The Phylogenomic Roots of Translation

Derek Caetano-Anollés and Gustavo Caetano-Anollés

> In my version the history of life is counterpoint music, a two-part invention with two voices, the voice of the replicators attempting to impose their selfish purposes upon the whole network and the voice of homeostasis tending to maximize diversity of structure and flexibility of function. The tyranny of the replicators was always mitigated by the more ancient cooperative structure of homeostasis that was inherent in every organism. The rule of the genes was like the government of the old Hapsburg Empire: Despotismus gemildert durch Schlamperei, or 'despotism tempered by sloppiness'.
>
> —Freeman Dyson [1]

1 Introduction

The mechanisms behind translation and the specificities of the genetic code are well understood and are dependent on both nucleic acids and proteins [2]. In particular, transfer RNAs, or tRNAs for short, are central L-shaped nucleic acid molecules that are necessary for the transfer of genetic information from genomes and its interpretation during protein biosynthesis. They play fundamental roles during the entire translation process and during other processes of the cell as well. tRNAs recognize cognate aminoacyl-tRNA synthetase (aaRS) enzymes, which help them charge specific amino acids to the 3′ ends protruding from their acceptor stems. In turn, 'anticodon' sequences in their anticodon loops recognize complementary 'codon' sequences in messenger RNA (mRNA), translating genetic information that was

D. Caetano-Anollés (✉)
Department of Evolutionary Genetics, Max-Planck-Institut für Evolutionsbiologie, 24306 Plön, Germany
e-mail: caetano@evolbio.mpg.de

G. Caetano-Anollés (✉)
Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA
e-mail: gca@illinois.edu

G. Hernández and R. Jagus (eds.), *Evolution of the Protein Synthesis Machinery and Its Regulation*, DOI 10.1007/978-3-319-39468-8_2

transcribed into RNA. The codon-anticodon recognition occurs within the confines of a complex ribonucleoprotein environment, the ribosome. tRNAs not only interact with mRNA but also with ribosomal RNA (rRNA) and proteins (r-proteins), as tRNAs are being ratcheted through the center of the biosynthetic complex and their amino acids unloaded during protein bond synthesis in the ribosomal peptidyl transferase center (PTC). The resulting polypeptides that are extruded through the ribosomal exit pore then fold according to hidden rules determined by interactions of tRNAs with all of its protein and nucleic acid partners. This 'structural code,' which holds deep historical information on the origin and evolution of proteins and life, differs from the 'genetic code.' It holds overarching specificities for the central molecular machinery that drives metabolism, translation, transcription and replication. Its vocabulary is currently unknown.

tRNAs are also very ancient molecules, a fact that is made evident by their universality and the many roles they play in translation and other biological processes [3]. For example, a recent study of the distribution of RNA molecules in a catalog of over a thousand RNA families revealed that tRNA molecules were part of only five families that were universally present in all biological organisms [4]. These families included rRNA and ribonuclease P (RNase P) RNA. The ubiquity and universality of the very central tRNA molecules have prompted their phylogenetic study using information in their sequences and structures [5–9]. Here we focus on the history of tRNA and its most fundamental interacting proteins and nucleic acid partners, aaRSs, elongation factors and ribosomal molecules, which are also part of a number of molecular complexes (e.g., ribosomes, multi-aaRS complexes). To unfold this history, we used phylogenomic information extracted from the sequence of millions of protein sequences and thousands of molecular structures to build a step-by-step timeline of accretion of their component parts, protein structural domains and RNA helical segments. We show that the gradual nucleation of these molecular modules, which behave as evolutionary units of proteins and nucleic acids, is ultimately responsible for the complexity of structures and molecular interactions unfolding in the biology of extant organisms.

2 A Structural Phylogenomic Method to Study the Evolution of Macromolecules

Phylogenetic analysis provides an objective criterion to study the natural history of biological entities of many kinds, beginning with the evolution of organisms, using information in specific features of those entities. The phylogenetic rationale of traveling back in time was made explicit by German entomologist Willi Hennig about half a century ago [10]. The systematization of evolutionary analyses gave rise to the fields of cladistics and systematic biology and provided background knowledge for the development of the field of molecular evolution and evolutionary genomics. It also resulted in the ongoing construction of a Tree of Life

(ToL) describing the evolution of organismal diversity at the planetary scale. Remarkably, no comparable community-driven effort is being pursued that would produce global views of the evolution of molecules of the kind advocated by Emile Zuckerkandl and Margaret Dayhoff in the early 1970s [11, 12]. Despite this shortcoming, the fields of structural biology and genomics have advanced considerably during past decades to provide wide-encompassing understanding of molecular diversity at atomic resolution [13]. As of 4 January 2016, there are 114,697 models of molecular structure deposited in the entries of the PROTEIN DATA BANK (PDB) [14], and their associated functions are encoded in the DNA of the 8,434 genomes and metagenomes that have been completely sequenced (GOLD DATABASE [15]). Genomic information has given rise to 0.55 million UNIPROTKB/SWISSPROT and ~50.4 million UNIPROTKB/TREMBL protein sequence entries and information in thousands of functional RNA molecules.

Phylogenetic analysis builds tree representations of genealogical relationships of the entities that are being studied, the *phylogenies*, by mining information in a number of biological features of interest, the phylogenetic *characters* [16]. Traditional characters that are useful include biochemical, morphological, physiological, developmental and molecular features with historical signal. The vast majority of molecular features that have been studied so far involve sequence information in alignments, i.e., sets of characters describing positions along a string of monomers that are homologous within groups of macromolecules. However, function impacts fitness and constrains evolution. Since molecular structures are the repositories of molecular functions, they are generally more resistant to change than sequences. They are therefore highly conserved at the evolutionary level and ideal candidates to study the history of life, from the very deep relationships to the most recent. For that reason, we have been studying the evolution of protein and nucleic acid structures for almost 2 decades using the wealth of information generated by the genomic revolution (first reviewed in [17]). We start by first summarizing the experimental strategies used to study molecular history (Fig. 1) and then describing some useful applications.

(1) *Evolution of proteins.* Advanced hidden Markov models (HMMs) of structural recognition assign fold structures to protein sequences with high accuracy and low error rates. These bioinformatic annotations permit the generation of a structural census of proteins, with structural domains defined at various levels of protein structural abstraction in the hierarchical classifications of SCOP [18] and CATH [19], the gold standards. We have computed the proteomic occurrence and abundance of each domain structure across a wide transect of organisms and used this proteomic census to construct data matrices (arrays) for phylogenetic analysis. Phylogenetic trees of domains (ToDs) and trees of proteomes (ToPs) were built from this census. The first study of this kind was published in 2003 and involved a proteomic analysis of only 32 organisms [20]. Recent analyses extended the approach to thousands of them and to viruses [21]. Since ToDs and ToPs can be rooted using direct methods of character polarization, the rooted trees describe the origin and evolution of parts and wholes, the structural domains (the evolutionary units of proteins) and the proteomes (the entire protein repertoire of an organism),

