J Mol Evol (2011) 72:14–33
DOI 10.1007/s00239-010-9400-9

Proteome Evolution and the Metabolic Origins of Translation and Cellular Life

Derek Caetano-Anollés · Kyung Mo Kim · Jay E. Mittenthal · Gustavo Caetano-Anollés

Received: 3 March 2010 / Accepted: 25 October 2010 / Published online: 17 November 2010

Abstract The origin of life has puzzled molecular scientists for over half a century. Yet fundamental questions remain unanswered, including which came first, the metabolic machinery or the encoding nucleic acids. In this study we take a protein-centric view and explore the ancestral origins of proteins. Protein domain structures in proteomes are highly conserved and embody molecular functions and interactions that are needed for cellular and organismal processes. Here we use domain structure to study the evolution of molecular function in the protein world. Timelines describing the age and function of protein domains at fold, fold superfamily, and fold family levels of structural complexity were derived from a structural phylogenomic census in hundreds of fully sequenced genomes. These timelines unfold congruent hourglass patterns in rates of appearance of domain structures and functions, functional diversity, and hierarchical complexity, and revealed a gradual build up of protein repertoires associated with metabolism, translation and DNA, in that order. The most ancient domain architectures were hydrolase enzymes and the first translation domains had catalytic functions for the aminoacylation and the molecular switch-driven transport of RNA. Remarkably, the most ancient domains had metabolic roles, did not interact with RNA, and preceded the gradual build-up of translation. In fact, the first translation domains had also a metabolic origin and were only later followed by specialized translation machinery. Our results explain how the generation of structure in the protein world and the concurrent crystallization of translation and diversified cellular life created further opportunities for proteomic diversification.

Keywords Origin of life · Phylogenetic analysis · Protein domain structure · Ribonucleoprotein world · RNA world

Abbreviations

aRS	Aminoacyl-tRNA synthetase
F	Fold
FSF	Fold superfamily
FF	Fold family
Nd	Node distance
r-Protein	Ribosomal protein
SCOP	Structural classification of proteins

Introduction

Proteins are fundamental components of molecular machinery. They carry genetic information and their structures embody the molecular functions that are needed for cellular and organismal processes (Lesk 2001). Protein structure is hierarchical, embedding several levels of

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9400-9) contains supplementary material, which is available to authorized users.

D. Caetano-Anollés · K. M. Kim · G. Caetano-Anollés (✉)
Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois, 332 NSRC, 1101 W Peabody Drive, Urbana, IL 61801, USA
e-mail: gca@illinois.edu

D. Caetano-Anollés
School of Molecular and Cellular Biology, University of Illinois, Urbana, IL 61801, USA

J. E. Mittenthal
Department of Cell and Developmental Biology, University of Illinois, Urbana, IL 61801, USA

molecular organization that range from primary sequence and supersecondary motifs to three-dimensional (3D) chain arrangements and the formation of macromolecular complexes (Caetano-Anollés et al. 2009a). Folding of a protein is dictated by its amino acid sequence and is influenced by the formation of compact and stable structures that are capable of concealing hydrophobic residues from the aqueous environment of the cell. These compact stable units, called *domains*, are endowed with a diversity of molecular functions and interact with a wide range of molecules, including metabolic substrates, nucleic acids, proteins, and lipid bilayers, and with other domains in multidomain proteins. Interactions link cellular components to each other and enable the many complicated biological processes that operate in a living cell, including metabolism, translation, and DNA replication. In general, the 3D folding arrangements of the domains are orders of magnitude more resilient to molecular change than domain sequence in the peptide chains. Furthermore, domains have local structure that is stable, and domains that are evolutionarily related can be grouped together in hierarchical classifications (Chothia and Gough 2009). For example, the structural classification of proteins (SCOP) groups domains that are closely related at the sequence level into fold families (FFs), domains with structural and functional features that suggest a common ancestry into fold superfamilies (FSFs), and those with common fold architectural designs into folds (Fs) (Murzin et al. 1995; Andreeva et al. 2008). Note that similarities in protein folding that define F architectures do not imply a necessary ancestral relationship and could instead have arisen simply because of the physics and chemistry of the domain packing arrangements. This argument also applies to F-linked molecular functions and biological processes, as these are recruited with their structures into modern proteins through for example processes of domain combination (Wang and Caetano-Anollés 2009).

One fundamental assumption in biology is that modern proteomes are part of an unbroken chain of ancestors, most of which are still somehow embodied in modern counterparts, and that these can be traced back in time to primordial protein repertoires in emergent cells. This is reasonable since proteomes are protein repertoires in modern organisms, which can also be traced back to ancient ancestors along tree-like or network-like lineages using the tools of phylogenetic analysis (Doolittle 2005). We can also extend the assumption to molecular functions and biological processes in proteomes, as these are the direct consequence of protein structure (Kim and Caetano-Anollés 2010). The history of protein domains and proteomes has been therefore reconstructed using a census of protein domain architecture in hundreds of genomes that have been completely sequenced (e.g., Caetano-Anollés and Caetano-Anollés 2003; Yang et al. 2005; Wang and Caetano-Anollés 2006, 2009). Currently, millions of proteins in over 1,000 completely sequenced genomes have been assigned to at least one of the ~2,000 SCOP FSFs by scanning with hidden Markov models (HMMs; Gough et al. 2001). The relatively small number of the FSFs that have been identified already implies that they are more conserved than corresponding sequences and are useful to explore evolution of life and biochemistries prior to organism diversification. The evolutionary conservation and deep phylogenetic signal of FSFs has been repeatedly and empirically verified by reliable phyletic patterns in the three superkingdoms of life (Caetano-Anollés and Caetano-Anollés 2003; Yang et al. 2005; Wang and Caetano-Anollés 2006) and by the successful exploration of the origin of modern metabolic networks (Caetano-Anollés et al. 2007) or metallomes (Dupont et al. 2010). The rationale and advances of these structural phylogenomic studies have been recently reviewed (Caetano-Anollés et al. 2009a). In some of these studies, global inferences about the history of biological functions associated with protein repertoires were possible by sorting out recruitment processes (Caetano-Anollés et al. 2007, 2009b) or by focusing on annotations at lower hierarchical levels of structural classification (Wang et al. 2007; Wang and Caetano-Anollés 2009; Caetano-Anollés et al. 2009a).

The origin of life and proteins has puzzled molecular scientists for over half a century. Yet fundamental questions remain unanswered. Which came first, the metabolic machinery of the cell or the encoding nucleic acids? An ancient world in which RNA was the sole catalyst and self-replicating molecule has been a prevailing theory (Gesteland et al. 2006), boosted notably by the existence of possible RNA-based enzyme relics in translation, the system that synthesizes proteins (Ellington et al. 2009). However, most proteins do not interact with RNA and some are believed to be very ancient (Caetano-Anollés et al. 2009a), and new views are emerging of the old idea of a peptide-dominated early world (Kauffmann 1993; Egel 2009). A recent phylogenomic analysis of hundreds of thousands of terminal ontological terms of molecular functions and biological processes associated with the sequence of 38 genomes revealed the origin of modern biochemistries in functions linked to metabolism (Kim and Caetano-Anollés 2010). These results confirm phylogenomic statements based on the structure of enzymes in metabolic networks and the protein repertoires of hundreds of genomes that reveal that the protein world originated in enzymes of nucleotide metabolism harboring the P-loop containing nucleoside triphosphate (NTP) hydrolase fold (Caetano-Anollés et al. 2007, 2009a, b). Remarkably, these studies also reveal an explosive

discovery of metabolic functions, which recapitulate well-defined prebiotic shells and involved the recruitment of structures and functions. A study of physical clustering of genes in bacterial genomes also reveals that the most ancient group of genes is related to metabolism (Danchin et al. 2007). These observations are based on the analysis of genomic repertoires and support a metabolism-first and protein-first scenario for the origins of life. More importantly, the ribosome, heralded to be a ribozyme and a stronghold of the RNA world hypothesis, appears more and more dependent on ribosomal proteins (r-proteins) for its protein biosynthetic function. While r-proteins were attributed only auxiliary roles, recent biochemical studies (Maguire et al. 2005) and higher resolution structures of intact ribosomes (Voorhees et al. 2009) have shown ribosomal proteins stabilize tRNA and facilitate amino-acyl-tRNA binding in the peptidyl transferase center (PTC). These and other findings raise doubts whether the ribosome is indeed a ribozyme (Hoogstraten and Sumita 2007). Instead, its protein biosynthetic function appears to result from the coordinated activity of proteins and RNA. We have also shown tight co-evolution of r-proteins and rRNA from the start in 5S rRNA (Sun and Caetano-Anollés 2009) and in the entire ribosomal ensemble (Harish and Caetano-Anollés, submitted). The existence of metabolism-driven protein synthesis in the absence of a modern ribosome is also likely. Non-ribosomal peptide synthesis is widespread and is mediated by multimodular megaenzymes that act in the absence of templating nucleic acids, the non-ribosomal peptide synthetases (Marahiel 2009). Precursors of these ancient enzymes that catalyze peptide bond formation could have synthesized peptides early in evolution, and even earlier, random polypeptides could have become autocatalytic sets (Bagley et al. 1991), doing work, undergoing processes of prebiotic mitosis, and evolving towards better and more efficient catalytic functions (Bagley et al. 1991; Kauffmann 1993; Egel 2009). All of these recent findings make a strong case for the likelihood of a protein or a ribonucleoprotein (RNP) world and weaken the feasibility of an RNA world.

In this paper, we focus on evolution of domains defined at F, FSF, and FF levels of structural complexity and test the ancestral origins of proteins and their links to RNA. We use functional annotations of FSF and FF structure to study the evolutionary emergence of molecular functions, including functions associated to crucial processes, such as metabolism, translation, and DNA replication. The analysis uncovers specular bimodal patterns in proteome evolution that resemble hourglass patterns. These hourglasses dissect the emergence of the protein world and provide a molecular scenario for the evolution of biological functions at the beginnings of life.

Materials and Methods

Phylogenomic Analysis of Protein Architectures

We conducted a census of genomic sequence in 185 organisms (DATASET A185: 19 Archaea, 129 Bacteria, and 37 Eukarya) and 584 free-living, parasitic, and obligate parasitic organisms (DATASET A584: 46 Archaea, 397 Bacteria, and 141 Eukarya), and 420 free-living organisms (DATASET FL420: 48 Archaea, 239 Bacteria, and 133 Eukarya) assigning protein structural domains corresponding to 1,259 FSFs (out of 1,447 in SCOP 1.67), 1,453 FSFs (out of 1,539 defined by SCOP 1.69), and 2,397 FFs (out of 3,464 defined by SCOP 1.73), respectively, to protein sequences using advanced linear HMMs of structural recognition in SUPERFAMILY (Gough et al. 2001) and probability cutoffs E of 10^{-2}, 10^{-4}, and 10^{-4}, respectively. Genome sequences are scanned against an HMM library generated using the iterative Sequence Alignment and Modeling System (SAM) method. Detecting remote homology of protein structures (i.e., at F, FSF, and FF levels of structural complexity) between distant species is always challenging. However, SAM provides top performance in CASP assessments along all classes of predictions (Karplus 2009). Furthermore, an internal calibration of the accuracy of HMM SAM-T02 model prediction against Protein Data Bank (PDB) records in the ASTRAL compendium (Brenner et al. 2000) showed that structures were identified correctly in 98% of sequences analyzed (Kim et al. 2006).

