Prokaryotic phylogenies inferred from protein structural domains

Eric J. Deeds,[1] Hooman Hennessey,[2] and Eugene I. Shakhnovich[3,4]

[1]*Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA;* [2]*McGill University, Montreal, Quebec H3A 2K6, Canada;* [3]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA*

The determination of the phylogenetic relationships among microorganisms has long relied primarily on gene sequence information. Given that prokaryotic organisms often lack morphological characteristics amenable to phylogenetic analysis, prokaryotic phylogenies, in particular, are often based on sequence data. In this work, we explore a new source of phylogenetic information, the distribution of protein structural domains within fully sequenced prokaryotic genomes. The evolution of the structural domains we use has been studied extensively, allowing us to base our phylogenetic methods on testable theoretical models of structural evolution. We find that the methods that produce reasonable phylogenetic relationships are indeed the methods that are most consistent with theoretical evolutionary models. This work represents, to our knowledge, the first such theoretically motivated phylogeny, as well as the first application of structural information to phylogeny on this scale. Our results have strong implications for the phylogenetic relationships among prokaryotic organisms and for the understanding of protein evolution as a whole.

[Supplemental material is available online at www.genome.org and http://paradox.harvard.edu/~eric/struct_phylo.htm.]

Our understanding of the evolution of protein structures has advanced considerably in the past several years (Dokholyan et al. 2002; Koonin et al. 2002; Deeds et al. 2004). This advance has relied, at least in part, on the application of graph theoretic methods to the representation and analysis of structural similarity between protein domains (Dokholyan et al. 2002; Koonin et al. 2002; Deeds et al. 2003, 2004). One such application is the Protein Domain Universe Graph (or PDUG), a graph in which a nonredundant set of all known protein structural domains (Holm and Sander 1996; Dietmann and Holm 2001) are represented as nodes, and the structural similarity between domains is used to define edges between them (Dokholyan et al. 2002). The distribution of edges per node in this graph, (known as the degree distribution or $p(k)$), was shown to follow a power law, i.e., $p(k) \sim k^{-\gamma}$ (Dokholyan et al. 2002). This degree distribution is markedly different from that of random graphs (Albert and Barabasi 2002) or structural similarity graphs based on complete sets of model polymer structures (Deeds et al. 2003). Graphs with degree distributions similar to that of the PDUG have been produced via evolutionary models that are divergent in nature (Dokholyan et al. 2002; Deeds et al. 2003, 2004), and these findings have provided further support for a divergent picture of protein structural evolution in the debate between divergent and convergent scenarios of protein structural evolution (Dokholyan et al. 2002; Koonin et al. 2002; Deeds et al. 2003, 2004).

The structural domains used to create the PDUG correspond to families of similar sequences that adopt highly similar structures (Dokholyan et al. 2002). Domains from the PDUG may thus be assigned to the proteomes of an organism based on the presence or absence of a protein sequence belonging to that domain's family within the genome of that organism (Deeds et al. 2004). The "structural proteomes" of fully sequenced prokaryotes were recently determined in this manner. These proteomes may be understood as subgraphs of the PDUG (see Fig. 1), and analysis of the results demonstrated that the organismal subgraphs of the PDUG were also scale free with power-law exponents similar to that of the PDUG (i.e., γ~1.6) (Deeds et al. 2004). It was also demonstrated that, for most of the organisms analyzed, the organismal subgraphs had a very low probability of being random subgraphs of the PDUG (with around 67% of the proteomes having a probability of being random of ~10^{-6} or less, see Table 1). Addition of a speciation mechanism to divergent models of structural evolution results in model proteomes that are highly nonrandom subgraphs of their respective model PDUGs (Dokholyan et al. 2002; Deeds et al. 2004). In this model, inheritance of a particular domain (i.e., a particular sequence-structure pair) occurs only through descent, and the fact that nonrandom model subgraphs are produced by this model implies that the strict partitioning of diverging structural characters into specific genomes explains the nonrandom quality of actual structural proteomes (Deeds et al. 2004). These findings, when taken together with the fact that a significant amount of domain overlap is found between the structural proteomes of even widely diverged prokaryotes (Deeds et al. 2004), indicate that structural proteomes may contain phylogenetically informative signals. Indeed, if the model discussed above represents the source of nonrandom behavior in structural proteomes, the extent of structural domains shared between organisms should represent the extent of shared descent between the two organisms, and thus lead to a very reasonable set of phylogenetic relationships under the correct set of phylogenetic assumptions.

Molecular sequence information is currently the most prevalent source of data for phylogenetic analysis (Lin and Gerstein 2000; Brown et al. 2001; Wolf et al. 2001, 2002; Giribet 2002; Korbel et al. 2002), especially in Prokaryotes where phylogenetically informative morphological characters are largely ab-

[4]**Corresponding author.**
E-mail: eugene@vodka.chem.harvard.edu; fax (617) 384-9228.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3033805.

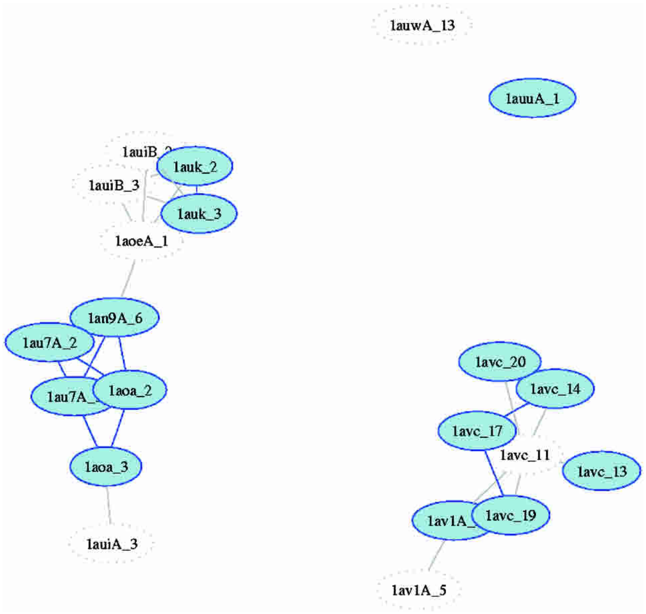

Figure 1. Organismal subgraphs. The set of domains in a particular structural proteome correspond to a subgraph of the PDUG. In this case, the domains and edges that exist in *Bacillus subtilis* are colored in blue, and those that do not are colored gray.

sent. Although sequence information has proved quite useful, the structural domains used in this work offer several potential advantages over sequence-based characters. For one, new structural domains are likely to evolve much more slowly than new sequences (in the case of sequence comparisons between widely conserved orthologs such as the 16S ribosomal RNA) and more slowly than new genes (given that new genes may be discovered through novel permutations of existing structural domains). The longer time scales characterizing domain evolution thus hold great promise for illuminating the “deeper” branches of prokaryotic phylogeny.

The second major advantage of structural domains has to do with the current theoretical understanding of structural evolution. There are many existing methods that allow for the inference of phylogenies from data such as the structural domains used in this work (Lin and Gerstein 2000; Brown et al. 2001; Wolf et al. 2001, 2002; Korbel et al. 2002; Mirkin et al. 2003). Each method rests on a distinct set of assumptions about the evolution of the characters used to infer the phylogeny. In this case, it is possible to derive these assumptions, and thus choose a particular method, on the basis of theoretical models for the evolution of protein structures. This allows us to base phylogenies on models that have been independently tested against statistical features of the PDUG and structural proteomes (Deeds et al. 2004). Given that there are no analogous models for sequence-based characters, the theoretical grounding of this analysis is currently limited to structural domains.

