Global Phylogeny Determined by the Combination of Protein Domains in Proteomes

Minglei Wang and Gustavo Caetano-Anollés[1]

Department of Crop Sciences, University of Illinois at Urbana-Champaign

The majority of proteins consist of multiple domains that are either repeated or combined in defined order. In this study, we survey the combination of protein domains defined at fold and fold superfamily levels in 185 genomes belonging to organisms that have been fully sequenced and introduce a method that reconstructs rooted phylogenomic trees from the content and arrangement of domains in proteins at a genomic level. We find that the majority of domain combinations were unique to Archaea, Bacteria, or Eukarya, suggesting most combinations originated after life had diversified. Domain repeat and domain repeat within multidomain proteins increased notably in eukaryotes, mainly at the expense of single-domain and domain-pair proteins. This increase was mostly confined to Metazoa. We also find an unbalanced sharing of domain combinations which suggests that Eukarya is more closely related to Bacteria than to Archaea, an observation that challenges the widely assumed eukaryote–archaebacterial sisterhood relationship. The occurrence and abundance of the molecular repertoire (interactome) of domain combinations was used to generate phylogenomic trees. These global interactome-based phylogenies described organismal histories satisfactorily, revealing the tripartite nature of life, and supporting controversial evolutionary patterns, such as the Coelomata hypothesis, the grouping of plants and animals, and the Gram-positive origin of bacteria. Results suggest strongly that the process of domain combination is not random but curved by evolution, rejecting the null hypothesis of domain modules combining in the absence of natural selection or an optimality criterion.

Introduction

Except for some disordered proteins, most proteins consist of one or more domains that fold into compact 3-dimensional (3D) structures known as protein folds. Domains are not only units of protein structure and function but also units of evolution (Murzin et al. 1995; Orengo et al. 1997; Riley and Labedan 1997). Taxonomies that attempt to provide a comprehensive description of structural and evolutionary relationship of proteins of known structure, such as the Structural Classification of Proteins (SCOP) (Murzin et al. 1995), use these building blocks as units of classification. In SCOP, proteins that are evolutionarily closely related at the sequence level (generally with $>30\%$ amino acid residue identities) are clustered together into protein families. Proteins belonging to different families that exhibit low sequence identities but share structural and functional features suggesting a common evolutionary origin are further unified into fold superfamilies. Finally, fold superfamilies sharing secondary structures that are similarly arranged and topologically connected are unified into protein folds (Murzin et al. 1995).

The number of available domains appears finite (Chothia 1992, Wolf et al. 2000) and so does the repertoire of domain combinations in proteins (Apic et al. 2001a). The majority of proteins are composed of multiple domains that are either repeated or combined in defined order. When creating new functions, redundancy appears to be a favored outcome. Consequently, protein building blocks are reused more often than discovered (Apic et al. 2001b). However, the topology of domain combinations is highly conserved. The orientation of domains and the type of neighboring domains in proteins are limited (Apic et al. 2001a; Bashton and Chothia 2002). For example, domain permutation may not materialize in all protein variants. These constraints define a molecular "interactome," that is, a collection of possible topologies depicting intramolecular interactions between protein domains. Because molecular interactions in protein architectures define functions that can impact fitness, this molecular interactome should be curved by natural selection and evolution.

In this paper, we survey and analyze interactomes phylogenetically at the genome level. The recent availability of an exponentially increasing number of whole-genome sequences has opened new possibilities in the study of evolution (Delsuc et al. 2005). Evolutionary history has been reconstructed using combined or concatenated genomic sequences (e.g., Baldauf et al. 2000; Brown et al. 2001; Ciccarelli et al. 2006) and genomic features describing the survey (genomic demography) (Gerstein 1998; Gerstein and Hegyi 1998; Snel et al. 1999; Tekaia et al. 1999; Wolf et al. 1999; House and Fitz-Gibbon 2002; Lin and Gerstein 2000; Wolf et al. 2002; Caetano-Anollés G and Caetano-Anollés D 2003; Dutilh et al. 2004; Yang et al. 2005) and arrangement (genomic topography) (Dandekar et al. 1998; Wolf et al. 2001, 2004; Korbel et al. 2002) of genomic component parts. In particular, phylogenomic (whole-genome) trees were built from features describing the occurrence and distribution of protein folds in proteomes (Gerstein 1998; Gerstein and Hegyi 1998; Wolf et al. 1999; Lin and Gerstein 2000; Caetano-Anollés G and Caetano-Anollés D 2003; Deeds et al. 2005; Yang et al. 2005). For example, we measured the popularity (number of occurrences) of each protein fold in sequenced genomes and used multistate phylogenetic characters to reconstruct intrinsically rooted proteome trees invoking the concept that being popular at the molecular level is a favored evolutionary outcome (Caetano-Anollés G and Caetano-Anollés D 2003, 2005). We here take this approach a step further and introduce a method that reconstructs phylogenies from features describing the content and arrangement of domains in proteins at a genomic level. Phylogenetic characters are therefore drawn from a molecular topography that describes how evolutionary units of structure arrange in protein molecules and how popular these arrangements are within each proteome. The reconstructed pan-domain

[1] Present address: 332 National Soybean Research Center, 1101 W. Peabody Drive, Urbana, IL.

Key words: protein domains, evolution, phylogenomics, combinations.

E-mail: gca@uiuc.edu

Mol. Biol. Evol. 23(12):2444–2454. 2006
doi:10.1093/molbev/msl117
Advance Access publication September 13, 2006

phylogenies confirm accepted lineage relationships within major organismal groups, support disputed or preliminary classifications, and reveal interesting evolutionary patterns.

Materials and Methods

Analysis of Domain Combinations

Domains are here defined according to SCOP (Murzin et al. 1995). SCOP release 1.67 classifies 24,037 Protein Data Bank (PDB) entries into 65,122 protein domains, which are then grouped into 2,630 domain families, 1,447 fold superfamilies, and 887 protein folds. Structural domains were assigned to proteins belonging to individual genomes at the fold superfamily level using linear hidden Markov models (HMMs) in SUPERFAMILY (version 1.67; http://www.supfam.org/SUPERFAMILY/) (Gough et al. 2001). This architectural hierarchy pools proteins for which there is structural and sequence evidence of a common evolutionary ancestor. Superfamilies were later assigned to folds using SCOP 1.67. Genome sequences were scanned against a linear HMM library generated using the iterative Sequence Alignment and Modeling (SAM) system method (http://www.cse.ucsc.edu/research/compbio/sam.html). Each model in the library (generated by SAM-T02) identifies each nonidentical SCOP domain. The HMM searching protocol uses a probability cutoff E of 0.02. Differences in topologies of trees reconstructed with more stringent cutoff values were found negligible (Yang et al. 2005). Consequently, we did not explore the role of this parameter. An internal calibration of the accuracy of HMM prediction against PDB sequence records in the ASTRAL compendium (http://astral.berkeley.edu) (Brenner et al. 2000) correctly identified 98% sequences from 21,173 enzyme-associated PDB entries (http://manet.uiuc.edu/download.php).

