A Universal Molecular Clock of Protein Folds and Its Power in Tracing the Early History of Aerobic Metabolism and Planet Oxygenation

Minglei Wang,[1] Ying-Ying Jiang,[2,3] Kyung Mo Kim,[1] Ge Qu,[3] Hong-Fang Ji,[3] Jay E. Mittenthal,[4] Hong-Yu Zhang,*[,2] and Gustavo Caetano-Anollés*[,1]

[1]Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois, Urbana-Champaign

[2]National Key Laboratory of Crop Genetic Improvement, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, People's Republic of China

[3]Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo, People's Republic Of China

[4]Department of Cell and Developmental Biology, University of Illinois, Urbana-Champaign

The first two authors contributed equally to this work.

***Corresponding author:** E-mail: zhy630@mail.hzau.edu.cn; gca@illinois.edu.

Associate editor: Jeffrey Thorne

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

Abstract

The standard molecular clock describes a constant rate of molecular evolution and provides a powerful framework for evolutionary timescales. Here, we describe the existence and implications of a molecular clock of folds, a universal recurrence in the discovery of new structures in the world of proteins. Using a phylogenomic structural census in hundreds of proteomes, we build phylogenies and time lines of domains at fold and fold superfamily levels of structural complexity. These time lines correlate approximately linearly with geological timescales and were here used to date two crucial events in life history, planet oxygenation and organism diversification. We first dissected the structures and functions of enzymes in simulated metabolic networks. The placement of anaerobic and aerobic enzymes in the time line revealed that aerobic metabolism emerged about 2.9 billion years (giga-annum; Ga) ago and expanded during a period of about 400 My, reaching what is known as the Great Oxidation Event. During this period, enzymes recruited old and new folds for oxygen-mediated enzymatic activities. Remarkably, the first fold lost by a superkingdom disappeared in Archaea 2.6 Ga ago, within the span of oxygen rise, suggesting that oxygen also triggered diversification of life. The implications of a molecular clock of folds are many and important for the neutral theory of molecular evolution and for understanding the growth and diversity of the protein world. The clock also extends the standard concept that was specific to molecules and their timescales and turns it into a universal timescale-generating tool.

Key words: enzyme, structure, function, evolution, aerobic metabolism.

Introduction

Protein domains are compact and more or less independent folding units of structure that recur in different molecular contexts, are conserved, and are considered functional and evolutionary units of classification (Caetano-Anollés et al. 2009; Chothia and Gough 2009). A number of popular protein domain classification schemes are available that are based on features that exist in sequence and structure. For example, the Structural Classification of Proteins (SCOP) is a high-quality taxonomical resource that groups domains that have known 3D structures into families, superfamilies, and folds (Murzin et al. 1995). Fold families (FFs) group domains that are closely related at the sequence level, generally with more than 30% pairwise amino acid identities. Fold superfamilies (FSFs) unify FFs that share functional and structural features and that share a common evolutionary origin. Finally, folds group FSFs that have similar arrangements of secondary structures in 3D space but that may not be evolutionarily related. SCOP has defined about 1,200 folds and about 2,000 FSFs, and over 10^7 proteins belonging to over 1,000 completely sequenced genomes have been assigned to FSFs by scanning with hidden Markov models (HMMs; Gough et al. 2001). The relatively small number of folds and FSFs implies that they are more conserved than protein sequences and are therefore useful to explore ancient evolutionary events.

The repertoire of proteins of an organism (its proteome) defines a collection of domains that is specific to that organism and represents the modern instantiation of an unbroken chain of protein ancestors. This is reasonable because modern proteomes are encoded in genomes that can also be traced back to ancient ancestors along tree-like or network-like lineages using the tools of phylogenomic analysis (Doolittle 2005). To study the history of the protein world, a census of domains in the proteomes of hundreds of fully sequenced organisms was conducted and used to build phylogenomic trees of protein domains at different levels of structural organization (Caetano-Anollés and

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Caetano-Anollés 2003). These trees define evolutionary chronologies of domain architecture, which can be used, for example, to explore the evolution of proteomes, proteins, and networks (recently reviewed in Caetano-Anollés et al. 2009). Here, we describe for the first time the existence of a universal molecular clock of folds, a temporal recurrence of architectural innovation in proteins. This clock extends the power of assigning absolute timescales to phylogenetic trees of protein sequences (Doolittle et al. 1996; Feng et al. 1997), a feature we illustrate by studying the early history of aerobic metabolism and its centrality in organismal diversification.

The rise of atmospheric oxygen is a critical event in the history of our planet and has significant implications for biological evolution. Some crucial evolutionary events, such as the birth of eukaryotes and the explosion of animal diversity in the Cambrian, have been linked to elevated atmospheric oxygen (Falkowski and Isozaki 2008). The time of appearance of oxygen in meaningful amounts in the atmosphere and the molecular mechanisms responsible for oxygen-facilitated evolution are therefore challenging and important problems that need to be addressed. Currently, questions related to planet oxygenation depend largely on geochemical methodologies (Sessions et al. 2009). However, contamination of sediments and biomarkers are sometimes responsible for analyses reaching divergent conclusions (Anbar et al. 2007; Kaufman et al. 2007; Rasmussen et al. 2008; Godfrey and Falkowski 2009; Hoashi et al. 2009; Kato et al. 2009). Questions related to oxygen-facilitated evolution are generally linked to aerobic respiration, a process that involves many enzymes and is about 16 times more efficient in generating ATP than anaerobic metabolic pathways (Catling et al. 2005). Moreover, a number of other fundamental innovations brought about by aerobic metabolism are also beneficial. For instance, sterols play a critical role in regulating membrane functions (e.g., endo- and exocytosis) in eukaryotes and their biosynthesis, which is dependent on oxygen, was likely beneficial for the emergence of the eukaryotic superkingdom (Summons et al. 2006; Chen et al. 2007). In a recent simulation of metabolic networks under anaerobic or aerobic conditions, molecular oxygen enabled over 1,000 metabolic reactions (Raymond and Segrè 2006). Because the structure of enzymes is tightly correlated and coevolves with associated biological functions (Caetano-Anollés et al. 2009), the functional makeup of aerobic enzymes must be linked to a structural set of protein domains that harbor the necessary metabolic functions. Here, we explore how the clock-like history of domain structure is linked to the history of aerobic metabolism, revealing the emergence and evolution of oxygen-facilitated biochemistries during planet oxygenation.

Materials and Methods

Phylogenomic Analysis of Protein Domain Structure

Phylogenomic analysis follows previously described methodology (Caetano-Anollés and Caetano-Anollés 2003). We first conducted a census of genomic sequence in 749 organisms (52 archaeal, 478 bacterial, and 219 eukaryal species) assigning protein structural domains at fold and FSF levels of structural complexity to protein sequences using advanced linear HMMs of structural recognition in Superfamily (Gough et al. 2001; Andreeva et al. 2008) and probability cutoffs *e* of 10^{-4}. Domains were defined by SCOP version 1.73 (Murzin et al. 1995) and described using SCOP concise classification strings (*ccs*). Fold domains were assigned to FSFs using SCOP identifiers (IDs). The census was then used to construct data matrices of genomic abundance (*g*) of fold or FSF. Here, *g* indicates the number of multiple occurrences of a fold or FSF per proteome. Empirically, *g* values range from 0 to thousands and resemble morphometric data with a large variance (Wang and Caetano-Anollés 2009; Wang et al. 2007). Because existing phylogenetic programs can process only tens of phylogenetic character states depending on user's CPU performance, the space of *g* values in the matrix was reduced using a standard gap-coding technique with the following formula:

$$g_{ab\_\text{norm}} = \text{Round}\left[\frac{\ln(g_{ab} + 1)}{\ln(g_{ab\_\text{max}} + 1)} \times 20\right]$$

In this equation, *a* and *b* denote a fold (or FSF) and a proteome, respectively. g_{ab} represents the *g* value of the fold (or FSF) *a* in proteome *b*. g_{ab_max} indicates the maximum g_{ab} value in all folds (or FSFs) in a individual proteome. This round function normalizes a genomic abundance value of a particular fold or FSF in a proteome regarding the maximum *g* value and standardizes the values to a 0–20 scale (g_{ab_norm}). The 21 normalized *g* values represent character states and are encoded as linearly ordered multistate phylogenetic characters using an alphanumeric format of numbers 0–9 and letters A–K that are compatible with PAUP* version 4.0b10 (Swofford 2002). The 21 character states are polarized from "K" to "0" using the ANCSTATES command in PAUP* based on two fundamental premises: 1) protein structure is far more conserved than sequence and carries considerable phylogenetic signal, especially at high levels of structural organization (e.g., fold level), and 2) folds and FSFs that are successful and popular in nature are generally more ancestral, making "K" the most ancient character state and "0" the most recent. Details and support for character argumentation and absence of circularity in assumptions have been described and discussed previously (Caetano-Anollés and Caetano-Anollés 2003; Wang et al. 2006, 2007; Wang and Caetano-Anollés 2009; Kim and Caetano-Anollés 2010). Also, because fold and FSF domains are retained over long evolutionary periods, their gain or loss constitutes important events that appear to be independent of horizontal gene transfer and other convergent evolutionary processes (Gough 2005; Forslund et al. 2008; Yang and Bourne 2009; Kim and Caetano-Anollés 2010). Universal phylogenetic trees of protein domain structure were then built from the matrices using maximum parsimony as the optimality criterion in PAUP* and rooted by the Lundberg method (Swofford 2002). Because trees are

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

large and the search of tree space is computationally hard, we used a combined parsimony ratchet and iterative search approach to facilitate tree reconstruction (Wang and Caetano-Anollés 2009). Multiple iterations avoid the risk of optimal trees being trapped by suboptimal regions of tree space (Nixon 1999). A recent review summarizes the general approach and the progression of census data and tree reconstruction in recent years (Caetano-Anollés et al. 2009). Tree balance statistics (N-bar and cherry count) that measure the symmetry of trees and statistics that describe the shape of trees (node height, E/I ratio, and treeness) were calculated using TreeStat version 1.2 (http://tree.bio.ed.ac.uk/software/treestat). N-bar is the number of internal nodes between the base and the tips of the tree (Kirkpatrick and Slatkin 1993), and cherry count is the number of internal nodes that have only terminal leaves as children (McKenzie and Steel 2000). Node height measures the height of each internal node in a tree; the E/I ratio describes the ratio between the total lengths of external branches and the length of internal branches; and "treeness" represents the proportion of the total length of the tree that correspond to internal branches, interpreted as a measure of how well the data fit the tree in phylogenetic reconstruction.

Because trees are rooted and are highly unbalanced, we unfolded the relative age of protein domains directly for each phylogeny as a distance in nodes (node distance, nd) from the hypothetical ancestral architecture at the base of the trees in a relative 0–1 scale. Given a rooted tree, we calculated nd by counting the number of internal nodes along a lineage from the root to a terminal node (a leaf) of the tree on a relative 0–1 scale with the following equation: $\text{nd}_a = (\text{\# of internal nodes between nodes } r \text{ and } a)/(\text{\# of internal nodes between nodes } r \text{ and } m)$, where a means a target leaf node, r is a hypothetical root node, and m is a leaf node that has the largest possible number of internal nodes from the node r. Consequently, the nd value of the most ancestral taxon is 0, whereas that of the most recent one is 1. Node distance can be a good measure of age given a rooted tree because the emergence of protein domains (i.e., taxa) is displayed by their ability to diverge (cladogenesis or molecular speciation) rather than by the amount of character state change that exists in branches of the tree (branch lengths). The fact that rates of genetic change were positively correlated with rates of divergence suggests that nongradual evolutionary phenomena are rare in nature (Webster et al. 2003) and supports the semipunctuated emergence of protein domains in the highly unbalanced trees of folds (fig. 1A) and FSFs (supplementary fig. S1, Supplementary Material online).