The third advantage of structural characters is based on the hypothesis that Lateral Gene Transfer (LGT) has had a lesser influence on structural domain distributions than on orthologous gene sets. LGT is a widespread phenomenon in prokaryotic evolution (Aravind et al. 1998; Doolittle 1999; Ochman et al. 2000; Gogarten et al. 2002), and the disruption of phylogenetic signals via LGT has lead to the claim that reliable phylogenies cannot be constructed for prokaryotes (Doolittle 1999). It is not clear, however, the extent to which LGT has influenced distributions of sequence-structure pairs. Only LGT events that involve the transfer of a novel sequence-structure pair (i.e., structural domain innovation) will influence our data set, indicating that only lateral structural domain transfer (LSDT) events would interfere with

Table 1. P-values for structural proteomes

Organism Name	Probability
Agrobacterium tumefaciens C58	6.74E-12
Agrobacterium tumefaciens C58 UWash	9.51E-12
Aquifex aeolicus	5.21E-04
Archaeoglobus fulgidus	3.81E-05
Bacillus halodurans	1.07E-11
Bacillus subtilis	2.78E-14
Brucella melitensis	4.13E-06
Campylobacter jejuni	3.33E-03
Caulobacter crescentus	1.06E-10
Clostridium acetobutylicum	3.18E-12
Clostridium perfringens	1.55E-04
Corynebacterium glutamicum	1.74E-11
Deinococcus radiodurans	4.14E-10
Escherichia coli K12	4.59E-11
Escherichia coli O157H7 EDL933	8.24E-10
Escherichia coli O157H7	9.96E-10
Fusobacterium nucleatum	9.23E-04
Haemophilus influenzae	7.46E-05
Halobacterium sp.	7.41E-06
Helicobacter pylori 26695	5.55E-03
Helicobacter pylori J99	2.56E-03
Lactococcus lactis	3.44E-13
Listeria innocua	1.23E-13
Listeria monocytogenes	8.37E-12
Mesorhizobium loti	7.80E-11
Methanosarcina acetivorans	1.63E-05
Methanosarcina mazei	6.36E-05
Mycobacterium leprae	3.88E-04
Mycobacterium tuberculosis CDC1551	1.75E-07
Mycobacterium tuberculosis H37Rv	4.62E-08
Neisseria meningitidis MC58	7.49E-04
Neisseria meningitidis Z2491	9.79E-04
Nostoc sp.	2.10E-13
Pasteurella multocida	5.27E-05
Pseudomonas aeruginosa	7.41E-06
Pyrococcus furiousus	5.22E-08
Ralstonia solanacearum	1.14E-08
Salmonella typhimurium LT2	4.19E-10
Salmonella typhi	1.01E-08
Sinorhizobium meliloti	1.24E-10
Staphylococcus aureus Mu50	8.67E-07
Staphylococcus aureus MW2	8.18E-07
Staphylococcus aureus N315	9.73E-07
Streptococcus pneumoniae R6	1.09E-09
Streptococcus pneumoniae TIGR4	1.43E-08
Streptococcus pyogenes MGAS8232	4.27E-05
Streptococcus pyogenes	2.13E-05
Streptomyces coelicolor	5.80E-17
Sulfolobus tokodaii	1.34E-05
Synechocystis PCC6803	1.10E-04
Thermoanaerobacter tengcongensis	7.11E-12
Thermoplasma acidophilum	1.86E-11
Thermoplasma volcanium	2.66E-12
Thermotoga maritima	2.16E-13
Vibrio cholerae	7.24E-08
Xanthomonas campestris	6.68E-16
Xanthomonas citri	1.55E-15
Xylella fastidiosa	4.51E-20
Yersinia pestis	4.79E-06

This table contains the probability that each structural proteome used in this work is a random subgraph of the PDUG. The probabilities are calculated according to the method of Deeds et al. (2004) as described in the Methods section.

the phylogenetic signal of protein domains. The transfer of structural domains from one lineage to another, in a sense, mimics the convergent discovery of domains and might tend to increase the probability that a structural proteome is a random subgraph of the PDUG. The fact that most structural proteomes have a low probability of being random subsets of the PDUG (see Table 1) implies that the balance of LGT events may not have represented LSDT to the acceptor lineage, and thus, might have exerted less influence on distributions of domains than on sequence or gene-content data.

The structural domains used in this work also have advantages when compared with structural information that has been used to infer phylogenies in previous studies (Wolf et al. 1999; Lin and Gerstein 2000). These studies were based on the "fold" level of structural classification, which correspond to clusters of structural domains on the PDUG (Wolf et al. 1999; Lin and Gerstein 2000; Dokholyan et al. 2002). Given that folds do not correspond to specific sequence structure-pairs, folds represent a more "coarse-grained" source of information than the structural domains on the PDUG. This implies that such structural clusters might not be as phylogenetically informative as structural domains, given that two organisms can share very few domains, even though each contains at least one representative of a large number of clusters. In accordance with these observations, folds as structural characters have not proved terribly successful as a basis for inferring phylogenies (Wolf et al. 1999; Lin and Gerstein 2000), although these efforts were also potentially impeded by the lack of available fully sequenced taxa at that time. Current theoretical models thus indicate that specific structural domains may overcome some of the difficulties inherent to a variety of both sequence-based and structural character sets, and thus make a significant contribution to our understanding of the phylogenetic relationships among prokaryotic organisms.

In the following sections, we show that when mechanisms of LSDT are added to models of structural evolution, the resulting model proteomes have higher probabilities of being random. We explore three separate methods of phylogenetic inference on the basis of structural domains, each with varying levels of consistency with current models of structural evolution. Analysis of the results indicates that those methods that are more consistent with current evolutionary models result in more biologically reasonable sets of evolutionary relationships.

Results

Organismal subgraphs

As described previously (Deeds et al. 2004), structural domains from the PDUG may be assigned to specific organisms on the basis of sequence comparisons. The resulting subgraphs (see Fig. 1) exhibit degree distributions similar to that of the PDUG (Deeds et al. 2004). Subsets of nodes from the PDUG chosen completely randomly, however, also have PDUG-like degree distributions. Comparison of an organismal subgraph to a random subgraph of the same size indicates that the two graphs do differ significantly with respect to the number of connections made by the most highly connected node on the graph (called the *Maxk* of the graph) (Deeds et al. 2004). One can quantify the probability that particular structural proteome is a random subgraph of the PDUG using an analytical approximation to the degree distributions of random subgraphs (Methods) and calculating the probability that a node with connectivity *Maxk* will be observed in such a random subgraph. The resulting probabilities are shown in Table 1. It is clear from this data that most structural proteomes are highly nonrandom subgraphs of the PDUG.

Lateral gene transfer and models of structural evolution

To test the impact of lateral transfer of structural domains on the probability that structural proteomes are random subgraphs of the PDUG, we create several modified versions of a previous speciation model (Deeds et al. 2004). In these models, LGT events occur with some frequency compared with the discovery of new domains, and an LGT event is modeled as the transfer of one domain from one organism to another. Given that the only LGT events that will influence structural proteomes involve the transfer of a domain that does not exist in the acceptor population, all of the LGT models we consider involve domains chosen from the set of domains that exist in the donor organism, but not in the acceptor, and thus represent LSDT. The first model we consider involves transfer between any two existing organisms chosen at random. In this model, the acceptor organism concomitantly loses one of its domains randomly so that the LSDT event does not result in a net increase in the number of domains constituting the model proteome of a given organism. The preservation of structural proteome size is, in this case, included only because of the dependence of P-value on proteome size (see below). A more detailed description of this model may be found in the Methods. The model is diagrammed in Figure 2.

