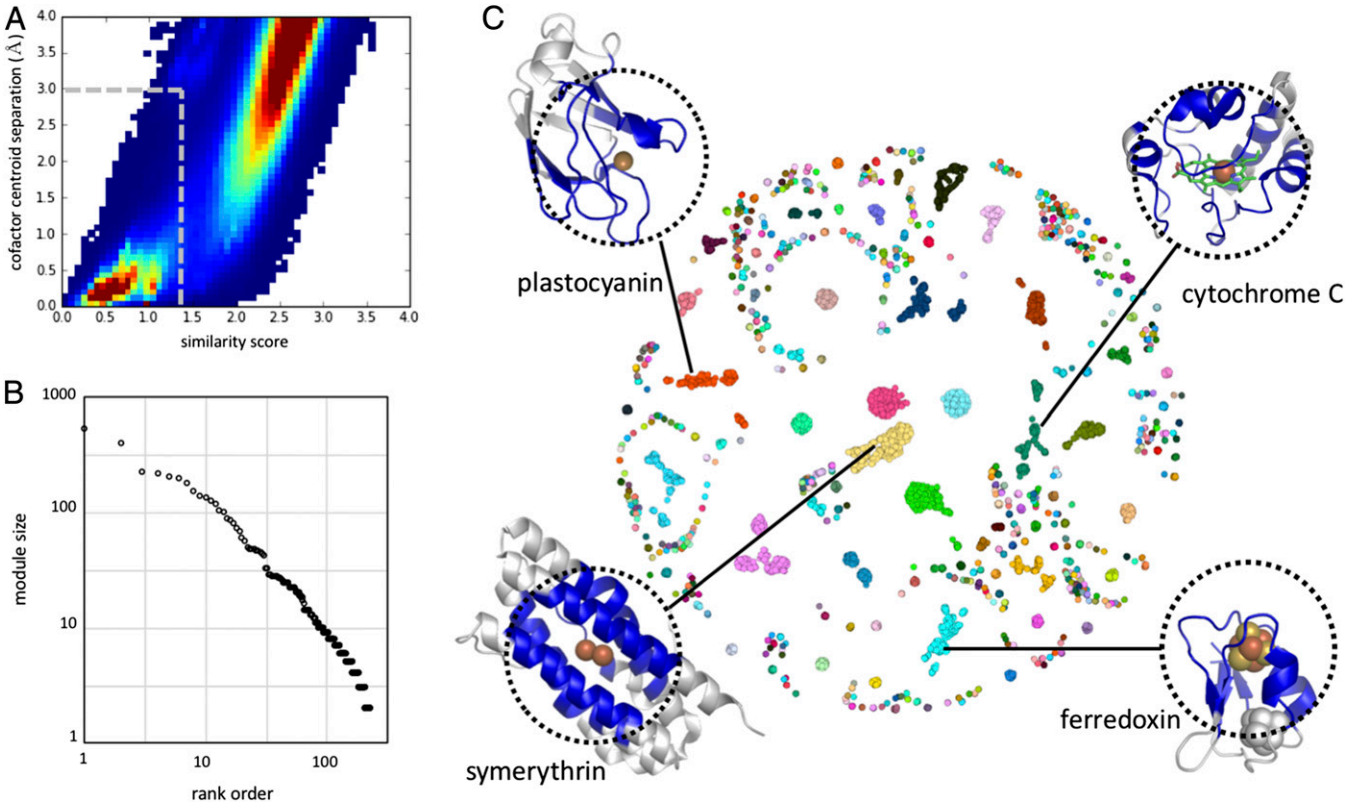

Fig. 2. Identifying cofactor modules. (*A*) Clustering of pairwise structural alignment of microenvironments based on overall similarity and cofactor distance. Accepted alignments are shown in the box in the lower left corner. (*B*) Number of microenvironments per module scales linearly with module size rank order on a log-log scale, supporting modules as functional domains. (*C*) Microenvironments that belong to connected components define the set modules. Structural exemplars of major modules are shown.

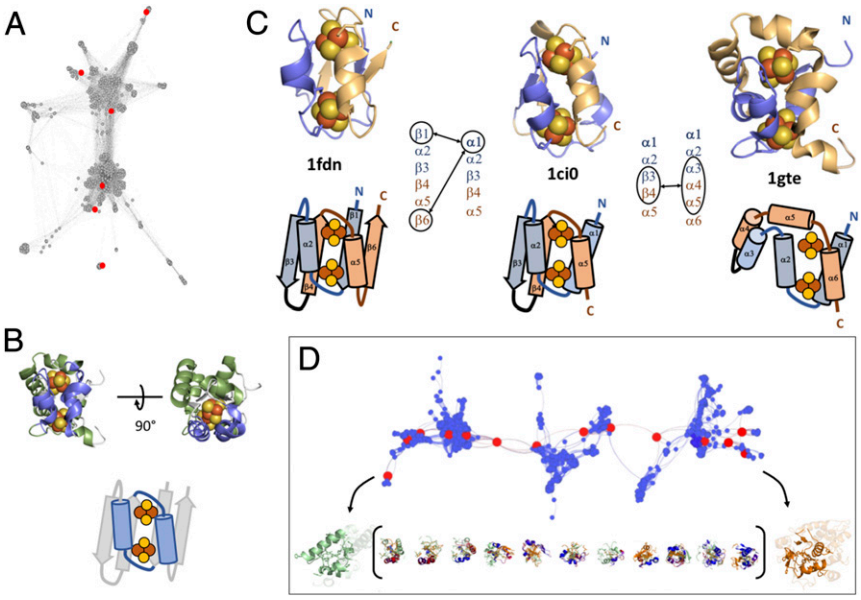

Fig. 3. Structural breadth of the modules. (*A*) Microenvironment network of the bacterial ferredoxin module, highlighting in red the 1FDN, 1CI0, and 1GTE microenvironment pairs. (*B*) Direct alignment of these two forms shows limited two-helix overlap. (*C*) Three ferredoxin-like folds can be related by secondary structure substitutions, indicating possible intermediates between $(\beta–\alpha–\beta)_2$ and all-helix forms. (*D*) Microenvironment network and the minimal path between two cytochrome c microenvironments shown for (from left to right) PDB ID codes 2j7a, 1fgj, 4rkn, 1ft6, 1sp3, 2ldo, 3ov0, 2b7r, 3x39, 4xxl, 1c6s, 2c1u,1eb7, and 2ozl.

