

Universal Sharing Patterns in Proteomes and Evolution of Protein Fold Architecture and Life

Gustavo Caetano-Anollés,¹ Derek Caetano-Anollés²

¹ Department of Crop Sciences, 332 NSRC, 1101 West Peabody Drive, University of Illinois, Urbana, IL 61801, USA

² Vital NRG, Knoxville, TN 37919, USA

Received: 16 June 2004 / Accepted: 11 October 2004 [Reviewing Editor: Dr. David Pollock]

Abstract. Protein evolution is imprinted in both the sequence and the structure of evolutionary building blocks known as protein domains. These domains share a common ancestry and can be unified into a comparatively small set of folding architectures, the protein folds. We have traced the distribution of protein folds between and within proteomes belonging to Eukarya, Archaea, and Bacteria along the branches of a universal phylogeny of protein architecture. This tree was reconstructed from global fold-usage statistics derived from a structural census of proteomes. We found that folds shared by the three organismal domains were placed almost exclusively at the base of the rooted tree and that there were marked heterogeneities in fold distribution and clear evolutionary patterns related to protein architecture and organismal diversification. These include 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 perhaps linked to multicellularity.

Key words: Archaea — Bacteria — Eukarya — Organismal diversification — Origins of life — Phylogenetic tracing — Protein structure — Proteome diversification

Introduction

Proteins are fundamental components of our biological world. Their function is determined by how amino acids are arranged in three-dimensional (3D) space and their emergence and diversification are imprinted in both the sequence and the structure of evolutionary building blocks known as protein domains (Aravind et al. 2002; Chothia et al. 2003). While the world of protein architecture can be considered uncharted (Kunin et al. 2003), domains have been unified into a comparatively small set of protein fold designs (Gerstein and Hegyi 1998). Consequently, the protein world appears to be finite ($< 10^4$ folds in size) and slowly curbed by progress in structural genomics and acquisition of the sequence of entire protein complements (proteomes) (Grant et al. 2004).

Most proteins have been formed by gene duplication, recombination, and divergence and the protein world can be studied by matching proteins of known structure to genome sequences (Aravind et al. 2002; Chothia et al. 2003). Protein folds are among the most conserved components in nature, making them good candidates for the study of distant evolutionary relationships. Folds were surveyed in a number of genomes (Gerstein and Levitt 1997; Gerstein 1997, 1998; Frishman and Mewes 1997; Wolf et al. 1999; Hegyi et al. 2002) and indexed in several databases (Lee et al. 2003). Fold composition, measured as presence-absence of individual folds, was used to reconstruct whole-genome trees based on the idea that closely related organisms must share significantly

more fold architectures than distantly related ones (Gerstein 1998; Wolf et al. 1999; Lin and Gerstein 2000). These phylogenetic trees generally resembled ribosomal RNA (rRNA) phylogenies, had generally well supported topologies, and were built using both distance and parsimony methods. While genome (proteome) trees depict organismal diversification (Wolf et al. 2002), a genomic census of architectures can also be used to reconstruct the history of architectural diversification of proteins.

Using a “Hennigian” cladistic approach based on shared and derived features in structure that are descriptive of common descent (Caetano-Anollés 2002), we recently reconstructed a phylogeny of protein architecture (Caetano-Anollés and Caetano-Anollés 2003). We counted the number of genes corresponding to fold architectures in genomes and used these measures of “genomic demography” to map the world of proteins and track both architectural and organismal history directly at the proteome level. Rooted phylogenetic trees of proteomes and fold architectures revealed clear evolutionary patterns, dramatic diversification events in the history of life, and a common ancestor with an architectural-rich and relatively modern organization. All structural classes of globular proteins appeared very early in evolution and in defined order, the α/β class being the first, followed by the $\alpha+\beta$, the all- α , and the all- β classes, and by small (S) and multidomain (M) proteins. This trend was also evident throughout the tree of architectures, supporting the idea that designs with interspersed α -helices and β -sheets were segregated in the course of evolution, first within their structure and then confined to separate molecules (as in the all- α and all- β classes), an observation that is consistent with the random origin hypothesis of proteins (White 1994) and patterns of modularity and simplification in molecular design (Hartwell et al. 1999; Ance and Fontana 1999). The most primitive protein folds shared a common architecture of barrels or interleaved β -sheets and α -helices, and clear trends of architectural transformation were evident in the tree of architectures, such as an evolutionary increase in the curl and stagger of β -barrels in the all- β class. While these patterns relate to global structural changes, they do not address how protein architectures are shared within an evolutionary context throughout the three domains of life.

We here extend our initial studies of the protein world by exploring how protein folds are distributed among and within organismal domains. These distributions were mapped (traced) along the branches of a phylogenetic tree of protein architecture that was reconstructed from measures of fold frequency in individual proteomes, charting organismal and genomic information in structural space. Our approach is based on two fundamental premises: (1)

that protein structure is far more conserved than sequence and carries considerably phylogenetic signal and (2) that biological designs that have been successfully implemented tend to be reused over and over again in nature. A number of considerations support these premises.

Protein structure is directly linked to function and is therefore the subject of natural selection and strong evolutionary constraint (Chothia and Lesk 1986). Consequently, 3D structure is less prone to mutation than sequence and the information embedded in structure persists longer than in primary sequence. In fact, the higher we go in the structural hierarchy of the protein world, the more resilient the architectural designs (Murzin et al. 1985; Orengo et al. 1997). While proteins in protein families are closely related, folds can pool distant relatives with <25% sequence homology (Chothia et al. 2003). These realizations are important. Theoretical considerations suggest that sequence data may be inherently limited in its ability to uncover deep phylogenetic signatures and ancient relationships (Sober and Steel 2002; Penny et al. 2003; Mossell 2003). Primary sequence runs out of useful information when the repeated accumulation of substitutions in nucleotide sites (site saturation) erases evolutionary history. Convergent evolution of nucleotide sites, differing substitution rates among sites and lineages, and nonindependent site substitutions are just few of many other contributing factors (Philippe and Laurent 1998). Consequently, phylogenetic hypotheses describing deep phylogenetic relationships can be sometimes equivocal or incorrect if solely based on primary sequence. Genomics now enables the use of other types of biological information that can be preserved for longer periods of time, such as rare genomic changes (e.g., intron indels, retrotransposon integrations, genome rearrangements [Rokas and Holland 2003]) or structural features in macromolecules (Caetano-Anollés 2002). However, it is difficult to find features that can be applied globally to a wide range of taxonomical levels. We contend that genomic demography of high levels of structural organization in proteins offers the right framework to uncover distant relationships.

From a perspective centered on fitness, architectural designs that had been successfully deployed will have more chances to be reused in other biological contexts, and are expected to become popular. Consequently, robust and well-evolved structures have more chances of withstanding the effects of time and becoming widely used. In this study, we use an initial model of structural evolution in which protein folds that are more prevalent in nature originate from innovations in structure occurring earlier in evolutionary time. The model is based on Okham’s principle of preferring simple explanations to complex ones and is supported by the survey and comparison

of protein folds, statistical analyses of genome sequences, and network behavior of protein fold occurrence (Caetano-Anollés and Caetano-Anollés 2003). The model is simple but is consistent with a process of fold acquisition that obeys a power law and is driven by the gain and loss of protein domains (Huynen and van Nimwegen 1998; Quian et al. 2001; Rzhetsky and Gomes 2001; Karev et al. 2002, 2003). In fact, power law distributions of folds in proteomes are better described by quasilinear birth–death–innovation models that take into account domain family size (Karev et al. 2002, 2003). These models suggest that small protein families are more dynamic, more prone to innovation and elimination, and more recent than large ones. This supports the “polarization” of phylogenetic characters and the rooting of trees in our study. The application of steady-state demographic techniques to the age distribution of duplicate genes in completely sequenced genomes has shown that the average rate of duplication of an eukaryotic gene (0.01/gene/MY) is of the same order of magnitude as the mutation rate per nucleotide site and that the half-life of a gene averaged over many species is small (~4 MY [Lynch and Conery 2000, 2003a]). Variation in duplication rates and genetic effective population size appear to be drivers of genome complexity and a direct consequence of the quasi-equilibrium birth-and-death process (Lynch and Conery 2003a, b). Despite being highly transient, some genes are preserved for long periods of time and contribute to both long-term phenotypic evolution (via subfunctionalization or neofunctionalization processes) and size increase of domain families and fold categories. These genes, together with processes such as domain “shuffling,” are probably the main contributors to innovation in protein architecture and are responsible for converting functionally versatile proteins into specialized forms.