We run this model with graph evolution parameters identical to that used in previous work (Deeds et al. 2004), with speciation events occurring every 500 domain discovery events from event 1 to event 2000. One run of the algorithm results in a total

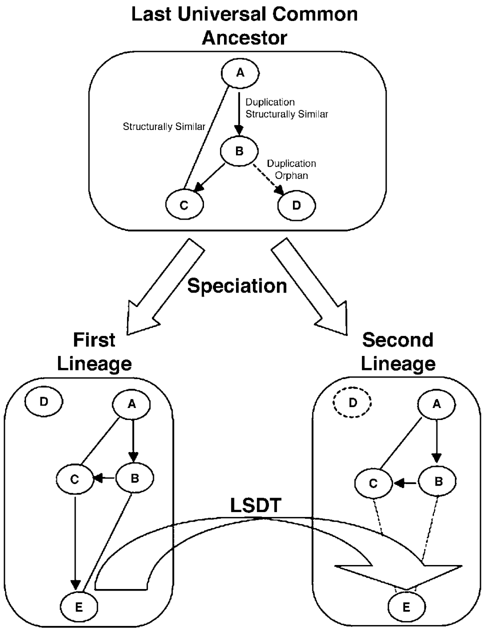

Figure 2. A schematic of the current model of structural evolution. The evolution of the structural characters occurs according to divergent rules (Dokholyan et al. 2002; Deeds et al. 2003, 2004), and the domains discovered as a result of this process are specific to a given lineage. Both LGT events and convergent discovery of domains introduce greater degrees of randomness into model structural proteomes (see Fig. 3). Domain D in the acceptor lineage is represented as a dashed line to indicate that it is concomitantly deleted in some LGT models but not in others.

of 3500 domains and 16 organisms. We run the model at two different LSDT frequencies, 1 LSDT event per every 1000 domain discovery events and 1 LSDT event per every domain discovery event. Four independent runs are conducted in each case, and the results are shown in Figure 3. The P-values of the model organisms are calculated according to the methodology introduced in Deeds et al. (2004) and represent the probability that a given model proteome is a random subset of its corresponding model PDUG. The level of nonrandomness for the lower frequency is nearly indistinguishable from that of the unmodified algorithm (Deeds et al. 2004). At higher frequencies, however, a much greater degree of randomness is observed in almost all of the model proteomes produced by the algorithm (see Fig. 3A). One LSDT event per domain discovery event represents a relatively moderate amount of LSDT (given between two and 16 organisms exist over the course of the simulation), indicating that LSDT frequencies need not be excessive to introduce significant randomness.

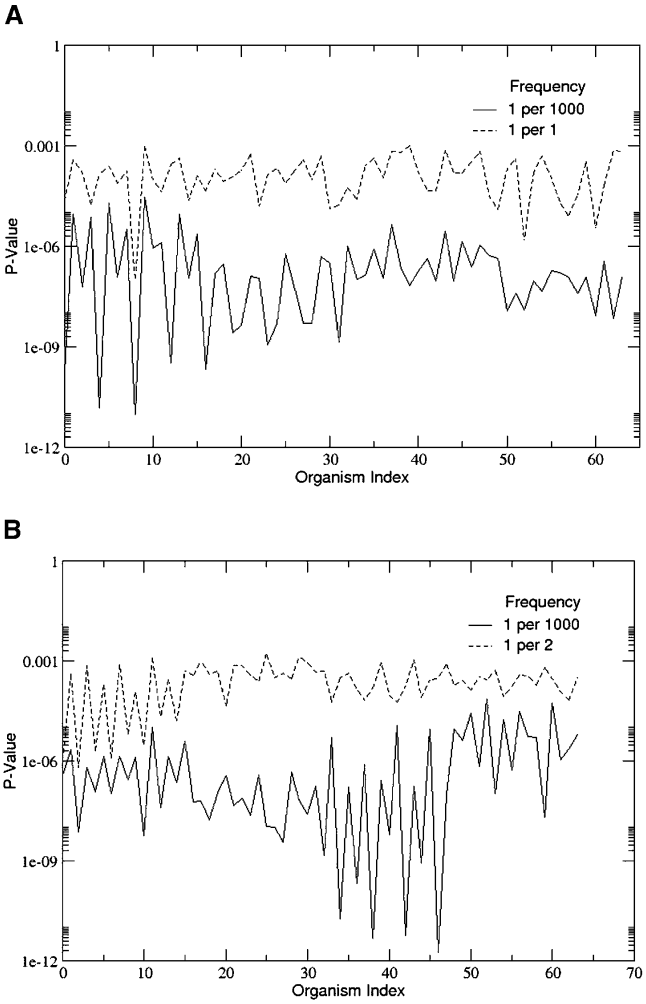

Figure 3. LGT and the randomness of structural proteomes. (*A*) The P-values of model proteomes created in eight independent runs with two different LSDT frequencies are displayed above. The legends "1 per 1000" and "1 per 1" indicate LSDT frequencies of 1 per every 1000 domain discovery event and 1 per every domain discovery event, respectively. In most cases, the P-values of organisms evolved at higher LSDT frequencies are significantly higher than those evolved at lower frequencies. (*B*) A plot similar to that in *A*, but involving an LSDT model in which LSDT may only occur between closely related organisms. As with the model in *A*, LGT tends to increase the probability that a model organismal subgraph will be a random subgraph of the model PDUG.

This result is admittedly specific to this particular implementation of LGT in the model, and in particular, the assumption of concomitant domain loss in the acceptor population may not be biologically realistic. Models in which domain loss never occurs show very similar behavior to the above model (data not shown), indicating that this feature of the model is not necessary to observe an increase in P-values via LSDT. Also, the assumption that LSDT can occur between any two organisms, regardless of phylogenetic distance, may also be biologically unreasonable. To test this feature of the model, we create a modified version in which LSDT can only occur between organisms that resulted from the last speciation event. In this model, LSDT may occur between any pair of organisms that split from one another in the most recent round of speciation. In this case, we also observe an increase in the probability that organismal subgraphs are random (see Fig. 3B), indicating that great phylogenetic distance between donor and acceptor organisms is not fundamental to LSDT-based randomness in structural proteomes.

Although the above results are encouraging, it is clear that we cannot possibly hope to implement all types of LSDT scenarios in our models. Despite this limitation, the above observations indicate that LSDT represents a very likely mechanism through which random-subgraph character may be introduced into the structural proteomes of organisms. Given that the balance of actual structural proteomes are quite nonrandom, the above model frames our hypothesis that LSDT may not have exerted a strong influence on the evolution of most structural proteomes.