To study how character state change distributes in phylogenetic trees, we also reconstructed trees of proteomes by transposing the data matrix of fold abundance. An unrooted phylogeny of 749 proteomes was obtained and character state changes were reconstructed on every single branch of the tree using PAUP*. These values were then plotted against the age of the individual folds. Because the abundance of domains in proteomes was range standardized to a 0–20 scale (this range is compatible with most phylogenetic analysis programs), total character state changes were adjusted to average fold abundance levels in all proteomes that were analyzed.

In contrast with standard applications that reconstruct phylogenies of molecules with taxa representing organisms out of a pool of millions of species, trees of folds (and associated FSFs) represent the evolution of a limited and finite set of fold architectures that have been uncovered by evolutionary change. This set is expected not to exceed about 1,600 folds (Levitt 2007). Undiscovered folds are also expected to be rare and consequently of recent origin. Our tree therefore applies to an entire repertoire and cannot be subjected to the procedure of taxon sampling without affecting significantly the validity of biological findings. All known taxa must be included in the analysis. Although tree statements generated from abundance or occurrence of domains in genomes were not significantly different (Kim KM and Caetano-Anollés G, unpublished data), phylogenetic analyses depend, for example, on the accuracy and balance of genomic databases (especially related to how representative they are of the biosphere), efficient and accurate assignment of structures to protein sequences, and methods of phylogenetic tree reconstruction. We do not expect that the effect of biases (e.g., faulty detection of FSFs with HMM, overrepresentation of superkingdoms; discussed in Caetano-Anollés and Caetano-Anollés 2003) will seriously affect the conclusions of this study.

Calibration of Protein Fold Chronology and Phylogenomic Inferences

Geological ages derived from fossils and geochemical, biochemical, and biomarker data were associated to the discovery of fold domain architectures and used to calibrate the chronology. When selecting proteins, we compared nd values for folds and related FSFs of corresponding domain architectures to dissect the effect of recruitment in protein evolution. The following folds were linked to geological ages:

(i) Boundary fold linked to the origin of proteins: The earliest evidence of biological activity is supported by ion microprobe analysis and carbon isotope composition of carbonaceous inclusions that exist in banded iron rock formations of Greenland (Mojzsis et al. 1996). These studies suggest life originated about 3.8 billion years (giga-annum; Ga) ago. We assume this time corresponds to the appearance of the first protein fold in our phylogeny, the P-loop containing nucleoside triphosphate hydrolase (c.37, nd = 0).

(ii) Fold linked to the biosynthesis of porphyrins: Porphyrins are essential metabolites to all living systems, even the most primitive ones. A recent advanced spectroscopic analysis identified vanadyl–porphyrin complexes in carbonaceous matter embedded in a 3.49-Ga-old polycrystalline rock that originated from silica sediments deposited on the floors of primitive seas (Gourier et al. 2010). These carbonaceous microstructures probably represent one of the oldest putative traces of life. This study reveals that metal ions of porphyrins were progressively replaced by vanadyl ions,

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

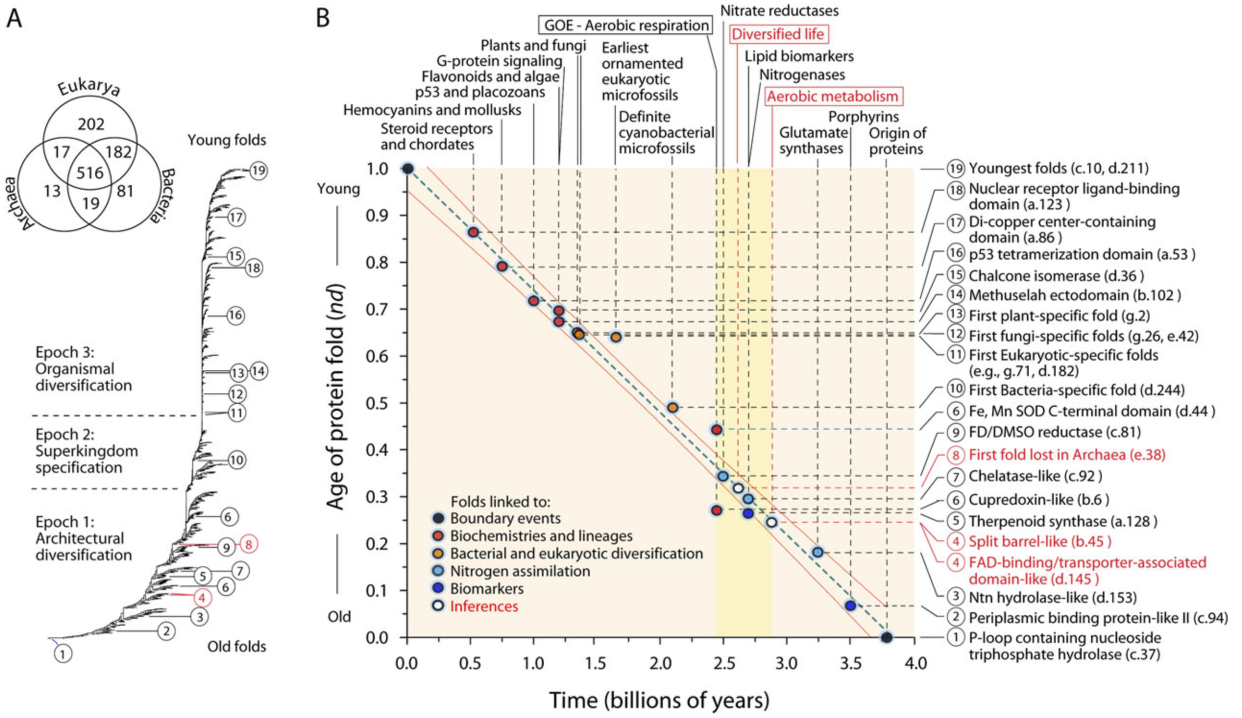

Fig. 1 A molecular clock of folds and recorded evolutionary events associated with the rise of planetary oxygen and organismal diversification. (*A*) Phylogenomic tree of folds (541,383 steps; CI = 0.028, RI = 0.783; $g_1 = -0.111$). Taxa are not labeled with fold *ccs* as labels would not be legible. The tree is highly unbalanced and asymmetric (N-bar = 111.7, expectation is 13.0; cherry count = 238, rejects Yule model null hypothesis at a significance of 0.05) and its shape is highly biased (ratio of external to internal branch lengths = 1.89; treeness = 0.346; also revealed by plots of node height of internal nodes that are not shown). Molecular fossils are labeled with numbers, and boundaries delimiting the three epochs of protein evolution (Wang et al. 2007) are indicated with dashed lines. The Venn diagram shows occurrence of folds in the three superkingdoms of life. Note that folds common to all life are located at the base of the tree. (*B*) Linear relationship between the age of fundamental evolutionary events and associated folds (color circles). Curve fit and 95% confidence belt are in red. Rationale for the linkage of folds and events is provided in the Materials and Methods. According to this linear correlation, b.45 and d.145 (white circle) were estimated to appear approximately 2.9 Ga ago, which implies an approximately 400 My time frame (yellow-shaded area) for planet oxygenation. Similarly, the first fold lost by a superkingdom (white circle) and the start of diversified life was estimated to occur 2.6 Ga ago. CI, consistency index; RI, retention index.

giving very stable vanadyl–porphyrin complexes, which are now universally found within biogenic terrestrial carbonaceous materials such as petroleum, bitumen, and coals. Therefore, vanadium porphyrins are not an artifact of breakdown of proteins and non-porphyrin metabolites. In a prior study (Caetano-Anollés et al. 2007), we revealed that some enzymes responsible for porphyrin biosynthesis (with EC numbers 4.2.1.24 and 4.1.1.37) indeed originated very early. These enzymes use the TIM β/α-barrel (c.1) and the periplasmic binding protein-like II (c.94) folds. The former comprises 33 FSFs; however, the latter comprises only one FSF and contains the catalytic site. The c.94 fold (nd = 0.0729) is therefore linked to the appearance of porphyrins.

(iii) Folds linked to nitrogen assimilation: To use the various nitrogen species in the environment, organisms have evolved a variety of nitrogen assimilation enzymes. Glutamate synthases appeared the earliest (3.25 Ga) (Glass et al. 2009), which consist of two domains and use the single-stranded right-handed β-helix fold (b.80, nd = 0.151) and Ntn hydrolase-like fold (d.153, nd = 0.182), respectively. Because d.153 appeared later than b.80 and contains the catalytic sites (according to the records in Catalytic Site Atlas; Porter et al. 2004), we used d.153 as landmark for glutamate synthases (this enzyme belongs to the earliest FSF of d.153). Nitrogenases appeared 2.7 Ga ago (Glass et al. 2009). Despite the diverse metallic cofactor usage of nitrogenases (Mo, V, and Fe), these enzymes use the same fold, the chelatase-like fold (c.92, nd = 0.297), and belong to the earliest FSF of this fold. The ferredoxin-nitrate reductase appeared 2.5 Ga ago (Garvin et al. 2009), which comprises four domains belonging to the ferredoxin-like (d.58; nd = 0.010), double-ψ β-barrel (b.52; nd = 0.245), formate dehydrogenase/dimethyl sulfoxide reductase domains 1–3 (c.81; nd = 0.344), and heme-binding four-helical bundle (f.21; nd = 0.349) folds. As the catalytic sites are located in domains of d.58 and c.81, and c.81 appeared later than d.58, we linked c.81 (comprising only one FSF) to the evolution of ferredoxin-nitrate reductase.

(iv) Fold linked to lipid biomarkers: Lipid biomarkers (e.g., hopanoids and biphytanes) result from structural and stereochemical transformations during diagenesis and are generally recovered from kerogen, bitumens, and hydrocarbons. Hopanoids are precursors of 2α-methylhopanes, the characteristic biomarkers of bacteria (including cyanobacteria) and indirectly of oxygenic photosynthesis (Sessions et al. 2009). Cyclic and acyclic phytanes and biphytanes, which are present in sediments and petroleum and can be generated diagenetically from the phytol side chain of chlorophylls, are biomarkers of methanotrophic pelagic microbes and are thought to derive from ether lipids of Archaea, such as archaeol and caldarchaeol (Ventura et al. 2007). Although the geological age of 2α-methylhopanes

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

determined by Brocks et al. (1999) (2.7 Ga) has been recently questioned (Rasmussen et al. 2008), the ancient history of lipid biomarkers is supported by the presence of hopanoids and biphytanes in 2.7-Ga-old metasedimentary rocks in several parts of the world (Ventura et al. 2007; Fischer 2008). Therefore, it is still reasonable to consider that these biomarkers of ancient microbial lipids appeared at that time. According to the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000), hopanoid biosynthesis depends on two enzymes, geranyltranstransferase and squalene synthase, both of which use the terpenoid synthase fold (a.128, nd = 0.265, comprising only one FSF). Folds linked to biphytane biosynthesis are sparse and very ancient, and because they have been probably co-opted during evolution (Kim et al. 2006), they are not used as molecular fossils in our study. The accuracy and evolutionary implications of lipid biomarkers are, however, affected by their synchronous deposit in the sediments where they are found and the link of their biological sources to relevant organisms and physiological conditions existing at a particular geological age (Sessions et al. 2009).

