[Frontiers in Bioscience 5212-5240, May 1, 2008]

Origins and evolution of modern biochemistry: insights from genomes and molecular structure

Gustavo Caetano-Anolles[1], Feng-Jie Sun[1], Minglei Wang[1], Liudmila S. Yafremava[1], Ajith Harish[1], Hee Shin Kim[1], Vegeir Knudsen[1], Derek Caetano-Anolles[1], Jay E. Mittenthal[2]

[1]Department of Crop Sciences and [2]Department of Cell and Developmental Biology, University of Illinois, Urbana, IL 61801, USA

TABLE OF CONTENTS

1. ABSTRACT

The survey of components in living systems at different levels of organization enables an evolutionary exploration of patterns and processes in macromolecules, networks, and genomic repertoires. Here we discuss how phylogenetic strategies that generate intrinsically rooted phylogenies impact the evolutionary study of RNA and protein components of the macromolecular machinery that is responsible for biological function. We used these methods to generate timelines of discovery of components in systems, such as substructures in RNA molecules, architectures in proteomes, domains in multi-domain proteins, enzymes in metabolic networks, and protein architectures in proteomes. These timelines unfolded remarkable patterns of origin and evolution of molecules, repertoires and networks, showing episodes of both functional specialization (e.g., rise of domains with specialized functions) and molecular simplification (e.g., reductive tendencies in molecules and proteomes). These observations have important evolutionary implications for origins of translation, the genetic code, modules in the protein world, and diversification of life, and suggest early evolution of modern biochemistry was driven by recruitment of both RNA and protein catalysts in an ancient community of complex organisms.

2. INTRODUCTION

The dynamic nature of life has two fundamental emerging properties, diversity and complexity. Diversity is fueled by the stochastic nature of change but is constrained by history, environment, and the architecture and thermodynamics of systems. Diversity in the long run increases complexity, taking evolving systems to new levels of architectural organization. Both diversity and complexity are difficult to define, especially as we leave the molecular realm. They represent relative concepts. For example, we tend to consider higher organisms more complex systems than microbes, yet the stylish molecular designs and the complex organismal communities uncovered in the microbial world are unrivaled. When did it all start to happen and how? The question is about origins and relates to the mapping of genotype, phenotype and fitness to each other. Here we focus on origins of biological macromolecules known to be important for change, because molecules are more tractable systems than those existing at higher levels of organization. In particular, we focus on proteins and RNA. They represent the molecular machinery that implements biological function in an organism, and they are clearly good starting material to uncover fundamental diversification patterns and processes

that could help explain more complex evolutionary phenomena. The recent 'omic' revolution has given us relatively accurate surveys of the macromolecular world, and progress in computational biology has provided the means to analyze its history and hierarchical complexity. Here we explore the origins and evolution of modern biochemistry, reaping the benefits of structural and evolutionary genomics and focusing on the process of discovery of components of molecular repertoires in the course of evolution.

3. EVOLUTION OF SYSTEM REPERTOIRES

3.1. Discovery of components in systems

Comparative and structural genomics offer unprecedented opportunities to understand genomic complexity. This is in part due to the massive and ongoing acquisition of nucleic acid sequence. Currently, ~800 genomes and metagenomes have been completely sequenced (December 2007) yielding millions of protein sequences and thousands of functional RNA molecules important for cell development and homeostasis. The number of ongoing genome sequencing projects (currently 2,327) is an indicator of exponential increase in years to come. This effort outpaces structural genomics (1-3) with ~47,000 three-dimensional (3D) models of molecular structure deposited in the Protein Data Bank (PDB). The benefit of all these developments is a census of component parts (e.g., genes, proteins, RNA, structures) and their interactions (e.g., networks, molecular machines, ensembles) that makeup molecular systems (e.g., genomes, proteomes, transcriptomes). Current research in bioinformatics attempts to mine, visualize and integrate these system repertoires. Any attempt to understand systems with their disparate components requires an evolutionary framework, because the structure and properties of these systems embed histories that reflect interaction with other systems and the environment. Consequently, insights into their structure and function can be gained by understanding how they evolved. This is a complex endeavor, one that is often at the interface of philosophy, mathematics and science. Here we use both standard and novel tools of phylogenetic reconstruction to generate timelines of discovery of components in molecular repertoires. We will describe methods to study the order of appearance of structural components in RNA molecules, architectures in proteomes, enzymes in subnetworks, pathways in networks, and proteomes in the protein world. Making sense of the discovery and use of these elements in life requires rooting of phylogenetic trees. Consequently, uncovering evolution's arrow is necessary for the success of such enterprise.

3.2. Evolution's arrow: reconstructing history from molecules

The numerical estimation of evolutionary relationships has been the subject of much study since the inception of bioinformatics in the 1960s. In general, most methods identify biological features (*characters*) that exhibit variation, give a numerically description to this variation, and then build graph representations (*phylogenetic trees*) that account for differences and similarities in the data. Phylogenetic trees depict ancestor-descendant relationships that illustrate the evolution of biological entities of any kind and build on the concept of *homology*, relationship between traits that are shared as a result of common ancestry (4). However, in recent years the study of morphological or biochemical features in organisms has been displaced by the study of molecular features, mostly nucleic acid and protein sequences. A number of approaches have been used to do this, including distance, maximum parsimony, maximum likelihood and Bayesian methods (5,6). Most of these methods rely on the use of *optimality criteria* that involve searching the space of all possible trees and identifying those that match both data and model of change according to some criteria, for example finding the simplest (most parsimonious) or the most probable (most likely) solution. However, the evolutionary history of individual sequences in certain occasions does not match that of organisms, usually because of differential loss, horizontal transfer of genetic information, reconstruction biases, or the use of incorrect models. Similarly, some sequences are not sufficiently conserved to uncover the desired phylogenetic relationships. With the advent of genomics, availability of other kinds of information beside sequences (e.g., gene content, gene order) prompted new methods and evolutionary models (6). Genomics also brought back important cladistic approaches that offer explicit and general definitions of biological relationship and had proven powerful for phylogenetic systematics and molecular evolution (7). Using these methods, genomic history has been reconstructed using combined or concatenated genomic sequences [e.g. (8-10)], and features describing the survey (genomic demography) (10-22) and arrangement (genomic topography) (16,23-27) of genomic component parts [reviewed in (28-30)]. In particular, whole-genome (phylogenomic) trees were built effectively from features describing the occurrence and distribution of protein fold architectures in proteomes (11-13,15,18,20-22,31). Figure 1 describes a tree reconstructed from abundance of protein domains at fold level encoded in genomes that have been fully sequenced, and Figure 2 a whole-genome tree built from the pair-wise combination of domains in proteins. Almost all universal phylogenies support the tripartite nature of life, with monophyletic groups corresponding to superkingdoms Archaea, Bacteria and Eukarya, already evident in trees reconstructed from ribosomal RNA molecules (32), confirm accepted lineage relationships within major organismal groups, support disputed or preliminary classifications, and reveal novel evolutionary patterns (30). In contrast, the use of phylogenetic methods to study the evolution of molecular repertoires has received limited attention. This stems probably from the absence of criteria with which to root phylogenetic trees that describe repertoires.

The current paradigm of phylogenetic reconstruction, illustrated with the construction of a universal tree of life with leaves (*taxa*) representing all known organisms and organismal groups (33), is to reconstruct unrooted trees using standard methods of phylogenetic reconstruction, and then use outlier groups (*outgroups*) to root the trees. Collections of these rooted

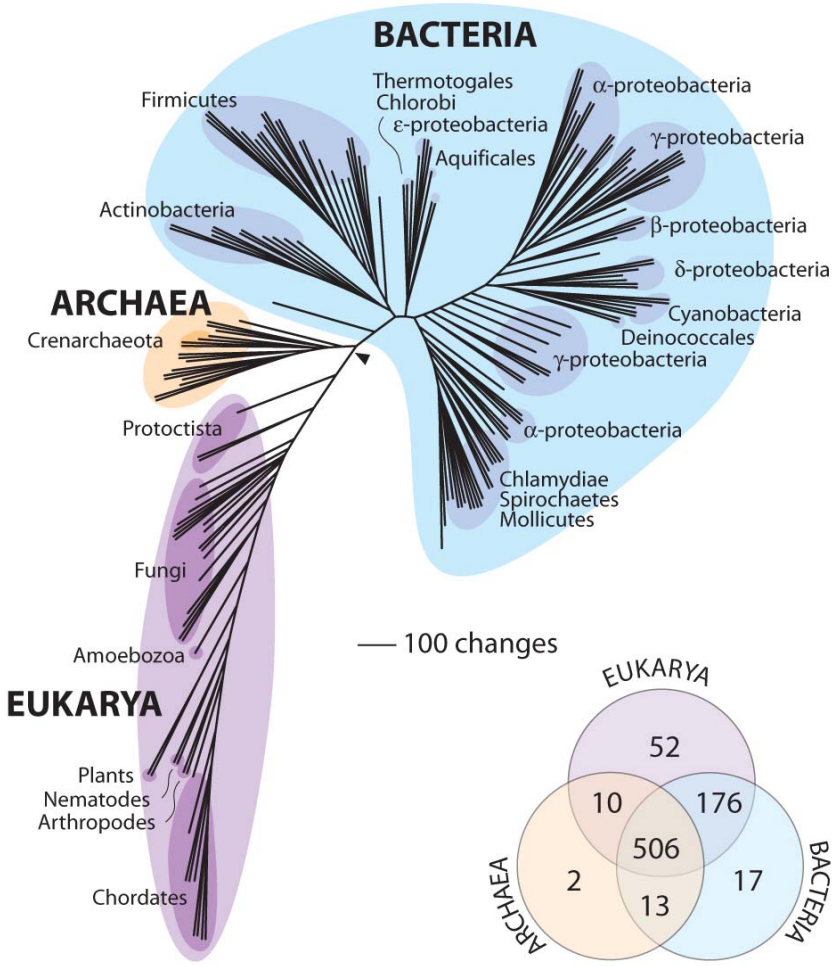

Figure 1. A universal phylogenomic *tree of proteomes* reconstructed from an analysis of protein domains at fold level belonging to 185 organisms that have been fully sequenced (21). Only one optimal tree of 115,639 steps was obtained using maximum parsimony as the optimality criterion (CI=0.133; RI=0.701; g-fit =–0.456). The tree shows three well-supported groups corresponding to the three superkingdoms of life. Terminal leaves are not labeled with organismal names as they would not be legible, and the arrowhead indicates the location of the root. Construction of the tree involves a structural census defined by advanced Hidden Markov Models (HMMs) that assigns domain structure to genomic sequences, as well as normalization of data, and phylogenetic analysis. The Venn diagram shows distribution of the 776 folds among superkingdoms of life.

trees are then assembled to generate a universal reconstruction. Outgroup taxa provide a direction to the evolutionary process and are generally chosen to be distantly related to the rest of taxa (*ingroup taxa*) by some external hypothesis of relationship. For example, in order to root organisms belonging to the grasses (Poaceae), one needs to identify organisms that are closely related (i.e. they are sufficiently distant not to warrant their inclusion as grasses but share a number of features with them). In this specific case outgroup taxa belonging to the Joinvillaceae have been used to establish an hypothesis of ancestral relationship with the grass ingroup. From a cladistic perspective, polarity of character states is based on the basic assumptions of the analysis and particular auxiliary assumptions relevant to the data (34). Examples include the use of ontological or palentological methods for outgroup determination and models of nucleic acid sequence evolution. Differences in basic assumptions can result in different approaches to polarity determination. The relevance of each method of polarization in particular instances depends on the accuracy of auxiliary assumptions. Most conventional hypothetical ancestors used to root trees have implicit analytical problems (35). They either represent character state information in one or more outgroup taxa, or use the ontogenetic and paleontological methods of polarizing character transformations. At times, inferred ancestors derived by combining these methods are used. However, the use of these hypothetical ancestors are invalid, as inferences regarding character states based on outgroup comparison apply to the outgroup node, whereas inferences based on either the ontogenetic or paleontological method apply to the ingroup node. These inferences cannot be combined into a single hypothetical construct. Inclusion of

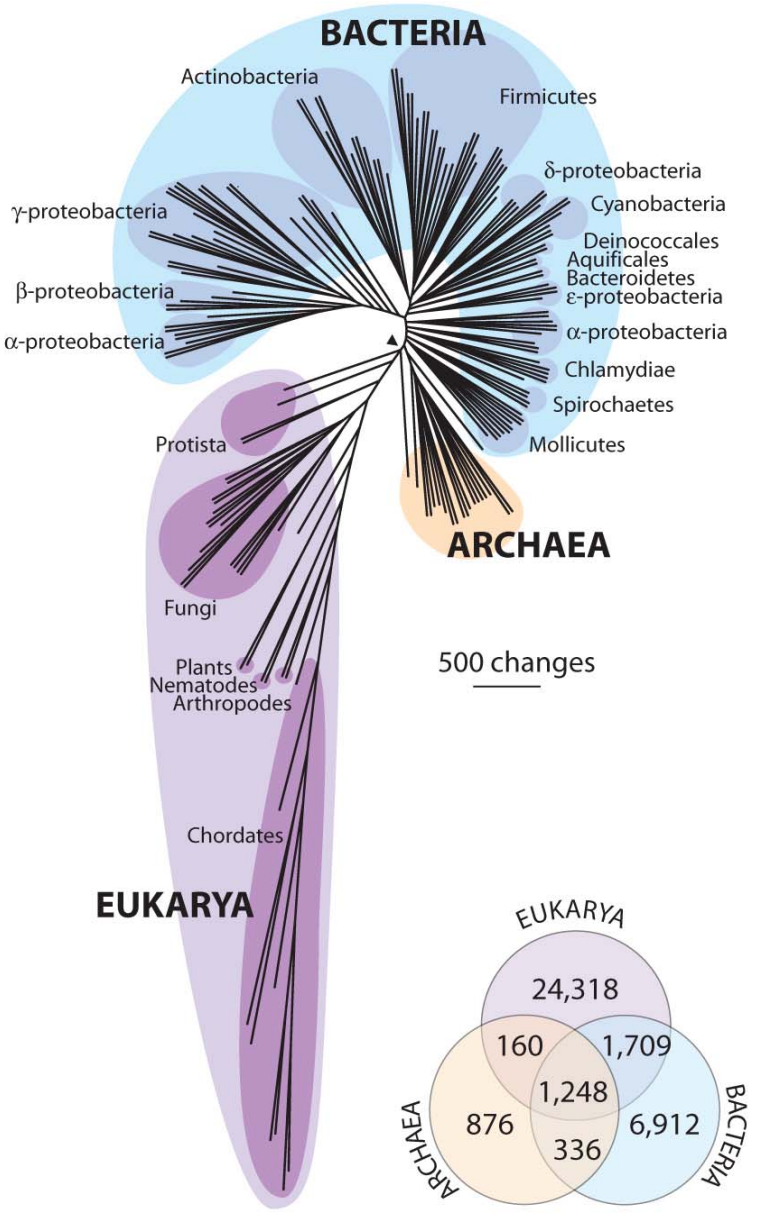

Figure 2. A universal phylogenomic *tree of proteomes* was reconstructed from an analysis of 35,559 pairwise domain combinations at fold superfamily (FSF) level in proteins belonging to 185 organisms. Only one optimal tree of 948,547 steps was obtained using maximum parsimony as the optimality criterion (CI=0.271; RI=0.537; g-fit=–1.033). Terminal leaves are not labeled with organismal names as they would not be legible. The arrowhead indicates the location of the root. The Venn diagram shows the occurrence of pairwise domain combinations at FSF level in proteomes.

hypothetical ancestors based on outgroup information in the data matrix to root trees can impose problematic constraints on the analysis. Thus, in most instances the actual outgroup taxa are preferable. The Lundberg rooting method in which the shortest ingroup network is rooted at the internode to which the hypothetical ancestor attaches most parsimoniously, is an appropriate use of a hypothetical ancestor inferred with the ontogenetic and paleontological methods. Under the current paradigm, reconstruction of a rooted phylogenetic tree requires data, a model of character change, an external rooting hypothesis, and generally a method to search for optimal trees. However, when taxa represent molecular features, it is often difficult if not impossible to find an appropriate outgroup and the only way to root the trees is to build a model of change that already provides an evolutionary direction. This is the most direct and powerful way to reconstruct rooted trees. Unfortunately, when studying nucleic acid and protein sequence there is no such model, because change at sequence level is highly idiosyncratic. In contrast, finding evolution's arrow at higher levels of

organization is possible and provides unanticipated benefits, as we will discuss below.

3.3. Fundamental assumptions for phylogenetic analysis

The function of molecules is shaped by evolution, generally (but not always) resulting from natural selection operating at high levels of structural organization. Because molecular information is ultimately structural (36), three general and fundamental assumptions relate to molecular structure and are important to approaches of phylogenetic reconstruction and rooting of trees: (i) molecular structure is far more conserved than sequence and carries considerable phylogenetic signal, (ii) there is a universal tendency towards molecular order, and (iii) successfully implemented biological designs tend to be reused in nature. A large body of evidence supports each and every one of these assumptions, selected aspects of which are described in the sections that follow.

We make evolutionary inferences from biological information that is in existence today, so our phylogenetic constructs reflect the myopic view of the present in relation to evolution of modern biochemistry. Any process that has the potential to erase history will bias our conclusions. Examples include evolutionary take-overs in which ancient molecules or processes are replaced completely by modern counterparts. This includes the replacement of pre-biotic chemistries, self-generating cycles, ribozymes, and ancient proteins by modern functional counterparts. Consequently, we consider our modern biochemistries as palimpsests that recapitulate earlier biochemistries (37) and prebiotic chemistries (36). In these cases, the overwritten history can only be sorted out by understanding the underlying processes of recruitment and the role of biological function in evolution.