the secondary structure elements around the heme vary significantly. For example, two very different microenvironments at opposite ends of this module require a minimum of 12 intermediates to be connected. These intermediates represent significant structural changes that cannot be ascribed to conservative changes in sequence (e.g., point mutations), and might not be detected by sequence-based alignment methods. As with ferredoxins, intermediates cannot be conclusively interpreted as homologous. Such connections may reflect fundamental constraints of metal coordination on local protein structure (26). Similar properties are observed in the plastocyanin and symerythrin network components.

Modules represent the minimal structural elements required for metal cofactor binding, but vectoral electron transfer requires multiple cofactors, and therefore multiple modules (5). To understand how modules are combined to produce ETCs, we next examined the spatial organization of cofactors in the structural dataset.

SPAN. All proteins containing multiple metallocofactors in the PDB were analyzed to identify pairs of modules where the edge-to-edge distance of the central cofactors was within the 14-Å limit specified by the Moser–Dutton ruler. This produced 8,978 pairs of microenvironments, which were used to construct a spatial adjacency network (SPAN). The SPAN is essentially a network of networks, where each node is a module containing a network of structurally similar microenvironments. The SPAN contained 93 connected components, with most consisting of two to five modules (Fig. S1). However, the largest connected component contained 105 modules consisting primarily of microenvironments extracted from oxidoreductases (Fig. 4 and Table S1).

Connections between modules in the SPAN are unevenly distributed and deviate significantly from what would be expected for a random graph. Four major modules (bacterial ferredoxin, cytochrome c, plastocyanin, and symerythrin) comprise the vast majority of connections, consistent with these being from the most populous modules in the dataset. These modules are repeatedly utilized across oxidoreductases of diverse function, suggesting their high functional utility in constructing ETCs, reinforcing a model in which modern oxidoreductases arose from modular assembly of reusable cofactor microenvironments.

Approximately one in five modules in the oxidoreductase SPAN has self-connections, where multiple instances of the same module are adjacent in structure. These are distributed between microenvironments in either the same or adjacent chains. Self-connections are expected for modules with internal symmetry, such as found in bacterial ferredoxin. The ferredoxin module consists of a symmetrical pair of microenvironments (24) thought to have arisen through gene duplication (27, 28). Dramatic examples of this include extended polyferredoxins seen in methanogen carbon fixation enzymes (29) or in the multiheme cytochrome c nanowires capable of long-distance electron transfer (30). Other self-connections arise from oligomeric self-assembly as in the case of ferritin and multidomain cupredoxins (Fig. 5). The abundance of self-connections across the SPAN indicates that duplication is a common strategy for the assembly of multiple cofactor ETCs.

Clustering of Cofactor Type in the SPAN. Another feature of the SPAN is the notable clustering of modules with the same metal type: 82% of all edges connect similar cofactor classes. This is highly unlikely for a random network ($P < 0.001$). Fe-S–containing modules, whether Fe_4S_4, FeS_4, or Fe_2S_2, are commonly found in spatial proximity. Similar clustering is seen for porphyrin, copper, mono-iron, and di-iron sites. There are several potential explanations for this. The first is that spatial proximity is constrained by functional properties of the modules (i.e., redox potential). Redox potential differences across adjacent modules should be energetically compatible to minimize the back-reactions and/or inversions (19, 31). However, for Fe-S and hemes, observed redox potentials can span a range of 1 V by changes in first- and second-shell amino acids in the cofactor microenvironment (14, 16, 17, 32). A plausible alternative is that cofactor type is constrained by the complexity associated with the incorporation of different cofactors into a single protein. In modern proteomes, metal incorporation is tightly regulated (33, 34), but in early stages of protein evolution, it is highly likely that metal selection was promiscuous (1).

It is also possible that all of the modules containing a similar cofactor originated from a common ancestor, despite having no detectable structural similarity. Spatial proximity in the SPAN may be a relic of evolutionary homology. For example, the wide variety of Fe-S modules from bacterial and plant-type ferredoxins to rubredoxins may have arisen through duplication and diversification of an Ur-bacterial ferredoxin ancestor. Highly connected modules in the SPAN may represent ancient cofactor-specific domains that served as last universal common ancestors for the modern family of oxidoreductases.

Discriminating Homology and Analogy. If spatial proximity in the SPAN is a signal for deep evolutionary homology, then the SPAN may provide useful information in discriminating structural homology from analogy. As shown earlier, all-helical ferredoxins share only scant structural similarity with the classic $(\beta–\alpha–\beta)_2$ structure (25).