We chose to define domains with SCOP since it represents a broad and conserved classification scheme hat has been used repeatedly as gold standard to benchmark structural prediction methods and describe the complexity of the protein world (Andreeva et al. 2008). SCOP partitions proteins into fewer and larger components taking into consideration both functional and evolutionary considerations (Holland et al. 2006). Thus, it is more useful for the analysis of ancient evolutionary history than other classifications. SCOP domains were identified with concise classification strings (ccs) (e.g., c.37.1.12, where c represents the protein class, 37 the F, 1 the FSF, and 12 the FF). F architectures were assigned to FSFs and FFs to FSFs using SCOP identifiers (IDs) and algorithmic implementations in SUPERFAMILY and the census was used to construct data matrices of genomic abundance of Fs, FSFs, and FFs that were coded as linearly ordered and polarized multistate phylogenetic characters. First, we counted how many times individual FSFs are assigned to each of the proteomes sampled. Here, the number of multiple occurrences of a FSF per proteome is defined as a genomic abundance value (g). We then calculated g values for all pair-wise combinations of proteomes and FSFs, and then

constructed a two-dimensional data matrix. Empirically, g values ranged from 0 to hundreds, resembling morphometic data with a large variance (Wang and Caetano-Anollés 2006; Wang et al. 2007). Because existing phylogenetic programs can digest only tens of phylogenetic character states depending on user's CPU performance, the space of g values in the matrix were reduced by the limited number of character states using a gap coding technique with the following formula (Wang and Caetano-Anollés 2006).

$$g_{ab\_\mathrm{norm}} = \mathrm{Round}[\frac{\ln(g_{ab}+1)}{\ln(g_{\max}+1)} \times 23]$$

In this equation, a and b denote a FSF and a proteome; g_{ab} describes the g value of the FSF a in the proteome b. $g_{\max}$ indicates the maximum g value in the matrix. This round function normalizes a genomic abundance value of a particular FSF in a proteome taking into account the maximum g value, and standardizes the values to a 0–23 scale (g_{ab_norm}). The 24 transformed values that represent character states are linearly ordered and encoded using an alphanumeric format of numbers 0–9 and letters A–N that are compatible with PAUP* ver. 4.0b10 (Swofford 2002). Phylogenomic trees of protein architectures were built by polarizing character states from 'N' to '0', with 'N' being the most ancient character state. Consequently, the most ancient architectures are positioned at the base of their corresponding trees. The model considers that abundance of individual architectures increases in nature depending on given time intervals from their discovery to the present time, although the extent of losses, expansions, and selective constraints can vary during evolution. Consequently, it is natural that more ancient architectures are more abundant and widely present in modern proteomes, supporting the character argumentation scheme. Universal trees of protein architectures were built from the transformed and polarized A184, A584, and FL420 matrices using the maximum parsimony (MP) method in PAUP*, with 1,000 replicates of random taxon addition, tree bisection reconnection (TBR) branch swapping, and maxtrees unrestricted. Because some of these trees are large and the search of tree space is computationally hard, we used a combined parsimony ratchet (PR) and iterative search approach to facilitate tree reconstruction (Wang and Caetano-Anollés 2009). In case of DATASET FL420, a single tree that had minimum tree length among over 300 MP trees derived from 300 ratchet iterations (10 × 30 chains) was chosen as the best one. Multiple chains and iterations avoid the risk of optimal trees being trapped by sub-optimal regions of tree space (Nixon 1999). DATASET A184 has been used previously in functional annotation efforts (Wang et al. 2007) and is here considered the reference set. The trees were rooted by the Lundberg method, which does not require the need of outgroup taxa. Phylogenetic reliability was evaluated by the nonparametric bootstrap method with 1,000 replicates, with resampling size being the same as the number of the genomes sampled, TBR, and maxtrees unrestricted. The structure of phylogenetic signal in the data was tested by the skewness (g_1) of the length distribution of 1,000 random trees (Hillis and Huelsenbeck 1992). Finally, the relative age of protein architectures (node distance, nd) was calculated directly for each phylogeny using a PERL script that counts the number of nodes from the ancestral architecture at the base of the tree to each leaf and provides it in a relative 0–1 scale. Since trees are highly unbalanced, nd values can quickly 'date' a domain at each level of structural classification (Wang and Caetano-Anollés 2009) and can be linearly proportional to time when trees are calibrated with geological evidence (Wang et al. 2010). A recent review summarizes the general approach and the progression of census data and tree reconstruction in recent years (Caetano-Anollés et al. 2009a).

Phylogenomic Analyses of Proteomes

In order to evaluate the extent of non-vertical inheritance of protein architectures (e.g., horizontal gene transfer, domain recruitment, convergent evolution) at low levels of structural organization, we reconstructed phylogenetic trees of proteomes in which FFs were considered characters. The matrix of DATASET FL420 was transposed and used to reconstruct unrooted trees describing the evolution of the proteomes of the 420 free-living organisms we analyzed using the MP method in PAUP*. Based on the MP tree obtained, homoplasy indexes for individual FF characters were calculated using the 'DIAG' option.

Annotations of Molecular Functions

Biological functions linked to FSFs were annotated using Vogel's hierarchical classification of molecular functions in SUPERFAMILY (Vogel 2005; Vogel and Chothia 2006) (retrieved January 2007). This classification assigns seven general functional categories and 50 subcategories to SCOP IDs based on information in SCOP, Interpro, Pfam, Swiss-Prot, and literature sources. Functions related to 'small molecule binding' were dissected using MANET (Kim et al. 2006). Domain architectures associated with PDB entries were queried in the PDB database (http://www.rcsb.org/pdb/home/) and annotated using Gene Ontology (GO) terms of molecular function (Ashburner et al. 2000). GO terms define a vocabulary of molecular functions, biological processes, and cellular component, establishing a hierarchical structure embedded in a directed acyclic graph (DAG)

that connects child nodes to one or more parent node terms. In particular, we examined GO terms linked to molecular functions in six biological processes: translation, DNA replication, and DNA recombination, and processes associated with proteasome complexes, nuclear pores, and spliceosomes. Manual annotations also involved queries in the UniProtKB (PROTEIN KNOWLEDGEBASE) database (http://www.uniprot.org/) and HMM-based structure assignments. Annotations were mapped onto the architectural chronology, generating a timeline that describes the evolution of biological functions. We also mapped the distribution of domain architectures in proteomes (distribution index, $f =$ number of proteomes that have a domain/total number of proteomes) and identified domains that were uniquely present in one or more than one of the three superkingdoms of life. We also mapped the distribution of FSFs in Fs as a measure of hierarchical structure (Coulson and Moult 2002) in the timeline.

Results and Discussion

Phylogenomic Trees of Protein Domain Architectures

We used established methodology to build intrinsically rooted phylogenomic trees from the structure of protein domains (Caetano-Anollés and Caetano-Anollés 2003). Figure 1 summarizes the experimental strategy. Linear HMMs of structural recognition are first used to survey protein sequences in hundreds of genomes that have been fully sequenced in the three superkingdoms of life, identifying associated domain structures. The survey establishes the number of copies of a domain that exist in a proteome. These 'domain abundances' are used as character states when constructing data matrices, with columns and rows representing proteomes or domain architectures (the characters and taxa of the phylogenomic analysis). Matrices (and their transposed derivatives) are then used to build trees of proteomes or trees of architectures that are most parsimonious and that are intrinsically rooted. Finally, to unfold the data in the trees of architectures, we calculate the relative age of individual domains, with time measured by a relative distance in nodes from a hypothetical ancestral architecture at the base of the trees. This node distance (nd) was used to construct timelines that describe the evolution of proteins, with time flowing from the origin of modern proteins (nd = 0) to the present (nd = 1). Remarkably, nd values have been shown to be proportional to geological time when trees of domain architectures are used as molecular clocks at F and FSF levels (Wang et al. 2010).

The genomic census involves the prediction of domain structures from sequence similarities using linear HMMs

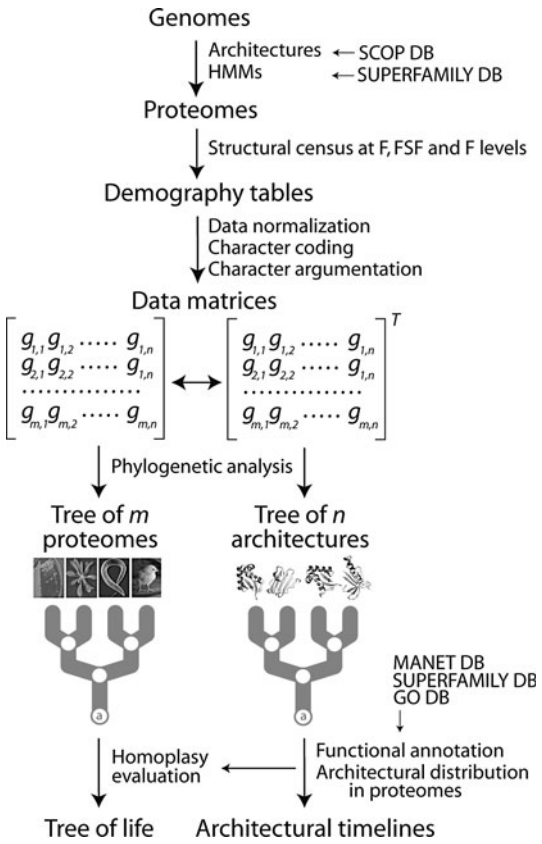


Fig. 1 Flow diagram describing the construction of phylogenomic trees. The structural census is used to compose data matrices (and their transposed derivatives), which are then used to build trees of proteomes (not described in this paper) and trees of domain architectures. Elements of the matrix (g_{mn}) represent genomic abundances of domains in proteomes, and different databases (DB) help assign function to evolving structures

methods of structural recognition. We use domains defined by the iterative SAM method, which inputs an amino acid sequence and outputs a domain structure in PDB format, together with multiple sequence alignments, local structural features, and other useful information. Protein structure can be predicted with high accuracy. In fact, the last Critical Assessment of Techniques for Structure Prediction (CASP) community-wide experiment that objectively tests the performance of prediction methods has arrived to the conclusion that comparative modeling can solve nearly all structural targets and that targets are becoming easier to predict (Tress et al. 2009). This probably results from having sampled protein structure in genomes quite exhaustively with sequence pattern profiles that correspond roughly to domains (Levitt 2009). Modern SAM versions provide top performance in CASP assessments along all classes of predictions (Karplus 2009). This guarantees prediction of domain structure from amino acid sequences with very high reliability, as long as structures are associated with experimentally verified structural models.

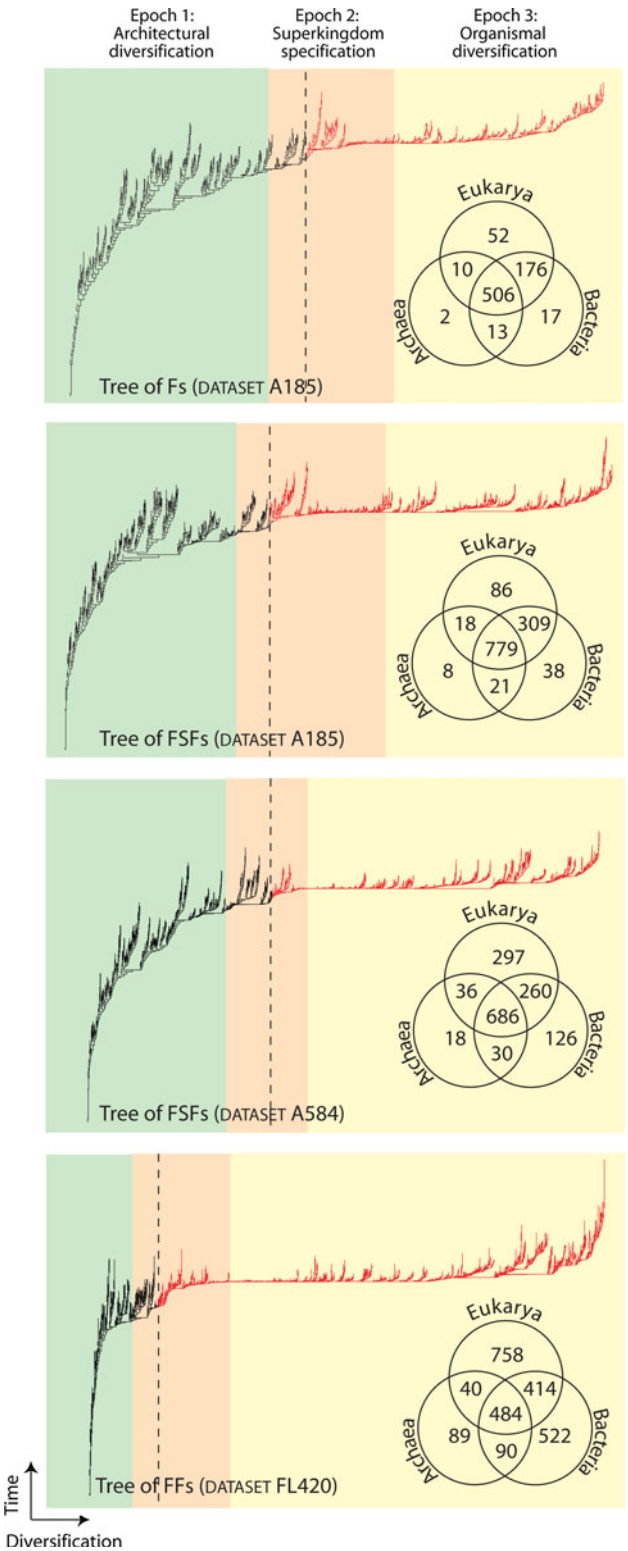
In this study we define protein domain structures with SCOP, the gold standard for protein classification (Murzin et al. 1995; Andreeva et al. 2008). We chose this scheme because it emphasizes manual curation of structural entries and preservation of functional and evolutionary relationships. Other classifications, including CATH (Greene et al. 2007), are automatic and partition domains artificially into a larger number of smaller components, splitting functional units that appear recurrently in proteins into domain segments without a rationale based on molecular evolution or function (Holland et al. 2006). While there are more similarities than differences between SCOP and CATH ($\sim$70% of domain definitions agree at an overlap threshold of 80%), inconsistencies occur preferentially at higher levels of the structural hierarchy (Csaba et al. 2009). Pairwise analysis of mappable domains in the two classifications showed that only 1.7% of FF entries in $\sim 8 \times 10^6$ comparisons or 8% of FSF entries in $\sim 9 \times 10^6$ comparisons were inconsistent. Most inconsistencies were associated with a small number of 'superfolds' that occur at structural levels higher than SCOP FSF and FF. Thus, results from phylogenetic analyses of SCOP domain definitions should not differ much from those that would be generated with CATH at the low hierarchical levels of structure we showcase in our study.