Based on these considerations, our study explores how the architectural protein repertoire was tailored during organismal diversification by gene genesis and innovation, gene loss, and horizontal gene transfer (HGT) events, uncovering patterns in the origin and diversification of protein molecules and life.

Materials and Methods

Genomic Demography and Character Coding. Protein entries matching fold categories in the Structural Classification of Proteins (SCOP) database (Murzin et al. 1995) were retrieved from PEDANT (Frishman et al. 2001) in a set of 32 species with finished genomic sequences encompassing Eukarya (*Arabidopsis thaliana* var. Columbia, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Neurospora crassa* OR74A, *Saccharomyces cerevisiae* S288C, and *Schizosaccharomyces pombe* 972h), Archaea (*Aeropyrum pernix* K1, *Archaeoglobus fulgidus* DSM4204, *Halobacterium* sp. NRC-1, *Methanobacterium thermoautotrophicum* delta H, *Methanococcus*

jannaschii DSM26621, *Pyrococcus horikoshii* OT3, *Sulfolobus solfataricus* P2, *Sulfolobus tokodaii* strain 7, and *Thermoplasma acidophilum*), and Bacteria (*Aquifex aeolicus* VF5, *Bacillus subtilis* 168, *Borrelia burgdorferi* B31, *Chlamydia pneumoniae* CWL029, *Clostridium acetobutylicum* ATCC824, *Deinococcus radiodurans* R1, *Escherichia coli* K12, *Haemophilus influenzae* Rd [KW20], *Helicobacter pylori* J99, *Mycobacterium tuberculosis* CDC1551, *Mycoplasma genitalium* G-37, *Mycoplasma pneumoniae* M129 [ATCC29342], *Rickettsia prowazekii* MadridE, *Staphylococcus aureus* N315, *Synechocystis* sp. PCC6803, *Thermotoga maritima* MSB8, and *Treponema pallidum* Nichols). SCOP release 1.59 cluster PDB structural entries into 686 fold categories depicting evolutionary and structural relationships (Lo Conte et al. 2002). Fold recognition in PEDANT involves 10 PSI-BLAST iterations using each SCOP domain against a nonredundant protein sequence database, the construction of an IMPALA-based SCOP profile library, and its use to search genomic sequences (Frishman et al. 2001).

For phylogenetic analysis, each genome was described by a set of linearly ordered multistate characters representing the frequency with which the fold occurs in the proteome. This frequency of fold architectures was termed *genomic abundance* (*G*). The gap-recoding technique of Thiele (1991) developed for the analysis of morphometric data was used to normalize *G* and compensate for differences in genome size and proteome representation. Briefly, a rescaling function was used to rescore character information while retaining information on both rank order and size of gaps between character states. Raw frequency data were ranked as ordered sets of states, standardized using a square root transformation to account for unequal variances, and range standardized to an arbitrary 0–20 scale (compatible with most phylogenetic programs). This ensured that differences within and between characters were equalized with the transformation and range-standardization procedure and that data would approximately fit a power-law with exponents of the order of 2, i.e., matching decay gradients for the vast majority of genomes analyzed. Once coded, characters were aligned in ordered columns and subjected to phylogenetic analysis.

Phylogenetic Analysis. A universal tree of protein architecture was reconstructed from the occurrence of folds in proteomes using maximum parsimony (MP) as the optimality criterion and directed character states in PAUP* (Swofford 1999). The method of tree reconstruction was described previously (Caetano-Anollés and Caetano-Anollés 2003). Characters were polarized with the ANCMETHODS command by assuming that the number of proteins in a genome that exhibits a particular fold increases in the course of evolution, with character state transformation proceeding from states of higher frequency to states of lower frequency (see Discussion for rationale). Trees were automatically rooted at the point where the hypothetical ancestor connects to the tree. Character argumentation is supported by the power-law behavior of fold occurrence (Huynen and van Nimwegen 1998) and is consistent with birth–death–innovation models that describe a preference for duplication of genes encoding already common features and the “rich-get-richer” growth of the protein world (Quian et al. 2001; Rzhetsky and Gomes 2001; Karev et al. 2002, 2003). When ranking genome components by their occurrence, we consider that those protein families that grew early in evolution will be prominent in many genomes and that the number of family members increases in single steps corresponding to the addition or removal of an 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 nonparametric bootstrap method (Felsenstein 1985), generally implemented using 2×10^3 pseudoreplicates in PAUP*, and by double decay (DD) analysis using RADCON (Thorley and Page 2000). Ensembles of DD-derived reduced cladistic consensus (RCC) support trees were

used to measure strengths and weaknesses of phylogenetic hypotheses (Wilkinson et al. 2000). The structure of phylogenetic signal in the data was tested by the skewness (g_1) of the length distribution of at least 10^4 random trees and permutation tail probability (PTP) tests of cladistic covariation using at least 10^3 replicates. Ensemble consistency (CI) and retention (RI) indices were used to measure homoplasy and synapomorphy.

Fold Distribution Among Organismal Domains and Cumulative Frequency Plots. We classified protein folds into categories that describe the spread of architectures across the three organismal domains of life, Eukarya, Archaea, and Bacteria. A fold that appears in at least one proteome but in all organismal domains was assigned to the EAB category, one that is only present in Archaea and Bacteria to the AB category, and so on. Cumulative frequency plots were used to depict order and rate of appearance of fold distribution categories. Cumulative fold number was given as a function of distance in nodes from a hypothetical ancestral fold on a relative 0–1 scale, termed “*node distance*” (nd). These plots can be considered time plots of lineages (Nee et al. 1994) with a time axis defined in relative units (cladogenic events). Fold distribution categories and a *distribution index* (f) that describes the distribution of individual folds among proteomes within each category were traced along the branches of the tree of protein architecture. The f index represents the fraction of proteomes harboring a fold within a category and ranges from absence ($f = 0$) to presence in all proteomes analyzed ($f = 1$). The f values were sometimes averaged over all folds within a category or over folds with a same nd (\bar{f}), and ANOVA used to evaluate statistical significant differences. The number of protein families associated with each fold in SCOP was also analyzed to study how unifolds, mesofolds, and superfolds (Coulson and Moulton 2002) were distributed in the tree of architectures. Values were averaged for folds in individual fold distribution categories or fold groups.

Character Tracing. Fold distribution categories and the distribution index were traced along the branches of the tree of architecture in MacClade (Maddison and Maddison 1999), using algorithms for Wagner and square-change parsimony, respectively. Wagner parsimony minimizes the sum of the absolute values of changes along the branches of the tree using a linear parsimony criterion (Swofford and Maddison 1987), and squared-change parsimony reconstructs the ancestral states of continuous-valued characters (Maddison 1991). Note that squared-change parsimony minimizes the sum of the squared changes on the branches and can be considered a Bayesian probability estimate under a Brownian motion model of evolution. Methods are “agnostic” in that they justify for example the existence of both increases and decreases in continuous-valued characters along branches of the tree.

