QRB Discovery

www.cambridge.org/qrd

Protein structures unravel the signatures and patterns of deep time evolution

Ajith Harish

Independent Researcher, Uppsala, Sweden

Perspective

Cite this article: Harish A (2024). Protein structures unravel the signatures and patterns of deep time evolution. *QRB Discovery*, **5**: e3, 1–13 https://doi.org/10.1017/qrd.2024.4.

Received: 18 August 2023
Revised: 13 November 2023
Accepted: 12 December 2023

Keywords:
rooting; phylogenetic tree; protein structure; two-domains tree of life; archaea

Corresponding author:
Ajith Harish;
Email: ajith.harish@gmail.com

Abstract

The formulation and testing of hypotheses using 'big biology data' often lie at the interface of computational biology and structural biology. The Protein Data Bank (PDB), which was established about 50 years ago, catalogs three-dimensional (3D) shapes of organic macromolecules and showcases a structural view of biology. The comparative analysis of the structures of homologs, particularly of proteins, from different species has significantly improved the in-depth analyses of molecular and cell biological questions. In addition, computational tools that were developed to analyze the 'protein universe' are providing the means for efficient resolution of longstanding debates in cell and molecular evolution. In celebrating the golden jubilee of the PDB, much has been written about the transformative impact of PDB on a broad range of fields of scientific inquiry and how structural biology transformed the study of the fundamental processes of life. Yet, the transforming influence of PDB on one field of inquiry of fundamental interest—the reconstruction of the distant biological past—has gone almost unnoticed. Here, I discuss the recent advances to highlight how insights and tools of structural biology are bearing on the data required for the empirical resolution of vigorously debated and apparently contradicting hypotheses in evolutionary biology. Specifically, I show that evolutionary characters defined by protein structure are superior compared to conventional sequence characters for reliable, data-driven resolution of competing hypotheses about the origins of the major clades of life and evolutionary relationship among those clades. Since the better quality data unequivocally support two primary domains of life, it is imperative that the primary classification of life be revised accordingly.

Introduction

The linear amino acid chains of most proteins fold into a specific three-dimensional (3D) structure to become stable and functional. Protein folding is determined by the biophysical and biochemical constraints on the amino acid sequences (Anfinsen, 1973). Misfolded proteins usually malfunction and can often be lethal to the cell if not degraded. The Protein Data Bank (PDB), established in 1971 with a handful of protein structures determined by X-ray crystallography, is one of the first open-source data repositories (Bank, 1971). The PDB now hosts more than 180,000 structures of proteins, nucleic acids, and assemblies of supramolecular complexes. The PDB has transformed many life science disciplines by enriching our understanding of the physical and chemical basis of the fundamental biological processes. In celebrating 50 years of the PDB, recent article collections recount how PDB changed biology (Berman and Gierasch, 2021; Gierasch and Berman, 2021; Zardecki *et al.,* 2021). On the heels of these celebrations, the latest computational tools for *ab initio* structure prediction joined the celebratory bandwagon. Considered to be a once-in-a-generation advance, the latest computational tools such as Alphafold (Tunyasuvunakool *et al.,* 2021) and RoseTTaFold (Baek *et al.,* 2021) are a major leap in *ab initio* protein structure prediction.

Ab initio structure prediction from protein sequences that have no representative structures is a notoriously hard problem. The new algorithms extract information in protein sequences that are 'trained' over eons by evolution for spontaneous folding into specific and complex 3D shapes. The significance of these new algorithms compared to their predecessors is (a) the speed of determining the best folded conformation of a given linear amino acid sequence among the numerous possible 3D conformations and (b) the unmatched accuracy of the predicted structure, which is comparable to structures determined by X-ray crystallography and other experimental methods (Baek *et al.,* 2021; Tunyasuvunakool *et al.,* 2021).

Tools such as Alphafold and RoseTTaFold are a shot in the arm in the efforts to map the 'protein universe'. The protein universe is the assortment of all proteins from all organisms that have evolved on Earth over ≈3.8 billion years (Levitt, 2009). Whether or not such computational *de novo* structure predictions can replace the many experimental methods is an intriguing prospect for the future. At any rate, the success of the algorithmic predictions and the boost in their predictive power rely on the thousands of experimentally vetted structures available in the

CAMBRIDGE UNIVERSITY PRESS

PDB. The training data from which the rules and physicochemical constraints of protein folding are learnt underscore the invaluable insights gained from these high-resolution structure data.

The availability of the large number of protein structures has significantly improved the efforts in resolving another remarkably hard problem—reconstructing evolution itself—specifically, looking back into the earliest stages of cellular evolution through 'evolutionary telescopes'. The tremendous advantages of a protein structure-based evolutionary telescope (i.e., a phylogeny) over its predecessor—the more commonly used sequence-based telescope, is underappreciated. Current 'sequence vs structure' debates (Kurland and Harish, 2015a; Harish, 2018; Harish and Morrison, 2020; Williams *et al.,* 2020) are reminiscent of the 'morphology vs molecules' disputes (Simpson, 1964), that arose about 50 years ago, about which data type is better to investigate evolutionary problems. In celebrating the transformative influence of the PDB and structure-based insights on resolving a myriad of biological problems, this essay puts the spotlight on the impact of structural biology in bringing longstanding debates in evolutionary biology to empirical resolution. Given the decades-long development, the historical context of the current thinking and how it can be re-evaluated in light of abundant new structural data is discussed. In so doing, the analyses and new evidence presented here decisively show that structure-based data are much superior to the widely used sequence data to reconstruct the earliest stages of evolution of life.

Evolution of the protein universe recapitulates the evolution of cellular universe

The vast majority of cellular life is microbial (Locey and Lennon, 2016), especially single-celled species populate the bulk of the universe of cellular organisms. Proteins are components of the molecular machinery involved in all cellular functions (Figure 1*a*), from the birth through death of cells. Proteins are not only the workhorses of cells that drive the molecular machinery, but they also make up the infrastructure that maintains the morphology and internal organization of cells (Figure 1*b*). Enzymatic proteins that catalyze the biochemical reactions are an example of the former and cytoskeletal proteins of the latter. Cells can be seen as membranous ensembles studded with proteins inside and out (Figure 1). Based on the extent of membranous organization observed in ultrastructures of cells, two basic types of cellular organisms are known (Figure 1*c*).

- Eukaryotes (Greek; eu, 'well' and karyon, 'kernel'): Organisms with a well-defined membrane-bound nucleus and other membrane-bound intracellular compartments.
- Akaryotes (Greek prefix 'a-' meaning 'without'): Organisms without a nucleus or other membrane-bound compartments.

The terms eukaryote and akaryote are comparative descriptions of cell ultrastructure, though the term 'prokaryote' is commonly used to represent organisms with akaryotic cell organization. However, the term 'prokaryote' is misleading (Pace, 2006) as it is based on a misconception that prokaryotes are ancestors of eukaryotes, which runs counter to Darwin's proposal that all species share a common ancestor (Darwin, 1859).

The concept of a protein universe was put forward to organize proteins in a natural hierarchical system using tools of protein taxonomy (Ladunga, 1992). The number of distinct and stable 3D structures possible is limited by the physical and chemical constraints on protein folding, and thus, the number of unique 3D conformations (or folds) possible was estimated to be finite. Furthermore, based on the relationship between sequence and structure divergence in proteins (Chothia and Lesk, 1986), it was

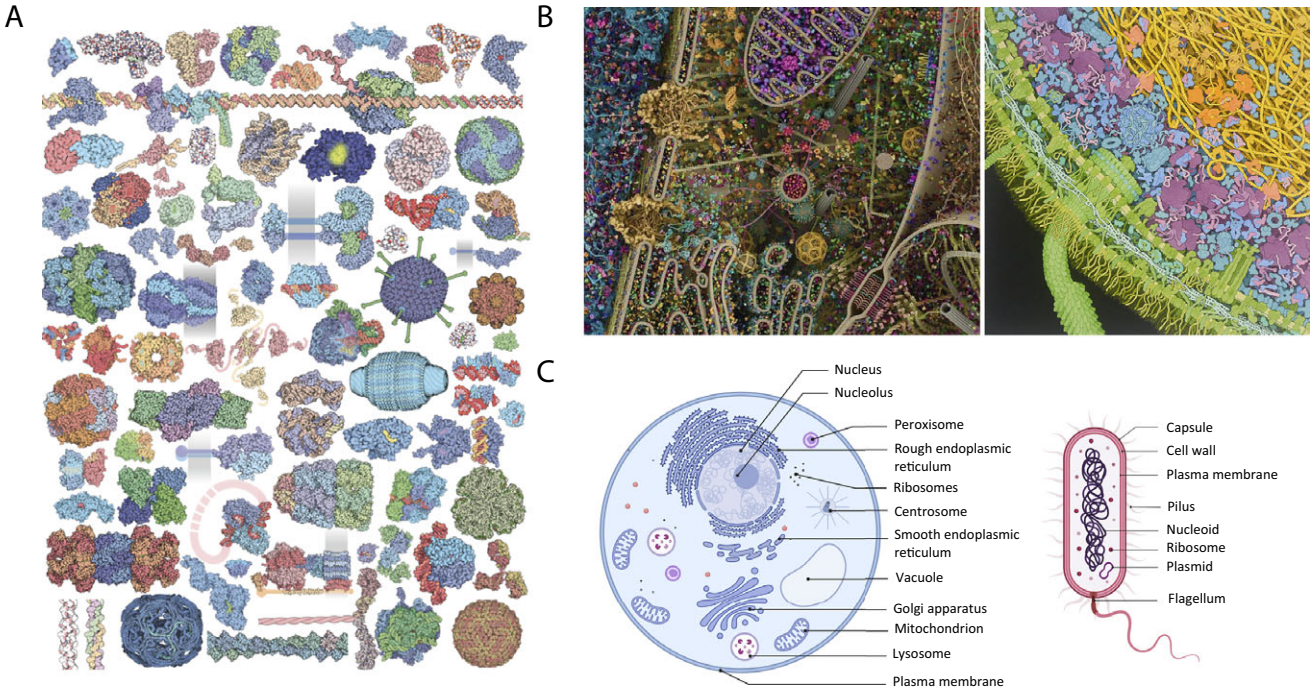


Figure 1. The molecular componentry of the cellular machinery. Most proteins fold into specific 3D shapes and form diverse supramolecular complexes to perform their biological functions (*a*). In addition to carrying out the biochemical reactions, proteins build and maintain the morphological features of cells. Cells are membranous ensembles studded with proteins, inside and out. The two basic cell types—eukaryotic (nucleated) and akaryotic (anucleate) cells—and the extent of membrane-bound compartmentation are shown as section of the ultrastructure (*b*) and the overall structure of an average eukaryotic cell and an average akaryotic cell (*c*).

