Structure
Article

The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World

Minglei Wang[1] and Gustavo Caetano-Anollés[1,*]
[1]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
*Correspondence: gca@illinois.edu
DOI 10.1016/j.str.2008.11.008

SUMMARY

Protein domains are compact evolutionary units of structure and function that usually combine in proteins to produce complex domain arrangements. In order to study their evolution, we reconstructed genome-based phylogenetic trees of architectures from a census of domain structure and organization conducted at protein fold and fold-superfamily levels in hundreds of fully sequenced genomes. These trees defined timelines of architectural discovery and revealed remarkable evolutionary patterns, including the explosive appearance of domain combinations during the rise of organismal lineages, the dominance of domain fusion processes throughout evolution, and the late appearance of a new class of multifunctional modules in Eukarya by fission of domain combinations. Our study provides a detailed account of the history and diversification of a molecular interactome and shows how the interplay of domain fusions and fissions defines an evolutionary mechanics of domain organization that is fundamentally responsible for the complexity of the protein world.

INTRODUCTION

The protein world has a modular and hierarchical organization at both molecular sequence and structural levels (Chothia et al., 2003; Grant et al., 2004). Protein molecules generally fold into compact architectures during a complex ''origami'' that arranges helical and strand elements of secondary structure in three-dimensional (3D) space. These tightly folded segments of the polypeptide chain constitute *protein domains*, structural and evolutionary ''modules'' that appear recursively singly or in combination with other domains in protein molecules. Structurally speaking, a module is here defined as a set of submolecular components that interact more extensively with each other than with other components outside the set and cooperate to perform a task. The module becomes a functional module when the task relates to a biological function (Hartwell et al., 1999). In the case of domains, modules occupy specific positions in the polypeptide chain and sometimes combine to produce homomultimeric (domain-repeat) or heteromultimeric (multidomain) proteins that contain a single or multiple types of domains, respectively (Vogel et al., 2004a). The combination and rearrangement of these modules during evolution defines a molecular ''interactome,'' a small-world and scale-free network of possible intramolecular interactions that delimit domain neighbor relationships in protein molecules (Apic et al., 2001a, 2001b; Wuchty, 2001). Our knowledge of how biological functions apportion within the modules of this interactome is incipient and is one important target of this study.

Domains embed biological function and are highly conserved (Bajaj and Blundell, 1984). Consequently, they are generally regarded as evolutionary units in structural classification schemes (Murzin et al., 1995; Orengo et al., 1997). For example, according to one popular taxonomy, the Structural Classification of Proteins (SCOP) (Murzin et al., 1995), domains that are closely related at the sequence level (generally expressing >30% amino acid residue identities), are pooled into fold families (FF). FFs sharing functional and structural features suggestive of a common evolutionary origin are further unified into fold superfamilies (FSF), and FSFs that share similarly arranged and topologically connected secondary structures are further grouped into protein folds (F) (Figure 1A). Both the atomic structure of domains (*domain structure*) and the way they are arranged along the sequence of multimeric proteins (*domain organization*) (herein collectively termed *domain architecture*; Figure 1B) are far more conserved than protein sequence. For example, tyrosine kinases exhibit highly conserved domain arrangements at the subfamily level along metazoan lineages (Shiu and Li, 2004). In fact, global phylogenetic trees based on a genomic census of domain structure (e.g., Caetano-Anollés and Caetano-Anollés, 2003; Yang et al., 2005; Wang et al., 2007) and organization (Wang and Caetano-Anollés, 2006; Fukami-Kobayashi et al., 2007) carry deep evolutionary information and are in good agreement with, for example, trees of life based on ribosomal RNA sequences. These ''phylogenomic trees'' have branches that represent organismal lineages and leaves that represent protein repertoires in organisms (proteomes). In particular, global phylogenies based on the combination of domains in proteins reveal the tripartite nature of the living world, describe organismal histories satisfactorily, and support the concept that the process of domain combination is not random but curved by natural selection or an optimality criterion (Wang and Caetano-Anollés, 2006).

Understanding the evolutionary history and functional roles of domains and their interaction constitutes a paramount endeavor (Bashton and Chothia, 2002; Vogel et al., 2004a). The topology

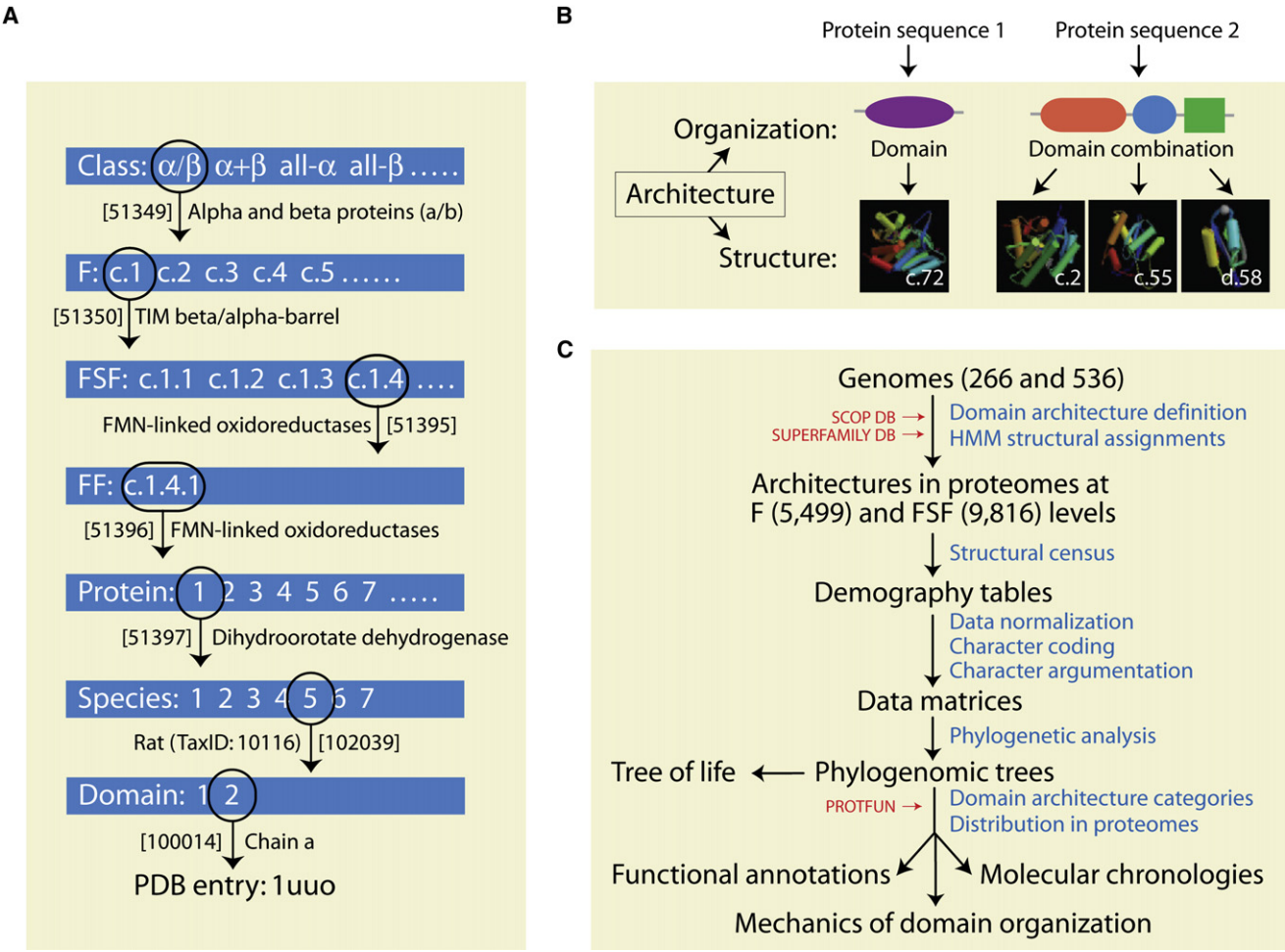


Figure 1. Analyzing the Evolutionary History of Protein Architecture

(A) Architectures can be defined at different levels of protein hierarchy using SCOP. Categories are described with alphanumeric labels and identifiers. Currently, a set of 1000 F, ~1800 FSF, and ~3500 families describes the world of proteins.

(B) Protein sequences have domains with architectures defined by the folding of the domain sequence in 3D space at F, FSF, or other levels of architectural hierarchy (domain structure) and by how domains combine with other domains in the polypeptide chain (domain organization).

(C) Flowchart describing data-mining strategies, including a structural census defined by advanced hidden Markov models (HMMs) that assign domain structure to genomic sequences, normalization of data, and phylogenetic analysis. Analyses of architectural distribution in proteomes, functional annotations, and domain categorization enable the reconstruction of architectural chronologies and the evolutionary study of biological function.

of domain combinations is usually highly conserved and the number of combinations limited (Bashton and Chothia, 2002; Wang and Caetano-Anollés, 2006). Interestingly, the module that embeds function is sometimes not the domain in multidomain proteins, but supradomains, two or three domain combinations that recur in different protein contexts (Vogel et al., 2004b). Many evolutionary studies have focused on domain organization, including conservation and variation of domain associations (Vogel et al., 2004a, 2004b, 2005; Bashton and Chothia, 2002; Apic et al., 2003), mechanisms of generation of new domain combinations (Enright et al., 1999; Marcotte et al., 1999), difference between fusion and fission mechanisms (Enright et al., 1999; Marcotte et al., 1999; Yanai et al., 2001), circular permutations in multidomain proteins (Jeltsch, 1999; Moore et al., 2008), and domain insertion and loss (Aroul-Selvam et al., 2004). Most of these studies use statistical or experimental approaches and generally take into account a limited number of domain-containing proteins. Consequently, they do not provide global evolutionary views.