(v) Fold linked to serine β-lactamases: Class A serine β-lactamases cleave the β-lactam ring of β-lactam antibiotics such as penicillin and are ancient enzymes that originated 2.4 Ga ago (Hall and Barlow 2004). They harbor the β-lactamase/transpeptidase-like fold (e.4, nd = 0.307, comprising only one FSF).

(vi) Folds linked to bacterial and eukaryotic diversification: The oldest unambiguous microfossils of cyanobacteria with detailed morphological features reminiscent of subsections I and II (unicellular cyanobacterial coccoids) and III (filamentous cyanobacteria with only vegetative cells) of cyanobacteria were found in 1.9-Ga tidal flat sedimentary rocks of the Canadian Belcher Supergroup (Hofmann 1976). The earliest known cyanobacterial resting cells correspond to the genus *Archaeoellipsoides* and are preserved in 1.5- to 2.1-Ga-old cherts throughout the world (Tomitani et al. 2006). The oldest are from the Franceville Group of Gabon. Integration of molecular, physiological, paleontological, and geochemical data suggests that a diversified clade of cyanobacteria with marked heterocyst and cell differentiation appeared no later than 2.1 Ga ago (Tomitani et al. 2006). This age can be used as an internal bacterial calibration point for studies of molecular evolution in early organisms and can be linked to the first Bacteria-specific fold, the cell division protein ZapA-like fold (d.244; nd = 0.490). The earliest ornamented and ultrastructurally complex microfossils (acritarchs) that exist in 1.5-Ga-old rocks of the Roper Group of Northern Australia constitute the most convincing evidence for the existence of early-diversified eukaryotes (Javaux et al. 2004). Recent evidence associated with controversial Cambrian-like fossils of the Lower Vindhyan basin in Central India push back eukaryotic origins to the Paleoproterozoic (~1.65 Ga) (Bengtson et al. 2009). The diversity and complexity of these eukaryotic microfossils show they belong to highly diversified organisms (Bengtson et al. 2009) and can be linked to the first Eukarya-specific domains, including g.71 and d.182 (nd = 0.640) that appear in fold trees and are believed to be markers of lineage diversification in Eukarya (Wang et al. 2007). The early diversification of the fungal and plant lineages based on protein alignments, estimated to have occurred on average 1.37 and 1.38 Ga ago (Hedges et al. 2006; Bhattacharya et al. 2009), corresponds to the first fungi-specific (e.42 and g.26; nd = 0.641) and plant-specific (the toxin hairpin, g.2; nd = 0.651) folds, respectively. The appearance of organic-walled Precambrian microfossils linked to ascomycetes in the Riphean supports the early age of fungi (Hermann and Podkovyrov 2008).

(vii) Folds linked to biological processes and lineages: (a) Biosynthesis of flavonoids and red algae—The biosynthesis of flavonoids depends on a unique enzyme, chalcone isomerase (Jez et al. 2000). The enzyme owns exclusively the chalcone isomerase fold (d.36, nd = 0.697), which is thus intimately coupled with the appearance of flavonoids. The existence of algaeal flavonoids (Yumiko et al. 2003) and a survey of the d.36 in the 749 genomes we analyzed suggest that flavonoid biosynthesis can be traced back to ancestors of unicellular red algae, which appeared 1.2 Ga ago (Payne et al. 2009). This coincides with the appearance of the earliest eukaryotic fossil that can be assigned to a living lineage 1.2 Ga ago, the sexual red algae *Bangiomorpha* (Butterfield 2000). (b) G-protein signaling and G-protein coupled receptors (GPCR)—This prominent receptor family shares a common molecular architecture of seven transmembrane domains connected by three intracellular and three extracellular loops, which is linked to cyclic AMP/pheromone-like receptors and probably arose 1.2 Ga ago (Römpler et al. 2007). The 3D structure of an extracellular domain of GPCR has been determined, which belongs to Methuselah ectodomain fold (b.102, nd = 0.667). (c) p53 protein and placozoans—The p53 protein is the most commonly mutated tumor suppressor. It is conserved from placozoans to man and appeared about 1 Ga ago with basal eumetazoans (Lane et al. 2010). p53 uses exclusively the p53 tetramerization domain fold (a.53, nd = 0.718). (d) Hemocyanins and mollusks—Atmospheric oxygen levels reached 10% of the present atmospheric level (PAL) in the Precambrian (Payne et al. 2009). To use oxygen more efficiently, oxygen transporters appeared in evolution, with hemocyanins being invented by mollusks (emerging 0.75 Ga ago) (van Holde et al. 2001). These proteins consist of three different domains, in which the di-copper center-containing domain is the functional unit and uses the latest fold (a.86, nd = 0.792, comprising only one FSF). (e) Steroid receptors and chordates—Steroid receptors originated with the advent of chordates (0.52 Ga) (Thornton 2001; Holland et al. 2008; Michael and David 2009). The six-type related steroid receptors (α and β estrogen, progesterone, androgen, glucocorticoid, and mineralocorticoid receptors) exclusively use the nuclear receptor ligand–binding domain fold (a.123, nd = 0.864).

(viii) Present day boundary folds: We assume the present (0 Ga ago) corresponds to the age of the youngest fold domains, c.10 and d.211 (nd = 1).

Linear correlations between age of fold or FSF domains and geological time and associated statistics were calculated using statistical pipelines in SuperANOVA 1.1 (Abacus Concepts, Berkeley, CA) and Aabel 1.5.8 (Gigawiz Ltd). Because measurements are uncorrelated and have different uncertainties, we used a weighted least squares plot approach. A negative correlation is expected for an ideal tree obtained through an exhaustive search that is free from the effects of homoplasy. However, under this ideal situation, many scenarios are possible, including those that introduce heterogeneities in the evolutionary process such as an

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

accelerated or a very slow discovery of folds early or late in evolution. In fact, we have identified the existence of an important heterogeneity in protein domain organization (the arrangement of domains in multidomain proteins) that manifests as a big bang of domain combinations half way in evolution (Wang and Caetano-Anollés 2009). In reality, the exact order of closely positioned folds is potentially debatable in phylogenetic reconstruction of large trees, even if general trends across the phylogeny are robust and informative (Caetano-Anollés et al. 2009). All these factors affect the shape of the tree of domain structure and could cause significant departures from a molecular clock.

Aerobic and Anaerobic Enzymes in Metabolic Networks

We also identified enzymatic activities, enzymes, metabolites, and modules in metabolic networks that could develop under different conditions and were simulated using the metabolic network expansion method of Raymond and Segrè (2006), dissecting reactions of aerobic and anaerobic metabolism. In this method, each simulation started with a set of randomly chosen metabolites, which were allowed to interact with each other according to the rules of enzymatic reactions detailed in KEGG (Kanehisa and Goto 2000). At the time the simulations were performed, KEGG contained information related to more than 6,000 different enzymatic reactions derived from 70 genomes. Networks (http://prelude.bu.edu/O2/networks.html) consist of 1,326 anoxic metabolites and 538 oxic metabolites (Raymond and Segrè 2006), from which 1,145 anaerobic and 454 aerobic enzymatic activities were identified. A total of 650 anaerobic and 223 aerobic enzymes linked to these enzymatic activities had structural records in the Protein Data Bank (PDB) (Berman et al. 2000), and these records were used to identify protein domains at fold and FSF levels of structural complexity using SCOP definitions. If target proteins did not have entries, their homologues with sequence identity more than 30%, *e* value less than 10^{-10}, and alignment coverage greater than 50% were considered. Finally, hierarchical modules were identified in pathways of KEGG from the simulated anaerobic and aerobic metabolic networks. The initial reactants were the major metabolites that kick off the reactions in the module. The initial reactants for aerobic modules were all anaerobic metabolites. The reaction degree of initial reactants was defined as the counts of reactions in which they participate.

Results and Discussion

Phylogenomic Trees of Protein Domain Structure

Molecular fossils (biomarkers) are organic remains of ancient organisms preserved in sediments over geologic timescales (Sessions et al. 2009). We here consider that structural features in modern molecules that are evolutionarily durable represent (living) molecular fossils. We focus on the most conserved molecular feature that exists in proteins, the 3D arrangement of helices and strands characteristic of the fold (Caetano-Anollés et al. 2009), a feature that has been used successfully to trace ancient evolutionary events (Caetano-Anollés and Caetano-Anollés 2003; Ma et al. 2008; Caetano-Anollés et al. 2009). In this study, we used these features as molecular fossils to define timescales. In this process, we first reconstruct phylogenomic trees that describe the evolution of 1,030 fold and 1,730 FSF domain structures in SCOP (Murzin et al. 1995). Figure 1A shows the tree of fold domains and supplementary figure S1 (Supplementary Material online) the tree of FSF domains. These trees were generated from a census of fold and FSF domains in 749 proteomes using established methodology (Caetano-Anollés and Caetano-Anollés 2003). Although trees of domain structures exploit the shared-and-derived tenet of cladistic analysis, they are atypical and some of their major properties (which have been discussed elsewhere; see Caetano-Anollés and Caetano-Anollés 2003; Wang et al. 2006, 2007; Wang and Caetano-Anollés 2009) deserve a brief description:

(i) Trees of domain structures are universal phylogenetic statements that are intrinsically rooted and are highly unbalanced: Phylogenomic trees are derived from a census of protein domain structures in genomes that have been sequenced. Phylogenies have branches that represent lines of direct descent and describe the natural history of domain structures that are known. Consequently, they are phylogenetic statements that apply to the entire world of proteins. The model of character state transformation used in phylogenetic reconstructions (see Materials and Methods) produces trees that are intrinsically rooted. Local external hypotheses of relationship, such as "outgroup" taxa, are not required to establish evolution's arrow. These phylogenetic statements relate only to modern biochemistry, as information in molecules that are modern is used to reconstruct the past. Consequently, any claims are necessarily linked to the design and structure of extant molecules and not to those of their predecessors that were perhaps lost in evolution. An analysis of tree shape (node height of internal nodes, E/I ratios, and treeness statistics) and symmetry (N-bar and cherry counts) indicates that trees of domain structures are highly unbalanced (e.g., see legend of fig. 1). The unbalanced nature of these trees suggests that semi-punctuated evolutionary processes (Webster et al. 2003) are important drivers of structural discovery and that evolution of protein structure does not fit, for example, null branching models of change (Kirkpatrick and Slatkin 1993). In a way, these results are expected. The total number of folds will probably not exceed approximately 1,600 (Levitt 2007). This indicates that the discovery of individual folds occurred at extraordinarily low rate (once every $\sim 10^6$–10^7 years; Caetano-Anollés et al. 2009) and that each new domain structure must be regarded as a unique and rare event (a "punctuation") in evolutionary history.

