

A Phylogenomic Reconstruction of the Protein World Based on a Genomic Census of Protein Fold Architecture

MINGLEI WANG,¹ SIMINA MARIA BOCA,¹⁻³ RAKHEE KALELKAR,^{1,2} JAY E. MITTENTHAL,²
AND GUSTAVO CAETANO-ANOLLÉS¹

¹Department of Crop Sciences, ²Department of Cell and Developmental Biology, and ³Department of Mathematics, University of Illinois, Urbana, Illinois 61801

This paper was submitted as an invited paper resulting from the "Understanding Complex Systems" conference held at the University of Illinois–Urbana Champaign, May 2005

Received November 12, 2005; revised June 7, 2006; accepted June 7, 2006

The protein world has a hierarchical and redundant organization that can be specified in terms of evolutionary units of molecular structure, the protein domains. The Structural Classification of Proteins (SCOP) has unified domains into a comparatively small set of folding architectures, the protein fold families and superfamilies, and these have been further grouped into protein folds. In this study, we reconstruct the evolution of the protein world using information embedded in a structural genomic census of fold architectures defined by a phylogenomic analysis of 185 completely sequenced genomes using advanced hidden Markov models and 776 folds described in SCOP release 1.67. Our study confirms the existence of defined evolutionary patterns of architectural diversification and explores how phylogenomic trees generated from folds relate to those reconstructed from fold superfamilies. Evolutionary patterns help us propose a general conceptual model that describes the growth of architectures in the protein world. © 2006 Wiley Periodicals, Inc. *Complexity* 12: 27–40, 2006

Key Words: evolutionary funnel; organismal diversification; origins of life; protein fold structure; architectural diversification

1. INTRODUCTION

The protein world is complex. Protein sequence is extraordinarily diverse and so is protein structure and function [1,2]. Interestingly, protein sequences ($\sim 10^{13}$)

encoded in the genomes of the millions of species that currently inhabit earth (believed $\sim 10^7$ – 10^8 ; [3]) cover necessarily only a minute fraction of the enormous permutational space defined by amino acid sequence ($\sim 10^{321}$ – 10^{469} variants, given recent estimates of average protein length in genomes; [4]). Yet, the tools of structural genomics and protein structure determination revealed that this limited evo-

Correspondence to: G. Caetano-Anollés, E-mail: gca@uiuc.edu

lutionary exploration of sequence space has uncovered considerable diversity in both three-dimensional (3D) structure [5] and biological function (e.g., enzymatic catalysis [6,7]). In this quest, crystallography has shown that proteins generally associate with unique 3D structures. However, the mapping between sequence and structure space appears more complex than anticipated. Proteins sometimes display conformational diversity independent of binding to ligands, use structures to “moonlight” different functions without involving their active sites, or become promiscuous by using the same active site for different functions [8].

The protein world is also hierarchical. The protein domain is considered the smallest evolutionary unit and the basis of structural classification schemes [9,10]. Domains are present as single entities in many proteins, but assemble with other domains in larger multidomain molecules [11]. Domains sharing a common evolutionary origin defined by sequence, structure, and function are unified into fold families and superfamilies, and fold superfamilies containing domains that share arrangements of secondary structure elements are further grouped into protein folds. Although the evolutionary relatedness (monophyly) of fold superfamilies in individual folds has yet to be established, most folds represent groups of families and few folds represent groups of superfamilies. The Structural Classification of Proteins (SCOP) database [9,12] release 1.67 (February 2005) defines 65,122 domains grouped into 2630 families, 1443 fold superfamilies, and 887 folds. The more recent SCOP release 1.69 (July 2005) defines 945 folds, and the number is expected to increase, but only to a limited extent. Although the number of folds is finite, the rate of discovery of protein families has remained constant over time; we should consider protein architecture incompletely charted at the family level [13].

It is generally assumed that the diversity of the protein world evolved from a handful of ancestral proteins [1,2], arising perhaps from short polypeptides with limited structural complexity and then by assembly of supersecondary structures into more complex fold architectures [14]. Although our knowledge of this process is limited, the massive acquisition of genomic sequences fueled by genomics provides a unique opportunity to study the evolution of the protein world [15]. Domains of known 3D structure have been matched to genome sequences with great success. For example, hidden Markov models (HMMs) [16,17] are now capable of assigning known structures to more than 60% of open reading frames in complete genome sequences [18,19]. This “structural demography” enables an evolutionary study of entire protein complements (proteomes) in genomes. In this regard, fold and fold superfamily genomic composition, measured as presence or absence of individual architectures, was used to reconstruct whole-genome phylogenies (hierarchical branching histories of inheritance) based on the idea that closely related organisms must

share significantly more architectures than distantly related ones [20–24]. Similarly, domain composition was used to reconstruct prokaryotic phylogenies [23]. These trees were built using both distance and parsimony methods of phylogenetic reconstruction, had generally well-supported topologies, and showed clear monophyletic groups depicting the three major domains of life, Archaea, Bacteria, and Eukarya. Tracing of domain architectures along trees showed that convergent evolution of protein architecture reflecting independent evolutionary outcomes is a rare event, suggesting that domain architecture arose in the protein world by evolutionary descent [25] and supporting parsimony reconstruction strategies. Protein fold composition measured as popularity (number of occurrences) of each protein fold in sequenced genomes was also used to reconstruct intrinsically rooted proteome trees, invoking the concept that being popular is a favored evolutionary outcome [26,27]. Again, these whole-genome trees showed the tripartite nature of the universal tree of life and revealed a sister-clade relationship between Archaea and Bacteria and a rooting in the Eukarya.

Although proteome trees depict organismal diversification, the genomic census of architectures can be similarly used to reconstruct a phylogeny of protein architecture. Using a strict *cladistic* approach, we counted the number of genes matching folds by iterative rounds of PSI-BLAST analysis in a group of 32 genomes encompassing the three domains of life and used these measures of “structural demography” to map the world of proteins and track architectural history directly at the proteome level [26,27]. Cladistics represents a powerful and well-established method of evolutionary classification that groups taxa or objects hierarchically into nested sets. Our phylogenetic trees therefore present for the first time an evolutionary view of the protein world. The tree of fold architectures showed that α/β proteins originated first and gave rise to other protein classes, supporting the idea that in evolution, interspersed α -helices and β -sheets were segregated within structures and then confined to separate molecules [26]. A similar conclusion was obtained when tracing fold occurrence along the branches of proteome trees [28]. We revealed other interesting evolutionary patterns, such as an evolutionary increase in the curl and stagger of β -barrels (i.e., frequency of partly open or open barrel structures and the tilt of β -sheets) in the all- β protein class, an α -to- β tendency of secondary structure replacement, and proposed structural transformation pathways matching the topologies of the universal tree. We also explored how protein folds crossed organismal domain boundaries and used this information to uncover patterns in the origin and diversification of protein molecules and life [27]. We found there were marked heterogeneities in fold distribution, such as the placement of widely shared folds at the base of the tree, and we were able to infer a relative timing for the emergence of

prokaryotes, congruent episodes of architectural loss and diversification in Archaea and Bacteria, and a late and quite massive rise of architectural novelties in Eukarya probably linked to the rise of multicellularity.

In the present study, we extend our analysis of the protein world to a larger sampling of genomes, using advanced HMMs that match sequences to protein architectures. We explore how phylogenies generated from protein folds relate to those reconstructed from fold superfamilies and reveal interesting evolutionary patterns that help us propose a general conceptual model capable of describing the growth of architectures in the protein world.