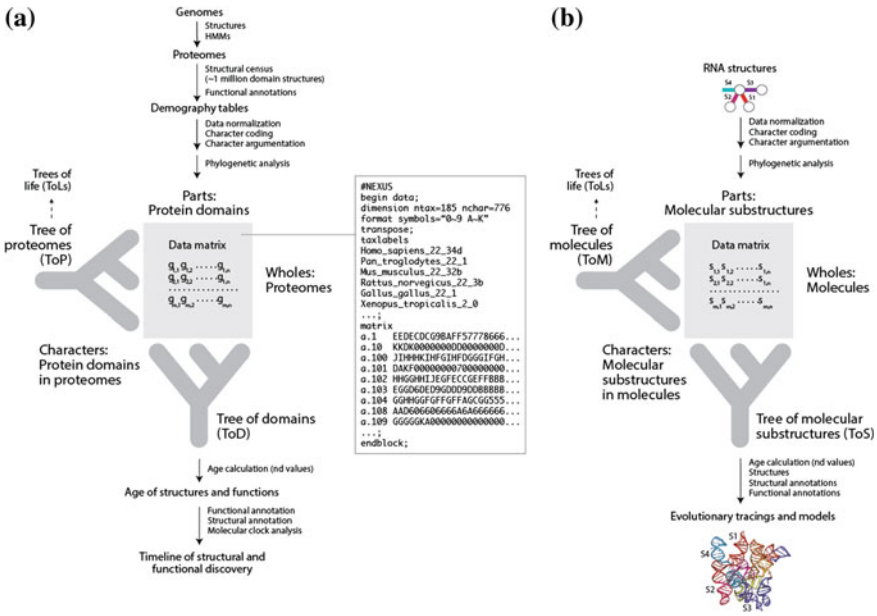
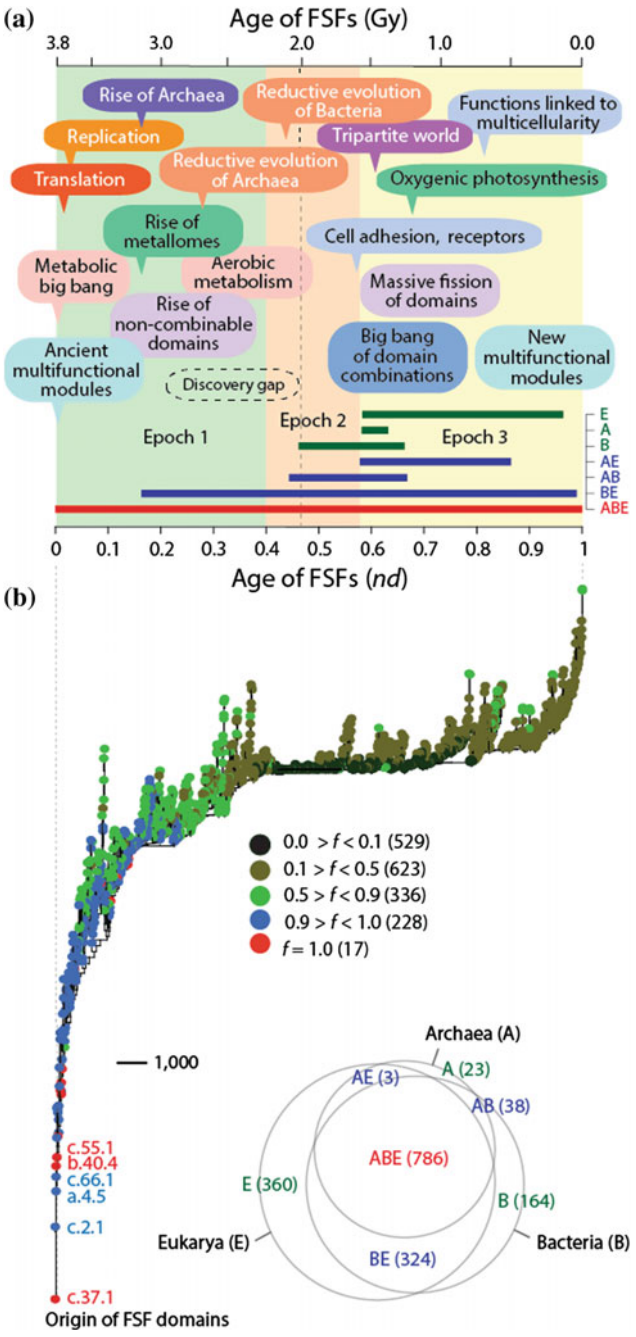


Fig. 1 Structural phylogenomic analyses of protein domains and RNA molecules. **a** The flow diagram describes the steps leading to the reconstruction of trees of domains (ToDs) and trees of proteomes (ToPs) and associated timelines of domain innovation. A census of domain structures in proteomes of thousands of completely sequenced organisms is used to compose data matrices (arrays with rows and columns corresponding to taxa/characters) for building phylogenomic trees. The trees describe the evolution of individual structural domains and proteomes, respectively. Elements of the array (g) represent genomic abundances of domains in proteomes, defined at different levels of classification of domain structure. The *inset* shows a very small segment of a NEXUS file holding readable information for tree computation. **b** The flow diagram to the right describes the phylogenetic reconstruction of trees of molecules (ToMs) and trees of substructures (ToSs) of RNAs. Molecular structures are first decomposed into substructures, including helical stem tracts and unpaired regions. Structural features of these substructures (e.g., length) are coded as phylogenetic characters and assigned character states according to an evolutionary model that polarizes character transformation toward an increase in conformational order (character argumentation). Coded characters (s) are arranged in data matrices. Phylogenetic analysis generates rooted phylogenetic trees. Embedded in ToDs and ToSs are timelines that assign age to molecular structures. These ages can be 'painted' onto 2D or 3D structural models of RNA or proteins, RNP complexes or protein complexes generating evolutionary heat maps

respectively. Furthermore, the fact that ToDs are comb-like enabled the construction of timelines of domain history by counting the number of nodes from the base of the tree to each taxon and expressing the value as a relative 'node distance' (*nd*). These *nd* values measure the relative age of each domain, which can be linked to the geological record through time calibration points. In fact, a remarkably linear relationship was observed between the *nd* and the age of biomarkers and geomarkers diagnostic of domain structures [22]. This relationship defined a molecular clock of domain structures, which effectively turned ToDs into 'timetrees' [23],

i.e., bona fide chronologies with time axes in billions of years (Gy). A ToD built from SCOP domain structures and its associated timeline is illustrated in Fig. 2. Note that the time of first appearance of a domain structure in a chronology records the time of origin of that structure and that the gradual evolutionary appearance of domains involves thousands of steps. Each step represents a domain structure with numerous and important annotations, including domain distribution in organisms and viruses, molecular functions, biochemical and biophysical properties, and biological network participation. Moreover, since the new discovery of domains has reached a plateau, the number of domains must be considered finite, and ToDs currently approach the highest level of universality that is possible in phylogenetic statements. The generation of timelines of structural innovation has already allowed exploration of a number of important questions. ToDs have been used to trace the origin and evolution of metabolic networks [27–29], study the rise of translation and the genetic code [30, 31], uncover the co-evolutionary history of the ribosome [32, 33], explore the evolution of metallomes and biological metal utilization [34], unfold the natural history of biocatalytic mechanisms [35] and protein flexibility [36], study the evolutionary dynamics of gain and loss of domains [37] and domain combinations [24], determine the makeup of the universal common ancestor of life [38], visualize a basal stem line of descent responsible for organismal biodiversity [25, 39] and generate a truly universal ToL that includes cellular organisms and viruses directly from the age of domains using multidimensional scaling approaches [21]. A recent encyclopedia entry summarizes some of the findings [40].

(2) *Evolution of nucleic acids.* Since RNA molecules carry deep phylogenetic signal and the arrow of time in their structures, we have been able to derive historical accounts of molecular evolution directly from structural topology and thermodynamics [8, 41–44]. The evolutionary signal that we mine exists because the secondary structure is closely linked to structural conformation and dynamics [45]. RNA folding is negatively correlated with chain length, and the frustrated energetics and dynamics of folding allows only few conformations to reach stable states [46]. This forces structures to collapse by quickly reaching local folding solutions, which result in the formation of a number of helical structural modules compatible with the length and history of the molecules. Since the folding process is frustrated, numerous folding conformations are possible in molecules with randomized sequences. However, the number of possible conformations is actually culled by evolution to ensure that their average life is sufficiently long for the molecules to hold durable molecular functions [47]. This link between molecular evolution and the biophysics of RNA provides a rationale for our phylogenetic methodology: (1) characters describe features of helical stem and non-paired segments of RNA, and (2) minimization of conformations in RNA provides a definition of ‘evolution’s arrow’ for rooting of trees. Our methods make use of a census of geometrical features that measures the length and topology of RNA substructures, including stem and non-paired segments, or statistical features portraying stability and conformational diversity. The census produces data matrices with rows and columns representing molecules and molecular parts and phylogenetic characters describing molecular length or statistical features of structure