predicted that a vast majority of proteins belong to no more than a thousand structural families (Chothia, 1992). At the time of this prediction almost 30 years ago, 866 structures were available in the PDB. In spite of the exponential growth of the number of structures available in PDB, the prediction has turned out to be largely true. Structure-based protein taxonomy developed by SCOP (Murzin *et al.,* 1995) and CATH (Orengo *et al.,* 1997) classification systems identify ≈1,500 and ≈1,400 folds, respectively. Despite being finite, and in spite of the remarkable advances in experimental 3D structure determination technologies, the protein universe is yet to be fully mapped (Levitt, 2009; Waman *et al.,* 2020). Due to the relative ease of DNA sequencing, mapping genomes has far outpaced protein structure determination. Tools such as Alphafold and RoseTTaFold could significantly speed up the efforts to map the protein universe.

At any rate, up to 70% of proteins of many species can be mapped to known structures (Kurland and Harish, 2015a; Waman *et al.,* 2020). This is already providing a substantial view of the distribution of proteomes in the cellular universe (Buchan *et al.,* 2002; Chothia *et al.,* 2003). In addition, the availability of such large numbers of protein structures has proved to be a new source of data as well as novel type of phylogenetic marker for (a) developing a new class of empirical models, namely nonstationary and nonreversible evolution models for statistical phylogenetics (Harish and Kurland, 2017a, 2017b), (b) empirical testing of competing hypotheses for the evolution of cellular life, including eukaryogenesis (Harish and Kurland, 2017a, 2017c), and (c) genome/proteome scale comparative analyses for reconstructing the major patterns of diversification of cellular life (Yang *et al.,* 2005; Fang *et al.,* 2013; Harish *et al.,* 2013). Importantly, the former two were previously not possible with commonly used nucleic acid and protein sequence data (Kurland and Harish, 2015b; Harish, 2018). In the following sections, I will discuss the pros and cons of both sequence and structure data and show why structure-based features are superior for a reliable reconstruction of the evolution of cellular universe or life as we know it.

The search for a perfect evolutionary character

Biologists utilize a variety of features or characters to describe and study organisms in a comparative framework. A character is any recognizable and heritable trait, feature, or property of an organism (Figure 2*a*) that can be employed for comparative analysis of character variances as a measure of evolutionary divergence of species (Figure 2*b*). Thus, 'characters' are fundamental data for

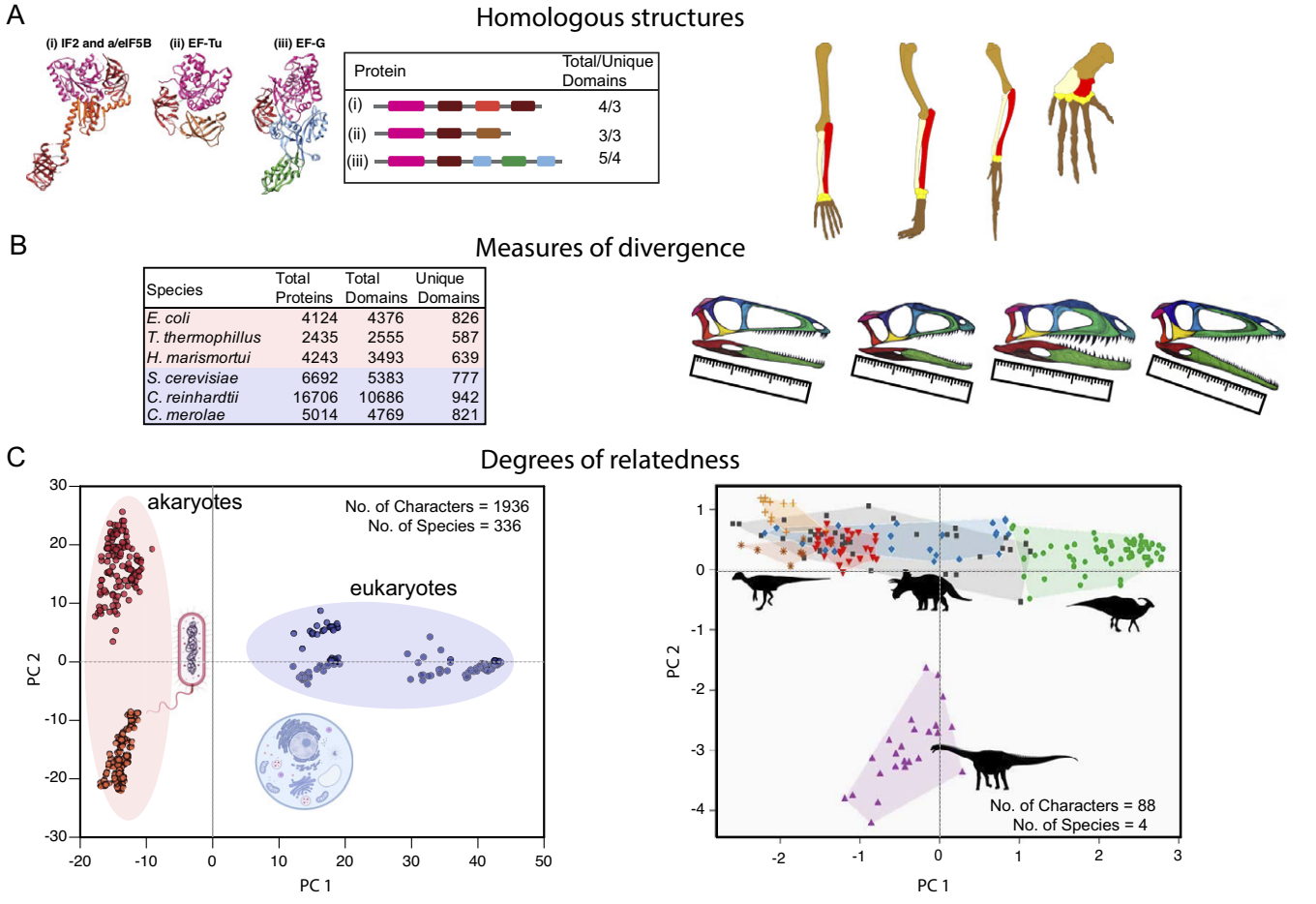


Figure 2. Protein domains are unique 'molecular phenotypes' to map patterns of species diversification. Structural domains are distinct homologous units that form complex protein morphologies (*a*; left) comparable to complex morphological structures (*a*; right). Since domains usually have characteristic functions, they are 'functional genomic signatures' (Harish, 2018). Measures of compositional variation of domains is a useful metric of divergence among organismal species groups (*b*; left). The number of unique occurrences defines a measure of 'intrinsic proteomic complexity' (Harish and Kurland, 2017b). Principal components analysis (PCA) projections show that the covariations of domain composition correspond to the two basic cell types (*c*; left). PC1 separates groups of eukaryote species from those of akaryote species, while PC2 separates species groups within each of eukaryotes and akaryotes. Domain-based metrics are comparable to measures of variance in fossil jaw and dental homologs (*b*; right), and the covariations of these features correspond to clades of dinosaur species (*c*; right panel taken from Nordén *et al.,* 2018).

evolutionary analyses and the character concept is central to evolutionary biology (Wagner, 2000). Initially, comparisons of morphological characters were used in taxonomic classification (Linnaeus, 1758) to study evolutionary processes (Darwin, 1859) and to determine phylogenetic relationships (Hennig, 1965). In principle, a single distinctive character is sufficient to distinguish species and groups of species from one another. For instance, the vertebral column is the defining feature of vertebrates—animals with a bony or cartilaginous backbone, which includes more than 60,000 species (Galbusera and Bassani, 2019). Likewise, the ascus, a microscopic structure that produces ascospores, is the defining feature of sac fungi or ascomycetes, with about 65,000 species (Schoch *et al.,* 2009). However, in practice, such defining features are not readily available for all species; hence, multiple characters are used for grouping related organisms into a clade or a monophyletic group, which is composed of a common ancestor and all its lineal descendants in a phylogenetic tree.

Distinctive characters that define clades are called synapomorphies or shared-derived characters that indicate a monophyletic origin of the character, i.e., synapomorphic characters represent a historically unique origin of an evolutionary novelty in the common ancestor of a clade (Hennig, 1965). Hennig reasoned that only synapomorphies should be used to diagnose common descent by tracing character evolution along a phylogeny. Complex developmental pathways and multiple genes are involved in the development of a morphological character. Complex characters like morphological features most likely arose only once and hence deemed to be homologous. In contrast, characters that evolve independently in multiple lineages are deemed to be homoplasious. The evidence to show that protein domains are superior characters while the sequence-based characters—nucleotides and amino acids—are poor quality data to resolve the deeper divergence of the tree of life is manifold. Here, I discuss the qualitative and quantitative evidence from recent studies that encourage the use of protein domains (Harish *et al.,* 2013; Harish and Kurland, 2017a; Harish, 2018). In addition, I present new quantitative evidence to show that sequence characters by themselves are unsuitable for confidently resolving the deeper branches of the universal phylogeny.