3.4. Evolutionary conservation of structure: complex interplay between genetic robustness and evolvability of molecules

Structure is directly linked to function and is therefore the subject of natural selection and strong evolutionary constraint (38,39). Recognition that knowledge of high order structure is fundamental to establishing structure-function relationships in biological macromolecules led to structural genomic initiatives that seek to create a complete inventory of protein folds from crystallographic data (1-3). Clearly, 3D structure is less prone to being affected by mutation than nucleotide or protein sequences, and the information in structure is expected to persist longer than in primary sequence (40). It is always advantageous to study biological systems at the phenotype level. Genotypes have a limited alphabet that changes constantly by mutation and thus serve as poor repositories of molecular history. One clear example is nucleotide sequence. Theoretical considerations suggest that the repeated accumulation of substitutions in nucleotide sites (site saturation) erases evolutionary history at intermediate and deep evolutionary time scales (41-43). Other factors can also complicate evolutionary interpretation of genotypes, including convergent evolution of nucleotide sites, differing substitution rates among sites and lineages, and non-independent substitutions among sites (29,44). In contrast, phenotypes have more complex alphabets, which impact higher levels of biological organization. For example, they generally result from interaction of numerous components (substructural, molecular, macromolecular, etc), which are often carefully culled by natural selection. They constrain the functionality of the system and are therefore generally left unchanged over short and intermediate time scales. The phenotype of molecular structure is particularly remarkable. This phenotype is expressed at low levels of biological organization but generally constitutes a fundamental repository of biological function. Structure delimits function directly or defines interactions that are collectively responsible for function (e.g., in molecular ensembles). The effects of selection are consequently stronger at this level than at the genotype level.

Schuster *et al.* (45) were the first to map the relationships between sequence and structure in RNA molecules, defining a computationally tractable molecular landscape that can test evolutionary hypotheses. Since then, the approach has been extended to proteins and used to study how the interplay of natural selection, self-organization, and environment is responsible for the phenotype (46). Because of their unique chemistries, the mapping of genotype (sequence) to phenotype (structure) in proteins and RNA biopolymers offers different challenges, with three shared properties: (i) there are many more sequences than structures (i.e., the sequence-to-structure map is highly degenerate); (ii) few common but many rare structures materialize in structure space; and (iii) extensive neutral networks that percolate sequence space define common structures and structural neighborhoods (47,48). The existence of these neutral networks illustrates how structure can be impervious to mutational change at the sequence level. Because the distribution of sequences that fold into the same structure within neutral networks in RNA is approximately random, the mapping has "space covering" properties. This means that all structures can materialize within relatively few mutational changes in sequence space. This property has been confirmed experimentally using RNA functional switches (49). Computational studies also predict the existence of neutral networks and space covering for polypeptides (50), and experiments support the model (51). However, the sequence-to-structure mapping of proteins is much more complex and its landscape "holey," with protein conformations missing in vast segments of sequence space. While the neutrality of protein sequence space is much higher than that of RNA [>90% of single amino acid substitutions are neutral (52)], protein structures appear to concentrate in dense clusters (53,54) while RNA structures spread through sparsely connected networks (55).

Robustness, the ability of a system to cope with genetic or environmental change (46), represents an important evolutionary concept that is highly relevant to conservation of structure. It is ultimately a measure of the optimality of a system. Robustness is particularly important when it is heritable and relates to the phenotype. A system is genetically robust when its function and structure withstands the effect of mutation on its component parts

(56). Robustness applies to all systems, regardless of their level or organization. For example, robustness takes the form of error tolerance in complex biological networks that have scale-free properties (57), gene dispensability in genomic repertoires (58), or tolerance to amino acid substitutions (59). It is generally believed to be an emergent property of systems. The origin of robustness is contentious and depends on the different systems that have been studied. It requires that we estimate phenotype frequencies and the number of neutral neighbors of each phenotype, which is computationally challenging. An analysis of microRNA, important endogenous RNA regulators that control expression of protein-encoding genes, showed for example that robustness in these molecules represents a direct adaptation, i.e., robustness evolves directly by natural selection (60). A similar conclusion comes from a study of plant viroid RNA (61), where robustness increased with evolutionary time in a phylogenetic analysis of viroid families. In these cases, robustness appears driven by links between structure and the biochemical function of molecules. In contrast, other studies support an alternative scenario in which robustness is a by-product of other selective pressures, such as the correlated selection for mutational robustness and thermodynamic stability of RNA molecules (62,63). Here molecules with robust structures sustain the effect of mutation and are resilient to environmental perturbations because they were selected for these traits in the course of evolution (64). These two forms of stability, structural/thermodynamic stability and mutational stability, appear correlated (52,62,63,65,66). The term 'plastogenetic congruence' was used to describe this biophysical correlation between the effects of environment and mutation (62). For RNA, evolution of robustness required a direct link between thermodynamic and genetic stability; i.e., the set of structures that are thermodynamically accessible to a molecule will overlap significantly in sequence space with the set of structures accessible by point mutations. Interestingly, a number of studies have shown that biological evolution has produced RNA molecules that are highly robust and at the same time significantly more stable (60,61,63,67). Evolution and physicochemical laws are therefore intertwined.

One fundamental consequence of robustness is paradoxical. Robust molecules are optimized to withstand the effect of mutation but at the same time are evolutionarily locked into their structures. They are *structurally canalized*. In other words, the evolved molecules are intrinsically less capable of generating heritable phenotypic variation, and consequently, they are less "evolvable". However, experimental and computational analysis has shown that robustness enhances evolvability. For example, cryptic genetic variation can become evident in certain environments or genetic backgrounds (68,69). Recently, the antagonistic nature of robustness and evolvability were reconciled (56). Using RNA as a model system, robustness was measured both at genotype (sequence) and phenotype (structure) levels. Genotype robustness was defined as the number of neutral neighbors of a genotype that differ by one nucleotide and fold into the same structure. Phenotype robustness was defined as the number of neutral neighbors averaged over all genotypes with a given structure. While genotypic robustness and genetic evolvability shared an antagonistic relationship, structural robustness promoted evolvability. Structurally robust molecules had access to increased amount of phenotypic variation as they spread through neutral networks. In summary, the system evolved towards order but at the same time increased its inherent ability to generate change.

3.5. Universal tendencies towards molecular order

The robustness and evolvability dichotomy relates to entropy in living systems, its connection to order, and its impact on molecules. There is no doubt that living organisms and their components can be regarded as highly ordered systems. Erwin Schrödinger, searching for links between biology and physics, proposed that life involved building order from disorder (70). This concept was based on the idea that biological systems embed highly complex structures at many levels of organization, and these are produced from simple systems in processes that seem to defy the second law of themodynamics. Schrödinger also recognized that living systems were in non-equilibrium and maintained their highly ordered states by degrading energy coming from larger encompassing (external) systems efficiently, decreasing their own entropy levels at the expense of increasing entropy of the surroundings. Thermodynamic functions of state such as internal energy (U), enthalpy (H), entropy (S), and Gibbs free energy (G), which are characteristic of systems in equilibrium and are independent of their history, have been used to study links between thermodynamics and living systems. In this regard, the expansion of G of a biological system into its component parts (e.g., using G of base pairs or stems in algorithmic folding implementations to identify more stable RNA molecules) has been used to propose a thermodynamic theory of evolution that could explain the basic principles underlying biological change (71-75). More recently, living systems have been defined using better thermodynamic descriptors of energy gradients, such as exergy, a quality measurement of energy that describes the ability to establish energy gradients in non-equilibrium systems that are open to flows of energy and matter. They have been viewed as "the solution to the thermodynamic problem of increasing the degradation of the incoming solar energy, while surviving in a changing and sometimes unpredictable environment" (76,77). More specifically, living systems were seen as chemical factories that degrade incoming energy by producing and degrading molecular structures through autocatalytic, self-assembly, reproduction, evolution and adaptation processes. All these processes enhance the order of the system, which in turn enhances the quality of energy dissipation. As evolution progresses, more complex structures degrade incoming energy more efficiently, in doing so decreasing the establishment of energy gradients and opposing disequilibrium (in accordance with the second law of thermodynamics).

The maximization of exergy in living systems has important consequences for evolution of molecular structure. Nucleic acid and protein molecules have low information entropy in sequence space, but in structure

space highly evolvable phenotypes are also more entropic (56). In other words, if we consider sequence and structure as two different levels of organization, increasing the order in one level has the consequence of decreasing the order in the next. This interplay is important because it allows diffusive walks in neutral space and exploration of more efficient structures. This exploration results in diminished optimization but is beneficial since it enhances discovery of structural variants (78). The thermodynamic view of life has been extended from organisms to entire ecosystems (76,77), and we have borrowed them to polarize change and root our phylogenetic trees with generalized trends that we apply to the structure of molecules. A considerable body of theoretical and experimental evidence supports a tendency towards order in RNA molecules. For example, the study of extant and randomized sequences has shown repeatedly that evolution enhances conformational order and diminishes conflicting molecular interactions (*frustration*) over what is intrinsically acquired by self-organization (63,78-84). A molecular tendency towards order and stability has been tested experimentally using thermodynamic principles generalized to account for non-equilibrium conditions (71,74). Assumptions are independently supported by analytical models based on the reconstruction of the structural repertoire of RNA sequences from energetic and kinetic perspectives (48,62,85). Some predictions of these models have been confirmed experimentally (49). Correlation between RNA folding and the occurrence of structural motifs in natural nucleic acids also supports tendencies towards order (86). Finally, tendencies towards order have been experimentally supported by phylogenetic congruence in the reconstruction of rooted trees generated from sequence, structure, and genomic rearrangements (see below) at different taxonomical levels (78,87-94). Note that order is seldom achieved in frustrated systems that are driven by the energetics of conformation and stability, such as RNA or proteins.

3.6. Redundancy and reuse of successful biological designs

Biological designs that had been successfully deployed will have more chances to be reused in other biological contexts, and consequently, are expected to become popular (95). This results in redundancy, a desirable characteristic for "backing-up" biological functions that are important and need to be preserved. From our above discussion, robust and well-evolved designs have more chances of withstanding the effects of time, and at the same time, propagating through systems. In fact, redundancy has been shown to increase genetic robustness (96), a tendency that has also been confirmed in an analysis of plant viroids (61). Redundancy is associated with another important property of evolving systems, modularity. A module is a set of structures or components that cooperate to perform a task and interact more extensively with each other than with other structures or components outside the modular set. The autonomous nature of the modules usually emerges from structural canalization in molecules (62), which as we have seen is linked to robustness. The formation of modules is ubiquitous in the biological world, is evident at different levels of biological organization, and has important evolutionary consequences. For example, modularity is embedded for example in the structure of proteins. The search for vestiges of ancient structural motifs revealed the existence of closed loops of nearly standard size that are omnipresent in proteomes and define putative components of structure, which are modular and very ancient (97,98). At higher levels of structural complexity, proteins fold into a limited set of 3D conformations, but elements of structure such as alpha-helices and beta-strands within the folds adopt a "vocabulary" that is both highly modular and redundant (99). For example, short peptide segments are repeated in fibrous proteins such as collagen and the highly ubiquitous coiled coil domains. Similarly, larger peptide segments that form supersecondary structural elements are also repeated. These elements include alpha-alpha-hairpins, beta-beta-hairpins, and beta-alpha-beta-elements. When these elements are repeated in tandem they give rise to open-ended solenoid structures exemplified in leucine-rich repeat proteins. However, reuse of these supersecondary structures sometimes results in globular arrangements with our without internal structural symmetry. For example, the repetition of beta-alpha-beta-elements is embedded in TIM barrels, and beta-beta-hairpins are reused to build beta-propellers, designs that sometimes lead to symmetric configurations. Other globular arrangements are non-symmetrical, such as the globin-like fold and the OB-fold. Modularity and reuse occurs also in protein domains when they combine to form multi-domain arrangements (100) or the ability of protein subunits to form quaternary complexes (101). At a genomic level, domains are reused to perform different tasks and hundreds of copies of some architectural designs (e.g. TIM-barrel, the Rossmann fold, or the P-loop hydrolase fold) can be found in proteomes. Even here, modularity expresses in networks of molecular interactions of proteins, nucleic acids and small molecules involving protein complexes and dynamic functional units (e.g., signaling cascades) (102,103). Successful biological designs are therefore clearly preserved in nature through redundancy and modularity.

4. EVOLUTION OF MODERN RNA

4.1. Functional RNA

RNA molecules are ubiquitous, they interact with many molecules (they are highly sociable), and they exhibit defined structural, enzymatic and regulatory activities. They are also very ancient, probably predecessors of DNA and perhaps protein. The RNA world hypothesis (104), first proposed by Woese (105), Crick (106) and Orgel (107), in which RNA was both a genetic and catalytic molecule, is now generally accepted. The discovery of catalytic RNA (ribozymes)(108,109) bolstered the hypothesis and modern RNA molecules are now considered relics (molecular fossils) of the RNA world and ancient components of a putative ribo-organism (110). One ribozyme that is universally found in all three superkingdoms of life and is absolutely essential for cellular function is ribosomal RNA (rRNA). Recent high-resolution crystal structures of the ribosome (111-114) and earlier biochemical studies (115) have convincingly shown

that the ribosome is a ribozyme (108,116). The rRNA performs the important functions of mRNA decoding and peptide bond synthesis. Ribosomal proteins mainly assist the folding and maintenance of the complex rRNA structure.

In recent years, the role of RNA in biology has expanded with the discovery of new classes of RNAs that harbor diverse structures and functions (117-119). Besides the classical three groups of molecules, tRNA, rRNA and mRNA, a repertoire of other RNA has been described. These molecules, collectively termed non-coding RNAs (ncRNAs), range in size from ~21-25 nucleotides for regulatory RNAs to ~10^3-10^4 nucleotides for molecules involved in the maintenance of chromatin structure. ncRNAs play important roles in a number of cellular processes, such as transcription, replication, RNA processing and modification, mRNA translation, and protein stability and translocation. Gene expression is modulated by miRNA and small interfering RNA (siRNA) (119), the translational tagging of proteins is mediated by tmRNA and post-transcriptional gene silencing by siRNA, and the targeted degradation of mRNA induced by interference RNA (RNAi) (120,121). RNA is also central for RNA processing, modification, and stability. For example, the catalytic core of the universally conserved RNase P enzyme (~300-500 nt) cleaves leader sequences from tRNA precursors (122). Similarly, the small nucleolar RNA (snoRNA) (~70-250 nt) is required for cleavage and processing of rRNA precursors (123). The ncRNA molecules are also involved in protein translocation across membranes [e.g., signal recognition particle (SRP) that targets nascent secretory and membrane proteins (124)]. Many other ncRNA molecules have been discovered that play structural roles, mimic the structure of other nucleic acids, or have very specific catalytic activities (118).

Global views of the universe of RNA are still missing. This is in part due to computational difficulties related to the study and discovery of RNA (117). However, novel systematic gene-discovery approaches have been utilized to uncover more effectively the RNA-encoding component of genomes [e.g., (125)] and first steps have been taken to produce a conceptual framework for an RNA ontology (126). While classification efforts have begun [e.g., 127] and RNA and RNA motifs can be catalogued using for example graph theoretical approaches (128-130), there are no phylogenetic-based RNA taxonomies and the evolutionary study of RNA structure is still incipient. In order to uncover evolutionary patterns and processes that are imprinted in the 3D structure of RNAs, we have devised phylogenetic methods that can reconstruct molecular history from molecular structure.

4.2. Evolution of RNA structure

There is great diversity in the structure of RNA molecules. However, the 3D conformations that molecules adopt are mainly structured by a set of short A-form helices, typically involving ~10 base pairs, which define elements of secondary structure. These elements are then arranged in higher-order structures through tertiary contacts sometimes delimited by motifs that can be identified at the sequence level (131,132). One important feature of the folding landscape of RNA is that secondary structure defines the overall fold, secondary elements being generally unaffected by formation of higher-order structures. However, the folding landscape is rugged and frustrated, and often molecules may adopt more than one optimal conformation (85). In order to study evolution of RNA structure, we focused on these elements of secondary structure and the properties of the folding landscape, searching for evolutionary patterns and processes embedded in functional RNA molecules. We used a tool for phylogenetic inference that we developed some time ago to reconstruct phyogenetic histories of molecular diversification directly from structure (89-91). The value of this methodological approach is that it unifies phylogenetics with structural biology (133). We have recently extended the strategy by generating phylogenies of both molecules and substructures that can illuminate processes related to structural evolution (94). This allowed us not only to study phylogenetic relationships at the structural level but also to build timelines describing how substructural components of molecules were added to the evolving molecules.

The underlying rationale is briefly illustrated with an analysis of tRNA (Figure 3). A set of character attributes that describe the geometry ("shape" characters that measure for example the length of stems, loops or unpaired segments) and the branching, stability, and uniqueness (*plasticity*) ["statistical" characters that measure statistical mechanic properties of branching, stability and uniqueness (55,82,134,135)] of the molecules are first identified. These are then encoded in alphanumeric format and used to construct data matrices for phylogenetic analysis. Using a model of character change (*transformation series*), we then impose evolution's arrow by invoking the tendency towards structural order in molecules that we discussed above. This tendency is supported by statistical mechanic, thermodynamic, and phylogenetic arguments. Structural features are consequently treated as ordered multi-state cladistic characters, and the transformation from one character state to another 'polarized' by identifying the ancestral state in the transformation series. Finally, phylogenetic trees of molecules or substructures are reconstructed using optimal tree search methods that minimize evolutionary change (e.g., maximum parsimony).