When surveying domain combinations, we considered the number of residues between 2 domains defined at fold or fold superfamily levels. We chose a threshold of 30 intervening residues to claim 2 domains were adjacent to each other. If a longer unmatched region was present between domains, we considered that the intervening region harbored an unknown neighboring domain. The 30-residue threshold level makes it unlikely that a transmembrane region or SCOP domain be present in the intervening sequence (Apic et al. 2001a). Domain combinations were described using the notation illustrated in the following example. Given the domain combination A|B-C=D, where "A, B, C, and D" stand for different fold or fold superfamily domains, connecting marks "|," "-," and "=" indicate there are no residue, $\leq$30 residues or $>$30 residues between adjacent domains, respectively. We wrote a computer script to scan every protein sequence of every organism fully sequenced to which structural domains had been assigned by using HMMs. Sequences were then transformed into expressions consisting of domain identifiers and connecting marks according to the rules described above. These expressions were compared with one another strictly, that is, expressions showing differences were considered as different expressions (e.g., domain combinations A|B-C, A-B|C, A-B-C, and A|B|C were all considered different). For example, the analysis of protein sequence ENSPTRP00000035659 in *Pan troglodytes* assigned the Chaperone J-domain (SCOP id 46565) to the region from the 5th to 70th residue, the DnaJ/Hsp40 cysteine-rich domain (SCOP id 57938) to the region from the 121st to 204th residue, and the HSP40/DnaJ peptide-binding domains (SCOP id 49493) to the regions from the 105th to 137th, from 204th to 248th, and from 256th to 330th residues. The expression describing ENSPTRP00000035659 was 46565=49493|57938|49493-49493. Expressions were classified into 5 categories of domain combinations (single domain, single domain in multidomain, domain repeat, domain repeat in multidomain, and domain pair; for a description, see Results) on a per-domain basis in which each domain contributes a score to each category depending on the domains that are combined with it. For example, given the expression A|B-B, A contributes 1 to the single domain in multidomain category and each B domain contributes 1 to the domain repeat in multidomain category. Categories are therefore tallied globally relative to the constituting domains of proteins analyzed.

In this study, "=" stands for a sequence region that is longer than 30 residues and does not have any assigned structural domain in the current data set, regardless of its length. Though it is unlikely that a long region with no assigned domain(s) will be unstructured, such regions become less frequent as they increase in length (Fig. S1 of Supplementary Material online) and their influence in the analysis diminishes considerably. A total of 229,786 intervening regions were longer than 30 residues. This represents 46.5% of all intervening regions identified. Out of these, about half (54%) were shorter than 90 residues and allowed space for only 1, or perhaps 2 additional putative domains. This is because almost all SCOP domains are longer than 30 residues and are generally no more than about 150 residues long (Liu and Rost 2003). The fact that HMMs are unable to match intervening sequences can be explained by the existence of unstructured protein regions or putative domains that have not been yet characterized (Apic et al. 2001a, 2001b). They can also be explained by pronounced structural differences of domains at fold and fold superfamily levels, in which domain cores harbor peripheral structural regions of weak sequence conservation that cannot be detected by homology modeling (Grishin 2001). These peripheral structures could occupy extensive regions of the intervening sequence and could be responsible for known difficulties in predicting domain boundaries (Liu and Rost 2003).

Character Coding and Argumentation

The frequencies with which individual domain combinations occur in genomes, termed domain combination abundance (G_{ab}), were used to describe at global levels the popularity of combinations of individual domains. Note that the sum of G_{ab} of a particular domain combination (i.e., expression) over all genomes is an indicator of how popular is this kind of domain combination in nature. In order to obtain G_{ab} values, we retrieved all expressions defining domain combinations and counted the frequency with which every expression occurred in individual genomes.

We analyzed the finished genome sequence of 185 organisms, encompassing 19 Archaea, 129 Bacteria, and

37 Eukarya (Table S1 of Supplementary Material online) and reconstructed phylogenies describing their evolution. Out of 1,015,140 protein-encoding sequences, an average of 49.0 ± 0.1% (SD) and a median of 52% entries could be assigned to structures. Assignments range from 15% in *Plasmodium falciparum* to 71% in *Blochmannia floridanus* (Table S2, Supplementary Material online). We also found that the structural HMM-based census was more effective when small prokaryotic genomes were surveyed (Fig. S2, Supplementary Material online).

Because the G_{ab} number of a particular domain combination in different genomes distributed widely, G_{ab} was normalized to compensate for differences in genome size and proteome representation and was subjected to logarithmic transformation to account for unequal variances. The following formula was used to do the transformation,

$$G_{ab\_norm} = \text{Round}\left[\frac{\ln(G_{ab}+1)}{\ln(G_{ab\_max}+1)}\times 20\right],$$

where G_{ab_norm} represents the G_{ab} value after normalization and G_{ab_max} the largest G_{ab} value of a particular domain combination in all genomes. "Round" and "ln" stand for rounding function and logarithm function, respectively.

Finally, the data were range standardized to a 0–20 scale, treated as linearly ordered multistate phylogenetic characters using an alphanumeric format with numbers 0–9 and letters A–K, aligned in ordered columns, encoded in the NEXUS format, and subjected to phylogenetic analysis. Characters are observable features that distinguish one object from another and constitute hypotheses of primary homology. They can display multiple numerical values and frequency distribution of values called character states. The ANCSTATES command was used to polarize characters assuming that the number of protein representatives in a genome exhibiting a particular fold or fold combination increases in the course of evolution. We consider that organisms with lineages that originated early in evolution have fewer folds and fold combinations than organisms that appeared late and that the number of folds and fold combinations increases in single steps corresponding to the addition or removal of a homologous gene in a family. Support for character argumentation has been described previously (Caetano-Anollés G and Caetano-Anollés D 2003, 2005).

We decided to include characters describing domain combination in proteins with intervening sequences of more than 30 residues. Their inclusion or exclusion could bias the analysis. Inclusion of these domain combinations ignores possible domains and structural elements embedded in the intervening sequence. If these putative structures are variable in different proteins, characters describing individual domain combinations will need to be split into alternative characters and this will bias phylogenetic analysis. Exclusion of domain combinations with long intervening sequences results in loss of phylogenetic information and can also bias phylogenetic reconstruction. Domain structures that have not been discovered to date are probably of low genomic abundance and are expected to be highly diverse (Gerstein and Hegyi 1998). Excluding proteins that contain these domains would favor representation of highly popular structures in domain combinations, which are generally old (Caetano-Anollés G and Caetano-Anollés D 2003). Future advances in structural genomics will help fill structural "gaps" in proteins with domain combinations and will decrease the bias introduced by unassigned domains and structural elements.