Results

Tracing Protein Fold Distribution Across Organismal Domains of Life

The monophyletic nature of Eukarya, Archaea, and Bacteria in a universal phylogeny of proteomes inferred from fold occurrence data (Caetano-Anollés and Caetano-Anollés 2003) supports the existence of three organismal domains in life (Woese et al. 1990). We therefore classified protein folds into categories that describe their distribution across these domains. Categories include folds common to all (EAB), a

subset of (EB, EA, and AB), or individual (E, B, and A) domains. Fold distribution categories and an index (f) describing the popularity of individual folds among proteomes in each category were traced in a phylogenetic tree of protein architecture and the timing and rates of appearance used to signal evolutionary patterns in the history of proteins and life. The tree used in this study was reconstructed from the occurrence of folds in 32 proteomes using MP and directed character states. Note that the tree is intrinsically rooted, that leaves correspond to folds and nodes to architectural diversification events, and that nodes close to the base of the tree reflect more ancient events than those close to the leaves.

Clear Fold Distribution Patterns in the Proteome World

Cumulative frequency plots were used to depict order and rate of appearance of fold distribution categories in the tree (Fig. 1). Cumulative fold number was given as a function of distance in nodes (nd) from a hypothetical ancestral fold. These representations were quite robust, as suboptimal tree reconstructions showed that systematic and random error did not affect the rates of fold accumulation substantially.

Frequency plots showed that the first folds to appear in evolution were those that were common to Eukarya, Archaea, and Bacteria (Fig. 1A). These early EAB architectures were highly represented in the 32 proteomes analyzed ($\bar{f} = 0.83$) and included most folds belonging to a genetic core of universal genes exhibiting the same phylogenetic pattern as rRNA (Harris et al. 2003) (Table 1). Most EAB folds were confined to the base of the universal tree (Fig. 2) and could be considered superfolds and mesofolds, based on the number of protein families associated with them (Fig. 3). The average number of protein families per fold was consequently high, 5.72 ± 0.56 (SE) for all EAB folds and 9.63 ± 1.34 (SE) for EAB folds at the base of the tree (Figs. 2 and 3). Figure 4 shows a subtree reconstructed from 17 of the most primitive folds (most of them superfolds) spanning the 0–0.12 nd range, using directed and undirected multistate characters. The example illustrates how tree topologies reconstructed from multistate characters are not significantly altered by character polarization. Interestingly, polarizing the character transformation series in the opposite direction, results in a less parsimonious reconstruction (1185 versus 1066 steps) (see legend to Fig. 4), an observation that supports character argumentation. This same behavior was obtained when reconstructing several other subtrees of the tree of architectures (data not shown). The number of protein families within individual fold architectures was traced in the subtree of

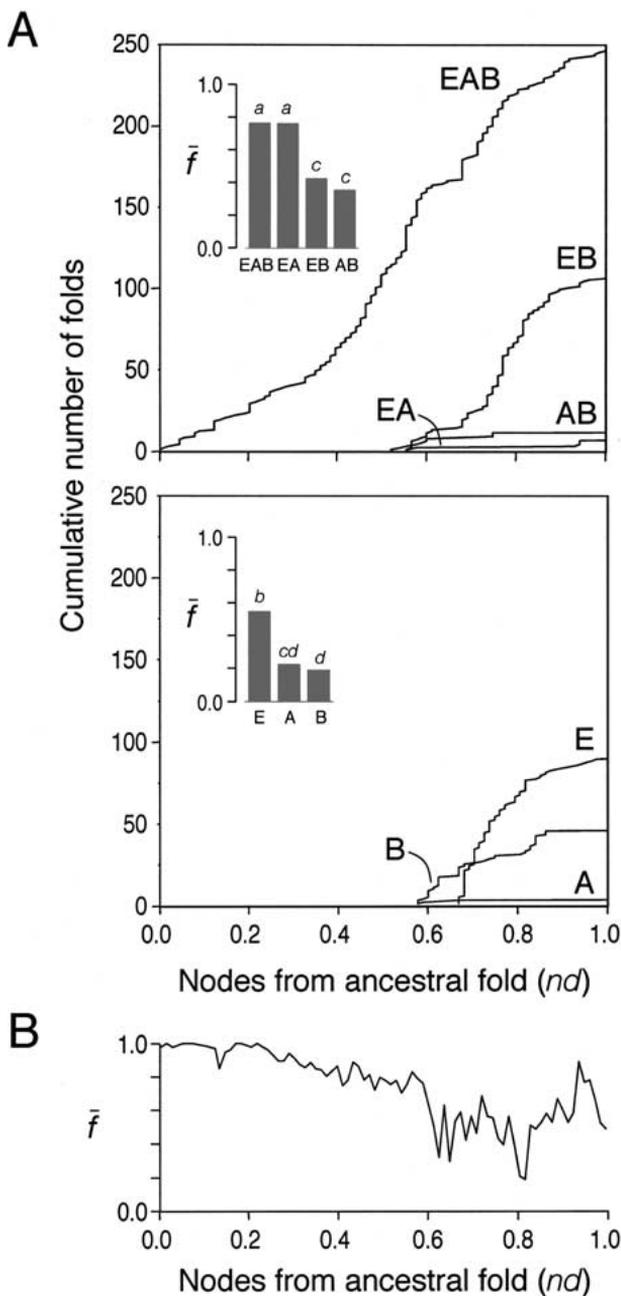


Fig. 1. Cumulative frequency distribution plots. **A** Plots illustrate the accumulation of folds belonging to different fold distribution categories along an optimal phylogenetic tree of protein architecture. Cumulative fold number is given as a function of distance (nd) in nodes from a hypothetical ancestral fold. Inset graphs show how architectures are widely distributed within each fold distribution category. Bars headed by the same letter are not significantly different ($p < 0.05$; Fisher's protected LSD). **B** Change of \bar{f} along the phylogenetic tree. \bar{f} was averaged over folds with the same nd value. The peak at $nd > 0.94$ is due to several widely distributed E and EA folds (including histone and basic HLH DNA-binding domains).

ancient architectures and the value for the ancestral fold reconstructed at the base of the tree (Fig. 4C). Based on extant protein information, the ancestor of the tree of architectures appears to have had 31 protein families associated with it.

Table 1. Folds linked to clusters of orthologous groups (COGs) believed associated with a core of universal genes

Group ^a	Fold	FD ^b	nd	COG group ^c	
I	c.23	EAB	0.045	COG1 COG5	
	a.4	EAB	0.045	COG1 COG2 COG3	
	d.58	EAB	0.045	COG1 COG3 COG5	
	c.3.1	EAB	0.057	COG5	
	c.26	EAB	0.080	COG3 COG4 COG5	
	b.40	EAB	0.080	COG1 COG5 COG6	
	c.67.1	EAB	0.080	COG2	
	c.69.1	EAB	0.125	COG1	
	c.47	EAB	6.125	COG6	
	c.55	EAB	0.125	COG1	
	d.104.1	EAB	0.170	COG5	
	b.43	EAB	0.193	COG1	
	c.8	EAB	0.204	COG5	
	c.68.1	EAB	0.205	COG5	
	a.2	EAB	0.239	COG5	
	g.41	EAB	0.239	COG5	
	c.51	EAB	0.250	COG5	
	a.27.1	EAB	0.250	COG5	
	a.60	EAB	0.205	COG3	
	II	d.14.1	EAB	0.261	COG1
d.41		EAB	0.511	COG1	
d.47.1		EAB	0.534	COG1	
b.39.1		EAB	0.557	COG1	
a.75.1		EAB	0.557	COG1	
d.55.1		EAB	0.557	COG1	
d.141.1		EAB	0.557	COG1	
d.140.1		EAB	0.557	COG1	
d.131.1		EAB	0.557	COG1 COG3	
a.16.1		EAB	0.568	COG1	
e.24.1		EAB	0.580	COG1	
c.84.1		EAB	0.614	COG5	
a.129.1		EAB	0.625	COG5	
d.56.1		EAB	0.648	COG5	
d.50		EAB	0.682	COG1	
d.67		EAB	0.682	COG5	
b.51.1		EAB	0.682	COG5	
d.127.1		EAB	0.682	COG2	
b.34		EAB	0.693	COG1 COG3	
d.26		EAB	0.716	COG5	
c.53	EAB	0.750	COG3		
d.66.1	EAB	0.761	COG1 COG5		
III	a.97.1	EAB	0.523	COG5	
IV	d.129	EAB	0.580	COG5	
VI	d.77.1	EB	0.773	COG1	
	d.12.1	EB	0.773	COG1	
	c.22.1	EB	0.773	COG1	
	d.28.1	EB	0.784	COG1	
	c.21.1	EB	0.784	COG1	
	c.20.1	EB	0.784	COG1	
	c.12	EB	0.784	COG5	
	VIII	a.26.1	B	0.818	COG5
	IX	g.39.1	EB	0.943	COG1
		d.17	EAB	0.909	COG6

^aGroups correspond to clades defined in Fig. 2.

