Supporting Information

Wang et al. 10.1073/pnas.1017361108

SI Materials and Methods

Protein sequences from the 749 proteomes listed in Table S1 and the corresponding structural assignments of these proteomes were downloaded from SUPERFAMILY (1) on June 12, 2009. Structural assignments were based on linear HMMs of structural recognition (2). Domains were defined using the Structural Classification of Proteins (SCOP) database, release 1.73 (3). A probability cutoff E value of 10^{-4} was applied to an initial cohort of 3,016,187 sequences. The final number of sequences was reduced to 2,983,835 after culling 32,352 sequences of four eukaryotic proteomes (*Drosophila yakuba* 1.3, *Drosophila simulans* 1.3, *Drosophila sechellia* 1.3, *Drosophila erecta* 1.3) with ID mismatches between sequences and structural assignments. In the final set, 64%, 34%, and 2% of sequences were of eukaryotic, bacterial, or archaeal origin, respectively. Lengths were calculated for single domain and multidomain proteins. Domain lengths were given as numbers of amino acids within a domain (defined at FSF level of structural complexity), or the sum of all lengths for domains composed of discontinuous sequences. When a sequence had two or more different domains with different E values or distinguishable domain models, the sequence was treated as a multidomain sequence and the domain length as the sum of all domains in the sequence. If domains in multidomain proteins had any overlaps (i.e., were completely or partially nested), the length of the overlap was calculated only once. Protein sequences without any domain assignment were excluded from the analysis (Table S1). In this study, we assume all genes in a genome are correctly predicted and we do not address shortcomings that may arise from gene splicing, incorrect fusions of genes, and other causes of potentially false domain predictions. Nevertheless, the consistency of our results and the stability of the patterns we have discovered lend some confidence to the present analysis.

Regions without domain assignments were treated as linkers, and their lengths were calculated as the differences between the length of proteins and their domains. Seed sequences for HMMs are based on the Protein Data Bank sequences found for each SCOP superfamily in the ASTRAL database (http://astral.berkeley.edu/) with less than 95% sequence identity. However, it has been shown that a model represents its superfamily first and foremost, rather than its seed sequence (3). For this reason, it is not likely that a model will detect domains based only on homology. This constrains domain boundaries within homologous regions of the seed and target sequences, which might be very short when they are remote homologues. To examine whether the origin of seed sequences affects our results, we selected the most ancient and abundant FSF, the P-loop containing the nucleoside triphosphate hydrolase superfamily (c.37.1), which is widely distributed in Eukarya, Bacteria, and Archaea (4). We examined only sequences with a single c.37.1 domain to avoid possible influences of other domains and c.37.1 models that hit five or more sequences in the three superkingdoms (101 of 393 models) to ensure good sequence representation in the model dataset. We then calculated the average length of the c.37.1 domain in the sequences hit by the same model within each superkingdom and identified the organismal species or virus linked to the seed sequence for every model. In this way, we could determine whether there were significant differences between the c.37.1 lengths recovered from the three superkingdoms for models in each superkingdom and in viruses by applying one-way ANOVA (Table S2). We found that there were no significant differences in the length of domains hit in individual superkingdoms by using models of Eukarya ($P = 0.2983$), Bacteria ($P = 0.1758$), Archaea ($P = 0.8865$), and viruses ($P = 0.7529$). This suggests that HMMs detect domain length equally well regardless of how remote the seeds are.

The fact that HMMs are unable to assign domains to entire sequences can be explained by the existence of unstructured protein regions (i.e., linkers) or putative domains that have not yet been characterized (5, 6). Alternatively, they can be explained by pronounced structural differences of fold superfamilies, in which domain cores harbor peripheral structural regions of weak sequence conservation that cannot be detected by homology modeling (7). These peripheral structures could occupy long regions in intervening sequences and could be responsible for known difficulties in predicting domain boundaries (8). However, and as we explain later, we determined that the probability that linkers contain putative domains that escape detection is in fact very low. We also carried out a detailed analysis of the length distribution of domains (or domain segments), linker segments scattering in the protein sequences, domain overlaps in multidomain sequences, and sequence length distributions of proteins with different numbers of domains. The data suggest that these factors do not significantly alter the global patterns of domains and linker lengths.

