Discovering the electronic circuit diagram of life: structural relationships among transition metal binding sites in oxidoreductases

J. Dongun Kim1,2, Stefan Senn3, Arye Harel3, Benjamin I. Jelen3,4 and Paul G. Falkowski1,2,3,4

1Department of Chemistry and Chemical Biology, and 2Department of Earth and Planetary Sciences, Rutgers University, Piscataway, NJ 08854, USA
3Environmental Biophysics and Molecular Ecology Program, Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA
4Department of Environmental Sciences, Rutgers University, New Brunswick, NJ 08901, USA

Oxidoreductases play a central role in catalysing enzymatic electron-transfer reactions across the tree of life. To first order, the equilibrium thermodynamic properties of these proteins are governed by protein folds associated with specific transition metals and ligands at the active site. A global analysis of holoenzyme structures and functions suggests that there are fewer than approximately 500 fundamental oxidoreductases, which can be further clustered into 35 unique groups. These catalysts evolved in prokaryotes early in the Earth’s history and are largely responsible for the emergence of non-equilibrium biogeochemical cycles on the planet’s surface. Although the evolutionary history of the amino acid sequences in the oxidoreductases is very difficult to reconstruct due to gene duplication and horizontal gene transfer, the evolution of the folds in the catalytic sites can potentially be used to infer the history of these enzymes. Using a novel, yet simple analysis of the secondary structures associated with the ligands in oxidoreductases, we developed a structural phylogeny of these enzymes. The results of this ‘composome’ analysis suggest an early split from a basal set of a small group of proteins dominated by loop structures into two families of oxidoreductases, one dominated by α-helices and the second by β-sheets. The structural evolutionary patterns in both clades trace redox gradients and increased hydrogen bond energy in the active sites. The overall pattern suggests that the evolution of the oxidoreductases led to decreased entropy in the transition metal folds over approximately 2.5 billion years, allowing the enzymes to use increasingly oxidized substrates with high specificity.

1. Introduction

Biologically driven electron-transfer reactions are the primary energy-transduction processes across the tree of life. These reactions depend upon energy sources external to the system. The two external energy sources on the Earth are solar radiation and geothermally derived heat. These create chemical redox gradients, which are coupled to biological redox reactions and ultimately drive the non-equilibrium thermodynamic reactions that make life possible. Indeed, the origin of life almost certainly began with the evolution of a small set of metabolic processes coupled to redox chemistry.

Oxidoreductases (enzyme commission 1, EC1) are a class of enzymes that facilitate these proton-coupled electron-transfer reactions. All of the core metabolic processes mediated by EC1 proteins evolved in prokaryotes and ultimately became coupled on local and planetary scales to facilitate an electron ‘market’ between the major light elements. This electron market ultimately led to a closed cycle between respiratory reactions and their biological oxidative analogues, and photosynthesis and their biological reductive analogues. These reactions, which
are far from thermodynamic equilibrium, allow gas exchanges across the tree of life and transformed the Earth’s biogeochemical cycles (figure 1a) [1].

Although the exact number of core EC1 proteins, including orthologues, paralogues and analogues is unknown, their functions appear to be encoded by fewer than 500 unique genes (i.e. genes that encode for unique functions including paralogues and analogues; figure 1a). One reason that such a small set of proteins plays such a large role in the Earth’s elemental cycles is that the different electron-transport chains found ubiquitously throughout life share common components. For example, chemosynthetic, anaerobic and aerobic respiration, as well as anaerobic and oxygenic photosynthesis all use similar proton-coupled electron-transport schemes. The basis of the scheme is separation of protons from electrons across a membrane. The electrons are inevitably ferried via carriers, across a set of membrane-bound proteins, ultimately arriving at a transient sink that allows a negatively charged carrier to be neutralized by a proton. The protons are initially segregated from the electrons by the membrane, thereby forming an asymmetric distribution of charge. The return flow of the protons (i.e. the proton motive force) is coupled to nanomachines, especially the ‘coupling factor,’ ATP synthase, which conserves the electrochemical energy as chemical bond energy.