^bFD, fold distribution category.

^cCOG1, proteins and translation factors with 3-domain phylogeny; COG2, ribosome associated proteins with 3-domain phylogeny; COG3, transcription and replication proteins with 3-domain phylogeny; COG4, proteins of unknown function with 3-domain phylogeny; COG5, conserved proteins that do not exhibit 3-domain phylogeny; COG6, universal COG protein families of unknown phylogeny (Harris et al. 2003).

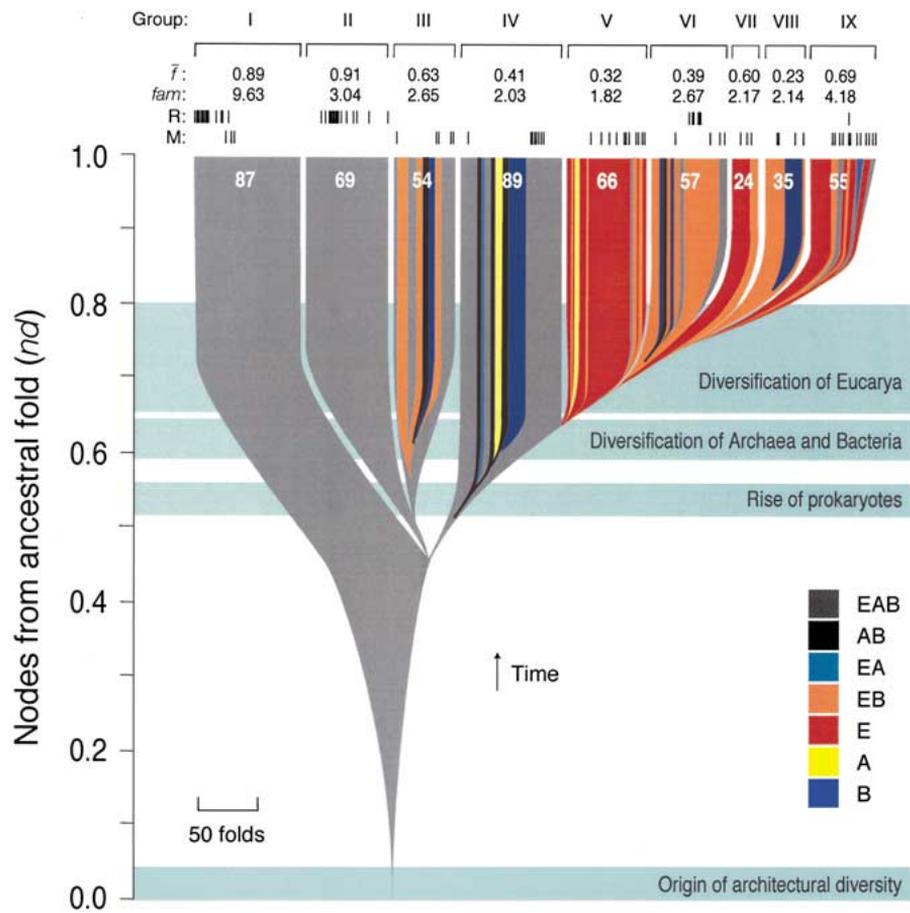


Fig. 2. Tracing fold distribution categories in the tree of architectures. The diagram depicts a phylogenetic tree of fold architectures in which the distribution of folds among different organismal domains is traced and labeled with different colors. Nine groups of folds are indicated in roman numerals together with corresponding \bar{f} values, average number of protein families per fold (f_{am}), and number of folds in each group. Bars on top of the tree indicate folds believed to be associated with a universal core of genes exhibiting the same phylogenetic pattern as rRNA (R) and with multicellularity (M) (see Tables 1 and 2).

In contrast, the second derived half of the inferred evolutionary tree was mainly composed of folds belonging to one or subsets of organismal domains (Fig. 2). These folds were patchily distributed among proteomes ($\bar{f} = 0.51$), perhaps resulting from architectural innovation, reductive evolution, and HGT. This trend toward patchiness was evident when plotting \bar{f} values as function of nd (Fig. 1B). Most folds in the second one half of the tree were unifolds, composed of only one or two protein families (Figs. 2 and 3). Interestingly, mesofold and superfolds common at the base of the tree recurred later in evolution, but at a low frequency.

Tracing distribution categories in the tree showed interesting patterns. First, noncommon categories appeared in defined order, AB being first ($nd \geq 0.522$), followed by EA and EB ($nd \geq 0.577$), then by A and B ($nd \geq 0.579$), and finally, by E ($nd \geq 0.670$) (Fig. 1). Rates of appearance were also illustrative; especially those related to the E and EB categories, which increased late but quite massively. Second, categories were represented in proteomes at varying levels (ANOVA, $p < 0.0001$). While EAB, EA, and, to a lesser degree, E folds were highly represented within their categories, all the rest were sparsely distributed (Fig. 1A, insets). Finally, folds in the different cate-

gories were clearly clustered in the tree, suggesting episodes of protein diversification (Fig. 2).

Topological features in the tree were used to classify folds into different groups (Fig. 2). Two major clades emerged about halfway in evolution ($nd \sim 0.43$), one defining two clear subclades (groups II and III) and the other defining several subclades (> 20 nodes each) composed mostly of folds shared by proteomes in one or two organismal domains (groups IV–IX). Note that the emergence of these two major clades coincides with the onset of organismal diversification.

The Rise of Prokaryotes

The very early appearance of architectures in the AB category (initially in group IV; Fig. 2) suggests a prokaryotic lineage common to Archaea and Eukarya. These AB folds are currently associated for example with iron-uptake systems (a.76), cobalamine synthesis (c.39), oxidation in specialized glycolytic pathways (d.152), DNA bending (a.55), and chemotaxis (c.40, a.58). Phylogenetic analysis of folds appearing within the 0.52–0.56 nd range is shown in Figure 5. The tree shows a topology that defines the