Information in the 3D structure of domains can be used to study the evolution of the modern protein world. However, problems associated with the systematic classification of architectures at a topological level make it difficult, if not impossible, to find a general metric of pairwise comparison (Taylor, 2007). In search of other approaches, we have generated phylogenomic trees from the occurrence and abundance of domain structures in proteomes at F and FSF levels (Caetano-Anollés and Caetano-Anollés, 2003; Wang et al., 2006, 2007). These global trees are rooted and have branches representing architectural lineages and leaves representing the structures of individual domains. They were used to uncover patterns and processes in protein evolution (Caetano-Anollés and Caetano-Anollés, 2003, 2005;

Wang et al., 2006), origins and evolution of metabolic networks (Caetano-Anollés et al., 2007), and reductive tendencies in architectural repertoires linked to origins of diversified life (Wang et al., 2007). For example, patterns of representation of F and FSF architectures over evolutionary history revealed three epochs in the evolution of the protein world: (1) architectural diversification, where a relatively complex and communal protein repertoire is developed, (2) superkingdom specification, an epoch that sets the pace of an emerging tripartite world, and (3) organismal diversification, where architectures diversify along lineages in superkingdoms of life (Wang et al., 2007). These epochs were congruent with observations derived from an analysis of the sequence and structure of tRNA molecules (Sun and Caetano-Anollés, 2008)

Here we take advantage of a similar approach to study the evolution of the architectural repertoire of domains and domain combinations. We assign architectures to proteins in proteomes at F and FSF levels and use this architectural census to build data matrices of architectural abundance and reconstruct phylogenies that embed chronologies of molecular discovery (Figure 1C). Because the fusion of domains responsible for multimeric proteins and the fission of domain combinations have different relative rates (Kummerfeld and Teichmann, 2005; Pasek et al., 2006), we trace the fate of architectures directly in our phylogenomic trees and study the role of these processes in protein evolution. Results highlight the evolutionary mechanics of the protein world, revealing an explosive expansion of domain combinations that occurred relatively late in evolution and clear evolutionary patterns related to the fusion of domains and fission of domain combinations. These patterns were responsible for the modular rearrangement of molecular structure in proteins and were particularly important in Eukarya. We also survey the functions of architectures in specific organisms, uncovering interesting evolutionary patterns related to multiplicity of biological function.

RESULTS AND DISCUSSION

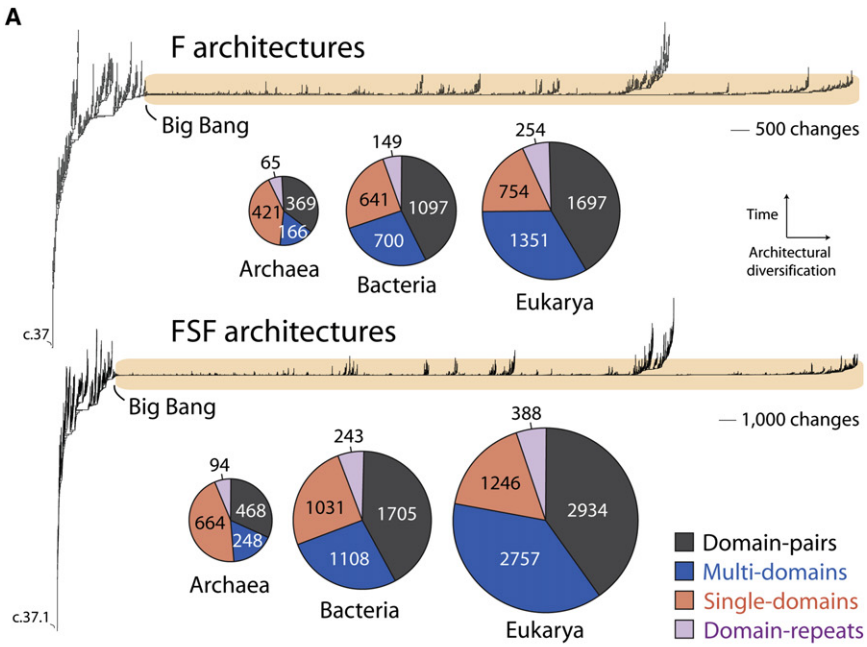
Reconstructing the Natural History of Domain Architecture in Proteins

We generated rooted phylogenomic trees using information embedded in a structural genomic census of domain architecture in organisms that have been fully sequenced (Figure 2A). We first reconstructed the optimal most parsimonious trees describing the history of 5499 architectures at F level that were present in 266 proteomes. We then extended the original analysis to 9816 architectures at FSF level in 536 proteomes. Because evolutionary patterns uncovered from these trees were congruent, we generally illustrate results with those generated at F level. The reconstruction of these large trees is computationally hard and tree visualization is challenging. We used a combined parsimony ratchet (PR) and iterative search approach to make reconstruction feasible (see Figure S1 available online). The trees represent timelines (chronologies) of architectural discovery and reveal fundamental patterns in the evolution of the protein world. The approach does not focus on protein-encoding genomic sequence, which is fast evolving, or domains defined at sequence profile level (e.g., Pfam; Finn et al., 2006), which can be subject to the vagaries imposed by mutation saturation and homoplastic confounding processes (Delsuc et al., 2005). Instead, our focus is on 3D architectural designs that are immutable over extended periods of time (Chothia et al., 2003). The discoveries of these architectures constitute important and rare events in the history of the protein world and are good repositories of deep phylogenetic signal in genomes (Caetano-Anollés and Caetano-Anollés, 2003). Our phylogenomic study also reveals how individual domains have combined with others in evolution to form domain combinations. To our knowledge, this represents the first direct phylogenetic reconstruction effort that describes the global history of a molecular interactome based on abundance of architectures in hundreds of proteomes. We note that recent studies have compared architectures found in different species, tracing for example Pfam domains onto NCBI taxonomy trees using substractive search methods (Pal and Guda, 2006) or architectural transformation pathways that are most parsimonious (Fong et al., 2007). These approaches, however, are entirely dependent on the taxonomy (species) tree that is used for the tracing exercise, the validity of which can be contested. More recently, Forslund et al. (2008) compared and traced the origin of architectures in a neighbor-joining tree from Pfam multidomain architectures, revealing that only 12.4% of these had multiple origins. These convergent evolutionary events at sequence profile levels were rare but were more numerous than those obtained (1.9%) by tracing SCOP structures in a species tree (Gough, 2005).

The phylogenetic relationships in the basal part of the trees were robust, displaying patterns that were consistent with a subtree that describes the evolution of the 100 most basal architectures and with global trees of domain structure and trees of domain organization that were reconstructed separately (e.g., Figure S2). Single-domain proteins (domains) appeared very early. In fact, only 17% of proteins harbored more than one domain in the subtree of basal architectures (Figure S2). The 20 most ancestral architectures belonged to the four major protein classes, α/β, $\alpha+\beta$, all-α, and all-β, as well as membrane and cell-surface proteins and peptides. Their basal placement was consistent with the most ancestral taxa in trees describing the evolution of the protein world that we reconstructed previously (Caetano-Anollés and Caetano-Anollés, 2003; Wang et al., 2006, 2007). Congruence in the appearance of protein classes in evolution provides further support to the proposal that architectural designs with interspersed α-helical and β strand elements were segregated, first within the structure and then confined to different molecules (Caetano-Anollés and Caetano-Anollés, 2003).