Phylogenetic trees reconstructed using geometrical and statistical characters and derived from several functional RNA molecules were congruent and similarly rooted (78,93,94). Consequently, geometrical considerations in structure appeared intimately linked to molecular order, supporting both the evolutionary model and arguments of character polarization. We also found the general approach is robust and can be applied to molecules from species that are closely or distantly related. We reconstructed phylogenies from the structure of many RNA molecules, including tRNA, rRNA, spacer rRNA, SRP RNA, RNase P, small signal mRNA, microRNA, and short interspersed element (SINE) RNA; we did this at different taxonomical levels. For example, we reconstructed

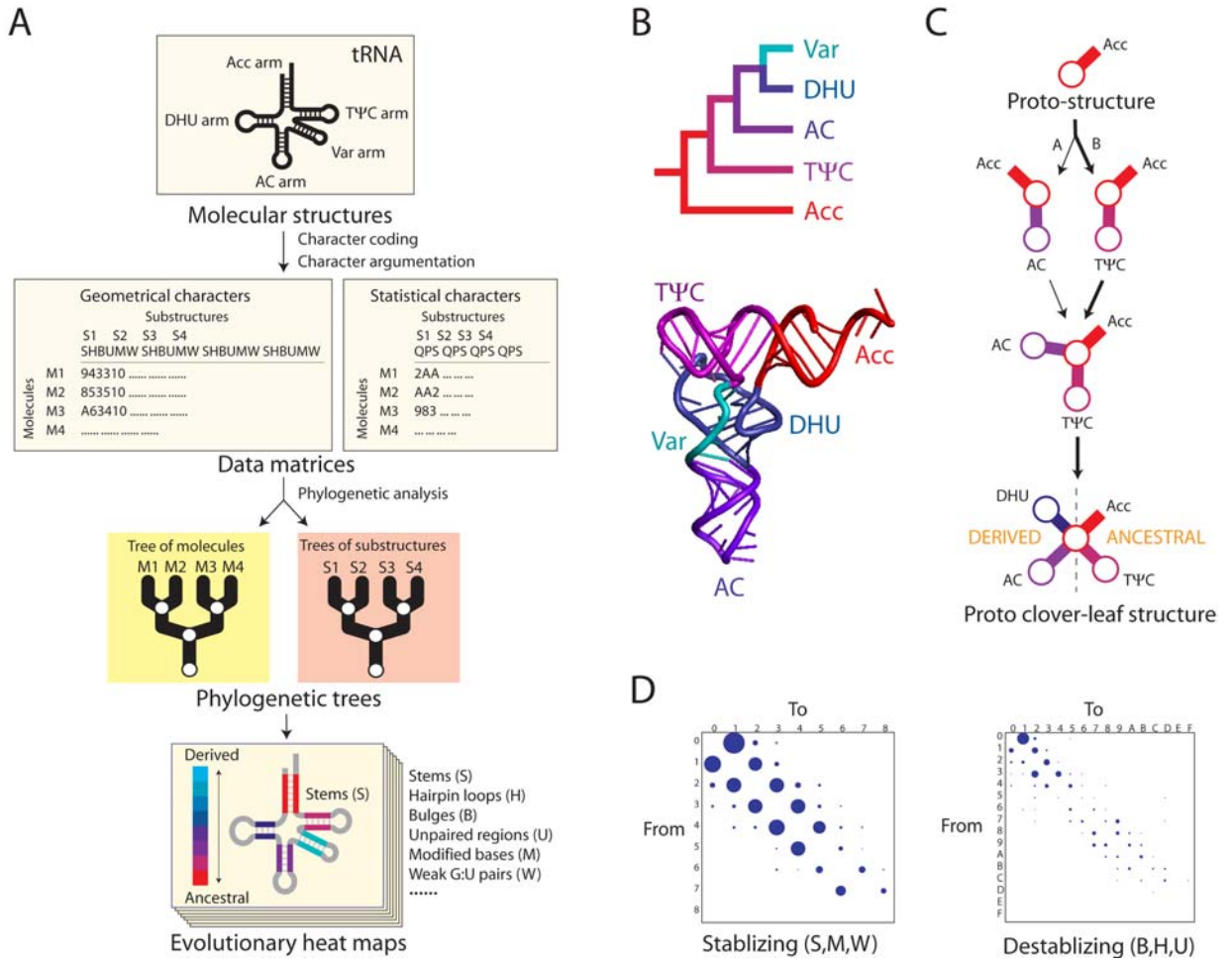

Figure 3. Construction of phylogenetic *trees of molecules* and *trees of substructures* of RNA. A, The structure of an RNA molecule, illustrated with tRNA, can be decomposed into segments (e.g., coaxial stem tracts and unpaired loop regions) or substructures (e.g., major structural and functional domains such as tRNA arms) that can be studied using features (*characters*) that describe their geometry [e.g. length of stems (S) and unpaired regions (H, B, and U)] or their stability and uniqueness (e.g. morphospace parameters *Q*, *P* and *S*) (78,90,91). These 'shape' and 'statistical' characters are coded and assigned "character states" according to an evolutionary model that polarizes character transformation towards an increase in molecular order (*character argumentation*). Coded characters are arranged in data matrices and subjected to cladistic analysis, generating phylogenies of molecules and substructures. Rooted trees can be used to color 2D or 3D structural models of RNA (*evolutionary heat maps*) that help infer models of structural evolution by providing 'timelines' of structural diversification. B, A tree of tRNA stem substructures obtained from tRNA from Bacteria and Eukarya revealed patterns of structural evolution that were used to build an evolutionary 3D heat map. Other trees of substructures provide interesting evolutionary patterns (not shown). C, A model of the early evolution of proto-tRNA molecules based on data derived from B. The model shows formation of substructures homologous to present-day acceptor (Acc), pseudouridine (TpsiC), anticodon (AC) and dihydrouridine (DHU) arms. Substructures may have had different functions than those of extant tRNA molecules. Data suggest the existence of two alternative evolutionary routes, with route A linked to ancestors of Archaea and route B linked to ancestors of Eukarya and Bacteria. D, Matrices of character transformation costs depict the frustrated energetics of base pairing in tRNA. The bubble charts describe the average frequency of changes between states in stabilizing (stem-related characters) and destabilizing segments (unpaired characters) of tRNA molecules. Charts summarize matrices for individual characters (not shown).

phylogenies describing the evolution of internal transcribed sequences (ITS) of rRNA in strains of phytopathogenic fungi that matched diversification patterns that followed continental pathogen introduction (89) or habitat adaptation (90), laboratory lines of *Chlamydomonas* that were congruent with pedigree histories (Caetano-Anollés, unpublished), and phylogenetic relationships of wild perennial relatives of soybeans that matched traditional classification (90). We also conducted a study of deep phylogenetic relationships in the grasses (Poaceae) based on the structure of mRNA molecules encoding small signal peptides (*enod40*) and non-protein coding RNA molecules such as SRP RNA and rRNA and on chromosomal rearrangements (78). In particular, *enod40* is a gene

involved in early regulation of nodulation in legumes that has homologues with putatively functional RNA structures that are highly conserved in plants (136). The study established rooting for the grasses, an order for the diversification of major grass lineages, and a basis for the study of evolution of genome size in the grasses (137). In all cases we found that trees from sequence and structure were congruent, revealing the direct evolutionary links that exist between genotype (sequence) and phenotype (structure) in RNA. We even extended our studies to molecules from all superkingdoms of life and reconstructed universal trees from the structure of several molecules, including the small and large subunits of rRNA (90,91). These ribosomal trees showed it was equally parsimonious to consider ancestral eukaryotes or prokaryotes as the organisms that gave rise to modern life.

We also focused on tRNA, a molecule that bridges fundamental components of the translation machinery (94,138,139). tRNA is an adaptor with an acceptor arm (Acc) that charges amino acids through the specific activity of aminoacyl-tRNA synthetases and an anticodon (AC) arm with triplets of bases that recognize complementary codon sequences in mRNA. This molecule historically considered as a static "adaptor" is gradually being recognized as an active and dynamic player in the process of protein synthesis (140), as more specific roles for tRNA are revealed (141). To investigate the controversial origin and evolution of tRNA we analyzed the entire set of 571 tRNA molecules deposited as RNA sequences in the Bayreuth database and generated global trees of molecules. Structural phylogenies placed tRNA molecules that coded for a group of four amino acids ($tRNA^{Sec}$, $tRNA^{Ser}$, $tRNA^{Leu}$, and $tRNA^{Tyr}$) at the base of the tree of tRNA structure (138). The basal placement suggests these four amino acids (Sec, Ser, Leu and Tyr) were probably the first charged or coupled by tRNA in processes related to translation and/or RNA world-based replication that occurred before organismal diversification. All basal tRNAs harbored a long variable arm (Var) which is known to carry important identity elements for the recognition of these molecules by their cognate aminoacyl-tRNA synthetases (142-146). The long variable arm therefore appears to be an ancestral structure in tRNAs that harbor modern functions. This observation suggests that functional diversification in tRNA developed once the cloverleaf structure was fully formed. This is consistent with the recent proposal that structural diversification preceded the establishment of amino acid and anticodon specificities and the diversified organismal world (147). Interestingly, some few $tRNA^{Ser}$, $tRNA^{Leu}$, and $tRNA^{Tyr}$ lacked the long variable arm and were derived, suggesting the existence of take-overs in evolution that involved the loss of these substructures. This tendency of structural simplification is remarkable but is not unique. As we will describe below, we have observed similar tendencies in rRNA and other molecules.

Because tRNA phylogenies did not reveal clearly the tripartite nature of life or clear patterns linked to anticodon or amino acid-charging functions, we used phylogenetic constraint to untangle confounding histories of recruitment in these molecules (139). We forced tRNA molecules into monophyletic groups in the trees to falsify competing hypotheses and generated timelines of amino acid charging specificities, codon discovery, and organismal diversification. Analyses confirmed the ancestral nature of Sec, Ser, Leu and Tyr charging, revealed the early role of the second and then the first codon base, and identified codons for Ala and Pro as the most ancient. Timelines of codon discovery showed patterns suggestive of an early code that was degenerate and later expanded by exploiting additional anticodon identity elements. This tendency towards enhancement of specificity in the genetic code is remarkable and paraphrases similar patterns in proteins related to biological function. Most importantly, our study showed a lack of correlation between timelines of amino acid charging and codon discovery, suggesting independent histories of recruitment of these two tRNA functions. This is consistent with evolutionary profiles related to aminoacyl-tRNA synthetases and the emerging phylogenetic picture that suggests these enzymes played a minimal role in the evolution of the genetic code (148). We believe these histories of recruitment were driven by co-options and important take-overs during early diversification of the protein world.

4.3. Evolution of components of RNA structure

We also generated phylogenetic trees of molecular substructures that describe histories of structural evolution (94). The method is summarized in Figure 3A. This method is novel in that it reconstructs trees that do not describe the evolution of organismal taxa or molecules (e.g. protein sequence or structure). Instead, trees describe the evolution of component parts of RNA molecules. Because trees are rooted using the central assumption of a tendency towards order, phylogenies embed relationships of ancestry of molecular substructures and describe histories of finite repertoires of molecular component parts. These histories define evolutionary timelines that for visualization can be superimposed directly onto 2D or 3D representations of simple or complex RNA molecules (*evolutionary heat maps*)(Figures 3 and 4). The method was used to study tRNA (94), tRNA-derived SINE retroelements (93), and rRNA (Harish and Caetano-Anollés, ms. in preparation).

The occasional and idiosyncratic recognition of the AC arm by variable protein domains in aminoacyl-tRNA synthetases has been given as a tell-tale sign of its derived nature (149). In contrast, the Acc arm and the top half of the molecule harbor roles in almost all macromolecular interactions that involve tRNA (149-151). The Acc arms is consistently recognized by aminoacyl-tRNA synthetases (which drive the aminoacylation of the molecule), RNase P, elongation factor Tu, and rRNA. It is therefore noteworthy that phylogenies describing the evolution of tRNA substructures showed the molecule had an origin in the Acc arm and the top half domain of the molecule (Figure 3B). This is remarkable and lends strong support to the 'two halves" hypothesis put forth by Maizels and Weiner (150) that postulates the AC/DHU domain was incorporated later in evolution. This hypothesis constitutes the cornerstone of the "genomic tag" hypothesis that considers tRNA as ancient telomeres of genomes in the

RNA world (152,153). Our trees also support a more detailed structural transformation sequence, in which the tRNA molecule evolves from a mini-hairpin by gradual addition of nucleotide pairs to its growing double-helical stems. This ultimately results in a molecular arrangement that favors multiloop conformations and molecular multifurcation, expected outcomes when seeking to maximize molecular order. Other interesting patterns were also identified. For example, modified bases were incorporated earlier than weak G:U base pairs. Interestingly, diversification of unpaired regions somehow followed the addition of stems in the evolving molecule, with the 5'-terminal free end being the most ancestral and the 3'-terminal free end (including the CCA terminus) the most derived. The suggestion that the 3'-terminal sequence was added only after the entire cloverleaf structure was formed matches inferences derived from statistical analyses of tRNA sequences (154). Finally, analyses of partitioned data matrices revealed fascinating patterns. Trees of substructures obtained from archaeal tRNAs showed that the AC arm predated the TpsiC arm. In contrast, trees of substructures from Bacteria, Eukarya and viruses showed the TpsiC arm being more ancient. These two evolutionary routes are compatible with whole-genome analysis of protein complements and domain combinations that suggest an early split of the archaeal lineage from a protein-rich communal world by reductive genomic tendencies in Archaea (22,26). We will discuss these observations in more detail section 7.2.

The evolutionary analysis of the structure of tRNA-derived SINEs in plants and eukaryotes was also revealing (93). SINEs are a class of dispersed mobile sequences that use RNA as an intermediate in a genomic dispersion process called retroposition (155). The exercise established a model of structural evolution of these transposable elements that explained the popularity of SINE sequence families in the plant genome. Furthermore, trees of substructures showed SINEs had an origin in tRNA-derived stem-loop structure and suggested proto-SINEs were rich in maximal hydrogen-bonding GC base pairs, two aspects that are also fundamental to support the idea that modern viruses are "molecular fossils" of ancient strategies of replication (153). Again, our observations provide support to the genomic tag hypothesis. They also show common evolutionary trends important for tRNA-derived SINE RNA structures.

4.4. Origin, evolution, and simplification tendencies in rRNA

One of the most powerful features of structure-based phylogenetic tree reconstruction is the ability to trace the evolution of functionally important components in nucleic acid molecules. Recent progress in the structural determination of the ribosomal ensemble (111-114) provided a scaffold for evolutionary tracings. A preliminary exercise showed patterns of evolution in inter-subunit bridge contacts and tRNA-binding sites that were remarkable (91). These patterns were consistent with the proposed coupling of tRNA translocation and subunit movement (111). Results also supported the concerted evolution of tRNA-binding sites in the two rRNA subunits and the ancestral nature of both the peptidyl site and the functional relay of the penultimate stem helix of the small subunit. We have extended these results by generating trees of substructures and evolutionary heat maps of the entire ribosomal ensemble (Harish and Caetano-Anollés, ms. in preparation). Figure 4 shows the evolution of stabilizing stem components of the entire ribosome. Analysis of the small subunit confirmed that the penultimate stem helix (S49 or helix 44 in the prokaryotic model), the dominant SSU rRNA component of the subunit interface (111) and the proposed ribosomal functional relay (156,157), was the most ancestral substructure of the ensemble. Analysis of the large subunit suggested a late and more complicated structural origin. Interestingly, ancient substructures were located in the middle of the subunit ensemble and were clearly linked to ratchet mechanics. Our observations are compatible with an origin of the ribosome in structures that were not linked with modern protein synthesis. While it is intuitive that an ensemble as complex as the ribosome did not evolve at once, our trees of substructures and timelines support the progressive evolution of this complex biosynthetic machinery from a much simpler proto-ribosomal structure. However, the molecule did not grow (in an evolutionary sense) in an ordered fashion, starting from the core and adding more and more derived layers of substructures to the molecule. Instead, ancient and derived substructures sometimes occupy the same regions of the molecular space. A complicated history of recruitment of substructures seems to dominate evolution of these complex molecules.

We also traced the complete repertoire of ribosomal structural characters, lineage-by-lineage, in the universal phylogenetic tree of rRNA molecules (91). This offered the opportunity to study how evolutionary change was distributed and constrained in rRNA and allowed to reconstruct hypothetical ancestral molecules. The exercise revealed a tendency towards molecular simplification, especially in highly variable regions of the molecules. This tendency was maximal in rRNA from *Encephalitozoon cuniculi*, an amitochondriate microsporidian endoparasite with a highly reduced genome and protein complement. The exercise also showed reduction of ribosomal structural change with time that occurred concomitantly in both ribosomal subunits, which is compatible with plastogenetic congruence and structural canalization (48,62). It also allowed the inference of a probabilistic model of character evolution in the form of a step matrix of transformation costs from one character state to another (Figure 3D). These matrices were described in bubble diagrams and as expected depicted the frustrated energetics of base pairing in RNA structure (85). The minimum free energy of a secondary structure can be considered the sum of its loop energies, which have been measured, tabulated, and used in folding algorithms (158). Energetically unfavorable loops destabilize the contribution of energetically stable helical stem regions, delimiting a frustrated landscape. For rRNA, step matrices showed tendencies to form molecules with pair segments of an optimal length of ~10 base pairs (91). For tRNA, the optimal length of stems was lower, ~5 base pairs (Figure 3D). This suggests an important evolutionary tendency to optimize molecular size.

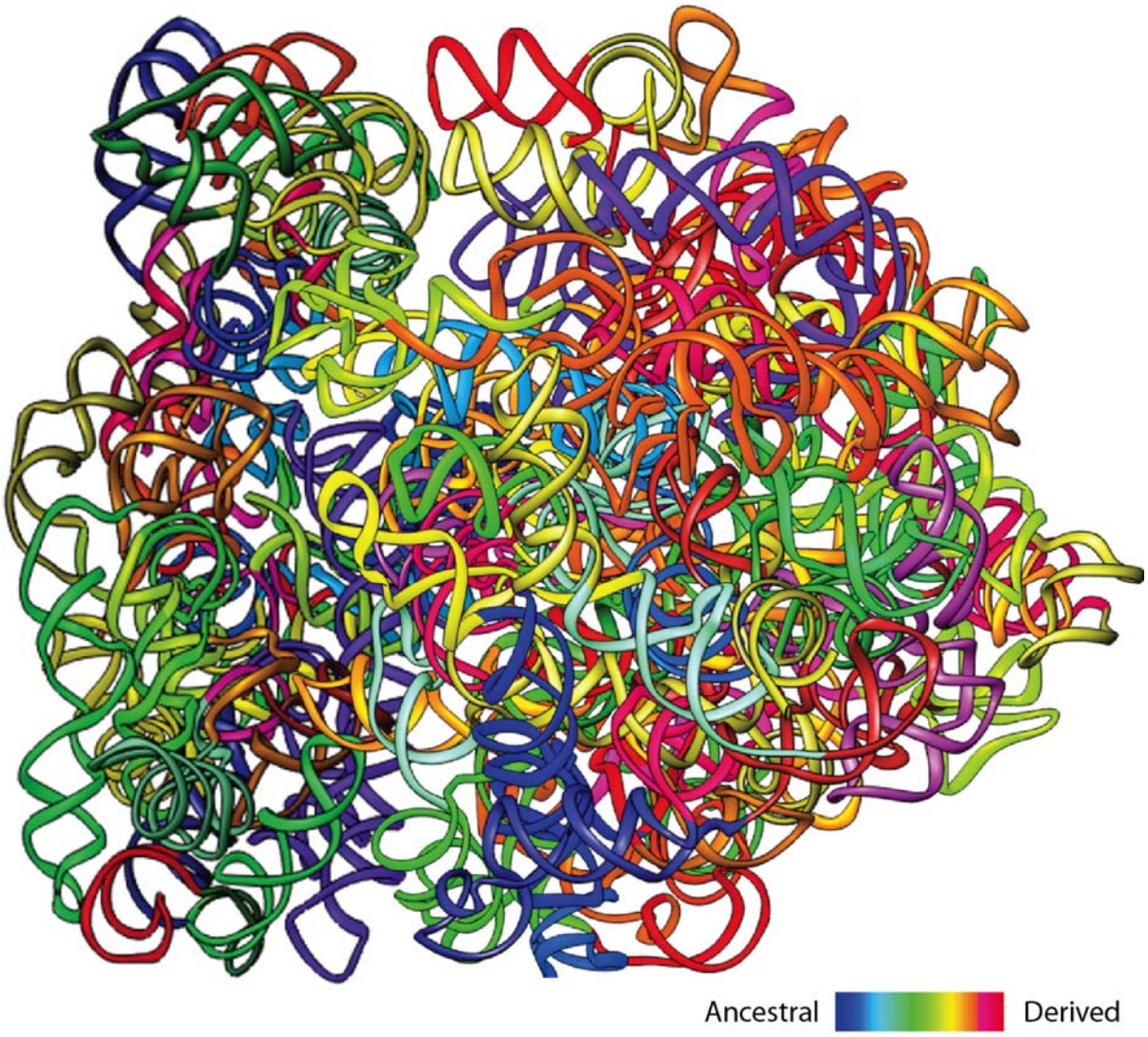

Figure 4. Evolutionary heat map describing the evolution of stabilizing stem components of the ribosome. Ancestries derived from trees of substructures were painted directly on a 3D structural model using a color purple-blue-cyan-green-yellow-red scale that describes relative ancestry values from ancestral to derived. The model was visualized in ribbons (239). The small subunit is located at the left hand side of the molecular ensemble and young components (painted with red and yellow) are all located in the periphery of the ensemble.

