Structure
Article

Protein Domain Structure Uncovers the Origin of Aerobic Metabolism and the Rise of Planetary Oxygen

Kyung Mo Kim,[1,2,5] Tao Qin,[3,4,5] Ying-Ying Jiang,[4,5,6] Ling-Ling Chen,[3] Min Xiong,[3] Derek Caetano-Anollés,[1] Hong-Yu Zhang,[3,*] and Gustavo Caetano-Anollés[1,*]
[1]Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA
[2]Korean Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology, 111 Gwahangno, Yuseong-gu, Daejeon 305-806, Korea
[3]National Key Laboratory of Crop Genetic Improvement, Center for Bioinformatics, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, People's Republic of China
[4]School of Life Sciences, Shandong University of Technology, Zibo 255049, People's Republic of China
[5]These authors contributed equally to this work
[6]Present address: Biochemical Engineering Institute, Saarland University, Campus A 1.5, 66123 Saarbrücken, Germany
*Correspondence: zhy630@mail.hzau.edu.cn (H.-Y.Z.), gca@illinois.edu (G.C.-A.)
DOI 10.1016/j.str.2011.11.003

SUMMARY

The origin and evolution of modern biochemistry remain a mystery despite advances in evolutionary bioinformatics. Here, we use a structural census in nearly 1,000 genomes and a molecular clock of folds to define a timeline of appearance of protein families linked to single-domain enzymes. The timeline sorts out enzymatic recruitment, validates patterns in metabolic history, and reveals that the most ancient reaction of aerobic metabolism involved the synthesis of pyridoxal 5′-phosphate or pyridoxal and appeared 2.9 Gyr ago. The oxygen source for this primordial reaction was probably Mn catalase, which appeared at the same time and could have generated oxygen as a side product of hydrogen peroxide detoxification. Finally, evolutionary analysis of transferred groups and metabolite fragments revealed that oxidized sulfur did not participate in metabolism until the rise of oxygen. The evolutionary patterns we uncover in molecules and chemistries provide strong support for the coevolution of biochemistry and geochemistry.

INTRODUCTION

The chemical and energetic workings of life are dictated by thousands of biochemical reactions that are catalyzed by enzymes. The evolutionary history of these reactions, however, eludes our understanding, fundamentally because the conventional tools of molecular biology and paleontology have been used almost exclusively to dissect the history of species, and not the evolution of biochemical repertoires. Because protein functions are tightly coupled with protein structures and structures are much more conserved than sequences (Caetano-Anollés et al., 2009a), structures can be used as molecular fossils to study the early evolution of modern biochemistry (Caetano-Anollés and Caetano-Anollés, 2003; Caetano-Anollés et al., 2007, 2009a, 2009b, 2012, 2011; Ma et al., 2008).

Protein domain structures are organized in the Structural Classification of Proteins (SCOP) into a hierarchy of levels of structural abstraction (Murzin et al., 1995). The fold family (FF) level describes domains with similar pairwise amino acid residue identities that are closely related evolutionarily. The fold superfamily (FSF) level pools domains with similar structural and functional features that are most likely unified by common ancestry. The fold (F) level defines domains that have common fold architectural designs, with similarities portraying the physics and chemistry of folding rather than an ancestral relationship. We calculated the evolutionary age of protein domain structures at FF (Caetano-Anollés et al., 2012, 2011), FSF (Wang et al., 2007), and F (Caetano-Anollés and Caetano-Anollés, 2003) levels from intrinsically rooted phylogenies reconstructed from a census of domain structures in hundreds of genomes that have been completely sequenced. The ages of domains are characterized by node distances (*nd*) along branches of the highly unbalanced phylogenetic trees (from root to leaf) on a relative zero to one scale and are used to build timelines of domain discovery. Because the *nd* values of F and FSF domains correlate well with their geological ages, the chronology of domains at these levels can be used as molecular clocks to trace critical evolutionary events, such as the birth of aerobic metabolism (Wang et al., 2011). Besides, domain age information can provide answers to other important questions. For example we revealed that metabolic networks evolve mostly by enzyme recruitment, and not sequentially in pathways (Caetano-Anollés et al., 2007; Kim et al., 2006). When applied to sets of proteins, this methodology has the advantage of directly inferring the time of origin of individual biochemical reactions, inferring the origin of structures linked to important pathways and geochemical events, and avoiding complications from contamination in biological and geological samples in paleobiological studies.

The rise of atmospheric oxygen is considered one if not the most critical event in planetary history (Campbell and Allen, 2008; Canfield, 2005; Sessions et al., 2009). Oxygen is also believed to be a key player in biological evolution. Some crucial evolutionary steps, such as the birth of eukaryotes, the Cambrian explosion of animal diversity, and the increase of body size, have been correlated with oxygen elevation (Dahl et al., 2010; Falkowski and Isozaki, 2008; Payne et al., 2009). The benefit of oxygen to the evolutionary progression is primarily attributed to aerobic respiration, which is about 16 times more efficient than its anaerobic counterpart in generating ATP (Catling et al., 2005). Other more sophisticated explanations resort to biochemical innovations induced by aerobic metabolism. For instance simulations of metabolic networks under anaerobic or aerobic conditions revealed that molecular oxygen enables thousands of metabolic reactions (Raymond and Segrè, 2006). These new reactions generated many novel metabolites, such as steroids, alkaloids, and isoflavonoids. Steroids crucially modulate membrane functions, especially endo- and exocytosis, which facilitate intra- and intercellular communications in multicellular organisms (Chen et al., 2007; Summons et al., 2006). Some steroids (e.g., estrone and testosterone) and isoflavonoids (e.g., genistein and daidzein) are signaling molecules that target nuclear receptors (Jiang et al., 2010), which are also indispensable for establishing complicated signaling systems in higher organisms.

Although the Earth was pervasively oxygenated 2.3–2.4 billion years (Gyr) ago during the Great Oxidation Event (GOE) (Campbell and Allen, 2008; Canfield, 2005; Sessions et al., 2009), the record of aerobic metabolism in the chronology of F and FSF suggests that aerobic respiration appeared 2.8 Gyr ago and that the most ancient aerobic biosynthesis occurred 2.9 Gyr ago (Wang et al., 2011). These findings are well supported by a series of recent geochemical and biochemical observations that suggest that aerobic metabolism appeared 300–400 million years earlier than the GOE (David and Alm, 2011; Stolper et al., 2010; Waldbauer et al., 2009). Protein domain structures can also record important geochemical events associated with the rise of oxygen, especially because biochemistry coevolves with geochemistry (Saito et al., 2003; Williams and Fraústo Da Silva, 2003; Anbar, 2008). For example recent investigations of metalloprotein evolution showed that F and FSF domain history reflects the bioavailability of metals in the geochemical record. The earliest manganese and heme iron protein F and FSF precede copper counterparts (Dupont et al., 2010; Ji et al., 2009). This supports geochemical evidence that manganese and iron were bioavailable on anaerobic Earth, whereas copper availability was restricted under oxygen limitation (Anbar, 2008; Saito et al., 2003; Williams and Fraústo Da Silva, 2003).

The details about origin and early evolution of aerobic metabolism recorded in the chronology of Fs and FSFs are, nevertheless, limited because the link between domains at these levels of structural abstraction and protein function is rather loose. It remains unknown which aerobic reactions were first developed, what was the origin of the oxygen fueling the earliest aerobic reactions, and whether geochemical imprints exist in metabolism other than the usage of metallic cofactors. Because protein functions can be generally unambiguously assigned to FF domains (Murzin et al., 1995) and FFs are conserved enough to survey ancient evolutionary history and dissect protein recruitment in biological networks (Caetano-Anollés et al., 2012, 2011), here, we use a chronology of FF domains to uncover the evolutionary details of aerobic metabolism and to address the intriguing origins of planetary oxygen and aerobic reactions.