An Evolutionary "Big Bang" of Domain Combinations

The most notable feature of the trees of architectures was the abrupt appearance of a large number of terminal leaves halfway through evolution, most of which had short branch lengths. The evolutionary pattern was congruently observed in trees reconstructed at F and FSF levels (Figure 2A) and in subtrees of these trees (e.g., Figure S3). This suggests strongly that they are not the result of tree-building artifacts. Furthermore, an analysis of tree shape (node heights of internal nodes, ratios of external-to-internal lengths, and treeness statistics) and symmetry (N-bar and cherry counts) (Figure S4) showed values exceeded 95% confidence expectations, confirming trees were highly



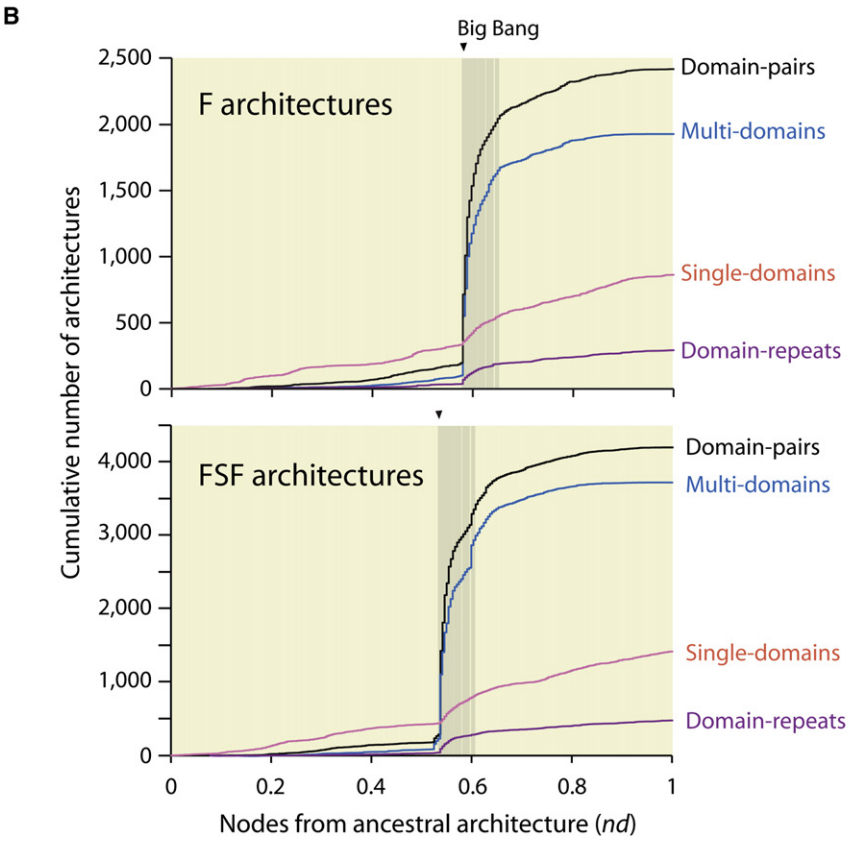


Figure 2. Phylogenomic Trees and Architectural Accumulation in Timelines

(A) Phylogenomic trees of domain architectures at F and FSF levels generated from a genomic census in 266 and 536 completely sequenced genomes, respectively. The optimal most parsimonious F tree (203,885 steps; CI = 0.026, RI = 0.732; g_1 = –0.329) and FSF tree (519,993 steps; CI = 0.021, RI = 0.711; g_1 = –0.329) were recovered from 11 PR searches using the strategy described in Figure S1 and were rooted using the Lundberg method. Fold nomenclature follows that given in SCOP 1.69 (June 2006). The 5499 terminal leaves were not labeled because they would not be legible. Pie charts show distribution of architectures belonging to four major categories in the three superkingdoms of life.

(B) Cumulative frequency distribution plots describing the accumulation of single-domain, domain-pair, domain-repeat, and multidomain architectures along the tree of architectures. Cumulative number is given as a function of distance (*nd*) in nodes from the hypothetical ancestral architecture, on a relative scale. The salmon shaded area shows the big bang of domain combinations resulting from a combinatorial burst.

unbalanced and were similarly shaped, even during the explosive appearance of lineages. These results support the idea that semipunctuated evolutionary processes are important drivers of architectural innovation in protein evolution and that evolution of protein architecture does not fit stochastic or null branching models (Kirkpatrick and Slatkin, 1993; McKenzie and Steel, 2000).

To unfold patterns of architectural discovery and explore their explosive appearance in the trees, we studied how architectures distributed among proteomes in cumulative frequency distribution plots (Figure 2B). These plots represent "architectural chronologies" (timelines) in which the accumulation of architectures was given as a function of relative distance in nodes from a hypothetical ancestral architecture in the tree (node distance; *nd*). We divided architectures according to four organizational schemes: (1) domains appearing singly (single domains), (2) domain combinations consisting of only two different domains (domain pairs), (3) domain combinations consisting of different domains, with domains sometimes repeated (multidomains), and (4) domains of one type that are repeated (domain repeats). For simplicity herein, we will refer to single-domain architectures as "domains" and to architectures in the other three categories as "domain combinations."

The evolutionary accumulation of these four architectural categories revealed that schemes of domain organization were established early in evolution (*nd* $\leq$ 0.135). The first architectures to emerge in the modern protein world were single-domain proteins that were omnipresent, but proteins harboring domain combinations (domain repeats, domain pairs, and multidomains, in that order) soon followed. However, the rate of accumulation of each

category differed notably. The cumulative rate of single domains and domain repeats was low and relatively constant. Their steady accumulation produced a relatively limited number of architectures in evolution when compared to those in other architectural categories. In contrast, the steady accumulation of domain-pair and multidomain architectures showed a marked and abrupt increase at *nd* ~0.58, which made these categories the most prevalent in the protein world, and slowed down again at *nd* ~0.65. The explosive exploration of domain organizational schemes is remarkable, coincides with the start of the organismal diversification epoch (Wang et al., 2007), was mostly restricted to superkingdom-specific architectures, and matched the topology of the reconstructed trees (Figure 2A). We have termed this phenomenon the "big bang" of domain combinations following an analogy related to the cosmological origin of the universe, and propose it is of fundamental evolutionary significance. In these studies, *nd* cannot be represented in a timescale of millions of years, because we have not identified architectures that can time important organismal diversification events in the history of the world and we do not know whether a molecular clock will apply to global evolution of the protein world. Nevertheless, time and *nd* are related by some function, and the big bang pattern characteristic of domain categories with complex organization schemes (domain-pair and multidomain) suggests these architectures were massively produced in a relatively short period following the initial rise of domains and domain repeats. Interestingly, the survey of domain organization in the three superkingdoms showed domain pairs and multidomains were particularly enhanced in the protein repertoire of Bacteria and Eukarya, mostly at the expense of single-domain architectures (Figure 2A). This analysis also revealed differences in the size of the architectural repertoires of individual superkingdoms that we observed earlier (Wang et al., 2007).

The late appearance of domain combinations in the protein world signals a major evolutionary transition (sensu Szathmáry and Smith, 1995). It involved a massive combinatorial exchange of modules and became evident only when organismal lineages were well established in all three superkingdoms of life. It is noteworthy that diversity of domain organization schemes increased with organismal complexity (Figure S5; Table S1). This trend was notable in multicellular organisms, particularly in metazoa, a finding that is in line with the elevated domain rearrangement levels observed in this group of organisms (Ekman et al., 2007). We propose that the pervasive movement of genes in chromosomes facilitated this combinatorial explosion, perhaps through the discovery or enhancement of chromosomal recombination (Vogel et al., 2005), intronic recombination of domain-encoding exons and faulty excision of introns (Patthy, 1999; Kaessmann et al., 2002), domain insertion and deletions at C and N termini (Björklund et al., 2005; Vibranovski et al., 2005; Weiner et al., 2006), and/or the activity of ancient retrotransposons (e.g., Moran et al., 1999). Mounting evidence also suggests that "exonization" of intron sequences can play an important role in the creation of domains (Schmidt and Davies, 2007), and these processes could be used to explain their fusion. Clearly, some mechanisms govern the creation of new architectures and other mechanisms facilitate the rearrangement of those in existence. Under these mechanistic scenarios, the placement of genes under new genomic contexts would sometimes result in recruitment of neighboring domains (fusional combinations) or "decoration" with neighboring sequences that would enhance the function of the embedded functional and evolutionary units (fusional/fissional combinations) (Figure 3A). Alternatively, the faulty excision of gene segments during chromosomal rearrangement would sometimes cause the fission of domains and domain combinations (fissional combinations) in processes that would give rise to new modules, some of which had the potential of enhancing the combinatorial interplay. The existence of each and every one of these processes (summarized in Figure 3B) is highly probable and is the subject of this study.

Timelines Support the Three Evolutionary Epochs of the Protein World

The distribution of architectures across the three superkingdoms of life, Archaea, Bacteria, and Eukarya (herein labeled A, B, and E, respectively), and along the evolutionary timeline of architectural discovery was congruent with results from a recent study of domain structure at F and FSF levels (Wang et al., 2007). Using cumulative frequency distribution plots and distribution indices (*f*), we revealed patterns emerging from the rooted tree that support the proposal made previously of three evolutionary epochs and reductive evolutionary tendencies embedded in the repertoire of architectures (Figures S6 and S7). Distribution patterns showed that ancient architectures were omnipresent or widely distributed in all organisms analyzed and, later, common to all superkingdoms (e.g., the 100 most basal; Figure S2). Omnipresent single-domain architectures were observed for the first time before omnipresent or widely distributed domain combinations (domain-repeat, domain-pair, and then multidomain). Architectures shared by the three superkingdoms (ABE domains and ABE combinations) maintained relatively constant rates of accumulation during the first half of the architectural timeline (Figure S6). All architectures shared by only two superkingdoms appeared later in evolution (BE, AB, and AE domains appeared in that order, as did BE, AB, and AE combinations), well after the emergence of ABE combinations, and already suggests the rise of organismal superkingdoms. Within this group, combinations were the only categories accumulating explosively during the big bang ($0.58 < nd < 0.65$).

BE architectures originated quite early, consistent with an observed close relationship between ancestors of Bacteria and Eukarya (Wang and Caetano-Anollés, 2006) and reductive tendencies in the ancient archaeal lineage (Wang et al., 2007). These BE architectures resulted fundamentally from a losing trend in the architectural repertoire of Archaea, which was clearly evident in patterns of representation of architectures in lineages (Figure S7). The losing trend represents the hallmark of the architectural diversification epoch and provides strong support for an early organismal divide in which the archaeal lineage was segregated from an ancient and architecturally rich community of organisms (Wang et al., 2007).