Protein structures are highly conserved and harbor strong phylogenetic signal in proteins (Caetano-Anollés et al. 2009a) and RNA (Caetano-Anollés 2002; Sun and Caetano-Anollés 2008a, b, 2009). An analysis of structure at these high levels of structural classification guarantees deep phylogenomic analysis of evolutionary relationships of proteins and proteomes. Character state changes (changes in domain abundance) occur along lineages of the tree of proteomes. Many of these evolutionary changes occur in branches at the base of the tree. These branches describe a period in which organism diversification was inexistent or was incipient. Similarly, character state changes occur along lineages of the tree of domain architectures, with changes also manifesting at the base in branches leading to domains that are universally distributed in life. These basal branches describe evolution of domains prior to organismal diversification. This indicates that phylogenetic signal spans both the world of diversified organisms and the ancient world of organisms that preceded the last universal cellular ancestor (LUCA) of diversified life, and that our method can dig deep into origins of early biochemistries, as long as protein molecules harbor enough phylogenetic signal in their structures. We note, however, that we trust modern molecular machinery carries the imprints of ancient counterparts and assume that biosynthetic infrastructure was not completely replaced in evolution. Any structural information lost in extant molecules will not contribute meaningful signal.

Here we focus on trees that describe the evolution of domains at three major levels of structural complexity, F, FSF, and FF (Fig. 2). Timelines derived from these trees define the evolutionary age of 776 Fs (DATASET A185), 1,259 FSFs (DATASET A185), 1,453 FSFs (DATASET A584), and 2,397 FFs (DATASET FL420). The global evolutionary model used to reconstruct the rooted trees rests on the central assumption that domain reuse in proteins increases with evolutionary time, with characters transforming from one state to another in pathways that are linear, directed, and polarized (Caetano-Anollés and Caetano-Anollés 2003). While simple, the model places the proteomic repertoire within a historical perspective and is validated by the reconstruction of reasonable trees of life from the same data (see Wang et al. 2007). The model supports gradual change in structural discovery, but can accommodate punctuated phenomena or selective loss (reduction) during for example speciation, duplications, or recruitment within and among species or ancestral lineages. Our experience with growing genomic and structural datasets has shown that evolutionary patterns are consistently recovered (Caetano-Anollés et al. 2009a) despite possible biases in the structural census, such as over- or under-representation of sequences and structures, definitions of fold space, and sampling limitations (Caetano-Anollés and Caetano-Anollés 2003). This enhances our confidence in phylogenetic statements. We recognize, however, that all historical work is complicated by problems with phylogenetic reconstruction. For example, the evolutionary effect that mutation, recombination, and duplication of genes on protein and nucleic acid sequences can result in differential evolutionary rates among organismal lineages, gene paralogs, and non-orthologous gene displacement, and can produce phylogenetic artifacts such as long-branch attraction and unrecognized paralogy (Philippe and Laurent 1998). These compounds with difficulties in assigning orthologous relationships among homologous sequences and saturation of protein and nucleic acid substitutions by purifying selection. Since molecular features that are highly conserved in evolution are less susceptible to some of these artifacts, a focus on structure is more appropriate than sequence when studying deep evolutionary phenomena (Caetano-Anollés and Caetano-Anollés 2003; Ranea et al. 2006).

The Effect of Horizontal Gene Transfer

At sequence level, the view that horizontal gene transfer (HGT) is rampant (Doolittle 1999) has been toned down by sound evolutionary considerations (Kurland et al. 2003). In particular, the evolutionary impact of HGT appears quite limited at higher levels of structural organization. For example, global analysis of sequence motifs in protein



◀ **Fig. 2** Universal phylogenomic trees of architectures reconstructed from a genomic census of protein domain structure at different levels of structural classification (F, FSF, and FFs) in 184 (DATASET A184), 584 (A584), and 420 (FL420) genomes. The three evolutionary epochs of the protein world are overlapped to the trees, labeled with different shades. Terminal leaves are not labeled in the tree since they would not be legible. Branches in *red* delimit the birth of architectures after the appearance of the first domain unique to a superkingdom (*dashed line*). The *Venn diagram* shows occurrence of domain architectures in the three superkingdoms of life. Note how trees at lower structural complexity (e.g., FF level) have more branches defining the organismal diversification epoch, and are less able to dissect deep phylogenetic processes. In turn, they are better fit to dissect molecular functions, diminishing the effects of recruitment on evolutionary statements

domains (Choi and Kim 2007) or domain architectures (Gough 2005; Forslund et al. 2008; Yang and Bourne 2009) reveals that HGT is relatively rare. Furthermore, analysis of entire genomic complements indicates that the assumption of massive HGT is not warranted and does not impair the phylogenetic reconstruction of a universal tree (Doolittle 2005). However, trees reconstructed at FF levels of structural complexity (e.g., from DATASET FL420; Fig. 2) are closer to sequence and could be more prone to HGT and other convergent evolutionary processes than those reconstructed at FSF and F levels. This prompted an analysis of the effect of convergent evolutionary processes on phylogenetic reconstruction.

When characters fail to fit a phylogeny perfectly they introduce conflicting phylogenetic signals. This conflict results in characters that are less consistent and have higher levels of homoplasy. Since the homoplasy index (H_i) is a good indicator of deviations from vertical inheritance (Kluge and Farris 1969), we constructed a universal tree of life derived from FF architectures and calculated H_i for the FF characters used in the analysis. We recovered a single most parsimonious tree of proteomes from DATASET FL420. In this reconstruction, 2,262 out of 2,397 FFs were parsimony-informative. The H_i of each FF character was calculated and plotted against the nd_{FF} value of the corresponding FF (Fig. 3). nd_{FF} were derived directly from the tree of FF architectures (Fig. 2). The H_i values were in average high. This is expected since H_i increases considerably with the number of taxa (Archie 1989) and our tree of life is large. We then identified that 173 out of the 2,262 FFs belong to the broad functional category 'information', using the hierarchical classification of biological functions of SUPERFAMILY, and that there were 435 bacterial-specific FFs in the set (Table S1). Remarkably, while the mean H_i value (0.78) of the 2,262 FFs was slightly larger than that of the 173 informational FFs (0.76), a one sample *t*-test failed to accept the alternative hypothesis that H_i of the informational FFs is significantly lower than the rest of FFs ($P = 0.24$). Similarly, the H_i mean of the 435 bacteria-specific FFs (0.58) was slightly smaller than that of 34 bacteria-specific informational FFs (0.59) in the set, but differences were again not significant. Since the extent of HGT is limited in informational genes that are involved in translation, transcription, and DNA replication (Jain et al.

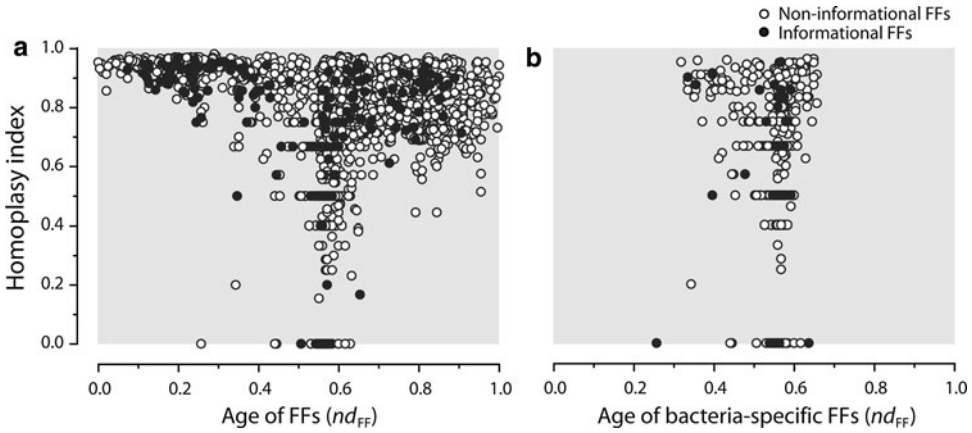


Fig. 3 Using homoplasy of phylogenomic characters to evaluate the role of horizontal gene transfer in protein evolution. To calculate the homoplasy index (H_i) of the FFs, when these are used as characters, we reconstructed a phylogenomic tree of life describing the evolution of 420 free-living proteomes (tree length = 128,371; consistency index = 0.103; retention index = 0.760). H_i values for 2,262 parsimony-informative FFs (**a**) and 435 bacteria-specific FFs (**b**) were plotted against nd_{FF} values of the FFs. These values were derived from the tree of FF architectures in Fig. 2. A total of 173 and 34 out of the 2,262 and 435 FFs, respectively, were involved in informational cellular processes such as transcription, translation, and DNA replication

1999), and since HGT levels are higher in Bacteria than in other superkingdoms, the expectation was that the H_i of informational FFs and the H_i of FFs that were not bacteria-specific had to be smaller than those of other FFs. However, our statistical analyses revealed that the H_i distributions were not significantly different (Fig. 3). These observations suggest HGT does not contribute significantly to evolution of FF architecture, over the expected contribution of other processes, including the gain and loss of domains, domain recruitment, and convergent processes induced by domain rearrangements (discussed in Wang et al. 2007). We therefore conclude that the impact of HGT in FF evolution is limited and is probably comparable to that of FSF and F (Gough 2005; Forslund et al. 2008; Yang and Bourne 2009).

Hourglass Patterns in Proteome Evolution

Timelines of protein evolution revealed consistent evolutionary patterns, regardless of the dataset that was used. We illustrate these patterns with trees of FSFs derived from DATASET A184 (Fig. 4), which is the best annotated of the three datasets used in this study (Wang et al. 2007; Caetano-Anollés et al. 2009a). The distribution of domains among superkingdoms of life was remarkable, with ancient domains being universally present in all organisms. With time, domains were first lost in primordial archaeal lineages (EB FSFs) and then in eukaryal and bacterial lineages (Fig. 4a). In turn, the rather late gain of Bacteria-specific domains (B FSFs), and then, Eukarya-specific and Archaea-specific domains (E and A FSFs), signal the emergence of superkingdoms (Fig. 4a). These patterns have been observed in previous studies and highlight three evolutionary epochs (Wang et al. 2007): epoch 1, an ancient 'architectural diversification' period in which ancient molecules emerged and diversified and proteomes were highly similar to each other, with archaeal lineages reducing their complements by domain loss towards the end; epoch 2, a 'superkingdom specification' period in which molecules sorted in emerging organismal lineages and some became specific to emerging superkingdoms; and epoch 3, a late 'organismal diversification' period in which molecular lineages diversified in an increasingly diversified tripartite world and notable proteome expansions occurred in Eukarya (Fig. 4). We note that Epoch 1 largely coincides with the Archean eon (4.8 to 2.5 billions of years ago) once nd values are calibrated with a molecular clock (Wang et al. 2010).