RESULTS AND DISCUSSION

Metabolic History Recorded in Protein Domain Families

We reconstructed an intrinsically rooted phylogenomic tree that describes the evolution of 3,513 FFs that are known, i.e., the evolution of parts (domains) of a system (proteomes), using a structural census of 989 completely sequenced genomes. Each proteome that is used as "phylogenetic character" represents a lineage of a tree of life that describes the evolution of systems, and not of parts, and has branches that are shared with more and more proteomes as one travels from the leaves to the root of the tree (K.M.K. and G.C.-A., unpublished data). The branch that is at the root of the tree of life represents a portion of history that is common to all proteomes (organisms) and impacts the genomic counts of the entire phylogenetic matrix. However, and by definition, each proteomic character contributes these portions of history to the global phylogeny of FF domains independently, avoiding constraints on phylogeny reconstruction.

The geological ages of selected FFs were estimated using molecular clocks of F and FSF domains (Wang et al., 2011), provided the FFs were the most ancient in each group. The fit (and confidence belt) that links the age of domains and geological timescales indicates that FF age can be obtained with confidence throughout the timeline. We note, however, that (1) the exact order of closely positioned FFs can be debatable in phylogenetic reconstructions of trees with thousands of leaves, despite robust evolutionary trends across phylogenies (Caetano-Anollés et al., 2009a), (2) the molecular clock derived from trees of Fs and FSFs is necessarily based on relatively few molecular fossils, and (3) rates of domain discovery and accumulation could be deviant for some domain structures (Wang et al., 2011). These factors could cause departures from a clock, with overdispersion sometimes resulting from changes in foldability and structural stability of domains (C. Debes, M. Wang, F. Graeter, and G.C.-A., unpublished data).

The accumulation patterns of total FFs along evolutionary timelines show that the growing universe of domains at this level of abstraction underwent a clear expansion ~1.4 Gyr ago (Figure 1). This expansion was most likely responsible for the rise of diversified multicellular life (Hedges et al., 2006) and the appearance of the first Eukarya-specific FFs (Caetano-Anollés et al., 2011; Wang et al., 2007). Two less-clear FF expansions ~2.9 and ~2.3 Gyr ago are also evident in the timeline (Figure 1). The former coincides with major events in organism diversification (Blank, 2009), including the rise of first lineages (Kim and Caetano-Anollés, 2011) and the loss of the first FF domains in Archaea (Caetano-Anollés et al., 2011; Wang et al., 2007), whereas the latter corresponds to bacterial diversification (Wang et al., 2011). These expansions establish a biphasic pattern of growth for the world of FF domains that resembles an hourglass (Caetano-Anollés et al., 2011) and has implications for the generation of modules and hierarchical complexity in life (J.E. Mittenthal, D.C.-A., and G.C.-A., unpublished data).

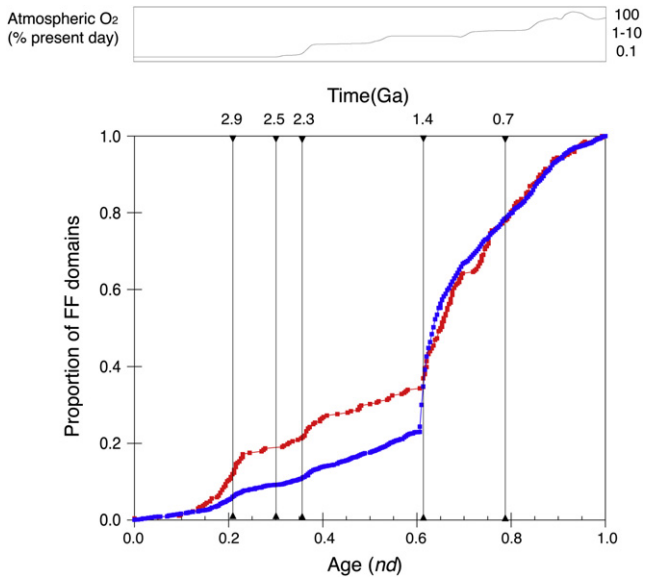

Figure 1. Accumulation of Total Protein Families (in Blue) and Single-Domain Enzyme Families (in Red)

Geological ages were calculated using the molecular clock of protein folds (Wang et al., 2011). Variation of oxygen concentration with geological ages is from Campbell and Allen (2008).

See also Tables S1–S3.

Here, we establish a temporal order of appearance of crucial protein enzyme-catalyzed metabolic reactions. We consider that chemical reactions of primordial metabolism were inherited from a prebiotic world (Morowitz, 1999) and that their piecemeal replacement by enzyme-catalyzed reactions provided strong selective advantages to primordial cells. Because metabolic history can be blurred by the accretion of domains in enzymes (Caetano-Anollés et al., 2007) and most multidomain proteins have been shown to appear quite late in evolutionary history (Wang and Caetano-Anollés, 2009), our analysis excluded all but single-domain enzymes. According to Protein Data Bank (PDB) (Berman et al., 2000) and KEGG (Kanehisa et al., 2010) records, single-domain enzymes embody 276 FFs with single enzyme functions (see Table S1 available online). These enzyme activities could be described by at least the same first three Enzyme Commission (EC) levels of classification (class, subclass, and subsubclass), guaranteeing a tight link between enzyme FF structure and function. When we linked enzymatic functions and KEGG pathways to the chronicled FFs, we found that early FFs are all involved in primary metabolism, with purine and pyrimidine metabolism appearing first, followed by amino acid, carbohydrate, cofactor, terpenoid, glycerolipid, and porphyrin metabolism (Table S2). This order is congruent with patterns of origin and evolution of metabolic networks that we revealed previously by studying enzymatic structure and domain recruitment at F level (Caetano-Anollés et al., 2007, 2009b). Remarkably, patterns recapitulate in part the early evolution of prebiotic metabolism (Caetano-Anollés et al., 2009b) embodied in the metabolic shells of Morowitz (1999), in which an energetic amphiphile shell that included energy-linked hydrolysis and biosynthesis of fatty acids precedes the biosynthesis of amino acids. Salient evolutionary patterns are revealed in the timeline of single-domain enzymes.

The First Metabolic Enzymes Were ATP Phosphohydrolases Linked to Nucleotide Metabolism

The most ancient enzymatic reactions that were identified support the origin of metabolism in energy interconversion reactions of nucleotide metabolism, as we showed earlier by studying enzymes at F level (Caetano-Anollés et al., 2007). The oldest enzyme, the ATP phosphohydrolase (EC 3.6.3.49), harbors the ABC transporter ATPase domain-like family structure (c.37.1.12, *nd* = 0), suggesting that the first primordial enzymes hydrolyzed ATP and had a P loop hydrolase fold (c.37) design. This again confirms previous observations (Caetano-Anollés and Caetano-Anollés, 2003; Kim and Caetano-Anollés, 2010; Caetano-Anollés et al., 2011). Because ATP-binding proteins selected from a random peptide library exhibit ATP-hydrolysis activity (e.g., Simmons et al., 2009) and have substructures that match the c.37.1 architecture (Caetano-Anollés et al., 2012), the present finding suggests that the birth of the most ancient proteins is likely a result of ATP selection from a pool of random peptides. Results are therefore consistent with the cofactor-selection model for protein origins (Ji et al., 2007; Tokuriki and Tawfik, 2009).