Dollo parsimony

As mentioned above, there are many existing methods that allow for the inference of phylogenies from data such as the structural domains used in this work (Lin and Gerstein 2000; Brown et al. 2001; Wolf et al. 2001, 2002; Mirkin et al. 2003). Each method rests on a distinct set of assumptions about the evolution of the characters used to infer the phylogeny. Phylogenies that are based on maximum parsimony (MP) methods (Lin and Gerstein 2000; Brown et al. 2001; Wolf et al. 2001, 2002; Mirkin et al. 2003) assume that the "best" tree is the one that requires the fewest possible evolutionary events to describe the patterns of characters observed in extant organisms. Moreover, one may use various character-type assumptions when building a tree on the basis of MP. The entirely divergent nature of domain evolution in theoretical models indicates that the Dollo criterion (Farris 1977; Swofford 2003) for character evolution is the most appropriate choice for structural domains. Under the Dollo criterion, characters (such as these structural domains) are effectively constrained to be monophyletic; that is, they arise only once on the tree. This is achieved by imposing the restraint that "reversions" cannot occur; that is, once a domain is lost from a given lineage, it cannot be regained. The Dollo assumption thus effectively prevents both convergent and LSDT events from "occurring" on the tree, and in the limit where the current model of structural evolution is exact, the true maximum parsimony tree based on Dollo parsimony represents an exact solution to the phylogenetic problem. Thus, although this form of parsimony may be completely unreasonable for other types of characters (such as certain sets of orthologous genes or detailed sequence changes (Brown et al.

2001; Mirkin et al. 2003), it is strongly indicated in the case of structural domains.

We construct a phylogeny on the basis of the structural characters and phylogenetic assumptions outlined above using version 4 (β 10) of the software program PAUP* (Swofford 2003). We use 59 prokaryotic taxa in this analysis (see Table 1 for a list of taxa), corresponding to those structural proteomes in the fully sequenced set that contain more than 550 domains (see Methods). The eukaryotic taxon *Saccharomyces cerivisiae* is included and is used as the outgroup to root the tree. A full description of the characters and taxa may be found in the Methods, and the entire data matrix used in this work is available from our Web site (http://paradox.harvard.edu/~eric/struct_phylo.htm). Bootstrap analysis is used to determine the statistical support for internal nodes and is based on 500 replicates. The results are shown in Figure 4, with bootstrap proportions as percentages labeling the relevant nodes (those cases where bootstrap support is <50% are presented as polytomies). Bootstrap support is relatively strong for most nodes (>80% for ~68% of the internal nodes), with notable exceptions in the archaea, proteobacteria, and some of the deeper branches of the gram-positive clade. In the case of both the proteobacteria and the gram-positive bacteria, bootstrap support was less than 50% for one or two internal nodes. The overall topology of the tree is similar to that of the 16S small-subunit rRNA tree (or SSU rRNA tree) (see Fig. 5 for a simple SSU rRNA tree of these taxa), and to that of some phylogenies based on other gene-content or whole-genome approaches (Brown et al. 2001; Wolf et al. 2001, 2002). As observed in some gene-content methodologies (Wolf et al. 2002), this tree exhibits some mixing among the α, β, and γ proteobacteria, as well as mixing between the Crenarchaeota and Euryarchaeota. Low bootstrap proportions in these regions of our trees indicate that structural information may not prove useful for determining the branching order of proteobacteria or the archaea. The root is placed similarly to the rRNA tree, providing some indication that eukaryotes and archaea are more closely related to each other than to the bacteria. Also, many of the well-recognized "deep" groupings, such as high G+C gram positives, low G+C gram positives, and proteobacteria, exhibit the "canonical" member taxa in this tree, although there are some notable exceptions to this observation.

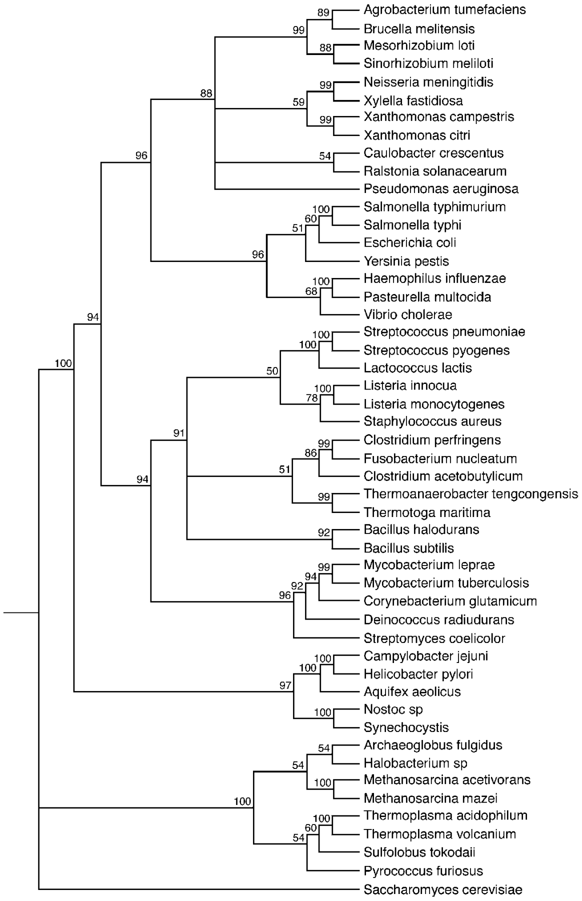

Figure 4. Bootstrapped phylogeny inferred from structural domains using the Dollo criterion. This consensus phylogeny was created from 500 bootstrap replicates. The nodes are labeled with bootstrap proportion as a percentage of replicates. In those cases where there exists more than one sequenced strain of a given bacterium in the data matrix, those strains are always most closely related (with near 100% bootstrap support), and so the individual strains are represented here as single taxa.

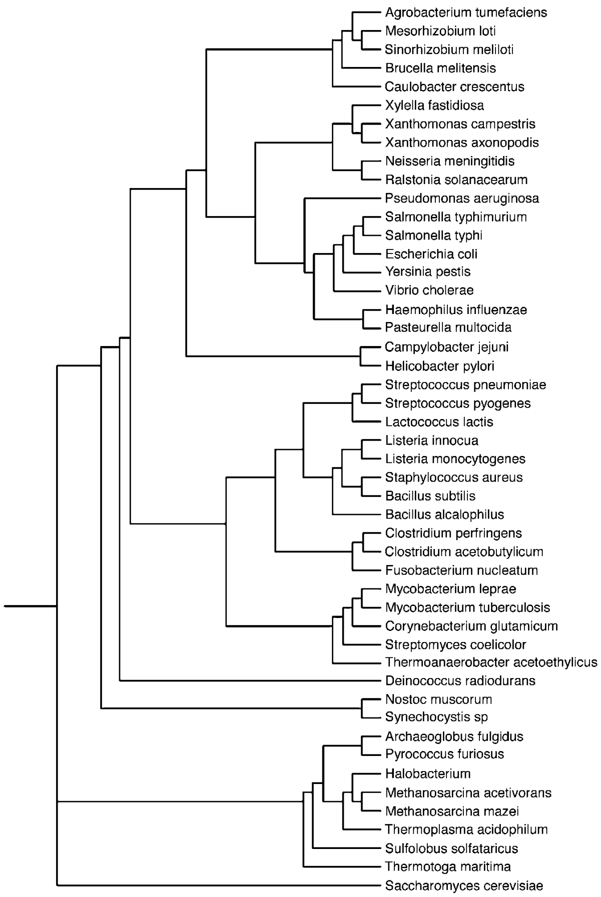

Figure 5. 16S small-subunit rRNA tree. This tree is a phylogeny created on the basis of the SSU rRNA sequences for a set of taxa similar to that used to create the structural phylogeny. The details of the calculation of this tree are given in the text of the Supplemental material.