Superkingdom-specific architectures appeared later in evolution (Figure S6). B and AB architectures signal the start of the superkingdom specification epoch and, later on, A and E architectures delimit the organismal diversification epoch. The onset of this last epoch coincides with the big bang of domain combinations, in which E and B (and to a lesser degree BE) domain combinations expanded explosively in the protein world and

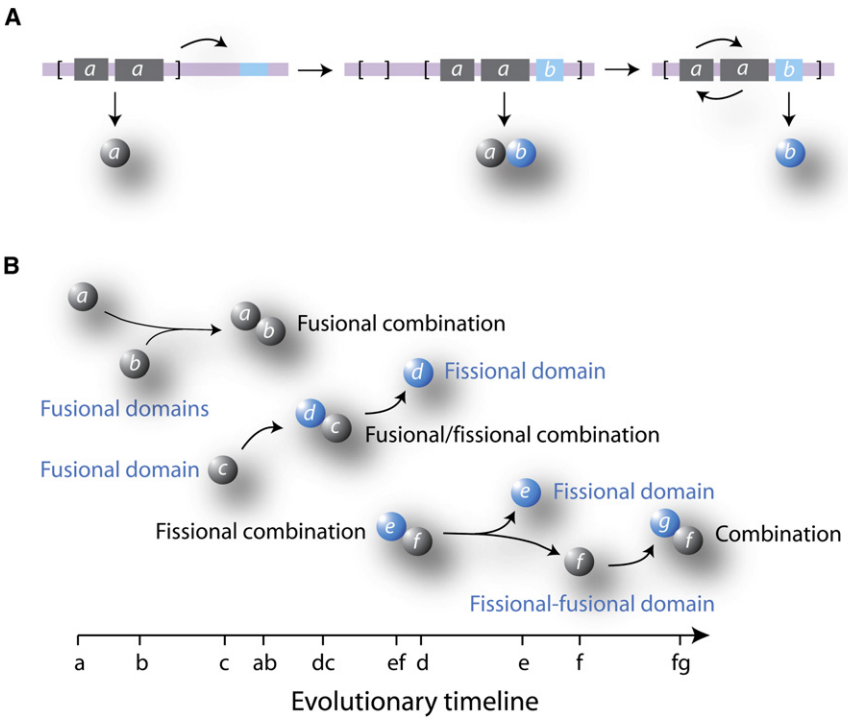


Figure 3. Processes Underlying the Combinatorial Repertoire of Domain Combinations

(A) An example of evolutionary recruitment and takeover, in which a sequence (blue segment) close to gene *a* (gray segments encoding domain *a*) is recruited from neighboring sequences (delimited by brackets) to form a new functional gene that encodes domain combination *ab*. This fusion process is followed by a fission that inactivates gene *a* by either a rearrangement or shuffling event. The inactivated gene later decays by mutation (not shown).

(B) The order of appearance of architectures along the evolutionary timeline defines six categories indicative of the evolutionary mechanics of domain organization: fusional domains, fissional domains, fissional/fusional domains, fusional combinations, fusional/fissional combinations, and fissional combinations. A seventh category includes domains that do not partake in the combinatorial game. These categories arise from three fundamental processes of domain organization that are illustrated in the diagram: (1) fusions: domains *a* and *b* (depicted with spheres) appear earlier than domain combination *ab*, so domain *a* and *b* are fusional domains and combination *ab* is a fusional combination resulting from the fusion of two domains; (2) fusions and fissions: domain *c* and *d* appear earlier and later than the combination *cd*, respectively, so domain *c* is a fusional domain, domain *d* is a fissional domain, and combination *cd* is a fusional/fissional combination resulting from a fusion with a newly discovered architecture which is destined to appear as single-domain later in evolution; (3) fissions and fusions: combination *ef* appeared earlier than both of its constituents, domains *e* and *f*, which result from fissions at times *e* and *f*, respectively. Combination *ef* is a fissional combination. However, domain *e* is a fissional domain, whereas domain *f* is a fissional/fusional domain because it fused to domain *g* later in evolution. Domains are depicted with gray spheres if they partake in fusions or with blue spheres if they involve only fissions.

made the highest contribution to the total number of extant architectures (65%). After *nd* = 0.65, the accumulation of A and B combinations reaches a plateau, whereas E domains and combinations continue to increase until the present, accounting for about half the architectural diversity of the protein world. These evolutionary trends were responsible for the make-up of architectural repertoires of present-day proteomes (Figure S5); repertoires were maximal in Eukarya and minimal in Archaea, with Bacteria in between. Total repertoires of superkingdom-specific architectures followed this pattern, matching global trends observed previously (Wang et al., 2007).

The appearance and rise of architectures across superkingdoms coincide almost perfectly with patterns we have previously observed (Wang et al., 2007). Phylogenetic congruence over such broad timescales provides strong evidence for a historical association between the architecture of functional proteins and the structure of embedded domains. This important phylogenetic link demonstrates that the combinatorial interactome of protein structure preserves accurately the evolutionary history of the protein world, despite protein recruitment processes (e.g., domain co-option). It also dispels the possibility that the big bang pattern in the tree arises solely from an artifact of tree reconstruction.

Timelines Provide Evolutionary Scenarios for Domain Rearrangement: An Example with Scaffolding Proteins

Timelines of discovery of domain modules and their assortment in combinations can help dissect the evolution of families of proteins that are related by function. For example, the membrane-associated guanylate kinases (MAGUKs) include proteins involved in cell-to-cell communication that are specific to metazoans (Funke et al., 2005). These scaffolding proteins tether adhesion molecules, receptors, and intracellular signaling enzymes, organizing macromolecular complexes at cellular junctions. Most MAGUK family members share a conserved core structure, which is composed of one or multiple PDZ domains, a Src homology 3 (SH3) domain, and a guanylate kinase (GK) domain (te Velthuis et al., 2007). A detailed phylogenetic analysis of MAGUK sequences in metazoan genomes provided indications that the core MAGUK structure originated from a GK-SH3 domain arrangement, which later combined with the PDZ domain (te Velthuis et al., 2007). This conclusion is consistent with the timeline of appearance of MAGUK domains and domain combinations in our tree (Figure 4). The GK domain has a P loop hydrolase fold (c.37) which appeared at the base of the tree (*nd* = 0), whereas the SH3 domain (b.34) arose later (*nd* = 0.264) from the fission of a domain combination (b.40|b.34; *nd* = 0.260). Note that F domain labels in this paper follow SCOP nomenclature (Murzin et al., 1995). The first two MAGUK families were discovered halfway through evolution, during the big bang (*nd* = 0.580), by incorporation of PDZ (b.36) and WW (b.72) domains into their architectures. These included the MAGI family (b.36|c.37|b.72&|b.36) and the DLG family (b.36&|b.34|c.37). Interestingly, both PDZ and WW domains appeared in single-domain proteins much later (b.36 at *nd* = 0.896 and b.72 at *nd* = 0.853), indicating their initial role was accessory. The core MAGUK structures SH3-GK

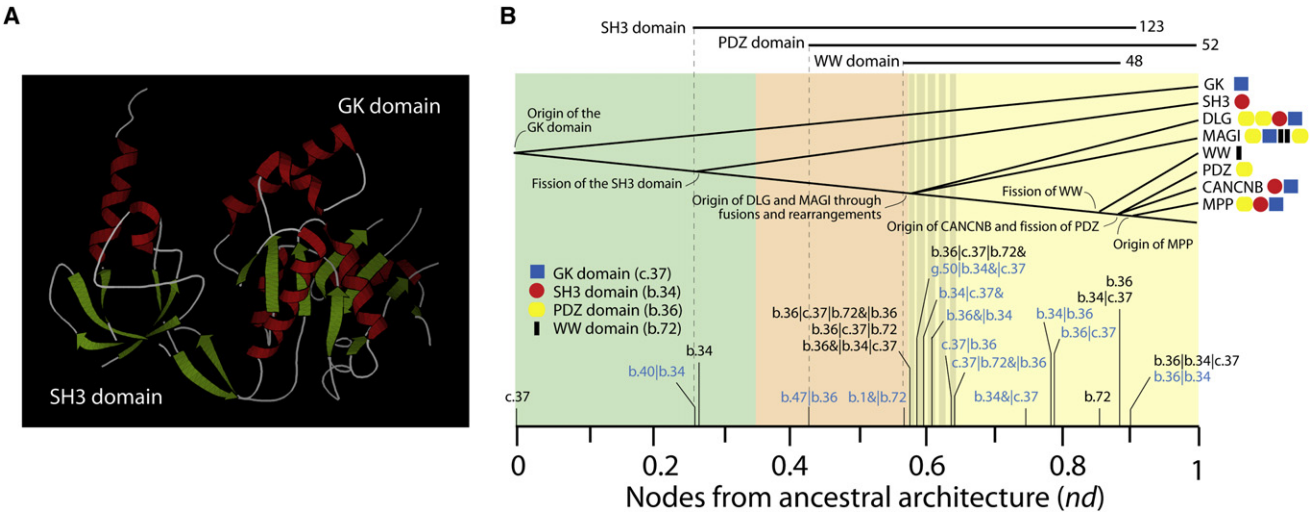


Figure 4. Evolution of Domain Organization in the MAGUK Family of Scaffolding Proteins

(A) The SH3-GK core defining a typical MAGUK protein shows the 3D arrangement of helices (red) and strands (green) defining the P loop hydrolase F for the GK domain and the SH3-like barrel F for the SH3 domain.