2. MATERIALS AND METHODS

The frequencies with which individual protein folds occur in an individual genome, termed *genomic abundance* (G), were used to describe at global levels the popularity of fold architectures. In order to calculate G , structural domains were assigned to proteins at the fold superfamily level using the HMMs in SUPERFAMILY (<http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY>). This architectural hierarchy pools proteins for which there is structural and sequence evidence of a common evolutionary ancestor [18]. 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 [24]. Consequently, we did not explore the role of this parameter. Superfamilies were assigned to folds using SCOP 1.67 [9]. This release classifies 24,037 PDB entries into 887 protein folds. We analyzed the genome sequence of 185 organisms, encompassing 37 from Eukarya (species with lineage specifications: *Homo sapiens* 22 34d, *Pan troglodytes* 22 1, *Mus musculus* 22 32b, *Rattus norvegicus* 22 3b, *Gallus gallus* 22 1, *Xenopus tropicalis* 2 0, *Fugu rubripes* 22 2c, *Danio rerio* 22 3b, *Ciona intestinalis* 1 0, *Drosophila melanogaster* 3 2, *Anopheles gambiae* 22 2b, *Caenorhabditis elegans* WS123, *Caenorhabditis briggsae* Aug03, *Ustilago maydis* 1 r2, *Aspergillus nidulans* 1 r3 1, *Neurospora crassa* 3, *Magnaporthe grisea* 7 r2 3, *Fusarium graminearum* 1, *Saccharomyces cerevisiae*, *Saccharomyces paradoxus* MIT, *Saccharomyces mikatae* MIT, *Saccharomyces bayanus* MIT, *Candida glabrata*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, *Ashbya gossypii* 1 0, *Debaromyces hansenii*, *Candida albicans*, *Yarrowia lipolytica*, *Schizosaccharomyces pombe*, *Encephalitozoon cuniculi*, *Dictyostelium discoideum* 2, *Arabidopsis thaliana* 5, *Oryza sativa* ssp. *japonica* 2 0, *Plasmodium falciparum* 1, *Plasmodium yoelii* ssp. *yoelii* 1, and *Trypanosoma brucei*), 19 from Archaea (*Methanosarcina acetivorans* C2A, *Methanosarcina mazei* Go1, *Sulfolobus solfataricus* P2, *Sulfolobus tokodaii* 7, *Pyrobaculum aerophilum* IM2, *Archaeoglobus fulgidus* DSM 4304, *Pyrococcus furiosus* DSM 3638, *Halobacterium* sp. *NRC* 1, *Pyrococcus horikoshii* OT3, *Pyrococcus abyssi* GE5, *Methanothermobacter thermautotrophicus* Delta H, *Aeropyrum pernix* K1, *Methanocaldococcus jannaschii*

DSM 2661, *Methanococcus maripaludis* S2, *Methanopyrus kandleri* AV19, *Picrophilus torridus* DSM 9790, *Thermoplasma volcanium* GSS1, *Thermoplasma acidophilum* DSM 1728, and *Nanoarchaeum equitans* Kin4 M), and 117 from Bacteria (*Bradyrhizobium japonicum* USDA 110, *Streptomyces coelicolor* A3 2, *Streptomyces avermitilis* MA 4680, *Pirellula* sp. 1, *Mesorhizobium loti* MAFF303099, *Pseudomonas aeruginosa* PAO1, *Pseudomonas syringae* pv. *tomato* DC3000, *Nostoc* sp. PCC 7120, *Pseudomonas putida* KT2440, *Bacillus anthracis* Ames, *Bacillus cereus* ATCC 14579, *Bacillus thuringiensis* ser. *konkukian* 97 27, *Bordetella bronchiseptica* RB50, *Vibrio vulnificus* YJ016, *Vibrio parahaemolyticus* RIMD 2210633, *Rhodopseudomonas palustris* CGA009, *Bacteroides thetaiotaomicron* VPI 5482, *Leptospira interrogans* ser. *Lai* 56601, *Photobacterium luminescens* ssp. *laumondii* TTO1, *Vibrio vulnificus* CMCP6, *Erwinia carotovora* ssp. *atroseptica* SCRI1043, *Gloeobacter violaceus* PCC 7421, *Salmonella typhimurium* LT2, *Chromobacterium violaceum* ATCC 12472, *Salmonella enterica* ssp. *enterica* ser. *Typhi* CT18, *Mycobacterium avium* ssp. *paratuberculosis* k10, *Shewanella oneidensis* MR 1, *Xanthomonas axonopodis* pv. *citri* 306, *Escherichia coli* K12, *Bordetella parapertussis* 12822, *Xanthomonas campestris* pv. *campestris* ATCC 33913, *Shigella flexneri* 2a 301, *Bacillus subtilis* ssp. *subtilis* 168, *Bacillus halodurans* C 125, *Rhodobacter sphaeroides*, *Mycobacterium tuberculosis* H37Rv, *Mycobacterium bovis* AF2122 97, *Yersinia pseudotuberculosis* IP 32953, *Yersinia pestis* CO92, *Vibrio cholerae* O1 biovar. *eltor* N16961, *Caulobacter crescentus* CB15, *Clostridium acetobutylicum* ATCC 824, *Bdellovibrio bacteriovorus* HD100, *Oceanobacillus iheyensis* HTE831, *Geobacter sulfurreducens* PCA, *Ralstonia solanacearum* GMI1000, *Bordetella pertussis* Tohama I, *Desulfovibrio vulgaris* ssp., *vulgaris* Hildenborough, *Sinorhizobium meliloti* 1021, *Symbiobacterium thermophilum* IAM 14863, *Acinetobacter* sp. ADP1, *Brucella suis* 1330, *Brucella melitensis* 16M, *Synechocystis* sp. PCC 6803, *Desulfotalea psychrophila* LSv54, *Enterococcus faecalis* V583, *Lactobacillus plantarum* WCFS1, *Deinococcus radiodurans* R1, *Corynebacterium glutamicum* ATCC 13032, *Listeria innocua* Clip11262, *Corynebacterium efficiens* YS 314, *Listeria monocytogenes* EGD e, *Treponema denticola* ATCC 35405, *Xylella fastidiosa* 9a5c, *Staphylococcus aureus* ssp. *aureus* Mu50, *Clostridium perfringens* 13, *Thermoanaerobacter tengcongensis*, *Synechococcus* sp. WH 8102, *Thermosynechococcus elongatus* BP 1, *Nitrosomonas europaea* ATCC 19718, *Staphylococcus epidermidis* ATCC 12228, *Clostridium tetani* E88, *Lactococcus lactis* ssp. *lactis* Il1403, *Propionibacterium acnes* KPA171202, *Corynebacterium diphtheriae* NCTC 13129, *Chlorobium tepidum* TLS, *Streptococcus agalactiae* 2603V R, *Streptococcus pneumoniae* TIGR4, *Neisseria meningitidis* MC58, *Fusobacterium nucleatum* ssp. *nucleatum* ATCC 25586, *Wolinella succinogenes* DSM 1740, *Parachlamydia* sp. UWE25, *Leifsonia xyli* ssp. *xyli* CTCB07, *Pasteurella multocida* Pm70, *Coxiella burnetii* RSA 493, *Thermus thermophi-*

lus HB27, *Streptococcus mutans* UA159, *Porphyromonas gingivalis* W83, *Streptococcus pyogenes* MGAS10394, *Prochlorococcus marinus* ssp. *marinus* CCMP1375, *Helicobacter hepaticus* ATCC 51449, *Thermotoga maritima* MSB8, *Agrobacterium tumefaciens* C58, *Lactobacillus johnsonii* NCC 533, *Bifidobacterium longum* NCC2705, *Haemophilus ducreyi* 35000HP, *Streptococcus pyogenes* M1 GAS, *Haemophilus influenzae* Rd KW20, *Campylobacter jejuni* ssp. *jejuni* NCTC 11168, *Mycobacterium leprae* TN, *Helicobacter pylori* 26695, *Aquifex aeolicus* VF5, *Bartonella henselae* Houston 1, *Rickettsia conorii* Malish 7, *Wolbachia* *Bartonella quintana* Toulouse, *Chlamydomydia pneumoniae* J138, *Mycoplasma penetrans* HF 2, *Treponema pallidum* ssp. *pallidum* Nichols, *Mycoplasma mycoides* ssp. *mycoides* SC PGI, *Chlamydomydia caviae* GPIC, *Chlamydia muridarum*, *Chlamydia trachomatis* D UW 3 CX, *Borrelia burgdorferi* B31, *Rickettsia typhi* Wilmington, *Rickettsia prowazekii* Madrid E, *Borrelia garinii* PBI, *Tropheryma whipplei* Twist, *Mycoplasma pulmonis* UAB CTIP, *Onion yellows phytoplasma* OY M, *Mycoplasma gallisepticum* R, *Mycoplasma pneumoniae* M129, *Mesoplasma florum* L1, *Mycoplasma mobile* 163K, *Ureaplasma parvum* ser. 3 ATCC 700970, *Wigglesworthia glossinidia*, *Candidatus Blochmannia floridanus*, *Buchnera aphidicola* Bp, and *Mycoplasma genitalium* G 37, and reconstructed phylogenies describing the evolution of 776 protein folds that were present in these organisms and carried phylogenetic signal.