We also reconstructed phylogenomic trees of proteomes based simply on the occurrence (G) of domain combinations in proteomes, that is, the existence or absence of a domain combination in the protein complement of organisms. Genome occurrence was encoded as a simple binary character state vector at protein fold and fold superfamily levels. Under this cladistic scheme, presence and absence of a domain combination in the proteome of an organism was coded as 1 or 0, respectively, and binary data matrices were analyzed with maximum parsimony and distance methods.

Data matrices with characters describing genomic occurrence and abundance can be found in Supplementary Material online.

Phylogenetic Analysis

Phylogenetic trees were reconstructed using heuristic searches and maximum parsimony as the optimality criterion in PAUP* (Swofford 2003). Generally, 100 heuristic searches were initiated using random addition starting taxa, with Tree Bisection-Reconnection branch swapping and multrees selected. One shortest tree was saved from each search. Phylogenetic reliability was evaluated by the bootstrap method with 5,000 pseudoreplicates, generally using "fast" stepwise addition of taxa. The structure of phylogenetic signal in the data was tested by the skewness (g_1) of the length distribution of 10^4 random trees. Ensemble consistency index and retention index were used to measure homoplasy and synapomorphy, confounding and desired phylogenetic characteristics, respectively. Phylogenetic trees were also reconstructed using the Neighbor-Joining (NJ) method (Saitou and Nei 1987).

Results

Categories of Domain Combinations and Their Distribution in Organismal Domains

We investigated the distribution of domain combinations in the proteomes of 185 organisms belonging to Archaea, Bacteria, and Eukarya that have been completely sequenced. Domains in proteins defined at the fold superfamily level were divided into 5 categories: 1) "single domain," in which only one domain is present in a peptide sequence, 2) "single domain in multidomain," in which one domain is present in a peptide sequence together with 2 or more types of domains, 3) "domain repeat," in which more than one domain of the same kind is present in a peptide sequence, 4) "domain repeat in multidomain," in which a domain repeat is present in a peptide sequence together with more than 2 types of domains, and 5) "domain pair," in which the peptide sequence contains only 2 different domains (fig. 1). A total of 22,902, 220,186, and 336,562 proteins in genomes belonging to Archaea, Bacteria, and Eukarya were analyzed, respectively. The different categories of domain combinations distributed differently within organismal domains (fig. 1). In Bacteria and

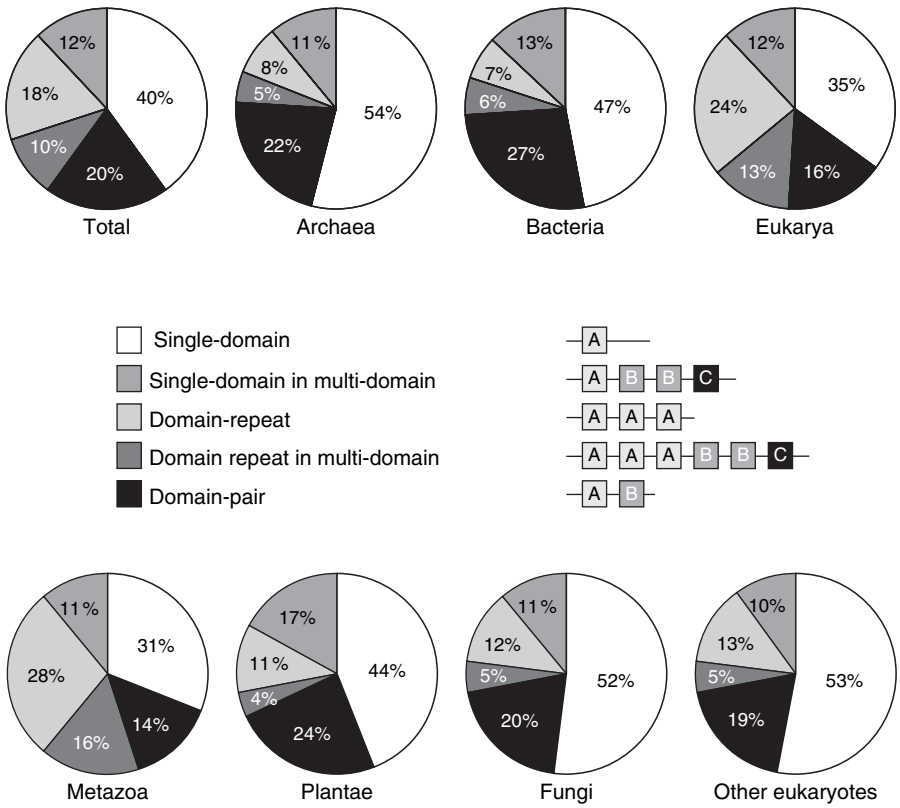

FIG. 1.—Distribution of superfamily domain combination categories in all 185 proteomes or proteomes belonging to Eukarya, Bacteria, and Archaea, as well as Metazoa, Plantae, Fungi and basal unicellular eukaryotes (other eukaryotes). Categories are defined relative to individual domains. Example domain combinations that are illustrated are categorized relative to domain A. Figure S11 in Supplementary Materials online shows a color version of this figure.

Archaea, single-domain and domain-pair categories were predominant, whereas in Eukarya, domain repeat and domain repeat in multidomain categories increased at the expense of single domain and single domain in multidomain categories. We also analyzed how categories distributed among major phyla in Eukarya (fig. 1). The distribution of categories in Metazoa resembled distribution patterns in Eukarya, whereas patterns in Plantae, Fungi, and basal unicellular eukaryotes resembled prokaryotic patterns.

When surveying domain combinations, we considered 2 domains being adjacent when separated from each other by less than 30–amino acid residues. This threshold level minimizes the likelihood of obtaining incorrect combinations (Apic et al. 2001a). According to the above definition, we found 32,001 and 35,559 different domain combinations present in the 185 proteomes analyzed, defined at fold and fold superfamily levels, respectively.

Multidomains can also be analyzed by defining sets of pairwise domain combinations from N- to C-terminus along the polypeptide chain, treating multidomain proteins as arrangements of domain pairs. All domain combinations reported at the fold superfamily level in the SUPERFAMILY database, including the multidomains mentioned above, are pairwise, that is, multidomain combination A-B-C are considered as 2 pairwise combinations, A-B and B-C. By parsing SUPERFAMILY domain combination files, we found 6,460 pairwise fold combinations, including domain pairs representing 2 different domain combinations and domain repeats representing combination of a same domain.

Sharing of Domain Combinations at Protein Fold and Fold Superfamily Levels Between Organismal Domains

A total of 776 folds and 1,259 fold superfamilies present were examined. Venn diagrams showed the distribution of domain combinations (as different combinations in proteins or as pairwise relationships), folds, or fold superfamilies among organismal domains (fig. 2 and Table S2, Supplementary Material online). These distributions exhibited opposing trends. Domain combinations that were unique to Archaea, Bacteria, or Eukarya were more prevalent than those that were shared. In contrast, folds and fold superfamilies shared by organismal domains were more prevalent than those that were unique. For example, about 85% of the 6,460 pairwise combinations at fold level were specific to organismal domains (58% for Eukarya, 24% for Bacteria, and 3% for Archaea). Conversely, more than 65% of the 776 folds and 62% of the 1,259 fold superfamilies were common to all domains of life. An analysis of the distribution of domain combinations among organismal domains showed that the numbers of domain combinations shared by Bacteria and Eukarya were consistently at least 10-fold higher than those shared by Archaea and Eukarya. Similarly, the numbers of domain combinations shared by Bacteria and Eukarya were consistently 5-fold higher than those shared by Archaea and Bacteria. Similar trends were found when studying distributions of fold and fold superfamily architectures.

