Comparison of Phylogenetic Trees and Search for a Central Trend in the "Forest of Life''

3 authors, including:

Eugene V Koonin
National Institutes of Health
1,012 PUBLICATIONS   **108,401** CITATIONS

SEE PROFILE

Pere Puigbo
University of Turku
69 PUBLICATIONS   **2,898** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Law of mass action in protein interaction networks View project

Project   Evolution of microbial enzymes and pathways View project

Comparison of Phylogenetic Trees and Search for a Central Trend in the "Forest of Life"

EUGENE V. KOONIN, PERE PUIGBÒ, and YURI I. WOLF

ABSTRACT

The widespread exchange of genes among prokaryotes, known as horizontal gene transfer (HGT), is often considered to "uproot" the Tree of Life (TOL). Indeed, it is by now fully clear that genes in general possess different evolutionary histories. However, the possibility remains that the TOL concept can be reformulated and remain valid as a statistical central trend in the phylogenetic "Forest of Life" (FOL). This article describes a computational pipeline developed to chart the FOL by comparative analysis of thousands of phylogenetic trees. This analysis reveals a distinct, consistent phylogenetic signal that is particularly strong among the Nearly Universal Trees (NUTs), which correspond to genes represented in all or most of the analyzed organisms. Despite the substantial amount of apparent HGT seen even among the NUTs, these gene transfers appear to be distributed randomly and do not obscure the central tree-like trend.

Key words: evolution, genomics.

1. INTRODUCTION

THE TREE OF LIFE (TOL) is one of the dominant concepts in biology, first seen in the famous single illustration in Darwin's "*Origin of Species*," and now in 21st century undergraduate textbooks. For approximately a century, beginning with the first, tentative trees published by Haeckel in the 1860s and up to the foundation of molecular evolutionary analysis by Zuckerkandl, Pauling, and Margoliash in the early 1960s, phylogenetic trees were constructed on the basis of comparing phenotypes of organisms. Thus, by design, every constructed tree was an "organismal" or "species" tree; that is, a tree was assumed to reflect the evolutionary history of the corresponding species. Even after the concepts and early methods of molecular phylogeny have been developed, for many years, it was used simply as another, perhaps, particularly powerful and accurate approach to the construction of species trees. The TOL concept remained intact, with the general belief that the TOL, at least in principle, would accurately represent the evolutionary relationships between all lineages of cellular life forms. The discovery of the universal conservation of rRNA and its use as the molecule of choice for phylogenetic analysis pioneered by Woese and others (Pace et al., 1986; Woese, 1987) resulted in the discovery of a new domain of life, the archaea, and boosted the hopes that the definitive topology of the TOL was within sight.

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

However, even before the era of complete genome sequencing and analysis, it has become clear that in prokaryotes some common and biologically important genes have experienced multiple exchanges between species known as horizontal gene transfer (HGT); hence, there is the idea of a "net of life" as an alternative to the TOL. The advances of comparative genomics have revealed that different genes very often have distinct tree topologies and, accordingly, that HGT appears to be the rule rather than an exception in the evolution of prokaryotes—bacteria and archaea (Dagan et al., 2008; Doolittle, 1999; Gogarten and Townsend, 2005).