Although peripheral components of these proton-coupled electron-transport reactions have been selected for specific reaction substrates and products, the basic architecture of all the core pathways shares similar protein structures and ligands, including iron–sulfur clusters, pterins, haems and quinones. These interchangeable structures and ligands have evolved into a metabolic network with overlapping functions across the tree of life (figure 1b). Additionally, many biological energy-transduction systems share a small subset of metabolic pathways such as glycolysis (the Embden–Meyerhof, the Entner–Doudoroff, including Archaean modifications of these pathways) or the reverse TCA cycle [2]. Thus, metabolism, using a variety of electron donors and acceptors, draws catalysts from a core set of similar components and pathways to enable a flow of electrons and protons. This modular approach to metabolism has provided great flexibility on a relatively small number of EC1 genes. Indeed, prokaryotes are often able to regulate major components of metabolism in accordance with environmental conditions [3], often at suboptimal efficiency.

Regardless of efficiency, the flux of electrons through the metabolic network is particularly dependent on, and sensitive to the availability of specific transition metals, especially iron (figure 2). The bioavailability of transition metals is, in turn, highly dependent on the redox state of the environment. A recent whole-genome analysis of phylogenetically diverse micro-organisms suggests that the earliest proteins incorporated metals, and that metal usage over time evolved in accordance with environmental availability [4]. The metals are invariably coordinated to the protein scaffolds via a small set of specific protein folds [1]. Identifying members and the evolutionary pattern of this set of folds is critical to understanding the evolution of metabolism across the tree of life, as well as the emergence of biogeochemical cycles, far from equilibrium.

In this paper, we present an analysis of the evolutionary history of metal usage, the structures of the protein folds and redox state of the oxidoreductases across the tree of life, which ultimately formed an electronic circuit on a planetary scale. Our results suggest that the redox processes connecting metabolism across the Earth’s surface underwent a secular trend in evolutionary transitions that led to successively greater complexity and thermodynamic efficiency in these critical enzymes over the first approximately 2.5 billion years of the Earth’s history.

2. Transition metals in oxidoreductases

Oxidoreductases catalyse electron-transfer reactions via prosthetic groups that usually contain transition metals whose ions have incompletely filled d or f orbitals and which can accept a electrons from protein side-chains [5,6]. Multiple oxidation states, especially of molybdenum and manganese, allow transition metals to be coupled with a wide range of the reduction–oxidation reactions. Further, a specific metal–ligand structure is ‘tuned’, or poised for specific redox reactions by the protein–metal microenvironment [7].

Among transition metals, specific elements were naturally selected for their physical–chemical properties, abundance, coordination bond strength, atomic radii, solubility and polarizability [8]. Iron is, by far, the most abundant transition metal on the Earth [9]. The abundance of this element, especially as a ferrous ion in the Archaean and early Protzozoic oceans, is reflected in its wide use as a catalysing cofactor for oxidoreductases (figure 2). In Swiss–Prot [10] and the protein data bank (PDB) [11], iron is identified as a catalytic component in more than 67 per cent of the EC1 proteins. Iron-containing oxidoreductases can use the metal in a metal enzyme of a sulfide, or coordinated to imidazole nitrogens in porphyrins, forming haems. Following iron, oxidoreductases contain metals in the following order of relative abundance: Cu > Mn > Ni > Mo > Co > V > W (figure 2). It should be noted that under anoxic and/or euxenic conditions, Cu and Mo are highly insoluble unless the ions are oxidized.

3. Biopolymer–metal interactions in the ‘ancient ocean’

Assuming that the conversion of energy via dissipation of redox gradients was an early bioinorganic reaction essential for the origin of life, it logically follows that transition metals played a key role. Transition metals can undergo stoichiometric reactions via photochemical processes or in solution phase with other redox couples [12–15]. Transition metals are also capable of carrying out catalytic reactions when surrounded by a biopolymer that provides a specific structural framework that allows reversible population of the metal–ligands with electrons [16].