(ii) The leaves of the trees (taxa) are domain structures: The structure of domains can be defined at different levels of the SCOP structural hierarchy—class, fold, FSF, and FF (Murzin et al. 1995). We selected the highest and more detailed hierarchical levels (folds or FSFs) as taxa because these levels coarse grain the 3D architectural design of proteins and are

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

evolutionarily highly conserved. Very much as with taxa of "trees of life" (i.e., trees of organismal species), each hierarchical level of structure was used separately in each tree reconstruction exercise to ensure homology. However, when reconstructing trees of species, the definition of what is a species is sometimes controversial (e.g., the definition of α-proteobacteria given pervasive horizontal gene transfer). Similarly, the reconstruction of trees of structures is also dependent on the validity and definition of terminal taxa. For example, a phylogenomic analysis of folds places trust on how SCOP assigns FSFs to folds. Although previous studies confirm the evolutionary relatedness of FSF in, for example, β/α-barrel folds (Copley and Bork 2000; Nagano et al. 2000), this important issue has not been explored for many other of the relatively few (mostly ancient) folds that harbor more than one FSF. In this regard, phylogenomic analysis at the FSF level offers a higher level of certainty that proteins belonging to this hierarchical level share a common evolutionary origin (Yang et al. 2005), especially because families unified by FSFs show good structural and functional evidence of common ancestry (Murzin et al. 1995). Previously, we reconstructed trees of folds and FSFs and traced the evolution of ancient folds along the branches of the trees of FSFs, showing folds and FSFs follow evolutionary pathways that are congruent and independent of the underlying hierarchical organization of domain structure (Wang et al. 2006). This is relevant because we also showed that folds were collections of evolving FSFs of different ages. In this study, we compare timescales derived from trees of folds and FSFs to determine possible biases introduced by the structural hierarchy and find that a molecular clock holds for the two structural levels. For simplicity, taxa (domains) were identified with *ccs* descriptors, widely used symbolic representations of domains within the hierarchy of structural classification. For example, the P-loop hydrolase FSF is named c.37.1, where c represents the protein class, 37 the fold, and 1 the FSF.

(iii) The internal nodes of the trees define chronologies of structural diversification: Whereas the leaves of the trees correspond to domain structures (folds or FSFs), the nodes represent structural diversification events that occur as changes in the popularity and spread of structures in the organismal world develop in time. Consequently, nodes close to the base of the tree reflect more ancient events than those close to the leaves. We interpret the birth of domain structures and their structural diversification in the context of degenerate mappings (and associated neutral nets) that exist between the space of protein sequences and the space of protein structures (reviewed in Caetano-Anollés et al. 2009 and Caetano-Anollés and Mittenthal 2010; discussed below). As protein sequences harboring a primordial fold drift by mutation in sequence space, seek stability, diversify, and populate the proteome of a primordial organism, new sequences are discovered that fold into new fold structures. Within this setting, the rare finding of a new set of sequences with a novel attribute of structure constitutes a molecular speciation event, which defines an internal node in the tree of domain structure. As evolution continues to build protein complements, variants of the primordial folds or the newly encountered structures may give rise to new molecular speciation events and new folds, and consequently new nodes in the evolving tree. Algorithmically and working down from the leaves to the root of the tree, a node that gives rise to two extant domains represents a hypothetical ancestor of these domains that, if it existed today, would have a greater genomic abundance than its children domains across the proteomes that are examined. Similarly, a hypothetical ancestor of an extant domain and of another hypothetical ancestor, or of two hypothetical ancestors, would have a greater abundance than its children. Without an effort of character state reconstruction and basic knowledge of change in protein structure, however, little can be said about the identity of hypothetical ancestors; the topology, thermodynamics, and function of the folds of hypothetical and extant molecules in a lineage of the tree must be necessarily compatible. Because trees are highly unbalanced and the timing of domain discovery is largely defined by molecular speciation and not by changes in domain abundance (see below), that is, by the shape of trees and not the lengths of branches, the number of internal nodes in lineages delimits structural diversification and is here used to define the relative age of domains and a molecular clock.

(iv) The features (phylogenetic characters) that are used to build trees are domain structures in proteomes and the values of data (character states) of their genomic abundance: Because proteomes define the growing protein complement of extant organisms and the complexity and diversity of the protein world, we counted the number of domain sequences that fold into domain structures in each proteome and use these numbers as character states to measure their genomic abundance. By definition, organisms and proteomes generally represent lineages that have distinct histories and fulfill the requirement of character independence that is necessary for phylogenetic analysis. We note, however, that crucial factors complicate the "model–tree–data" interplay of phylogenetic analysis, including the effect of convergent evolutionary processes (i.e., those that lead independently to a similar outcome) and horizontal transfer (Wang et al. 2006). Although these processes have the potential of obliterating patterns of descent, their effect on structure appears limited (Gough 2005; Forslund et al. 2008; Yang and Bourne 2009; Kim and Caetano-Anollés 2010). Similarly, the existence of lineages arising from multiple ancestors (e.g., ancestral architectures such as P-loop hydrolases and β/α-barrels arising independently from small peptides) could complicate phylogenetic interpretation (Wang et al. 2006). However, this scenario is less parsimonious and has been shown to be unlikely in evolution of life (Theobald 2010).

(v) The criterion of primary homology rests on genomic abundance levels of individual domains that exist in individual proteomes: Homology can be defined as correspondence arising from common ancestry, and its analysis represents a complex theoretical problem. Two criteria of homology can be distinguished: "primary" and "secondary" (de Pinna 1991). Primary homology assumes that parts of a set are the same by inheritance, whereas secondary homologies are primary homologies that materialize as shared and derived characters on a cladogram (synapomorphies), that is, they withstand the test of congruence. The criterion of primary homology in our study defines general patterns of correspondence in genomic abundance of domain structures in proteomes using topology and ontogenetic criteria (both of which are linked) that follow a transformation sequence of ordered and polarized multistate characters. The criterion rests on how domains distribute and are reused in proteomes.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

A Molecular Clock of Protein Folds

Given trees of domains structures, we defined timescales by calculating the relative age of each fold and FSF domain in the phylogenies. Time was measured by the relative number of branch splits (nodes) that exist in lineages of the phylogenomic tree in a relative 0–1 scale (nd), starting at the base of the tree (hypothetical ancestral structure) and ending in leaves (extant domain structures). The number was expressed as a fraction of the total possible number of nodes in a relative 0–1 scale, with time flowing from the origin of ancient proteins ($nd = 0$) to the origin of the most recent domain ($nd = 1$). Because trees are highly unbalanced (Wang and Caetano-Anollés 2009), nd values can quickly ''date'' a domain at each level of structural classification by defining time lines. Supplementary figure S2 (Supplementary Material online) illustrates the calculation of nd values with an example. Finally, we associated geological ages derived from fossil, geochemical, biochemical, and biomarker data to the discovery of fold domains and used it to calibrate the chronology. Comparing nd values for fold and related FSF dissected the possible effect of recruitment in protein evolution, such as the use of old folds to perform new functions or the replacement of old folds by new folds to perform a same function, and allowed to discard domains subjected, for example, to co-option events. A number of folds were linked to geological age, including folds linked to nitrogen assimilation, lipid biomarkers, bacterial and eukaryotic diversification, biological processes (e.g., biosynthesis of porphyrins and flavonoids, degradation of antibiotics, signaling, oxygen transporters) and lineages, and boundary events (origin of proteins and the present). The rationale for the assignment of geological ages to domains is described in the Materials and Methods.

Plotting fold age (nd) against geological time (in Ga) revealed a significant linear correlation ($y = -0.263x + 1.003$; $R^2 = 0.989$; $F = 1392.1$, $P < 0.0001$) (fig. 1B). This result is by itself remarkable (see rationale in the Materials and Methods) and supports the existence of a molecular clock of folds, a recurrence in molecular speciation that applies to the entire world of proteins. We note that the accuracy of the clock rests on the accuracy of geological time estimates, which depends on each and every assumption used to support molecular, physiological, paleontological, and geochemical inferences. It also depends on the accuracy of domain age assignments, which in turn rests on genomic data and phylogenetic considerations. However, our experience with growing genomic and structural data sets has shown that evolutionary patterns are consistently recovered (Caetano-Anollés et al. 2009) in spite of possible biases in the structural census, such as over- or underrepresentation of sequences and structures, definitions of fold space, and sampling limitations (Caetano-Anollés and Caetano-Anollés 2003). Given our confidence in phylogenetic statements, we do not expect marked departures from our estimates.

Plotting FSF age (nd_{FSF}) against geological time (in Ga) also revealed a significant linear correlation ($y = -0.261x + 0.947$; $R^2 = 0.956$; $F = 328.1$, $P < 0.0001$) (supplementary fig. S1, Supplementary Material online). However, data dispersion was larger, as exemplified by the wider 95% confidence belt. We reason that the poor performance of a clock of FSF rests on lower levels of deep phylogenomic signal necessarily embedded at lower levels of structural complexity (Caetano-Anollés et al. 2009). We therefore selected the molecular clock of folds for timescale applications we here report. We note that departures from a molecular clock are informative and can give clues about the accuracy of time estimates for some evolutionary events. For example, if the appearance of red algae and flavonoids complies with a molecular clock (fig. 1), then the lineage and the metabolic pathways appeared 1.3 Ga ago. This estimate is closer to predictions that consider the evolutionary appearance of chromalveolate plastids and dinoflagellate biomarkers (Porter 2004). Similarly, the departure of the time of bacterial and eukaryotic diversification is significant (fig. 1) and may indicate a general pattern for the distribution of domains in superkingdoms Bacteria and Eukarya along the branches of the trees of domain structures in which tendencies of organism diversification predate tendencies of architectural diversification. Significant molecular clock deviations should be investigated to uncover underlying evolutionary processes.

Implications of a Universal Molecular Clock of Domain Structure

Proteins change by accumulating mutations while continuing to fold into stable structures and while preserving function. The interplay between stability and function is important in molecular evolution (Caetano-Anollés and Mittenthal 2010). Experimental studies have shown that mutations in divergent lineages have only modest and additive effects on protein stability (Serrano et al. 1993) and that a substantial number of mutations promote mutational robustness and have no detectable effects on protein structure (Bloom et al. 2005). This suggests that much of protein evolution is neutral and provides strong support to the existence of a molecular clock.