In mapping the protein universe, a protein domain is the basic unit of structure, function, and evolution (Figure 2*a*). Domains are independently folding sectors of a polypeptide chain, with a unique 3D shape that is associated with a distinct amino acid sequence profile and a characteristic function (Murzin *et al.,* 1995; Orengo *et al.,* 1997). For these reasons, domains are excellent 'characters' to study many aspects of biological evolution. Therefore, tracing the history of the variation of domain composition in species is valuable for determining the evolutionary relationships and patterns of diversification among species groups. The species-specific composition of unique domains was termed as 'intrinsic proteomic complexity' (Harish and Kurland, 2017a).

The idiosyncratic assortment of domains in organisms corresponds to species groups (Figure 2*c*) and other levels of the taxonomic hierarchy of organismal classification (Harish *et al.,* 2013; Harish and Kurland, 2017b). Advantages of protein domains for assessing both qualitative and quantitative empirical evidence are summarized below. For details and incisive analyses, see the studies by Harish and Kurland, 2017a, 2017b; Harish, 2018; Harish and Morrison, 2020.

- Protein domains, unlike their component amino acids, provide for a large number of 'unique' characters. Latest updates of SCOP and CATH classification of PDB entries identify ≈2,750 and ≈5,500 homologous superfamilies. This translates to anywhere between 2,750 and 5,500 unique structure characters as opposed to only 20 distinct sequence characters (amino acids).
- Since each homologous domain has a distinct 3D structure, a unique sequence profile, and a characteristic function, substitution between domains is not known. In contrast, repeated substitution of amino acids at the same site is frequent, resulting in a rapid decay of historical signal.
- Independent evolution of complex structural domains in diversified species and *ab initio* evolution of new proteins by random mutations are both extremely rare. However, it is relatively easier to lose domains via multiple mechanisms. For example, a mutation causing a premature stop codon or loss of a genomic locus during genetic recombination. This naturally skewed propensity for loss (death) over gain (birth) of a new domain can be exploited to implement time nonreversible or directional evolution models, which are better suited to reconstruct the universal tree.
- Finally, the relatively lower variation of (a) compositional heterogeneity and (b) rate heterogeneity of birth/death of unique protein domains compared to point mutations in sequences supports statistically robust phylogenetic inferences.

Thus, structure-defined characters provide for a robust and reliable resolution of the deeper nodes of the universal tree. A caveat is that some of the recent divergences in certain clades are not as well supported as the deeper ones when only structure-based characters are used (Harish and Kurland, 2017b; Harish, 2018). In contrast to the underappreciated advantages of structure-based data, the deficiencies of sequence data have been hashed out in multiple studies over the past three decades. Here, I will describe key qualitative and quantitative evidence. Although the ribosome as a whole and the small subunit ribosomal RNA (SSU rRNA) in particular were thought to be the 'universal chronometer' of evolutionary analyses initially (Woese, 1987), it is now abundantly clear that focusing on the ribosome alone is a reductionist approach (Harish, 2018). The deficiencies of sequence characters, in general, and the resulting error prone analyses, specifically of the rRNA and r-proteins are rather pronounced, as shown in many studies during the past two decades (Tourasse and Gouy, 1999; Rokas and Carroll, 2008; Philippe *et al.,* 2011; Gouy *et al.,* 2015). For instance, inclusion (or exclusion) of different sets of ribosomal genes/proteins produces different relationships between Archaea, Bacteria, and Eukarya (Da Cunha *et al.,* 2017). In addition, application of different models of sequences evolution to the same dataset produces different results (Tourasse and Gouy, 1999; Harish, 2018).

These incongruences are due to many well-known deficiencies of sequence data, such as:

- higher incidence of homoplasy (or lack of homology) in sequence characters,
- large variation in rates of evolution among different genes/proteins and/or within different sections of the same gene/protein (e.g., in different domains of multi-domain proteins),
- sequence data are often limited to the application of time-reversible models of evolution.

To mention a few. Chief among these deficiencies is the inapplicability of time nonreversible models of character evolution for rRNA and r-protein sequences, and for sequence-based analyses in general. This serious deficiency of sequence data is demonstrated

Table 1. Results of model selection tests for sequence-based and structure-based characters

Primary sequences Character type: Amino acids				Tertiary structures Character type: Protein domains			
Model	–lnL	BIC	BF	Model	–lnL	BIC	BF
LG + R5	166404	333463	0	***Mk + G12 + NONREV***	78941	–	0
LG + R6	166396	333463	0	Mk + G12	79797	–	856
LG + R4	166468	333575	56	Mk	85817	–	6876
LG + I + G4	166521	333647	92				
LG + G4	166614	333824	181				
LG + F + R5	166601	334013	275				
LG + F + R6	166593	334016	276				
LG + R3	166697	334016	277				
NONREV + FO + R5	165164	334133	335				
NONREV + FO + R6	165161	334143	340				
NONREV + FO + R4	165208	334205	371				
NONREV + F + R5	165163	334289	413				
NONREV + FO + R3	165424	334620	579				
WAG + F + R5	167858	336529	1533				
WAG + F + R6	167854	336536	1537				
WAG + R5	168015	336684	1611				
LG + R2	168042	336688	1613				
WAG + R6	168010	336692	1615				
NONREV + FO + R2	166587	336930	1734				
LG + I	172062	344720	5629				
LG	174208	349005	7771				

Note: Best-fitting models for the sequence data were determined, in the present study, using ModelFinder (Kalyaanamoorthy *et al.,* 2017) implemented in IQ-TREE (v 2.1.3) (Nguyen *et al.,* 2015). The top ranked time reversible and nonreversible models are shown here; the complete list of models tested is in the Supplementary Material. Sequence alignments used to estimate a global ToL in an earlier study (Williams *et al.,* 2012) were employed. BF scores were estimated from BIC scores in bayestestR (v 0.13.1.7) (Makowski *et al.,* 2019). Model selection tests for the structure-based data are from a previous study (Harish, 2018) using tests implemented in MrBayes (Klopfstein *et al.,* 2015). Time nonreversible models of evolution are highlighted in bold and italicized. lnL, log-likelihood scores; BIC, Bayesian Information Criterion scores; BF, Bayes Factor scores.

here with quantitative evidence obtained from model selection tests, as shown in Table 1.

Bayes factor (BF) scores are useful to assess the relative merits of competing models, as BF is considered as the weight of the evidence coming from the data. A difference in BF scores in the range of 20–150 is typically treated as strong evidence in favor of the better model (and the resulting tree), while BF difference of above 150 is considered very strong empirical evidence (Kass and Raftery, 1995; Bergsten *et al.,* 2013). Thus, the quantitative evidence in Table 1 shows that time reversible models are better fitting models and time nonreversible models are worse fitting models for sequence data, by a large margin. In contrast, time nonreversible models are better fitting models for structure data, by a huge margin. Time reversible models can only produce unrooted trees, which has no evolutionary direction, for example, from ancestor to descendant (Figure 3*c*). However, in practice a rooted tree, which has an evolutionary direction (Figure 3*b*), is necessary for almost all evolutionarily relevant answers sought using phylogenetic analyses, such as (i) ancestor–descendant polarity, (ii) branching order in evolutionary history, and (iii) evolutionary groups that are clades (Morrison, 2006). Thus, the limited applicability of nonreversible models is a severe intrinsic deficiency of sequence data, which in and of itself, is the source of ambiguity, and often discord, among proponents of competing hypothesis for the relationships and origin of the major clades of life including origin of animals and eukaryotes (Kurland and Harish, 2015a; Harish, 2018; Harish and Morrison, 2020).

Many deficiencies of sequence-based analyses have been addressed over the past two decades, primarily through improved statistical modeling (Philippe *et al.,* 2011). However, since models are as good as the data on which they are based on, data quality is the most important aspect of building and testing statistical models. Thus, high-quality data (characters) is essential for the success of data-driven resolution of competing hypotheses. Here, data quality refers to the quality or the strength of the phylogenetic signal of homology that can be recovered from the data. The strength of the phylogenetic signal is proportional to the confidence with which unique state transitions can be determined for a given set of characters on a given tree (Harish, 2018). Ideally, historically unique character transitions that entail rare evolutionary innovations are desirable to identify patterns of uniquely shared innovations (synapomorphies) among lineages. Although improved modeling can correct errors of estimation and improve the fit of the data to the tree, it is not a solution to improve phylogenetic