In these analyses, we chose an increment of 30 aa residues to count the number of domains or linkers distributed in different length intervals. The 30-residue threshold level makes it unlikely that a transmembrane region or SCOP domain is present in a prospective linker interval (5). This criterion has been used to define domain combinations in previous studies (9, 10). The distributions of linker and domain lengths were staggered in the ten 30-residue and greater than 300-residue length intervals (Fig. S2). A total of approximately 83% of linkers were shorter than 60 residues, and approximately 90% domains were longer than 60 residues in Bacteria and Archaea. For Eukarya, the numbers for linkers and domains were 64% and 76%, respectively. We also considered the possibility that domains might fill relatively short linkers and overlap with an adjacent domain. Thus, we checked how domain overlaps distributed in all sequences of this study Fig. S3. We found that 266,338 (14%) sequences in Eukarya, 153,907 (15%) sequences in Bacteria, and 8,992 (13%) sequences in Archaea had overlapping domains. Among these, 207,870 (78%) sequences in Eukarya, 125,799 (82%) sequences in Bacteria, and 6,702 (75%) sequences in Archaea had overlapping domain segments shorter than 30 residues.

We assume that putative domains in linkers follow the same distribution pattern as known domains. To simplify calculations, we also assume that a putative domain matches a linker if it falls within the same length interval. We note, however, that a linker might still be shorter than a putative domain within the length interval, and this would decrease the probability of the match. We used the frequencies of domains and linkers in different length intervals as well as the frequencies of domain overlaps to estimate the expected probabilities of linkers, which might correspond to putative domain assignments that could influence the final average length of domains and linkers in the three superkingdoms, respectively (Table S3). We used the following formula:

$$P_i = P_{l\_i} \times \left(\sum_{j=1}^{i} P_{d\_j} + P_{d\_(i+1)} \times P_{o\_30} + P_{d\_(i+2)} \times P_{o\_60} + P_{d\_(i+3)} \times P_{o\_90} \right) \quad \textbf{[S1]}$$

where i equals approximately 1 to 11 for length ranges approximating 0 to 30, 30 to 60, 60 to 90, . . ., 270 to 300, and at least

300, respectively; P_i represents the probability of the linkers in a specific length range that may get domain assignments; P_{l_i} and P_{d_i} represents the frequencies of linkers and domains in a specific length range, respectively; and P_{o_30}, P_{o_60}, and P_{o_90} are the frequencies of sequences with domain overlaps shorter than 30, between 30 and 60, and between 60 and 90 residues, respectively. For simplicity, we assume that domain overlap frequencies reflect the global distributions of overlaps rather than their distributions among domains in a specific length range.

We also studied the distributions of average domain (or segment) lengths and linker lengths for every genome in the three superkingdoms by using the nonparametric Mann–Whitney test. We found that the differences were significant ($P < 0.0001$). To further compare the distributions of domain segments and linkers, χ^2 tests were carried out for the length distribution of domain segments and linkers in every specific length range. Here too the differences were significant in each length interval ($P < 0.0001$). The statistical tests indicate that the length distributions of domain segments and linkers were substantially different, which was consistent with tendencies observed in Fig. S2. The probability of linkers matching domains in every length range was calculated with Eq. **S1** (Table S3). General probabilities for linkers in each superkingdom were $P = 0.04155$ for Eukarya, $P = 0.02170$ for Bacteria, and $P = 0.02252$ for Archaea. The small probabilities so obtained indicate that only a very small proportion of cryptic domains match linkers, so their putative presence would not affect length estimates significantly. Calculations show that putative domains reduce the actual average length of linkers by only approximately 22% uniformly for all superkingdoms (54 of 250 residues for Eukarya, 18 of 86 for Bacteria, and 16 of 73 for Archaea; Table S3), and that the average lengths of domains and linkers are reduced to 335 and 196 for Eukarya, 298 and 68 for bacteria, and 272 and 57 for Archaea. Thus, even when HMMs fail to detect domains in linker regions, their maximum influence is approximately uniform for Eukarya, Bacteria, and Archaea, and at the same time, the patterns of length distributions are maintained. In fact, sequence length distributions of proteins with different numbers of domains showed patterns that were consistent in the three superkingdoms (Fig. S4). We therefore conclude that, in the present study, HMMs mine structural information in sequence without significant bias.

