ONLINE SUPPLEMENTARY MATERIAL:

The evolution of protein architecture and the rooting of the universal tree

Gustavo Caetano-Anollés,[1,2]* and Derek Caetano-Anollés[1]

Corresponding author:
Dr. Gustavo Caetano-Anolles
Department of Crop Sciences
University of Illinois
332 NSRC, 1101 West Peabody Drive
Urbana, IL 61801, USA
Phone and fax: 865-909-0232
e-mail: gcaetano@earthlink.net

ONLINE MATERIAL

Genome trees

The idea that closely related organisms must share significantly more genes than distantly related ones has been used as general phenetic principle in the construction of genome trees (reviewed by Wolf et al. 2002). Most of these trees depict relationships based on overall genome similarity and were generated using distance instead of maximum parsimony or maximum likelihood methods. Consequently, the evolutionary implications drawn from them have been questioned on technical grounds (Doolittle 1999). Protein folds are amongst the most conserved components in life and are therefore ideal for analyzing distant evolutionary relationships. Fold architectures were surveyed in a number of genomes (Gerstein and Levitt 1997; Gerstein 1997, 1998; Frishman and Mewes 1997; Wolf et al. 1999; Snel et al. 1999; Hegyi et al. 2002) and protein fold composition used to reconstruct genome trees (Gerstein 1998, Wolf et al. 1999; Lin and Gerstein 2000). Based on the strong foundation laid by the well-established SCOP classification of protein architecture (Murzin et al. 1995), Gerstein and colleagues produced phylogenetic trees using presence-absence data (binary characters) from eight genomes and both distance and parsimony methods (Gerstein 1998; Lin and Gerstein 2000). These trees resembled rRNA phylogenies and their topologies were relatively well supported. We similarly analyzed folds in protein complements from this same set of 8 genomes, and from an expanded set of 15 genomes (Fig. S1). We first generated most-parsimonious reconstructions from fold usage data that was scored as binary or multi-state characters, and found the resulting trees had congruent topologies and were comparably supported (Fig. S1-a-c). When compared with trees obtained from polarized multistate characters (Fig. S1-c), reconstructions showed that hypotheses of character polarity (unlike those of order; Kitching 1992) had negligible impact on tree topology. We then compared rooted trees depicting the phylogenetic

relationship of the expanded genome set (Fig. S1-d and e). These trees were reconstructed from datasets derived from two releases of SCOP (1.39 and 1.59). The topologies of the two trees were in good agreement, rejecting strongly the existence of taxonomic congruence by chance ($p<0.001$). Interestingly, the tree reconstructed using the latest SCOP release was better-supported (as revealed by tree measures, BS levels, and decay indexes), indicating that the enrichment of structural entries in the fold database improved tree reconstruction substantially. Therefore, we expect database enrichment will continue to increase phylogenetic signal in the future.

Transformation pathways of protein architecture. We studied well-described protein transformations (Grishin 2001) from an evolutionary perspective by establishing an evolutionary direction of the proposed structural change and inferring a relative time frame in which these transformation events took place. For example, the conversion of a α-helix into a three-stranded β-meander causes Rossmann fold proteins to change fold architecture (from c.2 to c.3; Fig. S2-a). Both folds are ancient and are closely related, so this putative transformation must have occurred already very early during evolution. In contrast, circular permutations in protein phosphatases resulted in changes that were quite derived and occurred probably within the last third of the history of protein diversification. Similarly, a strand re-arrangement by C-terminal and N-terminal extension in a widespread RNA-binding domain transforms a putative ancestor into the evolutionary distant d.51 and d.53 architectures, respectively. The gradual effect of indels and substitution in protein evolution was also analyzed in two putative pathways involving Rossmann-like folds (Fig. S2-b) or extensive changes that transform an all-α 3-helical bundle into an all-β barrel-like architecture (Fig. S2-c). The first pathway was quite ancestral and confined to the first fifth of the history of protein diversification, while the second extended throughout the tree of protein architectures.

REFERENCES

Doolittle, W.F. in response to Huynen, M., Snel, B., and Bork, P. 1999. Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science* **286:** 1443a.

Frishman, D. and Mewes, H-W. 1997. Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4:** 626-628.

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. and Levitt, M. 1997. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. USA* **94:** 11911-11916.

Lin, J. and 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.

Korbel, J.O., Snel, B., Huynen, M.A., and Bork, P. 2002. SHOT: a web server for the construction of genome phylogenies. *Trends. Genet.* **18:** 158-162.