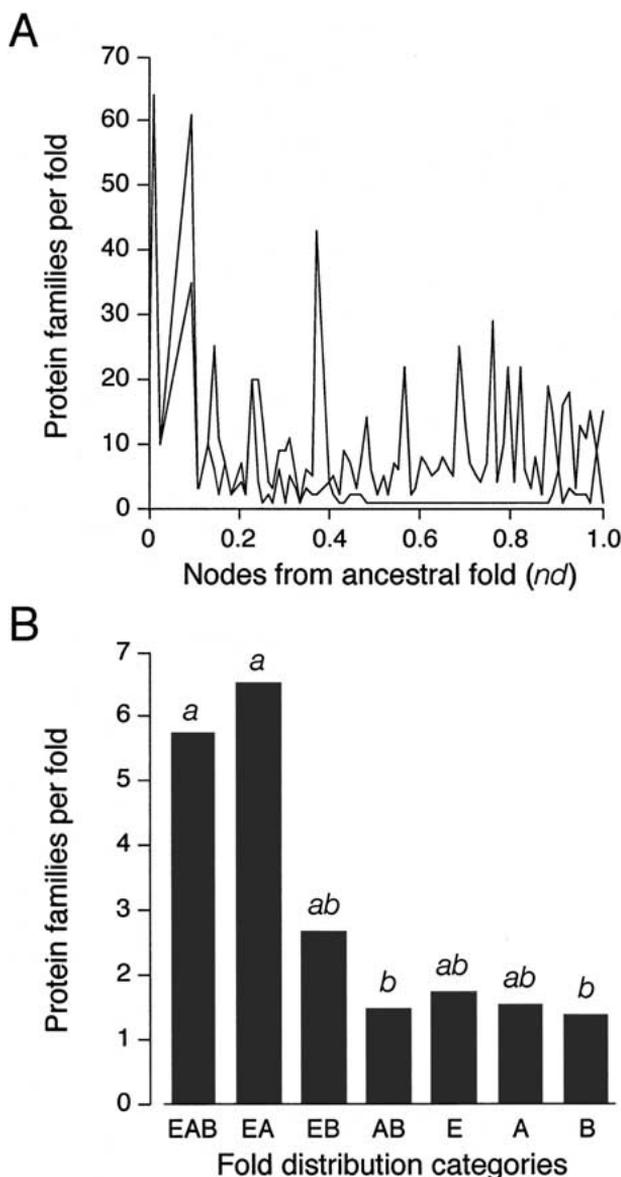


Fig. 3. The evolutionary expansion of protein families along the tree of architectures. The plot illustrates how the numbers of protein families in fold categories change along an optimal phylogenetic tree of protein architecture. The number of families per fold is given as a function of distance (*nd*) in nodes from a hypothetical ancestral fold and, when needed, was given as maximum and minimum values observed over folds with the same *nd* value. Most folds within the 0.4–1.0 *nd* range (time frame) had only one or two families associated with them and could be defined as unifolds. **B** Bar graphs show the average numbers of families per fold for individual fold distribution categories. Bars headed by the same letter are not significantly different ($p < 0.05$; Fisher's protected LSD). Most superfold and mesofolds were found in the EAB and EA categories.

onset of the major fold divide described above (groups II and III vs. group IV). The topology was well supported by the RCC trees derived from DD analysis. Curiously, the appearance of AB folds during this transition phase was accompanied by a sharp decrease in *f* values, suggesting a marked episode of architectural loss in the prokaryotic ancestors.

Architectural Loss and Diversification in Archaea and Bacteria

AB folds appearing in group IV, and subsequently in group III, had a very low collective \bar{f} value (0.37) compared with surrounding EAB folds. Furthermore, their gradual appearance was accompanied by the emergence of a group of 19 EA and EB folds. Interestingly, EB folds were confined to group III and EA folds to group IV. This could signal episodes of architectural loss in the prokaryotic ancestors of Archaea and Bacteria, respectively, induced perhaps by genome reduction. Figure 6 shows folds appearing within the 0.58–0.60 *nd* range. This diversification time frame includes the origin of first folds unique to Archaea and Bacteria and shows a marked decrease in the *f* values of associated folds. The pronounced dip of *f* values starting at about 0.6 *nd* unit in the tree (Fig. 1B) and the low \bar{f} value averaged for folds in clade IV (Fig. 2) are consistent with a marked diversification episode associated with gene loss. Folds unique to Archaea and Bacteria that followed immediately could also result from episodes of architectural innovation (presumably coupled with HGT). Note that folds in the A and B categories were those with the lowest \bar{f} (Fig. 1A, inset), showing the distinct behavior of prokaryotic proteomes. Interestingly, over half of all A and B folds arose during this period of diversification and B folds in clades III had significantly higher *f* values (0.65) than those appearing in clades IV and VIII (0.10). These differences may indicate different mechanisms of diversification in these two architectural groups.

Late Rise of Architectural Novelty in Eukarya

Folds unique to Eukarya appeared quite late in the evolutionary tree ($nd > 0.670$) (Fig. 1) and massively (especially in group V; Fig. 2). Proteins believed linked to multicellularity (Chervitz et al. 1998; Copley et al. 1999; Patthy 2003), including those involved in intra- and intercellular signaling and cell death programs (e.g., necrosis, apoptosis), generally contained multiple domains within only 50 fold categories (Table 2), 34 of which arose during the Eukaryal diversification phase (Fig. 2). Interestingly, 14 folds were eukaryotic novelties and 12 folds originated immediately after prokaryotic diversification events, most of which were EAB folds. One of these folds is the LysM domain, a fold that was recently linked to a Nod factor receptor in plants mediating the establishment of the *Rhizobium*–legume nodule symbiosis (Limpen et al. 2003). Note that 20 folds of the total fell within the EAB category, but only 4 were really ancestral (b.1, a.7, c.62, and d.144). Many of the folds associated with multicellularity, especially those in

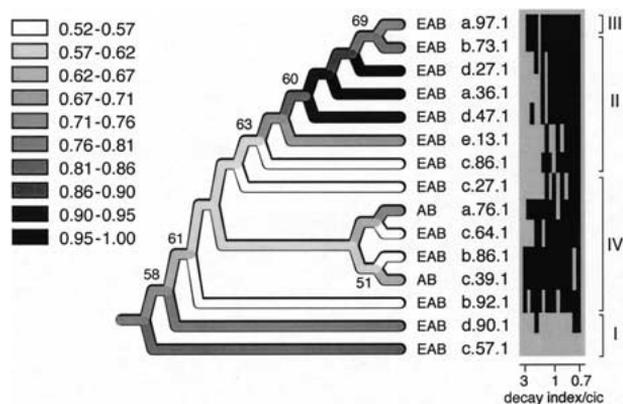


Fig. 5. Tracing the popularity of folds within each distribution category (f) during the rise of prokaryotic architectures. A phylogenetic tree of folds appearing within the 0.52–0.56 node distance (nd) range was reconstructed. Six optimal most-parsimonious trees (775 steps; CI = 0.823, RI = 0.549; $g_1 = -0.374$; PTP test, $p = 0.001$) were recovered from a branch-and-bound search. The tree shown is congruent with the 50% majority-rule consensus. Roman numerals describe fold groups. The graph describes the identity of terminal taxa joined by the reduced cladistic consensus (RCC) support trees (bars) derived from double decay (DD) analysis. Within the 15 RCC topologies, total decay ranged from 5 to 33 and cladistic information content (cic) values ranged from 1.5 to 43. RCC topologies are presented in order, starting with the most informative (i.e., with higher decay-to-cic ratios). Bootstrap values > 50% are shown for individual nodes. Square root parsimony was used to reconstruct ancestral f states as continuous characters in the phylogenetic tree, using McClade with the rooted tree option. Reconstructed f states are given as gray-scale shadings and as values for selected nodes.

(Murzin et al. 1995), cross organismal domain boundaries along the branches of a phylogenetic tree of protein architecture. The trees we use for phylogenetic “tracing” of these distribution patterns reconstruct the evolutionary history of protein diversification using data generated from a genomic census of architecture. Consequently, they depend on (1) the accuracy and balance of genomic databases, (2) efficient and accurate assignment of structures to proteins, (3) adequate structural classification schemes, and (4) the methods of phylogenetic tree and character state reconstruction. The survey of proteomes in terms of structural components can be influenced by biases in the databases and sampling error (Gerstein and Hegyi 1998). Biases include over- and underrepresentation of certain sequences and structures in the protein repositories. PDB databanks are especially biased by preferences of individual investigators for targets and organisms and physical constraints imposed by crystallography and NMR spectroscopy. Structural assignments have been quite empirical and cover only a fraction of the proteome. However, the use of advanced hidden Markov models (HMM) and threading techniques has advanced the mapping of protein folds to more than half of protein-encoding genes in currently available genome sequences (Grant et al. 2004). The number of “orphan” sequences awaiting structural assignments