In contrast with the standard molecular clock discovered by Zuckerkandl and Pauling (1962) and later elaborated by Margoliash (1963) in which the rate of change of a protein is characteristic and occurs uniformly along organism lineages (Kumar 2005), the molecular clock of folds depends on the tree of folds and applies globally to the growth of the entire protein world. This is advantageous. Standard clocks have important limitations, such as variable ''tick'' rates, biases from sequence mutational saturation, and uneven evolutionary rate variations within (the ''residual effect'') or between lineages (the ''lineage effect'') (Bromham and Penny 2003). In this regard, the molecular clock of folds appears free from rate variation of its components and from lineage effects; but very much as any molecular clock, the clock is expected to speed up and run down. Nevertheless, because of its universal nature, the

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

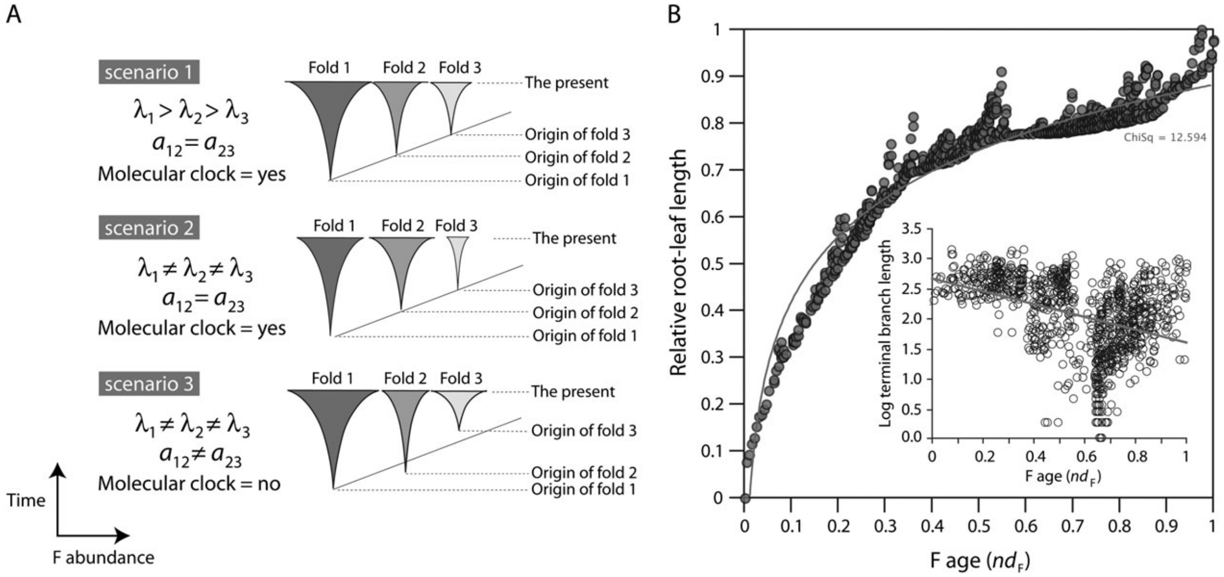

Fig. 2 The funnel paradigm and character state change along the branches of the tree of folds. (*A*) Different possible scenarios that describe the growth of fold abundance as folds are being discovered in protein evolution. Rate constants λ and a describe growth in fold abundance and in the discovery of folds, respectively. Flow of time and increases in fold abundance are indicated with arrows. Note, however, that given our model, fold abundance is also dependent on time. (*B*) Changes in root–leaf length/maximum root–leaf length along the evolutionary time line at fold level of structural organization. The inset shows character state change in the terminal leaves of the tree.

clock establishes a timescale that encompasses the entire history of life, including events that occurred in life's distant past. This has two fundamental advantages. First, it allows time inferences by interpolation in regions of the timescale that date back to times where lineages (as we know them) were inexistent and where fossil records are unavailable. In other words, the clock permits digging deep into origins of biochemistry and life. Second, because structure is linked to molecular function, the discovery of functions can be placed in a global timescale and linked to crucial innovations at other levels of organization (cellular, organismal, etc.). This new ability of the universal clock will be invaluable for evolutionary and functional genomics and for paleobiology.

A molecular clock of folds also informs about structural evolution. Comparison of fold and FSF phylogenies reveals that individual fold domains represent collections of FSFs of different ages, some being ancient and others of recent origin (Wang et al. 2006). A general dynamic model of evolution in which folds evolve in a stochastic branching process describes the multidimensional growth in abundance of domain structures (*G*, their popularity) in the world of proteins and supports these results (Boca 2006). The model behind our tree reconstruction exercise is driven by the accumulation of successful architectural variants within a structural neighborhood, very much as propagation in Galton–Watson branching processes responds to a delicate balance of survival and extinction (Harris 1963). Consequently, protein evolution is driven by the success (fitness) of architectures in the protein world (and later in the organismal world) and not by structural transformations reflecting individual sequence-to-structure mappings (Wang et al. 2006). A "funnel" metaphor helps illustrate the evolution of folds and the patterns of recent and ancient origin (fig. 2*A*). Once discovered, a fold will initiate a funnel of roughly exponential growth in the number of its variants, and in cases where the fold diversifies into different FSF architectures, these will also produce exponentially growing subfunnels within the fold funnel. As time progresses, *G* increases exponentially. The time course of changes in abundance for the *k*th fold, G_k, depends on transition probabilities per unit time. Two sets of these rate constants are parameters of the model—λs, the rate constants of growth of folds when sequence and structural variants are generated by mutation, and *a*s, the rate constants for transition from one fold to the next as protein evolution proceeds and new folds get discovered.

Our molecular clock shows that folds appear in the tree at times that are proportional to nd, that is, the rate of discovery of folds is proportional to the rate of molecular speciation (cladogenesis, i.e., the generation of splits in a tree) but not necessarily to the rate of growth of the fold. Under one scenario, λ values of ancient folds are always larger than λ values of recent folds, with *a* being equal (scenario 1; fig. 2*A*). In an alternative but related scenario, λ are allowed to take any value (scenario 2; fig. 2*A*). *k* defines the fold as folds appear in the time line of fold discovery. In both scenarios, $G_k = \exp \lambda_k(t - t_k)$; *t* is present time, t_k is specific to each fold, and for the specific example, $k = 1$–3. In the tree of folds, a number of character state changes along lineages from the root of the tree to the leaves (root–leaf length) expressed in a relative 0–1 scale appear to follow a log pattern, suggesting that change in fold popularity indeed follows an exponential trend (fig. 2*B*). Similarly and as suggested in scenario 1, a plot

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

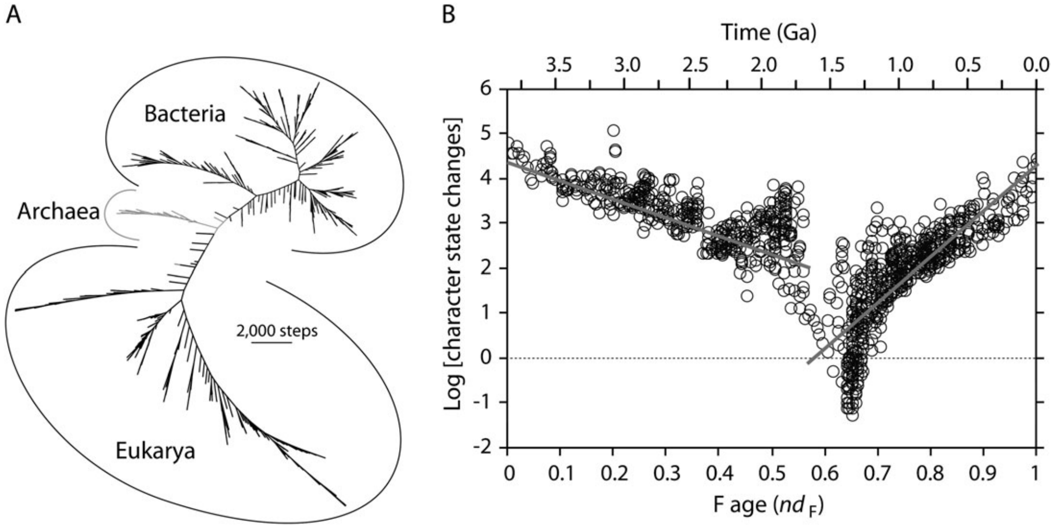

Fig. 3 Reconstructing character state changes in the branches of a universal tree of proteomes. (*A*) Phylogenomic tree of proteomes (371,239 steps; CI = 0.054, RI = 0.782; $g_1 = -0.146$). (*B*) Plot describing character state change in the tree of proteomes for every fold, with folds ordered according to their age. Regression lines show trends in two different phases of protein evolution. Model summary for nd < 0.56: log $y = -3.356x + 4.375$; $R^2 = 0.542$; $F = 517.1$, $P < 0.0001$). Model summary for nd > 0.56: log $y = -10.182x - 5.860$; $R^2 = 0.659$; $F = 1{,}100.2$, $P < 0.0001$).

of the logarithm of terminal branch length versus nd shows a significant decrease ($F = 155$, $P < 0.0001$) (fig. 2*B*, inset). However, the correlation is poorly supported ($R^2 = 0.134$; $P < 0.0001$).

Other scenarios are possible if funnels develop slowly or fast and if *a*s are different (scenario 3; fig. 2*A*). In some of these cases, there could be overdispersion of the molecular clock, that is, the ticking of the clock will not always be constant. To explore behaviors, we studied character state change in our data set (fig. 3). We first used the fold census to generate a tree of proteomes (fig. 3*A*). In this case, characters are folds and taxa are organisms. As expected (Caetano-Anollés et al. 2009), the tree reveals the three superkingdoms of life and groups some organisms according to established classification. We then traced character state change for each fold along the (organism) lineages of the tree, adjusting values to genome abundance levels (see Materials and Methods). Finally, we plotted the logarithm of overall change in the tree against the nd of the folds (fig. 4*B*). Remarkably, two behaviors compatible with scenario 2 are clearly evident. At nd < 0.56, character state change decreases exponentially with fold age, supporting our model. However, at nd > 0.56, the trend reverses and change increases with age. The cusp coincides with the appearance of Eukarya in the time line 1.7 Ga ago (fig. 1).

The biphasic behavior we observe that is apparently triggered by the rise of Eukarya is not new. We observed this trend in evolution of folds and FSFs architectures in eukaryotic genomes following clear reductive evolutionary trends (Wang et al. 2007), a pattern that was enhanced by the combination of domains in proteins (Wang and Caetano-Anollés 2009). It is possible that increases in λ during this second evolutionary phase may result from the increased capacity of Eukarya to encode genes and domain architectures as organisms abandon the microscopic realm of the microbial world. More likely is the enhanced ability of eukaryotic organisms (especially metazoa) to rearrange domains in proteins and the higher representations of these domains in their genomes, which increases genomic abundance of participating folds (Wang and Caetano-Anollés 2009). These two factors would set in motion a trend of proteomic increase while obeying the clocklike discovery

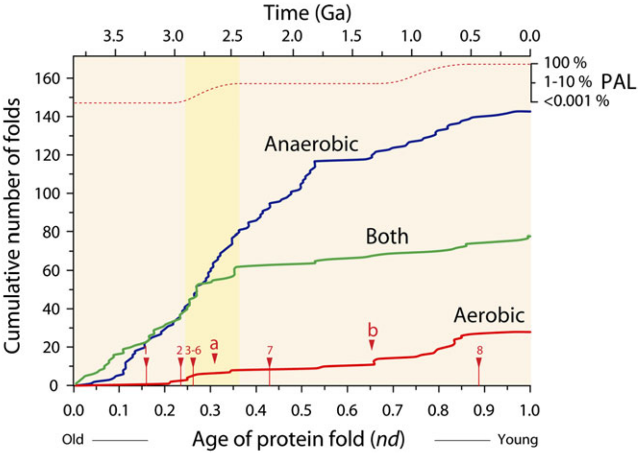

Fig. 4 Accumulation of fold domains linked to aerobic and anaerobic metabolism and those shared by both metabolisms in the evolutionary time line. Oxygen levels are indicated (Sessions et al. 2009) together with folds linked to initial anaerobic reactants for aerobic modules (arrowheads: 1, 4-hydroxyphenylacetate for Tyr metabolism; 2, L-tryptophan for Trp metabolism; 3, ent-kaurene for diterpenoid biosynthesis; 4, 4-coumarate and *trans*-cinnamate for phenylpropanoid biosynthesis; 5, squalene for steroid biosynthesis; 6, (−)-limonene for limonene and pinene degradation; 7, phenylpyruvate for Phe metabolism; 8, urate for purine metabolism). The start of diversification of organisms (a) and algae (b) are indicated.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

of folds and the exponential growth in fold abundance in the funnels.