One such exception is the placement of the pathogenic gut bacterium *Helicobacter pylori* and its relative *Campylobacter jejuni*. These bacteria are considered ϵ-proteobacteria and are often (but not always) placed as a deeply branched group within the proteobacteria (Brown et al. 2001; Wolf et al. 2001, 2002; Korbel et al. 2002) (see Fig. 5). In the Dollo tree, these bacteria are placed (with strong bootstrap support) together with *Aquifex aeolicus* and the cyanobacteria as the most deeply branched bacteria. This difference may be explained by the fact that the structural proteomes of *C. jejuni* and *H. pylori* have greatest probability of being random subgraphs of the PDUG in this entire set of taxa—the probability that the structural proteomes of these organisms represent a random subset of the PDUG is around 0.3%–0.6% (Deeds et al. 2004). As discussed above, this may indicate significant LSDT during the evolutionary history of these organisms, which would render them difficult to place on the basis of Dollo parsimony (or, indeed, potentially difficult with any method). The strong bootstrap support for this grouping indicates that these characters provide an unambiguous placement of these organisms under the Dollo constraint; the potential for error in this case arises from the fact that Dollo most likely does not represent the correct assumption for the structural characters found in these two organisms. This grouping may also indicate that some of the LSDT that has influenced the structural proteomes of *C. jejuni* and *H. pylori* has occurred between these organisms and the organisms with which they cluster on this tree, although the great degree of randomness in both proteomes may complicate understanding of structural evolution in this particular case. Although the randomness of *H. pylori* and *C. jejuni* is technically only a feature of their structural proteomes, this observation provides a strong clue as to why these organisms are often difficult to place for other phylogenetic methods (Wolf et al. 2001, 2002; House and Fitz-Gibbon 2002; Korbel et al. 2002).

Distance-based methods

Although Dollo parsimony is strongly implied for structural domains by theoretical models, it is not the only method that may be considered, consistent with our understanding of structural evolution (Deeds et al. 2004). Indeed, the nonrandom overlap that we observe between structural proteomes, both in real proteomes and in simulation (Deeds et al. 2004), indicates that distance-based methods may also produce reasonable phylogenies, even though such methods do not explicitly disallow convergent or LSDT events. Indeed, in the limit in which current structural evolution models are exact, these two methods should yield roughly equivalent results. To test this hypothesis, we create a distance matrix on the basis of the total number of character differences between each pair of taxa (i.e., the "Hamming distance" between each organism) (see Lin and Gerstein 2000) and create a phylogeny from this matrix using the neighbor-joining (NJ) algorithm (Saitou and Nei 1987; Swofford 2003). As with the Dollo tree, this phylogeny is bootstrapped with 500 replicates, and the results are shown in Figure 6.

This tree largely agrees with the results from Dollo parsimony (see Fig. 4), with most differences occurring as either statistical discrepancies (internal nodes that are supported by Dollo, but not by the distance method, and vice versa) or as relatively small differences in branching order within the major clades. One interesting difference is that support for the monophyly of the gram positives in this case is somewhat weak, with a bootstrap proportion of only 60% compared with the 94% support in the Dollo tree. Aside from this reduced support for the monophyly of the gram-postitive bacteria, the two trees support roughly the same phylogenetic conclusions and are of similar quality.

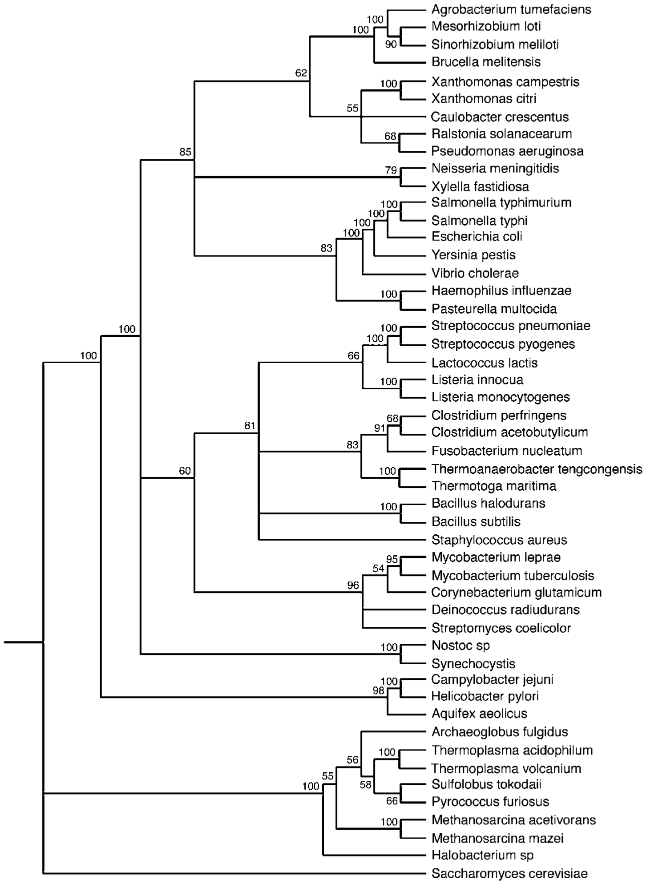

Figure 6. Neighbor-joining structural tree. This phylogeny is created using the neighbor-joining algorithm with the distances between organisms calculated as the total number of structural character differences between them. Bootstrap proportions are obtained from 500 replicates and are labeled as percentages near each internal node.

Unconstrained parsimony

One may also calculate maximum parsimony trees under the assumption that reversions may occur, i.e., that certain characters may be discovered independently multiple times on the tree (either through true convergence or through a mechanism such as LSDT). This method will tend to use independent discovery events whenever doing so reduces the total number of evolutionary events on the tree, and thus, this form of parsimony is much less consistent with theoretical models for structural evolution than the Dollo method.

In order to test such unconstrained methods, we create a bootstrapped phylogeny from the structural domains on the basis of this type of parsimony (see Fig. 7). On the whole, the resultant tree is significantly less reasonable than the tree obtained based on the Dollo assumption. For instance, the gram-positive bacteria are not monophyletic with respect to the proteobacteria in this tree; the high G+C gram-positive bacteria are clustered with the proteobacteria before this clade joins the other gram positives. Although the monophyly of the gram-positive bacteria is a matter of some debate, those methods that do not indicate monophyletic gram positives often associate the low G+C gram positives with the proteobacteria rather than the high G+C gram positives (Brown et al. 2001; Korbel et al. 2002). Ultimately, we consider this particular grouping as relatively unreasonable and

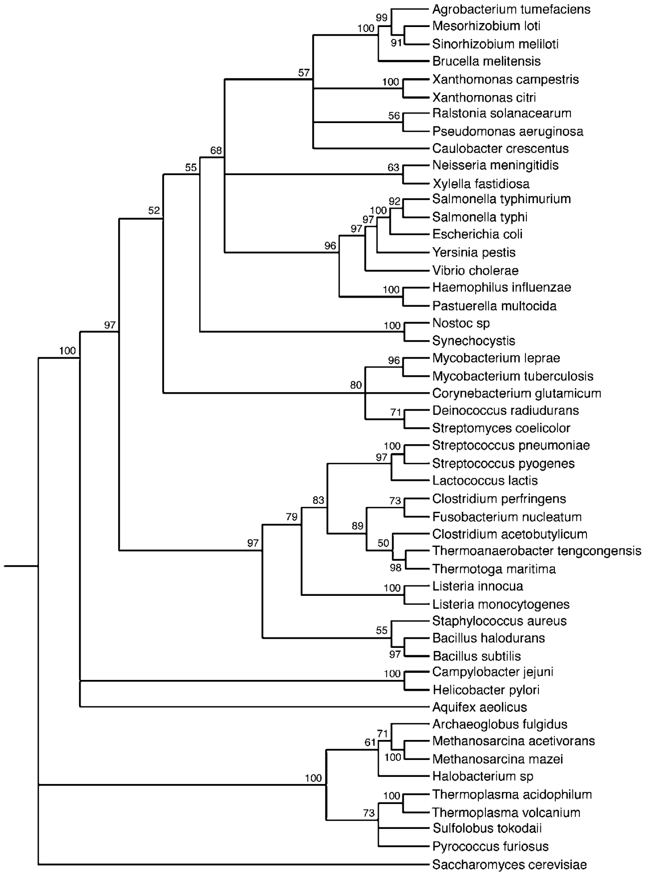

Figure 7. Unconstrained parsimony tree. This phylogeny is created using unconstrained parsimony based on the structural characters described in the text of the Supplemental material. Bootstrap proportions are obtained from 500 replicates and are labeled as percentages near each internal node.