The Origin of Amino Acid Biosynthetic Pathways Is Congruent with the Temporal Order of Codon to Amino Acid Assignments and Is Contemporary to Recruitment of Aminoacyl-tRNA Synthetase Editing and Anticodon-Binding Domains

The pathways for early FFs indicate that the evolution of amino acid biosynthesis proceeds according to the temporal order of codon to amino acid assignments established by other methods (Johnson and Wang, 2010; Trifonov et al., 2006; Wong, 2005). The very early appearance of 3-isopropylmalate dehydrogenase (EC 1.1.1.85 with its dimeric isocitrate and isopropylmalate dehydrogenase c.77.1.1 FF [*nd* = 0.095] and EC 4.2.1.33 with its LeuD-like c.8.2.1 FF [*nd* = 0.139]) of the Val, Leu, and Ileu biosynthesis metabolic subnetwork (AAC 00290), and ornithine carbamoyltransferase (EC 2.1.3.3; aspartate/ornithine carbamoyltransferase c.78.1.1 FF, *nd* = 0.099) of Arg and Pro metabolism (AAC 330), suggests a very early origin (*nd* = 0.095–0.099) of protein enzyme-mediated biosynthesis of some or all of these amino acids (Table S2). Both enzymes belong to the amino acid metabolism (AAC) mesonetwork of KEGG. They were followed by chorismate synthase (EC 4.2.3.5; chorismate synthase, AroC d.258.1.1 FF, *nd* = 0.179) of the Phe, Tyr, and Trp metabolism subnetwork (AAC 00400), histidinol dehydrogenase (EC 1.1.1.23; L-histidinol dehydrogenase HisD c.82.1.2 FF, *nd* = 0.186) of His metabolism (AAC 00340), and methionine adenosyltransferase (EC 2.5.1.6; S-adenosylmethionine synthase d.130.1.1 FF, *nd* = 0.190), formylmethionine deformylase (EC 3.5.1.31; peptide deformylase d.167.1.1 FF, *nd* = 0.208), and homoserine O-acetyltransferase (EC 2.3.1.31; O-acetyltransferase c.69.1.40 FF, *nd* = 0.212) of Cys and Met metabolism (AAC 00270). The mapping of these single-domain enzymes (molecular landmarks) in the timeline provides a temporal order of appearance of amino acid biosynthetic pathways, starting with those involved in the biosynthesis of Val, Leu, Ileu, Pro, and Arg, followed by those linked to Phe, Tyr, and Trp, and ending in the biosynthesis of His, Cys, and Met (Table S2). This

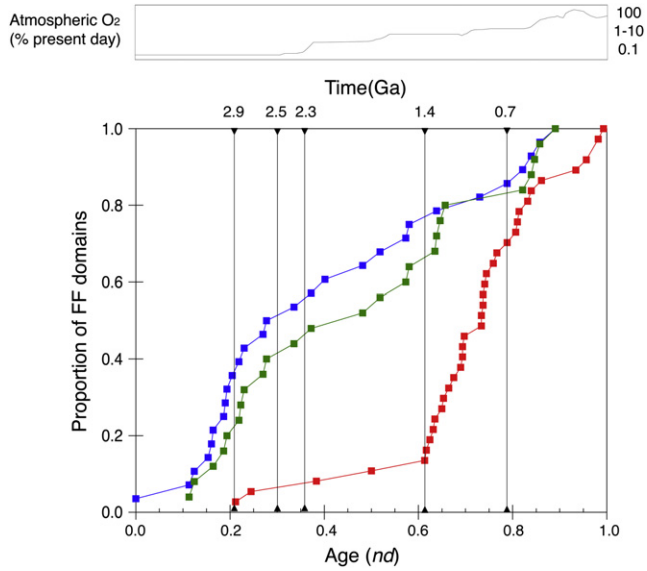

Figure 2. Accumulation of Oxygen-Consuming Families (in Red), ATP-Binding Families (in Blue), and Phosphotransferase Families (in Green)

Oxygen-consuming families contain ten proteins in the aerobic network simulated by Raymond and Segrè (2006), which do not involve oxygen explicitly. Geological ages were calculated using the molecular clock of protein folds (Wang et al., 2011). Variation of oxygen concentration with geological ages is from Campbell and Allen (2008).
See also Figure S1.

congruence manifests within an *nd* of 0.095–0.212 span and supports the stereochemical hypothesis in which interactions between polynucleotides containing codon/anticodon sequences and their corresponding amino acids are responsible for developing the genetic code (Woese et al., 1966). The amino acid biosynthesis progression is also consistent with evolution of amino acid usage in proteins, with Ala, Gly, Glu, Val, and Pro appearing earlier than Phe, Tyr, Trp, His, Cys, and Met. This possible connection of amino acid usage and the establishment of the genetic code was intimated by Zuckerkandl et al. (1971) and suggested by several genome-based studies (e.g., Brooks and Fresco, 2002; Jordan et al., 2005). The biosynthesis of His appears last in the timeline delimited by single-domain backbone enzymes. His is widely used in catalytic centers and metal-chelating sites, and its prebiotic synthesis is believed to have occurred with low efficiency (Plankensteiner et al., 2006; Shen et al., 1990). Histidinol dehydrogenase (EC 1.1.1.23) catalyzes the last two steps of histidine biosynthesis (Hernández-Montes et al., 2008). According to the forward evolution model for biosynthesis of His (Cunchillos and Lecointre, 2007), the last enzymatic steps represent the completion of the evolution of this pathway. The molecular clock of folds suggests that this enzyme (with the ALDH-like c.82 F domain; nd_F = 0.156) appeared 3.2 Gyr ago, and histidine was not highly available before this time. This explains why metalloproteins that bind transition metals emerged late (Dupont et al., 2010; Ji et al., 2009). In fact the first FSFs in metalloproteins (metallo-hydrolase/oxidoreductase d.157.1; nd_{FSF} = 0.094) appeared ~3.3 Gyr ago (Dupont et al., 2010) and have binding sites that are dominated by His (67.22%) (Table S3). No wonder they were absent in the early protein world.

The first recruitment of editing (ValRS/IleRS/LeuRS editing domain b.51.1.1 FF; *nd* = 0.126) and anticodon-binding domains (beginning with those for Val, Leu, Ileu, and Pro that harbor the anticodon-binding domains of class II aminoacyl-tRNA synthetase [aaRS] c.51.1.1 and class I a.27.1.1 FFs) follows the discovery of catalytic domains in aaRSs of class I (c.26.1.1) and class II (d.104.1.1) at a *nd* of 0.020–0.024 and manifests within the *nd* 0.126–0.240 span. These domains perform crucial editing operations and contact the anticodon arm of tRNA to define amino acid-charging specificities. Remarkably, their initial recruitment is contemporary to the origin of biosynthetic pathways described above (*nd* = 0.095–0.212) and occurred soon after the emergence of ribosomal proteins at a *nd* of 0.114 but before the emergence of the peptidyl transferase center (PTC) of the ribosome at a *nd* of 0.253 that is responsible for modern protein synthesis (Caetano-Anollés et al., 2012, 2011; Kim and Caetano-Anollés, 2011; A. Harish and G.C.-A., unpublished data). Thus, tRNA recognition and aaRS specificities unfolded as protein enzymes replaced prebiotic counterparts in primordial metabolic pathways.

Relatively Late Appearance of Secondary Metabolism

Signature FFs for secondary metabolism are only evident at *nd* >0.2 and are present in enzymes for tropane, piperidine and pyridine alkaloid biosynthesis (*nd* = 0.208), streptomycin biosynthesis (*nd* = 0.245), clavulanic acid biosynthesis (*nd* = 0.617), flavonoid biosynthesis (*nd* = 0.646), isoquinoline alkaloid biosynthesis (*nd* = 0.675), and ascorbate and aldarate metabolism (*nd* = 0.745). The functions of these secondary metabolites are associated with antibiotic (streptomycin, clavulanic acid), UV-radiation protection (flavonoids), defense (tropane, piperidine and pyridine alkaloids and isoquinoline alkaloids), and antioxidant (ascorbate) activities. Some of these activities involve assembly line of nonribosomal protein synthesis (NRPS), an enzyme complex that in our timeline appears earlier than the ribosome. NRPS-linked secondary metabolites are therefore relatively derived, suggesting that the NRPS assembly line was probably used to synthesize primordial proteins before its use for specialized cellular machinery (Caetano-Anollés et al., 2012).