We also determined whether there were any biases when applying the HMM method to sequences in the three superkingdoms. Our expectation was that protein sequences from any particular family should have the same number of domains, regardless of its superkingdom identity. We retrieved 5,523, 38,077, and 4,553 eukaryotic, bacterial, and archaeal sequences, respectively, with structural assignments to 352 orthologue families of a previous cohort (11). We then calculated the average number of assigned domains for sequences in every orthologue family for Eukarya, Bacteria, and Archaea, respectively. We recovered 246 families (70%) whose sequences had the same average number of domains in superkingdoms, and all of them had an SD of 0, which indicates that the individual numbers of domains in a family were exactly the same for Eukarya, Bacteria, and Archaea. The remaining families had equal numbers of domains with nonzero SDs (~4%) or occasionally contained equal number of domains in pairs of superkingdoms (~26%). Thus, HMMs capture the signal of domains without major bias in eukaryotic or microbial superkingdoms.

The entire set of protein sequences from the three superkingdoms was separated into six groups according to the number of domains that exist in proteins (i.e., proteins with one to five as well as more than five domains). For each group, we counted the total number of protein sequences analyzed and calculated its fraction in the entire sequence set, mean and median lengths, and corresponding SDs for proteins, domains, and linkers. Because length data in every group did not distribute normally ($P < 0.05$), we determined whether there were significant differences among protein, domain, or linker lengths by using Kruskal–Wallis one-way ANOVA on ranks. This test is a nonparametric analogue of a one-way ANOVA that is often used when there is one nominal variable and one measurement variable, and the measurement variable does not meet the normality assumption (12). Kruskal–Wallis H values were recorded to determine aggregate length differences among Eukarya, Bacteria, and Archaea. The aggregate values from the corresponding sets with equal data set sizes diverged significantly to different degrees as indicated by H, even if P values were not greater than 0.001 (i.e., differences in lengths among Eukarya, Bacteria, and Archaea were statistically significant). Box-and-whisker plots were generated to graphically illustrate the statistically descriptive data for protein, domain, and linker lengths in individual genomes.

Finally, we counted the numbers of the 20 aa as well as rare amino acids in the domains and linkers of all protein sequences, and then calculated the fractions of each amino acid in domains and linkers in proteins belonging to the three superkingdoms. We also used HMMs to assign domains to sequences in the DisProt database of intrinsically disordered proteins (13). We selected subsets that contained domain assignments and had low threshold levels of structurally undetermined regions. We then measured levels of order, disorder, and structurally undetermined regions in the set of DisProt entries.

1. Wilson D, Madera M, Vogel C, Chothia C, Gough J (2007) The SUPERFAMILY database in 2007: Families and functions. *Nucleic Acids Res* 35(Database issue):D308–D313.
2. 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.
3. Murzin AG, 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.
4. Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res* 17:1572–1585.
5. Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 310:311–325.
6. Apic G, Gough J, Teichmann SA (2001) An insight into domain combinations. *Bioinformatics* 17(suppl 1):S83–S89.
7. Grishin NV (2001) Fold change in evolution of protein structures. *J Struct Biol* 134: 167–185.
8. Liu J, Rost B (2003) Domains, motifs and clusters in the protein universe. *Curr Opin Chem Biol* 7:5–11.
9. Wang M, Caetano-Anollés G (2006) Global phylogeny determined by the combination of protein domains in proteomes. *Mol Biol Evol* 23:2444–2454.
10. Wang M, Caetano-Anollés G (2009) The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* 17:66–78.
11. Kurland CG, Canbäck B, Berg OG (2007) The origins of modern proteomes. *Biochimie* 89:1454–1463.
12. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 47:583–621.
13. Vucetic S, et al. (2005) DisProt: A database of protein disorder. *Bioinformatics* 21: 137–140.