quite contrary to the results of most other phylogenetic studies (Wolf et al. 2002). Of course, in the absence of an a priori reference phylogeny, it is impossible to make an absolute claim regarding the relative quality of this tree compared with the Dollo and NJ trees. Nonetheless, the phylogenetic relationships in the other regions of this tree are not improved when compared with the Dollo or NJ trees, and as a whole, the unconstrained tree represents a somewhat less-likely scenario of prokaryotic evolution when compared with the alternative methods above. As with the placement of *C. jejuni* and *H. pylori* in the Dollo tree, the assumption of unconstrained parsimony leads to a statistically robust, but nonetheless unreasonable placement (most likely) of the High G+C gram positives with the proteobacteria in this case.

A note on small genomes

As mentioned above, proteomes that contain fewer than 550 domains were omitted from the analysis due to a lack of statistical support for the power-law fits of the corresponding organismal subgraphs. These proteomes also display relatively high probability of being random subgraphs of the PDUG (i.e., have high P-values); in some cases (such as *Mycoplasma genitalium* and *Ureaplasma urealyticum*), the P-values are close to 0.5. As one might expect, when these proteomes are included for phylogenetic analysis with Dollo parsimony, they cluster together with one another and with *H. pylori* and *C. jejuni*, a grouping that has been observed in other contexts (Wolf et al. 2002). The only exceptions to this behavior are certain archaea (such as *Pyrococcus abyssi*) where the P-values are close to 2 × 10^{-4} (indicating less randomness in this proteome than *H. pylori* or *C. jejuni*). A similar placement of small-genome organisms occurs with the alternative methods (distance and unconstrained parsimony). In these cases, the lack of data combined with the large degree of randomness in the small proteomes results in unreliable placement of these organisms on the basis of structural domains.

Interesting groupings

Given that the Dollo and distance phylogenies largely support "known" associations between these taxa, one may wonder whether this analysis has provided anything more than support (however theoretically grounded) for well-known phylogenetic relationships. We will only discuss one of the potentially interesting groupings here, and that is the placement of *Thermotoga maritima* with *Thermoanaerobacter tengcongensis* in the low G+C region of the gram-positive clade. This grouping is quite robust statistically and methodologically, as evidenced by strong bootstrap support for this association in all of the trees examined in this work (see Figs. 4, 6, 7). This feature is notably absent from the SSU rRNA tree (Fig. 5), and has not been suggested in other approaches, partially due to the fact that the sequence of *T. tengcongensis* is relatively new (Brown et al. 2001; Bao et al. 2002; Wolf et al. 2002). Indeed, *T. maritima* is normally considered to be one of the most deeply branched bacteria (Barns et al. 1996; Bocchetta et al. 2000; Nesbo et al. 2001), although its placement in the rRNA tree may be an artifact of long-branch attraction (Gribaldo and Philippe 2002) (as is most likely the case with Fig. 5) or rapid evolution for G/C content to produce a thermally stable rRNA (Galtier and Lobry 1997). Several whole-genome methods have indicated a potential association between *T. maritima* and the low G+C gram positives (Wolf et al. 2001; Korbel et al. 2002), although other (limited) gene-content methods have implied a grouping with the proteobacteria (Brown et al. 2001).

We postulate two potential origins for this grouping in our data set, both of which have interesting implications. The first explanation suggests that there has been extensive LGT between these two lineages and that this LGT has involved the transfer of a large number of domains. The second explanation posits that *T. maritima* may have its ultimate origins within the gram-positive clade. Although it is difficult to conclusively distinguish between these possibilities at this juncture, some evidence may indicate that this behavior is the result of the descent of structural characters. Both proteomes are strongly nonrandom (*T. maritima* has a probability of being a random subgraph of the PDUG of 2 × 10^{-13}, *T. tengcongensis* a probability of 7 × 10^{-12}), which tends to argue against the LSDT explanation. Also, the mechanism of protein stabilization in *T. maritima* is quite distinct from the mechanisms used by Archaea (I. Berezovsky and E.I. Shakhnovich, unpubl.), which may argue against a very basal placement of this organism. Given that *T. maritima* was the bacterium with the fourth greatest extent of ORF overlap with *T. tengcongensis* when the *T. tengcongensis* genome was first sequenced and analyzed (Bao et al. 2002), it is clear that the relationship between these two organisms warrants further study.

Another statistically and methodologically robust feature of our trees is the association between the high G+C gram-positive bacteria and *Deinococcus radiudurans*. This particular grouping has been observed before (Brown et al. 2001; Wolf et al. 2001, 2002; Korbel et al. 2002), although it is absent in the rRNA tree (see Fig. 5). In many phylogenies, this grouping also includes the cyanobacteria (Wolf et al. 2001, 2002; Korbel et al. 2002), a feature not present in any of the trees presented in this study. Lack of association of the *D. radiudrurans*-actinobacteria group with the cyanobacteria in these trees has potentially interesting implications for the resolution of this particular area of the bacterial phylogenetic tree. Although much more work must be done to ensure the correct placement of the cyanobacteria within the phylogeny, it is interesting to note that *Synechtocystis* exhibits a relatively high probability of being a random subgraph, which, given its close association with *Nostoc sp.*, may overwhelm the low P-value of the other cyanobacterium, and thus make reliable placement in this (and other) phylogenies somewhat difficult.

Discussion

As mentioned in the introduction, protein structural domains represent an interesting source of phylogenetic information, not only because they may contain information complementary to that available from sequence data, but also because the theoretical exploration of structural evolution allows us to motivate the choice of a particular phylogenetic method. Consistent with this observation, we have found that methods that are less consistent with theoretical models of structural evolution result in less "reasonable" phylogenies. Indeed, we find that complete relaxation of the requirements suggested by these models gives somewhat unreliable results on a relatively large scale. It is interesting to note that the greater the degree to which the Dollo constraints are relaxed (i.e., the more the rules underlying evolutionary models are ignored), the more large-scale discrepancies occur between the tree based on structural domains and those based on other methods. This demonstrates that methods less consistent with these models (i.e., those that allow for convergence or LSDT) tend to misinterpret important phylogenetic signals revealed by more theoretically grounded methods. Thus, the Dollo tree not only provides a measure of theoretical support for the relationships derived using other phylogenetic methods and data, it also provides further evidence that the current evolutionary model is representative of the evolution of protein structural domains.