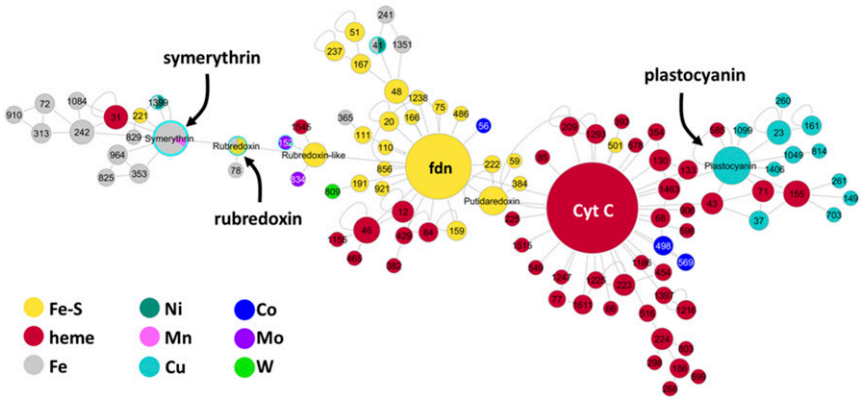

Fig. 4. Oxidoreductase SPAN, the largest connected component in the SPAN, representing the majority of modules found in EC1 proteins. Node size correlates with the number of edges to adjacent module types. Edge thickness correlates with the number of microenvironment pairs connecting module classes. Loopback edges indicate connections between microenvironments of the same module type. Details of module number features and associated microenvironment PDB identities are provided in Table S1 and Dataset S1. fdn, ferredoxin.

However, these two folds are structurally adjacent in the same protein (Fig. 6*A*), supporting a model of duplication and divergence rather than convergence. Similarly, within the cytochrome c module, we identify eight submodules using the Louvain method for community detection in the network (35). When the SPAN is recalculated for these submodules, two connected components (submodules 1,2 and 3–8) are observed, suggesting, at minimum, two classes of homologous cytochrome c-type folds. Combining structural similarity and spatial adjacency provides a tool for proposing modes of deep-time evolutionary connections across protein structures.

It is important to note that not all connections in the SPAN represent common ancestry. Another possible mechanism for the assembly of ETCs is independent evolution of distinct modules containing multiple cofactor types that subsequently were fused into a single protein. The modules that bridge well-connected subnetworks within the SPAN are both independent soluble electron carriers and domains fused to ETCs; for example, the rubredoxin and rubredoxin-like modules bridge the major ferredoxin and symerythrin subnetworks (Fig. 4). Similarly, the connections between ferredoxin and cytochrome c in the SPAN are unlikely due to a common ancestor.

Age of the Core Modules. The four major modules identified here have also been proposed to be ancient folds by other structural informatics methods. The ferredoxin-like and cupredoxin-like fold superfamilies were estimated by Dupont et al. (1) to have emerged early in protein evolution, before the Great Oxidation Event *ca.* 2.4–2.3 billion y ago. Edwards and Deane (36) determined ferredoxin to be a pivotal fold at the base of a phylogenetic tree of global structure space. The ferredoxin $(\beta–\alpha–\beta)_2$ topology, or "plaitfold," is one of 10 superfold structural classes that was proposed to give rise to extant proteins in general (21).

We estimated the age of emergence of ETC modules by combining functional annotations of oxidoreductases with the ages of metabolic pathways inferred from the geological record (4) (Fig. 7*A*). The earliest posited metabolisms (hydrogen, sulfur and sulfate reduction, methanogenesis, and anoxygenic photosynthesis) are overwhelmingly dominated by ferredoxin-like folds and other FeS-containing modules (Fig. 7*B*). Ferredoxin and cytochrome c are the most functionally diverse, found in metabolic pathways present in all stages of the Archean eon. Subsequent metabolisms, nitrogen fixation and oxygenic photosynthesis, use ferredoxin, cytochrome c, and plastocyanin modules. The symerythrin and plastocyanin modules are largely associated with later Archean metabolisms, which include oxidation or aerobic pathways (Fig. 7*A*). During the latter period, the emergence of molecular oxygen depleted soluble Fe in the ocean, necessitating the emergence of iron-storage proteins like ferritin, which belongs to the symerythrin module. The metal distribution shows FeS dominating the earliest metabolisms with heme and Cu increasing in prevalence with time (Fig. 7*B*), consistent with the availability of metals as a result of oxygen fugacity (37, 38). The SPAN contains the earliest and most functionally diverse modules at the center, with the peripheral modules evolving later and becoming more specialized. The structure of the SPAN recapitulates the expansion of metabolic pathways in the Archean eon.

Conclusions

We analyzed deep-time evolutionary connections within the oxidoreductase class of enzymes, extending previous sequence-based approaches (6), by considering protein structure similarities. Analysis of structure has its own challenges; protein folds do not change linearly with accumulating amino acid substitutions, precluding quantitative estimates of evolutionary distance based on structural similarity. In metalloproteins, evolutionary inference is further confounded by strong chemical constraints of metal coordination on the local protein environment, where observed structural similarity between proteins may be a result of convergence on a limited repertoire of metal-binding protein topologies. Even with evidence of spatial