The Origins of Aerobic Metabolic Reactions and Planetary Oxygen

We explore the first metabolic reactions of aerobic metabolism and the primordial sources of planetary oxygen. The history of aerobic metabolism unfolds in the timeline of FF discovery (Figure 2). Oxygen-consuming FFs increased slowly during the Archean eon and then abruptly ~1.4 Gyr ago. In contrast, ATP-binding families appeared at the very beginning of life and accumulated quite rapidly. These distinct accumulation patterns reflect the varying availability of ATP and oxygen in evolutionary history and agree well with their early and late participation in enzymatic reactions, respectively (Ma et al., 2008). Although prior studies suggest that aerobic metabolism appeared 2.8–2.9 Gyr ago (David and Alm, 2011; Stolper et al., 2010; Waldbauer et al., 2009; Wang et al., 2011), these studies did not characterize the possible ancient aerobic reactions. Our chronology reveals that the earliest oxygen-utilizing FF, the PNP-oxidase

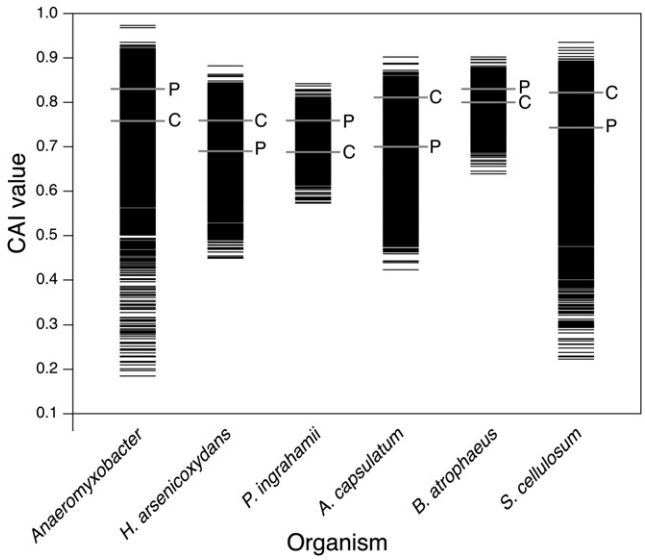

Figure 3. Gene Expression Levels of Six Bacteria Predicted Using the CAI

Three bacteria (*Anaeromyxobacter Fw* 109-5, *Herminiimonas arsenicoxydans*, *Psychromonas ingrahamii* 37) are anaerobic, and three (*Acidobacterium capsulatum* ATCC 51196, *Bacillus atrophaeus* 1942, *Sorangium cellulosum* so ce56 are aerobic. The expression levels for the genes of pyridoxal 5′-phosphate synthase (P) and catalase (C) are comparable in anaerobic and aerobic bacteria.

like b.45.1.1 FF (*nd* = 0.212), appeared ~2.9 Gyr ago (estimated by the split barrel-like b.45 F; nd_F = 0.245) with the pyridoxal 5′-phosphate synthase enzyme (EC 1.4.3.5) (Table S1). This enzyme is mainly used to produce the common cofactor pyridoxal 5′-phosphate or pyridoxal. Pyridoxal 5′-phosphate acts as crucial coenzyme in transamination reactions that transfer amino groups between amino acids and α-keto acids and in condensation reactions in heme synthesis. The cofactor participates in biosynthesis of cofactors and amino acids. Thus, it seems that the most primordial aerobic reactions were not for energy generation, but for biosynthesis. This can also be explained by cofactor availability. Copper is a predominant metal cofactor used by enzymes of aerobic respiration (Wang et al., 2011) and was poorly accessible at the beginning of the aerobic period (Anbar, 2008; Saito et al., 2003; Williams and Fraústo Da Silva, 2003). In turn the coenzymes of pyridoxal 5′-phosphate synthase are flavin adenine dinucleotide (FAD) or flavin mononucleotide (FMN), which are very ancient coenzymes and were available in the anaerobic period (White, 1976; Ji et al., 2007; Caetano-Anollés et al., 2012).

What were the oxygen sources for the earliest aerobic reactions? The K_M value (Michaelis-Menten constant) for oxygen of pyridoxal 5′-phosphate synthase (~85 μM, according to BRENDA; Chang et al., 2009) is relatively high. Although this could be taken as indication that oxygen was rich in local environments ~2.9 Gyr ago, high-enzymatic K_M values could simply reflect relatively small quantities of biosynthesis products rather than high-localized oxygen abundance. For example steroid biosynthesis also depends on oxygen, exhibiting K_M values for oxygen that are at micromolar levels. However, the process of steroid biosynthesis is microaerobic and can occur at oxygen concentrations as low as 7 nM (Waldbauer et al., 2011). Although oxygenic photosynthesis may have appeared at that time (Claire et al., 2006; Nisbet et al., 2007), other pathways may have been responsible for supplying oxygen. For instance the strictly anaerobic bacterium *Methylomirabilis oxyfera* has the ability to produce oxygen for its own aerobic enzymes, e.g., methane monooxygenase (with a K_M of ~17 μM, according to BRENDA) (Ettwig et al., 2010). This is of special interest because pyridoxal 5′-phosphate synthase is indeed present in some strictly anaerobic bacteria, such as *Anaeromyxobacter* sp., *Herminiimonas arsenicoxydans*, and *Psychromonas ingrahamii*, and gene expression levels of the enzyme predicted using the codon adaptation index (CAI) are comparable in anaerobic and aerobic bacteria (Figure 3). In these studies gene expression data were not available, and gene expression was predicted from genomic data. The CAI measures the relative adaptive capacity of each codon from its frequency within a selected highly expressed gene set, and combines these weights to define a value for each gene in a sequenced genome. We note that the mechanism by which these anaerobic bacteria provide oxygen for pyridoxal 5′-phosphate synthase activity remains unknown. Although the three anaerobic bacteria lack genes for methane monooxygenase, they contain genes for catalase, a tetrameric enzyme that uses porphyrin heme iron groups to catalyze the decomposition of hydrogen peroxide into water and oxygen. Remarkably, catalase is relatively highly expressed in these organisms (Figure 3). Catalase has also been observed in strictly anaerobic archaeal species such as *Methanosarcina barkeri* (Brioukhanov et al., 2006). Because Mn catalase is structurally similar to half of the tetranuclear Mn center that makes up the photosynthetic oxygen-evolving complex, this oxygen-producing pathway has been considered a precursor of oxygenic photosynthesis (Blankenship and Hartman, 1998). It is noteworthy that Mn catalase originated 3.0 Gyr ago (estimated by the age of the ferritin-like a.25.1 FSF; nd_{FSF} = 0.171) and is, therefore, contemporary to the birth of pyridoxal 5′-phosphate synthase. We therefore conclude that the oxygen generated during the Mn catalase-mediated detoxification of hydrogen peroxide likely fueled the primordial aerobic reactions. Remarkably, high concentrations of hydrogen peroxide can be produced in the prebiotic conditions of early Earth via a surface-mediated reaction between water and pyrite in the absence of dissolved oxygen (Borda et al., 2001). More likely, however, would be the massive release of UV-induced oxygen peroxide that accumulated in the ice of the Pongola Supergroup glaciation ~2.9 Gyr ago (and later glaciation events) in basal meltwaters (Kirschvink and Kopp, 2008). We crucially note that late planetary oxygenation (~2.3 Gyr ago) has been linked to the "snowball" planetary glaciation inferred from the Transvaal Supergroup and the massive deposition of Mn in the Kalahari field that occurred during that time (Kopp et al., 2005). These events suggest a crucial role of glacial peroxides and Mn in the development of oxygen-mediating enzymes that probably started with the Pongola glaciation.

If indeed hydrogen peroxide was increasingly available in Archean times, we must explain why Mn catalase detoxification appeared so late in protein evolution (3.0 Gyr ago). A structural analysis of the catalytic center of Mn catalase provides a clue.