Taken as a whole, the results discussed above tend to indicate that LGT may not have influenced structural innovation in prokaryotes to the extent that it has influenced the innovation of genes (Aravind et al. 1998; Doolittle 1999; Ochman et al. 2000; Gogarten et al. 2002). Although this is somewhat surprising, this may be explained in light of two observations. The first is that novel genes may be created through novel combinations of existing structural domains, indicating that every lateral "gene" transfer event is not necessarily a lateral "domain" transfer event. Also, the amount of domain overlap between even relatively widely diverged organisms is extensive; for instance, 873 of the domains in *Bacillus subtilis* are also found in *Escherichia coli*, an overlap that represents nearly 90% of the *B. subtilis* structural proteome. In this case, the probability that an LGT event will transfer a new domain from *E. coli* to *B. subtilis* or vice versa is relatively low.

Our results do not imply, however, that LSDT has not played any role at all in determining the distribution of protein structural domains observed in prokaryotic organisms. Indeed, certain organisms exhibit proteomes with comparatively high probabilities of being random subgraphs of the PDUG, and in those cases, LSDT may have played a crucial role in the development of the structural repertoires of these organisms. Also, it is important to note that no attempt has been made to remove LSDT domains or randomizing signals from our data, and thus, some of the above phylogenetic groupings, even in regions of the tree with low P-values, may have been influenced in some measure by LSDT. Future work may allow for the identification of LSDT events, and thus lead to improvements in the structural phylogeny of prokaryotic organisms.

Given the above caveat, our trees do support or suggest a number of interesting and important phylogenetic conclusions, not the least of which involves providing further evidence for many of the groupings present in the rRNA, gene sequence, and gene-content phylogenies (Fitz-Gibbon and House 1999; Brown et al. 2001; Wolf et al. 2001, 2002; House and Fitz-Gibbon 2002; Korbel et al. 2002) should subsequent methods and data sets bring these relationships into question. As discussed in the introduction, the structural domains used in this work give insight into some of the more deeply branched groupings; for instance, *T. maritima* is reclassified in our trees as a member of the gram-positive clade, while *A. aeolicus* maintains a relatively basal position. These groupings, when taken with the placement of *D. radiudurans* and the cyanobacteria, indicate that the structural information used in this study have the potential to shed light on ancient aspects of prokaryotic phylogeny. These results, however, in no way indicate that structural information should supplant sequence-based methods; rather, they indicate that protein structure can provide interesting additional insights into phylogenetic relationships.

Although the above results are suggestive and relatively robust, the structural domains discussed here currently have several flaws that indicate that the contribution of structural information to the determination of phylogenies is far from complete. For instance, scientific exploration of the structural universe is nowhere near as complete as the exploration of sequences (Zhang and DeLisi 1998; Wolf et al. 2000; Gough et al. 2001; Sali 2001; Chothia et al. 2003), and so the data discussed above is far from the final picture of the distribution of protein structures in prokaryotic proteomes. Structural biology has not sampled structures with equal probability from all of these organisms or even from all domains of life (Wolf et al. 2000; Gough et al. 2001; Chothia et al. 2003), and there are many domains (i.e., sequence-structure pairs) whose structures have yet to be determined. It is possible that the lack of equal structural sampling may introduce biases into the data as it currently stands. Given that the above results are relatively reasonable, the continuing discovery of new structural domains should simply allow for more precise and robust phylogenies on the basis of these characters in the future. One may also imagine that structural domains may be combined with other sources of information to assist with the creation of ever-more robust phylogenies on the basis of diverse data sets. Protein structures thus represent an important source of additional information with which sequence information may be augmented in the future.

Our findings also have profound implications for the understanding of structural evolution. The success of certain assumptions as the basis methods for phylogenetic inference gives some clues as to the appropriate algorithms for computing the evolution of domains on trees with fixed topologies (Mirkin et al. 2003). This will allow for the construction of a "likely" history for domain evolution, which could be used to build a "forward" picture of domain evolution. This picture may then be compared with models of domain evolution and may suggest further refinements to those models. One may envisage that this process could eventually converge on a self-consistent theory of domain evolution, and such a theory may represent the closest possible approach to a complete understanding of protein structural evolution.

Methods

Random subgraphs

In order to calculate the probability that an organismal subgraph is random, we developed an analytical approximation to the degree distribution of a random subgraph (Deeds et al. 2004). Given an "underlying" graph (such as the PDUG) with N_0 nodes and a degree distribution $p_{N_0}(k)$, the degree distribution of N nodes chosen from this graph completely randomly should follow:

$$p_N(k) = \sum_{s=k}^{Maxk_{N_0}} \binom{s}{k} \left(\frac{N}{N_0}\right)^k \left(1 - \frac{N}{N_0}\right)^{s-k}$$

where $p_N(k)$ is the degree distribution in the subgraph and $Maxk_{N_0}$ is the degree of the maximally connected node in the underlying graph. This approximation is very accurate (Deeds et al. 2004) and is used to estimate the probability that a particular organismal subgraph is a random subgraph.

LSDT model

The LGT model is based on divergent evolutionary models for the PDUG (Dokholyan et al. 2002; Deeds et al. 2003, 2004). In this case, the evolution of domains in the PDUG is unchanged—duplication and divergence occur as in the previous models, and the parameters (such as a structural cutoff of 0.5) are identical to those used in Deeds et al. (2004). Each duplication and divergence event occurs within a specific proteome, and so when new nodes are added to the graph, they are also added to the subgraph of a particular organism. Speciation events (which occur at some frequency relative to duplication events) create two identical but separate copies of each existing organism. The resulting organisms then evolve nodes independently until the next speciation event occurs.

Lateral gene transfer is modeled as the movement of a node from a proteome in which that node exists into a proteome in which it does not. Transfer does not remove the node from the "donor" organism, but it may replace (thus, "erase") one of the nodes in the acceptor organism in order to preserve proteome size. The donor and acceptor organisms are chosen randomly, and the transferred node is chosen randomly from the set of nodes in the donor proteome that do not exist in the acceptor proteome. The acceptor node that is replaced is also chosen at random. LSDT events occur with some frequency compared with duplication events after the first speciation event has occurred. The LSDT model is represented schematically in Figure 2. Models in which domains are simply added to proteomes without replacing currently existing domains show similar behavior to those displayed in Figure 3. As discussed in the text, models in which LSDT may only occur between closely related organisms also exhibit similar behavior.

Structural characters

The structural domains that we use in this work correspond to all of the PDUG domains (Dokholyan et al. 2002) that are found within the structural proteomes of the set of fully sequenced prokaryotes (Deeds et al. 2004), representing 1818 of the 3464 domains on the PDUG. When combined with the structural proteome of *S. cerevisiae*, this results in a set of 1577 parsimony-informative structural characters.

The data matrix used in this work may be obtained in the NEXUS file format from our Supplemental Web site http://paradox.harvard.edu/~eric/struct_phylo.htm. Structural proteomes of these and other organisms may be accessed at varying levels of BLAST certainty from the ELISA Web site (http://romi.bu.edu/elisa) (Shakhnovich et al. 2003). The E-value for domains used in this work was set at 10^{-6} as in Deeds et al. (2004).