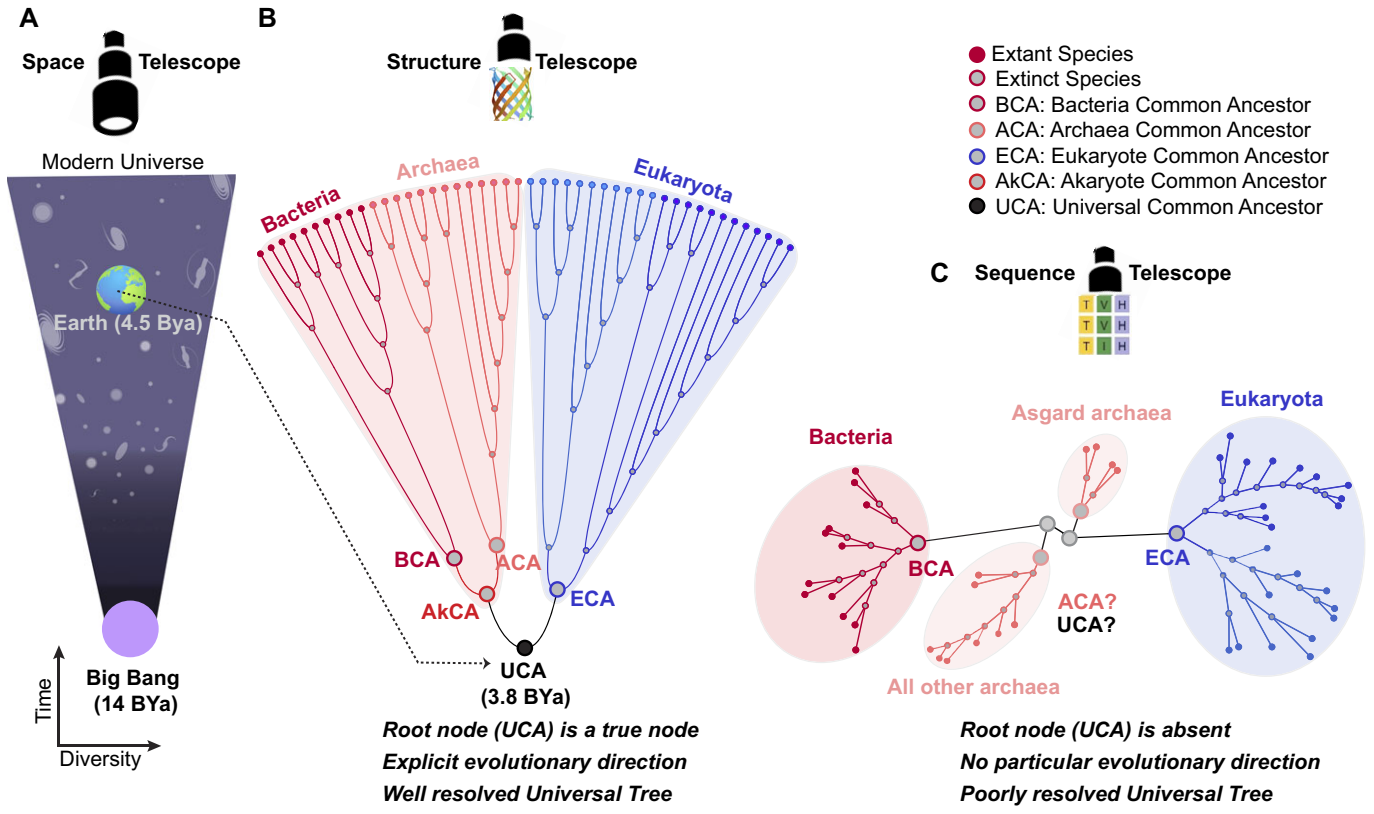


Figure 3. Evolutionary telescopes: Protein structure telescopes can look further back in time, however sequence telescopes cannot. Optical telescopes are used to look at the distant 'galactic fossils' of the cosmological universe estimated to have originated ≈14 BYa (*a*). Phylogenies are the 'evolutionary telescopes' used to look back into the distant biological past depicted as the 'universal tree' of life. The universal tree shown in (*b*) is a schematic of the phylogeny inferred from patterns of inheritance of 'functional genomic signatures' defined by unique protein domains (Harish, 2018). Protein structure telescopes can look further back in time due to their superior resolving power (*b*) compared to the commonly used sequence telescopes (*c*). Ancestral nodes including the root node of the universal tree (UCA) as well as the root node of the archaea tree (ACA) cannot be identified using sequence telescopes (e.g., Williams *et al.,* 2020); hence, the reconstructed picture of evolution is poorly resolved and incomplete (*c*).

signal, especially when the historical signal is exceedingly limited or absent in the source data. Thus, structure-based characters like protein domains are probably the closest to a perfect character that is currently available for evolutionary biologists.

Farsightedness and nearsightedness

Contemporary species are the evolutionary successors of long-gone ancestors. About 99% of species that evolved on Earth have gone extinct (Barnosky *et al.,* 2011) with little trace left as fossils, especially of microbial species. Therefore, reconstructing a detailed picture of the common ancestor of all extant life—the universal common ancestor (UCA)—is a daunting task (Harish and Morrison, 2020). The UCA was most likely a single-celled species estimated to have lived between 3.8 and 3.5 billion years ago (BYa) (Figure 3). Otherwise, the nature of UCA is still fuzzy and rife with speculation. When it comes to peering back into the distant past, astrophysicists and evolutionary biologists are faced with similar problems in collecting reliable data and building tools to analyze and interpret the data (Ade *et al.,* 2014; Krauss, 2014; Kurland and Harish, 2015a).

Studies to reconstruct the cosmological past initially relied on refracting telescopes. Telescopes are the fundamental tools to study the observable universe. Refracting telescopes are one of the two main types of optical telescopes, which operate by collecting light through a large lens and focusing the light on an eyepiece/camera (STScI, 2021). However, refracting telescopes suffer from a phenomenon called 'chromatic distortion'—a common optical problem resulting from an inability of the lens to bring all wavelengths of light to a sharp focus. As a result, the reconstructed images of distant galaxies—galactic ancestors of contemporary universe—are fuzzy. In a telescope, the function of a good lens is to minimize such optical aberrations as much as possible to produce an unblurred and high-fidelity view. The deficiencies of refracting telescopes were overcome by reflecting telescopes wherein the lens was replaced with a mirror to collect light and focus better for a clearer picture. The Hubble and Webb space telescopes are the largest reflecting telescopes that can collect high-quality data of the most distant 'galactic fossils' of the cosmological universe.

The observable cosmic universe converges into a singularity known as the cosmological light horizon, which represents beginnings of the universe and the boundary between the observable and unobservable universe (Figure 3*a*). In reconstructing the biological past, the UCA or Universal Common Ancestor represents a singularity—a phylogenetic event horizon (Figure 3*b*)—which is the root node of the universal tree of life (Harish *et al.,* 2013). Among the tools used to peer back into the galaxies of the cellular universe, protein structure–based evolutionary telescopes are like the reflecting telescopes with minimal distortions (Harish and Kurland, 2017a; Harish, 2018). Therefore, structure-based telescopes

produce a well-resolved and high-fidelity picture of the distant biological past (Figure 3*b*). In contrast, their predecessors—sequence-based telescopes produce a poorly resolved and low-fidelity picture of the past, which makes the identification of UCA and its immediate descendants unreliable (Figure 3*c*). It is worth noting that UCA is the 'most recent' UCA or the 'last' UCA of a lineage of cellular species since the origins of cells; evolutionary telescopes cannot peer into pre-UCA or pre-cellular epochs of evolution (Kurland and Harish, 2015a). However, plausible models of pre-UCA and pre-cellular evolution can be developed based on our knowledge of the biophysical and biochemical constraints that govern protein folding and protein evolution, independently of the use of evolutionary telescopes (Abroi and Gough, 2011; Norden, 2021; Kocher and Dill, 2023).

Resolving the deeper nodes of the universal tree of life (hereafter universal tree) in general, and the root node in particular, using sequence telescopes is fraught with distortions similar to the chromatic aberrations of refracting telescopes (Harish, 2018). This is because (1) the rates of substitution mutations in sequences show extreme variations and (2) the historical signal decays significantly with time due to repeated substitutions that overwrite the evolutionary record (Harish and Kurland, 2017a; Harish and Morrison, 2020). The decay of historical signal increases spurious signals and decreases the reliability of analyses. While distortions due rate variations can be corrected with mathematical models, decay and loss of signal cannot be compensated (Harish, 2018). Thus, distortions of evolutionary signal and high uncertainty in identifying the UCA are common with comparative analysis of primary sequence data (Harish, 2018; Harish and Morrison, 2020; Williams *et al.,* 2020; Liu *et al.,* 2021).

Speculative descriptions and theoretical predictions about the nature and origin of UCA are abundant, including the 'panspermia' hypothesis, which presumes that terrestrial life originated in outer space (Crick and Orgel, 1973). Regardless of an extraterrestrial or terrestrial origin, and whether life arose only once or multiple times, identifying the UCA using a data-driven and rigorous phylogenetic analysis boils down to determining the root of the universal phylogenetic tree (Theobald, 2010; Harish and Morrison, 2020). Determining the root node, which is the deepest node of the universal tree, is one of the hardest problems in phylogenetic analysis and thus far rooting using sequence characters has either (a) generally not been possible (Pace, 1997; Woese, 1987) or (b) has been ambiguous at best (Philippe and Forterre, 1999; Morrison, 2006; Harish *et al.,* 2013, 2016; Gouy *et al.,* 2015; Harish and Morrison, 2020). Hence, in practice rooting, the universal tree relies on assuming a false root (Woese, 1987; Pace, 1997; Spang *et al.,* 2015; Liu *et al.,* 2021) that is based on unverified suppositions that align with the traditional and falsifiable hypothesis that prokaryotes evolved before eukaryotes.

Nothing in the universal tree makes sense except in the light of the universal ancestor

During last two decades, the universal tree is routinely constructed as a composite tree of 30-50 phylogenetic 'marker' proteins (Ciccarelli *et al.,* 2006; Spang *et al.,* 2015; Liu *et al.,* 2021) rather than from a single marker gene: the SSU rRNA gene (Woese, 1987). Marker protein datasets are either solely or predominantly composed of ribosomal proteins (Harish, 2018). Standard sequence-based methods trace the history of substitution mutations using time-reversible substitution models, which are devised for computational convenience rather than to represent biological reality (Harish and Kurland, 2017b; Harish, 2018). Time-reversible models can only produce 'unrooted trees' and hence lack the ability to identify the root node, which in the case of the universal tree represents the UCA. This inherent deficiency of the sequence telescopes was noted early on (Woese, 1987). Thus, an unrooted universal tree (Figure 3*c*) is not only poorly resolved but is also an incomplete depiction of evolution. In addition, since unrooted trees do not present a clear evolutionary interpretation, they are prone to misreading and thus potentially misleading (Morrison, 2006; Harish, 2018; Harish and Morrison, 2020). Because of the total absence of a root node, the deficiency of standard sequence telescopes is, in fact, worse than chromatic distortion of refracting telescopes.

In general, a 'rooted tree' is a straightforward depiction of the principle of common ancestry with a clear branching order along a time axis of ancestor–descendant polarity (Figure 3*b*). In contrast, an unrooted tree is undirected with no particular direction for evolutionary time and thus with undefinable branching order (Figure 3*c*). This distinction between an unrooted and a rooted tree is of prime importance as most conclusions from phylogenetic analyses strictly depend on a rooted tree. The primary conclusions of significance include determining (a) ancestors and descendants, (b) branching order (i.e., tree topology), (c) evolutionary groups (clades), (d) degree of relatedness among clades, and (e) ancestral states of characters under study. Hence, an unrooted tree is not an evolutionary tree (phylogeny) in its true sense, even though it depicts relatedness among the organisms (Harish, 2018; Sánchez-Pacheco *et al.,* 2020).