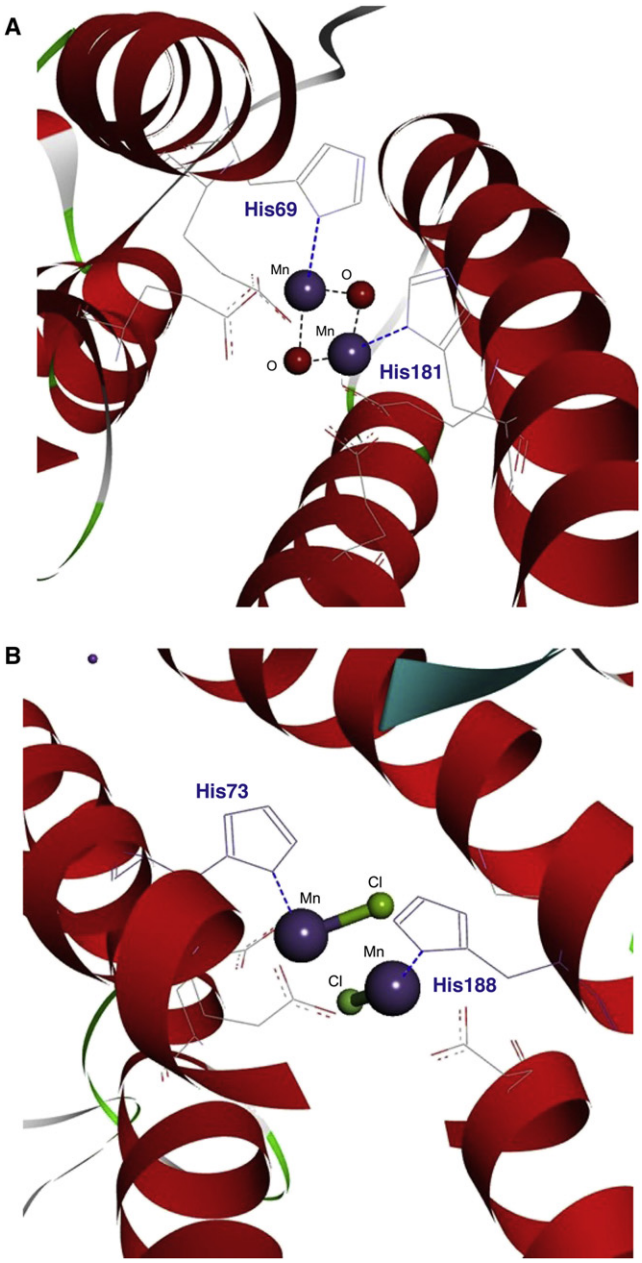

Figure 4. Schematic Representation of Mn-Chelating Histidine Residues in Mn Catalases

(A) Mn catalase from *Lactobacillus plantarum* (PDB entry: 1JKU) and (B) *Thermus thermophilus* (PDB entry: 2V8T). The histidine residues are highly conserved in extant Mn catalase sequences.

A crystallographic model of Mn catalase indicates that two His residues are indispensable to the chelation of Mn (Figure 4). As described above, His was not highly available to proteins through enzymatic biosynthesis until 3.2 Gyr ago. We therefore propose that the low availability of His (originally synthetized prebiotically) restricted the emergence of Mn catalase. This constraint implies the existence of a link between evolution of amino acid biosynthesis and the rise of oxygen.

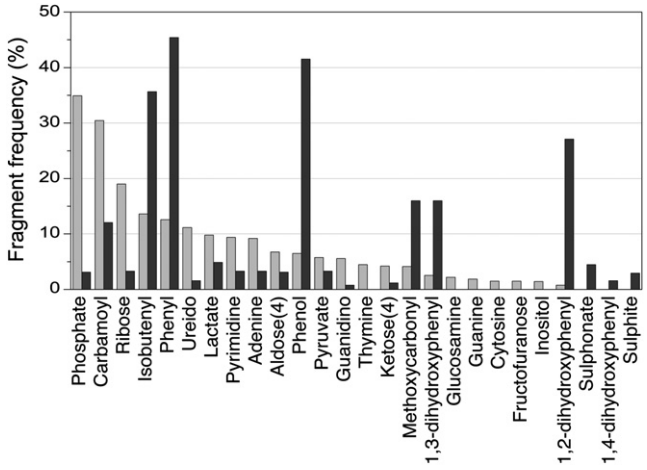

Figure 5. Fragment Usage in Anaerobic (Light Gray) and Aerobic (Dark Gray) Metabolites

It can be seen that carbamoyl, ureido, guanidino, and some basic building blocks of life, such as phosphate, ribose, pyrimidine, and purine, occur much more often in anaerobic than in aerobic metabolites ($p < 0.01$, chi-square test). In comparison, isobutenyl, phenyl, methoxycarbonyl, and phenols (including phenol, 1,2-dihydroxyphenyl and 1,3-dihydroxyphenyl) are much more frequently used in aerobic reactions ($p < 0.01$, chi-square test). More interestingly, some groups such as sulphonate, sulphite, and 1,4-dihydroxyphenyl are exclusively used by aerobic metabolites.

Geochemical Imprints of the Rise of Planetary Oxygen in Metabolic Evolution

The rise of oxygen not only brought novel biochemical reactions but also exerted big influences on the emerging chemistries of evolving Earth. Previous studies focused on metal imprints in proteins (Dupont et al., 2010; Ji et al., 2009). Our analysis now shows that groups transferred by enzymes also contain useful evolutionary and biogeochemical information associated with the rise of oxygen.

Functional annotations revealed that our enzyme data set is dominated by transferases (38%), in which phosphotransferases (EC 2.7.-.-) are the most abundant members (30%) (Figure S1). We find phosphotransferase FFs originated early and increased steadily until the present (Figure 2). In comparison the accumulation of transferases of sulfur-containing groups (EC 2.8.-.-) is quite complicated. Neutral or reduced sulfur groups were already transferred early in evolution starting at a *nd* of 0.197, whereas oxidized sulfur groups (e.g., sulfate and sulfite) were transferred much later within the *nd* 0.766–0.982 time frame (Table S1). It seems that enzymes did not use functional groups containing oxidized sulfur until oxygen levels were already elevated, which is supported by metabolite substructure (also termed fragment) usage in aerobic and anaerobic metabolites. We extracted and counted with Pipeline Pilot a total of 57 types of chemical substructures (according to the fragment library of Nobeli et al., 2003) in the 1,174 anaerobic and 520 aerobic metabolites with clearly defined structures obtained from the simulated metabolic networks (Raymond and Segrè, 2006). As shown in Figure 5, fragment compositions for anaerobic and aerobic metabolites were quite distinct. Carbamoyl, ureido, guanidino, and some basic building blocks of life, such

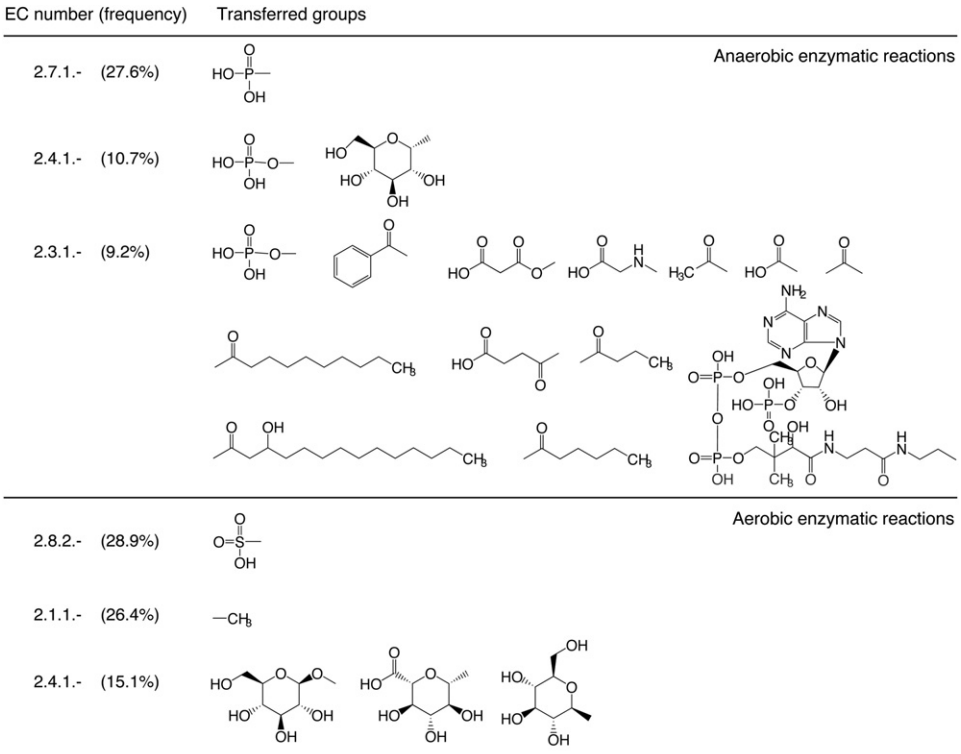

Figure 6. Some of the Most Frequently Transferred Groups in Anaerobic and Aerobic Enzymatic Reactions
It can be seen that phosphate and sulphonate are the most frequently transferred groups by anaerobic and aerobic transferases, respectively.