The rate of appearance of domain structures in evolution varied in an hourglass pattern, first increasing to a peak at about $nd_{FSF} \sim 0.3$ and decreasing to a minimum at $nd_{FSF} \sim 0.5$ (epoch 1), then increasing again to a peak at $nd_{FSF} \sim 0.6$ (epoch 2) and decreasing steadily to the present (epoch 3) (Fig. 4a). A congruent hourglass was determined by counting SUPERFAMILY subcategories of biological functions associated with FSFs, which also exhibited two major peaks at $nd_{FSF} \sim 0.3$ and $nd_{FSF} \sim 0.65$ (Fig. 4a). Hourglasses were also evident when levels of FSF multifunctionality and hierarchical complexity were examined. Sharing of six FSF-linked general categories of biological processes associated with cellular structure and function through GO term assignments was maximal very early and very late in the timeline (Fig. 4b). While FSFs with one or two functions were predominant, clear peaks at $nd_{FSF} = 0.3$ and at $nd_{FSF} \approx 0.8$ defined an hourglass that depicted the

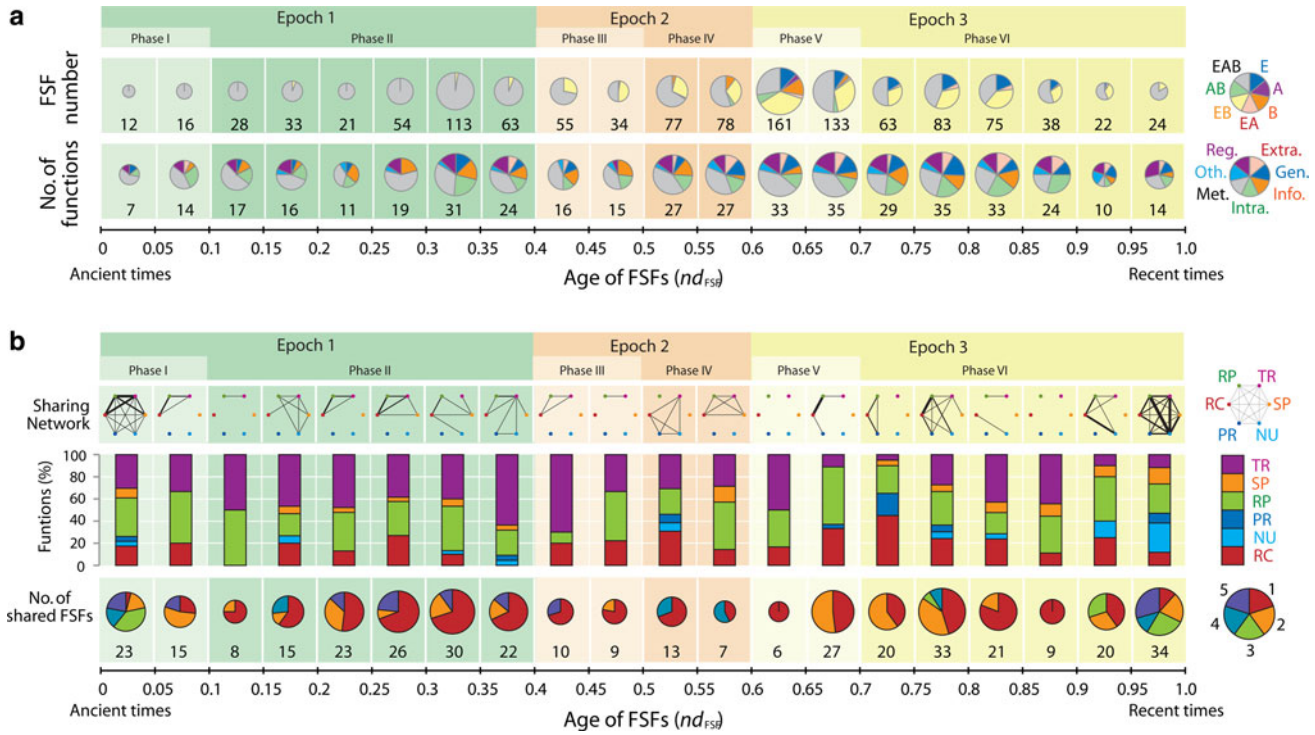


Fig. 4 Hourglasses in proteome evolution. **a** The evolutionary appearance of FSF domains, their distribution in superkingdoms of life, and discovery of associated biological functions. The number of FSFs and corresponding biological functions are plotted as function of their age (nd_{FSF}), with ages binned into 0.05 increments within a relative 0–1 scale. The different evolutionary epochs and phases of the protein world are shaded in the timeline and follow definitions by Wang et al. (2007). *Pie charts* above bins describe either the percent distribution of FSF architectures that are unique or shared between superkingdoms Archaea (*A*), Bacteria (*B*), and Eukarya (*E*), or the distribution of the 50 subcategories that are grouped into seven broad categories of biological functions in SUPERFAMILY (general, metabolism, information, regulation, intra-cellular processes, extra-cellular processes, and other), with *colors* described in the chart keys. The relative area of each pie is proportional to the number of FSFs or functional subcategories in each bin. **b** Levels of multifunctionality and recruitment in the evolutionary timeline. *Pie charts* describe the appearance of FSFs with PDB-linked GO associations to translation (TR), DNA replication (RP), DNA recombination (RC), and the spliceosome (SP), proteasome (PR), and nuclear pore (NU) complexes, and the number of FSFs that share one or more of these functions. *Bar diagrams* describe the relative representation of functions in each age bin. Note that translation, DNA replication, and DNA recombination contain many more appearances than the other three functions, and that only translation shows a clear hourglass pattern. Finally, sharing networks describe sharing patterns between the six functions, with width of edges proportional to the number of FSFs that share any given two functions in that bin. Phylogenomic reconstructions use dataset A184, the best curated in terms of function

recruitment of old structures to perform new roles. In all cases, a narrowing in the number of domains and functions at nd $\sim$ 0.5 defines the hourglass, which roughly coincides with the end of architectural diversification (epoch 1) and the start of lineage diversification (epochs 2 and 3). This is exactly the time when the first FSF unique to a superkingdom (Bacteria) appears in the timeline (e.g., $nd_{FSF} = 0.489$; d.229.1, TilS substrate-binding domain), signaling perhaps the end of the lineage of the last universal common ancestor.

The mapping of FSFs to Fs in two-dimensional plots also unfolds the evolutionary hourglass (Fig. S1). The plots compare positions (Fig. S1 A) or nd values (Fig. S1 B) of taxa in rooted trees of Fs and FSFs, with horizontal lines colored according to protein class and defining the evolutionary spread of FSFs within Fs. They reveal that both ancient and modern F domains are heterogeneous entities often composed of FSF domains of different age that are undergoing divergent evolution and harbor multiple functions. This property limits the value of Fs in dissecting the early evolution of proteins. However, Fs in domains of intermediate age ($nd_F = 0.55$–0.8) are overwhelmingly composed of single FSFs. These domain architectures probably forfeit hierarchical diversification to gain in specificity, as they are usually associated with single functional subcategories (Fig. 4). While the evolutionary relatedness of FSFs within an F needs to be proven for every F structure (only accomplished in a handful of cases), the pattern of phylogenetic relationship we uncovered shows hierarchical organization is constrained half-way in evolution of the protein world, at a time when organismal lineages were diversifying. This supports previous phylogenetic analyses and tracings of Fs along FSF timelines, which were used to propose a model of growth of the protein world (Wang et al. 2006).

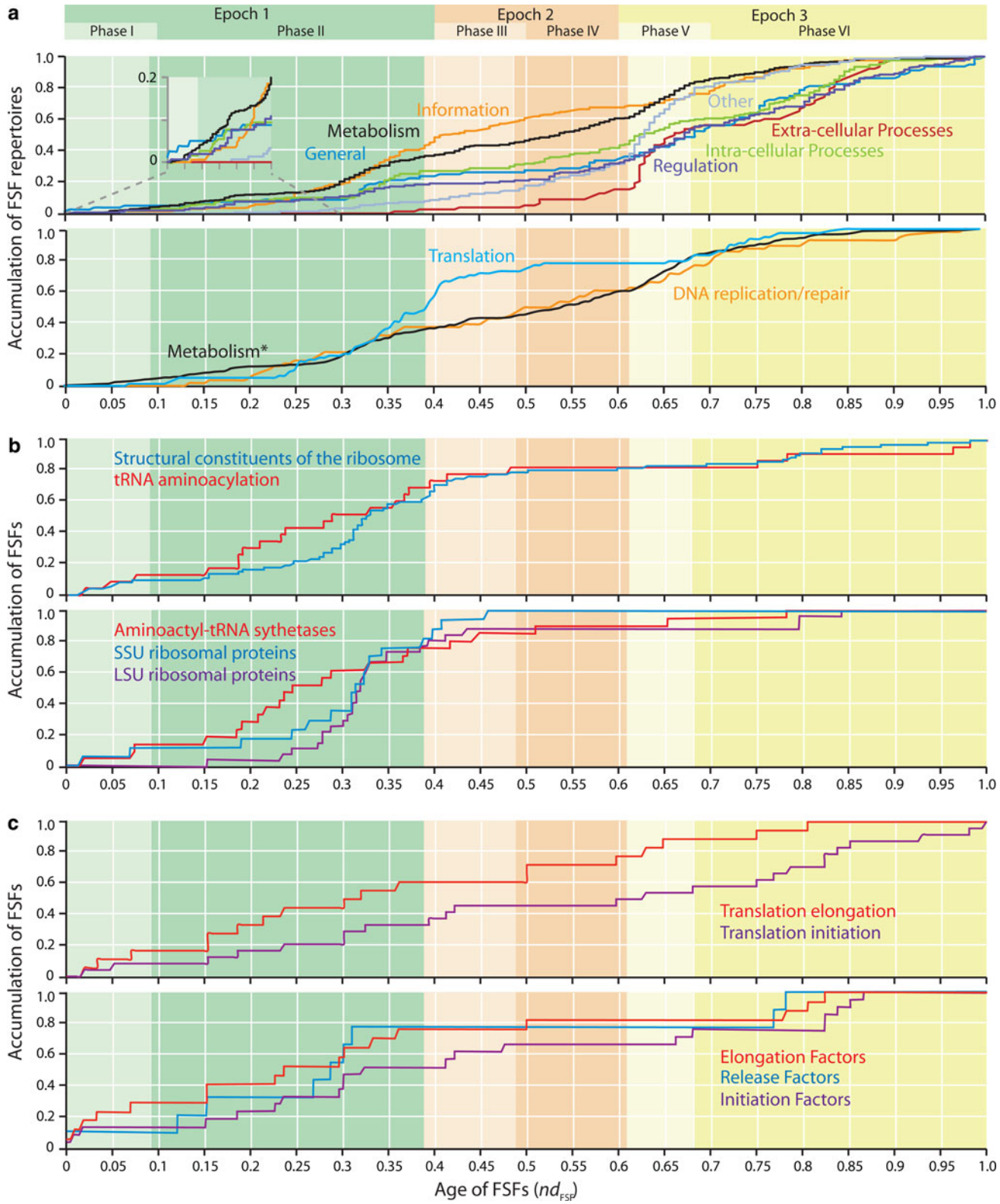
Developmental hourglasses have been interpreted in terms of linkage, the extent of interaction among processes in an embryo, and emergence of modularity in organogenesis (Raff 1996). The hourglass patterns we reveal describe the emergence and function of domain modules in protein molecules. A module can be defined as a set of parts that interact more strongly with each other than with other parts of the system (Hartwell et al. 1999). These parts can diversify and can be themselves modules, integrating into a hierarchy of modules in evolution (Caetano-Anollés et al. 2010). In proteins, domains diversify through mutation, often following gene duplication and divergence, and can sometimes change the folding pattern and give rise to both new FSFs and new functions (Caetano-Anollés et al. 2009a). Proteins that have multiple domains also diversify by adding new domains or rearranging their domain constituents (Moore et al. 2008). A recent structural phylogenomic analysis revealed that the combination of domains occurred explosively at the start of the organismal diversification epoch mediated fundamentally by fusion processes (Wang and Caetano-Anollés 2009). This coincides with the second phase of the hourglass. Here, domains become modules as they enter into a combinatorial game that enhances the functional repertoire, with molecular interactions establishing preferentially within sets of modules that perform common functions (Kummerfeld and Teichmann 2009). Remarkably, the emergence of fission processes produces evolutionarily derived multifunctional modules in Eukarya that enrich the functional repertoire of modern proteins (Wang and Caetano-Anollés 2009).