It seems worth mentioning some remarkable examples of massive HGT as an illustration of this key trend in the evolution of prokaryotes. The first case in point pertains to the most commonly used model of microbial genetics and molecular biology, the intestine bacterium *Escherichia coli*. Some basic information on the genome of *E. coli* and other sequenced microbial genomes is available on the website of the National Center for Biotechnology Information at the National Institute of Health (http://www.ncbi.nlm.nih.gov/sites/entrez?db = genome). The most well studied laboratory isolate of *E. coli* on which most of the classic experiments of molecular biology have been performed is known as K12. The K12 genome encompasses 4226 annotated protein-coding genes (there is always uncertainty as to the exact number of the genes in a sequenced genome, for instance, because it remains unclear whether or not some small genes actually encode proteins; however, the estimate suffices for the present discussion). Several other sequenced genomes of laboratory *E. coli* strains possess about the same number of genes. In contrast, genomes of pathogenic strains of *E. coli* typically are much larger, with one strain, O157:H7, encoding 5315 annotated proteins. The nucleotide sequences of the shared genes in all strains of *E. coli* are identical or differ by just one or two nucleotide substitutions. In a stark contrast, the differences between the genomes of laboratory and pathogenic strains concentrate in several "pathogenicity islands" that comprise up to 20% of the genome. The pathogenicity islands encompass genes typically involved in bacterial pathogenesis such as toxins, systems for their secretion, and components of prophages. One can imagine that the pathogenicity islands were present in the ancestral *E. coli* genome but have been deleted in K12 and other laboratory strains. However, the gene contents of the islands dramatically differ between the pathogenic strains, so that in three-way comparisons of *E. coli* genomes only about 40% of the genes are typically shared. Thus, the only possible conclusion is that the pathogenicity islands spread between bacterial genomes via rampant HGT, conceivably, driven by selection for survival and spread of the respective bacterial pathogens within the host organisms.

The second example involves apparent large-scale HGT across much greater evolutionary distances, namely, between the two "domains" of prokaryotes—bacteria and archaea (Pace et al., 1986; Woese, 1987). The distinction between these two distinct domains of microbes was established by phylogenetic analysis of rRNA sequences and the sequences of other conserved genes, and has been supported by major distinctions between the systems of DNA replication and the membrane apparatus of the respective organisms. Comparative analysis of the first few sequenced genomes of bacteria and archaea supported the dichotomy between the two domains: most of the protein sequences encoded in bacterial genomes show the greatest similarity to homologs from other bacteria and cluster with them in phylogenetic trees, and the same pattern of evolutionary relationships is seen for archaeal proteins. However, the analysis of the first sequenced genomes of hyperthermophilic bacteria, *Aquifex aeolicus* and *Thermotoga maritima*, yielded a striking departure from this pattern: the protein sets encoded in these genomes were shown to be "chimeric," i.e., they consist of about 80% typical bacterial proteins and about 20% proteins that appear distinctly "archaeal," by sequence similarity and phylogenetic analysis. The conclusion seems inevitable that these bacteria have acquired numerous archaeal genes via HGT. In retrospect, this finding might not appear so surprising because bacterial and archaeal hyperthermophiles coexist in the same habitats (e.g., hydrothermal vents on the ocean floor) and have ample opportunity to exchange genes. Similar chimeric genome composition, but with reversed proportions of archaeal and bacterial genes, has been subsequently discovered in mesophilic archaea such as *Methanosarcina*.

Beyond these and related observations made by comparative genomics of prokaryotes, HGT is thought to have been crucial also in the evolution of eukaryotes, especially, as a consequence of endosymbiotic events in which numerous genes from the genome of the ancestors of mitochondria and chloroplasts have been transferred to nuclear genomes (Embley and Martin, 2006). These findings indicate that no single gene tree (or any group of gene trees) can provide an accurate representation of the evolution of entire genomes; in other words, the results of comparative genomics indicate that a perfect TOL fully reflecting the evolution of cellular life forms does not exist. The realization that HGT is a major evolutionary phenomenon, at least,

among prokaryotes, led to a crisis of the TOL concept which is often viewed as a paradigm shift in evolutionary biology (Doolittle, 1999).

Of course, the inconsistency between gene phylogenies caused by HGT, however widespread, does not alter the fact that all cellular life forms are linked by an uninterrupted tree of cell divisions (*Omnis cellula e cellula,* according to the famous motto of Rudolf Virchow) that goes back to the earliest stages of evolution and is violated only by endosymbiosis events that were key to the evolution of eukaryotes but not pro-karyotes. Thus, the difficulties of the TOL concept in the era of comparative genomics concern the TOL as it can be derived by the phylogenetic analysis of multiple genes and genomes, an approach often denoted "phylogenomics", to emphasize that phylogenetic studies are now conducted on the scale of complete genomes. Accordingly, the claim that HGT "uproots the TOL" means that extensive HGT has the potential to completely decouple molecular phylogenies from the actual tree of cells. However, such decoupling has clear biological connotations given that the evolutionary history of genes also describes the evolution of the encoded molecular functions. In this article, the phylogenomic TOL is discussed with such an implicit understanding.