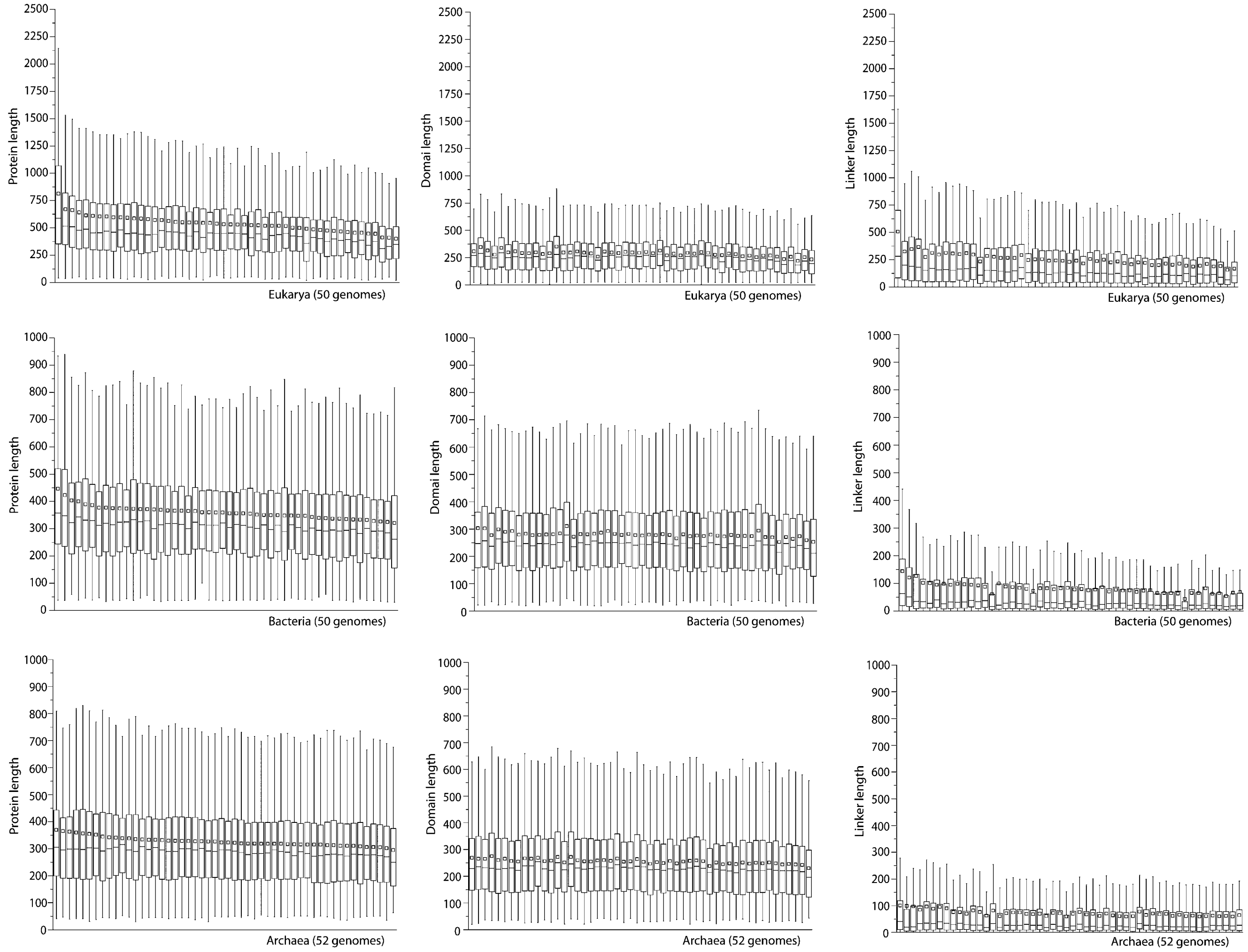

Fig. S1. Box-and-whisker plots illustrating key values from descriptive statistics of the lengths of proteins, domains, and linkers in individual genomes of Eukarya, Bacteria, and Archaea. To avoid illegibility, plots for sets of 50 genomes that were randomly selected from the 215 eukaryotic and 478 bacterial genomes that were analyzed (Table S1) are presented together with plots for the 52 archaeal genomes analyzed. Genomes were arranged in descending order according to average protein lengths. The lower and upper quartiles of box-and-whisker plots were set to 25% and 75%, respectively, as denoted by the range of the box. The short line at the center of the box stands for the 50% quartile (median value), and the small square in the box indicates the mean value. The whisker length defines the data's upper inner and lower inner fence values with a coefficient value equal to 1.5 using the following two formulas [upper inner fence = 75% percentile + (1.5 × IQR); lower inner fence = 25% percentile − (1.5 × IQR)]. Here, the interquartile range (IQR) = 75% quartile − 25% quartile. Because of the nonnormal distribution of the data used to construct the plots, Kruskal-Wallis tests were implemented to