Origins of Metabolism, Translation, and DNA

In order to explore the evolution of biological functions in emerging proteomes, we traced FSF repertoires that embody functional categories and subcategories in the SUPERFAMILY classification. FSFs in the 'general' and 'metabolism' categories evolved before all others, followed closely by FSFs in 'information', 'regulation', and 'intracellular processes' (Fig. 5a). As expected, functions typical of multicellular organisms ('extracellular processes', which include cell adhesion, immune response, toxins and defense, and blood clotting) or viruses ('viral proteins' in category 'other') increased substantially late in evolution, during the organismal diversification epoch. The 'general' category includes the 'small molecule binding' subcategory, a set of functions associated with FSFs typical of early metabolism. We therefore dissected these metabolic FSFs for a better definition of the metabolic repertoire and compared the set with 'translation' and 'DNA replication and repair' (both belonging to 'information') (Fig. 5a). Clearly, metabolic functions appeared first and accumulated steadily throughout the timeline, in accord with previous observations (Caetano-Anollés et al. 2007, 2009b). Translation emerged later, with FSFs accumulating considerably late during the architectural diversification epoch and reaching a plateau at the start of superkingdom specification ($nd_{FSF} \sim 0.41$) before the first Bacteria-specific FSF (Bacteria). Finally, structures important for general DNA-based processes appeared even later and accumulated steadily, supporting the accepted view that proteins and RNA predated DNA (Freeland et al. 1999; Joyce 2002). Since the late appearance of translation is highly significant we examined the accumulation of FSFs associated with two fundamental aspects of protein biosynthesis, the aminoacylation of tRNA and the structural makeup of the ribosome (Fig. 5b). While both FSF repertoires emerge together, domains linked to tRNA aminoacylation accumulated faster than structural constituents of the ribosome. Translation also requires protein factors, including initiation factors that get protein synthesis started, elongation factors that shepherd tRNA to the ribosome and help tRNA/mRNA translocation (the growing of peptides), and release factors that recognize termination codons and end nascent polypeptide chains. FSFs associated with these important translation molecules appeared concomitantly with other components of translational machinery (Fig. 5c). However, elongation factors accumulated first, followed by release and initiation factors, in that order. This highlights the importance of tRNA shepherding and the centrality of tRNA, offering insights into the molecular biosynthetic environment at the start of the protein world.

Metabolic Origins and Late Emergence of Translation

If our protein-centric view depicts accurately the complexities of early life, ancient protein domains did not harbor translation functions. Instead, they helped fulfill metabolic roles. Dissection of first appearance of fundamental innovations in molecular machinery (evolutionary landmarks) associated with metabolism and translation was particularly revealing (Fig. 6). In the analysis, reconstruction of trees at the lowest level of structural complexity that was possible (FF) provided sufficiently deep phylogenomic signal and avoided ambiguities in functional annotations and complexities imposed by recruitment. We note that timelines of FSF and F domain architectures are also informative (Table S2). However, a focus on evolution of F architectures to dissect origin and evolution of functions, as recently proposed (Goldman et al. 2010), can be misleading. F architectures harbor many FSFs of different age, each of which can associate with different functions (Wang et al. 2006). For example, the TIM β/α-barrel (c.1), one of the most ancient folds, contains 33 FSFs



with different functions, some ancient but many very young. Analysis of evolutionary landmarks at FF level is therefore appropriate, linking structure and function unambiguously. Our timelines reveal that metabolic functions were primordial and that interactions of protein and RNA appeared relatively late in evolution (Fig. 6).

◀**Fig. 5** Cumulative frequency distribution of FSFs associated with functional categories and molecular repertoires along the evolutionary timeline. **a** Accumulation of FSFs linked to the seven broad functional categories of SUPERFAMILY (*top graph*) and linked to curated metabolic FSF repertoires and translation and DNA replication/repair subcategories (*bottom graph*). **b** Accumulation of FSFs with PDB-linked GO associations to aminoacylation of tRNA and the structural constituents of the ribosome (*top graph*) and of FSFs present in aminoacyl-tRNA synthetases and ribosomal proteins (*bottom graph*). **c** Accumulation of FSF with PDB-linked GO associations to translation elongation and initiation (*top graph*) and to elongation, release and initiation factors (*bottom graph*). Accumulation of domains is given as a relative number scaled from 0–1. Timelines are derived from dataset A184. In contrast to SUPERFAMILY-based assignments, evolutionary statements derived from PDB-linked GO annotations in certain circumstances can be affected by co-option of domain structures to perform new functions (e.g., domains linked to ancient and multifunctional superfolds, such as c.37.1.8; see Fig. 6)

(1) First Proteins Were Hydrolases

The most ancient FSFs and FFs interact with small molecules and metabolites and are hydrolases. The most ancient FFs are P-loop containing domains harboring ATPase functions that belong to the P-loop containing NTP hydrolase (c.37) F architecture, a fold that contains only one FSF (c.37.1), but has 24 FFs that are highly diverse and multifunctional, some of which appear very early in the timeline. The three most ancient c.37.1 FSF domains were the ABC transporter ATPase domain-like (c.37.1.12), the extended AAA-ATPase domain (c.27.1.20) and the tandem AAA-ATPase domain c.37.1.19). Originally, P-loop containing enzymes probably performed functions of energy interconversion, distribution (storage and recycling) of chemical energy in acid-anhydride bonds of nucleotides, and terminal production of nucleotides and cofactors (Caetano-Anollés et al. 2007), taking advantage of the high reducing potential of primitive environments (e.g., exhalations of hydrothermal vents; Wächtershäuser 2007) and retrieving energy from the emergent chemical diversity of early life. The primordial appearance of hydrolases, and in particular of ATPase functions, is also supported by a recent phylogenomic systematization of molecular functions that explored the abundance and distribution of GO terms in 38 genomes (Kim and Caetano-Anollés 2010).

(2) The Function of Metabolic Proteins Soon Diversified

Oxidoreductases, transferases, and isomerases were added to the initial metabolic repertoire (Fig. 6; Table S2), including enzymes harboring the Rossmann and the TIM β/α-barrel folds, which are structures highly popular in modern metabolic networks (Caetano-Anollés et al. 2009b). The rather quick discovery of these general metabolic functions immediately after the primordial hydrolases is also supported by the phylogenomic analysis of gene ontology (Kim and Caetano-Anollés 2010) and is consistent with the proposed 'big bang' of metabolic discovery (Caetano-Anollés et al. 2007).

(3) Domains that Interact with RNA Appeared for the First Time Much Later and Were Involved in the Aminoacylation of RNA

The first translation domains were the catalytic domains of aminoacyl-tRNA synthetases (aRSs) ($nd_{FSF} \sim 0.07$, $nd_{FF} \sim 0.02$), including Class I (c.26.1.1, $nd_{FF} = 0.020$) and Class II (d.104.1.1, $nd_{FF} = 0.024$) aRSs that are responsible for the esterification reaction that links amino acids to RNA (Schimmel 2009). These domains brought innovations in fold design with alternative α-helices and β-strands in sandwiched conformations, and the ability of proteins to interact with RNA. However, the specificity of the genetic code was established much later (starting at $nd_{FF} = 0.12$) by matching aminoacylation reactions with anticodons in cognate tRNAs. This involved a process of accretion of editing and anticodon-binding domains of aRSs that are specific to tRNA isoacceptors, which spanned fundamentally the first phase of the superkingdom specification epoch ($nd_{FF} \sim 0.2$–0.25) and ended with the accretion of the specificity domain for Trp ($nd_{FF} = 0.65$). The emergence of the genetic code was therefore protracted but coincided with the bottleneck of the hourglass pattern. The appearance of the Class I and II aRS metabolic domains occurred much earlier than the first accessory domain, the ValRS/IleRS/LeuRS editing domain (b.51.1.1; $nd_{FSF} = 0.231$, $nd_{FF} = 0.126$) (Fig. 6; Table S2). This is one of the many that decorate and enhance the specificity of this large family of mutidomain proteins (Wolf et al. 1999). It is particularly noteworthy that this ancient editing domain has proof-reading hydrolase activity that increases the overall accuracy of the two-step aminoacylation reaction of LeuRS (e.g., Seiradake et al. 2009), especially because analysis of tRNA structure has shown leucine charging was one of the most ancient aminoacylation functions (Sun and Caetano-Anollés 2008a). The timing of this event is important. It indicates aminoacylation specificity developed concurrently with the ribosomal ensemble. In fact, the accretion of aRS domains continues throughout the timeline, in parallel with those of ribosomal proteins. For example, anticodon-binding domains appear at the beginning of superkingdom specification, starting with the 'putative anticodon-binding domain of AlaRS' FF (a.203.1.1; $nd_{FSF} = 0.356$, $nd_{FF} \sim 0.200$) and the 'anticodon-binding domain of a subclass of class I aRS' FF (a.27.1.1; $nd_{FSF} = 0.153$, $nd_{FF} \sim 0.241$) that is present in LeuRS. As expected, the evolutionary assembly of aRSs multidomain enzymes occurs in piecemeal fashion from single domain proteins (e.g., Fig. 6b). These domains have structures that are widely popular in many functional