Phylogenetic trees of protein architecture were reconstructed using maximum parsimony (MP) as the optimality criterion in PAUP* [29]. G was normalized using gap-recoding techniques to compensate for differences in genome size and proteome representation and was then subjected to logarithmic transformation to account for unequal variances. The data was range standardized to a 0–20 scale compatible with most phylogenetic analysis programs, 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 ANCMSTATES command was used to polarize characters, assuming that the number of protein representatives in a genome exhibiting a particular fold increases in the course of evolution. Character argumentation is supported by model and assumptions described previously [26,27]. We consider that those protein families that grew early in evolution are prominent in many genomes and that the number of family members increases in single steps corresponding to the addition or removal of a homologous gene in a family. We assume that this process is reversible and expresses an asymmetry with gene duplication being favored over gene

loss. Phylogenetic reliability was evaluated by the bootstrap method [30]. The structure of phylogenetic signal in the data was tested by the skewness (g_1) of the length distribution of $>10^4$ random trees and permutation tail probability (PTP) tests of cladistic covariation using $>10^3$ replicates. Ensemble consistency (CI) and retention (RI) indices were used to measure homoplasy and synapomorphy (confounding and desired phylogenetic characteristics, respectively).

Protein architectures were classified into *fold distribution categories* that describe their spread across the three organismal domains of life. Architectures appearing in at least one proteome but in all organismal domains were assigned to the EAB category; those present in two domains were assigned to the EA, EB, and AB categories; and those present in only one domain were assigned to the A, B, and E categories. A *distribution index* (f) that describes the distribution of individual architectures among proteomes was also calculated. The f index represents the fraction of proteomes harboring an architecture within a category and ranges from absence ($f = 0$) to presence in all proteomes considered ($f = 1$). The relative age (ancestry) of individual protein architectures was measured using an indicator of the number of branching (cladogenic) events along individual lineages. This “node distance” (nd) was calculated as the number of nodes from the hypothetical ancestral fold in the reconstructed tree and was given on a relative 0–1 scale. Cumulative frequency plots were used to depict order and rate of appearance of architectural distribution categories and the accumulation of folds belonging to a protein class along phylogenetic trees. These plots can be considered time plots of lineages [31] with a time axis representing nd . Perl and MatLab scripts were written to extract ancestries from phylogenetic trees, count architectures in different categories and construct cumulative frequency plots. Fold distribution categories were traced along the branches of

Figure 1 Phylogenomic tree of protein architecture generated from a protein domain census in 185 completely sequenced genomes. The optimal most-parsimonious tree (85,644 steps; CI = 0.043, RI = 0.770, PTP test, $p = 0.01$) recovered from an heuristic MP search with tree-bisection-reconnection (TBR) branch swapping and 100 replicates of random addition sequence after exclusion of uninformative characters, was well supported by measures of skewness in tree distribution (lower inset). To decrease search times during branch swapping of suboptimal trees not more than one tree was saved in each replicate. Terminal leaves were not labelled because they would not be legible. Fold nomenclature follows that given in SCOP 1.67. The structural census defined by advanced hidden Markov models assigned domain structure to about 60% of genomic sequences (upper inset). The most ancient fold architectures are also described in a phylogenomic tree that was reconstructed separately. A heuristic MP search with TBR branch swapping and 100 replicates of random addition sequence resulted in two trees of 11,518 steps (CI = 0.258, RI = 0.630; $g_1 = -0.512$; PTP test, $p = 0.001$). Branches with bootstrap support (BS) values $>50\%$ are shown above nodes. Folds labeled in bold represent fold architectures that are present in each of the 185 genomes analyzed. Most of these architectures appeared at the base of the tree and are the most ancient in the protein world.

FIGURE 1

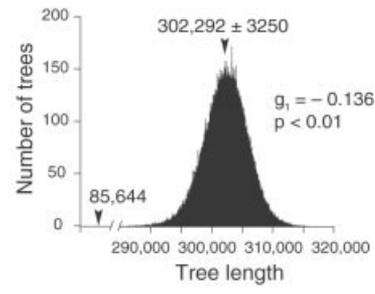
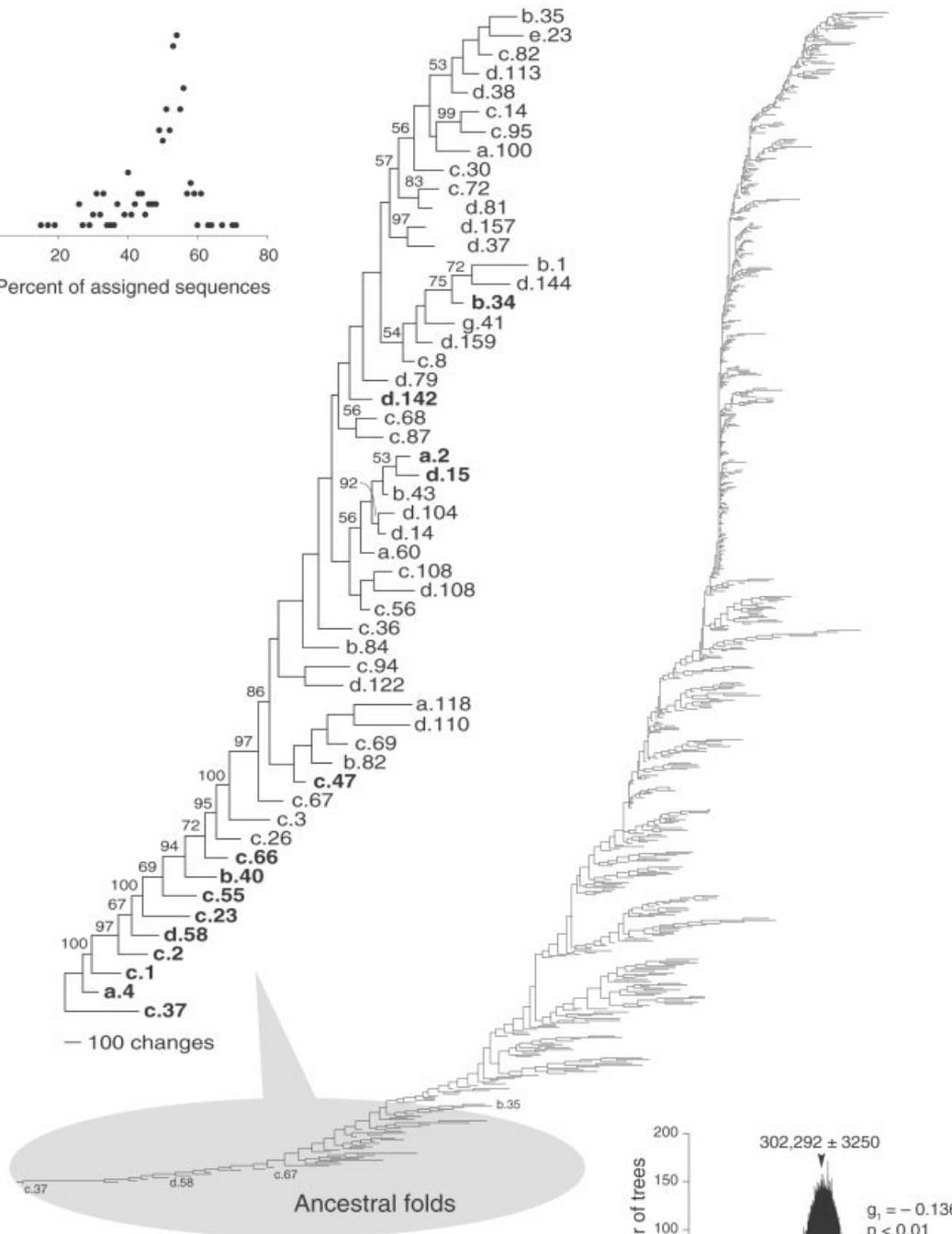
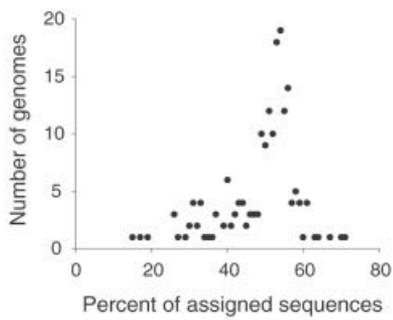
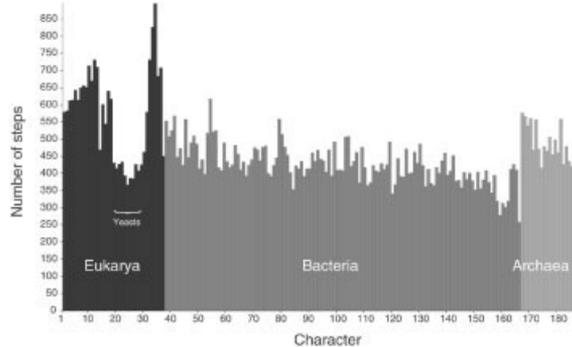