(B) Timeline describing the discovery of domain and domain combinations associated directly and indirectly with MAGUK proteins; these architectures are marked with lines along the architectural timeline and are indexed with black and blue labels, respectively. The tree illustrates the discovery and diversification of architectures associated with important MAGUK families. Shaded areas describe the architectural diversification (light green), superkingdom specification (salmon), and organismal diversification (light yellow) epochs (defined by Wang et al., 2007) and are defined according to landmarks described in Figure S3. Bars above the plot indicate the number of combinations and the range of *nd* values associated with architectures containing the SH3, PDZ, and WW domains in the timeline. F labels of architectures follow SCOP nomenclature and pipe and ampersand symbols denote domain junctions and domain repeats, respectively.

(b.34|c.37) and PDZ-SH3-GK (b.36|b.34|c.37) made their appearances quite late in evolution, at *nd* = 0.896 and *nd* = 0.900, respectively, and were subsequently accessorized with new domains or were subjected to PDZ duplications that ultimately gave rise to the complex MAGUK assortment now present in vertebrates (te Velthuis et al., 2007). The SH3, PDZ, and WW domains were quite promiscuous; they were involved in establishing 132, 52, and 48 domain combinations, respectively (Figure 4). Most of these occurred in the eukaryal lineage during or after the big bang and resulted in many MAGUK-like domain combinations. These results therefore support the model of MAGUK family evolution inferred from sequence analyses (te Velthuis et al., 2007).

Patterns of Modularity in the Protein World

The evolutionary mechanics of domain organization involve combining and splitting domains in fusion and fission processes to create, recruit, or enhance the biological functions that are needed to satisfy the growing complexity of life (Figure 3). Instances of these fundamental processes underlie the combinatorial rearrangement of domains and can be inferred directly from the order of appearance of domain combinations and their domain constituents along the universal tree of architectures, as we illustrated with MAGUK-related complements (Figure 4). Once domain combinations are generated, these can be further rearranged (e.g., circular domain permutations) through gene duplication, deletion, and rearrangement processes at DNA levels (Weiner et al., 2006) that may or may not involve fusion or fission processes.

The domain combinations we identified (4636 and 8397 at F and FSF levels) could therefore be divided into three categories describing fundamental mechanisms of domain organization that are based on historical happenings inferred directly from the tree (Figure 5; pie charts): (1) fusional combinations: combinations that appeared after all domain constituents had been discovered; (2) fissional combinations: combinations that appeared before the discovery of their domain constituents; and (3) fusional/fissional combinations: combinations that appeared before and after their domain constituents. Similarly, domains (863 and 1419 single-domain proteins at F and FSF levels) could be divided into four categories according to their role in domain organization (their ability to combine in evolution): (1) fusional domains: domains that fused with others to form combinations; (2) fissional domains: domains that resulted from fission of combinations; (3) fissional/fusional domains: domains that originated from fission processes but that later in evolution engaged in fusions; and (4) noncombinable domains: domains that did not partake in any combination. As we traced domains and domain combinations in these categories along molecular chronologies (Figure 5), we carefully annotated biological function (apportioned conservatively in 12 functional categories) in a representative genome (*Homo sapiens*) using a sequence-based (ab initio) prediction method (Figure 6). Our objective was to find links between fusion and fission processes and function in our trees as these developed in time. The exercise revealed remarkable evolutionary patterns, as follows.

Early Architectures Were Multifunctional Single Domains Poised to Combine by Fusions

Molecular chronologies showed that the first domains to be discovered had the potential to become modules (Figure 5). Each and every domain that appeared early during the

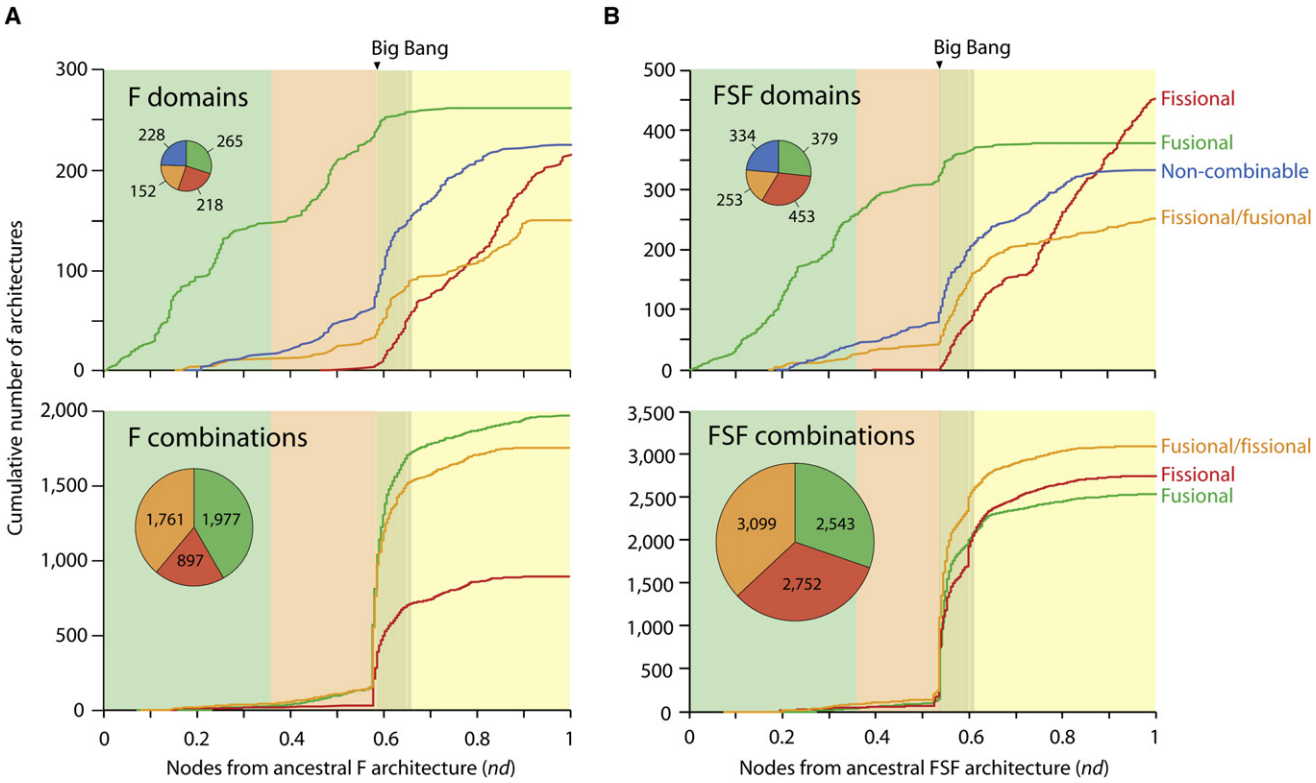


Figure 5. Accumulation of Architectures in Mechanistic Categories along the Universal Tree of Architectures

Cumulative frequency distribution plots describe architectural accumulation at F (A) and FSF levels (B). The background colors indicate the three evolutionary epochs of the protein world and the big bang, colored as in Figure 4. Pie charts categorize extant architectures, with the exception of one F (g.5&) and 3 FSF (a.8.1&, b120.1|b.1.20&, and d.58.33&) architectures that could not be assigned to categories. Besides these mechanistic categories, a group of 12 F and 32 FSF domains appeared only combined with others and never by themselves. These domains and associated domain combinations at F level (in parentheses) are the following: a.171 (a.170|b.40|a.171), a.49 (a.35|c.37|a.49), a.58 (a.58|c.66), a.89 (d.58|a.89), a.92 (c.30|d.142|a.92|c.30|d.142|c.24), b.114 (b.114|d.58|b.34), b.120 (b.120|b.1&), b.142 (b.142|c.52), b.48 (a.4|c.55|b.48), c.102 (c.102|b.80), c.105 (c.105|c.76), d.121 (d.121|d.163), g.5 (g.5&), and g.59 (d.241|g.59|b.40). Pipe and ampersand symbols denote domain junctions and domain repeats.

architectural diversification epoch (*nd* < 0.1) was a fusional domain, that is, it fused later with others to form fusional combinations (Figure 5). Most of these were highly multifunctional (Figure 6A). This was expected because ancient architectures located at the base of phylogenomic trees were found associated, for example, with many enzymatic functions (Caetano-Anollés and Caetano-Anollés, 2003). Moreover, a careful tracing exercise confirmed that the first nine F architectures delimited almost all major enzymatic activities that exist in cellular metabolism (Caetano-Anollés et al., 2007). Architectures emerging later (0.1 < *nd* < 0.3) had generally fewer functions, although several were highly multifunctional (e.g., a.60, a.118, d.144). However, the number of functions per architecture dropped precipitously after *nd* = 0.3, as the world entered into the superkingdom specification epoch.