will most surely decrease with further advances in structural genomics, and this benefits our approach. Similarly, progress in the number of completely sequenced genomes, now approaching 200 and yielding over 1 million protein sequences, will only enhance genomic demography estimates and widen taxonomical coverage. Our study is based on SCOP, a robust protein classification scheme supported by structural and evolutionary considerations (Murzin et al. 1995; LoConte et al. 2002). We do not expect that the operational definition of fold be seriously challenged in the near future, even though many folds may be better described by “continuous” rather than “discrete” distributions in structure space (Harrison et al. 2003). Comparison of releases 1.43 and 1.59 of SCOP showed that revisions did not affect substantially the reconstructed trees (Caetano-Anollés and Caetano-Anollés 2003). However, challenges to the monophily of specific fold architectures should be addressed appropriately with rigorous studies, such as those characterizing the TIM barrels (Copley and Bork 2000; Nagano et al. 2002), as well as the existence of convergent evolutionary events that could place unrelated protein families within nonnatural fold categories. Our study makes no effort to explore such instances and places trust on the accuracy of SCOP assignments. In this regard, our approach considers an individual fold as a collection of proteins undergoing different but concomitant evolutionary processes that translate into patterns of recent (close relationship) or ancient origin (distant relationship). Finally, phylogenetic methods affect the reconstruction of trees. We use MP as the optimality criterion, i.e., we prefer solutions for reconstruction of phylogenies and ancestral character states that require the least amount of change. The task is computationally demanding, as trees have over 500 leaves and are reconstructed from multistate characters. Consequently, it is unrealistic at this time to use even more demanding model-based methods such as maximum likelihood (ML) or Bayesian approaches.

MP works well when change is rare and tree branches are short but does not take into account branch length and can generate “long-branch attraction” artifacts (Felsenstein 2004). However, the use of large trees can mitigate some of these effects. Furthermore, MP can be an appropriate criterion that can outperform ML under certain circumstances (Steel and Penny 2000). MP is precisely ML when character changes occur with equal probability but rates vary freely between characters in each branch. This Poisson-like model with “no common mechanism” can be particularly useful when there is limited knowledge about underlying mechanisms linking characters to each other (Steel and Penny 2000). Furthermore, the use of a huge character state space,

Table 2. Folds linked to multicellularity

Group	Fold	FD	<i>nd</i>	Representative processes/proteins ^a	
I	b.1	EAB	0.398	EM, integrin, cadherin, fibronectin, tenascin, immunoglobulins	
	a.7	EAB	0.432	AP, silencer of death (BAG domain)	
	c.62	EAB	0.466	CC, von Willebrand factor	
III	a.20	EB	0.556	MM, gelatinases	
	d.7	EB	0.704	AR, receptor-like kinases (LysM domain), lysis motif	
	d.144	EAB	0.716	GF, glycogen synthase kinase, tyrosine kinase	
	b.47	EAB	0.75	CC, complement	
IV	b.36	EAB	0.761	AP, PDZ domain-containing proteins, interleukin	
	a.50	EAB	0.625	CC, complement	
	b.22	EAB	0.636	TNF-like, apoptosis, complement	
	g.1	EAB	0.648	Insulin-like growth factor-binding proteins	
	a.102	EAB	0.466	CC, complement	
	g.32	EAB	0.659	CC, prothrombin	
	g.19	EAB	0.659	Sea anemone toxin (ShK toxin)	
	g.24	EAB	0.659	AP, TNF (tumor necrosis factor), cytokines	
V	g.32	EAB	0.659	CC, coagulation factors (GLA domain)	
	g.7	E	0.682	GF, TGF- β 1-binding protein	
	d.40	E	0.704	CC, serine protease inhibitors	
	b.77	E	0.716	CC, mannose-binding protein-associated serine protease	
	b.8	E	0.727	AP, TRAF proteins	
	g.14	E	0.727	MM, CC, kringle-like, gelatinases, plasminogen	
	g.17	E	0.738	GF, nuclear hormone receptor	
	g.18	E	0.761	CC, complement	
	e.1	E	0.772	Serpins	
	g.8	E	0.772	EM, small Kunitz-type inhibitors	
	a.123	E	0.818	Nuclear hormone receptors	
	d.169	EB	0.818	EM, electins, collagen IV	
	VI	b.3	E	0.761	EM, prealbumin
		b.71	E	0.875	CC, decay-accelerating factor
c.44		EAB	0.863	Phosphotyrosine protein phosphatases	
b.68		EAB	0.909	Low-density lipoprotein receptor	
VII	g.52	E	0.773	AP, TAP repeat	
	b.66	E	0.806	MM, gelatinases	
	b.42	EB	0.795	Cytokines	
VIII	b.60	EB	0.818	MM, gelatinases	
	a.77	EB	0.818	AP, caspase recruitment, CARD, DED, DEATH domain	
	a.26	B	0.818	GF, interferons, short-chain cytokines	
	c.17	EB	0.898	GF, caspases, apoptosis	
IX	b.26	EB	0.852	Phosphotyrosine binding domain	
	a.91	EB	0.852	RGS regulator, regulator of G-protein signaling	
	d.92	EAB	0.92	AP, astacin, remodeling extracellular matrix	
	b.62	EAB	0.92	Cysteine endopeptidases, metalloproteases	
	g.3	E	0.932	CC, selectins, complement	
	b.55	EA	0.932	Phosphotyrosine-binding domain	
	g.39	EB	0.943	Glucocorticoid receptor-like	
	a.38	E	0.954	Basic helix–loop–helix (bHLH)-containing proteins	
	c.10	EB	0.966	CC, von Willebrand factor	
	b.29	EAB	0.977	AR, galectin/lectins, laminins	
	d.42	E	0.977	Elongin (BTB/POZ domain)	
d.3	EAB	0.987	AP, caspase-related Cys endopeptidases		

^aAP—apoptosis, cell death, and intracellular signaling; AR—adhesion and recognition; CC—coagulation and complement system; EM—extracellular matrix constituents; GF—growth factors and binding proteins; MM—matrix metalloproteases involved in remodeling of extracellular matrix.

2003), most of which contained many protein families and were mesofolds and superfolds (*sensu* Coulson and Moutl 2000) (Figs. 2 and 3). It is not surprising that ancient architectures were those commonly shared; they had more time to spread through vertical

descent in the proteome world. What is amazing is the incredible resilience of these architectural designs, capable of surviving billions of years of evolutionary change. In contrast, the second derived half of the tree of architectures was mainly composed of folds

that belonged to one or subsets of organismal domains, were patchily distributed among proteomes, and appeared in defined order. Most of these folds were unfolds associated with only one protein family. We believe they were the result of architectural innovation, reductive evolution, and HGT events.

The marked heterogeneities in fold distribution observed in the tree of architectures were unforeseen and suggest architectural innovation preceded organismal diversification. In fact, results suggest that 21% of folds had already been “discovered” by nature at the onset of organismal diversification and that an additional 10% were derivatives that evolved directly from them (in clades I and II) and remained widely distributed among organismal domains. Note that we defined the ancestral condition for folds in our model as being “popular” within a proteome (i.e., highly represented in relation to other folds) and not being “widely shared” between the proteomes of the organisms examined. Consequently, evolutionary patterns should not be considered artifacts stemming from tautology but true depictions of the spread of architectures through evolutionary transects. Note that all superfolds and highly represented mesofolds were expected to appear at the base of the tree and to be shared by all organismal domains. In fact, this is what was observed (Fig. 3). However, family expansions within folds occurred throughout the tree of architectures, suggesting that it is a phenomenon that is somehow unrelated to the representation of folds in proteomes and to the extent of fold distribution in life.