One additional noteworthy conclusion from the constraints imposed by a molecular clock is that the availability of niches in which domain architectures can diversify and survive is comparable for all folds, whenever they originate. Because the same patterns are found in evolution of FSF (data not shown), this same conclusion can be extended to all domains. We caution however that the linear plots of figure 1*B* and supplementary figure S1 (Supplementary Material online) are based on relatively few molecular fossils, and it is possible that a wider sampling would show points deviant from the straight line. We plan to study possible clock overdispersions in the near future (responsible for scenario 3; fig. 2*A*), especially because overdispersion can arise from fluctuations in protein structural stability (Bloom et al. 2005).

The mapping of mutating protein sequences into structures defines an evolutionary dynamic paradigm of protein change in which ensembles of sequences that fold into a given native fold structure delimit "neutral" sets (reviewed in Caetano-Anollés et al. 2009). The stochastic process of mutation causes sequences to drift along these sets (neutral networks, sequences linked by a series of single point mutations). These evolutionary trajectories are, however, tailored by thermodynamic and kinetic folding optimality, keeping sequences within sequence space attractors for individual folds. However, escapees from the nets occur, and our results suggest that these escapes occur at constant rate.

The key prediction of the neutral theory, the idea that most change is stochastic, is that given a constant mutation rate, a set of two protein sequences will diverge proportionally to the time of their divergence (Kimura 1983). The molecular clock of sequences (Zuckerkandl and Pauling 1962) constitutes the most important pillar supporting the neutral theory. There is, however, limited understanding of how sequence relates to structure (Schuster 2010) and what is the role of thermodynamics in the process (Bloom et al. 2005). The observation that there is indeed a molecular clock of domain structure now shows that the discovery of new structures in sequence space occurs at extraordinarily constant pace and over the entire timescale of life. The impact that the mutational diffusion of sequences in sequence space has on the structure of proteins (Caetano-Anollés et al. 2010) should be therefore regarded to be mostly stochastic, supporting and extending the neutral theory of evolution to higher levels of structural organization.

The Emergence of Aerobic Metabolism

We tested the power of the molecular clock of folds in a study of innovations and history of structures linked to aerobic metabolism. Because biochemistry coevolves with geochemistry (Saito et al. 2003; Williams and Fraústo Da Silva 2003; Dupont et al. 2006, 2010), fold domains are likely to record important geochemical events. This speculation is preliminarily supported by recent investigations of metalloprotein evolution that showed that fold domain history reflects the bioavailability of metals in the geochemical record (Dupont et al. 2006, 2010; Ji et al. 2009). For example, the earliest manganese and heme iron protein folds precede copper counterparts (Ji et al. 2009; Dupont et al. 2010), depicting geochemical evidence of manganese and iron being bioavailable on anaerobic Earth, whereas copper was being restricted under oxygen limitation (Saito et al. 2003; Williams and Fraústo Da Silva 2003; Dupont et al. 2006). Remarkably, the most ancient function of copper enzymes was the reduction of oxygen to water in the respiratory chain (enzymatic activity EC 1.9.3.1) (Ji et al. 2009). This suggests that the appearance of the first copper protein fold (b.6) at nd = 0.271 was tightly coupled with the emergence of aerobic respiration. Using the molecular clock of folds, we show by interpolation that the aerobic respiration–linked b.6 domain appeared 2.8 Ga ago and that the antioxidant defense system–linked oxygen radical–scavenging Fe, Mn SOD C-terminal domain (d.44; nd = 0.443) appeared much later (2.2 Ga). This suggests that aerobic respiration originated primordially 0.35 Ga before the Great Oxidation Event (GOE) that occurred 2.45 Ga ago (Sessions et al. 2009) and evolved gradually.

Because aerobic metabolism preceded aerobic respiration, we explored innovations in metabolites, enzymatic activities, and enzymes associated with aerobic metabolism, and we then used the molecular clock of folds to trace metabolic events linked to planet oxygenation. We assume that the discovery of domain structures in crucial enzymes was closely linked to oxygen-facilitated evolution of life and planet oxygenation. We were therefore interested in fold domains coupled with the first appearance of aerobic biosynthesis and the discovery of domains coupled with or sensitive to the rise of oxygen. We first identified 1,145 anaerobic and 454 aerobic enzymatic activities (~40% of aerobic activities used oxygen explicitly) in the simulated metabolic networks of Raymond and Segrè (2006). Out of these, 650 anaerobic and 223 aerobic enzymes (or in selected cases, closely related homologues) had PDB structural records. These enzymes defined 1,064 and 379 domains and 224 and 110 folds, respectively. A comparison of the fold structures revealed that aerobic enzymes made use of 31 folds exclusively (table 1), most of which (90%) use oxygen explicitly (some shown in table 2), indicating a functional selection of the unique architectures. These 31 fold domains were discovered quite late, only after the first 81 folds appeared in protein evolution. Accumulation of fold domains in the time line showed that whereas anaerobic domains accumulated steadily, aerobic domains appeared at nd > 0.2 (fig. 3). A substantial fraction of domains used by aerobic and anaerobic enzymes appear early, suggesting that aerobic enzymes also recruited ancient anaerobic domains.

An analysis of metabolic network structure reveals the existence of hierarchical modules of metabolic reactions in KEGG aerobic and anaerobic pathways, with some major modules being predominant in the networks. For instance, 19 major modules account for 80.8% of aerobic metabolites (supplementary table S1, Supplementary Material

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

Table 1. The Identity and Evolutionary Age of Fold Domains Exclusively Used by Aerobic Enzymes in Simulated Metabolic Networks.

Folds	nd Value of Fold
7-bladed beta-propeller (b.69)	0.203
Spectrin repeat-like (a.7)	0.208
Split barrel-like (b.45)	0.245
FAD-binding domain (d.145)[a]	0.245
Ferredoxin reductase-like, C-terminal NADP-linked domain (c.25)	0.255
Cupredoxin-like (b.6)	0.271
FMN-dependent nitroreductase-like (d.90)	0.339
Lipocalins (b.60)	0.526
Glutamyl tRNA-reductase dimerization domain (a.151)	0.547
Methylamine dehydrogenase, L chain (g.21)	0.656
Monooxygenase (hydroxylase) regulatory protein (d.137)	0.656
LigA subunit of an aromatic-ring-opening dioxygenase LigAB (a.88)	0.656
Hemocyanin, N-terminal domain (a.85)	0.667
Nitric oxide (NO) synthase oxygenase domain (d.174)	0.744
Coproporphyrinogen III oxidase (d.248)	0.760
Indolic compounds 2,3-dioxygenase-like (a.266)	0.786
Di-copper centre-containing domain (a.86)	0.792
Somatomedin B domain (g.64)	0.797
TAZ domain (g.53)	0.813
Nucleoplasmin-like/VP (viral coat and capsid proteins) (b.121)	0.828
Sterol carrier protein, SCP (d.106)	0.833
Heme-dependent catalase-like (e.5)	0.833
His-Me finger endonucleases (d.4)	0.833
Serpins (e.1)	0.849
Oxidoreductase molybdopterin-binding domain (d.176)	0.850
Heme-dependent peroxidases (a.93)	0.875
Cytochrome P450 (a.104)	0.958
Sulfite reductase hemoprotein (SiRHP), domains 2 and 4 (d.134)	—
E2 regulatory, transactivation domain (b.91)	—
C-terminal domain of mollusc hemocyanin (b.112)	—
Papillomavirus e6-interacting peptide of e6ap (j.74)	—

[a] D-lactate dehydrogenase covered by d.145 fold was defined as an anaerobic enzyme in metabolic simulations (Raymond and Segrè 2006). However, this enzyme uses oxygen explicitly (table 2).

online), in which 15 start from anaerobic metabolites (supplementary table S2, Supplementary Material online). The enzymes for the biosynthesis of these initial anaerobic reactants started to appear at nd approximately 0.2 (fig. 3) (supplementary table 2), except for those used in synthesis of urate and phenylpyruvate. Interestingly, urate biosynthesis is an anaerobic pathway that has been rewired late in evolution to take advantage of oxygen (Raymond and Segrè 2006). Whether the phenylpyruvate biosynthetic pathway suffered a similar fate remains to be determined.

Aerobic Metabolism, Planet Oxygenation, and the Diversification of Life

Only b.45 and d.145 folds (nd = 0.245), out of the most ancient aerobic folds (table 1), are rich in aerobic enzymes that use oxygen explicitly (table 2) and had FSFs that appeared early (table 3). Thus, b.45 and d.145 were clearly coupled with the emergence of aerobic metabolism, an inference that was further supported by the burst of discovery of aerobic enzymes that use oxygen explicitly and followed the appearance of these folds (fig. 3 and table 2). As expected, these folds preceded b.6 (the first copper protein fold) in the time line. Thus, aerobic biosynthesis appeared earlier than aerobic respiration. This matches the expected progression of oxygen levels in the planet (Saito 2009; Sessions et al. 2009), especially because experimental observations indicate that oxygen-dependent biosynthesis occurs at the very low oxygen concentration of approximately 0.1% of PAL, whereas aerobic respiration requires higher levels (~1% PAL) (Chapman and Schopf 1983). These higher levels were generally considered to have been reached 2.45 Ga ago during the GOE (Saito 2009; Sessions et al. 2009). Using our molecular clock, the age of the b.45 and d.145 fold domains tracked the early stages of planet oxygenation. Assuming life (and proteins) originated 3.8 Ga ago (Mojzsis et al. 1996), interpolation of data for both folds established a possible time of emergence of aerobic metabolism (~2.9 Ga) and a time frame of about 400 My for the rise of oxygen from 0.1% to 1% PAL (fig. 1*B*). This inference agrees with the opinion that oxygenic photosynthesis appeared at least 400 My prior to the GOE, that is, 2.9 Ga (Claire et al. 2006; Nisbet et al. 2007), and that atmospheric oxygen was greater than 0.01% PAL between 2.8 and 3.0 Ga (Holland 2006). Taken together, observations suggest that GOE was a gradual event.