as phosphate, ribose, pyrimidine, and purine, occur much more often in anaerobic than in aerobic metabolites ($p < 0.01$, chi-square test). In turn, isobutenyl, phenyl, methoxycarbonyl, and phenols (including phenol and 1,2-, 1,3-dihydroxyphenyl) are much more frequently used in aerobic reactions. More importantly, some groups such as sulphonate, sulphite, and 1,4-dihydroxyphenyl are exclusively used by aerobic metabolites ($p < 0.01$, chi-square test). Out of all fragments, we find that phosphate and sulphonate are indeed the most frequently transferred groups by anaerobic and aerobic transferases, respectively (Figure 6). We also find a prevalence of phenolic groups in aerobic metabolites that could be attributed to the emergence of oxygen-dependent hydroxylation reactions. The exclusive use of sulphonate and sulphite groups in aerobic reactions could be best explained by the high and exclusive availability of oxidized sulfur groups in the aerobic world (Canfield, 2005; Sessions et al., 2009). In contrast, phosphates were available during the entire history of Earth (Planavsky et al., 2010; Yamagata et al., 1991), explaining why they are highly used by anaerobic metabolites. Taken together, these results suggest that geochemical evolution has left imprints not only in protein compositions (metal cofactors) but also in metabolites and related reactions.

In summary the evolution of metabolism is overwhelmingly driven by the recruitment of protein domains and enzymes (Caetano-Anollés et al., 2007, 2009b). Protein FF domains are generally unambiguously linked to molecular functions and are therefore powerful in their ability to dissect recruitment in metabolic networks. In this study we linked the evolutionary age of FFs that are present in enzymes to metabolic reactions. We find that the timeline of FFs validates patterns in the history of metabolic evolution that were previously proposed, such as the very early appearance of ATP phosphohydrolases and transferases, the origin of pathways in metabolic shells, and the coevolution of amino acid biosynthesis, the genetic code, and aaRS recognition of identity elements in tRNA. The evolutionary timeline of FFs also revealed the origin and early evolution of aerobic metabolism, a critical invention in biological and planetary history. The aerobic synthesis of pyridoxal 5′-phosphate and pyridoxal was identified as the most ancient aerobic reaction. A molecular clock of folds established its appearance 2.9 Gyr ago. We also explored the possible oxygen source for this reaction and found that Mn catalase, which appeared at the same time as the most ancient aerobic reactions, could have generated oxygen as a side product of hydrogen peroxide detoxification in glaciation meltwaters. Finally, analysis of the evolutionary features of transferred groups in metabolic reactions and metabolite fragment usage in aerobic and anaerobic metabolites also revealed that geochemical evolution associated with the rise of planetary oxygen left imprints in metabolites and reactions. The evolutionary patterns in molecules and chemistries provide strong support for the coevolution of biochemistry and geochemistry. Moreover, our findings showcase the power of using protein structures as molecular fossils for reconstruction of ancient evolutionary history. Although technical advances have enabled the sequencing of fossilized

proteins (e.g., collagens) (Schweitzer et al., 2007, 2009), our phylogenomic strategies take advantage of the expanding data of genomic research. Rapid advances in functional, structural, and evolutionary genomics are expected to enhance the timeline of protein domain structures and provide unanticipated views of paleobiology.

EXPERIMENTAL PROCEDURES

Phylogenomic Methods

A timeline of domain discovery was derived from a universal phylogenomic tree of FF domain structure. Protein structural domains corresponding to 3,513 FFs (out of 3,902 defined by SCOP 1.75) were assigned to proteomes of 989 organisms whose genomes were completely sequenced (76 Archaea, 656 Bacteria, and 257 Eukarya). This structural genomic census used the iterative Sequence Alignment and Modeling System (SAM) method to scan genomic sequences (with probability cutoffs E of 10^{-4}) against a library of advanced linear hidden Markov models (HMMs) of structural recognition in SUPERFAMILY (Gough et al., 2001). The census produced a data matrix with columns representing proteomes (phylogenetic characters) and rows representing FFs (phylogenetic taxa). This matrix was used to build a phylogenetic tree of FF domain structure using the maximum parsimony (MP) method in PAUP* version 4.0b10 (Swofford, 2002) and a combined parsimony ratchet (PR) and iterative search approach (Wang and Caetano-Anollés, 2009) to facilitate tree reconstruction and avoid the risk for optimal trees being trapped in suboptimal regions of tree space. A single MP reconstruction was retained following 300 ratchet iterations (10 × 30 chains) with 1,000 replicates of random taxon addition, tree bisection reconnection (TBR) branch swapping, and maxtrees unrestricted. Concise classification strings (ccs) defined SCOP domains at FF level (e.g., c.37.1.12, where c represents the protein class, 37 the F, 1 the FSF, and 12 the FF) and were used to identify taxa in trees. Finally, the relative age of protein architectures (*nd*) was calculated directly from the phylogenomic tree using a PERL script that counts the number of nodes from the ancestral architecture at the base of the tree to each leaf and provides it in a relative zero to one scale. A recent review summarizes the general approach and the progression of census data and tree reconstruction in recent years (Caetano-Anollés et al., 2009a). In addition the phylogenomic approach based on a genomic census is robust against uneven sampling of genomes across the three superkingdoms (Kim and Caetano-Anollés, 2011).

Identification of Single-Domain Enzyme FFs

The 3,902 FFs defined by SCOP 1.75 (February 2009) cover 38,221 PDB entries. These proteins include 19,038 enzymes harboring 1,421 enzyme activities defined at 4 levels of EC classification (February 2011). Out of these enzymes, 4,138 consist of single-domain proteins corresponding to 416 FFs and 711 EC 4-level activities. To guarantee a tight link between FF and enzyme function, FFs with unambiguous enzymatic activities were identified by selecting enzymes with activities defined in at least three levels of EC classification. Out of the initial 416 FFs, 276 FFs were unambiguously linked to 347 EC numbers. These FFs catalyze 658 biochemical reactions recorded in KEGG (Kanehisa et al., 2010). Reaction directions and main reaction pairs were used to identify substrates and products of enzymatic activities.

Assigning Geological Ages for Protein Domains

The molecular clocks for protein domains at F ($t = -3.802\ nd + 3.814$, in Gyr) and FSF ($t = -3.831\ nd + 3.628$, in Gyr) levels (Wang et al., 2011) were used to calculate the geological ages of some families, provided that FFs were the most ancient in each group. These clocks were calibrated with geological ages derived from the study of fossils and geochemical, biochemical, and biomarker data, which are affected by the validity of the assumptions used in each and every one of the supporting studies. In few instances domain distribution in lineages, fossil, and standard molecular data confirms calibration data. For example the distribution of the chalcone isomerase fold (d.36) that is uniquely linked to flavonoid biosynthesis in 749 genomes suggests that the appearance of the pathway can be traced back to the appearance of red algae 1.2 Gyr ago, which coincides with the appearance of the earliest eukaryotic fossil assigned to a living lineage (the sexual red algae *Bangiomorpha*), and to dates of the red algal lineage inferred with standard molecular clocks (see Wang et al., 2011). A systematic comparison of dates (and constraints) for the rise of various metabolic pathways is needed and will be the subject of future studies.