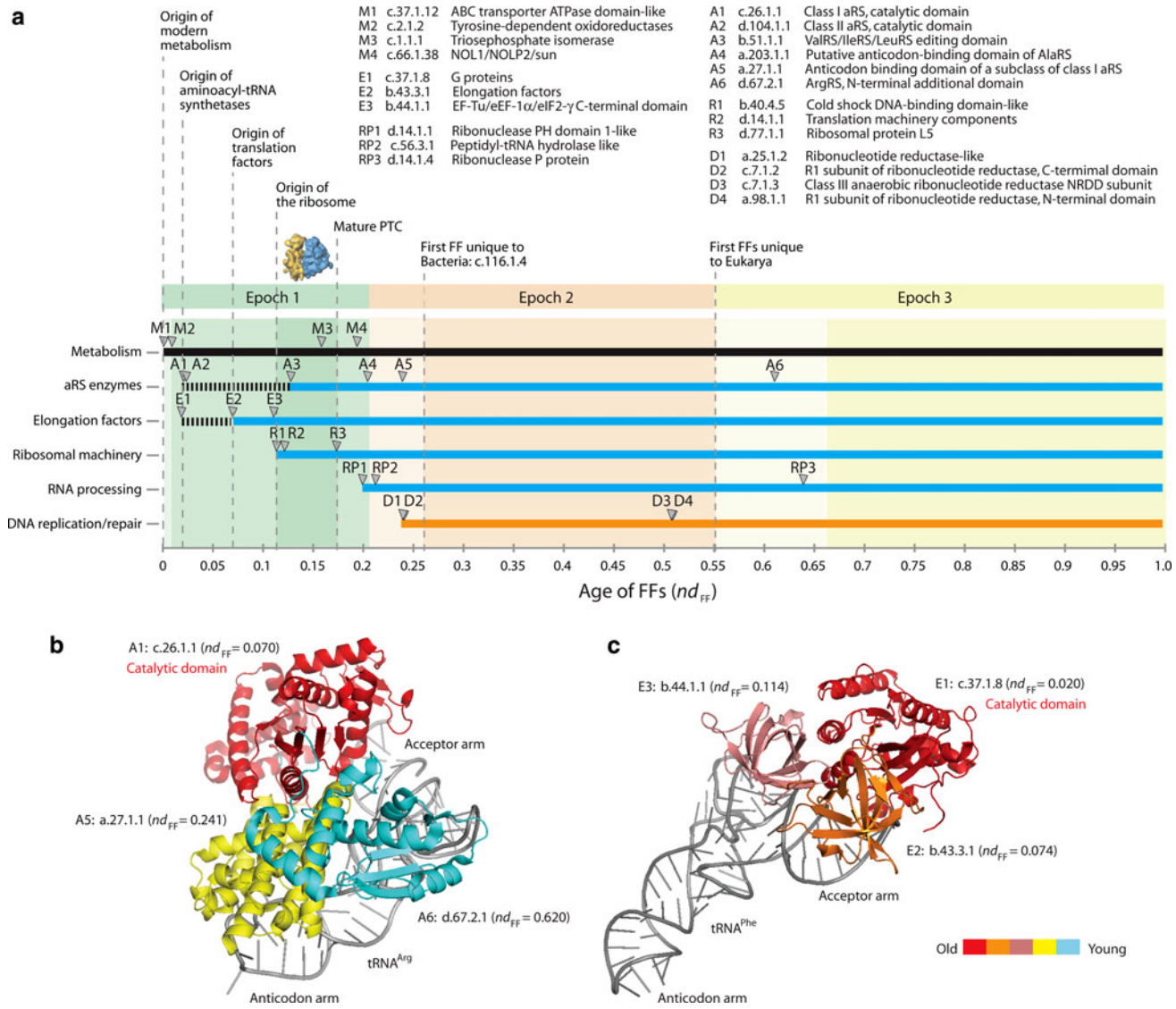


Fig. 6 Timeline of architectural landmarks in the early evolution of the protein world. **a** Landmark discoveries are identified with *arrows* in a timeline derived from a phylogenomic analysis of FF architectures (FL420). The metabolic origin of molecular functions linked to translation is indicated with *dashed black lines*. The emergence of ribonucleotide reductase enzymes responsible for producing the deoxyribonucleotide components necessary for DNA-linked functions at $nd_{FF} = 0.245$ is used as reference to show the late arrival of DNA, prior to proteins and RNA. See Table S2 for a more extended description of architectures and timelines. **b** Accretion of domains in yeast arginyl-tRNA synthetase complexed with $tRNA^{Arg}$ (PDB entry 1f7v), an example of evolution of a class Ia aminoacyl-tRNA synthetase (aRS). Note how the catalytic domain c.26.1.1, the most ancient protein domain of the ensemble (this work), makes contact with the acceptor arm of tRNA, the most ancient arm of the RNA moiety (Sun and Caetano-Anollés 2008b). Remarkably, the anticodon recognition domain is the most derived and appears during the emergence of superkingdoms in epoch 2. **c** Accretion of domains in elongation factor Tu in a ternary complex with $tRNA^{Phe}$ (1ttt), showing the catalytic GTPase domain is the most ancient but does not carry the tRNA acceptor arm binding capability. Instead, it provides the energetics necessary for the molecular switch that clamps tRNA. Protein domains are colored according to domain age (nd_{FF})

contexts. For example, the nucleotidyl transferase FSF (c.26.1; $nd_{FSF} \sim 0.07$) that defines the structure of class I aRS domains, makes up a number of metabolic enzymes, including transferases and synthetases involved in nucleotide and cofactor metabolism. Many of these enzymes preserved their single domain status and are younger ($nd_{FSF} = 0.171$–0.657). Similarly, the class II aRS and biotin synthetases FSF structures (d.104.1; $nd_{FSF} \sim 0.075$) make up enzymes involved in biotin biosynthesis. The patterns extend to other domain components of modular aRS assemblies. In fact, a number of proteins resemble aRSs, and many of these serve catalytic functions (Schimmel and Ribas de Pouplana 2000). Some of these 'evolutionary footprints' confirm the metabolic origin of

these first translation molecules, suggesting the early connection extends to enzymes important for the biosynthesis of nucleotides, amino acids, and macromolecules.

(4) Appearance of Molecular Switches and Regulatory Factors

The GTP-binding domain of elongation and initiation factors, the G protein domain (c.37.1.8; $nd_{FF} = 0.020$), originates concurrently with catalytic aRS subunits (Fig. 6). This domain is a GTPase with the ancient P-loop NTP hydrolase fold, which appears at the base of phylogenomic trees. It carries hydrolase activities that are today necessary for the formation of ternary complexes with tRNA and other proteins (exchange factors) that are required, among other things, for the loading of aminoacylated tRNA onto the ribosome (Rodnina and Wintermeyer 2009). The domain has clearly a metabolic origin but as part of translation factors acts as a conformational switch that clamps tRNA through a switch helix (Berchtold et al. 1993). However, the tRNA molecule binds to the elongation factor domain, which appears later in evolution (b.43.3.1; $nd_{FF} = 0.073$), indicating that the catalytic domain and RNA-binding domains combined to generate the primordial molecular mechanism at this later time. It is significant that the structure necessary for molecular switches that bind nucleic acids (Berchtold et al. 1993) recruited hydrolase activities from metabolism and RNA-binding abilities that were discovered much earlier with aRSs. Moreover, some factors were accessorized with further domains even later. For example, prokaryotic EF-Tu molecules have a third domain, the EF-Tu/eEF-1a/eIF2-g C-terminal domain (b.44.1.1), which appears together with the first ribosomal proteins ($nd_{FF} = 0.114$) (Fig. 6c). Some elongation and initiation factors add a fourth elongation domain to form molecules in the shape of 'chalices', such as the selenocysteine (Sec) tRNA-specific elongation factor (SelB). SelB is a structural chimera of initiation and elongation factors that delivers Sec-tRNASec to the ribosome in the presence of a *cis*-acting SECIS element in mRNA (Leibundgut et al. 2005).

(5) Discovery of Complex Ribosomal Machinery

aRSs and regulatory factors were followed by r-proteins, which appeared for the first time with the cold-shock DNA binding domain (b.40.4.5; $nd_{FF} = 0.114$). This FF harbors the OB fold, a closed or partly open β-barrel with Greek-key strand topology in the curled β-sheets. The OB fold is the first β-barrel like fold to appear in evolution (Caetano-Anollés and Caetano-Anollés 2003) and r-proteins containing this design are the first to interact with rRNA to define a more advanced molecular switch, the central ratcheting mechanism of the ribosome (Yusupov et al. 2001). These primordial r-proteins interact with helix 44 of SSU rRNA (the ribosomal functional relay), and as other r-proteins continue to accumulate in the timeline, they interact with expanding rRNA, first associated with regions important in RNA decoding, helicase, and translocation functions (translation machinery components, d.14.1.1, $nd_{FF} = 0.126$), and then to stabilize the structure and enhance the function of the growing biosynthetic machinery (A. Harish and G. Caetano-Anollés, submitted). Ribosomal proteins located in the small subunit (SSU) of the ribosome appeared earlier than those associated with the large subunit (LSU), though their accumulation became similar at an inflection point of $nd_{FSF} \sim 0.3$ (Fig. 5b). These observations are highly significant. They suggest SSU predates LSU, confirming results from an analysis of functional substructures in the ribosome (Caetano-Anollés 2002). More importantly, observations reveal r-proteins were added in concert to the two subunits throughout ribosomal history. Remarkably, a detailed phylogenetic analysis of rRNA structure and RNA-protein contacts suggests that the functional core of the ribosomal ensemble, the PTC, was added late in evolution but was already in place at $nd_{FSF} \sim 0.3$ (A. Harish and G. Caetano-Anollés, submitted), the inflection point in r-protein accumulation (Fig. 5b). The discovery of the oldest domain (r-protein L5, d.77.1.1 $nd_{FF} = 0.171$) that interacts with 5S rRNA (Sun and Caetano-Anollés 2009) signals the existence of a modern peptidyl transferase activity capable of decoding information in RNA. 5S rRNA is an integral component of the ribosome that functions as a signal transducer between the PTC and regions of LSU rRNA responsible for translocation (Bogdanov et al. 1995; Dokudovskaya et al. 1996). The placement of this molecular landmark in timelines of FSF domains [($nd_{FSF} = 0.352$ (A184); $nd_{FSF} = 0.328$ (A583)] indicates PTC-mediated translation was already operating at the end of the architectural diversification epoch (Fig. 6). In fact, most structural constituents of the ribosome were discovered when the narrowing of the protein hourglass was maximal and first Bacteria-specific domain architecture appeared in evolution [$nd_{FSF} = 0.489$ (A184), $nd_{FSF} = 0.513$ (A584); $nd_{FF} = 0.257$ (FL420)]. Patterns in the protein hourglass (Fig. 6) follow closely the emergence of the ribosomal ensemble (Fig. 6). Enhanced protein synthesis first increases protein diversity and then locks in change in the protein world. As translation components accumulate in the timeline, aRSs, regulatory factors, and ribosomal proteins were accessorized with new domains (see below), which establish new and increasing interactions. Domains become more and more modular in evolution and the long-lasting stability of the translation machinery consolidates the universal genetic code, crystallizing both the emergence of

a modular world of proteins (Wang and Caetano-Anollés 2009) and the rise of cellular lineages (Vetsigian et al. 2006). Increases in the number of interactions and the rise of modularity ultimately enhance recruitment processes. For example, ancient FSFs with translation functions are co-opted by modern processes, first by replication and recombination, and then by the spliceosome, proteasome, and the nuclear pore (Fig. 4b). This completes the hourglass pattern.

(6) Metabolic-Based Enhancements of the Ribosomal Machinery

While r-proteins, which appear later than regulatory factors in the timeline (Fig. 5), lack any apparent connection to metabolic enzymes, the interaction of elongation factor G (EF-G) with the ribosome as it induces translocation imposes a metabolic role on the entire RNP ensemble (Moore 2005) and increases the efficiency of protein synthesis more than 50-fold (Rodnina and Wintermeyer 2009). Remarkably, ribosomal interactions with EF-G appeared concurrently with 5S rRNA (A. Harish and G. Caetano-Anollés, submitted). EF-G assumes three different conformations during protein synthesis, first during preferential GDP binding in the cytosol, second by gaining a intermediary configuration when exchanging GDP by GTP during translocation, and third by hydrolyzing GTP and driving translocation to completion (Zavialov et al. 2005). Induced by EF-G, the ribosome acts effectively as a guanine nucleotide exchange factor. The EF-G-enhanced ribosome therefore represents an advanced molecular switch rather than a molecular motor (a device that uses energy from hydrolysis to gain movement). Another clue of the early metabolic origin of the ribosome is the important role of GTPases in ribosomal assembly (Britton 2009). Molecules such as the RgbA GTPase, the ObgE GTPase, and the *E. coli* ras (Era) are important for subunit assembly and other functions (e.g., stress response, cell wall metabolism, cell cycle). They carry the ancient signature of the P-loop NTP hydrolase fold and G protein FSF ($nd_{FF} = 0.020$), suggesting these ancient hydrolases were recruited for the assembly of the ribosomal machine at $nd_{FF} \sim 0.1$–0.2, which coincides with the emergence of the multidomain structure of elongation and initiation factors. Some of these GTPases participate in a guanine nucleotide-dependent interaction with the ribosome (e.g., Daigle and Brown 2004) that resembles the guanine nucleotide exchange properties of elongation factors.

(7) Late Emergence of RNA Processing

Finally, domains involved in RNA processing originated quite late in architectural diversification, starting with those that process tRNA (e.g., RNase PH proteins, $nd_{FF} = 0.200$; peptidyl-tRNA hydrolases, $nd_{FF} = 0.212$). RNase PH is an exoribonuclease, uses inorganic phosphate as cofactor (it is a phosphorolytic enzyme), and catalyzes the 3′ end processing of tRNA in Bacteria and Archaea (Deutscher 1984). Within tRNA processing proteins, catalytic functions continue to evolve along the timeline and are clearly derived. For example, RNase P is an important endonuclease that cleaves precursor tRNA and generally consists of a catalytic RNA subunit and one or more proteins (Altman 2009). Since the enzyme is a ribozyme, its origins are important and have been explored at RNA and protein levels (Sun and Caetano-Anollés 2010). The late appearance of RNase P protein domains ($nd_{FF} = 0.645$) suggests tRNA processing is a derived feature. However, the recent discovery of RNase P enzymes in organelles of Eukarya that do not require RNA cofactors (Holzmann et al. 2008) and carry the NAD(P)-binding Rossmann fold (c.2.1; $nd_{FSF} = 0.017$), suggests the origin of ribonuclease proteins is very ancient and that the catalytic RNA moiety was added quite late in evolution.