FIGURE 2

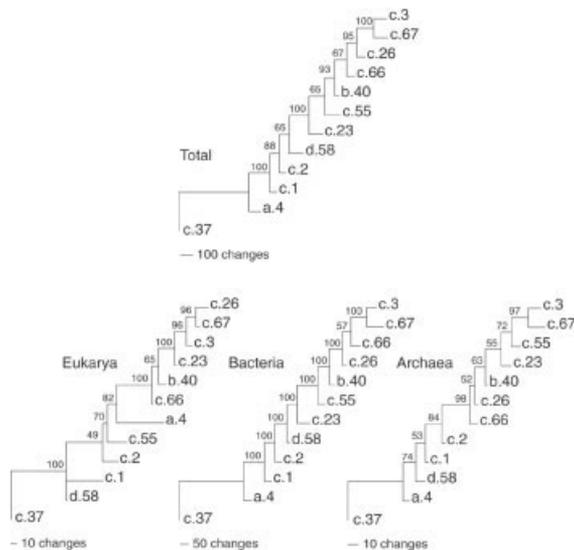
Plot describing character state change in each of the 185 genomes analyzed in this study. Phylogenetic characters (genomes) follow the order listed in Materials and Methods.

the tree of architecture using algorithms for Wagner and squared-change parsimony in MacClade [32].

3. RESULTS

3.1. A Universal Tree of Protein Fold Architecture

We reconstructed an intrinsically rooted phylogenetic tree describing the evolution of fold architectures of proteins encoded in 185 completely sequenced genomes belonging to organisms in Eukarya, Archaea, and Bacteria (Figure 1). The tree depicts the evolution of 776 folds defined in SCOP version 1.65 and identified in protein complements using advanced HMMs, expanding considerably our initial study of 536 folds and 32 genomes [26]. The study shows the evolutionary relationship of new folds that were not present in SCOP 1.59 or failed to be phylogenetically informative at that time. The average number of ORFs that could be assigned to a structure was $49.0 \pm 0.1\%$ (SD), and the median was 52%. Assignments ranged from 15% in *Plasmodium falciparum* to 71% in *Blochmannia floridanus* (Figure 1, top inset). These values were considerably higher than those achieved previously using PSI-BLAST (38.4%) [26], showing the superior performance of HMMs-based pattern recognition algorithms. The phylogenetic tree of protein architectures was well resolved (Figure 1). Tree distribution profiles and metrics of skewness were suggestive of strong cladistic structure ($p < 0.01$; Figure 1, bottom inset). In general, branches were poorly supported by bootstrap analysis, an expected outcome with trees of this size. The total length of internal branches (29,493 steps) was about half of the length of terminal leaves (56,151 steps) in the tree. However, character tracing showed that character state change was unequally distributed, increasing considerably towards the base of the tree (data not shown). Genomes from Eukarya, Archaea, and Bacteria contributed differently to character

FIGURE 3

Evolution of the 12 most ancestral protein architectures. One optimal most-parsimonious phylogenomic tree (2603 steps; CI = 0.648, RI = 0.675; $g_1 = -0.562$; PTP test, $p = 0.001$) was recovered by an exhaustive MP search using data embedded in all genomes analyzed. Analysis of genomes in Eukarya, Bacteria, and Archaea with exhaustive searches resulted in single optimal trees of 454 steps (CI = 0.729, RI = 0.803; $g_1 = -0.591$; PTP test, $p = 0.001$), 1676 steps (CI = 0.702, RI = 0.737; $g_1 = -0.586$; PTP test, $p = 0.001$), and 252 steps (CI = 0.706, RI = 0.748; $g_1 = -0.643$; PTP test, $p = 0.001$), respectively. BS values are shown above the nodes.

state change (Figure 2). Interestingly, *Ustilago maydis*, the yeasts (*Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikata*, *S. bayanus*, *Candida glabrata*, *Kluyveromyces lactis*, *K. waltii*, *Ashbya gossypii*, *Debaromyces hansenii*, *C. albicans*, and *Yarrowia lipolytica*), *Schizosaccharomyces pombe* and *Trypanosoma brucei* showed anomalously low character state change levels within Eukarya. Plants (*Oryza sativa* ssp. *japonica* and *Arabidopsis thaliana*) contributed the most to change and *Mycoplasma genitalium* the least.

An analysis of the 53 most ancestral protein folds appearing in the tree of architectures showed a well-supported phylogeny that preserved relationships without topological change (Figure 1). Interestingly, 9 of 16 folds found to be common in all 185 genomes analyzed appeared at the base of the reconstructed trees. In fact, 14 of these folds belonged to the set of 53 ancestral folds. A total of 19 folds within this set also contained fold superfamilies present in all organisms analyzed, 11 of which were placed at the base of the trees.