The distribution of domain architectures among superkingdoms of life is remarkably consistent (Wang et al. 2007; Caetano-Anollés et al. 2009; Wang and Caetano-Anollés 2009). The most ancient folds or FSFs are universally present in all organisms. With time, folds or FSFs are first lost in primordial archaeal lineages and then in eukaryal and bacterial lineages. In turn, the rather late gain of Bacteria-specific, and then, Eukarya-specific and Archaea-specific, structures signal the emergence of superkingdoms. These same patterns of diversification were also observed in the diversification of ancient RNA molecules such as transfer RNA, 5S ribosomal RNA, and ribonuclease P RNA, with RNA substructures specific to Archaea appearing before substructures specific to other superkingdoms (Sun and Caetano-Anollés 2008, 2009, 2010). These patterns highlight three evolutionary epochs (Wang et al. 2007): epoch 1, an ancient "architectural diversification" period in which ancient molecules emerged and diversified and proteomes were highly similar to each other, with archaeal lineages reducing their complements by domain loss toward the end; epoch 2, a "superkingdom specification" period in which molecules sorted in emerging organismal lineages and some became specific to emerging superkingdoms; and epoch 3, a late "organismal diversification" period in which molecular lineages diversified in an increasingly diverse tripartite world and notable proteome expansions occurred in Eukarya. Using these epoch definitions, we established boundaries in the phylogenomic tree (fig. 1) and used the first fold lost in Archaea, the release factor fold (e.38; nd = 0.318), to determine an upper boundary for the start of diversified life (2.6 Ga). This inference is consistent with the recent proposal that major diversification of lineages in Archaea began 2.8 Ga ago (Blank 2009).

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

Table 2. Some Folds Rich in Enzymes That Use Oxygen Explicitly.

Fold	nd Value of Fold	Enzyme	EC Number	Source
Split barrel-like (b.45)	0.245	Pyridoxal 5′-phosphate synthase	1.4.3.5	KEGG[i]
		Pyridoxal-dependent decarboxylase[a]	4.1.1.-	KEGG[i]
FAD-binding domain (d.145)	0.245	Cholesterol oxidase[b]	1.1.3.6	KEGG[i]
		Aryl-alcohol oxidase	1.1.3.7	KEGG[i]
		Alcohol oxidase	1.1.3.13	KEGG[i]
		Vanillyl-alcohol oxidase	1.1.3.38	KEGG[i]
		Cytokinin dehydrogenase	1.5.99.12	Brenda[j]
		D-Lactate dehydrogenase	1.1.1.28	van Berkel WJ et al. 2006
		p-Cresol methylhydroxylase	1.17.99.1	Dym O et al. 2000
		Xanthine dehydrogenase	1.17.1.4	Brenda[j]
		Xanthine oxidase	1.17.3.2	KEGG[i]
Acyl-CoA dehydrogenase NM domain-like (e.6)	0.250	Acyl-CoA oxidase	1.3.3.6	KEGG[i]
		Butyryl-CoA dehydrogenase	1.3.99.2	Brenda[i]
		Acyl-CoA dehydrogenase	1.3.99.3	Brenda[i]
Glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase (d.32)	0.250	Catechol 2,3-dioxygenase	1.13.11.2	KEGG[i]
		3,4-Dihydroxyphenylacetate 2,3-dioxygenase	1.13.11.15	KEGG[i]
		4-Hydroxyphenylpyruvate dioxygenase	1.13.11.27	KEGG[i]
		Biphenyl-2,3-diol 1,2-dioxygenase	1.13.11.39	KEGG[i]
		Pyridoxal-dependent decarboxylase[c]	4.1.1.-	KEGG[i]
Ferredoxin reductase-like, C-terminal NADP-linked domain (c.25)	0.255	Dihydroorotate oxidase	1.3.3.1	KEGG[i]
		Cytochrome-b5 reductase	1.6.2.2	Brenda[i]
		NADPH-hemoprotein reductase	1.6.2.4	Brenda[i]
		Methane monooxygenase	1.14.13.25	KEGG[i]
		Nitric oxide synthase	1.14.13.39	KEGG[i]
FAD-linked reductases, C-terminal domain (d.16)	0.260	Glucose oxidase	1.1.3.4	KEGG[i]
		Cholesterol oxidase[d]	1.1.3.6	KEGG[i]
		Pyranose oxidase	1.1.3.10	KEGG[i]
		Cellobiose dehydrogenase (acceptor)	1.1.99.18	KEGG[i]
		Protoporphyrinogen oxidase	1.3.3.4	KEGG[i]
		L-Amino-acid oxidase	1.4.3.2	KEGG[i]
		D-Amino-acid oxidase[e]	1.4.3.3	KEGG[i]
		Amine oxidase (flavin-containing)[f]	1.4.3.4	KEGG[i]
		Sarcosine oxidase	1.5.3.1	KEGG[i]
		Dimethylglycine oxidase	1.5.3.10	KEGG[i]
		Polyamine oxidase	1.5.3.11	KEGG[i]
		4-Hydroxybenzoate 3-monooxygenase	1.14.13.2	KEGG[i]
		Phenol 2-monooxygenase	1.14.13.7	KEGG[i]
Cystatin-like (d.17)	0.266	Amine oxidase (flavin-containing)[g]	1.4.3.4	KEGG[i]
		Amine oxidase (copper-containing)	1.4.3.6	Brenda[j]
		Protein-lysine 6-oxidase	1.4.3.13	KEGG[i]
		Naphthalene 1,2-dioxygenase	1.14.12.12	KEGG[i]
		Biphenyl 2,3-dioxygenase	1.14.12.18	KEGG[i]
Nucleotide-binding domain (c.4)	0.266	D-Amino-acid oxidase[h]	1.4.3.3	KEGG[i]

[a] 3-Octaprenyl-4-hydroxybenzoate carboxy-lyase UbiD domain.
[b] Domain of FAD-linked oxidases, N-terminal domain family.
[c] Glyoxalase I (lactoylglutathione lyase) domain.
[d] Domain of cholesterol oxidase family.
[e] D-Aminoacid oxidase domain.
[f] Monoamine oxidase B domain.
[g] Copper amine oxidase, domains 1 and 2.
[h] D-Aminoacid oxidase, N-terminal domain.
[i] Kanehisa and Goto (2000).
[j] Chang et al. (2009).

Remarkably, this boundary falls within the approximately 400 My span of oxygen rise. This result has important implications for understanding the relationship between biological diversification and oxygen. For example, metabolic changes induced by planet oxygenation may have had an important role in the generation of diversified lineages by providing, for example, selective advantages to organisms that adapted to the new oxygenic environments of Earth, perhaps crucially splitting the world of the universal common cellular ancestor into an heterogeneous ensemble.

Conclusions

We here describe for the first time the existence of a molecular clock of folds, a recurrence in the discovery of new structures in the world of proteins. Standard molecular clocks based on protein or nucleic acid sequence have

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

Table 3. The Ages of b.69, a.7, b.45, and d.145 Fold Domains and Their Associated FSFs.

Fold	nd Value of Fold Domain	nd Value of FSF Domain
7-bladed beta-propeller (b.69)	0.203	0.834 (b.69.2) 0.883 (b.69.1)
Spectrin repeat-like (a.7)	0.208	0.291 (a.7.3)
Split barrel-like (b.45)	0.245	0.206 (b.45.1)
FAD-binding domain (d.145)	0.245	0.247 (d.145.1)

revolutionized biology by providing a framework for timescales of population and species, gene families, and sequence variations (Doolittle et al. 1996; Kumar 2005). Once clocks have been calibrated, speciation events can be inferred even in the absence of fossils, biomarkers, or biogeographical records. Our clock now extends the standard concept that was specific to molecules and their timescales and turns it into a universal timescale-generating tool. To illustrate its power, we explored the process of planet oxygenation by identifying protein domains linked to aerobic metabolism along time lines of domain discovery. The study reveals that aerobic metabolism emerged during architectural diversification (epoch 1) and at the end of the Archean eon. Its inception and later expansion increased oxygen gradually from 0.1% to 1% PAL. During this period, enzymes recruited old and new folds for oxygen-dependent metabolic networks, as new aerobic enzymes brought innovations in reaction types (e.g., oxygen-dependent hydroxylation), metabolites, and metabolic pathways. This planetary phenomenon generated turmoil, perhaps triggering the emergence of the first diversified lineages of Earth.

Supplementary Material

Supplementary figs. S1 and S2 and supplementary tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Christopher Dupont (J. Craig Venter Institute) for comments and suggestions. This study was supported by the National Basic Research Program of China (973 project, grant 2010CB126100), the National Natural Science Foundation of China (grant 30870520), and Outstanding Youth Foundation of Shandong Province (grant JQ200812) (all to H.Y.Z.) and by the National Science Foundation (grant MCB-0749836 to G.C.A.). Any opinions, findings, and conclusions and recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Anbar AD, Duan Y, Lyons TW, et al. 2007. A whiff of oxygen before the great oxidation event? *Science* 317(5846):1903–1906.

Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36(Database issue):D419–D425.

Bengtson S, Belivanova V, Rasmussen B, Whitehouse M. 2009. The controversial "Cambrian" fossils of the Vindhyan are real but more than a billion years older. *Proc Natl Acad Sci U S A.* 106(19): 7729–7734.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28(1):235–242.

Bhattacharya D, Yoon HS, Hedges SB, Hackett JD. 2009. Eukaryotes (Eukaryota). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 116–120.

Blank CE. 2009. Not so old Archaea—the antiquity of biogeochemical processes in the archaeal domain of life. *Geobiology* 7(5):495–514.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci U S A.* 102(3):606–611.

Boca SM. 2006. A dynamical model for the evolution of protein folds [mathematics honors thesis]. [Urbana (IL)]: University of Illinois.

Brocks JJ, Logan GA, Buick R, Summons RE. 1999. Archean molecular fossils and the early rise of eukaryotes. *Science* 285(5430): 1033–1036.

Bromham L, Penny D. 2003. The modern molecular clock. *Nat Rev Genet.* 4(3):216–224.

Butterfield NJ. 2000. Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* 26(3):386–404.

Caetano-Anollés G, Caetano-Anollés D. 2003. An evolutionarily structured universe of protein architecture. *Genome Res.* 13(7): 1563–1571.

Caetano-Anollés G, Mittenthal JE. 2010. Exploring the interplay of stability and function in protein evolution. *BioEssays* 32(8): 655–658.

Caetano-Anollés G, Kim HS, Mittenthal JE. 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci U S A.* 104(22): 9358–9363.

Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. 2009. The origin, evolution and structure of the protein world. *Biochem J.* 417(3):621–637.

Caetano-Anollés G, Yafremava LS, Mittenthal JE. 2010. Modularity and dissipation in evolution of macromolecular structures, functions, and networks. In: Caetano-Anollés G, editor. Evolutionary genomics and systems biology. Hoboken (NJ): Wiley-Blackwell. p. 431–450.

Catling DC, Glein CR, Zahnle KJ, McKay CP. 2005. Why O2 is required by complex life on habitable planets and the concept of planetary "oxygenation time". *Astrobiology* 5(3): 415–438.

Chang A, Scheer M, Grote A, Schomburg I, 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.

Chapman DJ, Schopf JW. 1983. Biological and biochemical effects of the development of an aerobic environment. In: Schopf JW, editor. Earth's earliest biosphere: its origin and evolution. Princeton (NJ): Princeton University Press. p. 302–320 and references therein.

Chen LL, Wang GZ, Zhang HY. 2007. Sterol biosynthesis and prokaryotes-to-eukaryotes evolution. *Biochem Biophys Res Commun.* 363(4):885–888.