Translation and the Universal Cellular Ancestor of Life

In protein evolution, a number of innovations would have been particularly beneficial to the communal and emerging cellular entities that harbored the early proteomes. These innovations facilitated protein availability to primordial cells, enabled biocatalytic processes leading to their formation, and later on, allowed more reliable biosynthetic processes of primordial translation. We reveal in our timelines the gradual buildup of these innovations, especially prior to the narrowing of the hourglass during superkingdom specification (Fig. 6). The connection between the discovery of early biochemistries and the distribution of protein domains in life, the epochs of the protein world, and the emergence of LUCA is of significance. Autocatalysis seems a general feature of crystallization processes of many kinds, including the rise of prebiotic chemistries on Earth (Wächtershäuser 1990, 2007; Morowitz 1999; Orgel 2000). A popular model of emergence and autocatalysis posits that early cells were communities that evolved coordinately by sharing components through horizontal transfer (Woese 1998, 2002). These cells were highly dependent on the environment and were the subject of stochastic processes of change (mutation). While evolution was not Darwinian in the standard sense, sharing of components that improved function was of benefit, as these provided a selective advantage for the survival of the entire community. In particular, strategies that facilitated sharing of information, such as the genetic code, would have been preferentially selected (Vetsigian et al. 2006). These strategies promote faster spread of innovations. Under this

scenario, a sharing of vocabularies (information) present in the primordial components (a primordial genetic code) facilitates the actual sharing of components, and later on, the horizontal transfer of encoded proteins, biosynthetic machinery, and translation processes. Increased sharing also locks in useful interactions between molecules, increasing fidelity and the accuracy of emerging biological processes. This fosters growth of larger communities and accelerates the use of an emerging code. One fundamental outcome of the autocatalytic growth of primordial cellular communities is the increase of complexity and specificity of the interactions, which diminish change and lateral transfer at the expense of vertical inheritance and parallel cellular evolution. This drives early diversification further and further towards a bottleneck. The process resembles crystallization and can be used to explain the narrowing in our hourglass (Fig. 4), which coincides with the emergence of lineages in the superkingdom specification epoch, and the accretion of domains in evolution that enhance molecular specificity (Fig. 6).

Early Protein Evolution and the Principle of Continuity

Taken at face value, the results of our protein-centric phylogenomic analysis suggest proteins existed before RNPs. This questions the validity of an ancient protein-free RNA world, the common dogma that currently dominates origins of life research (Gesteland et al. 2006). The RNA world hypothesis has been also questioned by several important lines of evidence: (1) the poor catalytic performance of protein-free ribozymes that have been synthesized in vitro and are claimed to be doppelgängers of early RNA molecules, (2) the absence of naturally occurring protein-free ribozymes, (3) the absence of RNP mediating central metabolic reactions, and (4) the emergence of a complex protein-encoding apparatus (translation) in the absence of selective pressures that would favor the origination of proteins. For example, the crucial claim that the replicating center of the ribosomal machinery involves only RNA has been recently compromised by the discovery that two ribosomal proteins (L16 and L27) interact with tRNA in the PTC of the large ribosomal subunit (Maguire et al. 2005; Voorhees et al. 2009). Similarly, the recent finding that human mitochondrial RNAse P cleaves its substrate in the absence of an RNA subunit (Holzmann et al. 2008) questions this putative RNA fossil. These many lines of evidence challenge the principle of continuity that is needed to explain the transition of an ancient RNA world into a modern world of proteins and RNA.

In contrast, it is easier to explain how a protein world transitioned into the modern RNP world. The early origins of metabolic proteins explain why they are superior catalysts: change at sequence level directly impinges on their structure and on their functions (Schuster 2010). It also explains why protein enzymes drive central metabolic networks: proteins appear to be the first macromolecular catalysts of prebiotic reactions (Caetano-Anollés et al. 2009b). The gradual emergence of an increasingly complex protein biosynthetic apparatus can be better understood if the RNP machinery is built around proteins: proteins optimize conformations (Schuster 2010), recruit and rearrange domains (Moore et al. 2008; Wang and Caetano-Anollés 2009), and coevolve interactions with RNA (Sun and Caetano-Anollés 2009) that favor more efficient biosynthetic processes. The early origins of metabolic proteins also explain why protein enzymes (and not ribozymes) replaced the prebiotic chemical reactions: assembled around a self-organized citric acid cycle (Wächtershäuser 1990; Morowitz 1999), more efficient synthesis of organic compounds (e.g., amino acids) from cycle intermediates benefit the production of components needed to make proteins and cofactors (Caetano-Anollés et al. 2007; Danchin et al. 2007). This supports a metabolism-first versus a replication-first scenario for prebiotic origins (Trefil et al. 2009).

We note, however, that the relationship of ancient proteins supporting prebiotic chemistries and modern proteins is not known. Similarly, we do not know how structures of these prebiotic entities were 'inherited' in the primordial system or when and how a 'molecular container' could have arisen to accelerate primordial natural selection processes. Our phylogenetic extrapolations, though powerful, still use information in modern molecules and functions to infer the past. The question of the origin of life is therefore 'hard' and refractory to historical analysis. So the crucial question remains: How were protein catalysts replicated before the existence of nucleic acids? We speculate ancient proteins were probably polypeptides composed of few amino acid monomers that were more or less random and folded into few common structures. The abiogenic formation of amino acids monomers and peptides that are catalytic is relatively easy. For example, proteinoid microspheres can be easily produced by heating and desiccation (Fox 1980) and are currently used in nanotechnology applications as molecular containers for drug delivery. Since random mixtures of polypeptides of less than 50 amino acids exhibit a wide spectrum of weak catalytic activities, one could envision functions would emerge (crystallize) very early in protein evolution [reviewed in Kauffman (1993)]. Kauffmann (1986, 1993, 2007) suggests that as the complexity of the mix of polypeptide increases, a threshold is reached over which formation of peptides is catalyzed by some member of the set and that the process is autocatalytic. In sequence space, simulations show polypeptides seem to gain fold conformations in few generations and that selection for correct

local folding configurations in polypeptides leads in a few generations to folding ability, helicity, and compactness (Yomo et al. 1999). Consequently, short random polypeptides, perhaps composed of a limited set of amino acid monomers (Kauffmann 1993), could have quickly gained structural properties that were homogeneously advantageous for early environments, for example helping minerals and other catalysts promote the synthesis of organic molecules in autocatalytic cycles (Wächtershäuser 2007; Huber and Wächtershäuser 2007), and could have predated a modern genetic code (Ikehara 2009). The recent finding that prions, infectious proteins made of β-sheet rich conformers were capable of accumulating mutations and adapting to new tissues suggests interacting proteins can evolve in systems (Li et al. 2010). Despite lacking nucleic acids, prions showed the hallmarks of Darwinian evolution. They revealed both heritable changes in their phenotypes and produced distinct populations in different environments through selective amplification. The prion-like domain structure (d.6.1.1) of prions is clearly derived ($nd_{FF} = 0.641$; $nd_{FSF} = 0.605$; $nd_{F} = 0.688$). However, the properties of these structures uncover the ability of polypeptide chains to change in adaptive manner. Proteins per se can therefore fulfill the principle of continuity needed at the onset of life.

Finally, an RNA world in which small RNA molecules were the only encoded catalysts is less parsimonious; every putative ribozyme has to be replaced by protein derivatives that are catalytically superior (Jeffares et al. 1998). With the exception of few relatively large RNA molecules, all other ribozymes have seemingly vanished. Invoking an extinct world is problematic. Molecular fossil evidence must show unequivocally that ribozymes act in the absence of proteins, in vitro evolution experiments must replace ribozymic precursors with proteins, and the reasons why modern proteins have not preserved interactions with small RNA must be explained in light of pervasive interactions with nucleotides and ribotide cofactors. Clearly, nucleic acids are superior in their ability to encode biological information and act as genetic repository, but these traits could be derived. The late appearance of proteins that associate with RNA in our timeline suggest a more parsimonious scenario that also fulfills the principle of continuity, the gradual improvement of interactions between proteins, nucleotides and nucleic acids, first as substrates, then as docking-guides and cofactors, and finally as molecular switches and actuators.

Conclusions

Here we show the existence of hourglass patterns in proteome evolution, some of which are linked to translation and the emergence of diversified life, and we suggest they are associated with autocatalytic processes and the inevitable evolutionary increase of molecular interactions. We also reveal the gradual build-up of protein repertoires linked to metabolism, translation, and DNA. The progression of domain discovery in the timelines suggests metabolism preceded translation, and proteins and RNA preceded DNA. The most ancient domain architectures were hydrolases. The domain structures of these enzymes bind small organic cofactors, especially nucleotide derivatives that transport energy (e.g., ATP) or act as reducing agents (e.g., NADPH) (Ji et al. 2007). They do not bind or harbor RNA. Consequently, RNA biopolymers in their current form were either absent during the emergence of proteomes or were incapable of interacting with the primordial enzymes. The late and progressive appearance of domains that interact with RNA and the fact that modern metabolism lacks natural ribozyme relics support either of these two scenarios. Timelines also indicate a clear progression of enzymes that starts with catalysts and allosteric regulators (regulation by effector molecules) and continues with molecular switches (molecules that can be reversibly shifted from one stable configuration to another) and more complex molecular machinery. Ribozyme moieties such as the ribosomal PTC or RNase P RNA appear very late in the timeline and are clearly derived. The progression has two important players, aRSs and regulatory factors, both of which interact with tRNA and are crucial for the mechanics and specificity of modern translation. The discovery of further translation machinery appears to revolve around these classes of molecules. Remarkably, physical clustering of evolutionarily conserved genes that are shared by a clique of bacterial genomes defines three concentric rings of ancient proteins (Danchin et al. 2007). The discontinuous and loosely connected outer ring of gene neighbors, the most ancient, is mostly devoted to metabolism. This ring is made up of proteins known to be very ancient (e.g., they harbor the most ancient protein fold architectures that make up most of metabolism; Caetano-Anollés et al. 2007) and encircles the other two rings, which organize mainly around translation. The second ring organizes around aRSs, while the most inner circle comprises ribosomal and information transfer components. It is noteworthy that the organization of this core of ‘persistent’ genes reflects the evolution of functions during primordial life that we reveal in our timelines, from metabolism of small molecules, to aminoacylation of tRNA, to a world of molecules associated with functional RNA. The fact that results derived from an analysis of domain structure and physical clustering of genes in genomes reveal congruent and mutually supporting evolutionary scenarios is of utmost importance. Taken together, these results strongly support the late evolutionary arrival of translation.

Many have embraced the idea that macromolecules had a fuzzy start (Ycas 1974; Kacser and Beeby 1984). For example, recent studies suggest ancient aRSs may have lacked specificity, producing statistical peptides in the framework of a primitive genetic code (Schimmel 2009), a property that even provides a selective advantage to modern cells (Bacher et al. 2007). It is also likely that both proteins and RNA populated an ancient RNP world, since RNA is tightly entwined with protein function in modern biochemistry (Jeffares et al. 1998; Collins et al. 2009). The concerted evolution of proteins and RNA in 5S rRNA molecules (Sun and Caetano-Anollés 2009) and other functional RNA molecules (A. Harish and G. Caetano-Anollés, submitted) suggests proteins evolved hand-in-hand with RNA molecules. In our studies, it is evident that new and more sophisticated interactions occur increasingly between and within molecules in the timeline, beginning with proteins that were catalytically inefficient but promiscuous (Ycas 1974; Kacser and Beeby 1984). Continued evolution of these promiscuous architectures resulted in highly versatile folds that fulfill many metabolic roles, such as the Rossmann and the TIM β/α-barrel folds that populate metabolism. The discovery of new and more specific domains, the accretion of domains in proteins, and the establishment of complex macromolecular machinery later on exemplify the gradual build-up of interactions and increases in specificity. The outcome is autocatalysis and the crystallization process induced by modern replication machinery and a universal genetic code that appears at the end of the architectural diversification epoch. While the narrowing in the hourglass patterns constrains change and diversification, it enables the rise of domains as modules in cellular lineages. Remarkably, the outcome is the 'big bang' of domain combination in proteins that occurs at the start of the organismal diversification epoch (Wang and Caetano-Anollés 2009).