5. EVOLUTION OF MODERN PROTEINS

5.1. The hierarchical nature of the protein world

The majority of cellular functions involve protein molecules. Proteins are complex and extraordinarily diverse (159). There are about ~10^{13} protein sequences in the genomes of the estimated ~10^{7}-10^{8} species in the world, representing only a minute fraction of sequences from the ~10^{300}-10^{500} variants that are possible. The limited evolutionary exploration of this enormous permutational space has nevertheless produced considerable diversity at the structural level (160), generating for example great variation in enzymatic catalysis (161,162). The functions of proteins are embedded in structural, functional and evolutionary units called domains (163). Domains are compact folding arrangements of the polypeptide chain in 3D space that appear singly or in combination with other domains in a protein molecule and are evolutionarily conserved. Domains usually harbor pockets called "active sites" capable of hosting interacting molecules such as ligands and cofactors. Consequently, domains delimit the functional toolkit of proteins in a cell. Research in biology has been driven for years by the paradigm "one gene-one protein chain", but this has gradually changed with the discovery of alternative splicing, functional RNA molecules encoded in non-coding DNA segments, and developments in structural biology and genomics. The concept of "one protein-one function" has also been challenged by the existence of proteins that exhibit a multiplicity of functions. Proteins sometimes display conformational diversity independent of binding to ligands, use structures to "moonlight" different functions without

involving their active sites, or become promiscuous by using the same active site for different functions (164). The ubiquity of multidomain proteins now adds another complexity to the puzzle. It appears that genomes are highly fluid and evolution of genes believed to proceed mostly by gene duplication now seems tailored by rearrangement of the domain units (100,165). Domain combination is therefore pervasive and shows domains as sub-genic units of function and evolution. In this regard, domains can be viewed as the final units defining the identity of species.

The minimum protein configuration that is biologically active (*biological unit*) generally consists of one or more polypeptide chains, and these harbor one or more protein domains. Proteins are therefore modular, and modularity expresses at different hierarchical levels. In multidomain proteins, domains can be either repeated or combined in defined order (100,166-168). However, the topology of domain combinations is highly conserved. The orientation of domains and the type of neighboring domains in proteins is limited (26,27,166,169). For example, domain permutation may not materialize in all protein variants. These constraints define a molecular "interactome", i.e. a collection of possible topologies depicting intramolecular interactions between protein domains. The folding of individual domains into compact units is relatively independent from the folding of other domains in a protein, especially when the interfaces between domains are loosely packed (170). Interestingly, the biological unit that embeds function sometimes is not the domain in multidomain proteins, but supra-domains, two- or three-domain combinations that recur in different protein contexts (171). At a higher hierarchical level, proteins can interact in 3D by forming complexes through quaternary interactions. The evolution of the protein subunits in these complexes appears driven by duplication of homomeric interactions, for example present in homodimers through gene duplication and formation of paralogues (101).

Understanding function constitutes one of the greatest challenges in biological research. For example, a large number of functions of a cell depend on the *interactions* between proteins and small molecules, such as globular enzymatic catalysts and metabolites in cellular metabolism. However, interactions at higher hierarchical levels of protein organization also play important roles. These interactions are not random but are controlled by thermodynamic and kinetic laws that dictate for example the likelihood and rate of a chemical reaction. In turn, the cell and its compartments and regulatory processes provide the right environment for these reactions to operate in time and space. These interactions are the result of millions of years of evolutionary fine-tuning of the molecular structure of protein molecules. Consequently, finding links between structure and function represents a complex problem that requires incorporation of structural information of proteins in proteomic repertoires, regulatory differences that shape traits and physiology, and approaches in bioinformatics, biochemistry, molecular evolution and genomics. We have addressed some of these questions with a synthetic and systematic approach that merges phylogenetic analysis with structural biology and genomics. Below we summarize patterns and processes related to the evolution of the protein repertoire.

5.2. Evolution of domain structure and organization

Ever since the pioneering work of Jane Richardson (172), proteins of known structure have been grouped using taxonomies that attempt to provide a comprehensive description at structural and evolutionary levels. The Structural Classification of Proteins (SCOP) (173,174) and the CATH protein structure classification (175) are two examples that use expert and automated systems to classify proteins domains into hierarchies. In SCOP, proteins that are evolutionarily closely related at the sequence level are clustered together into protein fold families (FF). Proteins belonging to different families that exhibit low sequence identities but share structural and functional features suggesting a common evolutionary origin are further unified into fold superfamilies (FSF). Finally, FSFs sharing secondary structures that are similarly arranged and topologically connected are unified into protein folds. These folds sometimes have peripheral regions of secondary structure that differ in size and conformation and 'decorate' the central fold architecture distinctively. Consequently, fold categories should be regarded as "neighborhoods" defined by how much structural overlap exists between them (176). Some regions of the protein fold space represent a continuum for certain architectural arrangements (sometimes linked by supersecondary motifs) while in other regions clearly distinct non-overlapping topologies are observed.

While our knowledge of sequence space is far from complete (177), it is apparent that protein diversity originated from a limited set of architectural designs (178). Most proteins have been formed by gene duplication, recombination, and divergence. Proteome evolution can be tracked by matching proteins of known folding structure to genome sequences using Hidden Markov Models (HMMs) of structural recognition (179). However, it has become increasingly more difficult to find new folds in nature (3). Currently, a set of ~1,000 folds, ~1,800 FSFs, and ~3,500 FF describe the world of proteins in SCOP release 1.74 (November 2007). Clearly, the repertoire of protein architectures is finite making its study tractable. Unfortunately, there is no clear understanding of the evolutionary principles that drive the topology of protein structure at fold or FSF levels. A number of approaches have been used to characterize protein space and provide global views of the protein world directly from structure. This includes the generation of fold family trees (180,181), taxonomies based on secondary structure (182), metric distance comparison of structures (183), graph representations of domains based on scores of structural similarity (184,185), and a periodic table of structures (186), However, problems associated with the systematic classification of architectures at a topological level, make it difficult, if not impossible, to find a general metric of pairwise comparison that could be used for global analysis (187). Moreover, to be useful, strategies require of methods capable of organizing the comparative data within an evolutionary perspective. However, there is no reliable

procedure at present that can generate phylogenetic relationships at higher hierarchical levels of protein classification directly from the structure of proteins. Other approaches are needed.

We recently embarked on a systematic and global study of evolution of domain structure and organization in genomes that have been completely sequenced (18,21,22,26,27,188). In contrast with the approach we used to study RNA, our strategy did not require the direct use of elements of structure to reconstruct rooted phylogenetic trees. Instead, domain structures were assigned to protein sequences and the structural census was then used to generate phylogenomic *trees of proteomes* and *trees of architectures*, at fold and FSF levels of hierarchical classification. The census involved identifying fold and FSF architectures corresponding to individual domains but also arrangements of domains in biological units. Consequently, we were able to study both domain structure and domain organization in proteomic repertoires. Structural assignments corresponded to well-defined crystallographic 3D models that had been catalogued by SCOP, and were therefore restricted to proteins for which a known structure could be inferred (on average ~60% of the proteome). Since folds and FSFs are highly conserved, every instance of discovery or adoption of an architecture by a proteome represents a rare event in the history of the organismal lineage, and globally a rare event in the history of the protein world. The vast majority of these architectures should be viewed as highly successful architectural designs. They constitute historical imprints [modern *molecular fossils* (150)] preserved in nature by their successful retention and propagation in proteomic complements. Indeed, the occurrence, abundance, and combination of architectures in hundreds of proteomes have been used to build reasonable universal phylogenetic trees that describe the evolutionary history of major organismal lineages (18,20-22,26,189). All the features of "genomic demography" that we studied carried ancient phylogenetic signatures and could be used to uncover deep evolutionary phenomena related to the origins of the protein world and modern life. Examples of rooted trees of proteomes reconstructed from domain structure and pair-wise combination of domains in proteins are illustrated in Figures 1 and 2. These trees use entire repertoires of proteins or domain combinations (intra-molecular interactomes) to describe the evolution of organisms, revealing the tripartite nature of life and rejecting other evolutionary scenarios [e.g., (190)] often based on a limited set of features (molecular, biochemical, cellular, etc) that could be affected by horizontal gene transfer or recruitment processes. It is noteworthy that domain architectures have been traced along universal trees showing convergent evolutionary processes to be rare in protein structural evolution (191). This suggests that protein structure at these high levels of organization diversified mostly by vertical descent, empowering our phylogenetic reconstruction exercise.

Universal trees of architectures revealed remarkable patterns of molecular evolution (21,26,27,188). Figure 5A shows as an example a phylogenomic tree of FSF reconstructed from a genomic survey in 185 organisms (22). Architectures that were widely distributed in nature were found at the base of the tree and were only missing in parasitic organisms with highly reduced genomes (e.g., *Mycoplasma*, *Nanoarchaeum, Encephalitozoon*), known to have discarded enzymatic and cellular machinery in exchange for resources from their hosts. The first nine folds to emerge in evolution were common to every genome analyzed and included folds widespread in metabolism. It is noteworthy that only 16 folds were universally shared and all of them originated deep in the tree. Similarly, all classes of globular protein architecture appeared very early in evolution and in defined order, the alpha/beta class being the first, followed by the alpha+beta, the all-alpha, and the all-beta classes, and by small and multi-domain proteins. Patterns of origin and accumulation in the tree of folds suggest that architectural designs with interspersed alpha-helical and beta-sheet elements were segregated in the course of evolution, first within their structure (alpha+beta class) and then confined to separate molecules (all-alpha and all-beta classes) (18,21). This is consistent with the random origin hypothesis of proteins (192). A similar conclusion was recently reached when tracing fold occurrence along branches of proteome trees (193). Remarkably, the most ancestral folds harbored interleaved beta-sheets and alpha-helices and barrel structures, many important structural designs were derived in the tree (including polyhedral folds in the all-alpha class and beta-sandwiches, beta-propellers and beta-prisms in the all-beta class), and protein transformation pathways describing likely scenarios of structural evolution (194,195) and other patterns could be traced in the trees (18).

5.3. Timelines of architectural discovery

The trees of architectures that we reconstructed were intrinsically rooted. The trees generally described evolution of domains (e.g., Figure 5A), but we also reconstructed trees of pair-wise domain combinations, domain combinations, and domains and domain combinations. Evolution's arrow was established directly by the evolutionary model (discussed in section 3.3). As expected, the trees were also highly unbalanced suggesting architectural discovery involving semipunctuated evolutionary processes, similar to those recently suggested for substitutional change in nucleic acids (196). The rooted trees established by definition evolutionary timelines of architectural discovery, with time measured by a relative distance in nodes from a hypothetical ancestor at the base of the trees (*nd*). These timelines uncovered remarkable patterns. As mentioned above, architectures at the base of the tree were common and defined an "architectural diversification" epoch in protein evolution in which members of an ancestral community of organisms diversified their protein repertoires through differential loss (light green shaded area in Figure 5A). During this period, most architectures were shared by most organisms, with architectural loss occurring preferentially in organisms belonging to the archaeal lineages. Later in time, superkingdom-specific and lineage-specific architectures appeared in evolution as the world of organisms diversified, mostly resulting from mechanisms of discovery and loss, lineage diversification, and vertical and horizontal

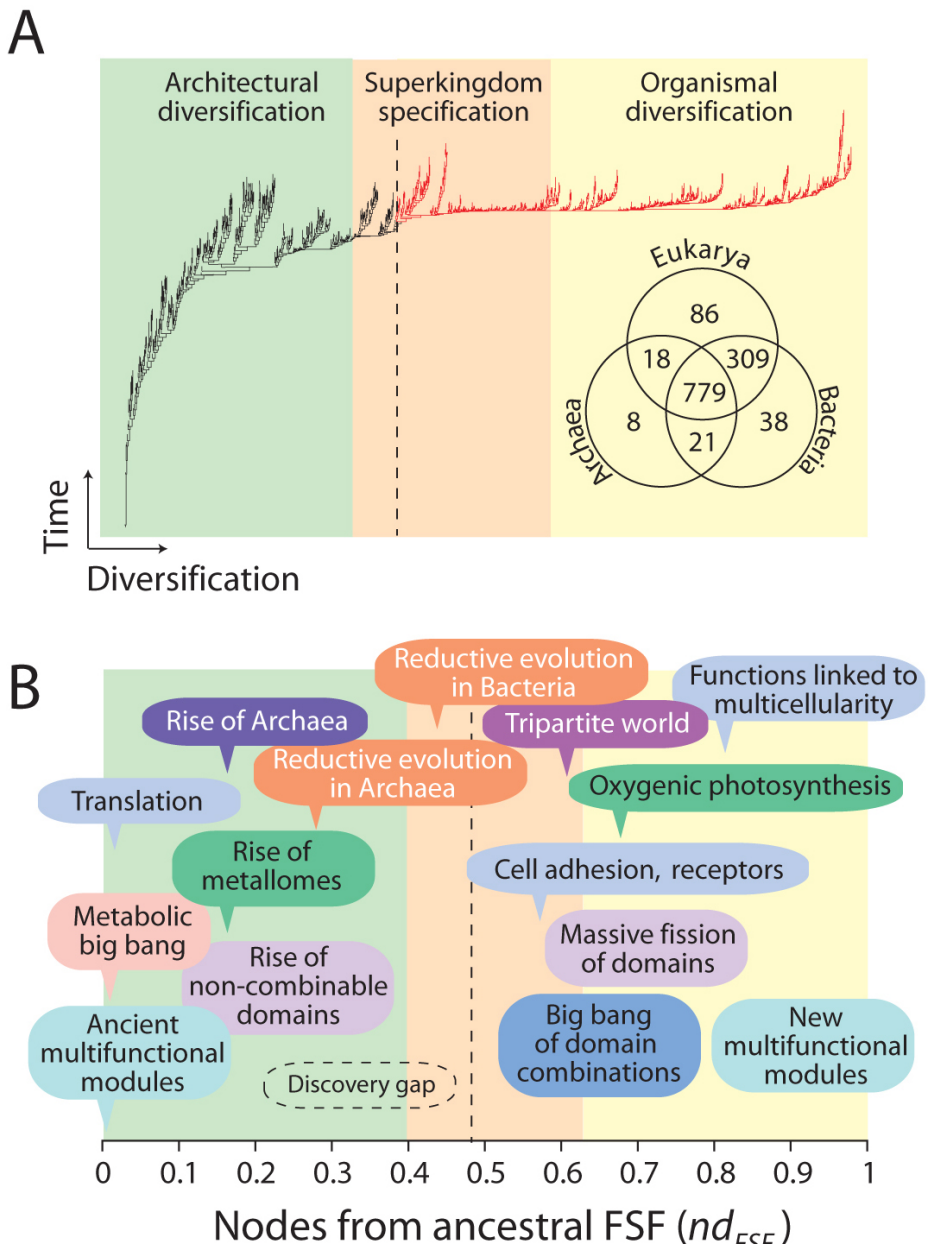

Figure 5. Evolution of the protein world. A, A phylogenomic *tree of domain architectures* was reconstructed from a census of domains at fold superfamily (FSF) level in 185 organisms that have been fully sequenced. The rooted optimal tree (118,119 steps; CI=0.031, RI=0.759) is well supported by measures of skewness in tree distribution (g-fit=–0.099; $P<0.01$). Terminal leaves are not labeled with FSF names as they would not be legible. Branches in the tree labeled in red occur after the appearance of the first architecture unique to a superkingdom (Bacteria, indicated by a dashed lined). The Venn diagram shows occurrence of FSFs in the three superkingdoms of life. B, Evolutionary timeline of architectural discovery in which the age of FSF architectures (nd_{FSF}, number of nodes from the ancestral FSF at the root/total number of nodes in the tree) is used to describe the relative timing of a number of important events in the history of life. Information in boxes without pointers were derived from trees of domain and domain combinations (27), and their relative location is approximate. The three evolutionary epochs of the protein world are shaded in light green (architectural diversification), salmon (superkingdom specification), and light yellow (organismal diversification) according to Wang *et al.* (22).

transfer (see section 7.2 below). Tracing biological function along the phylogenies revealed fundamental patterns, some of which are described in the timeline (Figure 5B). Trees of domain architectures showed that architectures at the base of the tree were multi-functional, and that the nine most ancient were responsible for most of the enzymatic functions present in modern metabolism (18,197). Observations provided further support for the proposal that

during metabolic evolution enzymatic multifunctionality was replaced by specialized function (198), and defined what we call a "metabolic big bang", an expansive tendency of recruitment of architectures to perform different functions, evident at the start of the protein world. The most ancient folds shared a common architecture of sheets and helices that formed either barrels or were interleaved and highly symmetrical. Proteins within these groups were generally large and architecturally complex. They interacted with organic cofactors, especially nucleotide-containing ligands such as ATP, ADP, GDP, NAD and FAD, all of which appeared to have originated early in evolution according to a power-law distribution of ligand-protein mapping (199). As time progressed, two important tendencies emerged in the protein world: (i) *increased specialization*, in which architectures harbored only one or very few functions, and (ii) *molecular simplification*, with structures that were increasingly smaller and more compact (e.g. increases in the tilt of strands or the frequency of open barrel structures in the popular b-barrels). At the same time, structures became more refined, as illustrated with barrel structures harboring increasingly more complex strand topologies).