Lewis basic peptide side-chains, such as thioclates (cysteine), imidazole nitrogens (e.g. histidines) and carboxylates (aspartic and glutamic acids), can form coordination bonds via d-orbitals in the transition metals. Metals bound by multiple side-chains are locked within the peptide/protein matrix. These interactions influence multiple physical properties of the holoprotein, including solvent accessibility, tuned redox potential, optimization of Gibbs free energy and enhanced substrate specificity. Understanding how the earliest biopolymer–metal interactions evolved is critical to understanding the origins of non-equilibrium bioenergetic reactions, and hence the origins of life.
Figure 1. Network of life’s biologically mediated cycles modified from [1], and a Venn diagram of the basic metabolic pathways. (a) The interconnected ‘circuit board’ of life’s electron-transfer reactions (for C, H, O, N, S and Fe) is labelled with pathway groups I to VI, corresponding to metabolic networks. (b) The metabolic pathway groups I to VI (large font) are shown to have redox components of electron-transport chains as well as auxiliary pathways in common (small font). Pathway group I: ‘aerobic sugar metabolism’ has components such as NAD(P)H reductases, as well as membrane-localized cytochromes and quinones, in common with pathway group VI: ‘oxygenic photosynthesis’. Pathway group II: ‘sulfate reduction’ utilizes oxidized sulfur compounds as terminal electron acceptors and carry out carbon metabolism via reductive acetyl CoA pathway or the TCA cycle. Pathway group III: ‘denitrification’ uses unique oxidoreductases to reduce nitrates and other oxidized nitrogen compounds but still requires the common components of the electron-transport chain as well as glycolysis and the tricarboxylic acid (TCA) cycle as a source of reductant. Pathway group IV: ‘hydrogen oxidation’ uses an electron-transport chain but is not dependent on glycolysis or the TCA cycle for producing reductant. Pathway group V: ‘methanogenesis’ does not use an electron-transport chain and is the only pathway using coenzyme F420. Size of circles is irrelevant.
modes of evolution such as gene insertion, duplication
subsequent structural innovations were accelerated by various
transfer, but ultimately diverged into several motifs. The
across the nascent tree of life primarily via horizontal gene
chemical cycles. The ancient folds were spread genetically
geological time with an increasing larger set of coupled biogeo-
tic functions. The specific folds almost certainly coevolved over
descent, and was inherited with modifications through a
myriad of organisms to form the observed protein fold. We
further hypothesize that these secondary structures are the out-
come of selection and provide a window into the processes in
which protein folds evolved. We assume that the composition
of the secondary structural motifs reflects the evolutionary his-
tory of the protoenzyme from which the extant motif is
descended, and was inherited with modifications through a
myriad of organisms to form the observed protein fold. We
further hypothesize that these secondary structures are the out-
come of selection and provide a window into the processes in
which protein folds evolved. We assume that the composition
of the secondary structural motifs reflects the evolutionary his-
tory of the protoenzyme from which the extant motif is
descended, and was inherited with modifications through a
myriad of organisms to form the observed protein fold. We
further hypothesize that these secondary structures are the out-
come of selection and provide a window into the processes in
which protein folds evolved. We assume that the composition
of the secondary structural motifs reflects the evolutionary his-
tory of the protoenzyme from which the extant motif is
descended, and was inherited with modifications through a
myriad of organisms to form the observed protein fold. We
further hypothesize that these secondary structures are the out-
come of selection and provide a window into the processes in
which protein folds evolved. We assume that the composition
of the secondary structural motifs reflects the evolutionary his-
tory of the protoenzyme from which the extant motif is
descended, and was inherited with modifications through a
myriad of organisms to form the observed protein fold. We
further hypothesize that these secondary structures are the out-
come of selection and provide a window into the processes in
which protein folds evolved. We assume that the composition
of the secondary structural motifs reflects the evolutionary his-
tory of the protoenzyme from which the extant motif is
descended, and was inherited with modifications through a
myriad of organisms to form the observed protein fold. We
further hypothesize that these secondary structures are the out-
come of selection and provide a window into the processes in
which protein folds evolved. We assume that the composition
of the secondary structural motifs reflects the evolutionary his-
4. Evolution of sequences and folds
Because of both their modularity and early spread across the tree
of life, it is extremely difficult to determine the evolutionary heri-
tage of folds in the oxidoreductases solely based on analysis of
organisms, sequences or synteny within highly conserved oper-
ons. While the core oxidoreductases in oxygenic photosynthesis
are extremely highly conserved, inspection reveals major sequence degeneracy in closely related structures. For example,
transmembrane-spanning helices of photosystem I and II have
highly divergent sequences, yet their structures are almost
identical [21]. This basic phenomenon was noted early on by pio-
ners in the field of bioinformatics. Indeed, Eck & Dayhoff [22, p.
363], noted ‘the processes of natural selection severely inhibit
cchange to a well-adapted system on which several other essential
components depend’. While their comments were based on the
highly conserved structure of ferredoxin, they apply to many
ancient proteins, including enzymes that do not catalyse redox
reactions. For example, ribulose-1,5-bisphosphate carboxylase
oxygenase (EC 4.1.1.39) is a carboxylase. The enzyme is respon-
sible for the fixation of CO2 in many photosynthetic and
chemoautotrophic organisms. This crucial enzyme cannot
easily distinguish between its ‘true’ substrate, CO2 and O2. The
result is that at present atmospheric levels of O2, the enzyme is
often remarkably inefficient. Moreover, the catalytic turnover of
the reaction, even under optimal conditions, is much slower
than reactions feeding the substrate (CO2) or removing the pro-
duct (3-phosphoglycerate). Regardless, the catalytic site of the
enzyme is highly conserved and the biological result of this con-
servation is that organisms often synthesize the enzyme in
excess to achieve maximum overall growth efficiency [23].
There are many other, similar examples.
The fundamental physico-chemical properties that govern
the major protein fold conformations have remained unchan-
ged. Reinvention of metalloenzyme folds is highly restricted,
given that geometrical and energetic selection processes
limit structural solutions. Let us examine a novel approach to
identifying and ordering the structural solutions found in
extant oxidoreductases.