The very early appearance of architectures in the AB category arising from a world of common folds suggests the birth of a prokaryotic lineage common to Archaea and Bacteria and a sister-group relationship of these prokaryotic domains. This matches the topology of the universal tree of proteomes (Caetano-Anollés and Caetano-Anollés 2003). Immediately following the AB folds, a group of EB and EA folds with moderate to low f values appeared to be confined to separate clades in the tree. The emergence of these folds signals episodes of architectural loss in the prokaryotic ancestors of Archaea and Bacteria, induced perhaps by genome reduction events affecting different architectural sets in the two lineages. Interestingly, folds unique to Archaea and Bacteria also appeared during this diversification time frame, always coupled with a marked decrease in the f values of associated folds. Gene loss has been cited as the most important factor shaping genome content in prokaryotes, followed by HGT and gene genesis (Snel et al. 2002; Kunin and Ouzounis 2003; Daubin et al. 2003). There are good reasons to believe that architectural loss was a dominant force during this first organismal diversification episode that started 0.52 nd unit

from the most ancestral fold. The episode was probably triggered by environmental influences and driven by the establishment of different selection strategies in the emerging organismal lineages. While organisms under K -selection probably took advantage of the carrying capacity of the environment, the early ancestors of prokaryotes resorted to r -selection pressures that favored rapid growth in periods of nutrient availability (Carlile 1982) or strategies that would favor colonization of hostile environments (Grime 1977). This forced the adoption of a more efficient and agile lifestyle, the expansion of effective population size by reducing organismal size, and the streamlining of genomic constitution (Conery and Lynch 2003b). Consequently, the patchiness of fold architectures observed during this period was probably the consequence of extensive and differential loss of genes in the emerging prokaryotic lineages, a process that occurred concomitantly with the discovery of new functions and architectures.

We found that protein novelties unique to organismal lineages emerged late and in defined order during evolution. Under this scenario, prokaryotes preceded eukaryotes. Note that folds unique to Eukarya originated quite late in the evolutionary tree, well after the diversification of Archaea and Bacteria. We believe that their massive appearance was linked to evolutionary novelties related to the emergence of multicellularity. The evolution of multicellularity probably involved modular assembly of domains from numerous extracellular matrix proteins and intracellular and extracellular signaling proteins (Chervitz et al. 1998). Most folds containing protein domains believed to be linked to multicellularity (Chervitz et al. 1998; Copley et al. 1999; Patthy 2003), such as those involved in programmed cell death (e.g., necrosis, apoptosis), appeared following prokaryotic diversification events and during the eukaryal diversification phase (Fig. 2). Interestingly, many folds that follow fold clusters unique to prokaryotes were common EAB architectures. Most of these folds describe domains that support extracellular signaling functions (Table 2) and could have been recruited to combine with domains unique to Eukarya. Perhaps these folds represent adaptations to innovations stemming from the prokaryotic world, and if so, they could support a symbiotic origin of multicellularity. There appears to be strong yet circumstantial evidence of microbial symbiosis influencing the evolution of multicellular organisms (McFall-Ngai 2001). Prokaryotes have the ability to behave as multicellular organisms, such as in quorum sensing and microbial cell differentiation phenomena, and multicellular eukaryotes can share pathways of responses, communicate, and establish symbiotic interactions with prokaryotes. For example, extensive responses can be triggered by bacterial

quorum sensing signals in plants and animals, and in turn, plant quorum sensing “mimics” can interfere with inter-bacterial communication (Bauer and Mathesius 2004). Interestingly, responses induced by quorum sensing *N*-acyl homoserine lactones in the legume *Medicago truncatula* (Mathesius et al. 2003) involved proteins that could be assigned to 30 fold architectures, 87% of which were EAB folds and most of which were basal (within groups I and II) in the tree of architectures (in preparation). Only 13 folds were derived and these were clearly associated with fold clusters unique to prokaryotes (groups IV and VI) or with folds linked to multicellularity (groups IX). Consequently, eukaryotic responses to bacterial communication signals involving folds of recent origin appear to follow instances of prokaryotic innovation.

The suggestion that the ancestors of Archaea and Bacteria preceded those of Eukarya tells little about the nature of the common ancestor of diversified life. Our results suggest however that the proteomes of these ancestors shared already a quite diverse arrangement of molecular architectures. These architectures appear functionally versatile, since folds at the base of the tree of architectures harbor many enzymatic functions (Caetano-Anollés and Caetano-Anollés 2003). Our view is consistent with a primitive proto-eukaryote (Glansdorff 2000; Poole et al. 1998) responsible for “crystallizing” diversified life (Woese 2000). It is also consistent with a limited role of HGT (Glansdorff 2000). This does not mean that HGT had not been rampant, especially within prokaryotes; organisms actually share conservatively only a very small proportion of gene sequences (Gogarten et al. 2002). Instead, the existence of clear evolutionary patterns in the data suggests that HGT had minimum homogenizing effect at the high levels of structural organization characteristic of protein folds. While fold architectures were lost or invented, HGT shuffling must have had little influence on the birth–death–innovation kinetics driving their accumulation.

Rivera and Lake (2004) recently proposed eukaryotes emerged from a fusion of archaeal and bacterial genomes. Our evolutionary tracings are compatible with this prokaryotic fusion hypothesis, since architectures unique to Archaea and Bacteria arose earlier than those unique to Eukarya. However, the ring graph used to infer the putative fusion event was unrooted, and alternative explanations should be considered. For example, differential loss of genetic repertoires could result in a proto-eukaryote “fission” that would open the “ring of life” to streamlined archaeal or bacterial prokaryotes. This could explain the rise of prokaryotic fold diversity associated with gene loss in our trees. In any event, fusions and fissions ultimately represent cataclysmic or progressive homogenizing forces that rival HGT and act as

genetic scaffolds for the generation of structural diversity.

Conclusions

The patterns of molecular and organismal diversification observed here are based on demography of fold architecture in proteomes. Consequently, evolution is not described as changes in structural character states (*sensu* Caetano-Anollés 2002) but rather as the accumulation of variants within a structural “neighborhood.” Within this framework, evolution of protein architecture can be explained by a verbal model that invokes the metaphor of a discrete multidimensional fitness landscape (Wright 1932; Kauffmann 1993; Gavrillets 1997) and describes the sequence space of the protein world. The ruggedness of this landscape (i.e., the frustration of the system) is dynamic and determines the nature of adaptive diffusive walks toward highly evolved molecular functions that occur during evolution and result from the general mapping of sequence into structure. These adaptive walks represent sets of character states that depict molecular transformations within the structural neighborhoods and lead to optimal phenotypes. Depending on the connectivity of the system, protein architectures can be trapped in local optima at different levels or can escape toward new adaptive peaks by changes for example in the phenotypic dimensionality of the landscape. This could occur when the landscape is altered, for example, by domain “shuffling” (*sensu* Lupas et al. 2001), HGT, architectural loss, fusions and fissions, or organismal diversification. Intuitively, shuffled protein segments expressing little evolutionary lock-in (probably common at the onset of structural diversification) will be less “evolvable” than those achieving higher lock-in levels and modularity (the ability to sustain integrity across varying genetic contexts) (Hansen 2003). Consequently, modularity embodied in domains capable of combining effectively to produce new proteins increases phenotypic dimensionality, reducing the ruggedness of the global landscape. HGT has a similar but less pronounced effect as entire genes are shuffled into different genetic contexts. In contrast, gene loss and organismal diversification (embodied in the emergence of major and minor lineages) decrease phenotypic dimensionality by forfeiting opportunities of innovation or by restricting the exchange of protein modules through recombination or genetic exchange, respectively. Our structural demography studies establish phylogenetic links between patterns describing molecular and organismal diversification that can be used to portray the complexities of the proposed adaptive landscape.

Acknowledgments. We would like to thank Jay Mittenthal (University of Illinois) and Dietz Bauer (Ohio State University) for valuable comments and suggestions.