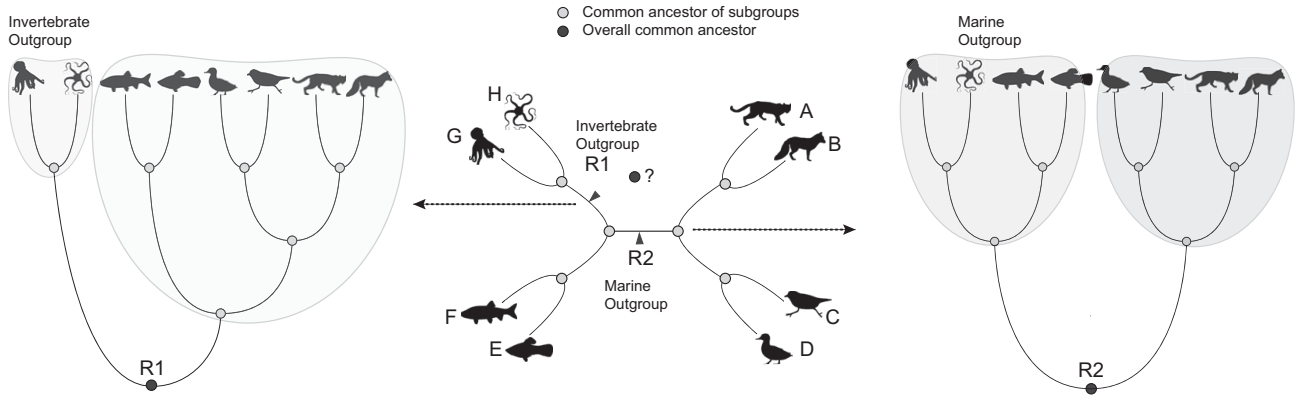
Evidently, the importance of identifying the UCA cannot be emphasized enough (Gouy *et al.,* 2015; Harish and Kurland, 2017b; Harish, 2018; Harish and Morrison, 2020). Yet, rooting is relegated as a secondary task and is often trivialized. As commonly used phylogenetic routines cannot identify the root, 'pseudo rooting' (see Box 1) based on external information and/or unverified conjectures that are independent of the data used to produce unrooted trees becomes necessary.

Pseudo-rooting converts undirected trees to directed trees so that evolutionarily meaningful conclusions can be drawn (Woese, 1987; Harish, 2018). Depending on the external information (e.g., outgroups) or conjectures about the UCA, pseudo-rooting of the universal tree is the common practice (Woese *et al.,* 1990; Spang *et al.,* 2015; Imachi *et al.,* 2020; Williams *et al.,* 2020; Liu *et al.,* 2021). Though widely accepted, the assumptions underpinning pseudo-rootings of the universal tree were rarely tested until recently (Harish and Kurland, 2017a; Harish, 2018). The studies provided, to my knowledge, the first formal test of the widely accepted conjectures about UCA and eukaryogenesis as well as the first empirical evidence that supports the independent evolution of eukaryotes and akaryotes (archaea and bacteria) and rejects the popular endosymbiotic origin of eukaryotes and origin of key eukaryotic features within archaea and bacteria.

The difficulty of locating the root node in sequence-based analysis and the importance of a statistically robust root inference were recently highlighted in efforts to trace the origin of SARS-CoV-2 (the human severe acute respiratory syndrome coronavirus 2) and the coronavirus 2019 (COVID-19) pandemic (Morel *et al.,* 2021; Pipes *et al.,* 2021). One study claimed to have identified the ancestors of the human SARS-CoV-2 lineages (Forster *et al.,* 2020) by including the bat corona virus as outgroup to root a median joining network (MJN). Rather than tracing the history of substitution mutations, MJNs estimate genetic distances, which is

Box 1. Using 'pseudo roots' to convert unrooted trees into evolutionary trees

Rooting trees in general is a difficult problem, conceptually and technically (Harish and Kurland, 2017b). The distinction between unrooted and rooted trees is nontrivial as shown in Box–Figure 1. The unrooted tree (Box–Figure 1; center) shows four groups of animal species A-H, of which A-F are vertebrates (species with a vertebral column) and G-H are invertebrates (species without a vertebral column). Species A-D are terrestrial (with lungs), while species E-H are marine (without lungs). The internal nodes represent common ancestors: the common ancestors of contemporary species, as well their common ancestors (gray circles). However, the overall common ancestor of all species (black circle) is unknown and unidentifiable in the unrooted tree.



Box–Figure 1. Examples of different rearrangements of the branching order following an outgroup rooting. Depending on the different positions of the root node, the different degrees of relatedness among species groups can be inferred. The nearest neighbor in an unrooted tree may not be the closest evolutionary relative (Harish, 2018). The degree of relatedness can only be determined with rooted trees. Trees are drawn as cladograms with emphasis on branching order and relative age of common ancestors of contemporary species; branch lengths have no evolutionary implications.

Standard time-reversible models of evolution produce 'unrooted trees', in which (a) the position of the overall common ancestor cannot be identified and (b) nor a particular direction for evolutionary time is implied. Unrooted trees are not only incomplete depictions of the hierarchy of descent but can potentially misrepresent the evolutionary kinships (Box–Figure 1; center). To complete the picture, an *outgroup* is usually chosen to assign a 'pseudo root'. The addition of the root node introduces a branching order by rearranging the tree around the root node (Box–Figure 1; left and right). Choice of an outgroup is based on assumptions about features (i.e., characters). For example, presence or absence of lungs/vertebrae is assumed to be the ancestral state. Since an artificial root node representing the overall common ancestor is introduced *after-the-fact*, it is designated as a 'pseudo-root'. Hence unrooted/undirected trees are not true evolutionary trees.

In the example shown above, choosing an invertebrate outgroup implies that the absence of vertebral column is the ancestral state. Choosing the invertebrate outgroup results in a clade wherein all the vertebrates (A-F) are grouped together. Similarly, a marine outgroup implies that the absence of lungs is the ancestral state. The choice of the marine outgroup results in a clade in which some of the vertebrates (E, F) as well as invertebrates (G, H) are grouped together. However, fossil evidence confirms that the invertebrate outgroup assumption is correct as far as the Animal tree is concerned (Donoghue and Purnell, 2009). Regardless of the fossil evidence, grouping together vertebrates and invertebrates in one clade is rather odd. Likewise, grouping some akaryotes (archaea) and all eukaryotes together is odd too. Thus, this rooting exercise shows that if assumptions about outgroups are wrong, as with the "marine outgroup", the results can be blatantly wrong. Besides, rooting with an outgroup is merely a tree drawing option, but the mathematically estimated tree remains unrooted as a mathematical entity. Therefore, though widely used, the outgroup rooting method is able to introduce only a false root, because a non-existent root node is artificially introduced to the unrooted tree.

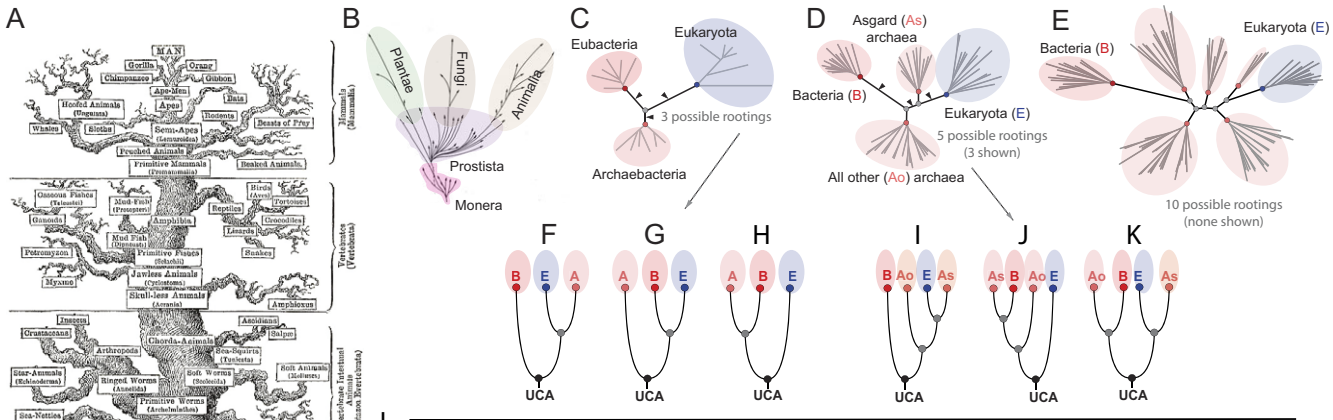
As for the Universal tree, in the absence of both organismal outgroups and reliable geological fossils, protein domains are perhaps the best characters available at present (Harish, 2018). Locating the root node (UCA) with a directional character evolution model is the most straightforward approach to infer evolutionary trees (Harish and Kurland, 2017a, 2017b). Accordingly, empirical evidence in favor of the descent of eukaryotes and akaryotes from UCA supports Eukarya and Akarya as the primary clades of life (see Figure 4 and associated discussion).

unsuitable to trace the history of mutations and to reconstruct ancestral sequence states (Sánchez-Pacheco *et al.,* 2020).

In contrast, several other studies employed suitable substitution models and more rigorous statistical phylogenetic methods to evaluate multiple rootings (Morel *et al.,* 2021; Pipes *et al.,* 2021). The SARS-CoV-2 tree was rooted with (1) nonreversible substitution models, (2) molecular clock models, and (3)(pseudo) rooted using the outgroup criterion with multiple outgroups. Yet, an unambiguous and statistically robust rooting was not possible using the best available methods of primary sequence analysis, in spite of the availability of massive whole-genome datasets. The difficulty and unreliability of rooting the SARS-CoV-2 tree was due to a rapid loss of evolutionary signal (Morel *et al.,* 2021; Pipes *et al.,* 2021). The shortcomings of unrooted MJNs and potential misinterpretations were pointed out in a sharp response as '*Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary*' (Sánchez-Pacheco *et al.,* 2020).