5. The ‘composome’ approach
We hypothesize that the ensemble of secondary structures in
the region surrounding the catalytically active metals has
been selected to facilitate catalysis of the holoenzyme. We
further hypothesize that these secondary structures are the out-
come of selection and provide a window into the processes in
which protein folds evolved. We assume that the composition
of the secondary structural motifs reflects the evolutionary his-
tory of the protoenzyme from which the extant motif is
descended, and was inherited with modifications through a
myriad of organisms to form the observed protein fold. We
further assume that the folds must obey the rules set by the
d-block metal coordination chemistry [5]. These underlying
hypotheses are extensions from our previous work where we
proposed that secondary structures around the metal or
metal–ligand in the active site would be more conserved
than elsewhere in the protein [24]. We call this quantitative
analysis of the secondary structure of the folds in active sites

(1) The electronic structure of transition metals must match
geometrical requirements for metal–ligand coordination
number and geometry. As a result, the emergent entatic
states constrain the subsequent evolution of the structures,
and topologies of the coordinately bound polypeptides or
other molecules required for catalysis.
(2) The mildly reducing environment of the ancient oceans
required a relatively low midpoint potential of the early
redox organocatalysts, especially compared with
more recently evolved oxidoreductases. The ligand
environment was modulated to tune the midpoint
potential to meet the functional requirements.

To meet these constraints, random polypeptides almost cer-
tainly evolved in association with the available metals and
continuously selected for sequences and folds that satisfied
both structural and functional constraints. Such a combinatori-
sal search to local minima would have persisted until the
polypeptide–metal complex acquired locally optimized catalytic
functions. The specific folds almost certainly coevolved over
geological time with an increasing larger set of coupled biogeo-
chemical cycles. The ancient folds were spread genetically
across the nascent tree of life primarily via horizontal gene
transfer, but ultimately diverged into several motifs. The
subsequent structural innovations were accelerated by various
modes of evolution such as gene insertion, duplication
and partial loss. Evolved core protein folds became
molecular ‘modules’ from which a variety of biomachines
could ultimately be built via a ‘mix and match’ set of motifs.
the elucidation of a ‘composome’. To the best of our knowledge, this is the first attempt to infer quantitative distances between distinctly different protein folds based solely on secondary structural composition. The resulting ‘phylogeny’ represents a linkage of fold relationships in structural space, and obviously is not intended to imply linear, monophyletic history of the evolution of EC1 proteins.