Chothia C, Gough J. 2009. Genomic and structural aspects of protein evolution. *Biochem J.* 419(1):15–28.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

Claire MW, Catling DC, Zahnle KJ. 2006. Biogeochemical modelling of the rise in atmospheric oxygen. *Geobiology* 4(4):239–269.

Copley RR, Bork P. 2000. Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J Mol Biol.* 303:627–640.

de Pinna MCC. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7:367–394.

Doolittle RF. 2005. Evolutionary aspects of whole-genome biology. *Curr Opin Struct Biol.* 15(3):248–253.

Doolittle RF, Feng DF, Tsang S, Cho G, Little E. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271(5248):470–477.

Dupont CL, Yang S, Palenik B, Bourne PE. 2006. Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc Natl Acad Sci U S A.* 103(47):17822–17827.

Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anollés G. 2010. History of biological metal utilization inferred through phylogenomic analysis of protein structure. *Proc Natl Acad Sci U S A.* 107:10567–10572.

Dym O, Pratt EA, Ho C, Eisenberg D. 2000. The crystal structure of D-lactate dehydrogenase, a peripheral membrane respiratory enzyme. *Proc Natl Acad Sci USA.* 97(17):9413–9418.

Falkowski PG, Isozaki Y. 2008. The story of O_2. *Science* 322(5901):540–542.

Feng DF, Cho G, Doolittle RF. 1997. Determining divergence times with a protein clock: update and reevaluation. *Proc Natl Acad Sci U S A.* 94(24):13028–13033.

Fischer WW. 2008. Life before the rise of oxygen. *Nature* 455: 1051–1052.

Forslund K, Henricson A, Hollich V, Sonnhammer E. 2008. Domain tree-based analysis of protein architecture evolution. *Mol Biol Evol.* 25:254–264.

Garvin J, Buick R, Anbar AD, Arnold GL, Kaufman AJ. 2009. Isotopic evidence for an aerobic nitrogen cycle in the latest Archean. *Science* 323(5917):1045–1048.

Glass JB, Wolfe-Simon F, Anbar AD. 2009. Coevolution of metal availability and nitrogen assimilation in cyanobacteria and algae. *Geobiology* 7(2):100–123.

Godfrey LV, Falkowski PG. 2009. The cycling and redox state of nitrogen in the Archaean ocean. *Nat Geosci.* 2(10): 725–729.

Gough J. 2005. Convergent evolution of domain structures (is rare). *Bioinformatics* 21:1464–1471.

Gough J, Karplus K, Hughey R, 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(4):903–919.

Gourier D, Delpoux O, Bonduelle A, Binet L, Ciofini I, Vezin H. 2010. EPR, ENDOR, and HYSCORE study of the structure and the stability of vanadyl-porphyrin complexes encapsulated in silica: potential paramagnetic biomarkers for the origin of life. *J Phys Chem B.* 114(10):3714–3725.

Hall BG, Barlow M. 2004. Evolution of the serine β-lactamases: past, present and future. *Drug Resist Updat.* 7(2):111–123.

Harris TA. 1963. The theory of branching processes. New York: Dover Publications.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22(23):2971–2972.

Hermann TN, Podkovyrov VN. 2008. On the nature of the Precambrian microfossils Arctacellularia and Glomovertella. *Paleontol J.* 42(6):655–664.

Hoashi M, David C, Bevacqua DC, Otake T, Watanabe Y, Hickman AH, Utsunomiya S, Ohmoto H. 2009. Primary haematite formation in an oxygenated sea 3.46 billion years ago. *Nat Geosci.* 2(4):301–306.

Hofmann HJ. 1976. Precambrian microflora, Belcher Islands, Canada: significance and systematics. *J Paleontol.* 50(6):1040–1073.

Holland HD. 2006. The oxygenation of the atmosphere and oceans. *Philos Trans R Soc Lond B Biol Sci.* 361(1470):903–915.

Holland LZ, Albalat R, Azumi K, et al. 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* 18(7):1100–1111.

Javaux EJ, Knoll AH, Walter MR. 2004. TEM evidence for eukaryotic diversity in mid-Proterozoic oceans. *Geobiology* 2(3):121–132.

Jez JM, Bowman ME, Dixon RA, Noel JP. 2000. Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase. *Nat Struct Biol.* 7:786–791.

Ji HF, Chen L, Jiang YY, Zhang HY. 2009. Evolutionary formation of new protein folds is linked to metallic cofactor recruitment. *BioEssays* 31(9):975–980.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28(1):27–30.

Kato Y, Suzuki K, Nakamura K, Hickman Ah, Nedachi M, Kusakabe M, Bevacqua DC, Ohmoto H. 2009. Hematite formation by oxygenated groundwater more than 2.76 billion years ago. *Earth Planet Sci Lett.* 278(1–2):40–49.

Kaufman AJ, Johnston DT, Farquhar J, et al. 2007. Late Archean biospheric oxygenation and atmospheric evolution. *Science* 317(5846):1900–1903.

Kim HS, Mittenthal JE, Caetano-Anollés G. 2006. MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics* 7:351.

Kim KM, Caetano-Anollés G. 2010. Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Mol Biol Evol.* 27(7):1710–1733.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kirkpatrick M, Slatkin M. 1993. Searching for evolutionary patterns in the shape of phylogenetic trees. *Evolution* 47:1171–1181.

Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet.* 6(8):654–662.

Lane DP, Cheok CF, Brown C, Madhumalar A, Ghadessy FJ, Verma C. 2010. Mdm2 and p53 are highly conserved from placozoans to man. *Cell Cycle.* 9(3):1–8.

Levitt M. 2007. Growth of novel protein structural data. *Proc Natl Acad Sci U S A.* 104:3183–3188.

Ma BG, Chen L, Ji HF, et al. 2008. Characters of very ancient proteins. *Biochem Biophys Res Commun.* 366(3):607–611.

McKenzie A, Steel M . 2000. Distributions of cherries for two models of trees. *Math Biosci.* 164:81–92.

Margoliash E. 1963. Primary structure and evolution of cytochrome c. *Proc Natl Acad Sci U S A.* 50(4):672–679.

Michael EB, David JC. 2009. 3D model of amphioxus steroid receptor complexed with estradiol. *Biochem Biophys Res Commun.* 386(3):516–520.

Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP, Friend CR. 1996. Evidence for life on Earth before 3,800 million years ago. *Nature* 384(6604):55–59.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.* 247(4):536–540.

Nagano N, Orengo CA, Thornton JM. 2000. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol.* 321:741–765.

Nisbet EG, Grassineau NV, Howe CJ, Abell PI, Regelous M, Nisbet RER. 2007. The age of Rubisco: the evolution of oxygenic photosynthesis. *Geobiology* 5(4):311–335.

Nixon KC. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15(4):407–414.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018

Payne JL, Boyer AG, Brown JH, 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 U S A*. 106(1):24–27.

Porter CT, Bartlett GJ, Thornton JM. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*. 32(Database issue): D129–D133.

Porter SM. 2004. The fossil record of early eukaryotic diversification. *Paleontol Soc Paper*. 10:35–50.

Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR. 2008. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* 455(7216):1101–1104.

Raymond J, Segrè D. 2006. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* 311(5768): 1764–1767.

Römpler H, Stäubert C, Thor D, Schulz A, Hofreiter M, Schöneberg T. 2007. G protein-coupled time travel: evolutionary aspects of GPCR research. *Mol Interv*. 7(1):17–25.

Saito MA. 2009. Less nickel for more oxygen. *Nature* 458(7239):714–715.

Saito MA, Sigman DM, Morel FMM. 2003. The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean-Proterozoic boundary? *Inorg Chim Acta*. 356: 308–318.

Schuster P. 2010. Genotypes and phenotypes in the evolution of molecules. In: Caetano-Anollés G, editor. Evolutionary genomics and systems biology. Hoboken (NJ): Wiley-Blackwell. p. 123–152.

Serrano LAG, Day G, Fersht AR. 1993. Step-wise mutation of barnase to binase: a procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J Mol Biol*. 233(2):305–312.

Sessions AL, Doughty DM, Welander PV, Summons RE, Newman DK. 2009. The continuing puzzle of the great oxidation event. *Curr Biol*. 19(14):R567–R574.

Summons RE, Bradley AS, Jahnke LL, Waldbauer JR. 2006. Steroids, triterpenoids and molecular oxygen. *Phil Trans R Soc B*. 361(1470):951–968.

Sun F-J, Caetano-Anollés G. 2008. The origin and evolution of tRNA inferred from phylogenetic analysis of structure. *J Mol Biol*. 66:1–35.

Sun F-J, Caetano-Anollés G. 2009. The evolutionary history of the structure of 5S ribosomal RNA. *J Mol Evol*. 69:430–443.

Sun F-J, Caetano-Anollés G. 2010. The ancient history of the structure of ribonuclease P and the early origins of Archaea. *BMC Bioinformatics*. 11:153.

Swofford DL. 2002. Phylogenetic analysis using parsimony and other programs (PAUP*), version 4. Sunderland (MA): Sinauer Associates.

Theobald DL. 2010. A formal test of the theory of common ancestry. *Nature* 465:219–222.

Thornton JW. 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci U S A*. 98(10):5671–5676.

Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. 2006. The evolutionary diversification of cyanobacteria: molecular–phylogenetic and paleontological perspectives. *Proc Natl Acad Sci U S A*. 103(14):5442–5447.

van Berkel WJ, Kamerbeek NM, Fraaije MW. 2006. Flavoprotein monooxygenases, a diverse class of oxidative biocatalysts. *J Biotechnol*. 124(4):670–689.

van Holde KE, Miller KI, Decker H. 2001. Hemocyanins and invertebrate evolution. *J Biol Chem*. 276(19):15563–15566.

Ventura GT, Kenig F, Reddy CM, Schieber J, Frysinger GS, Nelson RK, Dinel E, Gaines RB, Schaeffer P. 2007. Molecular evidence of Late Archean archaea and the presence of a subsurface hydrothermal biosphere. *Proc Natl Acad Sci U S A*. 104(36):14260–14265.

Wang M, Boca SM, Kalelkar R, Mittenthal JE, Caetano-Anollés G. 2006. A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. *Complexity* 12(1):27–40.

Wang M, Caetano-Anollés G. 2009. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* 17(1):66–78.

Wang M, Yafremava LS, Caetano-Anolles D, Mittenthal JE, Caetano-Anolles G. 2007. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res*. 17(11):1572–1585.

Webster AJ, Payne RJH, Pagel M. 2003. Molecular phylogenies link rates of evolution and speciation. *Science* 301:478.

Williams RJ. 2003. Fraústo Da Silva JJ. Evolution was chemically constrained. *J Theor Biol*. 220(3):323–343.

Yang S, Bourne PE. 2009. The evolutionary history of protein domains viewed by species phylogeny. *PLoS One*. 4:e8378.

Yang S, Doolittle RF, Bourne PE. 2005. Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A*. 102:373–378.

Yumiko Y-S, Ya-Pei H, Takeshi S. 2003. Distribution of flavonoids and related compounds from seaweeds in Japan. *J Tokyo Univ Fish*. 89:1–6.

Zuckerkandl E, Pauling LB. 1962. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in biochemistry. New York: Academic Press. p. 189–225.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/567/984545 by guest on 17 December 2018