Poor resolution due to a lack of historical signal of homology along with misinterpretation of unrooted trees can lead to profoundly misleading conclusions in many other evolutionary studies as well (Baum *et al.,* 2005; Harish, 2018). Recently, it was shown that significant loss of historical signal in standard sequence data is the basis of the problems and persistent ambiguities in resolving the deeper nodes of the universal tree (Harish, 2018). Thanks to structure telescopes, the deficiencies of sequence telescopes can now be overcome so that a well-supported and well-resolved universal tree can be reconstructed. The advantages of embracing structure-based characters for studying evolution are manifold (see next section). However, the routine use of sequence characters with weak evolutionary signal goes hand-in-hand with the standard practice of pseudo-rooting (Lake, 1986; Woese *et al.,* 1990; Liu *et al.,* 2021).

Pseudo-rootings are routinely used to assert that (a) archaea are the closest relatives of eukaryotes (Woese *et al.,* 1990), (b) eukaryotes evolved from a specific lineage of archaea (Spang



L

Tree	Rank	Primary Clades (Domain/Empire)	lnL Score	Bayes Factor ΔlnL	Odds Ratio $e^{\Delta lnL}$	Closest Relative of Eukaryotes	Monophyletic Archaea	Taxonomic Status of Archaea	Taxonomic Status of Eukaryota
H	1	Eukaryota Akaryota	-78941	-	Most Probable	Akaryotes	Yes	Kingdom	Domain
F	2	Eukaryota Bacteria Archaea	-79276	335	10^{145} times less likely	Archaea	Yes	Domain	Domain
G	3	Eukaryota Bacteria Archaea	-79334	393	10^{170} times less likely	Bacteria	Yes	Domain	Domain
I	4	Archaea Bacteria	-79696	755	Improbable	Asgard archaea	No	Domain	Undefined

Figure 4. Assessment of empirical evidence for or against alternative universal trees and for identifying the primary clades of life. The identification of the primary clades in the universal tree is basically linked to the identity of the root node (UCA), which is implicitly assumed ever since the earliest universal trees were put forward (*a*–*e*). The most popular assumption (Woese *et al.,* 1990) is that UCA is positioned on the stem branch leading to (Eu)bacteria (*c*) and the textbook universal tree (*f*). Several other phylogenetic positions for the UCA and the resulting phylogenies (*d*, *g*–*k*). Assessment of empirical evidence for or against these proposals (*l*) shows that the universal tree in which Eukarya and Akarya are the primary clades (*h*) is most likely to be correct (Harish, 2018).

et al., 2015; Imachi *et al.,* 2020), (c) archaea are intermediates on the evolutionary path to eukaryotes from bacteria (Imachi *et al.,* 2020), (d) archaeal origin of eukaryote protein homologs (Cotton and McInerney, 2010), and (e) bacterial origin of eukaryote protein homologs (Karlberg *et al.,* 2000; Martijn *et al.,* 2018). However, pseudo-rootings based on pseudo-outgroups or unverified assumptions are error prone and unreliable (Gouy *et al.,* 2015; Harish, 2018). Thus, the standard practice of using time-reversible evolution models and pseudo-rootings, solely for interpreting the results of an unrooted universal tree, is prone to faulty conclusions and can be misleading. Indeed, unsupported false-rootings along with poorly resolved trees tend to foster common misconceptions about evolution (Harish *et al.,* 2016; Harish and Kurland, 2017b; Sánchez-Pacheco *et al.,* 2020).

Misreading of even well-resolved, rooted trees is surprisingly common (Baum *et al.,* 2005). For instance, the universal tree in which eukaryotes and akaryotes descend and diverge from the UCA (Figure 3*b*) is often misinterpreted as a 'eukaryotes first' scenario or an 'upside down' tree of life, since it contradicts the common false rootings. Such rootings (a) conflate UCA with other common ancestors (Figure 3*c*) and (b) are usually based on the notion that archaea and bacteria are primitive and thus assumed to be ancestors of eukaryotes in a prokaryote-to-eukaryote progression (Harish and Kurland, 2017a; Harish and Morrison, 2020). Rather, the straightforward conclusion is that eukaryotes and akaryotes are sister clades and that the closest relative of the eukaryote common ancestor as well as the akaryote common ancestor is UCA (Figure 3*b*).

Roots of stability: The diminishing relevance of the three domains classification system

The Linnean system of organizing species into nested hierarchies, *Systema Naturae* (Linnaeus, 1758), first published in 1735, was developed a century before Darwin's oft-cited vision '*Our classifications will come to be, as far as they can be so made, genealogies*' (Darwin, 1859). The term 'phylogeny' was coined when one of the first genealogical tree of life was depicted (Haeckel, 1866), inspired by the principle that a 'natural system' and a true classification should be represented as an evolutionary tree (Darwin, 1859). Yet, some of the prominent genealogical trees (Figure 4*a*–*c*) are not Darwinian trees. In Darwinian phylogenetic trees, contemporary species are at the leaves (terminal nodes) and extinct ancestors at the internal nodes. Many notable hypotheses of phylogenetic progression assume, explicitly or implicitly, that some extant species groups are primitive (Figure 4*a*–*c*), much like the popular depictions of evolution as a linear progression from simple to complex forms. For instance, unicellular species with akaryotic cell organization were assumed to be primitive and placed near a 'virtual root' of the tree (Figure 4*a*; Haeckel, 1866), (Figure 4*b*; Whittaker, 1969), and (Figure 4*c*; Woese, 1987). However, the venerable ancestor, UCA in this case, was neither a distinct entity nor an empirically derived node on the phylogeny. Hence, neither the three-kingdom system (Figure 4*c*; Woese, 1987) nor its predecessor, the five-kingdom system (Figure 4*b*; Whittaker, 1969), is truly phylogenetic.

The poor resolution of archaea due to unreliable phylogenetic signal in routinely used 'marker sequences' is often seen as non-monophyly of archaea (Figure 4*d*,e). Regardless of rooting, poor resolution of archaea further confounds phylogenetic classification (Figure 4*i*–*k*). If we are to accept a poorly resolved universal tree, then some of the possibilities depending on different rootings are:

- Eukarya ceases to exist as an exclusive clade and a taxonomic domain (Figure 4*i*).
- Bacteria ceases to exist as an exclusive clade and a taxonomic domain (Figure 4*j*).
- All of eukarya, bacteria and archaea cease to exist as exclusive clades or domains of life (Figure 4*k*).

Arguably, resolving the root node of the universal tree is one of the hardest problems in evolutionary biology given the time depth (Harish, 2018). Fortunately, structural characters cut the Gordian knot by facilitating an empirical resolution of the rooting problem as well as diagnosis of the monophyly of the major species groups by allowing the assessment of empirical evidence in favor of the different suppositions and tentative hypotheses (Figure 4*l*). BFs provide a means to evaluate the strength of evidence in favor of the best hypothesis among a set of competing proposals (Harish and Kurland, 2017a; Harish, 2018). BF is the ratio of the likelihoods of the different hypotheses being compared. A BF of 5 or greater is considered as very strong empirical evidence in favor of the hypothesis with the better likelihood (Harish, 2018). Therefore, a BF of 335 means extremely strong empirical evidence for the two-domain universal tree (Figure 4*h*) and for primary clades Eukarya and Akarya (Bacteria and Archaea are sister clades), compared to other competing hypotheses (Harish, 2018). Likelihood scores are the log odds of the hypotheses; thus, the Eukarya-Akarya two-domain phylogeny is at least 10^{145} times more probable than the closest competing phylogeny (Figure 4*f*), which is the three-domain phylogeny (Eukarya and Archaea are sister clades). The alternative two-domain proposal is improbable and least supported. Put simply, the Eukarya-Akarya two-domain phylogeny is most likely to be correct.

The universal tree is primarily a phylogenetic classification. The basic requirements of phylogenetic classification are:

Monophyly: Only monophyletic groups (or clades) are true evolutionary groups. That is, groups comprising all the descendants of a given common ancestor should be identified.

Homology: Delineating clades is based on diagnosing patterns of descent of characters that evolved in the common ancestors and were inherited by the descendants (i.e., homologous characters). Nested configurations of sharing of such homologous characters in different species are used to group them in an order of common descent (i.e., branching order), which then diagnose the degree of relatedness among the clades.

This seemingly straightforward procedure of character analysis and the algorithmic logic to diagnose clades was developed so that phylogenetic classification is an objective exercise determined by the branching order (Hennig, 1965).

However, the common practice of pseudo-rooting is essentially based on unverified conjectures and unsupported assumptions, which not only encourages a subjective interpretation of the universal tree but also fosters the continued overlooking of evidence against popular conjectures (Kurland and Harish, 2015b; Harish and Kurland, 2017c). Though appealing and widely accepted, these assumptions were rarely tested until recently (Harish and Kurland, 2017a, 2017c; Harish, 2018), largely because it is not feasible to test the veracity of such assumptions with sequence characters and standard time-reversible models (Harish and Kurland, 2017a). In addition, perhaps because the pseudo-rooting aligns with another common traditional assumption: that simple is primitive (Harish *et al.,* 2016; Harish and Kurland, 2017b). This assumption is pervasive since the time of the earliest efforts to reconstruct a genealogical tree of life (Figure 4*a–c*). The current practice of pseudo-rooting the universal tree (Zaremba-Niedzwiedzka *et al.,* 2017; Liu *et al.,* 2021) goes back to the initial efforts of classification of life using molecular characters (Woese *et al.,* 1990). Thus, conclusions based on widely assumed pseudo-rooting are compromised both by a lack empirical evidence and their reliance on poor quality data (characters) (Harish and Kurland, 2017c; Harish, 2018; Harish and Morrison, 2020). Notably, the commonly acknowledged but rarely tested notions that eukaryotes evolved from within archaea (Spang *et al.,* 2015; Zaremba-Niedzwiedzka *et al.,* 2017; Imachi *et al.,* 2020; Williams *et al.,* 2020) and that eukaryotes evolved from a merger of archaea and bacteria (Karlberg *et al.,* 2000; Cotton and McInerney, 2010; Martijn *et al.,* 2018) are most seriously compromised due to (a) the strong evidence for the monophyly of archaea *and* akaryotes and (b) because the widely accepted false-rootings and their underlying assumptions lack support (Harish and Kurland, 2017c; Harish, 2018; Harish and Morrison, 2020).