Fission and Noncombinable Domains Gained Prevalence with Time

Fusional domains were followed by fissional/fusional domains with low-to-moderate numbers of functions (0.12 < *nd* < 0.91) and by noncombinable domains with single or very few associated functions (0.18 < *nd*) (Figures 5 and 6A). Finally, fissional domains appeared quite late (0.46 < *nd*), increased substantially in number during the big bang, and later on became highly multifunctional. Whereas fusional domains dominated evolution of single domains throughout most of early modern life (during architectural diversification and superkingdom specification), domains in all other categories (including those that did not combine) dominated evolution of the more recent protein world (during organismal diversification). Fissions became more and more popular as time progressed. The first domains produced from fission engaged later in fusions. Most domains of the fissional/fusional class appearing early originated by fission from fissional combinations or fusional/fissional combinations and later on engaged in fusion processes.

It is remarkable that a considerable number of domains (26% and 23% at F and FSF levels, respectively) failed to partake in the combinatorial game and that this phenomenon started to occur relatively early, during the architectural diversification epoch (Figure 5). Interestingly, 5 out of the first 17 noncombinable domains were missing entirely in Archaea despite being widely shared by most organisms (on average absent in 5–10 lineages) and 8 were ribosomal protein domains (Table S2). Noncombinable domains doubled during the big bang and reached a plateau at *nd* ~0.8, suggesting they represent a side product of the

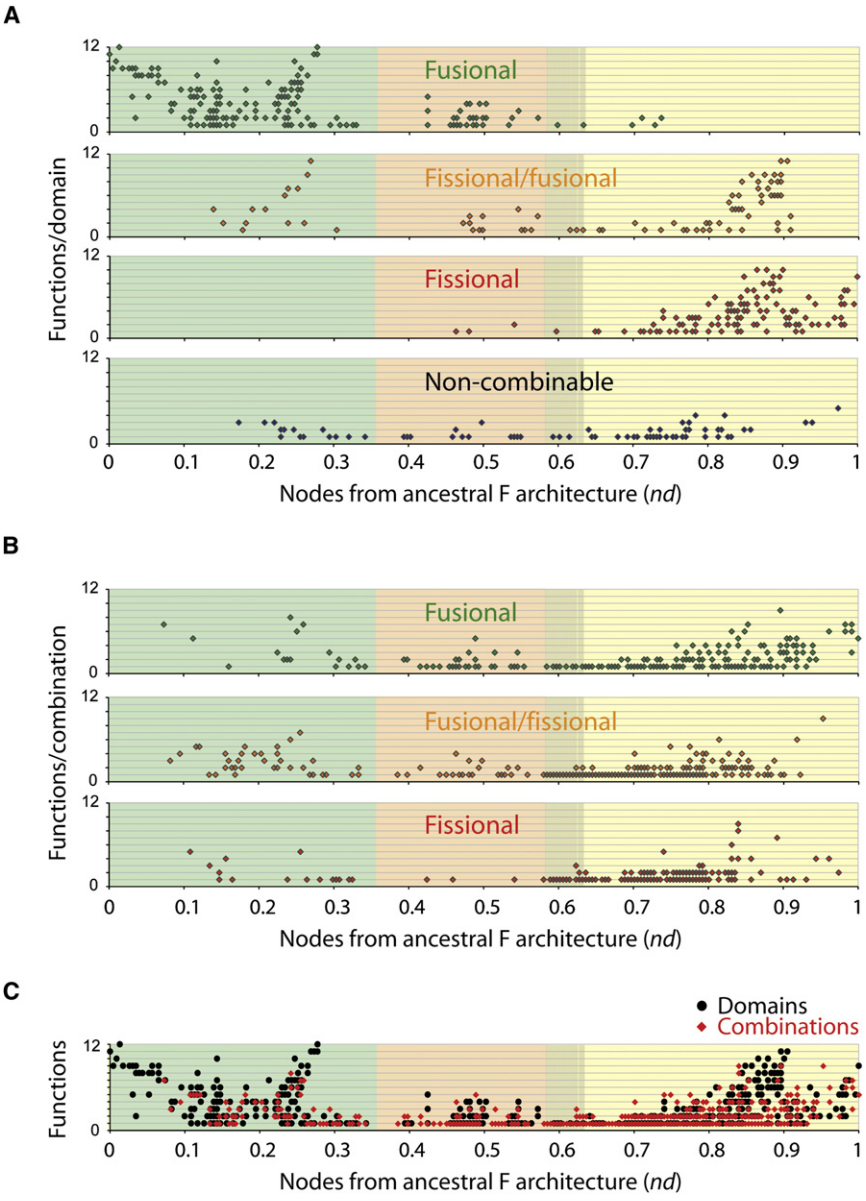


Figure 6. Evolution of Biological Function along the Tree of Architectures

A total of 15,029 protein sequences from man (*H. sapiens*) was assigned to 1,163 architectures, which were then annotated using ab initio prediction of protein function directly from sequence. The number of annotated functions per domain in architectural categories associated with domains (A), domain combinations (B), and all architectures (C) were plotted along the evolutionary timeline (with ancestries given in *nd* values).

combinatorial interplay. We believe these domains adapted to become specialists and represent evolutionary dead ends of domain organization resulting from processes of structural lock-in (structural "canalization"; sensu Ancel and Fontana, 2000). Also remarkable is the existence of only 14 domains at F level and 32 domains at FSF level that appeared only in combination but never as independent domains (listed at F level in Figure 5). They originated within the $0.43 < nd < 0.76$ range. These domains failed to split from domain combinations throughout history. However, they represent rare exceptions; domains that combine always appear as independent modules. Both these and the noncombinable domains may result from extreme cases of structural lock-in or functional specialization.

Unbiased Domain Adoption by Combinations

We did not find any particularly striking link between the adoption of domain architectures and the formation of multidomains as the result of domain fusions (Figure S8A). Most of the fusional combinations appeared during the big bang. Their youngest domain constituents generally ranged from being very ancient to contemporary with the combination, indicating there was no bias in the adoption of domains by domain combinations. However, we noticed that very few domains within the $0.32 < nd < 0.40$ range were fused into a combination. This suggests an evolutionary "gap" of domain adoption during this evolutionary period in which most domains that were discovered did not partake in fusion processes (Figure S8A). Only 16 domains appeared during this period (6.4% of all 250 at $nd = 0.4$), half of which were noncombinable. The numbers of associated proteins in human were also relatively small (Figure 6). It is noteworthy that several of these domains were linked to ribosomal function through SCOP assignments, and most peptides were assigned to "amino acid biosynthesis" and "translation" (Figure S9). These domains are therefore specific to protein synthesis. It is tempting to hypothesize that during this "gap" either a fundamental revision of the protein biosynthetic apparatus occurred or a cataclysmic event on Earth curtailed the expansion of the protein world.

Dominance of Fusion-Driven Combinations

The first fusions occurred relatively early in the tree (Figure S2) but at very low rates and produced at first domain-repeat and later domain-pair and multidomain arrangements. Most of these early domain combinations were associated with a moderate number of functions. Cumulative plots show clearly that fusional combinations appeared first, followed by fusional/fissional combinations and then fissional combinations (Figure 5). The modular combination and rearrangement of architectures were protracted and continued throughout evolution, making fusional domains and fusion-driven combinations the most abundant in the protein universe. Almost half of domains and about three fourths of combinations involved fusion processes. In fact, we observed dominance of fusional and fusional/fissional

combinations over fissional counterparts along the entire molecular chronology, with proportional increased representations of the three categories in the protein world during the big bang phase. These observations are compatible with previous studies that showed fusions were dominant contributors to evolution of domain architecture (Kummerfeld and Teichmann, 2005; Pasek et al., 2006).

Functional Specialization during the Big Bang

The explosive increase in domain combinations at the start of organismal diversification (Figure 5) coincides with a period in which multifunctional architectures were clearly replaced by single-function counterparts (Figure 6; red hues in heat maps of Figure S9). The combinatorial burst of domain combinations probably fulfilled the different functions needed by the emerging organismal lineages, replacing multifunctional proteins with highly specialized alternatives. This probably enhanced recruitment processes, which have been shown to be pervasive in metabolism (Teichmann et al., 2001; Kim et al., 2006; Caetano-Anollés et al., 2007).

Late Rise of Multifunctional Fissional Domains

After the big bang phase, Eukarya-specific architectures continued to accumulate, perhaps to fulfill the increasingly complicated needs of multicellular organisms (Figure S6). No such tendency was evident in prokaryotic microbes, which failed to enrich the architectural and functional repertoire to that level. The combinatorial interplay originally fueled by fusion was also revised during this period with a new reductive evolutionary process of architectural and functional diversification (fission) that atomized domain combinations to form new multifunctional modules. In particular, fissional domains became highly multifunctional well after the big bang ($0.8 < nd$) and were mostly confined to Eukarya (Figure 6A). Examples include the PDZ and WW domains of Figure 4. However, their domain constituents did not distribute widely, as did their fusional counterparts (Figure S8B). Careful analysis of the age of their domain constituents showed that the simultaneous appearance of single-domain architectures arising from fission of a fissional combination was extremely rare. Instead, the fission process resulted in a protracted "losing" trend in which different components of the combination were lost or excised at different times. Consequently, new single domains had different *nd* values. This fissional phenomenon may be of important evolutionary significance, especially for Eukarya, as it enhanced both the repertoire of modules and the repertoire of functions (Wang et al., 2007). We hypothesize fissions had the potential to produce new and more versatile architectures. These were needed to fulfill the functional demands of complex lifestyles and life in diverse environmental niches that are characteristic of eukaryal organisms (L.S. Yafremava, J.E. Mittenthal, and G.C.-A., unpublished), including functions related to intercellular communication, recognition of self, and multicellularity (Caetano-Anollés and Caetano-Anollés, 2005).