References

- Ancel LW, Fontana W (2000) Plasticity, evolvability, and modularity in RNA. *J Exp Zool (Mol Dev Evol)* 288:242–283
- Aravind L, Mazumder R, Vasudevan S, Koonin EV (2002) Trends in protein evolution inferred from sequence and structure analysis. *Curr Opin Struct Biol* 12:392–399
- Bauer WD, Mathesius U (2004) Plant responses to bacterial quorum sensing signals. *Curr Opin Plant Biol* 7:429–433
- Caetano-Anollés G (2002) Evolved RNA secondary structure and the rooting of the universal tree of life. *J Mol Evol* 54:333–345
- Caetano-Anollés G, Caetano-Anollés D (2003) An evolutionarily structured universe of protein architecture. *Genome Res* 13:1563–1571
- Carlile M (1982) Prokaryotes and eukaryotes: Strategies and successes. *Trends Biochem* 7:128–130
- Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM, Botstein D (1998) Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* 282:2022–2028
- Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823–826
- Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. *Science* 300:1701–1703
- Copley RR, Bork P (2000) Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J Mol Biol* 303:627–640
- Copley RR, Schultz J, Ponting CP, Bork P (1999) Protein families in multicellular organisms. *Curr Opin Struct Biol* 9:408–415
- Coulson AFW, Moulton J (2002) A unifold, mesofold, and superfold model of protein fold use. *Proteins* 46:61–71
- Daubina V, Moran NA, Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. *Science* 301:829–832
- Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791
- Felsenstein J (2004) *Inferring phylogenies*. Sinauer Associates, Sunderland, MA
- Frishman D, Mewes H-W (1997) Protein structural classes in five complete genomes. *Nature Struct Biol* 4:626–628
- Frishman D, Albermann K, Hani J, Heumann K, Metanowski A, Zollner A, Mewes H-W (2001) Functional and structural genomics using PEDANT. *Bioinformatics* 17:44–57
- Gavrilets S (1997) Evolution and speciation on holey adaptive landscapes. *Trends Ecol Evol* 12:307–312
- Gerstein M (1997) A structural census of genomes: Comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. *J Mol Biol* 274:562–576
- Gerstein M (1998) Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census. *Proteins* 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
- Gerstein M, Levitt M (1997) A structural census of the current population of protein sequences. *Proc Natl Acad Sci USA* 94:11911–11916
- Glandsdorff N (2000) About the last common ancestor, the universal life-tree and lateral gene transfer: A reappraisal. *Mol Microbiol* 38:177–185
- Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 19:2226–2238
- Grant A, Lee D, Orengo C (2004) Progress towards mapping the universe of protein folds. *Genome Biol* 5:107
- Grime JP (1977) Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. *Am Nat* 111:1169–1194
- Hansen TF (2003) Is modularity necessary for evolvability? Remarks on the relationship between pleiotropy and evolvability. *Biosystems* 69:83–94
- Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. *Genome Res* 13:407–412
- Harrison A, Pearl F, Mott R, Thornton J, Orengo C (2002) Quantifying the similarities within fold space. *J Mol Biol* 323:909–926
- Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* 402:C47–C52
- Hegyi H, Lin J, Greenbaum D, Gerstein M (2002) Structural genomics analysis: Characteristics of atypical, common, and horizontally transferred folds. *Proteins* 47:126–141
- Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family size in complete genomes. *Mol Biol Evol* 15:583–589
- Karev GP, Wolf Y, Rzhetsky AY, Berezhovskaya FS, Koonin EV (2002) Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol Biol* 2:18
- Karev GP, Wolf Y, Koonin EV (2003) Simple stochastic birth and death models of genome evolution: Was there enough time for us to evolve? *Bioinformatics* 19:1889–1990
- Kauffman SA (1993) *The origins of order*. Oxford University Press, New York
- Kunin V, Ouzounis CA (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res* 13:1589–1594
- Kunin V, Cases I, Enright AJ, de Lorenzo V, Ouzounis CA (2003) Myriads of protein families, and still counting. *Genome Biol* 4:401
- Lee D, Grant A, Buchan D, Orengo C (2003) A structural perspective on genome evolution. *Curr Opin Struct Biol* 13:359–369
- Limpens E, Franken C, Smit P, Willemsse J, Bisseling T, Geurts R (2003) LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. *Science* 302:630–633
- Lin J, Gerstein M (2000) Whole-genome trees based on the occurrence of fold and orthologs: Implications for comparing genomes on different levels. *Genome Res* 10:808–818
- Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin A (2002) SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Res* 30:264–267
- Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* 134:191–203
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 10:1151–1155
- Lynch M, Conery JS (2003a) The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 3:35–44
- Lynch M, Conery JS (2003b) The origins of genome complexity. *Science* 302:1401–1404
- Maddison WP (1991) Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. *Syst Zool* 40:304–314
- Maddison WP, Maddison DR (1999) *MacClade: Analysis of phylogeny and character evolution, version 3.08*. Sinauer Associates, Sunderland, MA
- Mathesius U, Mulders S, Gao M, Teplitski M, Caetano-Anollés G, Rolfe BG, Bauer WD (2003) Extensive and specific responses of a eukaryote to bacterial quorum sensing signals. *Proc Natl Acad Sci USA* 100:1444–1449

- McFall-Ngai MJ (2001) Identifying 'prime suspects': Symbioses and the evolution of multicellularity. *Comp Biochem Phys B Biochem Mol Biol* 129:711–723
- Mossell E (2003) On the impossibility of reconstructing ancestral data and phylogenies. *J Comp Biol* 10:669–678
- Murzin A, Brenner SE, Hubbard T, Clothia C (1995) SCOP: A structural classification of proteins for the investigation of sequences and structures. *J Mol Biol* 247:536–540
- Nagano N, Orengo CA, Thornton JM (2002) One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* 321:741–765
- Nee S, Holmes EC, May RM, Harvey PH (1994) Extinction rates can be estimated from molecular phylogenies. *Phil Trans R Soc Lond B Biol Sci* 344:77–82
- Orengo CA, Michie AD, Jones S, Jones DJ, Swindells MB, Thornton JM (1997) CATH: a hierarchic classification of protein domain structures. *Structure* 5:1093–1108
- Patthy L (2003) Modular assembly of genes and the evolution of new functions. *Genetica* 118:217–231
- Penny D, Hendy MD, Poole AM (2003) Testing fundamental evolutionary hypotheses. *J Theor Biol* 223:377–385
- Philippe H, Laurent J (1998) How good are deep phylogenetic trees? *Curr Opin Genet Dev* 8:6161–623
- Poole A, Jeffares DC, Penny D (1998) The path from the RNA world. *J Mol Evol* 46:1–17
- Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: Power-law behavior and evolutionary model. *J Mol Biol* 313:673–681
- Rivera MC, Lake JA (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152–155
- Rokas A, Holland PWK (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* 15:454–459
- Rzhetsky A, Gomez SM (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* 17:988–996
- Semple C, Steel M (2002) Tree reconstruction from multi-state characters. *Adv Appl Math* 28:169–184
- Snel B, Bork P, Huynen MA (2002) Genomes in flux: The evolution of Archaeal and Proteobacterial gene content. *Genome Res* 12:17–25
- Sober E, Steel M (2002) Testing the hypothesis of common ancestry. *J Theor Biol* 218:395–408
- Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol* 17:839–850
- Swofford DL (1999) *Phylogenetic analysis using parsimony and other programs (PAUP*)*, version 4. Sinauer Associates, Sunderland, MA
- Swofford DL, Maddison WP (1987) Reconstructing ancestral character states under Wagner parsimony. *Math Biosci* 87:199–229
- Thiele K (1993) The holy grail of the perfect character: The cladistic treatment of morphometric data. *Cladistics* 9:275–304
- Thorley JL, Page RDM (2000) RadCon: phylogenetic tree comparison and consensus. *Bioinformatics* 16:486–487
- White SH (1994) Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure. *Annu Rev Biophys Biomol Struct* 23:407–439
- Wilkinson M, Thorley JL, Upchurch P (2000) A chain is no longer than its weakest link: double decay analysis of phylogenetic hypotheses. *Syst Biol* 49:754–776
- Woese CR (2000) The universal ancestor. *Proc Natl Acad Sci USA* 95:6854–6859
- Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: Proposal 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, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. *Trends Genet* 18:472–479
- Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc Sixth Int Congr Genet* 1:356–366