The views of evolutionary biologists on the evolving status of the TOL in the age of comparative genomics span the entire spectrum of positions from: i) persisting denial of the major importance of HGT for evolutionary biology to ii) "moderate" overhaul of the TOL concept; to iii) genuine uprooting whereby the TOL is declared obsolete (Doolittle and Bapteste, 2007). The accumulating data on diverse HGT events are quickly making the first, "anti-HGT" position plainly untenable. Under the intermediate, moderate approach, despite all the differences between the topologies of individual gene trees, the TOL still makes sense as a representation of a central trend (consensus) that, at least, in principle, could be elucidated through a comprehensive comparison of trees for individual genes (Wolf et al., 2002). By contrast, under the radical "anti-TOL" view, rampant HGT eliminates the very distinction between the vertical and horizontal transmission of genetic information, so the TOL concept should be abandoned altogether in favor of some form of a network representation of evolution (Doolittle and Bapteste, 2007).

This article describes some of the methods that are used to compare topologies of numerous phylogenetic trees and the results of the application of these approaches to the analysis of approximately 7000 phylo-genetic trees of individual prokaryotic genes that collectively comprise the "Forest of Life" (FOL). This set of trees does gravitate to a single tree topology, suggesting that the "TOL as a central trend" concept is potentially viable.

2. BIOINFORMATIC PIPELINE FOR ANALYSIS OF THE FOREST OF LIFE

The realization that, owing to widespread HGT, the evolutionary history of each gene is in principle unique brings the emphasis on phylogenomics, that is, genome-wide comparative analysis of phylogenetic trees. This task depends on a bioinformatic pipeline which leads from protein sequences encoded in the analyzed genomes to a representative collection of phylogenetic trees (Fig. 1). The pipeline consists of several essential steps: (1) selection of genes for phylogenetic analysis, (2) multiple alignment of ortho-logous protein sequences, that is, amino acid sequences of proteins encoded by "the same" gene from different organisms (in evolutionary biology, such genes are usually called orthologs), (3) construction of phylogenetic trees, (4) calculation of the distances between trees and construction of a tree distance matrix, and (5) clustering and classification of trees on the basis of the distance matrix. Obviously, this pipeline incorporates a variety of computational methods, and it is impractical to present all of them in detail within a relatively short article. However, a brief outline of these methods is given below. The current collection of complete microbial genomes includes over 1000 organisms (http://www.ncbi.nlm.nih.gov/genomes/ MICROBES/microbial_taxtree.html), so it is impractical to use them all for phylogenetic analysis that quickly becomes prohibitively computationally expensive with the increase of the number of species. Therefore, the FOL was analyzed using a manually selected representative set of 100 prokaryotes (Puigbo et al., 2009).