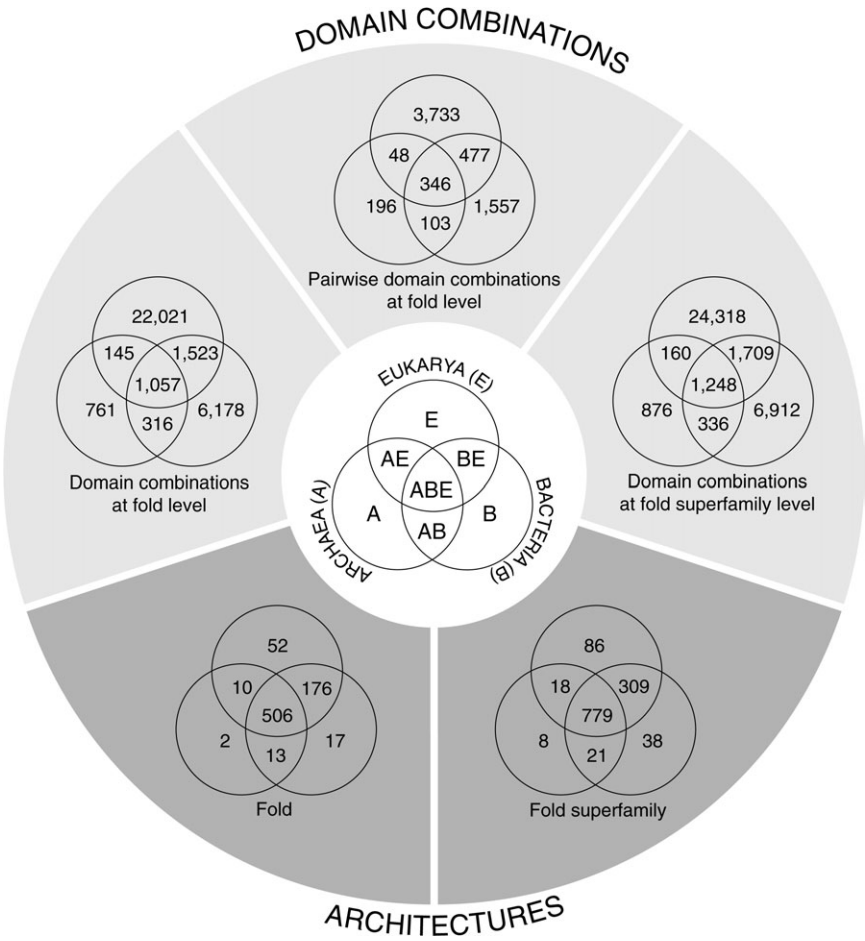

FIG. 2.—Venn diagrams showing the occurrence of domain combinations at fold and fold superfamily levels and of fold and fold superfamily architectures in proteomes belonging to the 3 major organismal domains.

Bias in the Permutation of Domains

In all the 6,460 pairwise domain combinations examined at the fold level, not all combinations exhibited the 2 possible orders of domains in the peptide sequence, that is, not all N-terminal and C-terminal domain permutations of possible domain pairs were observed. In fact, only 1,958 domain combinations (30%) showed both orders of domains, and these involved 979 domain pairs defined at the fold level, without taking into consideration the N- to C-terminal orientation. Moreover, the proportions of proteins exhibiting the 2 orders (as a permutation ratio) varied considerably among the 1,958 combinations (fig. 3). About 26% of these combinations showed permutation ratios that varied from 1251 to 10, about 57% of the combinations had ratios that varied from 10 to 1, and only 17% of the combinations showed absence of bias in domain orders, that is, the ratios were one. The frequency distribution of the permutation ratio over all pairwise domain combinations appears to follow a power distribution (fig. 3, inset).

Phylogenomic Trees Based on the Combination of Domains in Proteins at Fold and Fold Superfamily Levels

We reconstructed phylogenomic trees of proteomes based on the repertoire of domain combinations at protein fold and fold superfamily levels. Figure S3 (Supplementary Material online) shows a single most parsimonious tree reconstructed from the genomic abundance of different domain combinations in proteins at the fold superfamily level. The tree was well resolved and relatively well supported. Within the proteome tree, as many as 54% of branches were supported by at least 80% bootstrap proportions (BP). In fact, 28% were supported at 100% level. Poorly supported regions in the tree occurred at deep and intermediate taxonomical levels. Tree distribution profiles and metric of skewness were suggestive of strong cladistic structure ($P < 0.01$). The tree showed the tripartite nature of the living world, with well-supported clades defining Archaea (94% BP) and Eukarya (100% BP). One notable exception was *Nanoarchaeum equitans*, the archaeon that establishes parasitic relationships within Archaea and has a highly reduced genome. This proteome fell at the base of the eukaryal clade incorrectly. The tree was rooted in Eukarya, showing that Archaea and Bacteria were sister groups. However, support for this grouping was weak. The majority of character state change occurred in Eukarya, particularly in Metazoa, as shown by the long branch lengths associated with this segment of the tree. As many as 72% branches of the eukaryal subtree were supported at 80% BP. The tree reconstructed from domain combination at protein fold

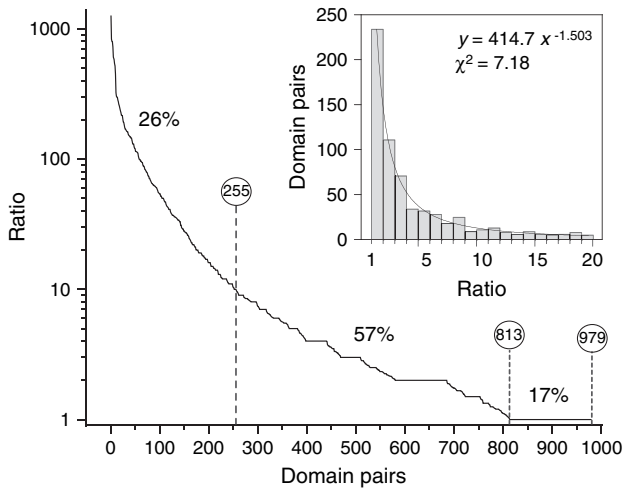

FIG. 3.—Domain permutation ratios of domain pairs sorted in descending order. The histogram summarizes the frequency distribution of the permutation ratio (binned to reduce noise) over all pairwise domain combinations.

level showed a very similar topology, indicating minimal bias related to the hierarchical level of structure used to define the domain in combinations (Fig. S4, Supplementary Material online).