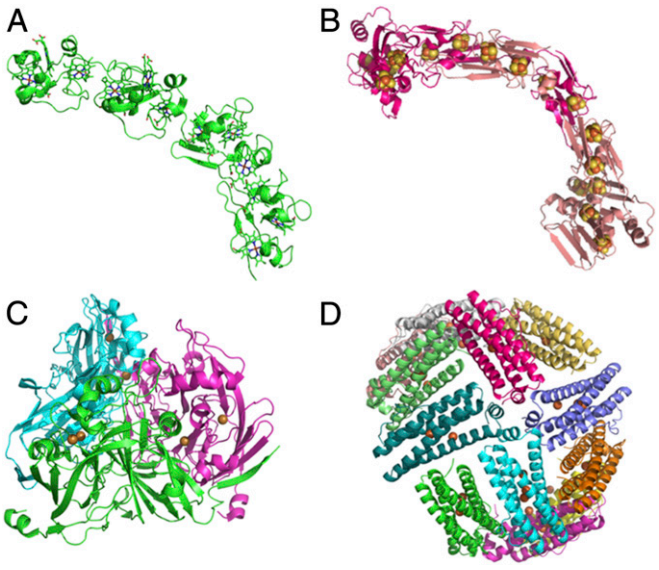

Fig. 5. Module repetition resulting in self-connections in the SPAN. (*A*) Dodeca-heme chain from *Geobacter* (PDB ID code 3ov0). (*B*) Polyferredoxin from the methanogen carbon fixation pathway (PDB ID code 5t5i). (*C*) Plastocyanin (PDB code ID 1j9q). (*D*) Bacterioferretin protein cage of symmerythin-like domains (PDB ID code 3e1m).

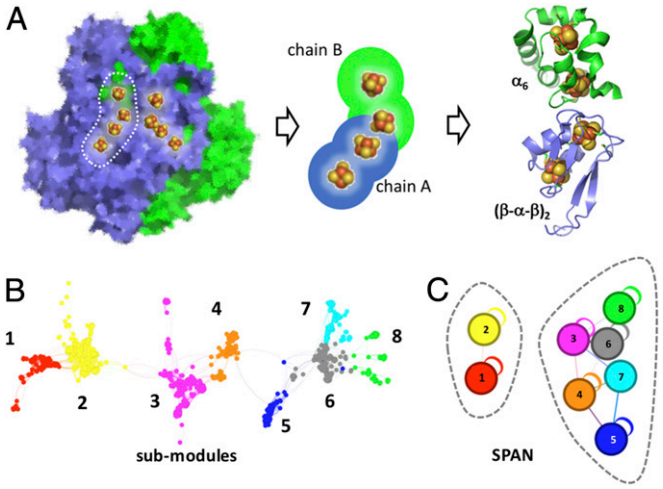

Fig. 6. Discriminating homology and analogy using spatial adjacency in an ETC. (*A*) Colocalization of $(\beta\text{–}\alpha\text{–}\beta)_2$ and all-helix forms of bacterial ferredoxin in a dehydrogenase structure (1GTE) suggests a homology based on duplication and diversification. (*B*) Cytochrome c module can be further divided using a community detection method. (*C*) The SPAN of these submodules suggests, at minimum, two evolutionarily related groups of cytochrome c folds.

coincidence in the SPAN, it would be challenging to prove that microenvironments with disparate structure arose from a common ancestor. This may be possible where pathways of intermediate sequences are found in observed genomic and metagenomic data (6).

Despite these caveats, a clear pattern emerges in the aggregate analysis of metal environments, where modules of similar metal cofactor types cluster in the SPAN. This strongly indicates a small contingent of modular structures were incorporated repeatedly in oxidoreductases across the tree of life. The emergence of complexity derives from two main modes of evolution: (*i*) gene duplication and diversification and (*ii*) recruitment and fusion of independent structures. This basic idea has been proposed for many other systems in biology (39). In the example of metabolic pathways, two fundamental mechanisms of evolution were proposed: (*i*) stepwise retrograde evolution (40), where enzymes diversified from a single substrate-product reaction, and (*ii*) the patchwork model (41, 42), where primordial catalytic functions were fused into specific pathways. The first likely module was ferredoxin, which gave rise to a number of specialized FeS-containing proteins. Independently, cytochrome c, plastocyanin, and symerythrin evolved, giving access to an increased variety of metabolic substrates and redox potentials. Ultimately, these electron transfer modules gave rise to a global electrical circuit that is a hallmark of life on Earth (2).

Methods

Structural Alignment. We generated PDB files of a 15-Å sphere around each transition metal-containing cofactor region (Fe, Cu, Mn, Ni, Mo, Co, V, and W) from the PDB. We found 36,787 spheres (i.e., "microenvironments") with 46 cofactors (PDB ID codes) that contain one of the above-mentioned transition metals, and have 20 or more entries in the PDB. We filtered out microenvironments from PDB files of de novo designed proteins. We used PyMOL (The PyMOL Molecular Graphics System, Version 1.7; Schrödinger, LLC) to perform pairwise alignments of the protein backbone in all of the microenvironments of each cofactor. To identify the structural similarities between the microenvironments of different cofactors, we also performed pairwise microenvironment alignments of all of the cofactors from a nonredundant set of protein chains, containing 6,924 microenvironments. The nonredundant set contained only chains with less than 90% sequence similarity generated by PISCES (43).

Dataset S1 is a Microsoft Excel-formatted spreadsheet that provides an annotated list of all microenvironments used in the analysis. The entries are indexed as follows:

(PDB ID).(cofactor name)_(cofactor chain)_(cofactor residue number)_(functional annotation),

(e.g., 4rvy.HEM_f_101_oxidoreductase).

Similarity Score. We estimated the similarity of pairs of microenvironments based on the rmsd of the alignments and the number of Cα atoms successfully aligned, together with the calculated structural distance (44). We also included the distance between the centroids of the cofactors in the microenvironment alignment (18) in our similarity score calculation. Centroids for cofactors containing multiple ions or organic ligands not contributed by the protein are determined using only the metal ions, excluding auxiliary atoms in the cofactor. The formula for similarity score is as follows:

$$\text{similarity score} = \frac{rmsd}{\max rmsd} + \frac{CA}{\max CA} + \frac{cd}{\max cd} + \frac{sd}{\max sd},$$

where *CA* is the number of backbone Cα atoms successfully aligned, *cd* is the distance between the centroids of the cofactors, and *sd* is the structural distance. In meaningful alignments, the cofactors were expected to be located at equivalent structural positions.