That said, it is worth emphasizing that given the hierarchy of descent, contextualizing the recent divergences is conditional on the resolution of the deeper divergences. In addition, character evolution models for protein domains and their component amino acids describe mutually exclusive processes that account for hierarchically different evolutionary timescales (Harish, 2018). Nevertheless, support for recent divergences can be improved in several ways: (1) by employing an expanded character set defined by the updated domain descriptions. While previous studies were based on ≈1,800 domains described at the time (Harish and Kurland, 2017b; Harish, 2018), the number of domain descriptions have tripled to 5,400 at present; (2) combining sequence and structural characters; and (3) a multiphasic approach to resolve different parts of the universal tree independently using either structure-based or sequence-based approaches, depending on the questions addressed, is a useful alternative (Harish, 2018).

Summary and implications

Perhaps, portraying the pros and cons of the sequence-based and structure-based reconstruction of the universal tree as a 'battle of characters' would make for an entertaining tale. However, both types of molecular features are complementary and are valuable for resolving different parts of the universal tree. That is, by melding together structure telescopes and sequence telescopes, both farsightedness and nearsightedness of evolutionary telescopes can be corrected. During the last two decades, neither increasing the sophistication of the substitution models nor aggregating more sequences has been productive in (a) reliably resolving contentious evolutionary relationships, (b) accurately determining the temporal order of key evolutionary innovations, and (c) describing the exceptional sister group differences of the major species groups. Embracing the well-defined structure-based characters will certainly prove to be beneficial.

Structure telescopes provide for a straightforward and objective means for identifying the UCA and to determine the major clades and key evolutionary transitions across the tree of life. Protein structural domains define unique molecular phenotypes, which are robust evolutionary characters that improve the level of confidence in resolving the deepest divergences of the universal tree. It is abundantly evident that, in addition to a richer representation of cellular and molecular phenotypes, protein domains offer a deeper perspective of the evolutionary history. Thus, they provide for a better means of (a) describing the key innovations and evolutionary transitions across the tree of life, (b) objective evaluation of competing hypotheses for the evolution of cellular life, and (c) a phylogenetic classification that is an accurate representation of the two basic types of cell organization.

The universal phylogeny wherein eukaryotes and akaryotes are sister clades is by far the best empirically supported universal tree of

life by any measure, qualitative and quantitative. Qualitative measures include both tree-independent assessment of character homology and the tree-based assessment of common ancestry—character homology and homology of clades. Quantitative measures include robust statistical support for (a) fit of character evolution model to a rooted tree, (b) the branching order, starting with the rooting, and (c) higher measures of confidence to reject alternative universal tree proposals. Hence, structure telescopes are better suited in contrast to sequence telescopes to look further back into the biological past. The serious limitation of sequence characters, especially with regard to the assessment of qualitative evidence, can be overcome with structure characters.

Outlook

The durability of Linnean classification rests on the choice of excellent 'diagnostic features' or characters used to group species into genera, families through Kingdoms. If only high-quality molecular characters such as protein domains were available early on, perhaps there never would have been a third domain of life. Hindsight is 20/20, after all. The Linnean hierarchical classification implicitly reflected common descent of the species thus classified and ultimately converged into an Empire, *Imperium Naturae*. Although the taxonomic grade Domain/Empire is warranted, in light of the new, stable rooting and well-supported branching order of the universal tree, grades for Archaea and Bacteria should be revised to Kingdoms, whether or not their respective initial nomenclatures Archaeabacteria and Eubacteria should take precedence. Eukarya and Akarya are the primary domains of life as well as the principal taxonomic ranks, both terms being descriptive of the two basic cell types.

Looking forward, as genome sequences and protein structures continue to accumulate, future efforts for a better resolved universal tree could employ a variety of new molecular features. In addition to primary sequences and known protein domains, many newer evolutionary characters can be identified by determining (a) novel protein domains for which structures are unknown and (b) new types of homologous features from quaternary assemblies of the protein complexes, among others. Tools like AlphaFold and RoseTTaFold seem to be primed for such undertakings. In this way, the evolution of morphological phenotypes as well as physiological phenotypes at the cellular level can be reconstructed in greater detail than what is possible at present.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/qrd.2024.4.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/qrd.2024.4.

Acknowledgements. I am grateful to Måns Ehrenberg and David Morrison for discussions and comments on an earlier version of the manuscript and for the continued support. I also thank the editor and reviewers for their comments, which improved the presentation. An earlier version of this manuscript, written on the occasion of the Golden Jubilee of PDB, can be found on OSF Preprints (https://doi.org/10.31219/osf.io/hyknj).

Image credits: Images available under public licenses for adaptation and sharing are acknowledged below. All other images were produced by the author for this article.

Figure 1: (*a*) Molecule of the month, illustration by David S. Goodsell, RCSB Protein Data Bank. (*b*, LHS) Eukaryotic cellular landscape, illustration by Evan Ingersoll and Gaël McGill, PhD (Digizyme Inc.) using Molecular Maya software. Created for Cell Signaling Technology, Inc., and inspired by the stunning art of David Goodsell, this 3D rendering of a eukaryotic cell is modeled using X-ray, nuclear magnetic resonance (NMR), and cryoelectron microscopy datasets for all of its molecular actors. (*b*, RHS) *Escherichia coli* bacterium, 2021, illustration by David S. Goodsell, RCSB Protein Data Bank. doi: 10.2210/rcsb_pdb/goodsell-gallery-028. (*c*) Created with BioRender.com.

Figure 2: (*a*, RHS) Volkov Vladislav Petrovich, Wikimedia Commons, CC BY-SA 4.0. B: RHS: Danny Cicchetti, Wikimedia Commons, CC BY-SA 4.0.

Figure 3: (*a*) First Stars: Timeline of the Universe, Space Telescope Science Institute (STSci).

Figure 4: (*a*) Tree of life by Ernst Haeckel, Wikipedia.org.

Box–Figure 1: Silhouette images of animals, by PhyloPic database (http://phylopic.org/).

Financial support. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Abroi A and Gough J (2011) Are viruses a source of new protein folds for organisms? – Virosphere structure space and evolution. *BioEssays* **33**(8), 626–635.

Ade PA, Aikin R, Barkats D, Benton S, Bischoff CA, Bock J, Brevik J, Buder I, Bullock E and Dowell C (2014) Detection of B-mode polarization at degree angular scales by BICEP2. *Physical Review Letters* **112**(24), 241101.

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* **181**(4096), 223–230.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ and Baker D (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876.

Bank PD (1971) Crystallography: Protein Data Bank. *Nature: New Biology* **233** (42), 223.

Barnosky AD, Matzke N, Tomiya S, Wogan GO, Swartz B, Quental TB, Marshall C, McGuire JL, Lindsey EL and Maguire KC (2011) Has the earth's sixth mass extinction already arrived? *Nature* **471**(7336), 51–57.

Baum DA, Smith SD and Donovan SSS (2005) The tree-thinking challenge. *Science* **310**(5750), 979–980.

Bergsten J, Nilsson AN and Ronquist F (2013) Bayesian tests of topology hypotheses with an example from diving beetles. *Systematic Biology* **62**(5), 660–673.

Berman HM and Gierasch LM (2021) How the Protein Data Bank changed biology: An introduction to the JBC reviews thematic series, part 1. *Journal of Biological Chemistry* **296**, 100608.

Buchan DWA, Shepherd AJ, Lee D, Pearl FMG, Rison SCG, Thornton JM and Orengo CA (2002) Gene 3D: Structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Research* **12** (3), 503–514.

Chothia C (1992) One thousand families for the molecular biologist. *Nature* **357** (6379), 543–544.

Chothia C, Gough J, Vogel C and Teichmann SA (2003) Evolution of the protein repertoire. *Science* **300**(5626), 1701–1703.

Chothia C and Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* **5**(4), 823–826.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B and Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311** (5765), 1283–1287.

Cotton JA and McInerney JO (2010) Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proceedings of the National Academy of Sciences* **107**(40), 17252–17255.

Crick FH and Orgel LE (1973) Directed panspermia. *Icarus* **19**(3), 341–346.

Da Cunha V, Gaia M, Gadelle D, Nasir A and Forterre P (2017) Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genetics* **13**(6), e1006810.

Darwin C (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Donoghue PC and Purnell MA (2009) The evolutionary emergence of vertebrates from among their spineless relatives. *Evolution: Education and Outreach* **2**(2), 204–212.

Fang H, **Oates ME**, **Pethica RB**, **Greenwood JM**, **Sardar AJ**, **Rackham OJL**, **Donoghue PCJ**, **Stamatakis A**, **De Lima Morais DA and Gough J** (2013) A daily-updated tree of (sequenced) life as a reference for genome research. *Scientific Reports* **3**, 2015.

Forster P, **Forster L**, **Renfrew C and Forster M** (2020) Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences* **117**(17), 9241–9243.

Galbusera F and Bassani T (2019) The spine: A strong, stable, and flexible structure with biomimetics potential. *Biomimetics* **4**(3), 60.

Gierasch LM and Berman HM (2021) How the Protein Data Bank changed biology: An introduction to the JBC reviews thematic series, part 2. *Journal of Biological Chemistry* **296**, 100748.