Trees of domains and domain combinations established an evolutionary mechanics for the protein world by mapping processes of fusion and fission of domains and tracing biological functions along the timeline (27). They revealed an explosive expansion (big bang) of domain combinations that occurred relatively late, at the onset of organismal diversification (Figure 5B). The trees showed that the first architectures to appear were multi-functional proteins with single domains, all of which produced fusion-driven combinations. These domain combinations arose early during the architectural diversification, were functionally specialized, and later dominated the protein world. In contrast, fission processes occurred late, were notable during the big bang of domain combinations and produced many derived multi-functional single-domain proteins in Eukarya. This 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.

Tracing functions along the timeline revealed remarkable patterns, including the very early (though protracted) discovery of proteins involved in translation (e.g., aminoacyl-tRNA synthetases, elongation factors, and ribosomal proteins), interrupted by a "discovery gap" that perhaps involved a revision of the translation apparatus, the relatively early rise of metallomes (the Zn-metallome appearing first) (C.L. Dupont, ms. in preparation), and a late rise of oxygenic photosynthesis, which was preceded and followed by the discovery of functions typical of Eukarya (cell adhesion, receptors, and chromatin structure, and functions linked to multicellularity). Timelines also suggest proteins were first associated with organic cofactors but later involved transition metals as ligands, perhaps mediated by the increasing energy demands of the ancient world. Some of these results are consistent with a recent proteomic analysis that suggest shifts in trace metal geochemistry related to the redox state of ancient oceans are imprinted in protein architecture and suggests prokaryotes evolved in anoxic marine environments while eukaryotes did so in oxic counterparts (200). Many of the basal structures in the trees were also involved in functions associated with ancient evolutionarily conserved genes that were identified by physical clustering in bacterial genomes (201).

6. EVOLUTION OF MODERN METABOLIC NETWORKS

Biological networks offer insight into the organization and evolution of life and have been the subject of intense study (202). One network that is particularly important is cellular metabolism. The existence of a core set of metabolic reactions common to life suggests that the global metabolic structure has been the subject of strong evolutionary constraint. Similarly, network connectivity properties suggest modular components typical of evolved systems (102,203) and emergence of hub metabolites involved in many reactions by enzyme specialization (204). At the same time, metabolic reconstruction exercises based on genomic information revealed that there is considerable diversity in pathways at the organismal level [e.g., (205)]. In many cases, the repertoire of networks can be considerably reduced, including reduction in central metabolic pathways, such as the citric acid cycle (206). Consequently, simplification tendencies are also revealed in metabolic networks. How metabolic networks grew as organisms increased in complexity remains an important question, making metabolism an interesting general model for the evolution of networks.

Metabolism is largely driven by the specificity of its enzymes. Consequently, the origin and evolution of metabolic networks can be best explored at protein rather than at metabolite levels. Metabolism is very ancient and parts of the metabolic network probably evolved prior to the origin of cellular life from reactions that could have proceeded without catalysis or with inorganic catalysts (207). This view is supported by *in vitro* experiments that try to simulate pre-biotic chemistry. An alternative view is that ribozymic catalysts preceded modern metabolic reactions. Under this scenario, the only palimpsest that is required relates to the pre-biotic creation of nucleotides (208,209). It is likely that polypeptides became metabolic catalysts through takeover processes (198). These involved ribozymes, pre-biotic reactions, or both. The earliest enzymes were probably weakly catalytic and multifunctional with broad specificities. Gradually, more numerous, effective, and specific enzymes evolved from the multifunctional enzymes through gene duplication, mutation and divergence (210). As enzymatic pathways became more complicated, new enzymatic functions and metabolic pathways could have been generated by recruitment of individual enzymes from the same or different pathways or by enzymatic recruitments *en masse* from entire pathways. In this regard, several possible scenarios for the evolution of enzymes in metabolic pathways have been proposed (211). In the retrograde evolution hypothesis, biosynthetic pathways evolve by recruitment of enzymes (from within or outside the

pathway) to host sites sequentially more remote from the end product of the pathway (212). By a symmetrical argument, catabolic pathways could have evolved sequentially from the metabolite being degraded (213). An alternative scenario is one in which new pathways evolve by recruitment from diverse donor sites throughout metabolism (214). This hypothesis assumes there is already an active enzymatic core with multifunctional and/or specialized enzymes from which new enzyme recruits are drawn for metabolic innovation. The result is a patchwork of homologous enzymes scattered over diverse pathways. Considerable evidence supports the patchwork recruitment scenario (211). For example, enzymes with alpha/beta barrel fold structure that catalyze similar reactions occur across metabolic pathways (215). These patterns of structural homology resulted in a mosaic when structural assignments and sequence comparisons were used to analyze the small-molecule metabolism in *Escherichia coli* (216,217).

We recently explored the origins and evolution of modern metabolism using phylogenomic information embedded in protein structure (197,218). We first painted the ancestries of enzymes derived from our trees of architectures directly onto over one hundred metabolic subnetworks in mesonetworks defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (219). In KEGG, subnetworks combine reactions and functions into categories by function, and mesonetworks pool together subnetworks with related functions. For example, the purine and pyrimidine metabolism are two subnetworks of the "nucleotide metabolism" mesonetwork. The evolutionary tracing exercise involved combining structural, functional and evolutionary information about enzymes in networks and the creation of the Molecular Ancestry Network (MANET) database (218). More specifically, MANET links structural information of enzymes (PDB entries), structural and evolutionary information embedded in the classification scheme of SCOP (SCOP IDs), integrated information about cellular metabolism in KEGG (diagrams and gene sequences), and ancestries derived from phylogenomic trees (*nd* values) (Figure 6). Metabolic enzymes without PDB entries were then linked to fold architectures using SUPERFAMILY HMMs (179) in almost a million genomic sequences. Finally, this information was used further to build trees of metabolic subnetworks, based on abundance of architectures in enzymes participating in each subnetwork. As a result, we treated protein domains in enzymes as structural "parts" and used them to derive evolutionary histories of enzymes. The exercise allowed an evolutionary exploration of how these modules were incorporated in metabolic networks. A preliminary analysis of evolutionarily painted subnetworks revealed patchy distribution patterns indicative of widespread enzymatic recruitment. These patterns were consistent with previous evidence (211,216,217). Interestingly, the distribution of abundance of folds with various ancestries shows that mesonetworks differed in mean ancestry, with amino acids oldest and lipids and glycans youngest.

Recruitment is not only limited by the history of enzymes but also depends on enzymatic characteristics associated with function and network location. Function relates to the catalytic specificity of an enzyme, given by its Enzyme Classification (EC) number, and both the catalytic activity and the substrate of an enzyme may change during evolution (211,215). Enzymes also differ in metabolite usage, the diversity of reactions using their substrates. Some metabolites (e.g. water, ATP, NADH) are used in many reactions. These metabolites affect significantly the topology of the metabolic network (220) and seem to facilitate recruitment of enzymes for new functions (209). Proximity of donor and host sites influence the probability of recruitment, with diversification to new host sites occurring mainly from nearby enzymes and varying with metabolite usage and enzyme class (221). In particular, sequence comparisons revealed homologous enzyme pairs occurring close to each other more often than expected by chance (222). Taking into consideration these limiting factors, we sorted out recruitment processes and established patterns of origin and evolution in modern metabolism, using a graph-based constructs we call "metabolic wheels" and MANET to mine the data on distribution of enzymes with certain folds across metabolic subnetworks (197). These wheels provided a compact and vivid summary of extensive information about relations between subnetworks in terms of enzyme sharing. For every fold, metabolic wheels considered the relative age of subnetworks inferred from trees of subnetworks, fold abundance in each subnetwork, and enzymatic activities shared at different levels of EC classification. This information established recruitment directionality when many enzymes were shared between subnetworks, and those subnetworks that shared particularly large number of their enzymes with other subnetworks became "hubs". Using this approach we discovered that most enzymatic activities were associated with the nine most ancient and widely distributed folds and that modern (protein-based) metabolism appeared explosively (197). This is remarkable and is compatible with arguments related to rapid duplication and divergence of genes in Archean times (223). We also discovered that modern metabolism originated in enzymes with the P-loop hydrolase fold in nucleotide metabolism, probably in pathways linked to the purine metabolic subnetwork (197). Consequently, the first enzymatic take-over of a ribozymic or prebiotic chemistry involved the synthesis of nucleotides for the RNA world. We believe this is a remarkable finding that links the onset of the protein world to the RNA world.

7. ORIGINS AND EVOLUTION OF THE TRIPARTITE WORLD

7.1. Genomic census and most parsimonious scenarios for the origin of diversified life

The survey of protein repertoires and interactomes at fold and FSF levels already reveals illuminating evolutionary patterns related to the tripartite world. For example, distribution of folds among organisms belonging to the three superkingdoms described in a Venn diagram showed that the majority of folds (506 in number) were shared by Archaea, Bacteria and Eukarya, and that 516-682 folds were shared by any two superkingdoms (being maximal in Eukarya and Bacteria) (Figure 1). Parsimony considerations based on these distributions

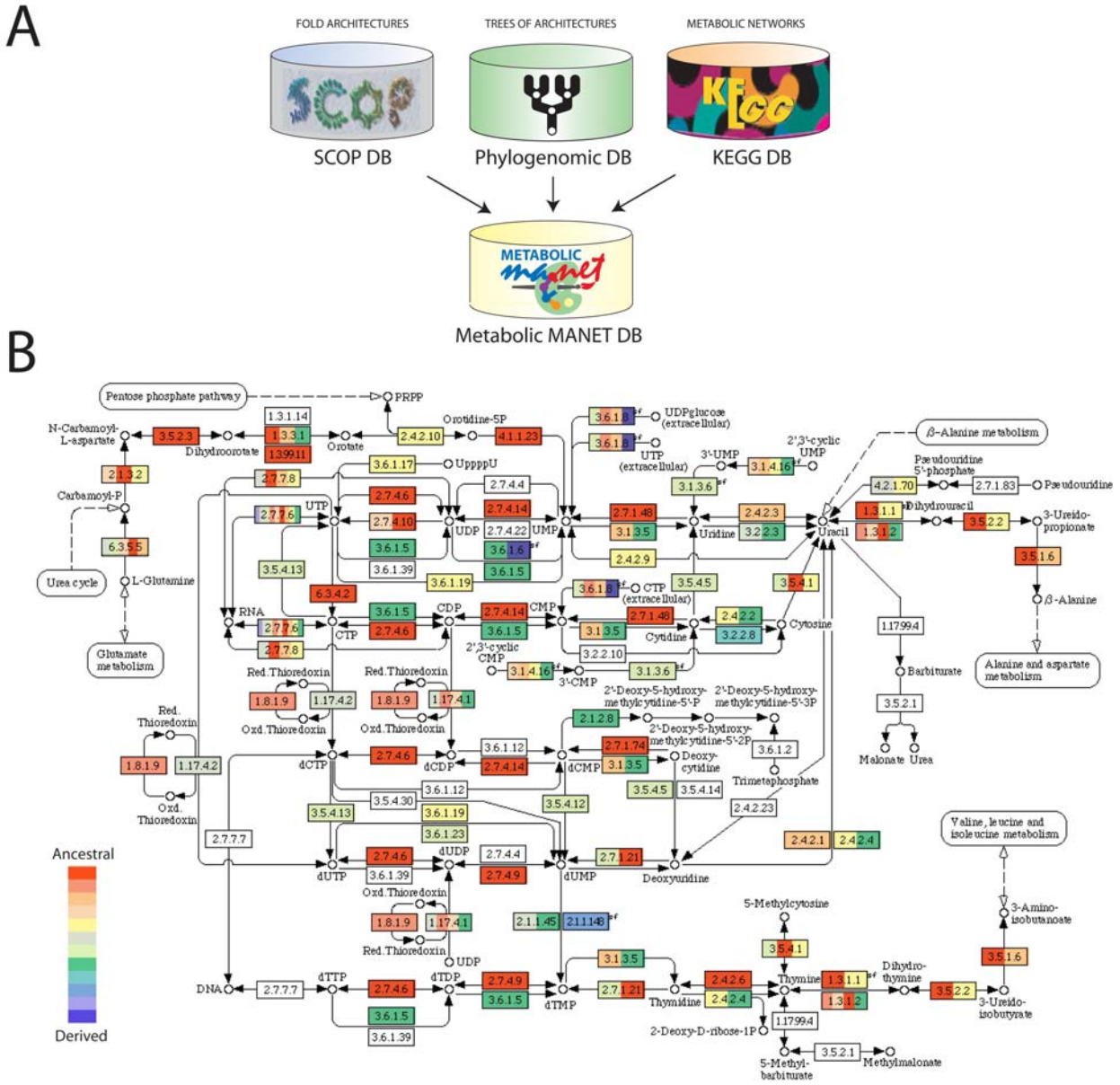

Figure 6. The MANET database. A, MANET links information in the metabolic pathways database of KEGG, the SCOP database, and phylogenomic trees reconstructed from a genome census of protein architectures. B, MANET traces the evolution of protein structure in biomolecular networks. The ancestry of each architecture was literally painted onto enzymes in metabolic subnetworks, based on the usage of folds in each enzyme. The resulting color code gives a lower bound on the time at which the architecture might have been adopted for a particular enzymatic activity.

alone suggest the ancestor to the three superkingdoms was endowed with a virtual genome akin to Eukarya. These considerations were recently well elaborated by Kurland *et al.* (224). Under such a scenario, the proteomes of prokaryotes had to suffer differential reductive tendencies in their repertoires to reflect the observed architectural distributions. Interestingly, reductive tendencies in Archaea and Bacteria were expressed also in protein length, with length variation localized in terminal sequences (224). In contrast with domain architectures, the majority of domain combinations were unique to Archaea, Bacteria and Eukarya (Figure 2), consistent with their late expansive appearance in evolution (27). However, there was again an unbalanced sharing of domain combinations that suggests Eukarya was more closely related to Bacteria than to Archaea. Reconstruction of trees of proteomes generally placed the Eukarya at the base of the tree, supporting the ancient nature of this lineage and pasimony considerations (18).

7.2. The rise of the tripartite world

Tracing features depicting organismal diversity along the branches of evolutionary trees of architectures allowed inference of the relative timing for the emergence

of the superkingdoms, identification of episodes of architectural loss and diversification in organismal superkingdoms, and identification of a late and quite massive rise of architectural novelties in Eukarya that was probably linked to the rise of multicellularity (21,22,188). Folds associated with processes related to multicellularity (e.g. apoptosis, cell death, adhesion and recognition) contained multiple domains and appeared both immediately after prokaryotic diversification (mostly folds common to organismal domains) and during eukaryotic diversification (mostly eukaryotic-specific). Our observations indicate that protein novelties unique to organismal lineages appeared late and in defined order during evolution. The proteomes of diversified organisms originated apparently from ancestors that already shared an arrangement of quite complicated molecular architectures and biological functions (22). This view is consistent with a proto-eukaryote (225,226) responsible for 'crystallizing' diversified life (227).

To unfold the data embedded in the trees of architectures (Figure 5A), we quantified the distribution of protein architectures among proteomes by a distribution index, defined as the relative number of species that use each architecture. This index, when plotted against ancestry values of an architecture derived from its position on a phylogenetic tree, revealed fascinating patterns (22). In particular, it allowed us to distinguish three epochs in protein evolution: (i) architectural diversification in which members of the ancestral community diversified their architectural repertoire through differential "loss" of folds; (ii) superkingdom specification, where superkingdoms Archaea, Bacteria and Eukarya are specified through invention of superkingdom-specific architectures; and (iii) organismal diversification, where protein folds and FSFs specific to relatively small sets of organisms appear as the result of diversification of organismal lineages (22) (Figure 5A). To explicate, the most ancient architectures were present in all proteomes, suggesting the existence of a universal communal ancestor that was complex and architecturally rich. Representation decreased with decreasing age until it approached cero about half way through evolution. At this point, a large number of new architectures were clustered, each specific to a small number of organisms. Later in evolutionary time an opposite trend takes place, in which architectures increase their representation in proteomes.

When the relationship between distribution index and ancestry was dissected for the three superkingdoms, the meaning of these patterns became apparent. The decrease of architectural representation in proteomes during the first third of the evolutionary timeline was mostly due to archaeal species. Thus, Archaea established the first organismal divide by losing a substantial number of architectures early in evolution. On the other hand, the increase in architectural representation during the last half of the evolutionary timeline was mostly due to Eukarya, through incorporation of most newly invented architectures into their proteomes. In addition, Eukarya retained more ancestral protein architectures compared to the prokaryotes. Bacteria take an intermediate position, displaying neither a massive reductive nor a prominent retentive tendency, sort of diversifying the distribution of architectures between species. Thus, emergence of the three superkingdoms of life seems to have been shaped by "reductive evolution", through reductive tendencies of prokaryotes relative to eukaryotes in their usage of architectures, which we think reflects their adaptation to the environment (see below). This is illustrated by the accumulation of architectures along the branches of a most-parsimonious tree of global repertoires corresponding to proteomes in superkingdoms (Figure 7A). Remarkably, the FSF complement of the universal ancestor had a fairly complex proteome with ~80 architectures. This matches a recent ancestral state reconstruction of the gene content of the universal ancestor that revealed a complex genome with a gene complement similar to that of extant free-living prokaryotes (228).