(a)
Age of FSFs (Gy)
3.8
3.0
2.0
1.0
0.0
Rise of Archaea
Reductive evolution of Bacteria
Functions linked to multicellularity
Replication
Tripartite world
Translation
Reductive evolution of Archaea
Oxygenic photosynthesis
Rise of metallomes
Cell adhesion, receptors
Metabolic big bang
Aerobic metabolism
Massive fission of domains
Rise of non-combinable domains
Big bang of domain combinations
New multifunctional modules
Ancient multifunctional modules
Discovery gap
Epoch 1
Epoch 2
Epoch 3
E
A
B
AE
AB
BE
ABE
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
(b)
Age of FSFs (nd)
0.0 > f < 0.1 (529)
0.1 > f < 0.5 (623)
0.5 > f < 0.9 (336)
0.9 > f < 1.0 (228)
f = 1.0 (17)
1,000
c.55.1
b.40.4
c.66.1
a.4.5
c.2.1
c.37.1
Origin of FSF domains
Archaea (A)
A (23)
AE (3)
AB (38)
ABE (786)
E (360)
B (164)
Eukarya (E)
BE (324)
Bacteria (B)

◀ **Fig. 2** The evolution of the protein world is visualized by studying its structural domain components. **a** Timeline of evolutionary appearance of fold superfamilies (FSFs) of structural domains describing the relative timing of important events in the history of life. Domain age was measured as a relative distance in the number of nodes from the base of the tree (*nd*) or was placed in a geological time scale of billions of years (Gy) using a molecular clock of domain structures [22]. Information in speech balloons without pointers was taken from trees of domain and domain combinations [24]. Their relative location is approximate. The three evolutionary epochs of the protein world are shaded in light green (Epoch 1, architectural diversification), salmon (Epoch 2, superkingdom specification) and light yellow (Epoch 3, organismal diversification) [25]. Boxplots display the FSF age distribution for the seven possible taxonomic groups. **b** Phylogenomic tree of domains (ToDs) describing the evolution of 1,733 FSFs reconstructed from structural domain abundance in the proteomes of 981 organisms. The tree was used to build the timeline of panel **a**. FSF taxa are colored according to FSF distribution (*f*) in the proteomes that were surveyed and used as characters to build the phylogenomic tree [26]. The most basal FSFs are labeled with SCOP concise classification strings (*ccs*; e.g., c.37.1 is the P-loop containing nucleoside triphosphate hydrolase FSF). The Venn diagram shows FSF distribution in superkingdoms

(e.g., branching, stability, diversity). Since the matrices can be transposed, the data can be used to build phylogenetic trees of molecules (ToMs) and trees of substructures (ToSs). ToMs and ToSs are data-driven models of the history of the molecular system or its component parts, respectively. The comb-like topologies of ToDs allow building timelines of the appearance of parts in molecules. These timelines define a 'natural history' of nucleic acids. The origin and evolution of the most ancient RNA molecules have been studied in this way, including tRNA [8, 48, 49], SINE elements [44], the large and small rRNA subunits [17, 32, 33, 42, 43], 5S rRNA [50] and RNase P RNA [51].

Note that the most parsimonious trees that describe the evolution of proteins and nucleic acids are retained after computational searches of tree space using the Wagner algorithm. Optimal trees are unrooted. They are only rooted a posteriori using phylogenetic process models that comply with Weston's generality criterion [52]. This criterion states that as long as ancestral characters are preponderantly retained in descendants, ancestral character states will always be more general than its derivatives given their nested hierarchical distribution in the rooted trees. Tracing the distribution of structural domains in proteomes (the *f* summary statistic) on the taxa of a ToD reveals compliance with the workings of Weston's rule (Fig. 2). When rooting a ToL, character change in domain abundance should be sequentially nested, with the most ancient structures being abundantly present in all or almost all of organismal lineages and more recent structures present at more moderate levels in increasingly more restricted groups of lineages. The ToD reflects that pattern; the most ancient domain structures (taxa) at the base of the tree are the most widely distributed in proteomes. A tracing of character state changes in the corresponding ToP (which is a ToL) shows that indeed these taxa (now characters) exhibit the widest distribution with change preponderantly restricted to the base of the tree. Weston's patterns also unfold by studying the distribution of domains across superkingdoms of life (Fig. 2). The Venn group of domains that are shared by all life (ABE) is the most ancient taxonomic group. Their domains span the entire time axis and are the most widely distributed in genomes. The evolutionary

appearance of the BE group shared by Bacteria and Eukarya occurs much later, coinciding with the first reductive loss of an FSF in Archaea. Domain structures specific to superkingdoms appear halfway in the timeline. These patterns also comply with the expected nesting of lineages.

Operationally, the direct character polarization method roots the trees of proteins by assuming domain structures accumulate in the evolution of the protein world and roots the trees of nucleic acids by assuming conformational stability increases in evolution as structures become canalized (reviewed in [17, 53, 54]). Biologically, domain structures spread by recruitment in evolution when genes duplicate and diversify, genomes rearrange, and genetic information is exchanged. Similarly, nucleic acid base pairs increase the stability and expand the size of RNA structures to match the increasing interactions with the expanding proteins and protein complexes that are responsible for cellular and functional makeup. This is a process of accumulation and retention of iterative homologies, such as serial homologs and paralogous genes, which is global, universal and largely unaffected by proteome or molecular size. The operational rooting (when made most parsimonious) complies with Weston's rule, and the axiomatic validity of character transformation can and has been tested using a number of approaches, including thermodynamics, phylogenetics and multidimensional scaling, proving its mettle.

3 The Early Emergence of Proteins and Metabolism

The structural domains are considered the evolutionary units of proteins. However, lower levels of structural granularity (abstraction) such as secondary structures (e.g., helix, strand, turns) or supersecondary structures (e.g., αα-hairpins, ββ-hairpins, βαβ-elements) could also hold evolutionary history. Remarkably, phylogenetic analyses, numerical approaches or machine learning techniques give no indication that these other levels hold strong phylogenetic signal or represent evolutionary modules (but see [55]). This may simply stem from our inability to suitably identify structural or non-structural lower level motifs that are responsible for molecular change. In contrast, domains have been carefully analyzed, unified into homologous groups and organized into a hierarchy in several classifications, including SCOP and CATH. For example, the SCOP classification groups domains into fold families (FFs), fold superfamilies (FSFs), folds and protein classes in a hierarchical classification system of decreasing granularity. Domains with pairwise amino acid sequence identities of more than 30 % are unified into FFs, and those FFs that share similar structures and functions are further unified into FSFs. FFs and FSFs have common evolutionary origins. FSFs sharing similar arrangements of secondary structures in three-dimensional space are further unified into folds, and those that share similar overall designs are further grouped into protein classes. The common evolutionary origin of FSFs in folds has not been systematically tested. In turn, classes unify large groups of folds that do not have a common evolutionary history.

Given these considerations, ToDs built at different hierarchical levels of protein classification should be considered phylogenetic statements solely related to structural domain history at those particular levels. Other possible structural modules at lower or higher levels of the hierarchical molecular system require separate exploration. The information gathered from ToDs has been however revealing since their inception [20]. The global emergent picture of molecular evolution derived from domain history is largely congruent regardless of the level of abstraction or the classification system. The global historical patterns obtained by tracing molecular functions annotated to domain structures in the timeline summarized almost a decade ago [17] still hold in updated timetrees and new studies. Here we highlight some of these patterns (summarized in Fig. 2):

1. The oldest domains are fully dependent on cellular membranes. It is therefore likely that the first proteins emerged enclosed in membrane containers forming primordial cells and evolved from there to form the wide diversity of globular proteins that today contribute to the complex make up of cellular organisms.
2. The very early proteins are first associated with organic cofactors but only later involve transition metals as ligands. This suggests an organismal response to increasing energy demands of the ancient world.
3. The very early, massive and then protracted appearance of domains with enzymatic functions indicates that the central metabolism played a primordial role in the early evolution of life.
4. The early but relatively late discovery of proteins involved in translation, including aaRSs, elongation factors and r-proteins, has a metabolic origin and is interrupted by a "discovery gap" that probably involves a historical revision of the translation apparatus.
5. The relatively early rise of metallomes (the Zn-metallome appearing first) and the late rise of oxygenic photosynthesis coincide with the late rise of aerobic metabolism. This explains the existence of the Great Oxygenation Event (GOE) $\sim$2.5 Gy ago, which is strongly supported by the geological record.
6. Domains involved in the synthesis of DNA precursors and replication complexes appear late. This indicates a late transition from storage of information in RNA genomes to storage in DNA genomes of cellular organisms.
7. Domains with functions that are typical of Eukarya, including cell adhesion, receptors, chromatin structure and functions linked to multicellularity, appear late and gradually and involve multidomain proteins. This suggests that modern Eukarya established as an organismal supergroup quite late in evolution.