Because genomes from Eukarya, Archaea, and Bacteria contributed differentially to phylogenetic change, we tested if topologies of reconstructed trees would change when selecting character subsets representing genomes belonging to individual organismal domains (Figure 3). The reconstruction of phy-

TABLE 1

The Twelve Most Ancestral Fold Architectures of the Phylogenomic Tree of Protein Architectures

SCOP label	Fold	Description
c.37	P-loop containing nucleoside triphosphate hydrolases	3 layers with $\alpha/\beta/\alpha$ arrangement, parallel or mixed β -sheets of variable sizes
a.4	DNA/RNA-binding 3-helical bundle	Core: 3-helices; closed or partly opened bundle, right-handed twist; up-and-down
c.1	TIM β/α -barrel	Closed barrel with parallel β -sheet and strand order 12345678; $n = 8$, $S = 8$
c.2	NAD(P)-binding Rossmann-fold domains	Core: 3 layers in $\alpha/\beta/\alpha$ arrangement; parallel β -sheet of 6 strands, order 321456
d.58	Ferredoxin-like	Core: 3 helices; closed or partly opened bundle, right-handed twist; up-and-down
c.23	Flavodoxin-like	3 layers with $\alpha/\beta/\alpha$ arrangement; parallel β -sheet of 5 strands, order 21345
c.55	Ribonuclease H-like motif	3 layers with $\alpha/\beta/\alpha$ arrangement; mixed β -sheet of 5 strands, order 32145 with strand 2 antiparallel to the rest
b.40	OB-fold	Closed or partly opened barrel, $n = 5$, $S = 10$ or $S = 8$ with Greek-key motif
c.66	S-adenosyl-L-methionine-dependent methyltransferases	Core: 3 layers with $\alpha/\beta/\alpha$ arrangement; mixed β -sheet of 7 strands, order 3214576 with strand 7 antiparallel to the rest
c.26	Adenine nucleotide alpha hydrolase-like	Core: 3 layers with $\alpha/\beta/\alpha$ arrangement; parallel β -sheet of 5 strands, order 32145
c.67	PLP-dependent transferases	Main domain: 3 layers with $\alpha/\beta/\alpha$ arrangement; mixed β -sheet of 7 strands, order 3245671 with strand 7 antiparallel to the rest
c.3	FAD/NAD(P)-binding domain	Core: 3 layers in $\beta/\beta/\alpha$ arrangement; central parallel β -sheet of 5 strands, order 32145; top antiparallel β -sheet of 3 strands, meander

logenetic trees describing the evolution of the 12 most ancestral folds showed discordant topologies in trees generated from Eukarya and Archaea, but not from Bacteria, when compared to the tree generated using all genomic characters. The identity of these 12 folds is described in Table 1. These folds had barrel (c.1, b.40) or interleaved β -sheets and α -helical architectures that packed helices to one face (c.3, d.58) or two faces (c.37, c.2, c.23, c.55, c.66, c.26, and c.67) of the central β -sheet arrangement.

3.2. Tracing the Evolution of Major Classes of Protein Architecture

Cumulative frequency distribution plots revealed clear evolutionary patterns in the appearance and accumulation of protein folds within the six major classes of globular proteins, namely, α/β , $\alpha+\beta$, all- α , all- β , small and multidomain (Figure 4). In these plots, cumulative fold number was given as a function of distance in nodes (nd) from a hypothetical ancestral fold. These patterns remained consistent with those uncovered previously in the tree generated from 32 proteomes [26]. All classes appeared very early in the tree of architectures ($nd \leq 0.127$), with small and multidomain proteins appearing last. However, folds accumulated at different levels. The α/β class appeared first and accumulated at high rates, assuring its prevalence in the bottom half of the tree. The $\alpha+\beta$ class accumulated significantly later but with increasing rates, and at about $nd = 0.4$, it became the most prevalent in the tree. Folds in the all- α , all- β , small and, to some point, the multidomain classes followed the same pattern of accumulation but with diminishing rates, in that order. The accumulation of multidomain proteins was substantial between 0.2 and 0.6 nd units, but multidomain folds were the most underrepresented in the tree.

3.3. Analysis of Distribution Patterns of Protein Fold Architectures across Organismal Domains of Life

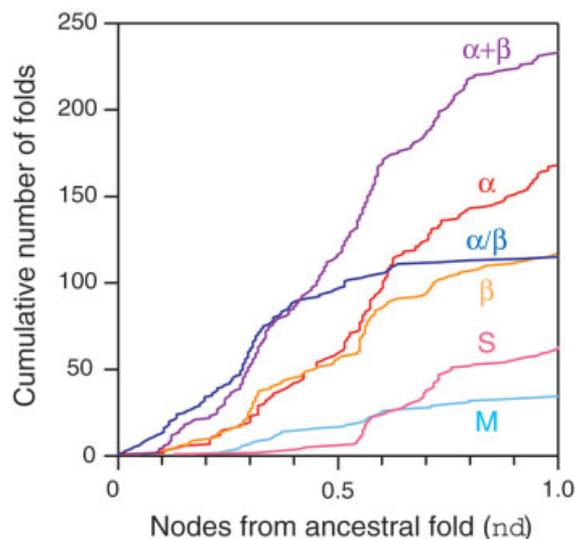
Distribution patterns of fold architectures can be indicative of how proteins and genomes have diversified in evolution [20,21,27]. The distribution of protein folds among organismal domains described in a Venn diagram (Figure 5) indicates that about two thirds of protein folds are common to all organismal domains (the EAB category). Within this set, 16 folds were omnipresent, i.e., they could be found in every genome analyzed. No other omnipresent fold appeared in any other category. Only 12 of the 16 omnipresent folds contained omnipresent fold superfamilies (a.4, b.34, b.40, c.1, c.120, c.2, c.23, c.37, c.55, c.66, d.142, and d.58), and all of them were of very ancestral origin (Figure 1).

There were two folds unique to Archaea (A), 17 unique to Bacteria (B), and 52 unique to Eukarya (E). Eukarya shared more folds with Bacteria (176 EB folds) than with Archaea (10 EA folds) and Archaea and Bacteria shared only 13 folds (AB). The tendencies in distribution patterns matched those seen previously [21,27]. However, new hidden Markov models and larger genome sampling resulted in an increase in the number of widely shared folds and an overall decrease in all other distribution categories, with an exception in EA folds, which increased in number.

3.4. Tracing the Evolution of Fold Distribution across Organismal Domains of Life

Fold distribution categories and an index (f) describing the popularity of individual folds among proteomes in each distribution category were traced in a phylogenomic tree of protein architecture, and tracings described as cumulative frequency distribution plots (Figure 5). Frequency plots showed that the first folds to appear in evolution were

FIGURE 4

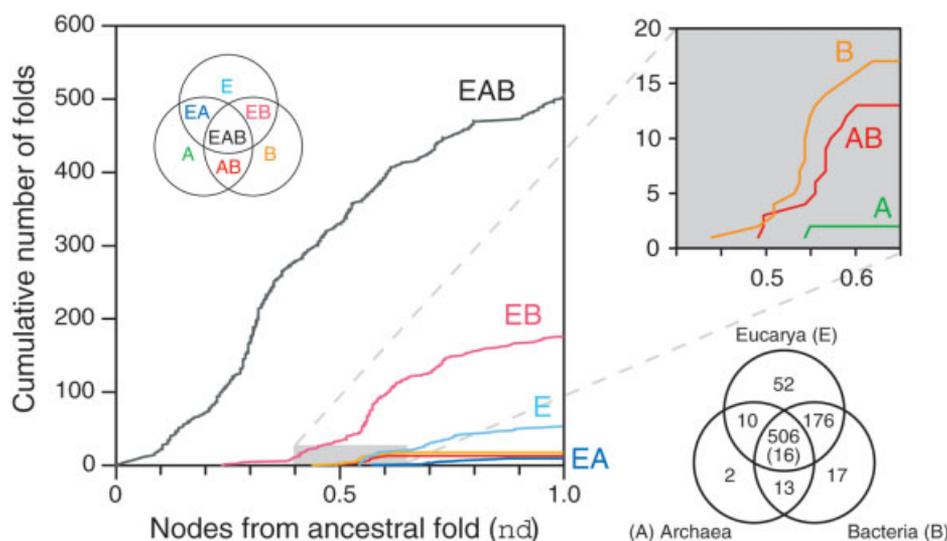


Cumulative frequency distribution plots describing the accumulation of folds belonging to different classes of globular proteins along the universal tree of fold architectures. Cumulative fold number is given as a function of distance (nd) in nodes from the hypothetical ancestral fold, in a relative scale.

common to all domains of life, with nine omnipresent folds appearing first and spanning the 0–0.046 nd range and seven more appearing sparsely within the 0.104–0.324 nd range (see also Figure 1, Table 1). Most of these folds were superfolds containing a large number of superfamilies. In fact, the first half of the tree of architectures was dominated almost exclusively by common EAB fold architectures, and these folds were generally well represented in all proteomes examined.