LEGENDS TO FIGURES

Figure S1 Phylogenetic reconstruction of genome trees based on the occurrence of fold architectures in protein complements. *(A-C)* We analyzed 420 fold categories from SCOP 1.39 using maximum parsimony as the optimality criterion in PAUP*. Phylogenetically uninformative characters representing invariant features and autapomorphies were excluded before tree reconstruction. A total of 8 genomes and 227-300 phylogenetically informative characters were examined with three coding schemes. In *A*, the presence or absence of individual folds was entered as binary characters states, 1 or 0 respectively, following the approach of Lin and Gerstein (2001). These undirected characters were arranged in data matrices and used for cladistic analysis. Two most-parsimonious unrooted

trees of 352 steps (CI=0.645, RI=0.555; g_1=-1.324; PTP, p=0.001) were retained after an exhaustive search. The tree shown is congruent with the 50% majority-rule consensus. In *B*, the actual number of protein sequences in each fold category was gap-recoded using a 0-20 scale and entered as ordered multi-state characters. A single most-parsimonious unrooted tree of 2044 steps (CI=0.668, RI=0.535; g_1=-0.935; PTP, p=0.001) was retained after an exhaustive search. In *C*, fold occurrence values were entered as multi-state characters, but these were polarized by assuming that the incidence of individual fold architectures increases in the course of evolution. A single most-parsimonious tree of 6957 steps (CI=0.835, RI=0.495; g_1=-0.739; PTP, p=0.001) was retained after an exhaustive search and it was automatically rooted at the point where the hypothetical ancestor connected to the tree. The reduced tree shows branches with less than 50% BS collapsed into polytomies and BS proportions. *(D-F)* We reconstructed genome trees depicting the phylogenetic relationship of a set of 15 genomes using maximum parsimony principles and datasets derived from two versions of the SCOP database. Rooted trees were derived as in *C* and were obtained by heuristic searches with TBR branch swapping and 50 replicates of random addition sequence. In *D*, we recovered a single tree of 8807 steps (CI=0.714, RI=0.455; g1=-0.616; PTP test, p=0.001) using fold categories in SCOP 1.39. A total of 320 characters were phylogenetically informative. In *E*, we recovered a single tree of 11435 steps (CI=0.764, RI=0.473; g1=-0.859; PTP test, p=0.001) using fold categories in SCOP 1.59. A total of 440 characters were phylogenetically informative. In *F*, a reference distance tree based on gene content was retrieved from the SHOT server (Korbel et al. 2002). Gene content was normalized by the weighted average of genome size, the evolutionary distance computed as $-\ln s$, with s being the normalized fraction of shared orthologous genes, and trees obtained using the neighbor-joining algorithm. Genomes and abbreviations used: *Aquifex aeolicus* (*Aae*), *Archaeoglobus fulgidus* (*Afu*), *Caenorhabditis elegans* (*Cel*), *Chlamydia pneumoniae* (*Cpn*), *Escherichia*

coli (*Eco*), *Haemophilus influenzae* (*Hin*), *Helicobacter pylori* (*Hpy*), *Methanobacterium thermoautotrophicum* (*Mth*), *Methanococcus jannaschii* (*Mja*), *Mycobacterium tuberculosis* (*Mtu*), *Mycoplasma genitalium* (*Mge*), *Mycoplasma pneumoniae* (*Mpn*), *Pyrococcus horikoshii* (*Pho*), *Rickettsia prowazekii* (*Rpo*), *Saccharomyces cerevisiae* (*Sce*), *Synechocystis* sp. (*Ssp*), *Treponema pallidum* (*Tpe*). BS values >50% are given above nodes of all trees.

Figure S2 Tracing the evolution of architectural change. Phylogenies were reconstructed from folds believed to have followed structural transformations during evolution. Trees of lengths ranging 536-680 steps [CI=0.938-0.959, RI=0.633-0.927; g_1=-(0.190-1.120); PTP tests, p=0.001] were retained after exhaustive searches. Branches were traced in the general tree of architectures (*see* Fig. 2) and the distance in nodes from the root given encircled at the leaves of the reconstructed trees. *(A)* The effect of insertion/deletions (indels) in lactate dehydrogenase (1ldn) and NADH peroxidase (1npx), of circular permutation in VH1-related phosphatase (1vhr) and phosphotyrosine protein phosphatase (1phr), and of strand invasion by terminal extension in the K homology (KH) domain of ribosomal protein S3 (1fjf) and hnRNP K (1khm). *(B)* Changes induced by indels and substitutions in proteins with Rossmann-like fold architecture. The c.47 fold in Synapsin (1auv) is transformed into the c.30 fold in glutathione synthetase (1gsa) by an extension of a β-hairpin, or alternatively into the c.56 fold of carboxypeptidase A (2ctc) by a meander-helix substitution. *(C)* Pathway from an all-β to an all-α architecture. The replacement of a winged helix-turn-helix (HTH) domain characteristic of nucleic acid-binding domains such as the C-terminal domain of the catabolite gene activator (CAP) protein (1cgp) by successive deletions of β-strands results in the N-terminal domain of the β-subunit of glycogen phosphorylase kinase (1phk), the sonic hedgehog N-terminal signal domain (1vhh), and finally the C-terminal domain of G4-amylase (2amg).