The great majority of orthologous gene clusters include a relatively small number of organisms. In the set of clusters selected for phylogenomic analysis of the FOL, the distribution of the number of species in trees showed exponential decay, with only about 2000 out of the approximately 7000 clusters including more than 20 species (Fig. 2). The truly universal gene core of cellular life is tiny and continues to shrink as new genomes are sequenced, owing to the loss of "essential" genes in some organisms with small genomes and
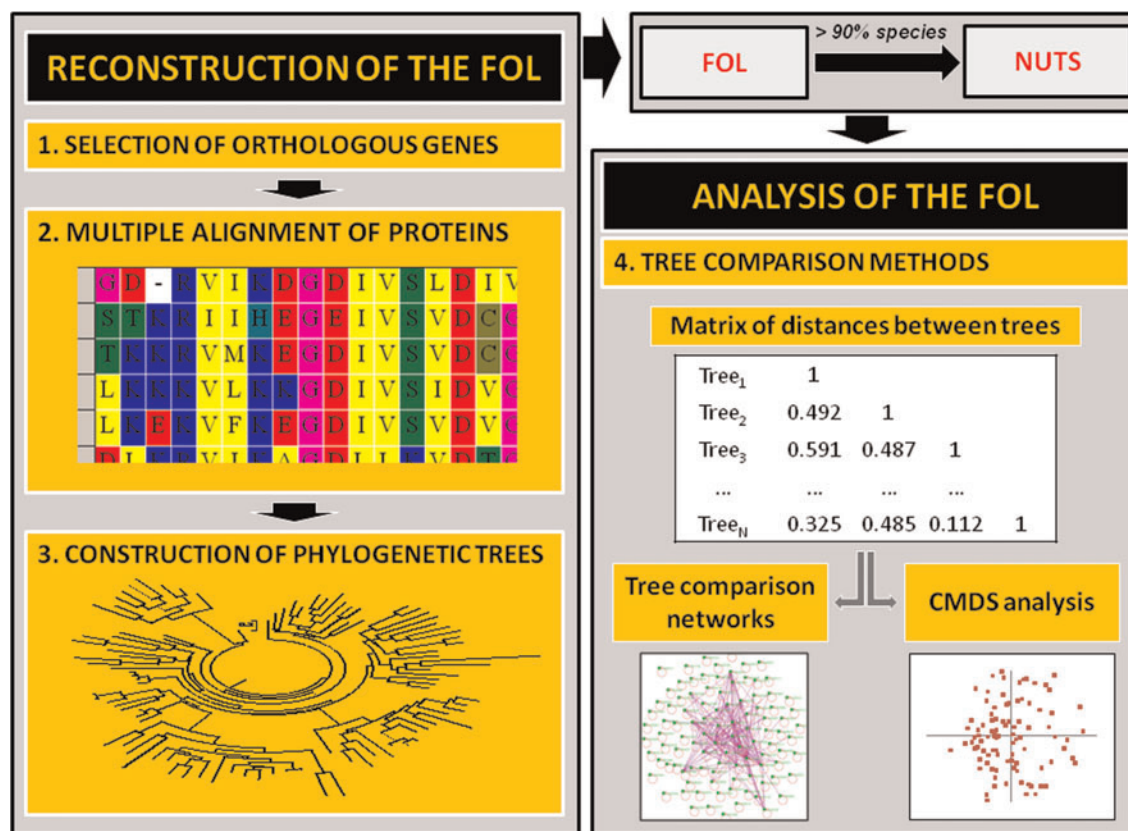
FIG. 1. The bioinformatic pipeline for the analysis of the Forest of Life (FOL).

to errors of genome annotation. Among the trees in the FOL, there were about 100 Nearly Universal Trees (NUTs), that is, trees for gene families represented in all or nearly all analyzed organisms; almost all NUTs correspond to genes encoding proteins involved in translation and transcription (Puigbo et al., 2009). The NUTs were analyzed in parallel with the complete set of trees in the FOL.

Before constructing a phylogenetic tree, the sequences of orthologous genes or proteins need to be aligned, that is, all homologous positions have to be identified and positioned one under another to allow subsequent comparative analysis of the sequences. For large evolutionary distances, as is the case between many members of the analyzed set of 100 microbial genomes, trees are constructed using multiple alignments of protein sequences (Fig. 1).

Once the sequences of orthologous proteins are aligned, the construction of phylogenetic trees becomes possible. Many diverse approaches and algorithms have been developed for building phylogenetic trees. There is no single "best" phylogenetic method that would be optimal for solving any problem in evolution,

FIG. 2. The distribution of the trees in the Forest of Life (FOL) by the number of species.
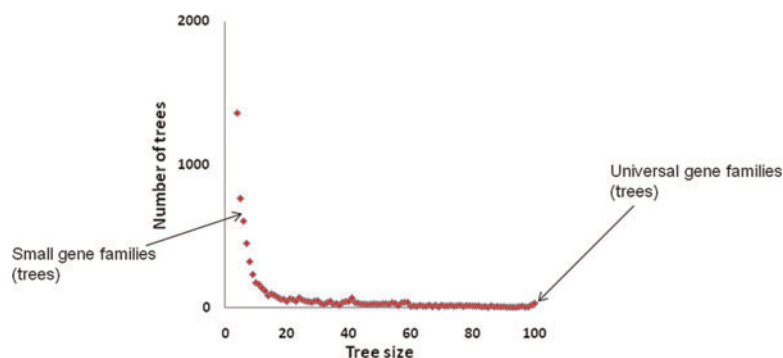
but in general the highest quality of phylogenetic reconstruction is achieved with maximum likelihood methods that employ sophisticated probabilistic models of gene evolution (Felsenstein, 2004).