We interpret the entire history of evolution of protein architecture in ecological terms. As we have seen, Archaea was the first superkingdom to segregate from the rest by adopting the minimalist approach to the molecular repertoire. Archaeal-like ancestor may have been defined by adaptation to physical extremes, because extreme environmental conditions limit the number of functional protein variants, thus reducing the number of viable protein folds in a cell (229). The eukaryal-like emerging lineage with its large and diverse architectural repertoire may have been better suited for K-selection by exploiting flexibility of use of environmental resources. Later, some lineages may have discovered the advantages of rapid growth in times when nutrients were accessible, entering into r-selection and a competitive strategy of survival, diversification, and streamlining (230), adopting a bacterial lifestyle. This decision encouraged genome reduction to shorten replication cycles (streamlining) and increasing the variety of metabolic functions to gain competitive advantage (diversification).

7.3. An ecological hypothesis of organismal origin

One of the present challenges in biology is linking genetic diversity to the physiological and ecological diversity that differentiates the species. In other words, what makes an organism be what it is? The physiology of an organism evolves in response to its environment, and is determined by the molecular repertoire of the organism and its uses in molecular networks. Thus, there must be a link between cellular physiology, ecology and evolution. We hypothesize that the three organismal superkingdoms emerged by committing to a certain pattern of usage of protein architectures as a result of adaptation to different niches within the primordial environment: *the primordial strategy hypothesis.* Specifically and as we discussed above, we suggest that Archaea were formed through adaptations to extreme physical conditions, Bacteria formed through interspecies competition, and Eukarya adapted to change and instability by developing predation.

We built an environmental chart by systematically considering two environmental variables: nutrient levels and physical state of the environment, and used the chart to make predictions about the physiological

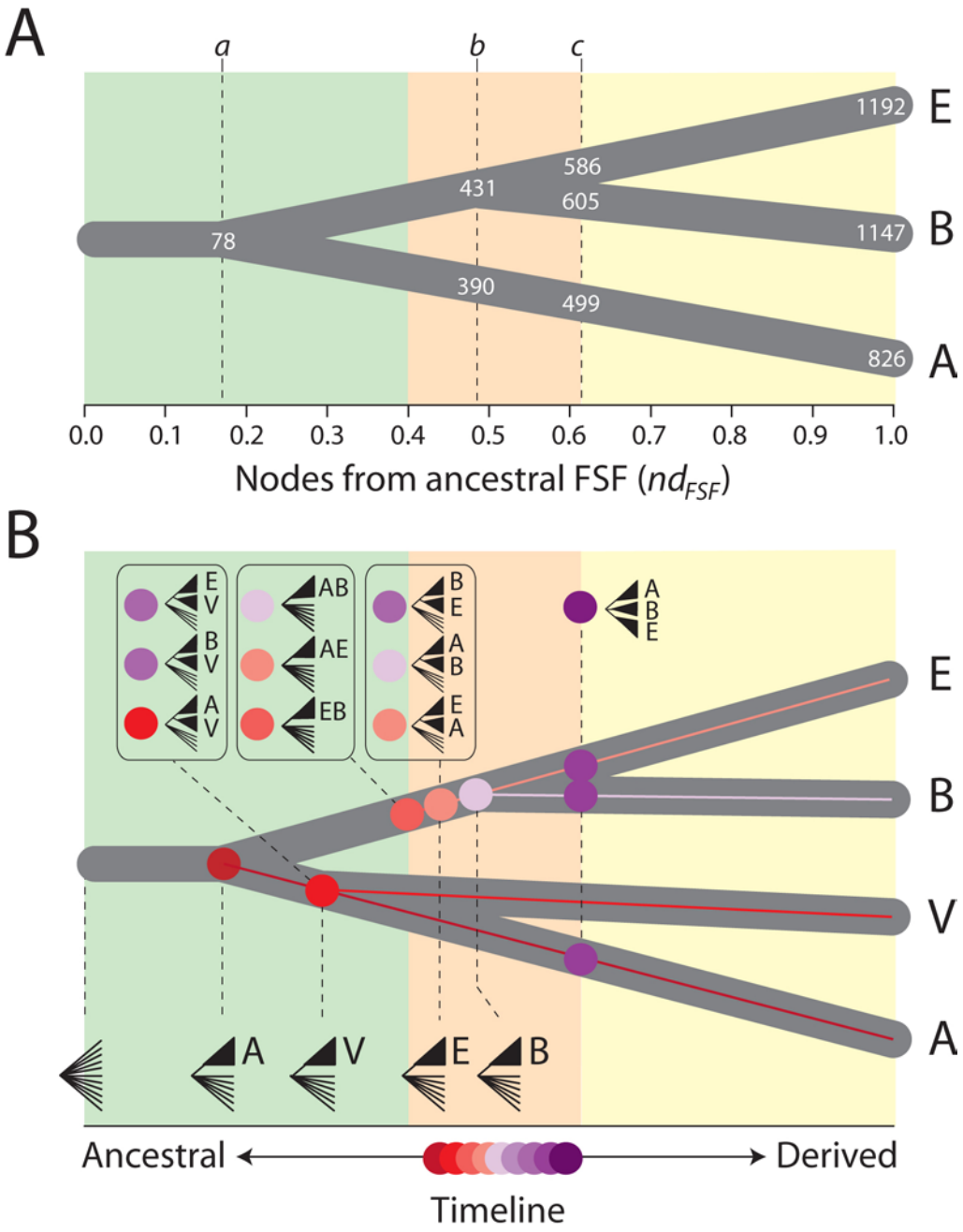

Figure 7. Evolution of the organismal world. A, Global most-parsimonious scenario for organismal diversification of proteomes based on architectural distribution patterns and trees of proteomes reconstructed from architectures of different age (Wang *et al.* 2007). Numbers in branches describe the FSF content of ancestors of modern superkingdoms at the time when the first architecture was lost in a superkingdom (*a*) and the first architectures unique to Bacteria (*b*) and to Archaea and Eukarya (*c*) arose in evolution. Numbers at the tip describe the extant FSF repertoires of superkingdoms. The tree overlaps an evolutionary timeline and shadings that corresponds to the three evolutionary epochs of the protein world (see Figure 5 for details). B, Global most-parsimonious scenario for organismal diversification based on information embedded in the history of tRNA. A total of 571 tRNAs with sequence, base modification, and structural information were used to reconstruct a global tree, which failed to show clear monophyletic groupings. Ancestries of lineages were then inferred by constraining sets of tRNA molecules into monophyletic groups representing competing (shown in boxes) or non-competing phylogenetic hypotheses and measuring tree sub-optimality and lineage coalescence (illustrated with color hues in circles) (139). Timelines and most parsimonious topologies were used to reconstruct universal trees that reveal the ancient nature of Archaea and the viral world.

features necessary for organisms to be adaptive in each environmental niche (Yafremava *et al.*, ms. in preparation)(Figure 8A). The most prominent predicted physiological distinctions between niche-adapted organisms include: (i) *limitations on the molecular repertoire* in the lower left corner of the chart, which corresponds to environments with extremely high temperature, salinity, pH, etc. that require special molecular adaptations for DNA, proteins and membrane to retain structural stability; (ii) *streamlining and specialization of physiology* in the center, which corresponds to comfortable environment with abundant resources that invite interspecies competition; and (iii) *flexibility, storage and resulting multicellularity* in the upper right corner of the chart, which corresponds to instabilities of environmental conditions and food supply.

To confirm these predictions on the molecular level, we quantified the molecular repertoires of organisms in the three domains. We found that the main physiological properties of Archaea, Bacteria and Eukarya (respectively) correspond well to those three categories, supporting the hypothesis that Bacteria are most adapted to competition, Archaea to physical extremes, and Eukarya to predation. Our survey of niche disposition of extant organisms also supports this assignment (Figure 8B).

We find it useful to compare protein fold usage by organisms with a Lego game. The complexity and variety of entities one is able to come up with when playing a combinatorial Lego game depends on the number and variety of Lego pieces that are available. Archaea are placed under the tightest constraints on availability of parts due to destructive influences of their environment, resulting in entities/organisms very similar to each other: they use the most limited set of folds, and use all folds roughly to the same extent. Bacteria live under much more relaxed conditions that allow for a great number of parts to exist. However, a single bacterial species, pressured by competition, is unable to use them all under the penalty of increasing the size of its genome and cell volume, and slowing down reproduction. A bacterial species "picks out" a fraction of the total architectural complement, choosing the fold combination most distinct from its immediate competitors. The disturbed environment of Eukarya favors flexibility of responses. Thus, they take the strategy of having as many protein folds as possible and using them in widely different combinations, taking advantage of the combinatorial power afforded by the large selection of building blocks, so as to cover most eventualities. Indeed, every fold is used by a higher fraction of eukaryotic species compared to the other two domains. The number of folds used by all species of a superkingdom is also the largest for Eukarya, and smallest for Archaea. Thus, Eukarya make the most complete use of the total molecular repertoire available, Archaea use the least, and Bacteria take an intermediate position.

8. THE EARLY EVOLUTIONARY APPEARANCE OF VIRUSES

Phylogenomic analyses established an origin of cellular life but did not address questions related to the origin and evolution of viruses. Viruses have long been considered fragments of cellular genomes and not living organisms and were generally excluded from consideration in evolutionary scenarios of the tripartite world. Consequently, currently available models for the universal tree of life uniformly exclude viruses, despite being important components of the biosphere. Recent studies have re-evaluated the importance of viruses and their potential roles in early cellular evolution (231). For example, it is shown that both RNA and DNA viruses may have been more ancient than previously thought, possibly even more ancient than the common ancestor of life (231). Comparative genomic analyses also suggested viruses could be the source of new proteins for cells (232). Overall, it is likely that many DNA informational proteins encoded today in cellular genomes originated first in the viral world and were transferred later on randomly into the three cellular domains. Patrick Forterre recently proposed that DNA itself appeared in ancestral viral lineages (233,234). He later on extended this proposal by suggesting that the DNA replication machineries of each superkingdom originated from three different viruses (235). In his latest proposal, each cellular domain originated independently from the fusion of an RNA-based cell and a large DNA virus (236).

We recently reconstructed phylogenies directly from the sequence and structure of tRNA and constrained sets of tRNAs belonging to different superkingdoms or viruses into monophyletic groups representing competing or non-competing hypotheses (139). The exercise resulted in timelines of organismal diversification and most parsimonious tree topologies that are illustrated in Figure 7B. Remarkably, organismal timelines showed Archaea was the most ancient superkingdom, followed by viruses, and superkingdoms Eukarya and Bacteria, a result that is congruent with our phylogenomic analysis of protein architecture (22). Most importantly, the viral lineage had an origin in the Archaeal lineage that was also quite ancient. The origin of the viral lineage in the Archaea is remarkable, especially if one considers the exceptional diversity and morphotype complexity of archaeal viruses (237). Such an origin is compatible with the proposal by Forterrre and colleagues that the transition from RNA to DNA genomes occurred in the viral world, and that cellular DNA and its replication machineries originated via transfers from DNA viruses to RNA cells.

9. SUMMARY AND PERSPECTIVE

The ongoing efforts in comparative, structural and evolutionary genomics provide exponentially growing repertoires of components of biological systems and the ability to study patterns and processes related to the evolution of macromolecules and genomes. Here we introduce phylogenetic methods that are capable of generating intrinsically rooted phylogenies and therefore provide the mean to identify evolution's arrow without the need to invoke local external hypotheses of relationship (outgroups). We also describe how these can be used to generate timelines of discovery of components in a number of biological systems, including RNA molecules,

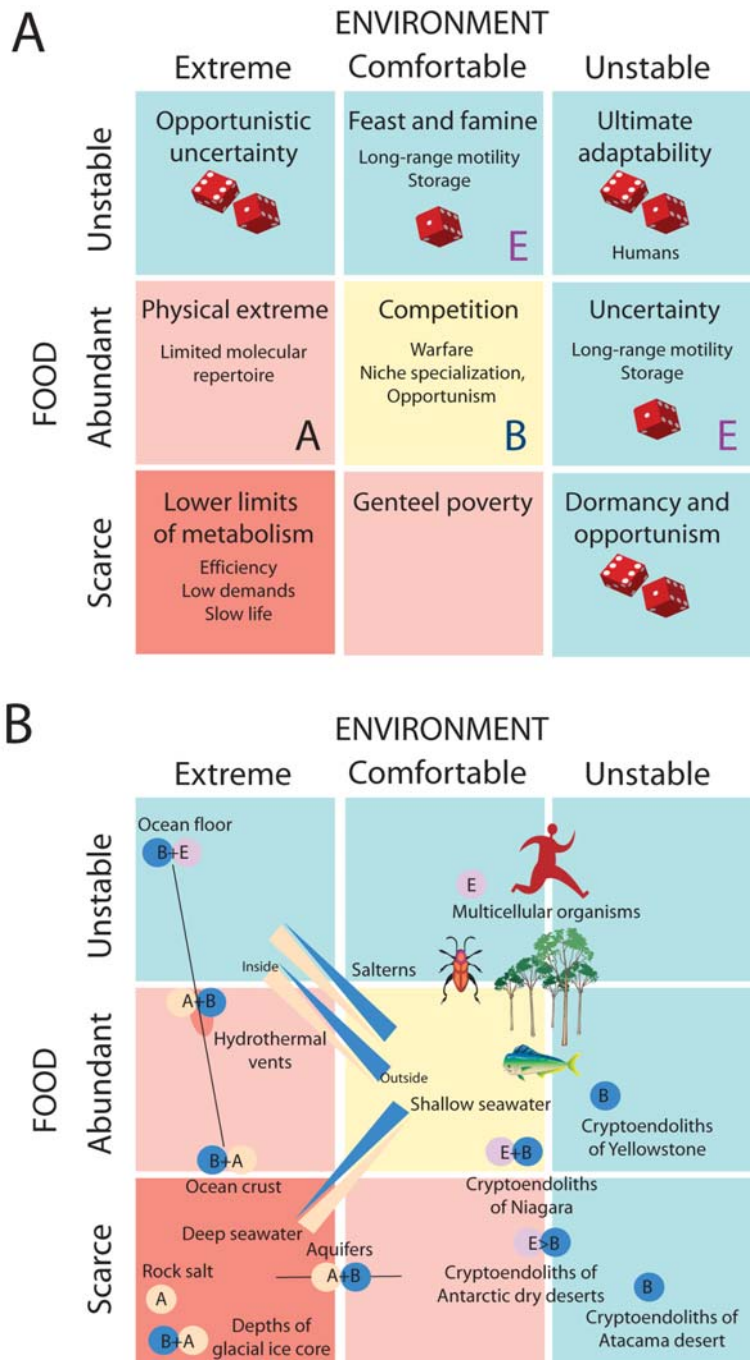

Figure 8. The tripartite world on the environmental chart. A, This environmental chart displays imaginable environmental niches by systematically considering two variables: nutrient levels and physical state of the environment. Food varies from scarce to abundant. Environmental conditions vary from extreme, such as extreme temperature and salinity, to comfortable. Each variable can also unpredictably fluctuate between those values (labeled 'unstable'). The three superkingdoms, Archaea (A), Bacteria (B) and Eukarya (E), were predicted to occupy these niches (quadrants) based on their physiology. B, Placement of extant organisms on the environmental chart based on ecological data. Blue color dots identifies B, tan A, and violet E. A+B means both superkingdoms are present and E>B means E are present to a greater extent than B. Bars and lines indicate placement of organisms along some known environmental gradients: salinity (salterns), temperature (hydrothermal vents), pressure (ocean floor and crust, aquifers, seawater). Triangles indicate known distribution of organisms along a gradient. When a distribution is not known, a line is drawn instead. Note: exceptions and overlaps between organismal niches exist, but are not depicted on this diagram.

proteomes, intra-molecular interactomes, and biological networks. Making sense of the order of discovery of individual components had unanticipated benefits. It unfolded remarkable patterns associated with the evolution of the individual systems. For example, timelines of discovery of substructures in RNA molecules allowed identification of molecular origins for tRNA, tRNA-derived SINEs, and rRNA, with evolutionary implications for translation, the genetic code, and life. Similarly, timelines of architectural discovery in proteins showed even more interesting patterns related to the origin and evolution of protein structure, enzymes in metabolism, and modules in the protein world. The emerging picture from these studies is the relatively quick formation of an ancient world populated with a community of relatively complex organisms in which both RNA and proteins played important catalytic roles. The evolution of this ancient world was probably mostly driven by recruitment, signs of which were left imprinted for example in tRNA and metabolic networks. This view is compatible with recent developments related to the theory of a chemo-autotrophic origin of life and the emergence of pioneer organisms within a flow of volcanic exhalations followed by enzymatization and cellularization processes (238). In particular, the concept of cellularization and the emergence of chiral phosphoglycerol lipids from racemic mixtures of ancestral molecules (that would generate two kinds of cellular envelopes, one of them defining the archaeal lineage) could explain an early emergence of Archaea and is compatible with our phylogenomic analyses. Timelines at different levels of organization (molecules, repertoires, networks) showed episodes of both specialization and simplification, illustrated for example with the loss of substructures in RNA molecules, the loss of degeneracy in the genetic code, the formation of domains with specialized functions, or the loss of enzymes and pathways in metabolic subnetworks. These observations have important consequences, especially when considering the fundamental role that recruitment plays in evolution.