The second derived half of the tree of architectures was fundamentally composed of folds belonging to the common EAB and the EB fold architectures (Figure 5). Folds belonging to one or subsets of organismal domains accumulated within the 0.5–0.6 nd range, mainly A, B, and AB folds. Folds unique to Eukarya (E folds) and those common to Eukarya and Archaea (EA folds) accumulated last at $nd > 0.55$. Most folds within the derived half of the tree were patchily distributed among proteomes, including those within the EAB category, and were unifolds with only one or two families. The patchiness of distribution of folds within proteomes in each category increased in the order $EAB < EA < E < A < EB < AB < B$, and ranged from $f = 0.64 \pm 0.31$ (SD) for EAB folds to $f = 0.26 \pm 0.07$ (SD) for B folds. Patterns of accumulation matched well those observed previously in the analysis of 32 genomes [27]. However, the discovery of new folds, the increase in the number of genomes analyzed, and the more efficient assignment of genes to fold architectures

FIGURE 5



Cumulative frequency distribution plots describing the accumulation of folds unique to one or to sets of organismal domains. The Venn diagram shows occurrence of 776 folds defined by SCOP 1.67 in the three organismal domains of life, based on the analysis of 185 genomes. The value in parenthesis indicates the number of folds that are omnipresent in that fold distribution category; there were no omnipresent folds in categories other than EAB. Cumulative fold number for each fold distribution category defined by the Venn diagram is given as a function of distance in nodes from the hypothetical ancestral fold. The inset shows a detail on the accumulation of A, B, and AB folds in the tree of architectures.

changed patterns of first appearance of folds and in some cases resulted in changes of fold distribution category. The first fold unique to Bacteria, the putative cell cycle protein MesJ (d.229), appeared at 0.439 *nd* units and was closely followed by the first fold common to prokaryotes, the AbrB/MazE/MraZ-like fold (b.129), appearing at 0.491 *nd* units, the first fold unique to Eukarya, the yeast killer toxin fold (d.70; 0.543 *nd* units), and the two folds unique to Archaea: the hypothetical protein MTH1880 fold (d.214; 0.543 *nd* units) and the methyl-coenzyme M reductase C-terminal domain fold (a.89; 0.549 *nd* units). All these folds had not been identified at the time of our previous analysis [27].

3.5. Phylogenetic Trees of Fold Superfamilies

The rise of fold superfamilies within a fold delimits fold accumulation in genomes. In order to explore evolutionary patterns defined by the phylogenetic relationship of fold superfamilies, we reconstructed a tree of superfamilies corresponding to the 12 most ancestral folds described in Table 1 (Figure 6). The appearance of fold superfamilies was not clustered in groups belonging to a fold, as could have been expected. Instead, fold superfamilies within a fold appeared interspersed with fold superfamilies belonging to other folds. Phylogenetic analysis of fold superfamilies within an individual fold resulted in trees with topologies that matched those delimiting fold superfamily relationship in the original tree (Figure 6). As expected, the resulting topologies preserved the branch distances observed in the original tree. Interestingly, cumulative frequency distribution plots showed that the number of fold superfamilies in folds within the basal 123 fold superfamilies examined increase in evolution (Figure 7). Moreover, the accumulation of fold superfamilies in the phylogenomic tree of fold superfamilies (Figure 7) matched the ancestry of folds revealed in the tree of fold architectures (Figure 3).

4. DISCUSSION

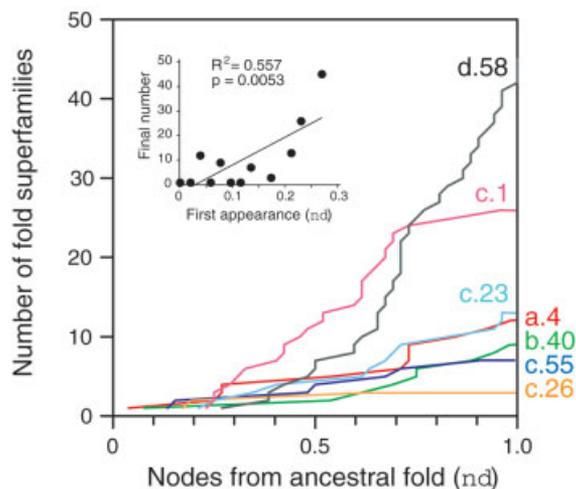
We here extend our initial studies of the protein world by exploring the evolutionary relationship of fold architectures encoded in a larger number of genomes. A phylogenomic tree describing the evolution of 776 protein folds defined in SCOP 1.67 was reconstructed using an efficient HMM-based census of protein structures in 185 genomes encompassing the three organismal domains of life. This tree is intrinsically rooted establishing evolution's arrow without the need of outgroups. The leaves in the tree correspond to folds and nodes represent architectural diversification events based on changes in popularity and sharing of folds in genomes, with nodes close to the base of the tree reflecting more ancient events than those close to the leaves.

Careful analysis of phylogenomic reconstructions showed interesting evolutionary patterns. As observed previously [26], all major classes of globular proteins appeared very early in evolution within the first 62 folds, and in

defined order, with the α/β protein class being first. Patterns of fold accumulation within these structural classes again suggest an evolutionary tendency of proteins that confine α -helices and β -sheets to separate locations in the molecules or to different molecules altogether. This tendency would favor the rise of modular and structurally canalized architectures, which are generally assumed to be favored outcomes in the evolution of molecular structure [33,34]. Note that we use the concepts of modularity and canalization in specific manner. A modular architecture uses a combination of building blocks such as domains and folds, and modularity entails an ability to sustain integrity of modules across varying environments and genetic contexts. Canalization describes the occurrence of lock-in mechanisms that limit the variation of modules so that they retain their ability to combine with other modules. The most ancestral folds appearing at the base of our tree were α/β barrel or interleaved architectures that were widely shared between organisms belonging to the three organismal domains of life. Interestingly, the most ancestral folds were omnipresent in all genomes analyzed. This is expected. Ancient architectures are by definition the most popular in genomes. They should have been present, perhaps in considerable number, in the genome(s) of the common ancestor(s) of all lineages that populate our living world, spreading efficiently in the protein world mostly by vertical descent. Their omnipresence should only have been compromised by architectural loss in exceptional cases. In this regard, it is noteworthy that the absence of common ancestral architectures in genomes occurs preferentially in organisms with parasitic lifestyles and gene complements that have been highly reduced in size during evolution (e.g., *Trypanosoma*, *Nanoarchaeum*, and *Mycoplasma*; S.M. Boca, unpublished results).

Character tracing along the branches of the phylogenomic tree revealed other interesting patterns. Although common fold architectures were confined to the bottom half of the tree, those that were unique to individual organismal domains or combinations of two domains appeared halfway in evolution and were prevalent in the latter half of our phylogeny. Character tracing also showed that common fold architectures accumulated in the derived half of the tree, but these generally failed to be widely shared among organisms within organismal domains. The appearance of folds unique to or shared by prokaryotes (i.e., Bacteria and Archaea) occurred mostly within a very narrow evolutionary window (<0.1 *nd* units; Figure 5). We believe that this window defines, at the fold level, the start of organismal diversification in the living world. Note that this window can be located elsewhere if fold architecture is defined and traced in trees describing the evolution of fold superfamily, families or domains, i.e., trees describing lower hierarchical levels of protein architecture. Interestingly, the only fold superfamily that is unique and omnipresent in Archaea (d.17.6) and is perhaps evolutionarily diagnostic to this group of organisms [24], belongs to a EAB fold that appears

FIGURE 7



Cumulative frequency distribution plots describing the accumulation of fold superfamilies in the phylogenomic tree of superfamilies in the 12 most ancestral folds. The cumulative number of fold superfamilies for superfamilies in each fold was given as a function of distance in nodes from the hypothetical ancestral fold. The inset shows a linear regression plot of the relationship between final number of fold superfamilies and the first appearance of a fold in the tree of fold superfamilies.

earlier in evolution at 0.295 *nd* units. This fold superfamily is found in an enzyme involved in the synthesis of archaeosine, a modified base found exclusively in Archaea [35]. It is noteworthy however that three fold superfamilies unique and omnipresent in Eukarya (g.41.10, b.34.10, and a.24.15) belong to EAB folds that appeared much earlier in our phylogenomic tree (within 0.123–0.231 *nd* units). g.41.10 is the Zn-finger domain of Sec23/24, a protein involved in GTP-dependent recruitment of a vesicular coat complex in budding yeast, b.34.10 is the Cap-Gly domain of a protein of unknown function in *Caenorhabditis elegans*, and a.24.15 the FAD-dependent thiol oxidase domain in, for example, an enzyme that promotes disulfide bond formation during protein biosynthesis in the yeast endoplasmic reticulum. It would be interesting to know how these fold superfamilies fare within fold distribution patterns in phylogenomic trees describing the evolution of fold superfamilies instead of folds.