The construction of the trees (about 7000 altogether) provides for an attempt to identify patterns in the FOL and address the question whether or not there exists a central trend among the trees that perhaps could be considered an approximation of a TOL. To perform such an analysis, it is necessary first to build a complete, all-against-all matrix of the topological distances between the trees; obviously, this matrix is a big, approximately 7000×7000 square table in which each cell contains a distance between two trees.

So how does one compare phylogenetic trees, and how are the distances in the matrix calculated? Comparison of trees is much less commonly used than phylogenetic analysis per se, but in the age of genomics, it is rapidly becoming a mainstream methodology. Essentially, what is typically compared are the topologies (that is, the branching order) of the trees, and the distance between the topologies can be captured as the fraction of the tree ''splits'' that are different (or common) between two compared trees (Fig. 3). An additional idea implemented in the method for tree topology comparison illustrated in Figure 3 is to take into account the reliability of the internal branches of the tree, so that the more reliable branches contribute more than the dubious ones to the distance estimates. The reliability or statistical support for tree branches is usually estimated in terms of the so called bootstrap values that vary from 0 (no support at all) to 1 (the strongest support). In the Boot Split Distance (BSD) method for tree topology comparison illustrated in Figure 3, the contribution of each split is weighted using the bootstrap values.

3. TRENDS IN THE FOREST OF LIFE

3.1. The NUTs contain a consistent phylogenetic signal, with independent HGT events

Figure 4 represents the NUTs as a network in which the edges are drawn on the basis of the topological distances between the trees (see the preceding section and Fig. 3). Clearly, the topologies of the NUTs are highly coherent, so that when a relatively short distance of 0.5 is used as the threshold to draw edges in the network, almost all the nodes in the network are connected (Fig. 4b). In 56% of the NUTs, representatives of the two prokaryotic domains, archaea and bacteria, are perfectly separated, whereas the remaining 44% of the NUTs showed indications of HGT between archaea and bacteria. Of course, even in the 56% of the NUTs that showed no sign of interdomain gene transfer, there were many probable HGT events within one or both domains, indicating that HGT is indeed common, even in this group of nearly universal genes.
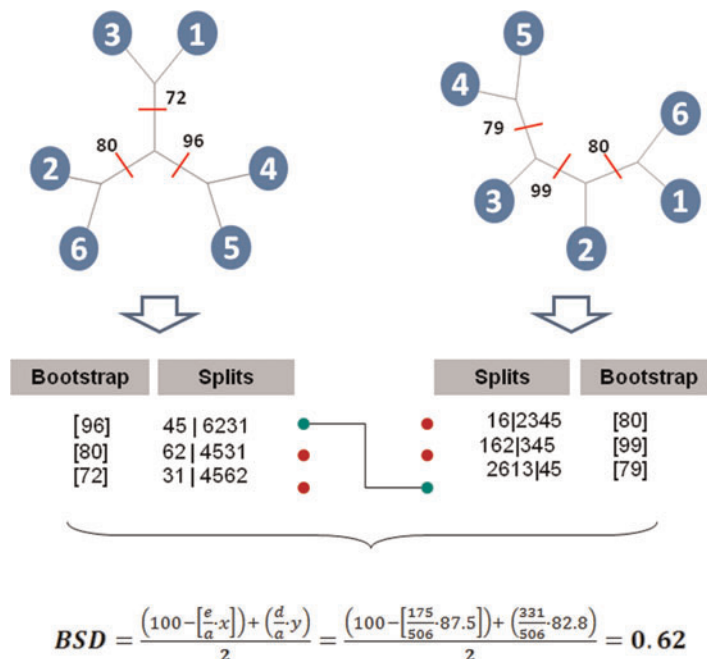