determine whether the average domain or linker lengths of individual genomes from Eukarya, Bacteria, and Archaea were significantly different. The results were summarized as follows. For domain lengths, median values were 252, 246, and 228 and mean values were 292 ± 262, 280 ± 189, and 257 ± 158 in the 50 eukaryotic, 50 bacterial, and 52 archaeal genomes analyzed, respectively; *H* was 788.184 ($P \leq 0.001$). For linker lengths, median values were 238, 76, and 67 and mean values were 259 ± 370, 858 ± 162, and 73 ± 131 in the 50 eukaryotic, 50 bacterial, and 52 archaeal genomes analyzed, respectively; the *H* value was 89,573.511 ($P \leq 0.001$). Accordingly, both the average domain and the average linker lengths of individual genomes are significantly different for the genomes of the three superkingdoms.

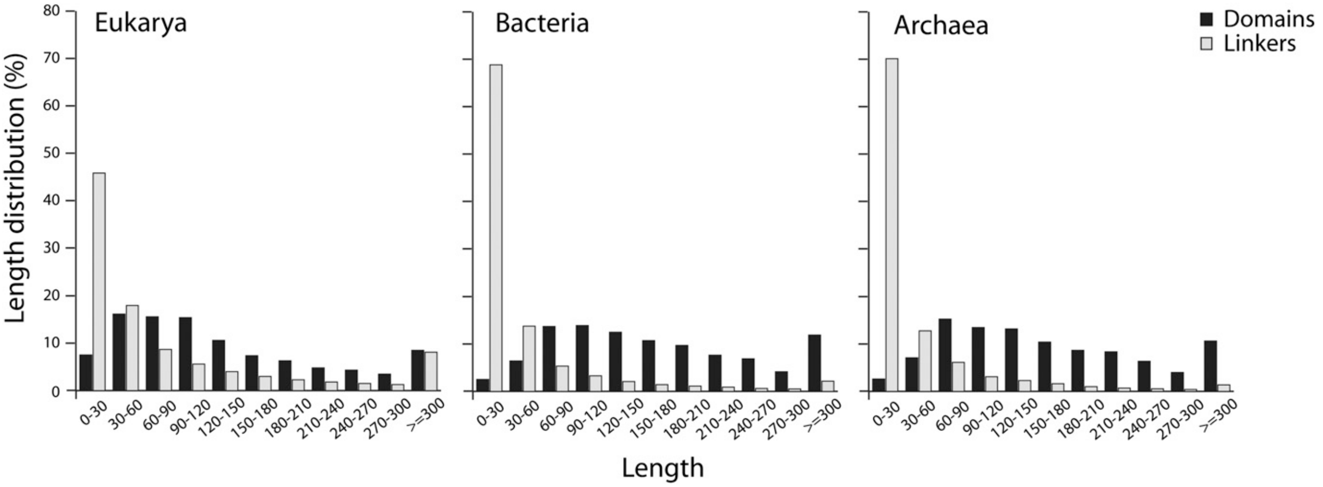

Fig. S2. Length distribution of domains and linker segments in the proteins of 745 organisms. The bar diagrams shows the distribution of domains or linker segments (given as percentage of total segments) in specific length intervals for organisms in each superkingdom. Because most linker segments are shorter than domain segments, the probability of cryptic domains existing in linkers is low.