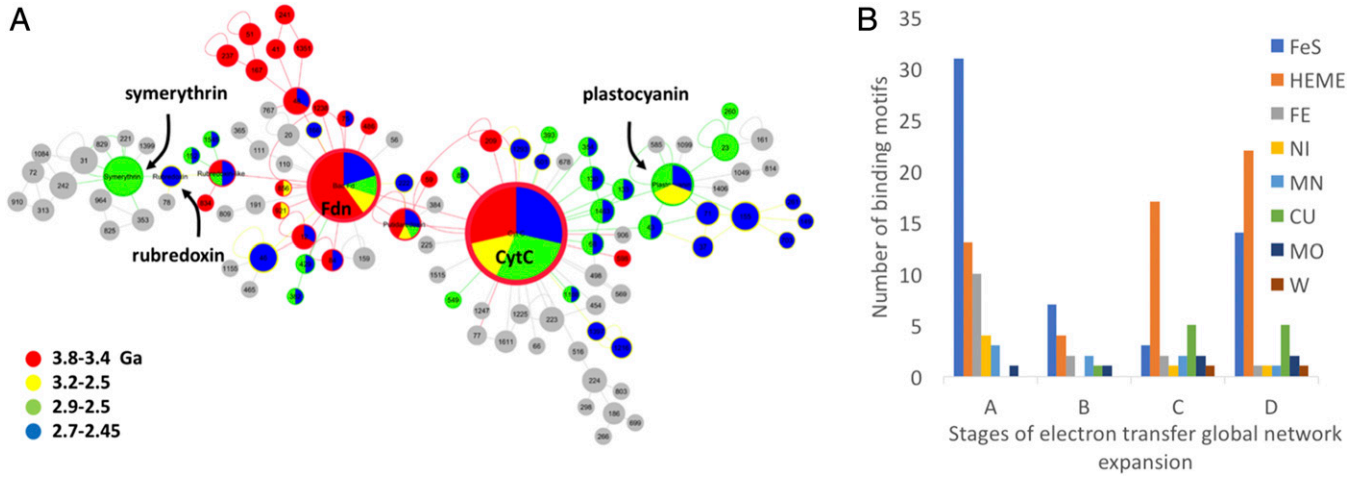

Fig. 7. Estimated age of modules relative to the four stages of global electron transfer network expansion, inferred from the geological record (4). (*A*) SPAN is colored according to the module involvement in core microbial metabolic pathway stages. Stage A (red): methanogenesis, sulfur reduction, sulfate reduction, anoxygenic photosynthesis, and heterotrophy and autotrophy. Stage B (yellow): nitrogen fixation and oxygenic photosynthesis. Stage C (green): methane oxidation, nitrification, denitrification, sulfur oxidation, and sulfide oxidation. Stage D (blue): aerobic respiration, ammonification, and oxidation/reduction of other elements. Node size is relative to the node degree. (*B*) Metallocofactor distribution in each of the four stages of global electron transfer network expansion.

The distribution of similarity scores versus cofactor distances of the alignments (*cd*) exhibit a distinct shape, with two clearly distinguished populations of structures (Fig. 2*A*). Similarity scores between 1.4 and 1.6 were tested. We chose a more conservative similarity score of 1.4 as the optimal threshold to minimize potential false positives. This threshold is >2σ from the mean similarity score (mean = 4.48, SD = 1.46).

Clustering of Microenvironments into Modules. We initially clustered 44,614 microenvironments from all metalloproteins in the PDB as of January 2016 into 1,626 modules based on a similarity score threshold of 1.4. Each module was a connected component in the network, and represented one consensus structure (i.e., module). The number of microenvironments in each module varied between 2 and >2,000. Subsequently, we removed modules where all microenvironments were from a single PDB file. The final dataset included 31,927 microenvironments from 9,531 proteins. These were clustered into 1,017 modules. Most of the microenvironments contained Fe [Fe (5,978), heme (9,237), and FeS (3,277)], followed in order of abundance by Mn (6,362), Cu (3,732), Ni (1,462), Co (1,321), Mo (216), W (210), and V (134).

Generating the SPAN. The SPAN constructed by assigning the modules as nodes in this network were connected by edges if at least one microenvironment from each module was located in the same protein (PDB file), where the cofactors are in electron transfer distance range. Every additional occurrence was counted to create the weight of that edge. We used edge-to-edge cofactor distances between 4.5 Å and 14 Å as relevant electron transfer ranges. The minimum distance of this range was determined to ignore self-connections in multi-ion microenvironments. The maximum distance of this range was determined by the observed maximum possible distance for electron tunneling in natural proteins (19). A χ^2 test was performed to show the statistical significance of the clustering of modules with the same metal type compared with random distribution.

Functional Diversity of Modules. We counted how many different EC numbers each module (clusters of microenvironments) contains. Partial EC numbers were counted only when no other protein from the same class (e.g., 1.7.x.x or 1.x.x.x) was found.

Core Microbial Metabolic Pathways. We used the list of the 392 EC1 homologs identified through a manually refined search of all Kyoto Encyclopedia of Genes and Genomes (www.genome.jp/kegg/) pathways directly involved in redox reactions of biogeochemical interest (3) to assign a list of EC numbers of enzymes involved in the core microbial metabolic pathways. We then assigned a metabolic pathway to all of the microenvironments of a protein according to its EC number. Finally, we assigned a list of pathways that each module was composed of. We further improved this list with a manually refined search of PDBs that are known to be involved in any of the metabolic pathways, and assigned it to the relevant modules.