Small-subunit rRNA tree

To allow direct comparison between the structural results and results of phylogenies created using the small-subunit (16S) ribosomal RNA sequence, we have provided a phylogeny of these taxa based on SSU rRNA data from the Ribosomal Database Project (RDP) (Cole et al. 2003). In some cases, the SSU rRNA sequence is not available from the RDP for an organism in this list of taxa, and in those cases, either that organism is omitted from the tree or a presumably closely related organism is included instead. Certain taxa are also denoted by alternative names in this data set; for instance, *Bacillus halodurans* is indicated by its synonym *Bacillus alcalophilus*. The phylogeny is calculated using the RDP Web site and is based on the Neighbor Joining algorithm and the default parameters provided by the RDP (Cole et al. 2003). This phylogeny is only provided as a rough guide to the set of phylogenetic relationships among these taxa that would be predicted on the basis of the SSU rRNA tree.

Phylogenetic inference

As mentioned in the text, all of the structural phylogenies used in this work were created using the PAUP* software package, version 4 (beta 10) (Swofford 2003). The Dollo parsimony tree is based on setting the character-type assumption to "dollo.up" and calculating the MP tree using the available heuristic algorithm. The neighbor-joining tree is created using the total number of distances (or Hamming distance) (Lin and Gerstein 2000) between two sets of characters as the distance between the two organisms. In the case shown in Figure 6, the distance criterion is set to minimum evolution in PAUP* (Swofford 2003); however, the results from least-squares methods are very similar. Internal nodes are labeled with bootstrap proportions that are obtained from 500 replicates. The unconstrained parsimony tree is created using the "unord" character type assumption in PAUP* (Swofford 2003). In every case, internal nodes are labeled with bootstrap proportions that are obtained from 500 replicates.

Acknowledgments

We thank Drs. G. Giribet and S. Baldauf for their comments on the manuscript. We are deeply indebted to Dr. Eugene Koonin for all of his advice. This work was supported by an HHMI predoctoral fellowship (to E.J.D.) and by the NIH.

Note added in proof

After this work was completed, Dutilh et al. (2004) published a study in which they attempted to remove LGT-induced noise in whole-genome gene-content data sets by removing genes from consideration on the basis of a given gene's level of discordance (or lack of monophyly) within the overall phylogeny (Dutilh et al. 2004). We note that the phylogeny produced via this method has many similarities with the Dollo parsimony structural domain tree, including the association between *T. maritima* and *T. tengcongensis* in the low G+C gram-positive clade (although this method does not directly link the two taxa aside from randomly initialized runs of their algorithm). These authors also find an association between *D. radiudurans* and the actinobacteria that does not include the cyanobacteria. These similarities imply that the "character-pruning" method used by these authors reveals signals similar to those present in the (mostly) nonrandom structural proteomes used in this work, although the lack of a P-value analog for COGs prevents an a priori assessment of the reduction of randomness in their case.

The Dutilh et al. (2004) (results do not support monophyly of the gram-positive bacteria, although the association in this case is between the low G+C gram positives and the proteobacteria, rather than the high G+C gram positives and the proteobacteria. It is possible that use of the Dollo method (rather than a distance-based method) on the "low-noise" orthologous groups produced by their algorithm might result in a monophyletic gram-positive clade. Application of a similar procedure to structural proteomes based not on phylogenetically discordant signals (which is an inherently circular method), but rather on minimization of P-values, might reveal an even more robust and informative prokaryotic phylogeny.

References

Albert, R. and Barabasi, A.-L. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74:** 47–97.

Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R., and Koonin, E.V. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14:** 442–444.

Bao, Q., Tian, Y., Li, W., Xu, Z., Xuan, Z., Hu, S., Dong, W., Yang, J., Chen, Y., Xue, Y., et al. 2002. A complete sequence of the *T. tengcongensis* genome. *Genome Res.* **12:** 689–700.

Barns, S.M., Delwiche, C.F., Palmer, J.D., and Pace, N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci.* **93:** 9188–9193.

Bocchetta, M., Gribaldo, S., Sanangelantoni, A., and Cammarano, P. 2000. Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J. Mol. Evol.* **50:** 366–380.

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., and Stanhope, M.J. 2001. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28:** 281–285.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. 2003. Evolution of the protein repertoire. *Science* **300:** 1701–1703.

Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., McGarrell, D.M., Schmidt, T.M., Garrity, G.M., et al. 2003. The Ribosomal Database Project (RDP-II): Previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.* **31:** 442–443.

Deeds, E.J., Dokholyan, N.V., and Shakhnovich, E.I. 2003. Protein evolution within a structural space. *Biophys. J.* **85:** 2962–2972.

Deeds, E.J., Shakhnovich, B., and Shakhnovich, E.I. 2004. Proteomic traces of speciation. *J. Mol. Biol.* **336:** 695–706.

Dietmann, S. and Holm, L. 2001. Identification of homology in protein structure classification. *Nat. Struct. Biol.* **8:** 953–957.

Dokholyan, N.V., Shakhnovich, B., and Shakhnovich, E.I. 2002. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci.* **99:** 14132–14136.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* **284:** 2124–2129.

Dutilh, B.E., Huynen, M.A., Bruno, W.J., and Snel, B. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* **58:** 527–539.

Farris, J.S. 1977. Phylogenetic analysis under Dollo's Law. *Syst. Zool.* **26:** 77–88.

Fitz-Gibbon, S.T. and House, C.H. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27:** 4218–4222.

Galtier, N. and Lobry, J.R. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44:** 632–636.

Giribet, G. 2002. Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion? *Mol. Phylogenet. Evol.* **24:** 345–357.

Gogarten, J.P., Doolittle, W.F., and Lawrence, J.G. 2002. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19:** 2226–2238.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313:** 903–919.

Gribaldo, S. and Philippe, H. 2002. Ancient phylogenetic relationships. *Theor. Popul. Biol.* **61:** 391–408.

Holm, L. and Sander, C. 1996. The FSSP database: Fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* **24:** 206–209.

House, C.H. and Fitz-Gibbon, S.T. 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: An update. *J. Mol. Evol.* **54:** 539–547.

Koonin, E.V., Wolf, Y.I., and Karev, G.P. 2002. The structure of the protein universe and genome evolution. *Nature* **420:** 218–223.

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.

Lin, J. and Gerstein, M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. *Genome Res.* **10:** 808–818.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3:** 2.

Nesbo, C.L., L'Haridon, S., Stetter, K.O., and Doolittle, W.F. 2001. Phylogenetic analyses of two "archaeal" genes in thermotoga maritima reveal multiple transfers between archaea and bacteria. *Mol. Biol. Evol.* **18:** 362–375.

Ochman, H., Lawrence, J.G., and Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:** 299–304.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Sali, A. 2001. Target practice. *Nat. Struct. Biol.* **8:** 482–484.

Shakhnovich, B.E., Harvey, J.M., Comeau, S., Lorenz, D., DeLisi, C., and Shakhnovich, E. 2003. ELISA: Structure-function inferences based on statistically significant and evolutionarily inspired observations. *BMC Bioinformatics* **4:** 34.

Swofford, D.L. 2003. Paup*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA.

Wolf, Y.I., Brenner, S.E., Bash, P.A., and Koonin, E.V. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9:** 17–26.

Wolf, Y.I., Grishin, N.V., and Koonin, E.V. 2000. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299:** 897–905.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L., and Koonin, E.V. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1:** 8.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., and Koonin, E.V. 2002. Genome trees and the tree of life. *Trends Genet.* **18:** 472–479.

Zhang, C. and DeLisi, C. 1998. Estimating the number of protein folds. *J. Mol. Biol.* **284:** 1301–1305.

Web site references

http://paradox.harvard.edu/~eric/struct_phylo.htm; Author's Web site.
http://romi.bu.edu/elisa; ELISA

Received July 21, 2004; accepted in revised form January 6, 2005.