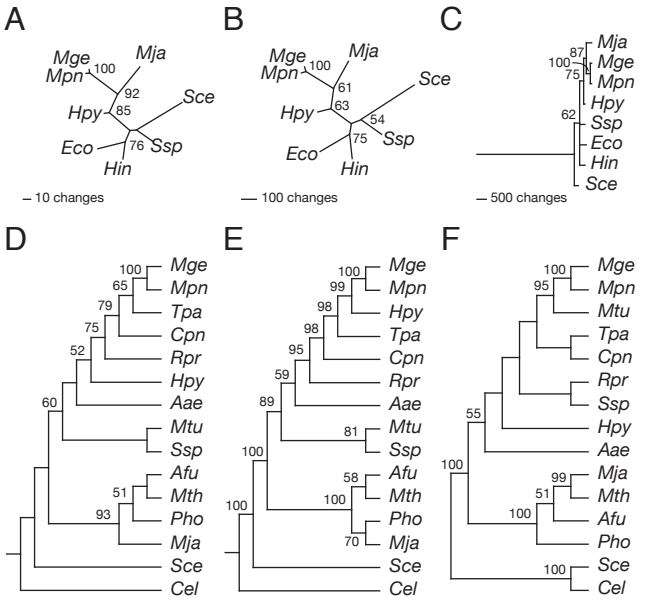
A
Mge
Mpn
100
Mja
92
85
Sce
Hpy
Eco
76
Ssp
Hin
10 changes
B
Mge
Mpn
100
Mja
61
63
Sce
Hpy
54
Ssp
75
Eco
Hin
100 changes
C
Mja
100
87
75
Mge
Mpn
Hpy
62
Ssp
Eco
Hin
Sce
500 changes
D
100
65
79
75
52
60
51
93
Mge
Mpn
Tpa
Cpn
Rpr
Hpy
Aae
Mtu
Ssp
Afu
Mth
Pho
Mja
Sce
Cel
E
100
99
98
98
95
59
89
100
81
100
58
100
100
70
Mge
Mpn
Hpy
Tpa
Cpn
Rpr
Aae
Mtu
Ssp
Afu
Mth
Pho
Mja
Sce
Cel
F
100
95
55
100
99
51
100
100
Mge
Mpn
Mtu
Tpa
Cpn
Rpr
Ssp
Hpy
Aae
Mja
Mth
Afu
Pho
Sce
Cel

Figure S1

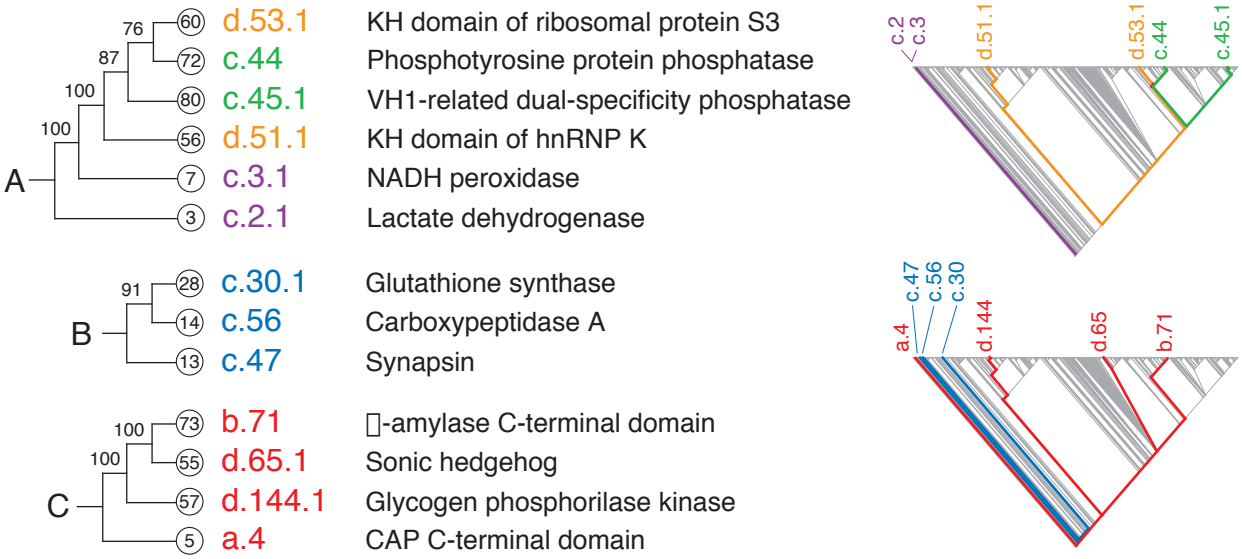
A
76
87
100
100
60
72
80
56
7
3
d.53.1
c.44
c.45.1
d.51.1
c.3.1
c.2.1
KH domain of ribosomal protein S3
Phosphotyrosine protein phosphatase
VH1-related dual-specificity phosphatase
KH domain of hnRNP K
NADH peroxidase
Lactate dehydrogenase
B
91
28
14
13
c.30.1
c.56
c.47
Glutathione synthase
Carboxypeptidase A
Synapsin
C
100
100
73
55
57
5
b.71
d.65.1
d.144.1
a.4
-amylase C-terminal domain
Sonic hedgehog
Glycogen phosphorilase kinase
CAP C-terminal domain
c.2
c.3
d.51.1
d.53.1
c.44
c.45.1
a.4
c.47
c.56
c.30
d.144
d.65
b.71

Figure S2