Furthermore, a careful study of the origin and evolution of domains and domain combinations in multidomain proteins indicates the existence of a 'big bang' of protein discovery coinciding with the rise of eukaryotic organisms [24]. The conclusions of this study still hold and explain biphasic evolutionary patterns that exist in proteins [56]. The trees showed that the first proteins had single domains and were multifunctional, all of which produced fusion-driven combinations. These domain combinations arose early in the timeline (during Epoch 1), were

functionally specialized and later dominated the protein world. In contrast, fission processes developed late, notably during the big bang of domain combinations. These fissions produced many derived multifunctional single-domain proteins in Eukarya. The cyclic pattern of distribution of biological function along the architectural timeline is remarkable and reveals the emergence of a new class of protein module in evolution [17].

A major corollary from our phylogenomic studies is that the process of accretion of domains in proteomes occurs pervasively in nature and is a driving force for the evolution of macromolecules and life. Accretion is gradual, follows a molecular clock, and reconciles biology and planetary history. This finding crucially supports the *principle of spatiotemporal continuity*, the fundamental axiomatic necessity of evolution. We note that one could argue that the mere reconstruction of phylogenetic trees implies per se the gradual appearance of biological entities in evolution, i.e., that well-resolved tree topology cannot test spatiotemporal continuity. This is not so. The existence of a comb-like tree is an outcome of the existence of phylogenetic signal in the data and the existence of a timeline of natural structural discovery. Absence of such historical information would collapse branches into 'hard' polytomies, i.e., nodes supporting more than three branches with splits that arise from natural phenomena. These polytomies would distort the unbalanced tree structures toward a 'star' tree topology, making the construction of timelines impossible. The fact that we detect strong phylogenetic signal in the data diffuses such concerns. Furthermore, the molecular clock of folds extends the timeframe of domain diversification to the vast majority of the geological record. This supports the gradual spread of domain innovation in evolution. The recent mathematical modeling of the accretion process now makes the entire evolutionary process of protein domain accumulation explicit and prompts an exploration of how protein diversity extends through sequence space [57].

4 Insights into the Generation of the First Protein Structures

In a relatively recent study, we mapped the first evolutionary appearance of the oldest 54 FFs, tracing a number of properties of these domain structures, including their ability to bind cofactors, interact with RNA, and display broad molecular movements and flexibility [58]. These primordial FFs are important since they are responsible for jumpstarting metabolism and translation. Remarkably, their order of appearance provided detailed information about which central biological processes of the cell came first, metabolism, translation or replication, and what sub-processes were involved. The first four FFs were the ABC transporter ATPase domain-like family (c.37.1.12), the extended and tandem AAA-ATPase domain families (c.37.1.20 and c.37.1.19) and the tyrosine-dependent oxidoreductase domain family (c.2.1.2). All of these FFs currently unfold in membrane-structured cellular

environments. A detailed tracing of these structures in metabolic sub-networks defined by the KEGG database showed that these FFs provide hydrolase and transferase functions needed for nucleotide interconversion, storage and phosphate transfer-mediated recycling of chemical energy [29]. They are ultimately responsible for seeding the pathways of purine biosynthesis and establishing the chemical currency of energy storage in the biological world, the ATP and then GTP families of cofactors. Note that three of the four FFs hold the P-loop containing the nucleotide triphosphate (NTP) hydrolase fold (c.37), which is placed at the very base of each and every one of the ToDs we have ever generated. In the timeline, it appeared for the first time associated with a primordial bundle, the predominant structure of proteins associated with membranes. The archaic association of the "Rossmann-like" α/β/α-layered design that is typical of the c.37 and c.2 folds and the bundle structure was even made explicit in ToDs generated using CATH domain definitions, which split the SCOP FFs structure into finer grained modules [59]. Thus, the structural phylogenomic statements derived from structures appearing at the base of ToDs establish that the origin of proteins was unequivocally associated with metabolism and membranes. Thus, Dyson's "*more ancient cooperative structure of homeostasis*" typical of protein enzymes of metabolism indeed preceded the "*tyranny of the replicators*" underlying a nucleic acid-based genetic system [1], debunking the widely held belief of an ancient RNA world. The consequence of this finding is that first proteins had to unfold in the absence of genetic memory within cellular compartments.

An early appearance of peptide and protein molecules in cellular compartments is not an alien concept. Prebiotic chemistry supports the facile production of amino acids (even in artificial spark discharge experiments) and short peptide molecules (even in simple cycle desiccation experiments), which is much simpler than the synthesis of nucleic acid precursors. Amphiphilic molecules capable of forming vesicle containers are even present in meteorites. These emerging molecular systems are prone to hold molecular and cellular memory. Cellular compartments that are stabilized by addition of peptides could be more persistent [58]. Similarly, biases in self-catalyzed ligations of short peptides could result in longer and more stable emergent structures [60, 61]. These are hallmarks of 'homeostasis,' 'competitive optimization' and 'compositional selection.' Such forces could impart archaic memories about the expanding cellular and molecular systems.

If these conjectures are true, then we must invoke an ancestral 'origami' responsible for the generation of the first stable structural domains, which assembled from ancient peptides [62, 63]. Would this origami point toward the primordial α/β/α-layered structure present in the c.37 and c.2 Rossmann-like folds? We already have an answer! The use of advanced bioinformatics methods to survey and classify modular-like arrangements of helix, strand and turn segments $\sim$25–30 amino acid residues long identified the most conserved loop-forming building blocks [64]. Remarkably, the most popular of these structural motif prototypes (known as 'elementary functional loops') in archaeal proteomes and the most widely spread in fold superfamily domain structures preferentially involved superfamilies holding the c.37 and c.2 folds. A tracing of the bipartite network of elementary functional

loops and domain superfamilies in time showed patterns of emergence of modular scale-free behavior [65]. The ancient link between peptides and structural domains is therefore established and must be further studied.

5 Late Evolutionary Appearance of First Structural Domains Interacting with RNA

The structural domains that consistently appear at the base of the ToDs do not interact with nucleic acid macromolecules. Instead, the first nucleic acid-interacting domains made their debut relatively late, had metabolic origins and associated with tRNA [30, 58]. When studying the timeline of FFs, a number of FF domain structures appear after the rise of metabolism ~3.7–3.6 Gy ago (nd_{FF} = 0.02–0.045). The first four FFs of this group involve the class I aaRS catalytic domain (c.26.1.1), class II aaRS and biotin synthetases (d.104.1.1), G proteins (c.37.1.8) and actin-like ATPase domain (c.55.1.1) FFs [58]. These structural domains, which are also part of the catalytic makeup of enzymes important for fatty acid biosynthesis, appear before r-proteins in the timeline. All of them have the α/β/α-layered Rossmann-like design, and three of them define the catalytic domains of aaRSs and structures of elongation factors that are central for translation and the specificity of the genetic code. They catalyze crucial acylation and condensation reactions involved in the aminoacylation of tRNA bound to the aaRSs or phosphopantetheinyl arms of carrier proteins that are part of non-ribosomal peptide synthetase (NRPS) complexes.

6 The Co-evolutionary History of Emerging tRNA, Ribosomes and Proteins

Having established that translation started late by laying down a foundation of interactions among tRNA, aaRSs and factors, can we explore patterns of molecular growth indicative of the processes behind the rise of translation and the specificities of the genetic code? Phylogenomic analysis of thousands of RNA molecules and millions of protein structural domains supports three crucial historical patterns: (1) the co-evolution of tRNA and aaRS enzymes during the rise of genetic code specificities, (2) the co-evolution of ribosomal RNA and proteins, and (3) the co-evolution of tRNA and the emerging ribonucleoprotein structure of the ribosomes. We here define co-evolution as a coordinated succession of structural changes occurring within the emerging molecular environment. These changes should be considered mutually induced by the increasing interactions between and among protein and nucleic acid molecules that were being recruited to perform the very initial molecular functions. In all cases, co-evolution's goal was to fold

macromolecules into more stable and functionally efficient structures capable of extending the persistence of the molecules and the emergent primordial cells that would contain them. In these phylogenomic studies, the relative ages of structures of tRNA, rRNA, aaRS domains and r-protein domains were calculated from the phylogenetic trees (ToSs, ToMs and ToDs), indexed with structural, functional and molecular contact information and mapped onto three-dimensional models of molecules and molecular complexes.