Phylogenetic Trees Based on Pairwise Combinations of Domains at Fold Level

We also reconstructed a phylogenomic tree of proteomes based on the repertoire of pairwise combinations of domains at fold level (Fig. S5, Supplementary Material online). Within the proteome tree, as many as 41% of branches were supported by at least 80% BP; 26% were supported at 100% level. Tree distribution profiles and metrics of skewness were suggestive of strong cladistic structure ($P < 0.01$). The tree revealed well-supported clades defining Archaea (91% BP) and Eukarya (82% BP). In this case, *N. equitans* fell within the archaeal clade correctly. Separate phylogenetic analyses of proteomes belonging to individual organismal domains showed the same general patterns observed in the global tree (Fig. S6, Supplementary Material online).

Phylogenetic Trees Based on Occurrence (Presence–Absence) of Domain Combinations and Pairwise Domain Combinations at Fold and Fold Superfamily Levels

Phylogenetic reconstructions described above use linearly ordered multistate characters that measure the abundance of domain combinations in proteomes. An alternative approach is to use binary characters that describe only the occurrence of individual domain combinations (Yang et al. 2005). Phylogenomic trees of proteomes were therefore reconstructed based on the presence or absence of domain combinations in proteomes at fold superfamily and fold levels (Figs. S7 and S8, Supplementary Material online) and the presence or absence of pairwise domain combinations at fold levels (Fig. S9, Supplementary Material online). However, we did not find major differences in the topologies and bootstrap support of branches of these trees when compared with those reconstructed from abundance of domain combinations. Except for minor differences in the placement of taxa within clades (e.g., *Spirochaetes*, fungi), trees were mostly congruent. Phylogenies again showed the tripartite nature of life and considerable character state change associated with Eukarya, particularly with Metazoa. Nevertheless, trees reconstructed from genomic occurrence of domain combinations at fold superfamily level showed that Archaea and Eukarya were sister groups (Fig. S7, Supplementary Material online), whereas those reconstructed at fold level (Fig. S8, Supplementary Material online) showed the sisterhood of Archaea and Bacteria that is typical of trees reconstructed from genomic abundance. Note, however, that support for all these groupings was weak.

Reconstruction of Global Phylogenies Using Distance Methods

We also reconstructed phylogenies using the computationally less demanding NJ distance method. For example, Figure S10 (Supplementary Material online) shows a phylogenomic tree reconstructed from occurrence of domain combinations at fold superfamily level. In this tree, the sisterhood of Archaea and Bacteria was well supported (89% BP), but the tree in general showed lower bootstrap support values and some topological differences, especially within Bacteria. Overall, distance and character state methods produced well-resolved trees that were concordant. However, maximum parsimony appears an appropriate criterion for an initial study of architectural evolution (see discussion in Caetano-Anollés G and Caetano-Anollés D 2005) and was chosen here to illustrate our phylogenomic approach.

Phylogenomic Patterns

Phylogenomic trees of proteomes reconstructed from the repertoire of domain combinations in proteins defined at fold and fold superfamily levels or from the repertoire of pairwise combination of domains showed common evolutionary patterns summarized in the circle consensus tree described in figure 4. Analyses of tree-length distributions show that character data from domain combinations in proteins were more structured ($g_1 = -1.038$ and -1.033, for fold and fold superfamily, respectively) than character data derived from pairwise combinations of domains ($g_1 = -0.878$), and these were significantly more structured than random data. However, these approaches produced trees with similar resolution and clade support, despite notable differences in the number of characters resulting from each coding scheme (Figs. S3–S5, Supplementary Material online).

1. Eukarya: Major clades consisted of animal (100% BP), plant (100% BP), and fungal proteomes, respectively. Animals and plants were sister taxa (53–97% BP). Within the animal clade, proteomes were well positioned except for amphibian (*Xenopus tropicalis*), normally assumed more derived than fish (*Fugu rubripes*, *Danio rerio*). This may be partly due to the preliminary nature of its genome sequence (for a detailed discussion,