FIG. 3. Comparison of phylogenetic tree topologies. Identical (equal) splits are shown by connected green circles, and different splits are shown by red circles. Bootstrap values are shown as percent. The Boot Split Distance (BSD) between the trees was calculated using the formula shown in the figure. The designations are:

$e = \sum Bootstrap\ of\ equal\ splits$

$d = \sum Bootstrap\ of\ different\ splits$

$a = \sum Bootstrap\ of\ all\ splits$

$x = Mean\ Bootstrap\ of\ equal\ splits$

$y = Mean\ Bootstrap\ of\ different\ splits$

$$BSD = \frac{\left(100 - \left[\frac{e}{a} \cdot x\right]\right) + \left(\frac{d}{a} \cdot y\right)}{2} = \frac{\left(100 - \left[\frac{175}{506} \cdot 87.5\right]\right) + \left(\frac{331}{506} \cdot 82.8\right)}{2} = 0.62$$
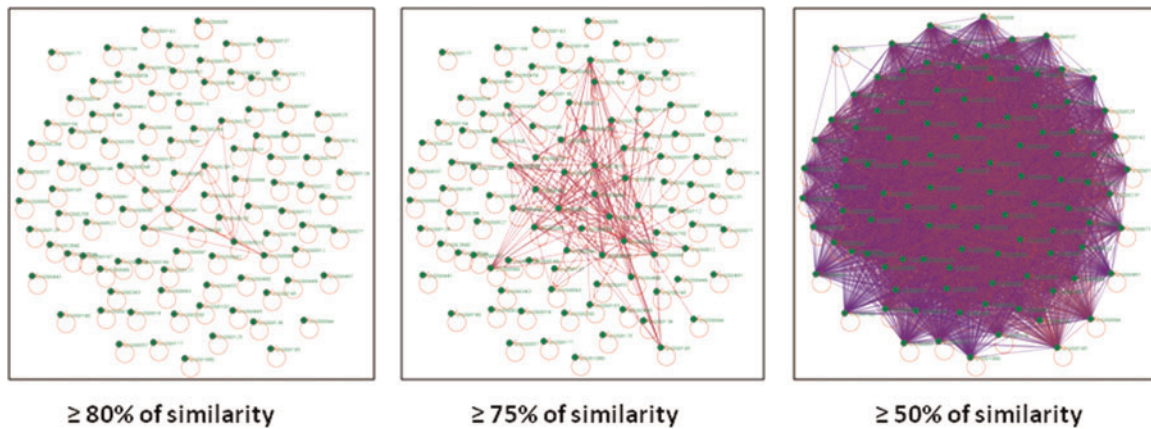
FIG. 4. The network of similarities among the Nearly Universal Trees (NUTs). Each node denotes a NUT, and nodes are connected by edges if the topological similarity between the respective trees exceeds the indicated threshold (in other words, if the distance between these trees is sufficiently low). The circular arrows show that each node is connected with itself.