Network analysis was performed using the NetworkX library for Python (45) and Gephi 0.9.1 (46). We used Gephi and Cytoscape 3.4.0 (47).

ACKNOWLEDGMENTS. We thank John D. Kim and Andrew C. Mutter for useful discussions. This work was supported by a grant from the Gordon and Betty Moore Foundation on "Design and Construction of Life's Transistors" (GBMF-4742). E.K.M. was supported by the Keck Foundation Research Award.

1. Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anollés G (2010) History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proc Natl Acad Sci USA* 107:10567–10572.
2. Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320:1034–1039.
3. Jelen BI, Giovannelli D, Falkowski PG (2016) The role of microbial electron transfer in the coevolution of the biosphere and geosphere. *Annu Rev Microbiol* 70:45–62.
4. Moore E, Jelen B, Giovannelli D, Raanan H, Falkowski P (2017) Metal availability and the expanding redox network of microbial metabolism in the Archean eon. *Nat Geosci* 10:629–636.
5. Gunner MR, Koder R (2017) The design features cells use to build their transmembrane proton gradient. *Phys Biol* 14:013001.
6. Harel A, Bromberg Y, Falkowski PG, Bhattacharya D (2014) Evolutionary history of redox metal-binding domains across the tree of life. *Proc Natl Acad Sci USA* 111:7042–7047.
7. Kim JD, Senn S, Harel A, Jelen BI, Falkowski PG (2013) Discovering the electronic circuit diagram of life: Structural relationships among transition metal binding sites in oxidoreductases. *Philos Trans R Soc Lond B Biol Sci* 368:20120257.
8. Nisbet EG (1995) Archaean ecology: A review of evidence for the early development of bacterial biomes, and speculations on the development of a global-scale biosphere. *Geol Soc Spec Publ* 95:27–51.
9. David LA, Alm EJ (2011) Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 469:93–96.
10. Baymann F, et al. (2003) The redox protein construction kit: Pre-last universal common ancestor evolution of energy-conserving enzymes. *Philos Trans R Soc Lond B Biol Sci* 358:267–274.
11. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27:401–410.
12. Holm RH, Kennepohl P, Solomon EI (1996) Structural and functional aspects of metal sites in biology. *Chem Rev* 96:2239–2314.
13. Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM (2008) Metal ions in biological catalysis: From enzyme databases to general principles. *J Biol Inorg Chem* 13:1205–1218.
14. Liu J, et al. (2014) Metalloproteins containing cytochrome, iron-sulfur, or copper redox centers. *Chem Rev* 114:4366–4469.
15. Aizman A, Case DA (1982) Electronic-structure calculations on active-site models for 4-Fe, 4-S iron sulfur proteins. *J Am Chem Soc* 104:3269–3279.
16. Dey A, et al. (2007) Solvent tuning of electrochemical potentials in the active sites of HiPIP versus ferredoxin. *Science* 318:1464–1468.
17. Hosseinzadeh P, et al. (2016) Design of a single protein that spans the entire 2-V range of physiological redox potentials. *Proc Natl Acad Sci USA* 113:262–267.
18. Senn S, Nanda V, Falkowski P, Bromberg Y (2014) Function-based assessment of structural similarity measurements using metal co-factor orientation. *Proteins* 82:648–656.
19. Page CC, Moser CC, Chen X, Dutton PL (1999) Natural engineering principles of electron tunnelling in biological oxidation-reduction. *Nature* 402:47–52.
20. Bagley SC, Altman RB (1995) Characterizing the microenvironment surrounding protein sites. *Protein Sci* 4:622–635.
21. Orengo CA, Thornton JM (2005) Protein families and their evolution-a structural perspective. *Annu Rev Biochem* 74:867–900.
22. Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. *J Mol Biol* 313:673–681.
23. Molina N, van Nimwegen E (2008) The evolution of domain-content in bacterial genomes. *Biol Direct* 3:51.
24. Adman ET, Sieker LC, Jensen LH (1973) Structure of a bacterial ferredoxin. *J Biol Chem* 248:3987–3996.
25. Krishna SS, Sadreyev RI, Grishin NV (2006) A tale of two ferredoxins: Sequence similarity and structural differences. *BMC Struct Biol* 6:8.
26. Andreini C, Bertini I, Cavallaro G, Najmanovich RJ, Thornton JM (2009) Structural analysis of metal sites in proteins: Non-heme iron sites as a case study. *J Mol Biol* 388:356–380.
27. Eck RV, Dayhoff MO (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* 152:363–366.
28. Davis BK (2002) Molecular evolution before the origin of species. *Prog Biophys Mol Biol* 79:77–133.
29. Wagner T, Ermler U, Shima S (2016) The methanogenic CO2 reducing-and-fixing enzyme is bifunctional and contains 46 [4Fe-4S] clusters. *Science* 354:114–117.
30. Pokkuluri PR, et al. (2011) Structure of a novel dodecaheme cytochrome c from Geobacter sulfurreducens reveals an extended 12 nm protein with interacting hemes. *J Struct Biol* 174:223–233.
31. Punnoose A, McConnell LA, Liu W, Mutter AC, Koder RL (2012) Fundamental limits on wavelength, efficiency and yield of the charge separation triad. *PLoS One* 7:e36065, and erratum (2012) 7: 10.1371/annotation/7db1b9ef-c93c-4b02-b721-a8473d2bb4e2.
32. Zheng Z, Gunner MR (2009) Analysis of the electrochemistry of hemes with E(m)s spanning 800 mV. *Proteins* 75:719–734.
33. Radisky D, Kaplan J (1999) Regulation of transition metal transport across the yeast plasma membrane. *J Biol Chem* 274:4481–4484.
34. O'Halloran TV, Culotta VC (2000) Metallochaperones, an intracellular shuttle service for metal ions. *J Biol Chem* 275:25057–25060.
35. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008:P10008.
36. Edwards H, Deane CM (2015) Structural bridges through fold space. *PLOS Comput Biol* 11:e1004466.
37. Williams R (1981) The Bakerian lecture, 1981: Natural selection of the chemical elements. *Proc R Soc Lond B Biol Sci* 213:361–397.
38. Anbar AD, Knoll A (2002) Proterozoic ocean chemistry and evolution: A bioinorganic bridge? *Science* 297:1137–1142.
39. Rison SCG, Thornton JM (2002) Pathway evolution, structurally speaking. *Curr Opin Struct Biol* 12:374–382.
40. Horowitz NH (1945) On the evolution of biochemical syntheses. *Proc Natl Acad Sci USA* 31:153–157.
41. Ycas M (1974) On earlier states of the biochemical system. *J Theor Biol* 44:145–160.
42. Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* 30:409–425.
43. Wang G, Dunbrack RL (2005) PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Res* 33(Suppl 2):W94–W98.
44. Sippl MJ (2008) On distance and similarity in fold space. *Bioinformatics* 24:872–873.
45. Hagberg AA, Schult DA, Swart PJ (2008) *Exploring Network Structure, Dynamics, and Function Using NetworkX* (Los Alamos National Laboratory, Los Alamos, NM).
46. Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. *Proc Int AAAI Conf Weblogs Soc Media* 8:361–362.
47. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504.
48. Bertero MG, et al. (2003) Insights into the respiratory electron transfer pathway from the structure of nitrate reductase A. *Nat Struct Biol* 10:681–687.