Evolutionary patterns of fold accumulation resembled those previously observed in an analysis of 32 genomes [27]. However, the discovery of new folds, the increase in the number of genomes analyzed, and the more efficient assignment of genes to fold architectures with HMMs sometimes changed patterns, especially those of first appearance of folds. For example, the first fold appearing in the tree that was unique to an organismal domain (d.229) was found to be unique to Bacteria. This fold was first described in SCOP

1.63 (May 2003) and was consequently absent in our previous SCOP 1.59 (May 2002)-based study. The same applies to several AB folds unique to prokaryotes that closely follow d.229 in the evolutionary tree (see Results). However, AB folds accumulated earlier than those that were unique to Bacteria and Archaea (Figure 5). AB fold accumulation started at about 0.5 *nd* units and preceded B and A fold accumulation at about 0.55 *nd* units. The early appearance of folds shared by prokaryotes suggest the emergence of a prokaryotic lineage and structural diversification in Bacteria and Archaea, and the late accumulation of folds unique to Eukarya suggests a late rise of architectural novelties perhaps linked to multicellularity. These findings support previous observations [27].

It is clear that extending our analysis of the protein world to a larger number of genomes and fold architectures has not changed considerably the evolutionary patterns of structural and organismal diversification we previously uncovered [26,27]. However, we still do not understand how the hierarchical organization of the protein world impacts evolutionary relationships at the structural level.

The reconstruction of phylogenomic trees based on the occurrence of fold architectures in genomes is driven by how architectures accumulate in time and in the different lineages of a universal tree of life. One fundamental assumption is that all architectures arose from a common ancestor at the time of the “big bang” of architectural diversification, i.e., that protein evolution can in fact be adequately described with an evolutionary tree. A similar assumption about organismal diversification is used when building the tree of life with phylogenetic or phylogenomic methods. However, two detracting arguments can challenge this view of vertical descent:

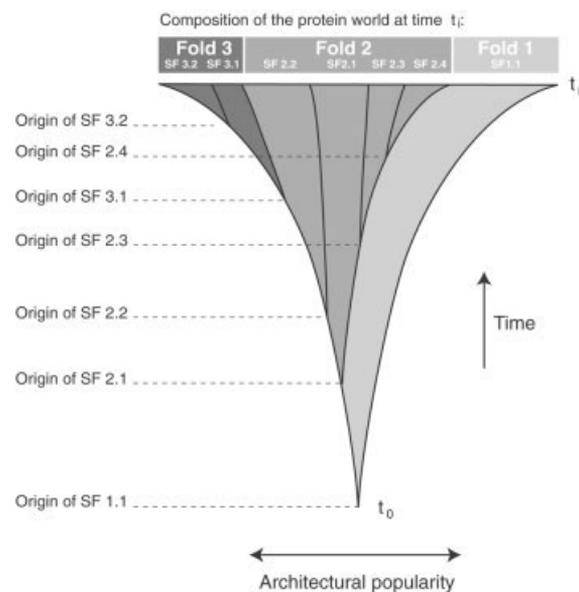
1. It could be argued that convergent evolutionary processes (i.e., those that lead independently to a similar outcome) and horizontal transfer have been pervasive phenomena. These processes have the potential of obliterating patterns of descent. However, a recent tracing of domains of known structure along whole-genome phylogenies reveals that convergent evolution is rare in domain architecture [25]. This is an important finding that links processes of architectural and organismal diversification. The considerable phylogenetic signal embedded in trees of architectures [26,27] and genomes (e.g., inferred from domains [23]) and the patterns of fold sharing observed at the base of our trees [26,27] are also consistent with absence of processes of convergence and horizontal transfer. Taken together, findings support the idea that evolution of fold architecture is driven by vertical descent, diversification of architectural components, and architectural loss rather than by independent inventions or horizontal transfer.

2. It could be similarly argued that lineages arose from multiple ancestors and that reconstructed trees are sets of overlapping lineages that evolved independently. Ancestral architectures such as P-loop hydrolases and β/α -barrels might have arisen independently from small peptides, yet still be universally shared and placed at the base of our tree of architectures. This is because our phylogenomic approach is inconclusive with respect to the hypothesis of multiple ancestors. Trees reconstructed from architectural popularity describe relationships based on demography of structures and not architectural topology. The model behind our tree reconstruction exercise is driven by the success (fitness) of architectures in the protein world (and later in the organismal world) and not by structural transformations reflecting individual sequence-to-structure mappings. It is driven by the accumulation of successful architectural variants within a structural neighborhood, very much as propagation in Galton-Watson branching processes responds to a delicate balance of survival and extinction [36]. Our study of architectural popularity therefore mutes the argument of one or many lineages by focusing on global evolutionary relationships that describe the entire world. In our case, lineages are used as devices that “sample” architectural diversity (i.e., genomes are used as phylogenetic characters) and trees are used as evolutionary clocks fueled by the successful accumulation of structural designs.

The reconstruction of our phylogenomic trees is also dependent on the validity and definition of terminal taxa. A phylogenomic analysis of fold architectures places trust on how SCOP assigns fold superfamilies to folds. While previous studies confirm the evolutionary relatedness of fold superfamilies in, for example, α/β -barrels folds [37,38], this important issue has not been explored extensively. Phylogenomic analysis at the fold superfamily level offers a higher level of certainty that proteins belonging to this hierarchical level share a same origin [24]. Families unified by fold superfamilies show good structural and functional evidence of common ancestry. However, fold superfamily trees such as those of Figure 6 reveal but fail to fully capture the evolutionary complexities of higher order structural organization. We should strive to establish the monophyletic nature (i.e., evolutionary relatedness) of protein folds. This would enhance our views of the protein world.

Protein architectures defined at different hierarchical levels should follow similar evolutionary pathways independently of the underlying hierarchical organization of the protein world. Patterns observed in trees of folds should be congruent with those of fold superfamilies, and so on. We test this concept by reconstructing phylogenomic trees of fold superfamilies and tracing folds along the reconstructed trees (Figures 6 and 7). The reconstruction of a fold superfamily tree describing the evolution of the most ancestral

FIGURE 8



A “funnel” can be used to illustrate the evolution of the world of fold architectures. The funnel describes in two dimensions the multidimensional growth in popularity of fold architectures describing the protein world after the “big bang.”

fold architectures shows superfamilies of different folds appearing interspersed in the tree and not clustered in groups (Figure 6). This suggests strongly that individual folds represent collections of proteins undergoing different but concomitant evolutionary processes that translate into patterns of recent (close relationship) or ancient origin (distant relationship). A gamut of different popularities is therefore expressed at the fold superfamily level, and this gamut ultimately defines the global popularity measure of the fold in question. Tracing of fold accumulation in the fold superfamily trees reveals accumulation patterns that match the evolutionary relationships of folds analyzed (Figure 7). These results suggest that folds are heterogeneous entities that evolve through increases in genomic representation of fold superfamily variants.