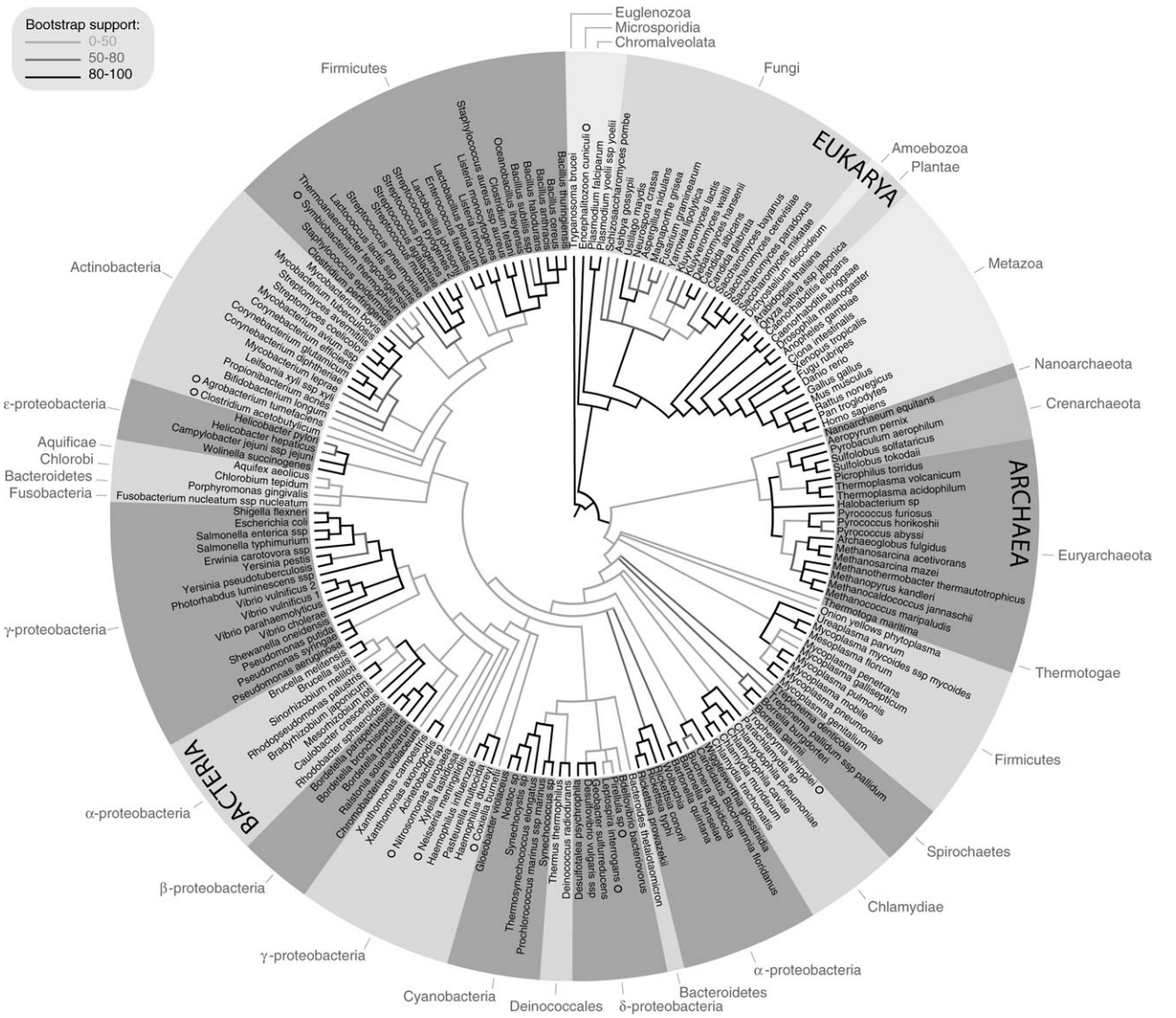

FIG. 4.—Global phylogeny of fully sequenced organisms. The circle tree describes a consensus tree generated from the abundance of domain combination in proteins using 3 different coding strategies. Labels and color shadings indicate frequently used organismal subdivisions. BP support levels for branches are indicated with different shades. Open circles denote anomalously positioned taxa. Figure S12 in Supplementary Materials online shows a color version of this figure.

see Yang et al. 2005). The clade consisting of Arthropoda (*Drosophila* and *Anopheles*; 94–99% BP) was sister to Deuterostomia (chordate lineages; 92–96% BP) with supports of 76–99% BP. Amebozoa with its single entry (*Dictyostelium discoideum*) was derived to fungi but ancestral to plants and animals (92–98% BP). As expected, proteomes from unicellular eukaryotes appeared at the base of the tree.

2. Archaea: The archaeal proteomes were generally divided into 2 groups, the Crenarchaeota (4 taxa) and the Euryachaeota (14 taxa), with *N. equitans* generally appearing at the base of the archaeal tree and close to Crenarchaeota. The methanogens, the 3 *Pyrococci* (100% BP), and the 3 *Thermoplasmata* (99% BP) fell into single clades and were satisfactorily positioned.
3. Bacteria: Our study included 129 bacterial proteomes belonging to 13 different classical phyla, including 5 classes of *Proteobacteria* and 4 classes of *Firmicutes*. Many of the bacteria had parasitic lifestyles and greatly reduced genomes, features known to bias whole-genome analyses (Yang et al. 2005). Proteomes from *Cyanobacteria* (80–97% BP), *Bacteroidetes* (92–100% BP), *Chlamydiae* (99–100% BP), *Actinobacteria*, and the *Spirochaetes* formed distinct clades. The *Cyanobacteria* and *Deinococcales* formed a clade, but it was poorly supported. The Firmicutes formed 2 separate groups, one basally located in the tree containing 9 species of Mollicutes (72–90% BP), which were closely associated and derived to onion yellow phytoplasma (52–53% BP), and the other derived, with *Bacillales*, *Lactobacillales*, and *Clostridia* (with clades containing *Bacillus* [91–99% BP], *Staphylococcus* [96% BP], *Streptococcus* [82–98% BP], and *Clostridiae* [78% BP]) and 2 unrelated taxa (*Symbiobacterium* and *Thermotoga*). The *Proteobacteria* was generally split into a major clade containing α-, β-, and γ-proteobacteria

and small clades containing the ε-proteobacteria (95–100% BP) and the δ-proteobacteria (sometimes associated with the spirochaete *Leptospira*). The phylogeny of pairwise domain combinations oddly placed *Agrobacterium* at the base of the tree. In turn, the phylogeny of domain combinations placed *Thermotoga* at the base of the bacterial clade. Except for the split of *Firmicutes* and *Proteobacteria*, the overall topology of the tree of Bacteria was reasonable.

Discussion

We here survey the combination of protein domains in 185 genomes belonging to organisms in Eukarya, Bacteria, and Archaea that have been fully sequenced. We then use the repertoire of domain combinations to reconstruct phylogenomic trees that describe histories of organismal diversification. These pan-domain phylogenies showed the tripartite nature of life, correctly grouped major organismal lineages within Archaea and Eukarya, and generally clustered major bacterial groups of organisms appropriately. To our knowledge, this is the first time that a molecular interactome is used to draw global inferences about organismal evolution.

Protein domains combine in defined patterns depending on the number of domains involved and the existence of tandem repeats (Apic et al. 2001a). Domains belonging to one family can be found repeated in "domain-repeat proteins." These proteins may have evolved from single-domain proteins by recombination or by internal duplication to form tandem repeats. They account for only 7–8% of prokaryotic proteomes, but their number is substantial in eukaryotes (24%), especially in Metazoa (28%) (fig. 1). Domains that combine with others can do this with a single partner in domain-pair proteins or with multiple partners in "multidomain proteins." Moreover, domains combining with multiple partners can also be repeated in domain-repeat patterns within the multidomain proteins. Domain-pair proteins account for 20% of domains, their relative number being higher in prokaryotes than in eukaryotes. In contrast, domain-repeat proteins by themselves or within multidomains proteins account for 13% of proteins in prokaryotes and 37% in eukaryotes. The fact that domain repeats by themselves or combined with multidomain proteins increased notably in eukaryotes, especially in Metazoa, mainly at the expense of the single-domain and domain-pair categories, is of importance. The larger unicellular and multicellular eukaryotic organisms have domains that combine more often and do so with multiple partners, perhaps to fulfill more complicated functions. A large-scale survey of annotation transfer in multidomain proteins showed significantly less conservation of function when these are compared with single-domain proteins (Hegyi and Gerstein 2001). This suggests that functional complexity in proteins is enhanced by multidomain structure. There may be also other explanations. Several important evolutionary mechanisms shape the genomes of eukaryotes, including population parameters that are conducive to genome-wide repatterning of gene structure (Lynch 2005) and consequently of protein architecture. In eukaryotes, exon boundaries may coincide with domain boundaries suggesting mechanisms of intronic recombination (Patthy 1999). Domain combinations are also believed to be the result of recombination through gene fusion rather than fission events (Snel et al. 2000; Kummerfeld and Teichmann 2005). In fact, a recent study showed that fusion was 4 times more common than fission in multidomain proteins and that in most cases these events occurred once in the course of evolution (Kummerfeld and Teichmann 2005). Previous observations had already hinted on the increase of domain repeats in eukaryotes (Ekman et al. 2005). However, our study shows that this increase is notable and is mostly due to evolutionary processes linked to Metazoa. In this regard, a study of domain evolution across 62 genomes of known phylogeny revealed that convergent evolution of domain architectures was rare (Gough 2005). It also showed that the number of tandem repeat domain architectures evolved more rapidly, and was less functionally constrained, than changes involving other types of recombination events, such as loss or gain of different domains. Our phylogenomic study (see below) indicates that the majority of phylogenetic change related to domain combination occurred within the eukaryal clade, especially within Metazoa, which is in line with the increase of the fraction of domain-repeat proteins observed in eukaryotic genomes. Consequently, the emergence of domain repeats in the protein world appears driven by very specific evolutionary processes responsible for the combination of domains in proteins that is confined to specific organismal lineages.