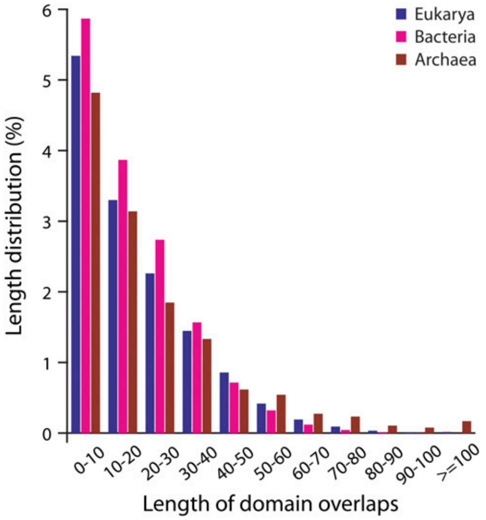

Fig. S3. Length distribution of domain overlaps in the proteins of 745 organisms. The bar diagrams the percentage of protein sequences with domain overlaps in the three superkingdoms. The existence of domain overlaps makes it possible for a relatively short sequence region to possess a long domain. However, the percentage of such sequences is small (<6% for the length intervals analyzed). For simplicity, domain overlaps were treated as being continuous. However, this was seldom the case. Broken overlaps will make it more difficult for a linker to match a domain.

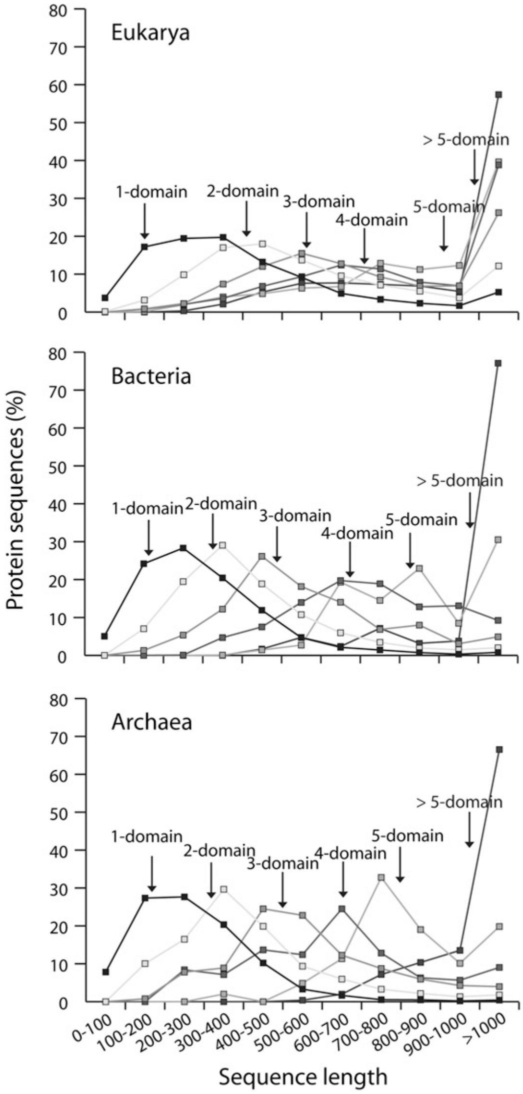

Fig. S4. Length distributions of protein sequences with 1–5 or more domains in 745 organisms. The arrows mark the curves of corresponding proteins, and the squares represent the fractions (%) of sequences within specific length ranges.