The rise of the genetic code. The specificity of translation and the 'memory' of genetics is ultimately controlled by the specificities that define the genetic code. In vitro studies have shown that discrimination against non-cognate substrates is maximal in aminoacyl-tRNA synthesis, unknown but probably significant for EF binding and minimal for aaRS editing, aaRS resampling and ribosomal tRNA recognition and proofreading [2]. The rate of misincorporated amino acids in aaRSs is 1 in 200–10,000, at least an order of magnitude lower than other specificities, and the rate of misincorporated tRNA is 1 in more than 10,000. It is therefore clear that genetic code safekeeping has been entrusted to aaRSs and not the ribosomes. Reconstruction of phylogenies and evolutionary timelines showed that the history of catalytic, editing and anticodon-binding domains of aaRSs matched the history of tRNA charging and encoding [31]. The catalytic domains, which are the most ancient of the aaRSs molecules [30], interact with the acceptor arm of the tRNA that charges specific amino acids, which is the most ancient of the nucleic acid molecule [8] (Fig. 3). Similarly, the more recent anticodon-binding domain of aaRSs interacts with the more recent anticodon-binding arm of tRNAs. These co-evolutionary patterns that are derived from ToDs and ToSs can be complemented with more powerful tools that couple ToMs and phylogenetic constraint analysis to fine-grain the evolutionary history of the charging and encoding functions of translation [48, 49]. This allowed making historical inferences of the progression of specificities for both the 'operational' genetic code of the acceptor arm of tRNA [66] and the more derived 'standard' genetic code of the anticodon-binding stem of tRNA. The rise of the aminoacylation specificities of tRNA isoacceptors is described in the timelines of Fig. 4. The first specificities unfold by pre-transfer and post-transfer editing and trans-editing activities of aaRSs. These molecular activities are responsible for sieving amino acids by size in the active sites of the catalytic domains. They involve 11 of the 20 standard amino acids. Specificities are however split into two groups. Group 1 specificities associate with the older 'type II' tRNA structures holding a variable arm. Group 2 specificities associate with the standard 'type I' tRNA cloverleaf structures that lack the variable segment of the structure. These interactions, which unfolded $\sim$3.7–3.0 Gy ago, involve the acceptor stem of the tRNA molecule and probably defined the 'operational' genetic code in the absence of a fully functional ribosome and a full cloverleaf structure. In turn, codon specificities unfolded $\sim$3 Gy ago with the first anticodon (AC) binding domains, which interact with the more modern anticodon stem of tRNA. The evolution of this more modern 'standard' genetic code produced its own timeline of codon specificities that sometimes overlapped and enhanced the specificities of the 'operational' code (Fig. 4). Separate timelines of amino acid charging and codon

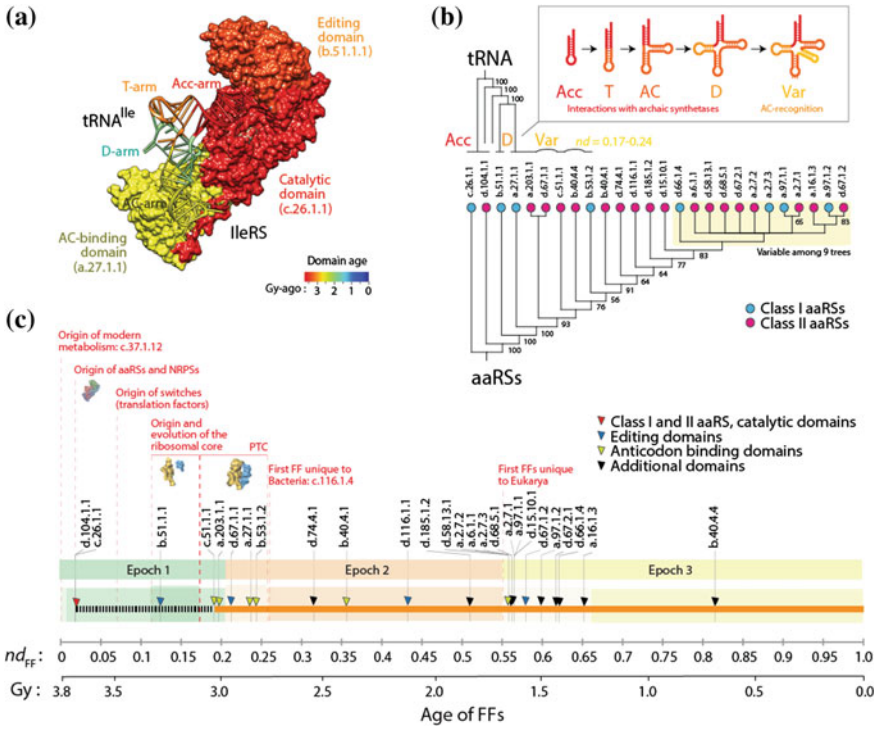


Fig. 3 The co-evolutionary history of tRNA and aaRSs. **a** The age of aaRS domains, exemplified by IleRS (PDB entry 1qu2), matches the age of the interacting arms of their tRNA isoacceptors. The oldest acceptor (*Acc*) arm interacts with the oldest catalytic domain and the more recent anticodon (*AC*) arm interacts with the more recent AC-binding domain. **b** One of nine most parsimonious phylogenomic tree reconstructions describing the history of aaRS domains [31]. Terminal leaves are colored according to aaRS class and indexed with domain *ccs* labels. The tree matches the corresponding subtree in the global tree of FFs described in the next panel. A tree of tRNA substructures describing the evolutionary growth of tRNA (made explicit in the *inset*; [8]) is mapped to the domains that interact with the unfolding tRNA substructures, showing tight co-evolution. **c** Timeline of FF domains directly obtained from a ToD reconstructed from information in the proteomes of 420 free-living organisms [31]. FFs (*indexed arrowheads*) are mapped along a timeline with landmarks derived from the domain history. The *dashed black segment* of the timeline indicates the aaRS history prior to the appearance of AC-binding domains and modern genetics. The three epochs of the protein world (described in Fig. 2) are *shaded*

recognition are therefore evident in the phylogenomic chronologies. Remarkably, a recent study shows that the acceptor and anticodon stem determinants encode the size and polarity of amino acid residues, respectively [67]. This matches the differential encoding of information in the top and bottom half of the tRNA molecule and the role of editing and anticodon binding recognition that differentiate these two sequential and apparently redundant codes [31]. This congruence supports the separate development of two genetic codes in evolution. A comparison of amino acid and dipeptide compositions of single-domain proteins appearing in the

Fig. 4 The history of the operational and standard genetic codes unfolds sequentially but the codes act redundantly. The operational code delimits amino acid charging, and the standard code delimits codon specificity. Phylogenomic analysis dissects their history [31, 49]

timelines before and after the first anticodon binding domains (i.e., the standard code) revealed enrichment of dipeptides with amino acids that are subject to aaRS editing (groups 1 and 2) [31]. Results uncover a hidden link between the emergence and expansion of the classic genetic code and protein flexibility [31].