To analyze the structure of a distance matrix between any objects, including phylogenetic trees, researchers often use so-called multidimensional scaling that reveals clustering of the compared objects. Cluster analysis of the NUTs using the Classical MultiDimensional Scaling (CMDS) method shows lack of significant clustering: all the NUTs formed a single, unstructured cloud of points (Fig. 5a). This organization of the tree space is best compatible with random deviation of individual NUTs from a single, dominant topology, mostly as a result of HGT but also in part due to random errors of the tree-construction procedure. The results of this analysis indicate that the topologies of the NUTs are scattered within a close vicinity of a consensus tree, with the HGT events distributed at least approximately randomly, a finding that is compatible with the idea of a "TOL as a central trend."
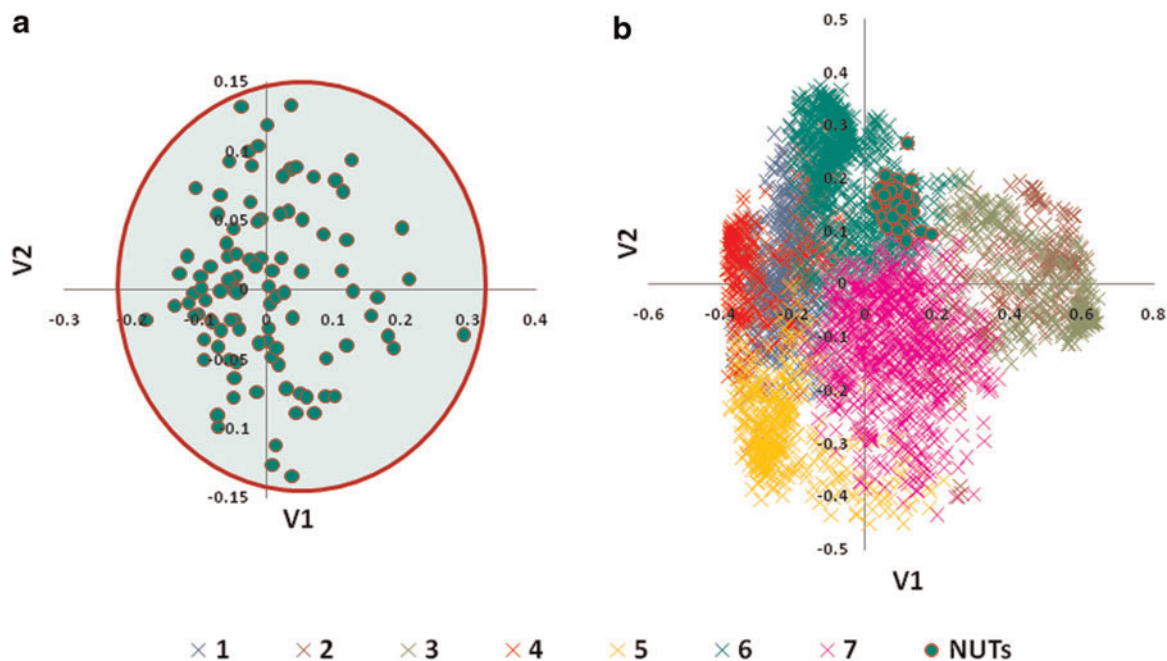

FIG. 5. Clustering of the Nearly Universal Trees (NUTs) and the entire Forest of Life (FOL) using the Classical MultiDimensional Scaling (CMDS) method. (**a**) The best two-dimensional projection of the clustering of the 102 NUTs in a 30-dimensional space. (**b**) The best two-dimensional projection of the clustering of 3789 largest trees from the FOL in a 669-dimensional space. The seven clusters are color-coded, and the NUTs are shown by circles.
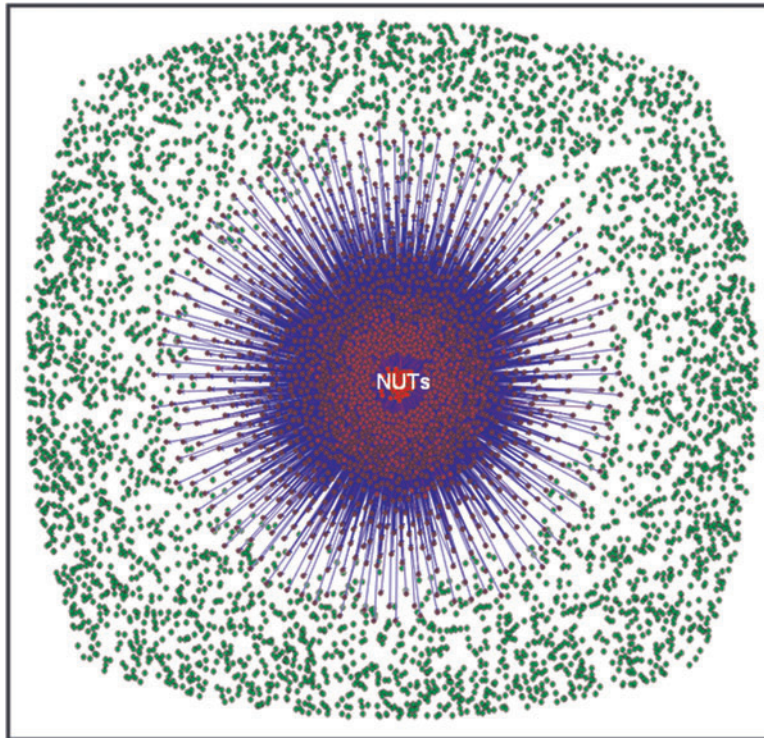
FIG. 6. The Forest of Life (FOL) network and the Nearly Universal Trees (NUTs). The figure shows a network representation of the 6,901 trees in the FOL. The 102 NUTs are shown as red circles in the middle. The NUTs are connected to trees with similar topologies: trees that show at least 50% of similarity with at least one NUT are shown as purple circles and are connected to the NUTs. The rest of the trees are denoted by green circles.