Number of Domains in Domain Combinations and Distribution of Architectures among Superkingdoms of Life

One remarkable feature of the combinatorial game was the preference for combinations of two domains over multidomain arrangements (Figure 7). For example, the survey of the number of domains in a domain combination in the 4636 combinations analyzed at F level revealed that 64% of them had two domains, 18% had three domains, and the rest had either more than three domains or one domain constituent in a domain repeat (Figure 7A). We also noticed that the preference for two-domain organization was not significantly biased by the mechanics of domain organization, although fusional combinations showed a slight tendency to form domain pairs. From an evolutionary standpoint and if domain organization is shaped by selection, it appears costly for domains to engage in the combinatorics of more than two modules. This may relate to limitations imposed by chain length and environment (Brocchieri and Karlin, 2005; Kurland et al., 2007). Multidomain arrangements may require exceedingly long protein sequences, perhaps encoded in genes with many introns. This could enhance domain shuffling and could increase fission propensity. Remarkably, the number of multidomains became significant halfway during superkingdom specification, at a time that coincides with the rise of fissional domains in evolution (Figures 2 and 5). This observation supports the idea that long multidomain proteins are generally prone to fission, shedding segments or forming new domains in the process. In this regard, it is particularly noteworthy that fissional domains became popular quite late and during the big bang. These fissional domains include domains generated by fission from fusional and fusional/fissional combinations (generally at $0.8 < nd$; Figure S8) and domains that were generated by fission but that later engaged in fusion processes (fissional/fusional domains) (Figure 5; Figure S8). All of these processes appeared particularly active in multicellular eukaryal species.

Analysis of how domain combinations at F level distributed among the three superkingdoms showed there were distinct differences among the fusional, fusional/fissional, and fissional combination categories (Figure 7B). Fusional combinations that were most common were specific to Bacteria (38%) and Eukarya (29%). In contrast, fusional/fissional combinations were mostly specific to Eukarya (64%). This trend was maximal with fissional combinations, where most combinations were Eukarya specific (90%). These results suggest fission processes were selectively enhanced in the eukaryotic superkingdom. A similar analysis of domain distribution showed that most fusional domains were common to all life (77%), whereas fissional/fusional, fissional, and noncombinable domains were mostly shared by all superkingdoms, shared by Eukarya and Bacteria, or specific to Eukarya (Figure 7C). A substantial number of noncombinable domains were also Bacteria specific.

Conclusions

Over 1000 genomes and metagenomes have been completely sequenced to date yielding millions of protein sequences and thousands of functional RNA molecules important for cell development and homeostasis. Structural genomics has also produced ~60,000 models of atomic structure embedded in Protein Data Bank entries, and advances in structural bioinformatics have extended structural information to macromolecules encoded by more than half of the gene complement identified in fully sequenced genomes (Grant et al., 2004). Our phylogenomic study has used this wealth of information to unravel the evolutionary mechanics of the protein world, showing the interplay of processes that combine, split, and rearrange domains. We

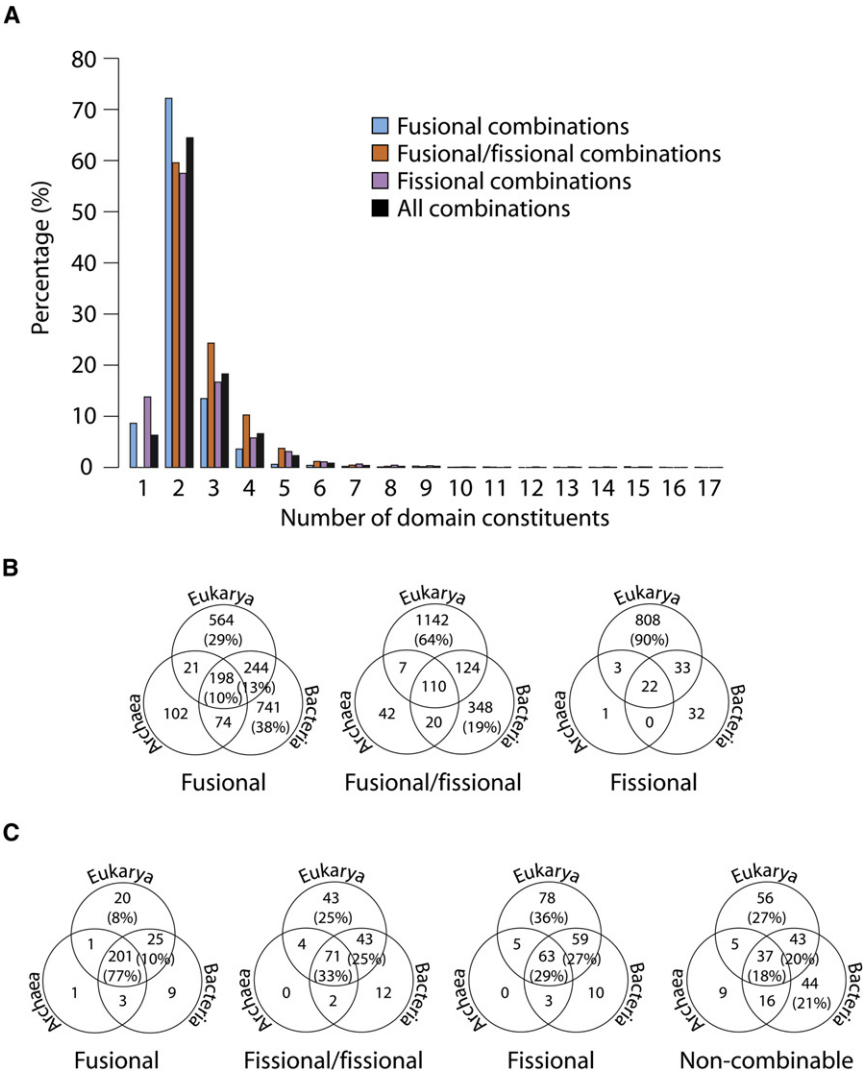


Figure 7. Patterns of Domain Use in Domain Combinations and Their Distribution in Superkingdoms of Life

(A) Bar diagram describing the average number of domains present in each of the 4,636 domain combinations analyzed according to their distribution in fusional (blue), fissional (purple), fusional/fissional (orange), and total combinations (black). Domain repeats were treated here as having only one constituent domain.

(B) Venn diagrams describing the occurrence of fusional, fusional/fissional, and fissional combinations at F level in proteomes belonging to the three superkingdoms of life.

(C) Venn diagrams showing the occurrence of fusional, fissional/fusional, fissional, and noncombinable domains.

show that fusions and fissions of domains (which can be explained by known biological phenomena) have enriched the protein world through a combinatorial game, fundamentally during an explosive phase that coincided with the creation of organismal lineages. Our results underscore the importance of modularity in evolution and reveal a cyclic pattern in the distribution of function among protein architectures (Figure 6C). This cycle began with few multifunctional domains capable of engaging effectively in fusion processes, was followed by the creation of many domain combinations with specialized function, and ended with highly multifunctional single domains arising from fission of domain combinations specific to Eukarya. The multifunctionality of the relatively few domain architectures at the onset of the protein world was probably the consequence of an exploration of structural variants within the same architectural design. In other words, ancient domains needed to accommodate the functional needs of an expanding, complex, and communal world. In contrast, late architectures exploited the diversity embedded in the "big bang" and used fission processes to produce evolutionarily derived multifunctional modules. We postulate this functional cycle relates to overarching molecular processes of diversification and unification and to adaptation to lifestyles as new niches became available for discovery on Earth.

EXPERIMENTAL PROCEDURES

A genomic census of protein architecture was conducted at F level in 266 genomes (64 Eukarya, 178 Bacteria, and 24 Archaea) and at FSF level in 536 genomes (134 Eukarya, 359 Bacteria, and 43 Archaea). Genome sequences were scanned with linear hidden Markov models (HMMs) (Gough et al., 2001) in SUPERFAMILY (Wilson et al., 2007) and structures of nonidentical SCOP 1.69 (Murzin et al., 1995) domains were assigned to proteins sequences using a probability cutoff *E* of 0.02 and boundaries of domain combinations that considered domain length distributions (Apic et al., 2001a; Liu and Rost, 2003). Genomic abundance data in demography tables were first normalized to compensate for differences in proteome representation and were then subjected to logarithmic transformation to account for unequal variance (Thiele, 1993). The data were finally coded as linearly ordered multistate phylogenetic characters and analyzed using maximum parsimony in PAUP* (Nixon, 1999; Goloboff, 1999; Swofford, 2002) and the PR search strategy (Sikes and Lewis, 2001). The structure of phylogenetic signal in the data was tested by the skewness (g_1) of the length distribution of 5×10^3 random trees. Ensemble consistency and retention indices were used to measure homoplasy and synapomorphy, confounding and desired phylogenetic characteristics, respectively. Domain architectures were categorized based on domain structure, domain organization, and the relative appearance of domain combinations and their constituent parts in the trees. The fraction of proteomes containing individual architectures (*f*) and their relative age (*nd*) were calculated for all architectures and given on a relative 0–1 scale. The functions of architectures were annotated using ProtFun 2.2 (Jensen et al., 2002, 2003) and indexed in heat maps that link function to age of architectures. Methodological details and definitions of domain architecture can be found in Supplemental Data.