The rise of the ribosome. Domain history indicates that r-proteins appeared 3.3–3.4 Gy ago, later than aaRSs and factors but earlier than anticodon binding specificities. The ribosome was therefore present while the 'operational code' was being developed. Since the small (SSU) and large (LSU) subunits of the ribosome contain 30–40 and 30–45 proteins, respectively, r-protein history unfolds considerable detail about the origin and evolution of the ribosome. Similarly, SSU and LSU hold about 50 and 100 universal helical segments, respectively, which can also provide details about the evolutionary growth of the RNA molecules. Indeed, ToDs and ToSs enabled construction of detailed timelines of the history of r-proteins and nucleic acids, respectively [17, 32, 33, 43, 50]. More importantly, the structural interactions present in models of the atomic structure of the ribosome permitted mapping interactions in both timelines, effectively linking the two. Remarkably, the exercise showed strong co-evolutionary relationships between the age of r-proteins and the age of interacting rRNA helices in the universal ribosomal core [32, 50], which were expressed as a significant correlation (Fig. 5). The oldest proteins (S12, S17, S9, L3) appeared together with the oldest rRNA substructures responsible for decoding and ribosomal dynamics. These structures include the ratchets and two hinges of SSU rRNA and the L1 and L7/L12 stalks important for ribosomal movement of tRNA in the complex. As the ribosome continued to unfold in evolution, the age of rRNA helical regions in both subunits (see Fig. 5) and interacting domains of r-proteins co-evolved simultaneously to form a fully functional ribosomal core. Importantly, the appearance of RNA substructures at first occurred in orderly fashion until the formation of five-way LSU and ten-way SSU junctions in SSU and LSU, respectively, at which point a 'major transition' in ribosomal evolution occurred 2.8–3.1 Gy ago (Fig. 6). This transition, which coincided with the start of planet oxygenation [28], brought ribosomal subunits together through inter-subunit bridge contacts [32]. It also stabilized loosely evolving ribosomal

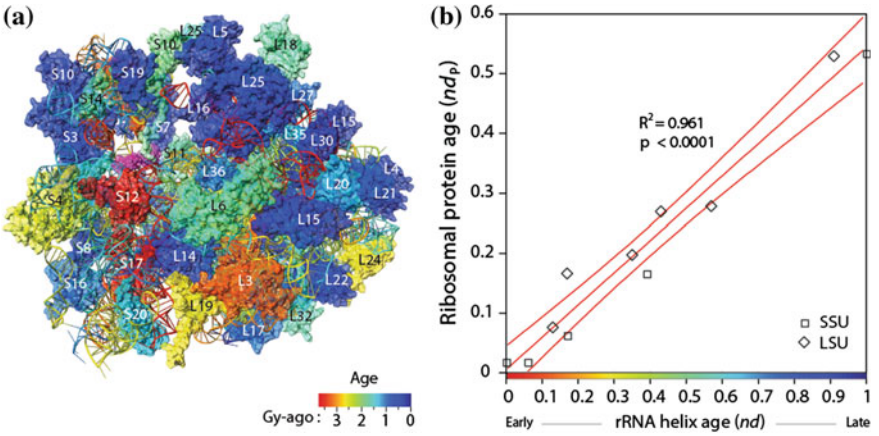


Fig. 5 The molecular evolution of the ribosome. **a** Secondary structure models of the small (*SSU*) and large (*LSU*) subunits of the *Escherichia coli* ribosome with rRNA helical segments and proteins colored according to their age (in Gy). Note the very ancient and central translocation core of helix 44 and r-proteins S12 and S17 (colored red) develops into a complex patchwork of molecular ages. **b** Tight co-evolution between r-proteins and rRNA helical segments. The relative ages are expressed as node distances (*nd*) derived from ToDs and ToSs. Figure modified from [32]

components and developed tRNA-interacting structures and a fully-fledged PTC with exit pores capable of protein biosynthesis. The implications of these co-evolutionary patterns of ribosomal history are profound. They debunk the idea of an origin of the ribosome in an ancient 'RNA world' since the growth of RNA and protein structure occurred in close interaction.

tRNA is at the center of ribosomal evolution. The timelines of ribosomal history showed that tRNAs were the centerpiece of important structures that were being accreted [32]. The gradual development of tRNA-rRNA molecular interfaces revealed that known interactions occurring before the major transition involved contacts between ancient SSU helices and the anticodon arm of tRNA. After the transition, most contacts involved newer LSU helices and the older half of tRNA. Contacts with the T-arm of tRNA formed soon after the transition. The T-arm is the only tRNA substructure that interacts with the two major subunits of the ribosome. Importantly, all tRNA contacts with the PTC unfolded abruptly during the major transition. Coupling the evolutionary timelines of tRNA and rRNA structure with annotations of their interactions with protein domains revealed that the tRNA cloverleaf structure was already fully formed when the PTC made its appearance [68]. Thus, fully formed tRNA molecules played other roles before being recruited for protein biosynthesis, perhaps both as cofactors of peptide-producing dipeptidases and ligases [31, 58] and as primordial genomes [69].

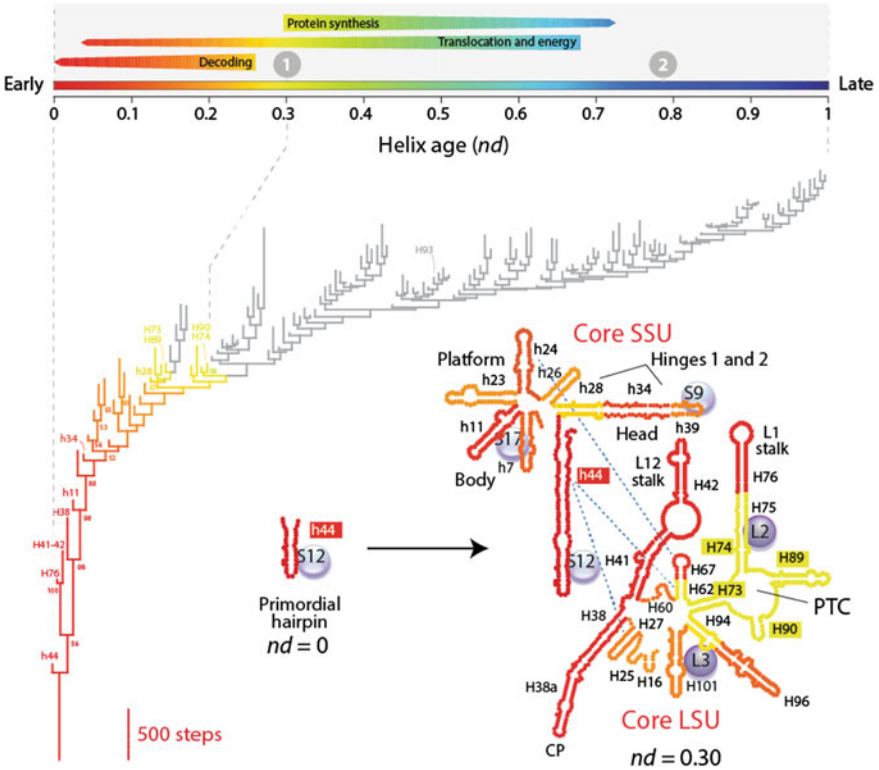


Fig. 6 Timeline of rRNA history. The first (major) and second transitions are indicated with *encircled numbers* in the timeline of rRNA substructures, which unfolds in time from *left* to *right* and is indexed with molecular functions. The timeline was inferred from a ToS, which is shown below. The branches of the ToS leading to the major transition are colored according to the age of evolving substructures. The major transition occurred once the decoding apparatus was in place, the H73, H74, H89 and H90 of the LSU formed the PTC responsible for protein synthesis (*yellow region* of the tree), and inter-subunit bridges (*dashed lines* in the model) were brought together and stabilized the SSU and LSU subunits. A model of the ribosome at the time of the major transition (*nd* = 0.3) is shown below the ToS with secondary structures colored according to their age. r-Proteins are indicated with *labeled buttons*. The growth of helical segments was modeled with growth rates of 100 base pairs/nd (~26 base pairs/Gy) and an average start length of 15.9 ± 11 (SD) bp to assume recruitment

7 tRNAs Are Evolutionary Building Blocks of Ribosomes and Genomes

A recent study generated lists of non-overlapping pairwise global alignments between tRNA and rRNA molecules that identified a number of remote homologies, which were often overlapping [70]. Similarly, sequential and overlapping remote homologies were detected between reconstructed tRNA and the PTC core of

LSU rRNA [71, 72]. These results suggest that both subunits of the ribosome were built piecemeal from primordial tRNA molecules. They also support an early proposal that the PTC originated from two tRNA halves by ancestral duplication [73] and even an earlier proposal supported by early bioinformatics analyses that tRNA and rRNA shared a common history [74]. Remarkably, we recently explored how the putative tRNA accretion process gave rise to functional rRNA by tracing the age of rRNA regions associated with the isoacceptor tRNA relics [75]. The ages of rRNA were taken directly from the work of Harish and Caetano-Anollés [32]. Remarkably, tRNA relics were enriched in older regions of the rRNA molecules, and these older regions harbored isoacceptor tRNA homologies that were also enriched in the oldest group 1 and 2 editing specificities for amino acid charging [31]. Thus, it appears tRNA relics preserve information about charging functions developed during the rise of the 'operational' code. What is even more remarkable is the existence of remote homologies to genes encoding very old proteins of metabolism, translation and replication that are also hidden in rRNA [70]. Thus, ancient rRNA had dual roles. It acted as a macromolecular machine or as a genome capable of encoding the information that the machine translated into proteins.