The approach uses PDB files that possess previously determined ‘gold standard’ domains [25]. From this dataset, we extracted a subset of representatives using the best resolution structure for each organism per ‘gold standard’ domain (see the electronic supplementary material). We included all structures of orthologous proteins with high resolution for every gold standard domain. For every domain, the corresponding metal–ligand was treated as the catalytic site.

For each catalytic site, we first collected a list of amino acid residues that are within 15 Å from a catalytic metal, based on carbon alpha (Cα) coordinates. Secondary structures of the residues were assigned using the Define Secondary Structure (DSSP) database [26]. The overall secondary structure composition (i.e. the ‘composome’) around each metal-bearing catalytic site was determined and further adjusted based on a residue-metal distance by a factor of 1/r². These secondary structural compositions, from each catalytic site, were plotted in a ternary vector space (figure 3). Structures that contained both identical metals and a Euclidean distance in composome space (ternary vector space (figure 3). Structures that contained both identical metals and a Euclidean distance in composome space (ternary vector space (figure 3)). Structures that contained both identical metals and a Euclidean distance in composome space (ternary vector space (figure 3)). Structures that contained both identical metals and a Euclidean distance in composome space (ternary vector space) were collapsed into a single representative structure for each organism per ‘gold standard’ domain (see the electronic supplementary material). We included all structures of orthologous proteins with high resolution for every gold standard domain. For every domain, the corresponding metal–ligand was treated as the catalytic site.

For each catalytic site, we first collected a list of amino acid residues that are within 15 Å from a catalytic metal, based on carbon alpha (Cα) coordinates. Secondary structures of the residues were assigned using the Define Secondary Structure (DSSP) database [26]. The overall secondary structure composition (i.e. the ‘composome’) around each metal-bearing catalytic site was determined and further adjusted based on a residue-metal distance by a factor of 1/r². These secondary structural compositions, from each catalytic site, were plotted in a ternary vector space (figure 3). Structures that contained both identical metals and a Euclidean distance in composome space (ternary vector space (figure 3)). Structures that contained both identical metals and a Euclidean distance in composome space (ternary vector space (figure 3)). Structures that contained both identical metals and a Euclidean distance in composome space (ternary vector space) were collapsed into a single representative structure for each organism per ‘gold standard’ domain (see the electronic supplementary material). We included all structures of orthologous proteins with high resolution for every gold standard domain. For every domain, the corresponding metal–ligand was treated as the catalytic site.

The sequences of the metal-binding sites were aligned using the ClustalX package [26] and compared in a dot plot of 3240 alignments of 82 metal catalytic sites. The 447 alignments between the same metal cofactors are coloured red. Alignments that are less than 40 per cent structurally similar (grey-shaded area) can be considered inconclusive in terms of structural similarity. The inset shows an alignment between the manganese-binding motif in oxalate oxidase (2et1, EC 1.2.3.4; orange for matched, blue for unmatched) and the iron-binding motif of clavaminate synthase (1ds1, EC 1.14.11.21; red for matched, green for unmatched). The folds are very similar in secondary structure content and overall fold (52.64% structurally similar, r.m.s.d. 1.86 Å). The sequence identity derived from the 50 residues’ alignment is 6%, or three residues, two of which are the metal-coordinating histidines (illustrated in stick representation). This analysis strongly suggests that the ‘composome’ approach based solely on secondary structures can differentiate between closely related folds with a high degree of accuracy.