10. ACKNOWLEDGMENTS

Research described here was supported by grants from the National Science Foundation (MCB-0343126), the Office of Naval Research (TRECC A6538-A76), the C-FAR Sentinel Program, and the Critical Research Initiative of the University of Illinois. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

11. REFERENCES

1. Burley, S.K. and J.B. Bonanno: Structuring the universe of proteins. *Annu Rev Genomics Hum Genet 3*, 243-262 (2002)
2. Zhang, C. and S.H. Kim: Overview of structural genomics: from structure to function. *Curr Opin Chem Biol 7*, 28-32 (2003)
3. Grant, A., D. Lee, and C. Orengo: Progress towards mapping the universe of protein folds. *Genome Biol 5*, 107 (2004)
4. Mindell, D.P. and A. Meyer: Homology evolving. *Trends Ecol Evol 16,* 434-440 (2001)
5. Holder, M. and P.O. Lewis: Phylogeny estimation: traditional and Bayesian approaches. *Nature Rev 4*, 275-284 (2003)
6. Savva, G., J. Dicks, and I.N. Roberts: Current approaches to whole genome phylogenetic analysis. *Brief Bioinform 4,* 63-74 (2003)
7. Albert, V.A.: Parsimony, phylogeny and genomics. *Oxford University Press*, NY (2005)
8. Baldauf, S.L., A.J. Rogers, I. Wenk-Siefert, and W.F. Doolittle: A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science 290,* 972-977 (2000)
9. Brown, J.R., C.J. Douady, M.J. Italia, W.E. Marshall, and M.J. Stanhope: Universal trees based on large combined protein sequence data sets. *Nature Genet 28,* 281-285 (2001)
10. Ciccarelli, F.D., T. Doerks, C.V. Mering, C.J. Creevey, B. Snel, and P. Bork: Towards automatic reconstruction of a highly resolved tree of life. *Science 311,* 1283-1287 (2006)
11. Gerstein, M.: Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct Funct Genet 33*, 518-534 (1998)
12. Tekaia, F., A. Lazcano, and B. Dujon: The genomic tree as revealed from whole proteome comparisons. *Genome Res 9,* 550-557 (1999)
13. Wolf, Y.I., S.E. Brenner, P.A. Bash, and E.V. Koonin: Distribution of protein folds in the three superkingdoms of life. *Genome Res 9*, 17-26 (1999)
14. House, C.H. and S.T. Fitz-Gibbons: Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J Mol Evol 54,* 539-547 (2002)
15. Lin, J. and M. Gerstein: Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res 10*, 808-818 (2000)
16. Wolf, Y.I., I.B. Rogozin, N.V. Grishin, R.L. Tatusov, and E.V. Koonin: Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol 1,* 8 (2001)
17. Snel, B., P. Bork, and M. Huynen: Genome evolution: gene fusion versus gene fission. *Trends Genet 16*, 9-11 (2000)
18. Caetano-Anollés, G. and D. Caetano-Anollés: An evolutionarily structured universe of protein architecture. *Genome Res 13*, 1563-1571 (2003)
19. Dutilh, B.E., M.A. Huynen, W.J. Bruno, and B. Snel: The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J Mol Evol* 58, 527-539 (2004)
20. Yang, S., R.F. Doolittle, and P.E. Bourne: Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA 102*, 373-378 (2005)
21. Wang, M., S.M. Boca, R. Kalelkar, J.E. Mittenthal, and G. Caetano-Anollés: A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. *Complexity 12*, 27-40 (2006)
22. Wang, M., L.S. Yafremava, D. Caetano-Anollés, J.E. Mittenthal, and G. Caetano-Anollés: Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res 17*, 1572-1585 (2007)

23. Dadenkar, T., B. Snel, M. Huynen, and P. Bork: Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23, 324-328 (1998)
24. Wolf, Y.I., I.B. Rogozin, and E.V. Koonin: Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res 14,* 29-36 (2004)
25. Korbel, J.O., B. Snel, M.A. Huynen, and P. Bork: SHOT: a web server for the construction of genome phylogenies. *Trends Genet 18,* 158-162 (2002)
26. Wang, M. and G. Caetano-Anollés: Global phylogeny determined by the combination of protein domains in proteomes. *Mol Biol Evol 23*, 2444-2454 (2006)
27. Wang, M. and G. Caetano-Anollés: The evolutionary mechanics of protein domain organization in proteomes. Ms. submitted (2008)
28. Wolf, Y.I., I.B. Rogozin, N.V. Grishin, and E.V. Koonin: Genome trees and the tree of life. *Trends Genet 18*, 472-479 (2002)
29. Delsuc, F., H. Brinkmann, and H. Philippe: Phylogenomics and the reconstruction of the tree of life. *Nature Rev Genet 6*, 361-375 (2005)
30. Doolittle, R.F.: Evolutionary aspects of whole-genome biology. *Curr Opin Struct Biol 15*, 248-253 (2005)
31. Deeds, E.J., H. Hennessey, and E.I. Shakhnovich: Prokaryotic phylogenies inferred from protein structural domains. *Genome Res 15*, 393-402 (2005)
32. Woese, C.R., O. Kandler, and M.L. Wheelis: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. *Proc Natl Acad Sci USA 87*, 4576-4579 (1990)
33. Assembling the Tree of Life (ATOL). http://atol.sdsc.edu
34. Bryant, H.N.: The polarization of character transformations in phylogenetic systematics - Role of axiomatic and auxiliary assumptions. *Syst Zool 40,* 433-445 (1991)
35. Bryant, H.N.: Hypothetical ancestors and rooting in cladistic analysis. *Cladistics 13,* 337-348 (1997)
36. Morowitz, J.H.: Beginning of cellular life. *Yale University Press*, CT (1992)
37. Benner, S.A., A.D. Ellington, and A. Tauer: Modern metabolism as a palimpsest of the RNA world. *Proc Natl Acad Sci USA 86,* 7054-7058 (1989)
38. Bajaj, M. and T. Blundell: Evolution and the tertiary structure of proteins. *Annu Rev Biophys Bioeng 13*, 453-492 (1984)
39. Vukmirovic, O.G. and S.M. Tilghman: Exploring genome space. *Nature 405*, 820-822 (2000)
40. Rokas, A. and P.W.K. Holland: Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol 15*, 454-459 (2000)
41. Sober, E. and M. Steel: Testing the hypothesis of common ancestry. *J Theor Biol 218*, 395-408 (2002)
42. Penny, D., M.D. Hendy, and A.M. Poole: Testing fundamental evolutionary hypotheses. *J Theor Biol 223*, 377-385 (2003)
43. Mossell, E.: On the impossibility of reconstructing ancestral data and phylogenies. *J Comp Biol 10*, 669-678 (2003)
44. Philippe, H. and J. Laurent: How good are deep phylogenetic trees? *Curr Opin Genet Dev 8*, 6161-623 (1998)
45. Schuster, P., W. Fontana, P. Stadler, and I. Hofacker: From sequences to shapes and back: a case study in RNA secondary structures. *Proc R Soc Lond B 255,* 279–284 (1994)
46. de Visser, J.A.G.M., J. Hermisson, G.P. Wagner, L. Ancel Meyers, *et al.*: Perspective: Evolution and detection of genetic robustness. *Evolution 57,* 1959-1972 (2003)
47. Schuster, P. and P.F. Stadler: Networks in molecular evolution. *Complexity 8*, 34-42 (2003)
48. Fontana, W.: Modelling 'evo-devo' with RNA. *BioEssays 24*, 1164-1177 (2002)
49. Schultes, E.A. and D.P. Bartel: One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science 289*, 448-452 (2000)
50. Babajilde, A., R. Farber, I.L. Hofacker, J. Inman, A.S. Lapedes, and P.F. Stadler: Exploring protein sequence space using knowledge based potentials. *J Theor Biol 212*, 35-46 (2001)
51. Keefe, A.D. and J.W. Szostak: Functional proteins from a random-sequence library. *Nature 410*, 715-718 (2001)
52. Bornberg-Bauer, E. and H. S. Chang: Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci USA 96,* 10689-10694 (1999)
53. Taverna, D.M. and R.A Goldstein: Why are proteins so robust to site mutations? *J Mol Biol 315,* 479-484 (2002)
54. Wroe, R., E. Bornberg-Bauer, and H.S. Chan: Comparing folding codes in simple heteropolymer models of protein evolutionary landscapes: robustness of the superfunnel paradigm. *Biophys J* 88, 118-131 (2005)
55. Huynen, M., R. Gutell, and D. Konings: Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol 267*, 1104-1112 (1997)
56. Wagner, A.: Robustness and evolvability: a paradox resolved. *Proc R Soc Lond B 275,* 91-100 (2008)
57. Albert, R., H. Jeong, and A.L. Barabási: Error and attack of complex networks. *Nature 406,* 378-382 (2000)
58. Giaever, G., A.M. Chu, L. Ni, C. Connelly, R. Riles *et al.*: Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature 418,* 387-391 (2002)
59. Bowie, J.U., J.F. Reidhaar-Olson, W.A. Lim, and R.T. Sauer: Deciphering the message in protein sequences: tolerance to amino acid substitution. *Science 247,* 1306-1310 (1990)
60. Borenstein, E. and E. Ruppin: Direct evolution of genetic robustness in microRNA. *Proc Natl Acad Sci USA 103,* 6593-6598 (2006)
61. Sanjuán, R., J. Forment, and S.F. Elena: In silico redicted robustness of viroids RNA secondary structures. I. The effect of single mutations. *Mol Biol Evol 23,* 1427-1436 (2006)
62. Ancel, L.W. and W. Fontana: Plasticity, evolvability, and modularity in RNA. *J Exp Zool (Mol Dev Evol) 288,* 242-283 (2000)
63. Ancel Meyers, L., J.F. Lee, M. Cowperthwaite, and A.D. Ellington: The robustness of naturally and artificially selected nucleic acid secondary structures. *J Mol Evol 58,* 681-691 (2004)
64. van Nimwegen, E., J. Crutchfield, and M. Huynen: Neutral evolution of mutational robustness. *Proc Natl Acad Sci USA 96*, 9716–9720 (1999)

65. Bussemaker, H.J., D. Thirumalai, and J.K. Bhattacharjee: Thermodynamic stability of folded proteins against mutations. *Phys Rev Lett 79,* 3530-3533 (1997)
66. Vendruscolo, M., A. Maritan, and J.R. Banavar: Stability threshold as a selection principle for protein design. *Phys Rev Lett 78,* 3967-3970 (1997)
67. Wagner, A. and P. Stadler: Viral RNA and evolved mutational robustness. *J Exp Zool (Mol Dev Evol) 285,* 119-127 (1999)
68. Rutherford, S.L. and S. Lindquist: Hsp90 as a capacitor for morphological evolution. *Nature 396,* 336-342 (1998)
69. Queitsch, C., T. Sangster, and S. Lindquist: Hsp90 as a capacitor of phenotypic variation. *Nature 417,* 618-624 (2002)
70. Schrödinger, E.: What is life? *John Wiley and Sons,* NY (1944)
71. Gladyshev, G.P.: On the thermodynamics of biological evolution. *J Theor Biol 75*, 425-441 (1978)
72. Black, S.: On the thermodynamics of evolution. *Persp Biol Med 21,* 348-356 (1978)
73. Wicken, J.S.: A thermodynamic theory of evolution. *J Theor Biol 87*, 9-23 (1980)
74. Gladyshev, G.P. and Y.A. Ershov: Principles of the thermodynamics of biological systems. *J Theor Biol 94*, 301-343 (1982)
75. Nicolis, G. and I. Prigogine: Exploring complexity. *W.M. Freeman,* NY (1989)
76. Schneider, E.D. and J.J. Kay: Complexity and thermodynamics: towards a new ecology. *Futures 26,* 626-647 (1994)
77. Schneider, E.D. and J.J. Kay: Life as a manifestation of the second law of thermodynamics. *Math Comp Modelling 19*, 25-48 (1994)
78. Caetano-Anollés, G.: Grass evolution inferred from chromosomal rearrangements and geometrical and statistical features in RNA structure. *J Mol Evol 60*, 635-652 (2005)
79. Stegger, G., H. Hofman, J. Fortsch, H.J. Gross, J.W. Randles, H.L. Sanger, and D. Riesner: Conformational transitions in viroids and virusoids: comparison of results from energy minimization algorithm and from experimental data. *J Biomol Struct Dynam 2*, 543-571 (1984)
80. Higgs, P.G.: RNA secondary structure: a comparison of real and random sequences. *J Phys I France 3*, 43-59 (1993)
81. Higgs, P.G.: Thermodynamic properties of transfer RNA: a computational study. *J Chem Soc Faraday Trans 91*, 2531-2540 (1995)
82. Schultes, E.A., P.T. Hraber, and T.H. LaBean: Estimating the contributions of selection and self-organization in RNA secondary structure. *J Mol Evol 49*, 76-83 (1999)
83. Steffens, W. and D. Digby: mRNA have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res 27*, 1578-1584 (1999)
84. Gultyaev, P.A., F.H.D. van Batenburg, and C.W.A. Pleij: Selective pressures on RNA hairpins *in vivo* and *in vitro*. *J Mol Evol 54*, 1-8 (2002)
85. Higgs, P.G.: RNA secondary structure: physical and computational aspects. *Quaterly Rev Biophys 33*, 199-253 (2000)
86. Kierzek, E., E. Biala, and R. Kierzek: Elements of thermodynamics in RNA evolution. *Acta Biochim Polonica 48*, 485-493 (2001)
87. Billoud, B., M.A. Guerrucci, M. Masselot, and J.S. Deutsch: Cirripede phylogeny using a novel approach: molecular morphometrics. *Mol Biol Evol 17*, 1435-1445 (2000)
88. Collins, L.J., V. Moulton, and D. Penny: Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J Mol Evol 51*, 194-2004 (2000)
89. Caetano-Anollés, G.: Novel strategies to study the role of mutation and nucleic acid structure in evolution. *Plant Cell Tissue Org Culture 67*, 115-132 (2001)
90. Caetano-Anollés, G.: Evolved RNA secondary structure and the rooting of the universal tree of life. *J Mol Evol 54*, 333-345 (2002)
91. Caetano-Anollés, G.: Tracing the evolution of RNA structure in ribosomes. *Nucleic Acids Res 30*, 2527-2587 (2002)
92. Swain, T.D. and D.J. Taylor: Structural rRNA characters support monophyly of raptorial limbs and paraphyly of limb specialization in water fleas. *Proc R Soc Lond B 270*, 887-896 (2003)
93. Sun, F.-J., S. Fleurdépine, C. Bousquet-Antonelli, G. Caetano-Anollés, and J.-M. Deragon: Common evolutionary trends for tRNA-derived SINE RNA structures. *Trends Genet 23*, 26-33 (2007)
94. Sun, F.-J. and G. Caetano-Anollés: The origin and evolution of tRNA inferred from phylogenetic analysis of structure. *J Mol Evol 66,* 21-23 (2008)
95. Hartwell, L.H., J.J. Hopfield, S. Leibler, and A.W. Murray: From molecular to modular cell biology. *Nature 401*, C47-C52 (1999)
96. Krakauer, D.C. and J.B. Plotkin: Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci USA 99,* 1405-1409 (2002)
97. Berezovsky, I.N., A.Y. Grosberg, and E.N. Trifonov: Closed loops of nearly standard size: common basic elements of protein structure. *FEBS Lett 446,* 283-286 (2000).
98. Sobolevsky, Y., Z.M. Frenkel, and E.N. Trifonov: Combinations of ancestral modules in proteins. *J Mol Evol* doi 10.1007/s00239-007-9032-x (2007)
99. Söding, J. and A.N. Lupas: More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays 25*, 837-846 (2003)
100. Vogel, C., M. Bashton, N.D. Kerrison, C. Chothia, and S.A. Teichmann: Structure, function and evolution of multidomain proteins. *Curr Opin Struc Biol 14*, 208–216 (2004)
101. Pereira-Leal, J.B., E.D. Levy, C. Kamp, and S.A. Teichmann: Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol 8*, R51 (2007)
102. Ravasz, E., A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási: Hierarchical organization of modularity in metabolic networks. *Science* 297:1551–1555 (2002)
103. Spirin, V. and L.A. Mirny: Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100, 12123-12128 (2003)
104. Gilbert, W.: The RNA world. *Nature 319*, 618 (1986)
105. Woese, C.R.: The genetic code: The molecular basis for genetic expression. *Harper & Row,* NY (1967)