8 Conclusions

Translation is a biological process of interpretation of genetic information for the biosynthesis of proteins. Structural phylogenomic analysis suggests translation is ancient but developed later than the most primordial enzymatic functions of metabolism. Interactions with tRNA involve domains that were not at the base of the phylogenomic trees. Even the most ancient translation-related domains had metabolic functions (e.g., amino acylation of tRNA in catalytic domains), which preceded ribosomal-mediated protein biosynthesis. This has profound consequences for our understanding of how the molecular machinery of the cell originated. In current efforts to jumpstart a cellular system in vitro with the tools of synthetic biology, the "*cooperative structure of homeostasis*," which is embedded in proteins and cellular structure, must be established first, before ever attempting to impose a "*tyranny of replicators*" on the emerging system. Bioengineering should interface with knowledge from evolutionary history.

We note that the historical explorations we here describe started almost two decades ago. Their premise is that phylogenetic history exists in the structure of extant molecules. Its approach is grounded in cladistic methodology widely applied to the systematic survey of organismal biodiversity. Inferences about molecular structure are made with state-of-the-art HMM methods taking advantage of genomic information that is increasingly available. Phylogenetic trees are built using algorithmic implementations that extract deep phylogenetic signal from protein and nucleic acid molecules. Our studies have been followed by a handful of explorations from other laboratories, including building trees of life [76, 77], tracing domain changes in their branches [78] or constructing databases of structures

present in the last universal common ancestor of life [79]. Some explorations have been misguided by the use of unrealistic evolutionary models [80]. Since cladistics offers an objective criterion to reconstructing history, explorations follow the hypothetico-deductive method for overthrowing theories that supports scientific growth [81]. The strength of relationships of homology is tested at every stage of the exploration. The goal is to enhance the breath and scope, universality and degree of precision of the evidence that supports the historical conjectures. The effort increases explanatory power, empirical content and degree of corroboration. In the process, phylogenomics has repeatedly falsified the ancient "RNA world" theory in favor of other alternatives. The exercise attempts avoidance of recently highlighted fallacies that exist in the origin-of-life research field [82]. The experimental research of this field, which is predicated on deductive logic, appears largely immature, lacks "*patterns of progress*," and cannot integrate empirical evidence and theory from many domains of inquiry. Uncertainties in origin-of-life research are a "*breeding ground for a proclivity to combine wild speculation with dogmatic defense*" [82]. This explains a number of pernicious tendencies, including the adoption of extreme skepticism, collapse into metaphysics, and retreat to aprioristic narration and mythology. Phylogenomics provides one avenue out of the impasse. This avenue can systematize knowledge about the natural history of biological molecules and life.

Acknowledgments Computational biology is supported by grants from NSF (OISE-1172791 and DBI-1041233) and USDA (ILLU-802-909) to GCA. DCA is the recipient of NSF postdoctoral fellowship award 1523549.

References

1. Dyson F. Origins of life. Cambridge: Cambridge University Press; 1999.
2. Reynolds NM, Lazazzera BA, Ibba M. Cellular mechanisms that control mistranslation. Nature Rev Microbiol. 2010;8:849–56.
3. Francklyn CS, Minajigi A. tRNA as active chemical scaffold for diverse chemical transformations. FEBS Lett. 2010;584:366–75.
4. Hoeppner MP, Gardner PP, Poole AM. Comparative analysis of RNA families reveals distinct repertoires for each domain of life. PLoS Comput Biol. 2012;8:e1002752.
5. Fitch WM, Upper K. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harb Symp Quant Biol. 1987;52:759–67.
6. Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress A, von Haeseler A. How old is the genetic code? Science. 1989;244:673–9.
7. Di Giulio M. The phylogeny of tRNA molecules and the origin of the genetic code. Orig Life Evol Biosph. 1994;24:425–34.
8. Sun F-J, Caetano-Anollés G. The origin and evolution of tRNA inferred from phylogenetic analysis of structure. J Mol Evol. 2008;66:21–35.
9. Farias ST. Suggested phylogeny of tRNAs based on the construction of ancestral sequences. J Theor Biol. 2013;335:245–8.
10. Hennig W. Phylogenetic systematics. Urbana: University of Illinois Press; 1966.

11. Zuckerkandl E. The appearance of new structures and functions in proteins during evolution. J Mol Evol. 1975;7:1–57.
12. Dayhoff MO. The origin and evolution of protein superfamilies. Fed Proc. 1976;35:2132–8.
13. Almo SC, Garforth SJ, Hillerich BS, Love JD, Seidel RD, Burley SK. Protein production from the structural genomics perspective: achievements and future needs. Curr Opin Struct Biol. 2013;23:335–44.
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28:235–42.
15. Reddy TBK, Thomas A, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos E, Kyrpides N. The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 2014;. doi:10.1093/nar/gku950.
16. Wheeler WC. Systematics: a course of lectures. Hoboken: John Wiley & Sons; 2012.
17. Caetano-Anollés G, Sun F-J, Wang M, Yafremava LS, Harish A, Kim HS, Knudsen V, Caetano-Anollés D, Mittenthal JE. Origin and evolution of modern biochemistry: insights from genomes and molecular structure. Front Biosci. 2008;13:5212–40.
18. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247:536–40.
19. Orengo CA, Michie A, Jones S, Jones DT, Swindells M, Thornton JM. CATH–a hierarchic classification of protein domain structures. Structure. 1997;5:1093–109.
20. Caetano-Anollés G, Caetano-Anollés D. An evolutionarily structured universe of protein architecture. Genome Res. 2003;13:1563–71.
21. Nasir A, Caetano-Anollés G. A phylogenomic data-driven exploration of viral origins and evolution. Science Adv. 2015;1:e1500527.
22. Wang M, Jiang Y-Y, Kim KM, Qu G, Ji H-F, Zhang H-Y, Caetano-Anollés G. A molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Mol Biol Evol. 2011;28:567–82.
23. Laurin M. Recent progress in paleontological methods for dating the Tree of Life. Front Genet. 2012;3:130.
24. Wang M, Caetano-Anollés G. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. Structure. 2009;17:66–78.
25. Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal LE, Caetano-Anollés G. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res. 2007;17:1572–85.
26. Nasir A, Caetano-Anollés G. Comparative analysis of proteomes and functionomes provides insights into origins of cellular diversification. Archaea. 2013;2013:648746.
27. Caetano-Anollés G, Kim HS, Mittenthal JE. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci USA. 2007;104:9358–63.
28. Kim KM, Qin T, Jiang Y-Y, Chen L-L, Xiong M, Caetano-Anollés D, Zhang H-Y, Caetano-Anollés G. Protein domain structure uncovers the origin of aerobic metabolism and the rise of planetary oxygen. Structure. 2012;20:67–76.
29. Caetano-Anollés K, Caetano-Anollés G. Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism. PLoS ONE. 2013;8:e59300.
30. Caetano-Anollés D, Kim KM, Mittenthal JE, Caetano-Anollés G. Proteome evolution and metabolic origins of translation and cellular life. J Mol Evol. 2011;72:14–33.
31. Caetano-Anollés G, Wang M, Caetano-Anollés D. Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. PLoS ONE. 2013;8:e72225.
32. Harish A, Caetano-Anollés G. Ribosomal history reveals origins of modern protein synthesis. PLoS ONE. 2012;7:e32776.