Figure 4. Comparison of the composome approach with a tertiary structural analysis. The structural similarity and composome distance (described in text) are compared in a dot plot of 3240 alignments of 82 metal catalytic sites. The 447 alignments between the same metal cofactors are coloured red. Alignments that are less than 40 per cent structurally similar (grey-shaded area) can be considered inconclusive in terms of structural similarity. The inset shows an alignment between the manganese-binding motif in oxalate oxidase (2et1, EC 1.2.3.4; orange for matched, blue for unmatched) and the iron-binding motif of clavaminate synthase (1ds1, EC 1.14.11.21; red for matched, green for unmatched). The folds are very similar in secondary structure content and overall fold (52.64% structurally similar, r.m.s.d. 1.86 Å). The sequence identity derived from the 50 residues’ alignment is 6%, or three residues, two of which are the metal-coordinating histidines (illustrated in stick representation). This analysis strongly suggests that the ‘composome’ approach based solely on secondary structures can differentiate between closely related folds with a high degree of accuracy.

Figure 3. Ternary vector space used for the secondary structure composition (i.e. the ‘composome’) and representative structures. Representative structures are shown for each of the major folds in the ternary space.

The compositional finger printing method described for calculating relations in folds across all known EC1 structures collapses three-dimensional information into a one-dimensional matrix. To assess the effects of this mathematical simplification and compression, the Cα backbone environments for the 82 structurally different catalytic sites were superimposed in a pairwise, all versus all fashion. We used a structural alignment protocol that calculates a novel similarity score, which combines alignment length and spatial deviation into one measurement [27]. The similarity values of highly similar structure pairs (more than 40% structurally similar) correlate with low compositional profile distances (figure 4). This result strongly
suggests that, in principle, secondary structures retain sufficient topological information such that a quantitative analysis of the fold can retrieve conformational relationships with sufficient resolution and confidence to derive the structural history of the folds in catalytic sites.

Indeed, for dissimilar structural pairs, secondary structure composition maintains a high degree of predictive performance. We find significantly lower secondary structure composition distances for environments of same ligands compared with different ligands (figure 4), especially for pairs where the calculated Ca backbone structure similarity is less than 40 per cent and therefore inconclusive. Very high composome distances are generally not observed for same ligand pairs, suggesting a limit of the differences between same ligand environment secondary structures. This phenomenon appears to be related to hydrogen bonding patterns that are required by certain electronic structures around metal or metal–ligand catalytic sites, leading to similar secondary structure composition signatures.

We kept orthologous structures with detectable sequence similarity in our dataset for two reasons: (i) to show the structural variability of a motif and get the full spectrum of secondary structure composition and (ii) to highlight the sensitivity of secondary structure composition analysis. This redundancy also includes phylogenetic information and represents a positive control for our analysis. One example is the molybdenum cluster (figure 3) where related structure domains (SCOP family: molybdenum cofactor-binding domain) are sampled over a wide phylogenetic space as well as different functions according to EC classifications.

Let us now examine the fold ‘phylogeny’ derived from the composome analysis.

### 6. Fold phylogeny based on composome

A polypeptide chain has an intrinsic property of having dynamic conformational variability, but specific structures with relatively rigid folds are often conferred for specific biological functions. Fold evolution is achieved when a protein structure with a functional promiscuity has a flexible chain, which is ‘evolvable’ [28].

To check the evolvability of each protein fold, we calculated the average hydrogen bond energy per residue across the composome surface. Hydrogen bond energy is a proxy for fold ‘evolvability’. The evolvability contour plot is generated based on average hydrogen bond energy per residue from 82 data points and overlayed on the ternary plot (figure 3). As protein folds acquire more helix or sheet secondary structures, they become less flexible and their folds are increasingly fixed. The evolution of fold occurred starting from simple disordered state (red) to complex organized state (blue).

Figure 5. Average hydrogen bond energy per amino acid residue associated with the metal/metal–ligand centre across the composome surface. Hydrogen bond energy is a proxy for fold ‘evolvability’. The evolvability contour plot is generated based on average hydrogen bond energy per residue from 82 data points and overlayed on the ternary plot (figure 3). As protein folds acquire more helix or sheet secondary structures, they become less flexible and their folds are increasingly fixed. The evolution of fold occurred starting from simple disordered state (red) to complex organized state (blue).
While evolution can obviously still occur in these folds, the resulting structures are highly conserved. Mixed usage of both helices and sheets requires adjoining loop regions. The result is a variable degree of hydrogen bonding across the peptide backbone. Protein folds with both helices and sheets are expected to be relatively highly evolvable. In our analysis, we observe two ‘evolvability hot spots’ (figure 5). One such area is occupied by a ferredoxin fold, which might be one of the ancient extant oxidoreductase structures [22], whereas other obvious hotspots of origins do not exist among folds in our dataset (see the electronic supplementary material).