Supporting Information

Raanan et al. 10.1073/pnas.1714225115

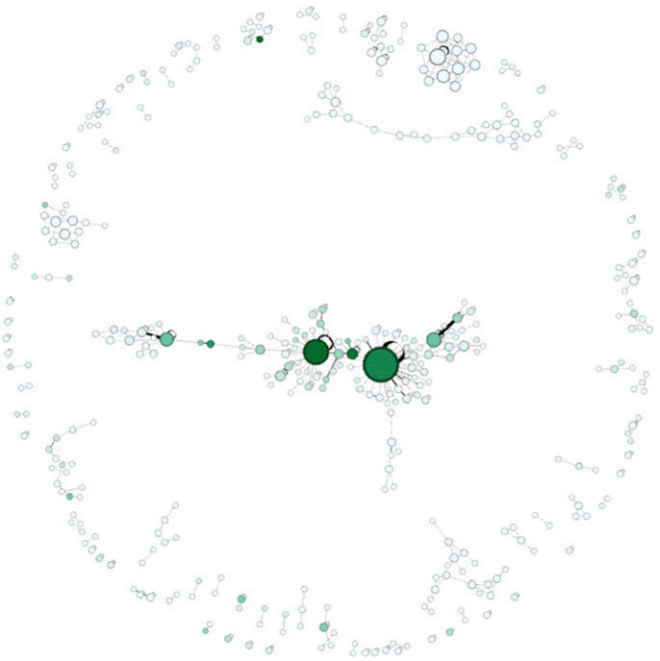

Fig. S1. Complete SPAN for metalloprotein microenvironments. Each node represents one module containing structurally related microenvironments. Edges indicate metal cofactor separations within electron transfer distance between adjacent microenvironments. Node size and intensity of color reflect the degree of the node (the number of connections to structurally adjacent modules). The large connected component in the center consists primarily of modules with microenvironments from oxidoreductase EC1 proteins.