Using the CAI to Predict Gene Expression Levels

The genomes of three anaerobic bacteria, *Anaeromyxobacter Fw* 109-5 (NC_009675), *Herminiimonas arsenicoxydans* (NC_009138), and *Psychromonas ingrahamii* 37 (NC_008709), and three aerobic bacteria, *Acidobacterium capsulatum* ATCC 51196 (NC_012483), *Bacillus atrophaeus* 1942 (NC_014639), and *Sorangium cellulosum so ce 56* (NC_010162), were downloaded from the NCBI genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Expression levels for the genes in these bacteria were predicted using the CAI (Sharp and Li, 1987). The relative adaptiveness of a codon is defined as its frequency relative to the most often used synonymous codon. In this study the "cai" program in the EMBOSS software package was used to calculate CAI values, which range from zero to one. Genes with higher CAI values are more likely to be highly expressed (Ishihama et al., 2008; Jansen et al., 2003; Sharp and Li, 1987).

Identification of Aerobic and Anaerobic Metabolites

The metabolic networks simulated by Raymond and Segrè (http://prelude.bu.edu/O2/networks.html) consist of 1,326 anaerobic metabolites (blue nodes) and 538 aerobic metabolites (red nodes) (Raymond and Segrè, 2006). We identified 1,174 anaerobic metabolites and 520 aerobic metabolites with clearly defined structures (without R group or polymeric form) in this set using the following protocol: (i) for multicomponent records the small fragments (counter-ions in salts, solvent molecules) were removed, and only the largest fragments were retained; (ii) hydrogen atoms were added to fill the valences of heavy atoms and to neutralize the molecular charges; and (iii) three-dimensional structures were generated for all of the compounds with Pipeline Pilot (Student Edition, Version 6.1.5; SciTegic Accelrys, San Diego, CA, USA).

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure and three tables and can be found with this article online at doi:10.1016/j.str.2011.11.003.

ACKNOWLEDGMENTS

We thank Mak Saito for constructive comments during review. This study was supported by the National Basic Research Program of China (973 project, Grant 2010CB126100), National Natural Science Foundation of China (Grant 30870520), Fundamental Research Funds for the Central Universities (Grants 2011PY027 and 2011PY142 to H.-Y.Z.), and National Science Foundation (Grant MCB-0749836 to G.C.-A.).

Received: August 9, 2011
Revised: November 6, 2011
Accepted: November 8, 2011
Published: January 10, 2012

REFERENCES

Anbar, A.D. (2008). Oceans. Elements and evolution. Science *322*, 1481–1483.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Blank, C.E. (2009). Not so old Archaea—the antiquity of biogeochemical processes in the archaeal domain of life. Geobiology *7*, 495–514.

Blankenship, R.E., and Hartman, H. (1998). The origin and evolution of oxygenic photosynthesis. Trends Biochem. Sci. *23*, 94–97.

Borda, M.J., Elsetinow, A.R., Schoonen, M.A., and Strongin, D.R. (2001). Pyrite-induced hydrogen peroxide formation as a driving force in the evolution of photosynthetic organisms on an early earth. Astrobiology *1*, 283–288.

Brioukhanov, A.L., Netrusov, A.I., and Eggen, R.I. (2006). The catalase and superoxide dismutase genes are transcriptionally up-regulated upon oxidative

stress in the strictly anaerobic archaeon *Methanosarcina barkeri*. Microbiology *152*, 1671–1677.

Brooks, D.J., and Fresco, J.R. (2002). Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol. Cell. Proteomics *1*, 125–131.

Caetano-Anollés, G., and Caetano-Anollés, D. (2003). An evolutionarily structured universe of protein architecture. Genome Res. *13*, 1563–1571.

Caetano-Anollés, G., Kim, H.S., and Mittenthal, J.E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc. Natl. Acad. Sci. USA *104*, 9358–9363.

Caetano-Anollés, G., Wang, M., Caetano-Anollés, D., and Mittenthal, J.E. (2009a). The origin, evolution and structure of the protein world. Biochem. J. *417*, 621–637.

Caetano-Anollés, G., Yafremava, L.S., Gee, H., Caetano-Anollés, D., Kim, H.S., and Mittenthal, J.E. (2009b). The origin and evolution of modern metabolism. Int. J. Biochem. Cell Biol. *41*, 285–297.

Caetano-Anollés, D., Kim, K.M., Mittenthal, J.E., and Caetano-Anollés, G. (2011). Proteome evolution and the metabolic origins of translation and cellular life. J. Mol. Evol. *72*, 14–33.

Caetano-Anollés, G., Kim, K.M., and Caetano-Anollés, G. (2012). The phylogenomic roots of modern biochemistry: origins of proteins, cofactors and protein biosynthesis. J. Mol. Evol., in press.

Campbell, I.H., and Allen, C.M. (2008). Formation of supercontinents linked to increases in atmospheric oxygen. Nat. Geosci. *1*, 554–558.

Canfield, D.E. (2005). The early history of atmospheric oxygen: homage to Robert M. Garrels. Annu. Rev. Earth Planet. Sci. *33*, 1–36.

Catling, D.C., Glein, C.R., Zahnle, K.J., and McKay, C.P. (2005). Why O^2 is required by complex life on habitable planets and the concept of planetary "oxygenation time". Astrobiology *5*, 415–438.

Chang, A., Scheer, M., Grote, A., Schomburg, I., and Schomburg, D. (2009). BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res. *37* (Database issue), D588–D592.

Chen, L.L., Wang, G.Z., and Zhang, H.Y. (2007). Sterol biosynthesis and prokaryotes-to-eukaryotes evolution. Biochem. Biophys. Res. Commun. *363*, 885–888.

Claire, M.W., Catling, D.C., and Zahnle, K.J. (2006). Biogeochemical modelling of the rise in atmospheric oxygen. Geobiology *4*, 239–269.

Cunchillos, C., and Lecointre, G. (2007). Ordering events of biochemical evolution. Biochimie *89*, 555–573.

Dahl, T.W., Hammarlund, E.U., Anbar, A.D., Bond, D.P., Gill, B.C., Gordon, G.W., Knoll, A.H., Nielsen, A.T., Schovsbo, N.H., and Canfield, D.E. (2010). Devonian rise in atmospheric oxygen correlated to the radiations of terrestrial plants and large predatory fish. Proc. Natl. Acad. Sci. USA *107*, 17911–17915.

David, L.A., and Alm, E.J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. Nature *469*, 93–96.

Dupont, C.L., Butcher, A., Valas, R.E., Bourne, P.E., and Caetano-Anollés, G. (2010). History of biological metal utilization inferred through phylogenomic analysis of protein structures. Proc. Natl. Acad. Sci. USA *107*, 10567–10572.

Ettwig, K.F., Butler, M.K., Le Paslier, D., Pelletier, E., Mangenot, S., Kuypers, M.M., Schreiber, F., Dutilh, B.E., Zedelius, J., de Beer, D., et al. (2010). Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. Nature *464*, 543–548.

Falkowski, P.G., and Isozaki, Y. (2008). Geology. The story of O2. Science *322*, 540–542.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J. Mol. Biol. *313*, 903–919.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics *22*, 2971–2972.

Hernández-Montes, G., Díaz-Mejía, J.J., Pérez-Rueda, E., and Segovia, L. (2008). The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. Genome Biol. *9*, R95.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J., and Frishman, D. (2008). Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics *9*, 102.

Jansen, R., Bussemaker, H.J., and Gerstein, M. (2003). Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res. *31*, 2242–2251.

Ji, H.F., Kong, D.X., Shen, L., Chen, L.L., Ma, B.G., and Zhang, H.Y. (2007). Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. Genome Biol. *8*, R176.