A “funnel” metaphor can illustrate the evolution of the world of fold architectures (Figure 8). The funnel describes in two dimensions the multidimensional growth in popularity of fold architectures after the “big bang” of architectural diversification. Embedded subfunnels describe the discovery and accumulation of architectural variants at different levels of hierarchical organization (Figure 8). The global funnel represents the combination of many subfunnels defining the contributions of each individual genome to the architectural diversification process, which are not shown in the figure. These contributions become meaning-

ful to the architectural diversification process at the time of organismal diversification, a process that becomes prevalent about halfway in evolution (Figure 5), as lineages of organisms compartmentalize the diversification process. In this model, we assume that rates of funnel growth are universal. Please note that the patterns of emergence of embedded subfunnels in the funnel reflect patterns of global architectural accumulation but they do not necessarily reflect patterns of phylogeny. However, phylogenies are the consequence of the funnel. Consequently, a mathematical model depicting the verbal model here proposed can be used to refine the hypotheses of character argumentation that support the reconstruction of phylogenomic trees.

5. CONCLUSIONS

The phylogenomic reconstruction of the protein world presented here establishes evolutionary links between patterns

of molecular and organismal diversification. These links portray the complexities and hierarchical organization that is embedded in protein architecture. They are here used to define a conceptual model of global accumulation of architectures in the living world. We believe this model is amenable to mathematical elaboration and will be useful for phylogenomic reconstruction. Ultimately, the model will help understand basic processes driving the molecular evolution of protein molecules.

ACKNOWLEDGEMENTS

We thank Seungwoo Hwang for Perl scripts, Derek Caetano-Anollés for help with data analysis and the Office of Naval Research, Department of Navy (TRECC A6538-A76) and the University of Illinois for financial support.

REFERENCES

1. Aravind, L.; Mazumder, R.; Vasudevan, S.; Koonin, E.V. Trends in protein evolution inferred from sequence and structure analysis. *Curr Opin Struct Biol* 2002, 12, 392–399.
2. Chothia, C.; Gough, J.; Vogel, C.; Teichmann, S.A. Evolution of the protein repertoire. *Science* 2003, 300, 1701–1703.
3. Bull, A.T.; Goodfellow, M.; Slater, J.H. Biodiversity as a source of innovation in biotechnology. *Annu Rev Microbiol* 1992, 46, 219–252.
4. Brocchieri, L.; Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res* 2005, 33, 3390–3400.
5. Todd, A.E.; Marsden, R.L.; Thornton, J.M.; Orengo, C.A. Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol* 2005, 348, 1235–1260.
6. Todd, A.E.; Orengo, C.A.; Thornton, J.M. Plasticity of enzyme active sites. *Trends Biochem Sci* 2002, 27, 419–426.
7. Gutteridge, A.; Thornton, J.M. Understanding nature's catalytic toolkit. *Trends Biochem Sci* 2005, 30, 622–629.
8. James, L.C.; Tawfik, D.S. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem Sci* 2003, 28, 361–368.
9. Murzin, A.; Brenner, S.E.; Hubbard, T.; Chothia C SCOP: A structural classification of proteins for the investigation of sequences and structures. *J Mol Biol* 1995, 247, 536–540.
10. Orengo, C.A.; Michie, A.D.; Jones, S.; Jones, D.J.; Swindells, M.B.; Thornton, J.M. CATH: A hierarchic classification of protein domain structures. *Structure* 1997, 5, 1093–1108.
11. Rossmann, M.G.; Moras, D.; Olsen, K.W. Chemical and biological evolution of nucleotide-binding protein. *Nature* 1974, 250, 194–199.
12. Andreeva, A.; Howorth, D.; Brenner, S.E.; Hubbard, T.J.P.; Chothia C.; Murzin A.G. SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acid Res* 2004, 32, D226–D229.
13. Kunin, V.; Ouzounis, C.A. The balance of driving forces during genome evolution in prokaryotes. *Genome Res* 2003, 13, 1589–1594.
14. Söding, J.; Lupas, A.N. More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays* 2003, 25, 837–846.
15. Doolittle, R.F. Evolutionary aspects of whole-genome biology. *Curr Opin Struct Biol* 2005, 15, 248–253.
16. Krogh, A.; Brown, M.; Mian, I.S.; Sjolander, K.; Haussler, D. Hidden Markov models in computational biology. Application to protein modeling. *J Mol Biol* 1994, 235, 1501–1531.
17. Karplus, K.; Barrett, C.; Hughey, R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 1998, 14, 846–856.
18. Gough, J.; Karplus, K.; Hughey, R.; Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 2001, 313, 903–919.
19. Madera, M.; Vogel, C.; Kummerfield, S.K.; Chothia, C.; Gough, J. The superfamily database in 2004: Additions and improvements. *Nucleic Acids Res* 2004, 32, D235–D239.
20. Gerstein, M. Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census. *Proteins* 1998, 33, 518–534.
21. Wolf, Y.I.; Brenner, S.E.; Bash, P.A.; Koonin, E.V. Distribution of protein folds in the three superkingdoms of life. *Genome Res* 1999, 9, 17–26.
22. Lin, J.; Gerstein, M. Whole-genome trees based on the occurrence of fold and orthologs: Implications for comparing genomes on different levels. *Genome Res* 2000, 10, 808–818.
23. Deeds, E.J.; Hennessey, H.; Shakhnovich, E.I. Prokaryotic phylogenies inferred from protein structural domains. *Genome Res* 2005, 15, 393–402.
24. Yang, S.; Doolittle, R.F.; Bourne, P.E. Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA* 2005, 102, 373–378.
25. Gough, J. Convergent evolution of domain architectures (is rare). *Bioinformatics* 2005, 21, 1464–1471.
26. Caetano-Anollés, G.; Caetano-Anollés, D. An evolutionarily structured universe of protein architecture. *Genome Res* 2003, 13, 1563–1571.
27. Caetano-Anollés, G.; Caetano-Anollés, D. Universal sharing patterns in proteomes and evolution of protein fold architecture and life. *J Mol Evol* 2005, 60, 484–498.
28. Winstanley, H.F.; Abeln, S.; Deane, C.M. How old is your fold? *Bioinformatics* 2005, 21, i449–i458.
29. Swofford, D.L. *Phylogenetic analysis using parsimony and other programs (PAUP*)*, version 4. Sinauer Associates: Sunderland, MA, 1999.
30. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985; 39, 783–791.

31. Nee, S.; Holmes, E.C.; May, R.M.; Harvey, P.H. Extinction rates can be estimated from molecular phylogenies. *Phil Trans R Soc Lond B Biol Sci* 1994, 344, 77–82.
32. Maddison, W.P.; Maddison, D.R. *MacClade: Analysis of phylogeny and character evolution*, version 3.08. Sinauer Associates: Sunderland, MA, 1999.
33. Hartwell, L.H.; Hopfield, J.J.; Leibler, S.; Murray, A.W. From molecular to modular cell biology. *Nature* 1999, 402, C47–C52.
34. Ancel, L.W.; Fontana, W. Plasticity, evolvability, and modularity in RNA. *J Exp Zool (Mol Dev Evol)* 2000, 288, 242–283.
35. Ishitani, R.; Nureki, O.; Fukai, S.; Kijimoto, T.; Nameki, N.; Watanabe, M.; Kondo, H.; Sekine, M.; Okada, N.; Nishimura, S.; Yokoyama, S.J. Crystal structure of archaeosine tRNA-guanine transglycosylase. *J Mol Biol* 2002, 318, 665–667.
36. Harris, T.A. *The theory of branching processes*. Dover Publications: New York, 1963.
37. Copley, R.R.; Bork, P. Homology among $(\beta\alpha)_8$ barrels: Implications for the evolution of metabolic pathways. *J Mol Biol* 2000, 303, 627–640.
38. Nagano, N.; Orengo, C.A.; Thornton, J.M. One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* 2000, 321, 741–765.