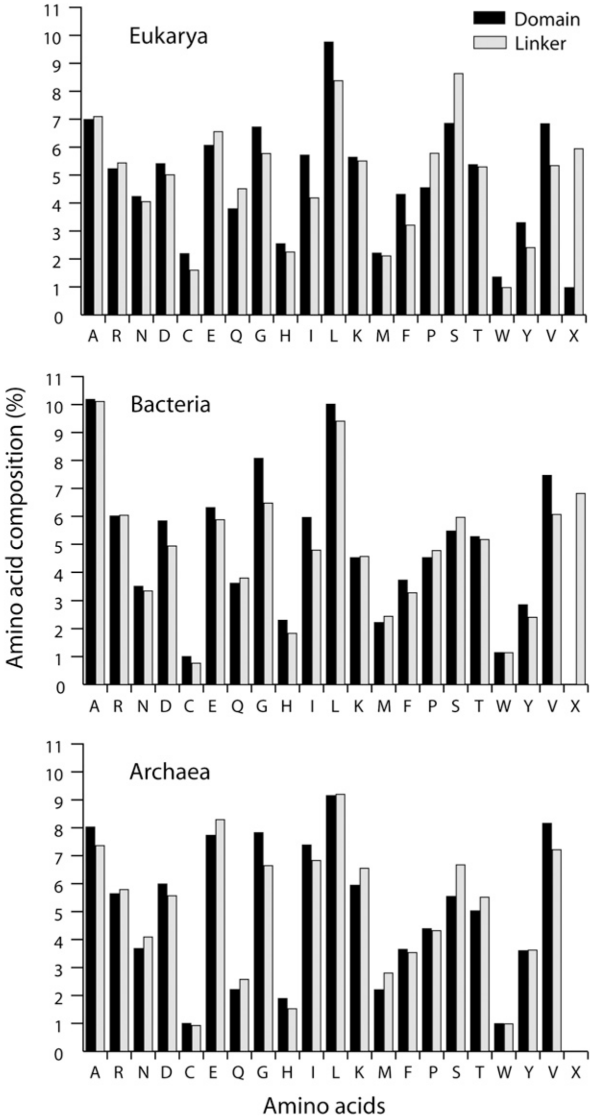

Fig. S5. Amino acid composition (%) of domains and linkers in Eukarya, Bacteria, and Archaea. One-letter amino acid abbreviations were used in the figure. "X" denotes a rare amino acid. We separated the amino acids into two groups according to the ratio of amino acid fraction in domains to fraction in linkers (>1 or <1) in the three superkingdoms, respectively. Paired *t* tests were carried out for the six resulting groups of amino acids identified in domains and linkers, Significant differences were found in all groups except for the two groups in Eukarya and Bacteria with amino acid fraction ratios <1.

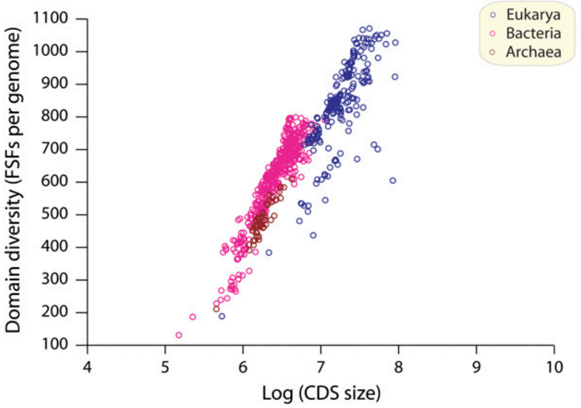

Fig. S6. The linear dependency between the logarithm of CDS size and the domain diversity (the number of FSF per genome) in the three superkingdoms of life. Significant linear correlations are obtained for Eukarya ($y = 378.4x - 1905.1$; $R^2 = 0.609$; $F = 331.166$, $P < 0.001$), Bacteria ($y = 431.150x - 2155.6$; $R^2 = 0.874$; $F = 3303.139$, $P < 0.001$), and Archaea ($y = 400.2x - 2019.9$; $R^2 = 0.860$; $F = 307.175$, $P < 0.001$).

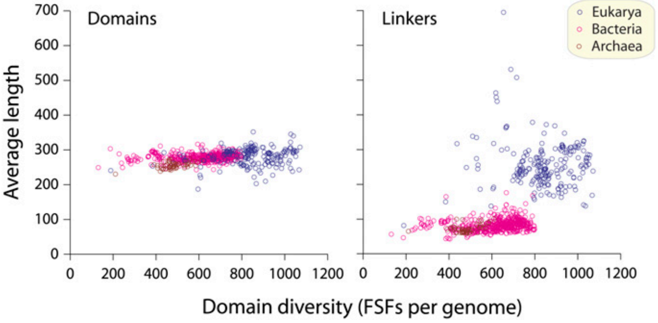

Fig. S7. Average domain and linker length plotted against domain diversity (number of unique FSFs) for individual genomes analyzed.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)