Acknowledgments A substantial portion of this work is part of DCA's undergraduate thesis. We thank Ajith Harish and Feng-Jie Sun for providing data on RNA-protein interactions, Minglei Wang for phylogenomic reconstruction, and Rakhee Kalelkar for help with construction of Z-diagrams. Research was supported by the National Science Foundation (MCB-0749836), the Illinois C-FAR program, CREES-USDA, and the International Atomic Energy Agency in Vienna. Any opinions, findings, and conclusions and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Altman S (2009) A view of RNase P. Mol Biosys 3:604–607

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

Archie JW (1989) Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. Syst Zool 38:253–269

Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

Bacher JM, Waas WF, Metzgar D, de Crecy-Lagard V, Schimmel P (2007) Genetic code ambiguity confers a selective advantage on *Acinetobacter baybili*. J Bacteriol 189:6469–6496

Bagley RJ, Farmer JD, Fontana W (1991) Evolution of metabolism. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial life II. Studies in the science of complexity, vol X. Addison-Wesley, Princeton, pp 141–158

Berchtold H, Reshetnikova L, Reiser COA, Schirmer NK, Sprinzl M, Hilgenfeld R (1993) Crystal structure of active elongation factor Tu reveals major domain rearrangements. Nature 365:126–132

Bogdanov AA, Dontsova OA, Dokudovskaya SS, Lavrik IN (1995) Structure and function of 5S rRNA in the ribosome. Biochem Cell Biol 73:869–876

Brenner SE, Kohl P, Levitt M (2000) The ASTRAL compendium for protein and sequence analysis. Nucleic Acids Res 29:254–256

Britton RA (2009) Role of GTPases in bacterial ribosome assembly. Annu Rev Microbiol 63:155–176

Caetano-Anollés G (2002) Tracing the evolution of RNA structure in ribosomes. Nucleic Acids Res 30:2575–2587

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

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

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

Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE (2009b) The origin and evolution of modern metabolism. Intl J Biochem Cell Biol 41:285–297

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

Choi I-G, Kim S-H (2007) Global extent of horizontal gene transfer. Proc Natl Acad Sci USA 104:4489–4494

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

Collins LJ, Kurland CG, Biggs P, Penny D (2009) The modern RNP world of eukaryotes. J Hered 100:597–604

Coulson AFW, Moult J (2002) A unifold, mesofold, and superfold model of protein fold use. Proteins 46:61–71

Csaba G, Birzele F, Zimmer R (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Struct Biol 9:23

Daigle DM, Brown ED (2004) Studies of the interaction of *Escherichia coli* YjeQ with the ribosome in vitro. J Bacteriol 186:1381–1387

Danchin A, Fang G, Noria S (2007) The extant core bacterial proteome is an archive of the origin of life. Proteomics 7:875–889

Deutscher MP (1984) Processing of tRNA in prokaryotes and eukaryotes. CRC Crit Rev Biochem 17:45–71

Dokudovskaya S, Dontsova O, Shpanchenko O, Bogdanov A, Brimacombe R (1996) Loop IV of 5S ribosomal RNA has contacts both to domain II and to domain V of the 23S RNA. RNA 2:146–152

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2129
Doolittle RF (2005) Evolutionary aspects of whole-genome biology. Curr Opin Struct Biol 15:248–253
Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anollés G (2010) History of biological metal utilization inferred through phylogenomic analysis of protein structure. Proc Natl Acad Sci USA 107:10567–10572
Egel R (2009) Peptide-dominated membranes preceding the genetic takeover by RNA: latest thinking on a classic controversy. BioEssays 31:1100–1109
Ellington AD, Chen X, Robertson M, Syrett A (2009) Evolutionary origins and directed evolution of RNA. Intl J Biochem Cell Biol 41:254–265
Forslund K, Henricson A, Hollich V, Sonnhammer E (2008) Domain tree-based analysis of protein architecture evolution. Mol Biol Evol 25:254–264
Fox SW (1980) Metabolic microspheres. Naturwissenschaften 67:378–383
Freeland SJ, Knight RD, Landweber LF (1999) Do proteins predate DNA. Science 286:690–692
Gesteland RF, Cech TR, Atkins JF (2006) The RNA world, 3rd edn. Cold Spring Harbor Laboratory Press, New York
Goldman AD, Samudrala R, Baross JA (2010) The evolution and functional repertoire of translation proteins following the origin of life. Biol Direct 5:15
Gough J (2005) Convergent evolution of domain architectures (is rare). Bioinformatics 21:1464–1471
Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure. J Mol Biol 313:903–919
Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291–D297
Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:C47–C52
Hillis DM, Huelsenbeck JP (1992) Signal, noise, and reliability in molecular phylogenetic analysis. J Hered 83:189–195
Holland T, Veretnik S, Shindyalov I, Bourne P (2006) Partitioning protein structures into domains: why is it so difficult? J Mol Biol 361:562–590
Holzmann J, Frank P, Löfler E, Bennett KL, Gerner C, Rossmanith W (2008) RNase P without RNA: identification and functional reconstitution of the human mitochondrial tRNA processing enzyme. Cell 135:462–474
Hoogstraten CG, Sumita M (2007) Structure-function relationships in RNA and RNP enzymes: recent advances. Biopolymers 87:317–328
Huber C, Wächtershäuser G (2007) α-Hydroxy and α-amino acids under possible Hadean, volcanic origin-of-life conditions. Science 314:630–632
Ikehara K (2009) Pseudo-replication of [GADV]-proteins and origin of life. Int J Mol Sci 10:1525–1537
Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806
Jeffares DC, Poole AM, Penny D (1998) Relics from the RNA world. J Mol Evol 46:18–36
Ji HF, Kong DX, Shen L, Chen LL, Ma BG, Zhang HY (2007) Distribution patterns of small molecule ligands in the protein universe and implications for origins of life and drug discovery. Genome Biol 8:R176
Joyce GF (2002) The antiquity of RNA-based evolution. Nature 418:214–221
Kacser H, Beeby R (1984) On the origin of enzyme species by means of natural selection. J Mol Evol 20:38–51
Karplus K (2009) SAM-T08, HHM-based protein structure prediction. Nucleic Acids Res 37:W492–W497
Kauffmann SA (1986) Autocatalytic sets of proteins. J Theor Biol 119:1–24
Kauffmann SA (1993) The origins of order. Oxford University Press, New York
Kauffmann SA (2007) Question 1: origin of life and the living state. Orig Life Evol Biosph 37:315–322
Kim KM, Caetano-Anollés G (2010) Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. Mol Biol Evol 27:1710–1733
Kim HS, Mittenthal JE, Caetano-Anollés G (2006) MANET: tracing evolution of protein architecture in metabolic networks. BMC Bioinformatics 7:351
Kluge AG, Farris JS (1969) Quantitative phyletics and the evolution of anurans. Syst Zool 30:1–32
Kummerfeld SK, Teichmann SA (2009) Protein domain organization: adding order. BMC Bioinformatics 10:39
Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. Proc Natl Acad Sci USA 100:9658–9662
Leibundgut M, Frick C, Thanbichler M, Böck A, Ban N (2005) Selenocysteine tRNA-specific elongation factor SelB is a structural chimaera of elongation and initiation factors. EMBO J 24:11–22
Lesk AM (2001) Introduction to protein architecture. Oxford University Press, New York, USA
Levitt M (2009) Nature of the protein universe. Proc Natl Acad Sci USA 106:11079–11084
Li J, Browning S, Mahal SP, Oelschiegel AM, Weissmann C (2010) Darwinian evolution of prions in cell culture. Science 327:869–872
Maguire BA, Beniaminov AD, Ramu H, Mankin AS, Zimmermann RA (2005) A protein component at the heart of an RNA machine: the importance of protein L27 for the function of the bacterial ribosome. Molecular Cell 20:427–435
Marahiel MA (2009) Working outside the protein-synthesis rules: insights into non-ribosomal peptide synthesis. J Pept Sci 15:799–807
Moore P (2005) The GTPase switch in ribosomal translocation. J Biol 4:7
Moore AD, Björklund ÅK, Ekman D, Bornberg-Buer E, Elofsson A (2008) Arrangements in the modular evolution of proteins. Trends Biochem Sci 33:444–451
Morowitz HJ (1999) A theory of biochemical organization, metabolic pathways, and evolution. Complexity 4:39–53
Murzin AG, Brenner SE, Hubbard TH, Chothia C (1995) SCOP: the structural classification of proteins database. J Mol Biol 247:536–540
Nixon KC (1999) The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15:407–414
Orgel LE (2000) Self-organizing biochemical cycles. Proc Natl Acad Sci USA 97:12503–12507
Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616–623
Raff R (1996) The shape of life. University of Chicago Press, Chicago
Ranea JAG, Sillero A, Thornton JM, Orengo CA (2006) Protein superfamily evolution and the last universal common ancestor (LUCA). J Mol Evol 63:513–525
Rodnina MV, Wintermeyer W (2009) Recent mechanistic insights into eukaryotic ribosomes. Curr Opin Cell Biol 21:435–443
Schimmel P (2009) Development of tRNA synthetases and connection to genetic code and disease. Protein Sci 17:1643–1652

Schimmel P, Ribas de Pouplana L (2000) Footprints of aminoacyl-tRNA synthetases are everywhere. Trends Genet 25:207–209
Schuster P (2010) Genotypes and phenotypes in the evolution of molecules. In: Caetano-Anollés G (ed) Evolutionary genomics systems biology. Wiley, Hoboken, pp 123–152
Seiradake E, Mao W, Hernandez V, Baker SJ, Plattner JJ, Alley MRK, Cusack S (2009) Structure of the human cytosolic leucyl-tRNA synthetase editing domain. J Mol Biol 390:196–207
Sun F-J, Caetano-Anollés G (2008a) Evolutionary patterns in the sequence and structure of transfer RNA: a window into early translation and the genetic code. PLoS ONE 3:e2799
Sun F-J, Caetano-Anollés G (2008b) The origin and evolution of tRNA inferred from phylogenetic analysis of structure. J Mol Evol 66:21–35
Sun F-J, Caetano-Anollés G (2009) The evolutionary history of the structure of 5S ribosomal RNA. J Mol Evol 69:430–443
Sun F-J, Caetano-Anollés G (2010) The ancient history of the structure of ribonuclease P and the early origins of Archaea. BMC Bioinformatics 11:153
Swofford DL (2002) Phylogenetic Analysis Using Parsimony and Other Programs (PAUP*). Ver 4.0b10. Sinauer, Sunderland, MA
Trefil J, Morowitz HJ, Smith E (2009) The origins of life. Am Sci 97:206–213
Tress ML, Ezkurdia A, Richardson JS (2009) Target domain definition and classification in CAP8. Proteins 77:10–17
Vetsigian K, Woese CR, Goldenfeld N (2006) Collective evolution and the genetic code. Proc Natl Acad Sci USA 103:10696–10701
Vogel C (2005) Function annotation of SCOP domain superfamilies 1.69. Superfamily—HMM library and genome assignments server. http://supfam.mrc-lmb.cam.ac.uk/beta_SUPERFAMILY/function.html
Vogel C, Chothia C (2006) Protein family expansions and biological complexity. PLoS Comp Biol 2:e48
Voorhees RM, Weixlbaumer A, Loakes D, Kelley AC, Ramakrishnan V (2009) Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome. Nat Struct Mol Biol 16:528–533
Wächtershäuser G (1990) Evolution of the first metabolic cycles. Proc Natl Acad Sci USA 87:200–204
Wächtershäuser G (2007) On the chemistry and evolution of the pioneer organism. Chem Biodivers 4:584–602
Wang M, Caetano-Anollés G (2006) Global phylogeny determined by the combination of protein domains in proteomes. Mol Biol Evol 23:2444–2454
Wang M, Caetano-Anollés G (2009) The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. Structure 17:66–78
Wang M, Boca SM, Kalelkar R, Mittenthal JE, Caetano-Anollés G (2006) A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. Complexity 12:27–40
Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res 17:1572–1585
Wang M, Jiang Y-Y, Kim KM, Qu G, Ji HF, Mittenthal JE, Zhang H-Y, Caetano-Anollés G (2010) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Mol Biol Evol [Epub ahead of print]
Woese CR (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859
Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci USA 99:8742–8747
Wolf YI, Aravind L, Grishin NV, Koonnin EV (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 9:689–710
Yang S, Bourne PE (2009) The evolutionary history of protein domains viewed by species phylogeny. PLoS ONE 4:e8378
Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined based on protein domain content. Proc Natl Acad Sci USA 102:373–378
Ycas M (1974) On earlier states of the biochemical system. J Theor Biol 44:145–160
Yomo T, Saito S, Sasai M (1999) Gradual development of protein-like global structures through functional selection. Nat Struct Biol 6:743–746
Yusupov MM, Yusupov GZ, Baucom A, Lieberman L, Earnest TN, Cate JHD, Noller HF (2001) Crystal structure of the ribosome at 5.5 Å resolution. Science 292:883–896
Zavialov AV, Hauryliuk VV, Ehrenberg M (2005) Guanine-nucleotide exchange on ribosome-bound elongation factor G initiates the translocation of tRNAs. J Biol 4:9