Table S1. Modules of the largest connected component of the SPAN

Module ID	Name	No. of microenvironments	Metals	Cofactors (PDB id)	Degree	Betweenness	EC no.	Metabolic pathways
6	Cyt C	1,908	HEME	HEC, HEM	34	3,823	19	D, OP, AR, STR, AP
7	Bac Fd	817	FeS	F3S, SF4	23	3,634	22	D, M, OP, SR, STR
14	Plastocyanin	1,012	CU	CU, CU1, CUA	11	1,194	11	D, AR, OP
15	Symerythrin	2,362	FE, MN	FE, FE2, FEO, MN, MN3, OFE	10	1,152	11	MO
4	Putidaredoxin	368	FeS	FES	7	231	21	M, MO, OP
46	46	100	HEME	HEM	6	205	4	
155	155	74	HEME	HEA	6	309	2	AR
23	23	882	CU	C2O, CU, CU1	5	205	7	D
31	31	336	HEME, FE	FE, HEM	5	50	2	
48	48	151	FeS	SF3, SF4	5	695	7	M, SR
58	58	76	FeS	F3S, SF4	5	1,596	7	D, SR
242	242	28	FE	FE2	5	353	2	
12	12	143	HEME	HEM	4	0	5	
43	43	47	HEME	HEM	4	137	3	D, AR
71	71	111	HEME	HAS, HEA	4	4	2	AR
130	130	44	HEME	HEC, HEM	4	1	2	D
223	223	82	HEME	HEC, HEM	4	0	2	
224	224	18	HEME	HEC	4	502	1	
1463	1,463	2	HEME	HEM	4	1	2	D
20	20	20	FeS	SF4	3	0	1	
24	Rubredoxin	514	FeS, FE, NI	FE, FE2, FES, NI	3	1,273	17	
37	37	120	CU	CU, CU1	3	1	4	AR
51	51	124	FeS	SF4	3	0	6	SR
68	68	106	HEME	HEC, HEM	3	1	2	D
72	72	135	FE	FE2	3	51	2	
84	84	78	HEME	SRM	3	0	4	STR
133	133	8	HEME	HEM	3	0	2	D
159	159	16	FeS	SF4	3	0	1	
161	161	16	CU	CU	3	0	2	
167	167	148	FeS	F3S, SF4	3	205	6	SR
186	186	18	HEME	HEC	3	205	1	
209	209	16	HEME	HEM	3	0	1	STR
237	237	24	FeS	SF4	3	0	3	SR
313	313	28	FE	FE2	3	51	1	
353	353	8	FE	FE	3	51	1	
498	498	4	CO	CO	3	1	1	
964	964	8	FE	FE	3	51	1	
1,216	1,216	6	HEME	HEC	3	0	2	
1,293	1,293	4	HEME	HEC	3	0	1	
1,611	1,611	2	HEME	HEM	3	1	1	
41	41	204	FE, NI	FCO, FE2, NI	2	51	6	M, SR
77	77	11	HEME	HEC	2	0	2	
111	111	22	FeS	SF4	2	103	3	
191	191	30	FeS	SF4	2	103	1	
222	222	15	FeS	SF4	2	0	1	
241	241	11	FE	FE2	2	1	4	M
354	354	8	HEME	HEM	2	0	2	D
429	429	10	HEME	HEM	2	103	1	D
454	454	15	HEME	HEC, HEM	2	0	1	
516	516	13	HEME	HEC	2	588	1	
569	569	4	CO	CO	2	0	1	
825	825	8	FE	FE	2	0	1	
910	910	39	FE	FE2	2	0	1	
1,049	1,049	2	CU	CU1	2	103	1	
1,084	1,084	3	FE	FE2	2	0	2	
1,225	1,225	2	HEME	HEM	2	0	1	
1,351	1,351	2	FE	FE	2	51	2	M
1,397	1,397	2	HEME	HEC	2	103	2	
56	56	4	CO	NCO	1	0	1	
59	59	133	FeS	FES	1	0	10	M
66	66	27	HEME	HEM	1	0	1	
75	75	101	FeS	SF4	1	0	5	STR

Table S1. Cont.

Module ID	Name	No. of microenvironments	Metals	Cofactors (PDB id)	Degree	Betweenness	EC no.	Metabolic pathways
78	78	208	FE	FE, FE2	1	0	9	
85	85	52	HEME	HEC	1	0	2	D
110	110	15	FeS	F3S	1	0	2	
149	149	38	CU	CUA	1	0	1	
152	152	21	MO, W	MO, MOS, W	1	0	3	D
166	166	30	FeS	SF4	1	0	4	
221	221	4	FeS	FES	1	0	1	
225	225	11	HEME	HEM	1	0	1	
260	260	4	CU	CU	1	0	1	
261	261	9	CU	CU	1	0	2	
266	266	18	HEME	HEC	1	0	1	
298	298	13	HEME	HEC	1	0	1	
365	365	6	FE	FE2	1	0	1	
382	382	9	HEME	HEM	1	0	1	D
384	384	9	FeS	SF4	1	0	1	
393	393	42	HEME	DHE	1	0	3	D
465	465	12	HEME	HEM	1	0	1	
486	486	20	FeS	SF4	1	0	1	
501	501	12	FeS	FES	1	0	1	
549	549	8	HEME	HEC	1	0	1	
585	585	33	HEME	HEC, HEM	1	0	1	
598	598	13	HEME	HEM	1	0	1	STR
678	678	2	HEME	HEC	1	0	1	
699	699	3	HEME	HEC	1	0	1	
703	703	9	CU	CU	1	0	2	
767	767	8	NI	3NI	1	0	2	
803	803	3	HEME	HEC	1	0	1	
809	809	20	W	W	1	0	1	
814	814	2	CU	CU1	1	0	1	
829	829	12	FE	FE	1	0	1	
834	834	8	MO	MO	1	0	1	SR
856	856	16	FeS	SF4	1	0	2	OP
906	906	3	HEME	HEM	1	0	1	
921	921	2	FeS	SF4	1	0	2	OP
1,099	1,099	4	CU	CU	1	0	1	
1,155	1,155	2	HEME	HEM	1	0	1	
1,186	1,186	10	HEME	HEM	1	0	2	D, AR
1,238	1,238	2	FeS	SF4	1	0	1	M
1,247	1,247	2	HEME	HEC	1	0	2	
1,399	1,399	7	NI, FE	FE2, NI	1	0	2	
1,406	1,406	2	CU	CU	1	0	1	
1,515	1,515	2	HEME	HEM	1	0	1	
1,545	1,545	3	HEME	HEC, HEM	1	0	2	

Degree represents the number of connections to other modules. Betweenness represents the importance of a module as a bridge between different parts in the network. The EC no. column lists the number of different EC annotations that each module is associated with. Module ID is the identifier used in Figs. 4 and 7. The metabolic pathways that each module is involved in are listed as follows: AP, anoxygenic photosynthesis (3.8–3.4 Ga); AR, aerobic respiration (2.72–2.45 Ga); D, denitrification (2.7–2.5 Ga); M, methanogenesis (3.8–3.45 Ga); MO, methane oxidation (2.9–2.7 Ga); NF, nitrogen fixation (3.2–2.9 Ga); OP, oxygenic photosynthesis (3.0–2.5 Ga); SDO, sulfide oxidation; SO, sulfur oxidation; SR, sulfur reduction (3.8–3.45 Ga); STR, sulfate reduction (3.8–3.45 Ga).

Other Supporting Information Files

Dataset S1 (XLSX)