SUPPLEMENTAL DATA

Supplemental data include nine figures, three tables, and Supplemental Experimental Procedures and can be found with this article online at http://www.cell.com/structure/supplemental/S0969-2126(08)00456-5.

ACKNOWLEDGMENTS

We thank Jay E. Mittenthal for encouraging discussions and Liudmila S. Yafremava for comments on the manuscript. Research was supported in part by grants from the NSF (MCB-0343126 and MCB-0749836), the C-FAR Sentinel Program, and the USDA through HATCH Illu-802-314 and the Soybean Disease Biotechnology Center. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Received: August 17, 2008
Revised: October 27, 2008
Accepted: November 13, 2008
Published: January 13, 2009

REFERENCES

Ancel, L.W., and Fontana, W. (2000). Plasticity, evolvability, and modularity in RNA. J. Exp. Zool. *288*, 242–283.

Apic, G., Gough, J., and Teichmann, S.A. (2001a). An insight into domain combinations. Bioinformatics *17* (*Suppl 3*), S83–S89.

Apic, G., Gough, J., and Teichmann, S.A. (2001b). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J. Mol. Biol. *310*, 311–325.

Apic, G., Huber, W., and Teichmann, S. (2003). Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. J. Struct. Funct. Genomics *4*, 67–78.

Aroul-Selvam, R., Hubbard, T., and Sasidharan, R. (2004). Domain insertions in protein structures. J. Mol. Biol. *338*, 633–641.

Bajaj, M., and Blundell, T. (1984). Evolution and the tertiary structure of proteins. Annu. Rev. Biophys. Bioeng. *13*, 453–492.

Bashton, M., and Chothia, C. (2002). The geometry of domain combination in proteins. J. Mol. Biol. *315*, 927–939.

Björklund, A.K., Ekman, D., Light, S., Frey-Skött, J., and Elofsson, A. (2005). Domain rearrangements in protein evolution. J. Mol. Biol. *353*, 911–923.

Brocchieri, L., and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. *33*, 3390–3400.

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

Caetano-Anollés, G., and Caetano-Anollés, D. (2005). Universal sharing in proteomes and evolution of protein fold architecture and life. J. Mol. Evol. *60*, 484–498.

Caetano-Anollés, G., Kim, H.S., and Mittenthal, J.E. (2007). The origins of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc. Natl. Acad. Sci. USA *104*, 9358–9363.

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

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. *6*, 361–375.

Ekman, D., Björklund, Å.K., and Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. J. Mol. Biol. *372*, 1337–1348.

Enright, A., Iliopoulos, I., Kyrpides, N., and Ouzounis, C. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86–90.

Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. (2006). Pfam: clams, web tools and services. Nucleic Acids Res. *34*, D247–D251.

Fong, J.H., Geer, L.Y., Panchenko, A.R., and Bryant, S.H. (2007). Modeling the evolution of protein domain architectures using maximum parsimony. J. Mol. Biol. *366*, 307–315.

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

Fukami-Kobayashi, K., Minezaki, Y., Tateno, Y., and Nishikawa, K. (2007). A tree of life based on protein domain organizations. Mol. Biol. Evol. *24*, 1181–1189.

Funke, L., Dakoji, S., and Bredt, D.S. (2005). Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. Annu. Rev. Biochem. *74*, 219–245.

Goloboff, P. (1999). Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics *15*, 415–428.

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

Gough, J., Karplus, K., Hughey, R., and 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.

Grant, A., Lee, D., and Orengo, C. (2004). Progress towards mapping the universe of protein folds. Genome Biol. *5*, 107.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47–C52.

Jeltsch, A. (1999). Circular permutations in the molecular evolution of DNA methyltransferases. J. Mol. Evol. *49*, 161–164.

Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H.H., Rapacki, K., Workman, C., et al. (2002). Prediction of human protein function from post-translational modifications and localization features. J. Mol. Biol. *319*, 1257–1265.

Jensen, L.J., Gupta, R., Stærfeldt, H.H., and Brunak, S. (2003). Prediction of human function according to Gene Ontology categories. Bioinformatics *19*, 635–642.

Kaessmann, H., Zöllner, S., Nekrutenko, A., and Li, W.H. (2002). Signatures of domain shuffling in the human genome. Genome Res. *12*, 1642–1650.

Kim, H.S., Mittenthal, J.E., and Caetano-Anollés, G. (2006). MANET: tracing evolution of protein architecture in metabolic networks. BMC Bioinformatics *7*, 351.

Kirkpatrick, M., and Slatkin, M. (1993). Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution Int. J. Org. Evolution *47*, 1171–1181.

Kummerfeld, S.K., and Teichmann, S.A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. *21*, 25–30.

Kurland, C.G., Canbäck, B., and Berg, O.G. (2007). The origins of modern proteomes. Biochimie *89*, 1454–1463.

Liu, J., and Rost, B. (2003). Domains, motifs and clusters in the protein universe. Curr. Opin. Chem. Biol. *7*, 5–11.

Marcotte, E., Pellegrini, M., Yeates, T., and Eisenberg, D. (1999). A census of protein repeats. J. Mol. Biol. *293*, 151–160.

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

Moore, A.D., Björklund, Å.K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. (2008). Arrangements in the modular evolution of proteins. Trends Biochem. Sci. *33*, 444–451.

Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. Science *283*, 1530–1534.

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

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

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997). CATH: a hierarchic classification of protein structure. Structure *5*, 1093–1098.

Pal, L.R., and Guda, C. (2006). Tracing the origin of functional and conserved domains in the human proteome: implications for protein evolution at the modular level. BMC Evol. Biol. *6*, 91.

Pasek, S., Rister, J.-L., and Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics *22*, 1418–1423.

Patthy, L. (1999). Genome evolution and the evolution of exon shuffling—a review. Gene *238*, 103–114.

Schmidt, E.E., and Davies, C.J. (2007). The origins of polypeptide domains. Bioessays *29*, 262–270.

Shiu, S.-H., and Li, W.-H. (2004). Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in Eukaryotes. Mol. Biol. Evol. *21*, 828–840.

Sikes, D.S., and Lewis, P.O. (2001). PAUPRat: PAUP Implementation of the Parsimony Ratchet, Version 1 (computer program). Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut.

Sun, F.-J., and Caetano-Anollés, G. (2008). Evolutionary patterns in the sequence and structure of transfer RNA: early origins of Archaea and viruses. PLoS Comput. Biol. *4*, e1000018.

Swofford, D.L. (2002). Phylogenetic Analysis Using Parsimony and Other Programs (PAUP*), Version 4 (Sunderland, MA: Sinauer Associates).

Szathmáry, E., and Smith, J.M. (1995). The major evolutionary transitions. Nature *374*, 227–232.

Taylor, W.R. (2007). Evolutionary transitions in protein fold space. Curr. Opin. Struct. Biol. *17*, 354–361.

Teichmann, S.A., Rison, S.C.G., Thornton, J.M., Riley, M., Gough, J., and Chothia, C. (2001). Small-molecule metabolism: an enzyme mosaic. Trends Biotechnol. *19*, 482–486.

te Velthuis, A.J.W., Admiraal, J.F., and Bagowski, C.P. (2007). Molecular evolution of the MAGUK family in metazoan genomes. BMC Evol. Biol. *7*, 129.

Thiele, K. (1993). The holy grail of the perfect character: the cladistic treatment of morphometric data. Cladistics *9*, 275–304.

Vibranovski, M., Sakabe, N.J., de Oliveira, R.S., and de Souza, S.J. (2005). Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. J. Mol. Evol. *61*, 341–350.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A. (2004a). Structure, function and evolution of multidomain proteins. Curr. Opin. Struct. Biol. *14*, 208–216.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., and Teichmann, S.A. (2004b). Supra-domains: evolutionary units larger than single protein domains. J. Mol. Biol. *336*, 809–823.

Vogel, C., Teichmann, S.A., and Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. J. Mol. Biol. *346*, 355–365.

Wang, M., and Caetano-Anollés, G. (2006). Evolution inferred from domain combination in proteins. Mol. Biol. Evol. *23*, 2444–2454.

Wang, M., Boca, S.M., Kalelkar, R., Mittenthal, J.E., and 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, L.S., Caetano-Anollés, D., Mittenthal, J.E., and Caetano-Anollés, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res. *17*, 1572–1585.

Weiner, J., III, Beaussart, F., and Bornberg-Bauer, E. (2006). Domain deletions and substitutions in modular protein evolution. FEBS J. *273*, 2037–2047.

Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J. (2007). The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res. *35*, D308–D313.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. Mol. Biol. Evol. *18*, 1694–1702.

Yanai, I., Derti, A., and DeLisi, C. (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc. Natl. Acad. Sci. USA *98*, 7940–7945.

Yang, S., Doolittle, R.F., and Bourne, P.E. (2005). Phylogeny determined by protein domain content. Proc. Natl. Acad. Sci. USA *102*, 373–378.