33. Caetano-Anollés G, Caetano-Anollés D. Computing the origin and evolution of the ribosome from its structure—uncovering processes of macromolecular accretion benefiting synthetic biology. Comp Struct Biotech J. 2015;13:427–47.
34. Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anollés G. History of biological metal utilization inferred through phylogenomic analysis of protein structures. Proc Natl Acad Sci USA. 2010;107:10567–72.
35. Nath N, Mitchel JOB, Caetano-Anollés G. The natural history of biocatalytic mechanisms. PLoS Comput Biol. 2014;10:e1003642.
36. Debès C, Wang M, Caetano-Anollés G, Gratër F. Evolutionary optimization of protein folding. PLoS Comput Biol. 2013;9:e1002861.
37. Nasir A, Kim KM, Caetano-Anollés G. Global patterns of domain gain and loss in superkingdoms. PLoS Comput Biol. 2014;10:e1003452.
38. Kim KM, Caetano-Anollés G. The proteomic complexity and rise of the primordial ancestor of diversified life. BMC Evol Biol. 2011;11:140.
39. Caetano-Anollés G, Mittenthal JE, Caetano-Anollés D, Kim KM. A calibrated chronology of biochemistry reveals a stem line of descent responsible for planetary biodiversity. Front Genet. 2014;5:306.
40. Vandergon TL. Protein domain structure evolution. Molecular Life Sciences. New York: Springer; 2014. doi:10.1007/978-1-4614-6436-5_19-2 .
41. Caetano-Anollés G. Novel strategies to study the role of mutation and nucleic acid structure in evolution. Plant Cell Tissue Org Cult. 2001;67:115–32.
42. Caetano-Anollés G. Evolved RNA secondary structure and the rooting of the universal tree of life. J Mol Evol. 2002;4:333–45.
43. Caetano-Anollés G. Tracing the evolution of RNA structure in ribosomes. Nucleic Acids Res. 2002;30:2575–87.
44. Sun F-J, Fleurdépine S, Bousquet-Antonelli C, Caetano-Anollés G, Deragon J-M. Common evolutionary trends for SINE RNA structures. Trends Genet. 2007;23:26–33.
45. Bailor MH, Sun X, Al-Hashimi HM. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. Science. 2010;327:202–6.
46. Hyeon C, Thirumalai D. Chain length determines the folding rates of RNA. Biopys J. 2012;102:L11–3.
47. Fontana W. Modeling 'evo-devo' with RNA. BioEssays. 2002;24:1164–77.
48. Sun F-J, Caetano-Anollés G. Evolutionary patterns in the sequence and structure of transfer RNA: Early origins of Archaea and viruses. PLoS Comput Biol. 2008;4:e1000018.
49. Sun F-J, Caetano-Anollés G. Evolutionary patterns in the sequence and structure of transfer RNA: A window into early translation and the genetic code. PLoS ONE. 2008;3:e2799.
50. Sun F-J, Caetano-Anollés G. The evolutionary history of the structure of 5S ribosomal RNA. J Mol Evol. 2009;69:430–43.
51. Sun F-J, Caetano-Anollés G. The ancient history of the structure of ribonuclease P and the early origins of Archaea. BMC Bioinformatics. 2010;11:153.
52. Weston PH. Indirect and direct methods in systematics. In: Humphries CJ, editor. Ontogeny and Systematics. New York: Columbia University Press; 1988. p. 27–56.
53. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. The origin, evolution and structure of the protein world. Biochem J. 2009;417:621–37.
54. Sun F-J, Harish A, Caetano-Anolles G. Phylogenetic utility of RNA structure: evolution's arrow and emergence of modern biochemistry and diversified life. In: Caetano-Anollés G, editor. Evolutionary bioinformatics and systems biology. Hoboken: Wiley-Blackwell; 2010. p. 329–60.
55. Przytycka T, Aurora R, Rose GD. A protein taxonomy based on secondary structure. Nature Struct Biol. 1999;6:672–82.
56. Mittenthal JE, Caetano-Anollés D, Caetano-Anollés G. Biphasic patterns of diversification and the emergence of modules. Front Genet. 2012;3:147.
57. Tal G, Boca SM, Mittenthal JE, Caetano-Anollés G. A dynamic model for evolution of protein structure. J Mol Evol. 2016;82:230–243.

58. Caetano-Anollés G, Kim KM, Caetano-Anollés D. The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. J Mol Evol. 2012;74:1–34.
59. Bukhari SA, Caetano-Anollés G. Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. PLoS Comput Biol. 2013;9: e1003009.
60. Ikehara K. Possible steps to the emergence of life: The [GADV]-protein world hypothesis. Chem Rec. 2005;5:107–18.
61. Jakschitz T, Rode BM. Evolution from simple in- organic compounds to chiral peptides. Chem Soc Rev. 2012;41:5484–9.
62. Söding J, Lupas AN. More than the sum of their parts: On the evolution of proteins from peptides. BioEssays. 2003;25:837–46.
63. Trifonov EN, Frenkel ZM. Evolution of protein modularity. Curr Op Struct Biol. 2009;18:335–40.
64. Goncearenco A, Berezovsky IN. Protein function from its emergence to diversity in contemporary proteins. Phys Biol. 2015;12:045002.
65. Aziz MF, Caetano-Anollés G. The early history and emergence of molecular functions and modular scale-free behavior. Sci Rep. 2016;6:25058.
66. Schimmel P, Giege R, Moras D, Yokoyama S. An operational RNA code for amino acids and possible relation to the genetic code. Proc Natl Acad Sci USA. 1993;90:8763–8.
67. Carter CW Jr, Wolfenden R. tRNA acceptor stem and anticodon bases form independent codes related to protein folding. Proc Natl Acad Sci USA. 2015;112:7489–94.
68. Caetano-Anollés G, Sun F-J. The natural history of transfer RNA and its interactions with the ribosome. Front Genet. 2014;5:127.
69. Rodin SN, Rodin AS. On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. Heredity. 2008;100:341–55.
70. Root-Bernstein M, Root-Bernstein R. The ribosome as a missing link in the evolution of life. J Theor Biol. 2015;367:130–58.
71. Farias ST, Rêgo TG, José MV. Origin and evolution of the peptidyl transferase center from proto-tRNAs. FEBS Open Bio. 2014;4:175–8.
72. Farias ST, Rêgo TG, José MV. tRNA core hypothesis for the transition between the RNA world to the ribonucleoprotein world. 2016 (submitted).
73. Agmon I, Bashan A, Yonath A. On ribosome conservation and evolution. Israel J Ecol Evol. 2006;52:359–74.
74. Bloch D, McArthur B, Widdowson R, Spector D, Guimarães RC, Smith J. tRNA-rRNA sequence homologies: a model for the origin of a common ancestral molecule, and prospects for its reconstruction. Orig Life. 1984;14:571–8.
75. Caetano-Anollés G, Root-Bernstein R, Caetano-Anollés G. tRNA: building blocks of ribosomes and genomes. 2016 (submitted).
76. Yang S, Doolittle RF, Bourne PE. Phylogeny determined by protein domain content. Proc Natl Acad Sci USA. 2005;102:373–8.
77. Fang H, Oates ME, Pethica RB, Greenwood JM, Sardar AJ, Rackham OJ, Donoghue PC, Stamatakis A, de Lima Morais DA, Gough J. A daily-updated tree of (sequenced) life as a reference for genome research. Sci Rep. 2013;3:2015.
78. Edwards H, Abeln S, Deane CM. Exploring fold preferences of new-born and ancient protein superfamilies. PLoS Comput Biol. 2013;9:e1003325.
79. Goldman AD, Bernhard TM, Dolzhenko E, Landweber LF. LUCApedia: a database for the study of ancient life. Nucleic Acids Res. 2013;41:D1079–82.
80. Kim KM, Nasir A, Caetano-Anollés G. The importance of using realistic evolutionary models for retrodicting proteomes. Biochimie. 2014;99:129–37.
81. Farris JS. Parsimony and explanatory power. Cladistics. 2008;24:1–23.
82. Wächtershäuser G. In praise of error. J Mol Evol. 2016. doi:10.1007/s00239-015-9727-3.