Gouy R, **Baurain D and Philippe H** (2015) Rooting the tree of life: The phylogenetic jury is still out. *Philosophical Transactions of the Royal Society B* **370**(1678), 20140329.

Haeckel E (1866) *Generelle Morphologie der Organismen: Bd. Allgemeine Entwickelungsgeschichte der Organismen* (Vol. **2**). Berlin: G. Reimer.

Harish A (2018) What is an archaeon and are the archaea really unique? *PeerJ* **6**, e5770.

Harish A, **Abroi A**, **Gough J and Kurland C** (2016) Did viruses evolve as a distinct supergroup from common ancestors of cells? *Genome Biology and Evolution* **8**(8), 2474–2481.

Harish A and Kurland CG (2017a) Akaryotes and eukaryotes are independent descendants of a universal common ancestor. *Biochimie* **138**, 168–183.

Harish A and Kurland CG (2017b) Empirical genome evolution models root the tree of life. *Biochimie* **138**, 137–155.

Harish A and Kurland CG (2017c) Mitochondria are not captive bacteria. *Journal of Theoretical Biology* **434**, 88–98.

Harish A and Morrison D (2020) The deep(er) roots of eukaryotes and Akaryotes [version 2; peer review: 2 approved, 1 approved with reservations]. *F1000Research* **9**, 112.

Harish A, **Tunlid A and Kurland CG** (2013) Rooted phylogeny of the three superkingdoms. *Biochimie* **95**(8), 1593–1604.

Hennig W (1965) Phylogenetic systematics. *Annual Review of Entomology* **10** (1), 97–116.

Imachi H, **Nobu MK**, **Nakahara N**, **Morono Y**, **Ogawara M**, **Takaki Y**, **Takano Y**, **Uematsu K**, **Ikuta T**, **Ito M**, **Matsui Y**, **Miyazaki M**, **Murata K**, **Saito Y**, **Sakai S**, **Song C**, **Tasumi E**, **Yamanaka Y**, **Yamaguchi T**, **Kamagata Y**, **Tamaki H and Takai K** (2020) Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* **577**(7791), 519–525.

Kalyaanamoorthy S, **Minh BQ**, **Wong TK**, **Von Haeseler A and Jermiin LS** (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**(6), 587–589.

Karlberg O, **Canbäck B**, **Kurland CG and Andersson SGE** (2000) The dual origin of the yeast mitochondrial proteome. *Yeast* **17**(3), 170–187.

Kass RE and Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association* **90**(430), 773–795.

Klopfstein S, **Vilhelmsen L and Ronquist F** (2015) A nonstationary Markov model detects directional evolution in hymenopteran morphology. *Systematic Biology* **64**(6), 1089–1103.

Kocher C and Dill KA (2023) Origins of life: First came evolutionary dynamics. *QRB Discovery* **4**, e4.

Krauss LM (2014) Peering Back to the beginning of time. *Physics* **7**, 64.

Kurland CG and Harish A (2015a) The phylogenomics of protein structures: The backstory. *Biochimie* **119**, 284–302.

Kurland CG and Harish A (2015b) Structural biology and genome evolution: An introduction. *Biochimie* **119**, 205–208.

Ladunga I (1992) Phylogenetic continuum indicates "galaxies" in the protein universe: Preliminary results on the natural group structures of proteins. *Journal of Molecular Evolution* **34**(4), 358–375.

Lake JA (1986) An alternative to archaebacterial dogma. *Nature* **319**(6055), 626–626.

Levitt M (2009) Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America* **106**(27), 11079–11084.

Linnaeus Cv (1758) *Systema naturae*, Vol. **1**. Stockholm.

Liu Y, **Makarova KS**, **Huang W-C**, **Wolf YI**, **Nikolskaya AN**, **Zhang X**, **Cai M**, **Zhang C-J**, **Xu W and Luo Z** (2021) Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**(7860), 553–557.

Locey KJ and Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* **113**(21), 5970–5975.

Makowski D, **Ben-Shachar MS and Lüdecke D** (2019) bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software* **4**(40), 1541.

Martijn J, **Vosseberg J**, **Guy L**, **Offre P and Ettema TJ** (2018) Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**(7703), 101.

Morel B, **Barbera P**, **Czech L**, **Bettisworth B**, **Hübner L**, **Lutteropp S**, **Serdari D**, **Kostaki E-G**, **Mamais I and Kozlov AM** (2021) Phylogenetic analysis of SARS-CoV-2 data is difficult. *Molecular Biology and Evolution* **38**(5), 1777–1791.

Morrison DA (2006) Phylogenetic analyses of parasites in the new millennium. *Advances in Parasitology* **63**, 1–124.

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

Nguyen L-T, **Schmidt HA**, **Von Haeseler A and Minh BQ** (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**(1), 268–274.

Norden B (2021) Which are the 'Hilbert problems' of biophysics? *QRB Discovery* **2**, e1.

Nordén KK, **Stubbs TL**, **Prieto-Márquez A and Benton MJ** (2018) Multifaceted disparity approach reveals dinosaur herbivory flourished before the end-cretaceous mass extinction. *Paleobiology* **44**(4), 620–637.

Orengo CA, **Michie AD**, **Jones S**, **Jones DT**, **Swindells MB and Thornton JM** (1997) CATH – a hierarchic classification of protein domain structures. *Structure* **5**(8), 1093–1108.

Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**(5313), 734–740.

Pace NR (2006) Time for a change. *Nature* **441**, 289.

Philippe H, **Brinkmann H**, **Lavrov DV**, **Littlewood DTJ**, **Manuel M**, **Wörheide G and Baurain D** (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology* **9**(3), e1000602.

Philippe H and Forterre P (1999) The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution* **49**(4), 509–523.

Pipes L, **Wang H**, **Huelsenbeck JP and Nielsen R** (2021) Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *Molecular Biology and Evolution* **38**(4), 1537–1543.

Rokas A and Carroll SB (2008) Frequent and widespread parallel evolution of protein sequences. *Molecular Biology and Evolution* **25**(9), 1943–1953.

Sánchez-Pacheco SJ, **Kong S**, **Pulido-Santacruz P**, **Murphy RW and Kubatko L** (2020) Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary. *Proceedings of the National Academy of Sciences* **117**(23), 12518–12519.

Schoch CL, **Sung G-H**, **López-Giráldez F**, **Townsend JP**, **Miadlikowska J**, **Hofstetter V**, **Robbertse B**, **Matheny PB**, **Kauff F and Wang Z** (2009) The Ascomycota tree of life: A phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic Biology* **58**(2), 224–239.

Simpson GG (1964) Organisms and molecules in evolution. *Science* **146**(3651), 1535–1538.

Spang A, **Saw JH**, **Jørgensen SL**, **Zaremba-Niedzwiedzka K**, **Martijn J**, **Lind AE**, **van Eijk R**, **Schleper C**, **Guy L and Ettema TJ** (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**(7551), 173–179.

STScI (2021) *Webb Telescope*. Available at https://www.stsci.edu/communications-and-outreach (accessed April 8, 2021).

Theobald DL (2010) A formal test of the theory of universal common ancestry. *Nature* **465**(7295), 219–222.

Tourasse NJ and Gouy M (1999) Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced

from rRNA and protein-coding genes. *Molecular Phylogenetics and Evolution* **13**(1), 159–168.

Tunyasuvunakool K, **Adler J**, **Wu Z**, **Green T**, **Zielinski M**, **Žídek A**, **Bridgland A**, **Cowie A**, **Meyer C**, **Laydon A**, **Velankar S**, **Kleywegt GJ**, **Bateman A**, **Evans R**, **Pritzel A**, **Figurnov M**, **Ronneberger O**, **Bates R**, **Kohl SAA**, **Potapenko A**, **Ballard AJ**, **Romera-Paredes B**, **Nikolov S**, **Jain R**, **Clancy E**, **Reiman D**, **Petersen S**, **Senior AW**, **Kavukcuoglu K**, **Birney E**, **Kohli P**, **Jumper J and Hassabis D** (2021) Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596.

Wagner GP (2000) *The Character Concept in Evolutionary Biology*. New Haven, CT, USA: Elsevier.

Waman VP, **Blundell TL**, **Buchan DW**, **Gough J**, **Jones D**, **Kelley L**, **Murzin A**, **Pandurangan AP**, **Sillitoe I and Sternberg M** (2020) The Genome3D consortium for structural annotations of selected model organisms. In Kihara D (ed.), *Protein Structure Prediction*. New York: Springer, pp. 27–67.

Whittaker RH (1969) New concepts of kingdoms of organisms. *Science* **163** (3863), 150–160.

Williams TA, **Cox CJ**, **Foster PG**, **Szöllősi GJ and Embley TM** (2020) Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution* **4**(1), 138–147.

Williams TA, **Foster PG**, **Nye TMW**, **Cox CJ and Embley TM** (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. *Proceedings of the Royal Society B: Biological Sciences* **279**, 4870–4879.

Woese CR (1987) Bacterial evolution. *Microbiological Reviews* **51**(2), 221.

Woese CR, **Kandler O and Wheelis ML** (1990) Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences of the United States of America* **87**(12), 4576–4579.

Yang S, **Doolittle RF and Bourne PE** (2005) Phylogeny determined by protein domain content. *Proceedings of the National Academy of Sciences of the United States of America* **102**(2), 373–378.

Zardecki C, **Shao C**, **Voigt M and Burley S** (2021) Protein Data Bank: 50 years of macromolecular structures enabling research and education. *The FASEB Journal* **35**, D464–D474.

Zaremba-Niedzwiedzka K, **Caceres EF**, **Saw JH**, **Bäckström D**, **Juzokaite L**, **Vancaester E**, **Seitz KW**, **Anantharaman K**, **Starnawski P**, **Kjeldsen KU**, **Stott MB**, **Nunoura T**, **Banfield JF**, **Schramm A**, **Baker BJ**, **Spang A and Ettema TJG** (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**(7637), 353–358.