106. Crick, F.H.C.: The origin of the genetic code. *J Mol Biol 38,* 367-379 (1968)
107. Orgel, L.E.: Evolution of the genetic apparatus. *J Mol Biol* 38, 381-393 (1968)
108. Cech, T.R., A.J. Zaug, and P.J. Grabowski: *In vitro* splicing of the ribosomal RNA precursor of *Tetrahymena*: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell 27,* 487-496 (1981)
109. Guerrier-Takada, C., K. Gardiner, T. Marsh, N. Pace, and S. Altman: The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell 35,* 849-857 (1983)
110. Jeffares, D.C., A.M. Poole, and D. Penny: Relics from the RNA World. *J Mol Evol 46,* 18-36 (1998)
111. Yusupov, M.M., G.Z. Yusupova, A. Baucom, K. Lieberman, T.N. Earnest, J.H.D. Cate, and H.F. Noller: Crystal structure of the ribosome at 5.5 Å resolution. *Science 292*, 883-896 (2001)
112. Nissen, P., J. Hansen, N. Ban, P.B. Moore, and T.A. Steitz: The structural basis of ribosome activity in peptide bond synthesis. *Science 289,* 920-930 (2000)
113. Ogle, J.M., A.P. Carter, and V. Ramakrishnan: Insights into the decoding mechanism from recent ribosome structures. *Trends Biochem Sci 28,* 259-266 (2003)
114. Schuwirth, B.S., M.A. Borovinskaya, C.W. Hau, W. Zhang, A. Vila-Sanjurjo, J.M. Holton, and J.H.D. Cate: Structures of the bacterial ribosome at 3.5 Å resolution. *Science 310,* 827-834 (2005)
115. Noller, H., V. Hoffarth, and L. Zimniak: Unusual resistance of peptidyl transferase to protein extraction procedures. *Science 256,* 1416-1419 (1992)
116. Steitz, T.A. and P.B. Moore: RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem Sci 28,* 411-418 (2003)
117. Eddy, S.R.: Non-coding RNA genes and the modern RNA world. *Nature Rev Genet 2*, 919-929 (2001)
118. Storz, G.: An expanding universe of noncoding RNAs. *Science 296*,1260-1263 (2002)
119. Bartel, D.P.: MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell 116*, 281-297 (2004).
120. Hutvágner, G. and P.D. Zamore: RNAi: nature abhors a double-strand. *Curr Opin Genet Develop 12*, 225-232 (2002)
121. Baulcombe, D.: RNA silencing in plants. *Nature 431*, 356-363 (2004)
122. Frank, D.N. and N.R. Pace: Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu Rev Biochem 67*, 153-180 (1998)
123. Eliceiri, G.L.: Small nucleolar RNAs. *Cell Mol Life Sci 56*, 22-31 9 (1999)
124. Keenan, R.J., D.M. Freymann, R.M. Stroud, and P. Walter: The signal recognition particle. *Annu Rev Biochem 70*, 755-775 (2001)
125. Washietl, S., I.L. Hofacker, M. Lukasser, A. Hüttenhofer, and P.F. Stadler: Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol 23*, 1383-1389 (2005)
126. Leontis, N.B., R.B. Altman, H.M. Berman, S.E. Brenner, D.R. Engelke, *et al*: The RNA Ontology Consortium: an open invitation to the RNA community. *Nucleic Acids Res 12,* 533-541 (2006)
127. Klosterman, P.S., M. Tamura, S.R. Holbrook, and S.E. Brenner: SCOR: a structural classification of RNA database. *Nucleic Acids Res 30*, 392-394 (2002)
128. Gan, H.H., S. Pasquali, and T. Schlick: Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res 31*, 2926-2943 (2003)
129. Zorn, J., H.H. Gan, N. Schiffeldrim, and T. Schlick: Structural motifs in ribosomal RNAs: implications for RNA design and genomics. *Biopolymers 73*, 340-347 (2004)
130. Kim, N., N. Schiffeldrim, H.H. Gan, and T. Schlick: Candidates for novel RNA topologies. *J Mol Biol 341*, 1129-1144 (2004)
131. Hermann, T. and D.J. Patel: Stitching together RNA tertiary structures. *J Mol Biol 294*, 829-849 (1999)
132. Leontis, N.B. and E. Westhof: Analysis of RNA motifs. *Curr Opin Struct Biol 13,* 300-308 (2003)
133. Pollock, D.D.: The Zuckerkandl Prize: structure and evolution. *J Mol Evol 56*, 375-376 (2003)
134. Fontana, W., D.A. Konings, P.F. Stadler, and P. Schuster: Statistics of RNA secondary structures. *Biopolymers 33*, 1389-1404 (1993)
135. Gardner, P.P., B.R. Holland, V. Moulton, M. Hendy, and D. Penny: Optimal alphabets for an RNA world. *Proc R Soc Lond B 270*, 1177-1182 (2003)
136. Gultyaev, P.A. and A. Roussis: Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants. Nucleic Acids Res 35, 3144-3152 (2007)
137. Caetano-Anollés, G.: Evolution of genome size in the grasses. *Crop Sci 45*, 1809-1816 (2005)
138. Sun, F.-J. and G. Caetano-Anollés: The evolutionary significance of the long variable arm in transfer RNA. Ms. submitted (2008)
139. Sun, F.-J. and G. Caetano-Anollés: Evolutionary patterns in the sequence and structure of transfer RNA: early origins of Archaea and viruses. *PLoS Comp Biol*, in press (2008)
140. Woese, C.R.: Translation: in retrospect and prospect. *RNA 7*, 1055-1067 (2001)
141. Cochella, L. and R. Green: An active role for tRNA in decoding beyond codon:anticodon pairing. Science 308, 1178-1180 (2005)
142. Himeno, H., T. Hasegawa, T. Ueda, K. Watanabe, and M. Shimizu: Conversion of aminoacylation specificity from $tRNA^{Tyr}$ to $tRNA^{Ser}$ *in vitro*. *Nucleic Acids Res 18,* 6815-6819 (1990)
143. Biou, V., A. Yaremchuk, M. Tukalo, and S. Cusack: The 2.9 Å crystal structure of *T. thermophilus* seryl-tRNA synthetase complexed with $tRNA^{Ser}$. *Science 263,* 1404-1410 (1994)
144. Wu, X.-Q. and H.J. Gross: The long extra arms of human $tRNA^{(Ser)Sec}$ and $tRNA^{Ser}$ function as major identity elements for serylation in an orientation-dependent, but not sequence-specific manner. *Nucleic Acids Res 21,* 5589-5594 (1993)
145. Breitschopf, K., T. Achsel, K. Busch, and H.J. Gross: Identity elements of human $tRNA^{Leu}$: structural requirements for converting human $tRNA^{Ser}$ into a leucine acceptor *in vitro*. *Nucleic Acids Res 23,* 3633-3637 (1995)

146. Soma, A., K. Uchiyama, T. Sakamoto, M. Maeda, and H. Himeno: Unique recognition style of $tRNA^{Leu}$ by *Haloferax volcanii* Leucyl-tRNA synthetase. *J Mol Biol 293,* 1029-1038 (1999)
147. Widmann, J., M. Di Giulio, M. Yarus, and R. Knight: tRNA creation by hairpin duplication. *J Mol Evol 61*, 524-535 (2005)
148. Woese, C.R., G.J. Olsen, M. Ibba, and D. Söll: Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev 64,* 202-236 (2000)
149. Schimmel P., R. Giegé, D. Moras, and S. Yokohama: An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci USA 90,* 8765-8768 (1993)
150. Maizels, N. and A.M. Weiner: Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc Natl Acad Sci USA 91*, 6729-6734 (1994)
151. Schimmel P. and L. Ribas de Pouplana: Transfer RNA: from minihelix to genetic code. *Cell 81,* 983-986 (1995)
152. Weiner, A.M. and N. Maizels: tRNA-like structures tag the 3' ends of genomic RNA molecules for replication: implications for the origin of protein synthesis. *Proc Natl Acad Sci USA 84*, 7383-7387 (1987)
153. Weiner, A.M. and N. Maizels: The genomic tag hypothesis: Modern viruses as molecular fossils of ancient strategies for genomic replication, and clues regarding the origin of protein synthesis. *Biol Bull 196*, 327-330 (1999)
154. Tanaka, T. and Y. Kikuchi: Origin of the cloverleaf shape of transfer RNAthe double-hairpin model: implication for the role of tRNA intron and the long extra loop. *Viva Origino 29,* 134-142 (2001)
155. Weiner, A.M.: SINEs and LINEs: the art of biting the hand that feeds you. *Curr Opin Cell Biol 14*, 343-350 (2002)
156. Cate, J., M. Yusupov, G. Yusupova, T. Earnest, and H. Noller: X-ray crystal structure of 70S ribosome functional complexes. *Science 285*, 2095-2104 (1999)
157. Frank, J. and R.K. Agrawal: A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature 406*, 318-322 (2000)
158. Mathews, D.W., J. Sabina, M. Zuker, and D,H, Turner: Expanded sequence dependence of thermodynamic parameters providesrobust prediction of RNA seconsdary structure. *J Mol Biol 288,* 911-940 (1999)
159. Ponting, C.P. and R.R. Russell: The natural history of protein domains. *Annu Rev Biophys Biomol Struct 31*, 45-71 (2002)
160. Todd, A.E., L.R. Marsden, J.M. Thornton and C.A. Orengo: Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol 348,* 1235-1260 (2005)
161. Todd, A.E., C.A. Orengo, and J.M. Thornton: Plasticityof enzyme active sites. *Trends Biochem Sci 27,* 419-426 (2002)
162. Gutteridge , A. and J.M. Thornton: Understanding nature's catalytic toolkit. *Trends Biochem Sci 30,* 622-629 (2005)
163. Chothia, C., J. Gough, C. Vogel, and S.A. Teichmann: Evolution of the protein repertoire. *Science 300*, 1701-1703 (2003)
164. James, L.C. and D.S. Tawfick: Conformational diversity and protein evolution. *Trends Biochem Sci 30*, 622-629 (2003)
165. Vogel, C., S.A. Teichmann, and J. Pereira-Leal: The relationship between domain duplication and recombination. *J Mol Biol 346*, 355-365 (2005)
166. Apic, G., J. Gough, and S.A. Teichmann: An insight into domain combinations. *Bioinformatics 17*, suppl. 1, S83-9 (2001)
167. Apic, G., J. Gough, and S.A. Teichmann: Domain combinations in Archaeal, Eubacterial and Eukaryotic proteomes. *J Mol Biol 310*, 311-25 (2001)
168. Bornberg-Bauer, E., F. Beausart, S.K. Kummerfeld, S.A. Teichmann, and J. Weiner: The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci 62,* 435-445 (2005)
169. Bashton, M. and C. Chothia: The geometry of domain combination in proteins. *J Mol Biol 315*, 927-39 (2002)
170. Han, J.H., S. Batey, A.A. Nickson, S.A. Teichmann, and J. Clarke: The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol 8,* 319-330 (2007)
171. Vogel, C., C. Berzuini, M. Bashton, J. Gough, and S.A. Teichmann: Supra-domains—Evolutionary units larger than single protein domains. *J Mol Biol 336,* 809–823 (2004)
172. Richardson, J.S.: The anatomy and taxonomy of protein structure. *Adv Protein Chem 34,* 167-339 (1981)
173. Murzin, A., S.E. Brenner, T. Hubbard, and C. Clothia: SCOP: a structural classification of proteins for the investigation of sequences and structures. *J Mol Biol 247*, 536-540 (1995)
174. Andreeva, A., D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia, and A.G. Murzin: SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Res 32,* D226-D229 (2004)
175. Orengo, C.A., A.D. Michie, S. Jones, D.J. Jones, M.B. Swindells, and J.M. Thornton: CATH: a hierarchic classification of protein domain structures. *Structure 5*, 1093-1108 (1997)
176. Harrison, A., F. Pearl, R. Mott, J. Thornton, and C. Orengo: Quantifying the similarities within fold space. *J Mol Biol 323*, 909-926 (2002)
177. Kunin, V., I. Cases, A.J. Enright, V. de Lorenzo, and C.A. Ouzounis: Myriads of protein families, and still counting. *Genome Biol 4*, 401 (2003)
178. Koonin, E.V., L. Aravind, and A.S. Kondrashov: The impact of comparative genomics on our understanding of evolution. *Cell 101*, 573-576 (2000)
179. Gough, J., K. Karplus, R. Hughey, and C. Chothia: Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol 31*, 903-919 (2001)
180. Efimov, A.V.: Structural trees for protein superfamilies. *Proteins 28*, 241-260 (1997)
181. Zhang, C. and S.H. Kim: A comprehensive analysis of the Greek key motifs in protein beta-barrels and beta-sandwiches. *Proteins 40*, 409-419 (2000)

182. Przytycka, T., R. Aurora, and G.D. Rose: A protein taxonomy based on secondary structure. *Nature Struct Biol 6*, 672-682 (1999)
183. Hou, J., G.E. Sims, C. Zhang, S.-H. Kim: A global representation of the protein fold space. *Proc Natl Acad Sci USA 100*, 2386-2390 (2003)
184. Dokholyan, N.V., B. Shakhnovich, and E.I. Shakhnovich: Expanding protein universe and its origin from the biological Big Bang. *Proc Natl Acad Sci USA 99*, 14132-14136 (2002)
185. Shakhnovich, B.E.: Improving the precision of the structure-function relationship by considering phylogenetic context. *PLoS Comp Biol 1*, e9 (2005)
186. Taylor, W.R.: A 'periodic table' for protein structures. *Nature 416*, 657-660 (2002)
187. Taylor, W.R.: Evolutionary transitions in protein fold space. *Curr Opin Struct Biol 17,* 354-361 (2007).
188. Caetano-Anollés, G. and D. Caetano-Anollés: Universal sharing patterns in proteomes and evolution of protein fold architecture and life. *J Mol Evol 60*, 484-498 (2005)
189. Fukami-Kobayashi, K., Y. Minezaki, Y. Tateno, and K. Nishikawa: A tree of life based on protein domain organizations. *Mol Biol Evol 24,* 1181-1189 (2007).
190. Cavalier-Smith, T.: Only six kingdoms of life. *Proc Biol Sci 271,* 1251-1262 (2004)
191. Gough, J.: Convergent evolution of domain architectures (is rare). *Bioinformatics 21*, 1464-1471 (2005)
192. White, S.H.: Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure. *Annu Rev Biophys Biomol Struct 23*, 407-439 (1994)
193. Winstanley, H,F, S. Abeln, and C.M. Deane: How old is your fold? *Bioinformatics 21*, i449-i458 (2005)
194. Murzin, A.: How far divergent evolution goes in proteins. *Curr Opin Struct Biol 8*, 380-387 (1998)
195. Grishin, N.V.: Fold change in evolution of protein structures. *J Struct Biol 134*,167-185 (2001)
196. Pagel, M., C. Venditti, and A. Meade: Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science 314,* 119-121 (2006)
197. Caetano-Anollés, G., H.S. Kim, and J.E. Mittenthal: The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci USA 104*, 9358-9363 (2007)
198. Kacser, H. and R. Beeby: On the origin of enzyme species by means of natural selection. *J Mol Evol 20*, 38-51 (1984)
199. Ji, H.-F., D.-X Kong, L. Shen, L.-L. Chen, B.-G. Ma, and H.-Y. Zhang: Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol 8,* R176 (2007)
200. Dupont, C.L., S. Yang, B. Palenik, and P.E. Bourne: Modern proteomes contain putative imprints of of ancient shifts in trace metal geochemistry. *Proc Natl Acad Sci USA 103*, 17822-17827 (2006)
201. Danchin, A., G. Fang, and S. Noria: The extant core bacterial proteome is an archive of the origin of life. Proteomics 7, 875-889 (2007).
202. Barabási, A.L. and Z.N. Oltvai: Network biology: understanding the cell's functional organization. *Nature Rev 5,* 101-113 (2004)
203. Rives, A.W. and T. Galitski: Modular organization of cellular networks. *Proc Natl Acad Sci USA 100,* 1128–1133 (2003)
204. Pfeiffer, T., O.S. Soyer, S. Bonhoeffer: The evolution of connectivity in metabolic networks. *PLoS Biology 3,* 1269-1275 (2005)
205. Peregrin-Alves, J., S. Tsoka, and C.A. Ouzounis: The phylogenetic extnt of metabolic enzymes in pathways. *Genome Res 13,* 422-427 (2003)
206. Huynen, M.A., T. dadenkar, and P. Bork: variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol 7,* 281-291 (1999)
207. Maden, B.E.H.: No soup for starters? Autotrophy and the origins of metabolism. *Trends Biochem Sci 20,* 337-341 (1995)
208. Orgel, L.E.: Self-organizing biochemical cycles. *Proc Natl Acad Sci USA 97,* 12503-12507 (2000)
209. Orgel, L.E.: Some consequences of the RNA world hypothesis. *Orig Life Evol Biosphere 33,* 211-218 (2003)
210. Lazcano, A. and S.L. Miller: On the origin of metabolic pathways. *J Mol Evol 49,* 424-431 (1999)
211. Schmidt, S., S. Sunyaev, P. Bork, and T. Dandekar: Metabolites: a helping hand for pathway evolution? *Trends Biochem Sci 28*, 336-341 (2003)
212. Horowitz, N.H.: On the evolution of biochemical syntheses. *Proc Natl Acad Sci USA 31,* 153-157 (1945)
213. Cordón, F. : *Tratado Evolucionista de Biologia*, Aguilar, Madrid, Spain (1990)
214. Jensen, R.A.: Enzyme recruitment in evolution of new function. *Ann Rev Microbiol 30,* 409-425 (1976)
215. Copley, R.R. and P. Bork: Homology among $(ba)_8$ barrels: Implications for the evolution of metabolic pathways. *J Mol Biol 303,* 627-41 (2000)
216. Teichmann, S.A., S.C.G. Rison, J.M. Thornton, M. Riley, J. Gough, and C. Chothia: Small-molecule metabolism: an enzyme mosaic. *Trends Biotech 19*, 482-486 (2001)
217. Teichmann, S.A., S.C.G. Rison, J.M. Thornton, M. Riley, J. Gough, and C. Chothia: The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J Mol Biol 311*, 693-708 (2001)
218. Kim, H.S., J. Mittenthal, and G. Caetano-Anollés: MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics 7*, 351 (2006)
219. Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa: KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acid Res 27*, 29-34 (1999)
220. Ma, H. and A.-P. Zeng: Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics 19,* 270-277 (2003)
221. Alves, R., R.A.G. Chaleil, and M.J.E. Sternberg: Evolution of enzymes in metabolism: A network perspective. *J Mol Biol 320,* 751-770 (2002)
222. Light, S., and P. Kraulis: Network analysis of metabolic enzyme evolution in *Escherichia coli*. *BMC Bioinformatics 5,* 15 (2004)

223. Lazcano, A. and S.L. Miller: The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell 85,* 793-798 (1996)
224. Kurland, C.G., B. Canbäck, and O.G. Berg: The origins of modern proteomes. *Biochimie 89,* 1454-1463 (2007)
225. Poole, A. D.C. Jeffares, and D. Penny: The path from the RNA world. *J Mol Evol 46*, 1-17 (1998)
226. Kurland, C.G., L.J. Collins, and D. Penny: Genomics and the irreducible nature of eukaryote cells. *Science 312*, 1011-1014 (2006)
227. Woese, C.R.: The universal ancestor. *Proc Natl Acad Sci USA 95*, 6854-6859 (2000)
228. Ouzounis, C.A., V. Kunin, N. Darzentas, and L. Goldovsky: A minimal estimate for the gene content of the lst universal common ancestor – exobiology from a terrestrial perspective. *Res Microbiol 157,* 57-68 (2006)
229. Berezovsky, I.N. and E.I. Shakhnovich: Physics and evolution of thermophylic adaptation. *Proc Natl Acad Sci USA 102,* 12742-12747 (2005)
230. Penny, D. and A. Poole: The nature of the last universal common ancestor. *Curr Opin Genet Dev 9,* 672-677 (1999)
231. Bamford, D.H.: Do viruses form lineages across different domains of life? *Res Microbiol 154*, 231-236 (2003)
232. Daubin, V. and H. Ochman: Start-up entities in the evolution of new genes. *Curr Opin Genet Devel* 14, 616-619 (2004)
233. Forterre, P.: The origin of DNA genomes and DNA replication proteins. *Curr Opin Microbiol 5,* 525-532 (2002)
234. Forterre, P., J. Filé, and H. Myllykallio: Origin and evolution of DNA and DNA replication machineries. In: The Genetic Code and the Origin of Life. Ed: Ribas de Pouplana L, *Springer,* NY (2004)
235. Forterre, P.: The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie 87,* 793-803 (2005)
236. Forterre, P.: Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA 103*, 3669-3674 (2006)
237. Prangishvili, D., P. Forterre, and R. A. Garrett: Viruses of the Archaea: a unifying view. *Nature Rev Microbiol 4,* 837-848 (2006)
238. Wächtershäuser, G.: On the chemistry and evolution of the pioneer organism. *Chem Biodiversity 4,* 584-602 (2007)
239. Carson, M.: Ribbons. *Meth Enzymol 277*, 493-505 (1997)

Key Words: Archaea, Evolution, Metabolism, Proteins, Proteome, RNA, tRNA, rRNA, Structure, Review

Send correspondence to: Dr. Gustavo Caetano-Anollés, Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA, Tel: 217-333-8172, Fax: 217-333-8046, E-mail: gca@uiuc.edu