Based on the evolvability, we hypothesize a parsimonious condition (analogous to a Bayesian prior) for fold evolution: that is, folds evolved from conditions with fewer to more hydrogen bond energies. Hence, loop- and coil-rich protein structures, lacking secondary structure, would be located close to the root of the phylogenetic tree of electron-transfer folds. Using the secondary structure compositional vectors for each protein fold, a matrix of Euclidian distances was generated, and a phylogenetic tree of protein folds (figure 6) was calculated using the Fitch–Margoliash algorithm and global tree optimization, as provided by the PHYLIP package [30]. The Rieske fold was chosen as a root for building a monophyletic tree; however, it is impossible to prove that the actual evolution of protein folds occurred in a monophyletic fashion, and it most likely did not. For example, the higher midpoint potential of Rieske proteins suggests that these structures evolved after ferredoxin. Clearly, simple folds could have been recruited from other functions to become a catalytic part of oxidoreductases. However, the basic

Figure 6. A tree of protein folds based on evolvability of metal binding sites (average hydrogen bond energy per residue, figure 5) and a Euclidean distance matrix based on the composome analysis (i.e. figure 3). Full names of proteins that include metal binding folds are provided and the metal types are labelled with different colours (light blue: FeS2 cluster; purple: molybdopterin; pink: Fe4S4 cluster; red: iron; yellow: haem; cyan: copper; green: molybdenum–iron–sulfur). A complete list of examined protein structures is provided in the electronic supplementary data. During early evolution, polypeptide chains gained catalytic ability through random mutations. (a) The earliest protein folds would have resembled extant Rieske or molybdopterin folds, which have high evolvability. (b) Conformational searches yields folds with both α-helix and β-sheet structures, such as ferredoxins. (c) Further conformational searches result in the divergence of protein structure, giving rise to α-helix-rich and (d) β-sheet folds.
topology of the fold tree is potentially informative about the evolutionary history of electron-transfer reactions.

7. Iron–sulfur proteins

Iron–sulfur-containing ferredoxin folds are located near the centre of the ternary plot, indicating relatively equal amounts of helix, sheet and loop composition. The location of iron–sulfur proteins on a ternary plot coincides with a high evolvability hotspot, suggesting ferredoxin folds might be the common ancestor of many extant protein folds for a number of reasons.

First, iron–sulfur minerals were thought to be relatively abundant in the early Archaean ocean and it has been speculated that iron–sulfur clusters played an important role in the evolution of bioenergetic redox transduction systems [14]. These hypotheses postulate that the earliest biologically relevant redox reactions occurred on iron–sulfur mineral deposits associated with hydrothermal vents. Second, ferredoxin is found across the tree of life, and is found alone or as a domain in larger proteins, many of which are encoded by the core redox genes of life. Third, sequence and structural symmetry of ferredoxins suggests that they may have evolved from a gene duplication event of a 28–30 amino acid sequence, each capable of binding one iron–sulfur cluster. Sequence analyses by Eck–Dayhoff revealed even shorter repeats of four amino acids, suggesting a prebiotic ‘protoferredoxin’ that was potentially composed of a primaeval subset of the 20 amino acids [22]. Fourth, all ferredoxins have a simple, conserved fold that binds two Fe₅S₄ clusters and is composed of thirty to sixty amino acids.

8. Molybdopterin and Rieske proteins

The lack of a helix/sheet is characteristic of Rieske iron–sulfur–containing proteins and molybdopterin proteins. Although pterins may have been formed very early in the Earth’s history (as they are derived from GTP), it is unclear how these proteins carry out specific biological functions without a fixed helix or sheet secondary structure. Regardless, our composome analysis suggests that these folds have the highest potential to evolve (figure 5). Unlike nitrogenases, molybdopterin proteins, such as dimethyl sulfoxide (DMSO) reductase or nitrate reductase, have a molybdenum atom chelated by pterins, which is surrounded by a protein environment that lack either α-helix or β-sheet. In the absence of Mo, tungsten can serve as a replacement owing to its similar chemical properties, and it is possible that tungsten-containing proteins gave rise to molybdenum-containing protein folds in a prebiotic world [31].