3.2. The NUTs versus the FOL

The structure of the FOL was analyzed using the CMDS procedure, with the results being very different from those seen with the NUTs: in this case, seven distinct clusters of trees were revealed (Fig. 5b). The clusters significantly differed with respect to the distribution of the trees by the number of species, the partitioning of archaea-only and bacteria-only trees, and the functional classification of the respective genes (Puigbo et al., 2009). Notably, all the NUTs formed a compact group within one of the clusters and were roughly equidistant from the rest of the clusters (Fig. 5b). Thus, the FOL seems to contains several distinct ''groves'' of trees with different evolutionary histories. The critical observation is that all the NUTs occupy a compact and contiguous region of the tree space and, unlike the complete set of the trees, are not partitioned into distinct clusters by the CMDS procedure (Fig. 5a). Moreover, the NUTs are, on average, highly similar to the rest of the trees in the FOL as shown in Figure 6. Taken together, these findings suggest that the NUTs collectively could represent a central trend in the FOL.

4. CONCLUSION

Prokaryotic genomics revealed the wide spread of HGT in the prokaryotic world and is often claimed to ''uproot'' the TOL (Doolittle, 1999). Indeed, it is now well established that HGT spares virtually no genes at some stages in their history (Gogarten and Townsend, 2005), and these findings make obsolete a ''strong'' TOL concept under which all (or the substantial majority) of the genes would tell a consistent story of genome evolution (the species tree, or the TOL) when analyzed using appropriate data sets and methods. However, is there any hope of salvaging the TOL as a statistical central trend (Wolf et al., 2002)? Comprehensive comparative analysis of the ''forest'' of phylogenetic trees for prokaryotic genes outlined here suggests a positive answer to this crucial question of evolutionary biology (Puigbo et al., 2009).

This analysis results in two complementary conclusions. On the one hand, there is a high level of inconsistency among the trees comprising the FOL, owing primarily to extensive HGT, a conclusion that is supported by more direct observations of numerous likely transfers of genes between archaea and bacteria. However, there is also a distinct signal of a consensus topology that was particularly strong among the

NUTs. Although the NUTs show a substantial amount of apparent HGT, these transfers seem to be distributed randomly and did not obscure the vertical signal. Moreover, the topologies of the NUTs are quite similar to those of numerous other trees in the FOL, so although the NUTs cannot represent the FOL completely, this set of largely consistent, nearly universal trees is a good candidate for representing a central trend.

ACKNOWLEDGMENTS

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Dagan, T., Artzy-Randrup, Y., and Martin, W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. USA* 105, 10039–10044.

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

Doolittle, W.F., and Bapteste, E. 2007. Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. USA* 104, 2043–2049.

Embley, T.M., and Martin, W. 2006. Eukaryotic evolution, changes and challenges. *Nature* 440, 623–630.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Gogarten, J.P., and Townsend, J.P. 2005. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687.

Pace, N.R., Olsen, G.J., and Woese, C.R. 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell* 45, 325–326.

Puigbo, P., Wolf, Y.I., and Koonin, E.V. 2009. Search for a Tree of Life in the thicket of the phylogenetic forest. *J. Biol.* 8, 59.

Woese, C.R. 1987. Bacterial evolution. *Microbiol. Rev.* 51, 221–271.

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

Address correspondence to:
Dr. Eugene V. Koonin
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894

E-mail: koonin@ncbi.nlm.nih.gov