Although the presence of domain combinations in genomes is pervasive, we found that a substantial portion of domains do not combine with others. In the study of proteomes belonging to 185 fully sequenced organisms, single-domain proteins account for 40% of all described proteins, ranging 35–54% for genomes belonging to individual organismal domains (fig. 1). These values are higher than those reported earlier (20–35%) in an analysis of 40 genomes (Apic et al. 2001a) but comparable to a more recent study (Ekman et al. 2005), probably resulting from the wider and more effective survey of genomic sequences. If domain combination provides a selective advantage (Apic et al. 2003), then the existence of a substantial number of proteins still harboring single domains suggests that the rise of domain combinations occurred relatively late in evolution. This view is supported by the study of how domain combinations distribute between organismal domains. The Venn diagrams show that domain combinations unique to Archaea, Bacteria, or Eukarya were considerably more prevalent than those shared between organismal domains (fig. 2). This is in sharp contrast with the distribution of folds and fold superfamilies in genomes, where most architectures were common to organismal domains (fig. 2) (Caetano-Anollés G and Caetano-Anollés D 2003, 2005; Yang et al. 2005) and were placed at the base of a rooted phylogenomic tree describing the evolution of the protein world (Caetano-Anollés G and Caetano-Anollés D 2003, 2005). An evolutionary tendency to combine domains that materialized during organismal diversification and relatively late in the evolutionary history of protein structure is also supported by a recent phylogenomic analysis of domain combinations (Wang M and Caetano-Anollés G, unpublished data).

It is noteworthy that Bacteria and Eukarya shared more domains and domain combinations at fold and fold superfamily levels than those shared by Archaea and Eukarya (fig. 2). This suggests strongly that eukaryotes are much more closely related to eubacteria than they are to archaebacteria. This finding is in line with recent studies. For example, 75% of genes in yeast that have homologues in prokaryotic genomes share greater amino acid sequence identity and gene content with Bacteria than with Archaea (Esser et al. 2004). This is clearly incompatible with the widely assumed Archaea–Eukarya sisterhood relationship inferred fundamentally from rooted trees generated from the small subunit of rRNA (Woese et al. 1990) and from genes of the informational class (Walsh and Doolittle 2005). Consequently, the sister group relationship between Bacteria and Eukarya revealed here by studying the distribution of domain combinations supports new views of organismal diversification (Esser et al. 2004; Rivera and Lake 2004; Walsh and Doolittle 2005; Embley and Martin 2006; Kurland et al. 2006) that differ notably from the established paradigm based on rRNA.

The repertoire of domain combinations is limited, and combinations are not generated by random recombination (Apic et al. 2003). Instead, domain pairs are duplicated more frequently than expected by chance and domain sequential order is conserved, favoring selected 3D arrangements or geometries (Bashton and Chothia 2002). An analysis of our data set shows that the repertoire of domain combination defined within the 185 genomes we examined was limited to only a small fraction of possible combinations (e.g., the 6,460 observed represent ~2% of possible pairwise combinations) and supports the existence of bias and geometrical constraints in domain permutation (fig. 3). This limited repertoire was used to generate a pan-domain phylogeny of proteomes. For this purpose, we introduced new kinds of phylogenetic characters depicting the content (occurrence), popularity (abundance), and arrangement of domains in domain combinations within the repertoire. The criterion of primary homology underlying the use of these characters was simply the sharing of features describing the combination of domains in proteomes at protein fold and fold superfamily levels. Characters were defined by coding each individual instance of domain combination. Alternatively, multidomain proteins were treated as arrangements of domain pairs, and each pairwise combination of domains in a protein was coded separately. We found that both strategies produced phylogenomic trees with similar topologies (Figs. S3–S6, Supplementary Material online). Topologies were also similar when domains were defined at either fold or fold superfamily levels (compare for example Figs. S3 and S4, Supplementary Material online) or when characters described the abundance (Figs. S3–S6, Supplementary Material online) or the occurrence (presence/absence) (Figs. S7–S9, Supplementary Material online) of domains in proteins. Combined evidence suggests that global trees had a cohesive phylogenetic signal that was somehow independent of the coding strategy and the hierarchical level of protein classification used to define domain architecture.

Careful analysis of phylogenomic reconstructions showed interesting evolutionary patterns related to the last universal ancestor, the Coelomata hypothesis, and the topology of clades. At global level, phylogenies revealed the tripartite nature of the living world (fig. 4). In a previous study, we generated a phylogenomic tree of proteomes from the number of occurrences of domains at fold level in 32 genomes that also showed the tripartite nature of life (Caetano-Anollés G and Caetano-Anollés D 2003). Clear clades defining Archaea, Bacteria, and Eukarya were also observed by others (Yang et al. 2005) and when we extended this approach to the study of the 185 genomes analyzed here (Wang M and Caetano-Anollés G, unpublished data). Proteome tree topologies also revealed a sister clade relationship between Archaea and Bacteria, suggesting that the last universal ancestor was eukaryote like (Caetano-Anollés G and Caetano-Anollés D 2003, 2005). However, support for this topology was weak in our study. In general, branches occurring at deep and intermediate taxonomical levels were poorly supported. In contrast, lineage relationships within organismal domains were generally well supported by bootstrap analysis. In Eukarya, 46–64% of branches of trees reconstructed using different coding schemes had at least 90% BP and relationships showed proteomes well positioned and establishing accepted organismal groupings. The animals, plants, and fungi were monophyletic, and unicellular eukaryotes (Euglenozoa and Chromoalveolata) appeared at the base of the tree. The microsporidian *Encephalitozoon cuniculi* remained anomalously positioned at the base of the tree. It is noteworthy that eukaryal relationships support strongly the Coelomata hypothesis that groups chordates with arthropods (organisms with true body cavities) and the sister clade relationship of animals and plants. This is in line with evidence from previous phylogenomic studies that rejected the popular ecdysozoan and opisthokontal clades (Wolf et al. 2004; Philip et al. 2005; Ciccarelli et al. 2006). Amebozoa was basal to plants and animals and derived to fungi. In Archaea, 32–50% of branches had at least 90% BP, and relationships supported the Crenarchaeota and Euryachaeota, with *N. equitans* generally appearing at the base of the archaeal clade close to the crenarchaeal species. The methanogens, *Pyrococci*, and *Thermoplasmata* fell into single clades. The topology of lineages in Archaea was in agreement with trees reconstructed using other strategies (House and Fitz-Gibbon 2002; Korbel et al. 2002; Yang et al. 2005; Ciccarelli et al. 2006). In Bacteria, 26–40% of branches had at least 90% BP. Major bacterial groups were revealed, but statistical support for deeper branches was generally poor. Mollicutes, *Spirochaetes*, and *Chlamydiae* were basal in the bacterial clade. *Thermotogae* was also basal in trees reconstructed from protein domain combinations. The *Proteobacteria* was split into a major clade containing the α-, β-, and γ-proteobacteria and small clades containing the ε- and δ-proteobacteria. This proteobacterial split is not an unexpected outcome as it has been observed in other whole-genome studies (Wolf et al. 2002; Deeds et al. 2005; Yang et al. 2005). The Firmicutes formed 2 separate groups, one basally located in the tree and containing the Mollicutes. It is noteworthy that the Firmicutes comprise the earliest branching clade (sometimes together with *Thermotoga maritima*). The basal placement of this bacterial group agrees with recent results using sequence concatenation