9. Molybdenum–iron–sulfur proteins

Reduction of N₂ to NH₃ is catalysed by nitrogenases, the modern forms of which contain molybdenum–iron–sulfur clusters. Molybdenum-containing nitrogenase folds are closely located near ferredoxin folds in a ternary composome space, indicating their similarity in the secondary structure composition with mixed α-helix and β-sheet. Our composome analysis suggests that the Mo–Fe–S cluster-containing folds may have evolved from Fe–S folds. Both iron–sulfur folds and molybdenum–iron–sulfur folds have α-helix and β-sheet elements with a significant loop content, adding flexibility to the core structure. As a result, these folds may have diversified by exploring different conformations and compositions, mainly diverging into two distinct clades: an α-helix-rich clade and a β-sheet-rich clade.

10. Rubredoxin, Mn and Cu proteins

The Fe atom is found alone or as a ligand complex with a porphyrin ring in most extant folds. Rubredoxin is a single iron-containing fold with high β-sheet content. Manganese and copper folds also appear at the far edges of the composome space in the β-sheet clade. These folds form extensive hydrogen bond networks across the backbone amide and carbonyl, making the fold less evolvable. In return, these folds gained high catalytic specificity and accuracy with rigid structure, but sacrificed evolvability.

11. Haem proteins

Haem-containing proteins are a hallmark of many electron-transfer reactions. Unlike rubredoxins, haem folds have an iron atom residing in a porphyrin ring surrounded by α-helices. The iron atom has octahedral coordination geometry, where axial positions are ligated to histidine side-chains. A surrounding porphyrin ring allows iron atoms to be incorporated into a protein scaffold with just one amino acid residue binding the iron directly. The usage of porphyrin rings may have relieved the stringent geometric requirement of the iron coordination (i.e. rubredoxin), allowing an explosive diversification of haem folds. Haems are among the most abundant protein cofactors in oxidoreductases (figure 2) and widely covered in composome ternary space (figure 3).

12. Concluding remarks

Life is dependent on the catalytic function of a relatively small set of core proteins. Within this core set, oxidoreductases play an outsized role, but their evolutionary history is poorly understood. Our brief analysis conforms to the hypothesis that the evolution of the oxidoreductases is, like for all proteins, a trade-off between enzyme specificity and evolvability. Assuming that the evolutionary trajectory followed from a disordered state to an ordered state, proteins appear to have evolved from loop-rich into either α-helix-rich or β-sheet-rich structures, becoming more specific but less evolvable. The early split between the two major clades of oxidoreductases continued to follow a decrease in the free energy within the active site—a long-term displacement of internal entropy that coincides with an increased contact order of the core structure. The selection pressure that led to the increased hydrogen bonding energy concordant with increased redox potential in oxidoreductases may be an outcome of evolution, or a result of a feedback between the initial ignition of a non-equilibrium thermodynamic system that is self-sustaining with a positive feedback. Regardless, the redox space explored by biology over the past two billion years appears to be extremely limited.

We conclude that life has evolved a very small set of key core catalysts, the genes of which are transmitted across vast expanses of geological time by microbes, allowing a fundamental set of electron-transfer reactions to permit energy extraction from an open system. The microbes themselves are temporary, disposable carriers of the core genes. They go extinct but
transfer the functions onwards. How these catalysts came to be coupled in specific sequences and with other chemical reactions to obviate non-equilibrium systems remains a major challenge to understanding the origins and continuity of life on the Earth.

This research is funded by the Gordon and Betty Moore Foundation through Grant GBMF2807 to Paul Falkowski. We thank Doron Lancet for suggesting the term ‘composome’. We thank Yana Bromberg, Vikas Nanda, David A. Case and two anonymous reviewers for constructive comments.

References