Ji, H.F., Chen, L., Jiang, Y.Y., and Zhang, H.Y. (2009). Evolutionary formation of new protein folds is linked to metallic cofactor recruitment. Bioessays *31*, 975–980.

Jiang, Y.Y., Kong, D.X., Qin, T., and Zhang, H.Y. (2010). How does oxygen rise drive evolution? Clues from oxygen-dependent biosynthesis of nuclear receptor ligands. Biochem. Biophys. Res. Commun. *391*, 1158–1160.

Johnson, D.B.F., and Wang, L. (2010). Imprints of the genetic code in the ribosome. Proc. Natl. Acad. Sci. USA *107*, 8298–8303.

Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., Wolf, Y.I., Koonin, E.V., Kondrashov, A.S., and Sunyaev, S. (2005). A universal trend of amino acid gain and loss in protein evolution. Nature *433*, 633–638.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. *38* (Database issue), D355–D360.

Kim, H.S., Mittenthal, J.E., and Caetano-Anollés, G. (2006). MANET: tracing evolution of protein architecture in metabolic networks. BMC Bioinformatics *7*, 351.

Kim, K.M., and Caetano-Anollés, G. (2010). Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. Mol. Biol. Evol. *27*, 1710–1733.

Kim, K.M., and Caetano-Anollés, G. (2011). The proteomic complexity and rise of the primordial ancestor of diversified life. BMC Evol. Biol. *11*, 140.

Kirschvink, J.L., and Kopp, R.E. (2008). Palaeoproterozoic ice houses and the evolution of oxygen-mediating enzymes: the case for a late origin of photosystem II. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 2755–2765.

Kopp, R.E., Kirschvink, J.L., Hilburn, I.A., and Nash, C.Z. (2005). The Paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. Proc. Natl. Acad. Sci. USA *102*, 11131–11136.

Ma, B.G., Chen, L., Ji, H.F., Chen, Z.H., Yang, F.R., Wang, L., Qu, G., Jiang, Y.Y., Ji, C., and Zhang, H.Y. (2008). Characters of very ancient proteins. Biochem. Biophys. Res. Commun. *366*, 607–611.

Morowitz, H.J. (1999). A theory of biochemical organization, metabolic pathways, and evolution. Complexity *4*, 39–53.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Nisbet, E.G., Grassineau, N., Howe, C.J., Abell, P.I., Regelous, M., and Nisbet, R.E.R. (2007). The age of Rubisco: the evolution of oxygenic photosynthesis. Geobiology *5*, 311–335.

Nobeli, I., Ponstingl, H., Krissinel, E.B., and Thornton, J.M. (2003). A structure-based anatomy of the *E. coli* metabolome. J. Mol. Biol. *334*, 697–719.

Payne, J.L., Boyer, A.G., Brown, J.H., Finnegan, S., Kowalewski, M., Krause, R.A., Jr., Lyons, S.K., McClain, C.R., McShea, D.W., Novack-Gottshall, P.M., et al. (2009). Two-phase increase in the maximum size of life over 3.5 billion years reflects biological innovation and environmental opportunity. Proc. Natl. Acad. Sci. USA *106*, 24–27.

Planavsky, N.J., Rouxel, O.J., Bekker, A., Lalonde, S.V., Konhauser, K.O., Reinhard, C.T., and Lyons, T.W. (2010). The evolution of the marine phosphate reservoir. Nature *467*, 1088–1090.

Plankensteiner, K., Reiner, H., and Rode, B.M. (2006). Amino acids on the rampant primordial Earth: electric discharges and the hot salty ocean. Mol. Divers. *10*, 3–7.

Raymond, J., and Segrè, D. (2006). The effect of oxygen on biochemical networks and the evolution of complex life. Science *311*, 1764–1767.

Saito, M.A., Sigman, D.M., and Morel, F.M.M. (2003). The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean-Proterozoic boundary? Inorganica Chim. Acta *356*, 308–318.

Schweitzer, M.H., Suo, Z., Avci, R., Asara, J.M., Allen, M.A., Arce, F.T., and Horner, J.R. (2007). Analyses of soft tissue from Tyrannosaurus rex suggest the presence of protein. Science *316*, 277–280.

Schweitzer, M.H., Zheng, W., Organ, C.L., Avci, R., Suo, Z., Freimark, L.M., Lebleu, V.S., Duncan, M.B., Vander Heiden, M.G., Neveu, J.M., et al. (2009). Biomolecular characterization and protein sequences of the Campanian hadrosaur B. canadensis. Science *324*, 626–631.

Sessions, A.L., Doughty, D.M., Welander, P.V., Summons, R.E., and Newman, D.K. (2009). The continuing puzzle of the great oxidation event. Curr. Biol. *19*, R567–R574.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.

Shen, C., Yang, L., Miller, S.L., and Oró, J. (1990). Prebiotic synthesis of histidine. J. Mol. Evol. *31*, 167–174.

Simmons, C.R., Stomel, J.M., McConnell, M.D., Smith, D.A., Watkins, J.L., Allen, J.P., and Chaput, J.C. (2009). A synthetic protein selected for ligand binding affinity mediates ATP hydrolysis. ACS Chem. Biol. *4*, 649–658.

Stolper, D.A., Revsbech, N.P., and Canfield, D.E. (2010). Aerobic growth at nanomolar oxygen concentrations. Proc. Natl. Acad. Sci. USA *107*, 18755–18760.

Summons, R.E., Bradley, A.S., Jahnke, L.L., and Waldbauer, J.R. (2006). Steroids, triterpenoids and molecular oxygen. Philos. Trans. R. Soc. Lond. B Biol. Sci. *361*, 951–968.

Swofford, D.L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10 (Sunderland, MA: Sinauer Associates).

Tokuriki, N., and Tawfik, D.S. (2009). Protein dynamism and evolvability. Science *324*, 203–207.

Trifonov, E.N., Gabdank, I., Barash, D., and Sobolevsky, Y. (2006). Primordia vita. Deconvolution from modern sequences. Orig. Life Evol. Biosph. *36*, 559–565.

Waldbauer, J.R., Sherman, L.S., Sumner, D.Y., and Summons, R.E. (2009). Late Archean molecular fossils from the Transvaal Supergroup record the antiquity of microbial diversity and aerobiosis. Precambrian Res. *169*, 28–47.

Waldbauer, J.R., Newman, D.K., and Summons, R.E. (2011). Microaerobic steroid biosynthesis and the molecular fossil record of Archean life. Proc. Natl. Acad. Sci. USA *108*, 13409–13414.

Wang, M., and Caetano-Anollés, G. (2009). The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. Structure *17*, 66–78.

Wang, M., Yafremava, L.S., Caetano-Anollés, D., Mittenthal, J.E., and Caetano-Anollés, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res. *17*, 1572–1585.

Wang, M., Jiang, Y.Y., Kim, K.M., Qu, G., Ji, H.F., Mittenthal, J.E., Zhang, H.Y., and Caetano-Anollés, G. (2011). A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Mol. Biol. Evol. *28*, 567–582.

White, H.B., 3rd. (1976). Coenzymes as fossils of an earlier metabolic state. J. Mol. Evol. *7*, 101–104.

Williams, R.J., and Fraústo Da Silva, J.J. (2003). Evolution was chemically constrained. J. Theor. Biol. *220*, 323–343.

Woese, C.R., Dugre, D.H., Saxinger, W.C., and Dugre, S.A. (1966). The molecular basis for the genetic code. Proc. Natl. Acad. Sci. USA *55*, 966–974.

Wong, J.T. (2005). Coevolution theory of the genetic code at age thirty. Bioessays *27*, 416–425.

Yamagata, Y., Watanabe, H., Saitoh, M., and Namba, T. (1991). Volcanic production of polyphosphates and its relevance to prebiotic evolution. Nature *352*, 516–519.

Zuckerkandl, E., Derancourt, J., and Vogel, H. (1971). Mutational trends and random processes in the evolution of informational macromolecules. J. Mol. Biol. *59*, 473–490.