of 31 universal orthologues (Ciccarelli et al. 2006) and support the theory of a Gram-positive bacterial ancestor (Koch 2003). In contrast, the derived placement of the other Firmicutes was unexpected. Interestingly, our results fail to support a thermophilic last universal common ancestor; unicellular eukaryotes were basal in our rooted trees and thermophilic Firmicutes were derived within the bacterial group. Overall, phylogenetic patterns obtained using domain combinations were reasonable and seem to support inferences from rRNA and whole-genome phylogenies (Yang et al. 2005) and relationships recently established by sequence concatenation (Brown et al. 2001; Ciccarelli et al. 2006). Advances in structural genomics and better knowledge related to definition of domain boundaries and domains and structural elements embedded in intervening sequences (Liu and Rost 2003) will enhance our database of domain combinations. We expect this will also enhance statistical support of phylogenetic patterns at deep and intermediate taxonomical levels, reinforcing the conclusions of this study.

The observation that the combination of domains in proteins represents a limited repertoire that carries phylogenetic information useful for the reconstruction of a reasonable supported tree of life suggests strongly that the process of domain combination is not random. Instead, it appears curved by evolution and fundamentally important for the formation of the protein repertoire. Protein structure is modular and probably arose from peptides by combination of supersecondary structural elements (Söding and Lupas 2003). In this regard, domains can be considered emergent modules, and the interactome of domain combinations a small-world and scale-free network defined by intramolecular interactions (Apic et al. 2001b; Wuchty 2001). The discovery of phylogenetic signal embedded in this interactome rejects the null hypothesis of modules combining in the absence of natural selection or an optimality criterion and supports the evolvability of intramolecular networks.

Supplementary Material

Tables S1 and S2, Figs. S1–S12, and a total of 6 Nexus files containing data matrices with characters describing genomic abundance (Gab_F.nex, Gab_FSF.nex and Gab_pw_F.nex) and occurrence (G_F.nex, G_FSF.nex and G_pw_F.nex) are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We gratefully acknowledge W. Martin and anonymous reviewers for their valuable comments. This work was supported in part by the Office of Naval Research, Department of Navy (TRECC A6538-A76) and the University of Illinois.

Literature Cited

Apic G, Gough J, Teichmann SA. 2001a. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol. 310:311–325.

Apic G, Gough J, Teichmann SA. 2001b. An insight into domain combinations. Bioinformatics. 17 (Suppl 1):S83–S89.

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

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science. 290:972–977.

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

Brenner SE, Koehl P, Levitt M. 2000. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res. 28:254–256.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nat Genet. 28:281–285.

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

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

Chothia C. 1992. One thousand families for the molecular biologist. Nature. 357:543–544.

Ciccarelli FD, Doerks T, Mering CV, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 23:324–328.

Deeds EJ, Hennessey H, Shakhnovich EI. 2005. Prokaryotic phylogenies inferred from protein structural domains. Genome Res. 15:393–402.

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

Dutilh BE, Huynen MA, Bruno WJ, Snel B. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol. 58:527–539.

Ekman D, Björklund ÅK, Frey-Sköt J, Elofsson A. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. J Mol Biol. 348: 231–243.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature. 440:623–630.

Esser C, Ahmadinejad N, Wiegand C, et al. (15 co-authors). 2004. A genome phylogeny for mitochondria among a-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 21:1643–1660.

Gerstein M. 1998. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. Proteins Struct Funct Genet. 33:518–534.

Gerstein M, Hegyi H. 1998. Comparing genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev. 22:277–304.

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.

Grishin NV. 2001. Fold change in evolution of protein structures. J Struct Biol. 134:167–185.

Hegyi H, Gerstein M. 2001. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. Genome Res. 11:1632–1640.

House CH, Fitz-Gibbon ST. 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J Mol Evol. 54:539–547.

Koch AL. 2003. Were gram-positive rods the first bacteria? Trends Microbiol. 11:166–170.

Korbel JO, Snel B, Huynen MA, Bork P. 2002. SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18:158–162.
Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21:25–30.
Kurland CG, Collins LJ, Penny D. 2006. Genomics and the irreducible nature of eukaryote cells. Science. 312:1011–1014.
Lin J, Gerstein M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10:808–818.
Liu J, Rost B. 2003. Domains, motifs and clusters in the protein universe. Curr Opin Chem Biol. 7:5–11.
Lynch M. 2005. The origins of eukaryotic gene structure. Mol Biol Evol. 23:450–468.
Murzin A, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 247: 536–540.
Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH: a hierarchic classification of protein structure. Structure. 5:1093–1098.
Patthy L. 1999. Genome evolution and the evolution of exon-shuffling-a review. Gene. 238:103–114.
Philip GK, Creevey CJ, McInerney JO. 2005. The opisthokonta and the ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the coelomata than ecdysozoa. Mol Biol Evol. 22:1175–1184.
Riley M, Labedan B. 1997. Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module. J Mol Biol. 268:857–868.
Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature. 431:152–155.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Söding J, Lupas AN. 2003. More than the sum of their parts: on the evolution of proteins from peptides. BioEssays. 25:837–846.
Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nat Genet. 21:108–110.
Snel B, Bork P, Huynen M. 2000. Genome evolution: gene fussion versus gene fission. Trends Genet. 16:9–11.
Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sunderland (MA): Sinauer Associates.
Tekaia F, Lazcano A, Dujon B. 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. 9:550–557.
Walsh DA, Doolittle WF. 2005. The real 'domains' of life. Curr Biol. 15:R237–R240.
Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposals for the domains Archaea, bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579.
Wolf YI, Brenner SE, Bash PA, Koonin EV. 1999. Distribution of protein folds in the three superkingdoms of life. Genome Res. 9:17–26.
Wolf YI, Grishin NV, Koonin EV. 2000. Estimating the number of protein folds and families from complete genome data. J Mol Biol. 299:897–905.
Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. Trends Genet. 18:472–479.
Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol. 1:8.
Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14:29–36.
Wuchty S. 2001. Scale-free behavior in protein domain networks. Mol Biol Evol. 18:1694–1702.
Yang S, Doolittle RF, Bourne PE. 2005. Phylogeny determined by protein domain content. Proc Natl Acad Sci USA 102: 373–378.

Michele Vendruscolo, Associate Editor

Accepted September 8, 2006