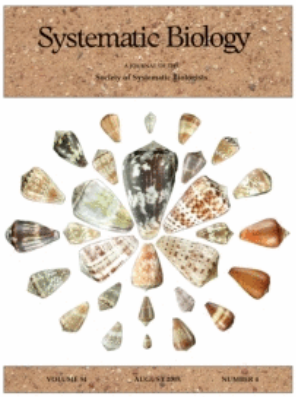

Systematic Biology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713658732

Evidence for a New Root of the Tree of Life

James A. Lake [abcd]; Jacqueline A. Servin [bd]; Craig W. Herbold [bd]; Ryan G. Skophammer [ad]
[a] Department of Molecular, Cellular, and Developmental Biology, University of California, Los Angeles, California, USA [b] Molecular Biology Institute, University of California, Los Angeles, California, USA [c] Department of Human Genetics, University of California, Los Angeles, California, USA [d] UCLA Astrobiology Institute, University of California, Los Angeles, California, USA

First Published on: 01 December 2008

To cite this Article Lake, James A., Servin, Jacqueline A., Herbold, Craig W. and Skophammer, Ryan G.(2008)'Evidence for a New Root of the Tree of Life',Systematic Biology,57:6,835 — 843

To link to this Article: DOI: 10.1080/10635150802555933

URL: http://dx.doi.org/10.1080/10635150802555933

Syst. Biol. 57(6):835–843, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802555933

Evidence for a New Root of the Tree of Life

JAMES A. LAKE,[1,2,3,4] JACQUELINE A. SERVIN,[2,4] CRAIG W. HERBOLD,[2,4] AND RYAN G. SKOPHAMMER[1,4]

[1]*Department of Molecular, Cellular, and Developmental Biology,* [2]*Molecular Biology Institute,* [3]*Department of Human Genetics, and* [4]*UCLA Astrobiology Institute, University of California, Los Angeles, California 90095, USA; E-mail: Lake@mbi.ucla.edu (J.A.L.)*

Abstract.—Directed indels, insertions or deletions within paralogous genes, have the potential to root the tree of life. Here we apply the top-down rooting algorithm to indels found in PyrD (dihydroorotate dehydrogenase), a key enzyme involved in the de novo biosynthesis of pyrimidines, and HisA (*P*-ribosylformimino-AICAR-P-isomerase), an essential enzyme in the histidine biosynthesis pathway. Through the comparison of each indel with its two paralogous outgroups, we exclude the root of the tree of life from the clade that encompasses the Actinobacteria, the double-membrane prokaryotes, and their last common ancestor. In combination with previous indel rooting studies excluding the root from a clade consisting of the Firmicutes, the Archaea, and their last common ancestor, this provides evidence for a unique eubacterial root for the tree of life located between the actinobacterial–double-membrane clade and the archaeal-firmicute clade. Mapping the phylogenetic distributions of genes involved in peptidoglycan and lipid synthesis onto this rooted tree parsimoniously implies that the cenancestral prokaryotic population consisted of organisms enclosed by a single, ester-linked lipid membrane, covered by a peptidoglycan layer. [Actinobacteria; double-membrane; Firmicutes; indels; prokaryotes; root; tree of life.]

A rooted tree allows one to parsimoniously integrate past geological, paleontological, and climatological events into a comprehensive picture of organismal and genomic evolution and thereby obtain a better understanding of the evolution of life on Earth. Dayhoff and Schwartz (1980) first demonstrated that duplicated genes could be used to root the tree of life by reconstructing phylogenetic trees from ortholog/paralog gene sets of ubiquitous sequences. This led the way for Gogarten and Iwabe to determine the first universal rooted tree (Gogarten et al., 1989; Iwabe, 1989). Today, however, there is a growing awareness that the traditional root might be an artifact of phylogenetic reconstruction resulting from long-branch attraction and frequent gene transfers (Felsenstein, 1978; Philippe and Forterre, 1999; Doolittle, 1999; Esser et al., 2004; Zhaxybayeva et al., 2005). Thus there is increasing interest in alternative methods for rooting the tree of life.

ANALYSES

Taxon Selection

Directed indels, insertions or deletions in paralogous genes, can persist over long time scales without changing length or position and thus have practical applications for rooting. To date, they have excluded the root, or cenancestor (Fitch and Upper, 1987), from the eukaryotes, from the Archaea, from the double-membrane prokaryotes, from the clade consisting of the Firmicutes and the Archaea, and from the Actinobacteria (Rivera and Lake, 1992; Skophammer et al., 2006, 2007; Lake et al., 2007; Servin et al., 2008). Through the analysis of indels contained within the enzymes PyrD, HisA, and HisF, we localize the root to an internal branch between a clade consisting of the Firmicutes plus the Archaea and a clade consisting of the Actinobacteria plus the double-membrane prokaryotes.

Taxon selection is an important aspect of indel analyses because it can be used to help reduce the effects of inter-group horizontal/lateral gene transfers. Several factors, in addition to phylogeny, are known to affect the frequency of gene transfers. These include genome G/C composition, carbon utilization, and oxygen tolerance (Jain, 2003). Thus, the four natural, phylogenetically well-separated taxa analyzed here, the Archaea, the Firmicutes, the Actinobacteria, and the double-membrane (gram-negative) prokaryotes are chosen, in part, to maximize differences between them and thereby minimize gene transfers that have occurred between them (Skophammer et al., 2006, 2007; Lake et al., 2007). Each of these taxa are relatively homogeneous, and together they contain all known prokaryotic life (Boone and Castenholz, 2001).

The Actinobacteria are characterized by high guanine-cytosine (GC) genomic compositions, which limit their gene exchanges with other groups (Jain, 2003). They are morphologically diverse and contain many human pathogens, most notably those that cause leprosy and tuberculosis. Likewise, the low GC compositions of the Firmicutes, which includes the Clostridia, the Bacilli, and other groups, are presumed to reduce gene exchange between them and other prokaryotic taxa, although clostridial exceptions to the low GC rule exist (Ueda et al., 2004). A characteristic of firmicutes is that many can form endospores during times of stress.

The double-membrane prokaryotes are the most speciose of the four groups. This taxon encompasses prokaryotes surrounded by double membranes and contains almost all known photosynthetic groups, including the Cyanobacteria, the Chlorobi, the Chloroflexi, and the Proteobacteria, suggesting that it is possibly a primitively photosynthetic taxon. (The heliobacterial clostridia are the only photosynthetic prokaryotes except for the double-membrane prokaryotes.) The fourth group consists of the Archaea, which contains extreme halophiles, methanogens, hyperthermophiles, and other unique phenotypes that exchange genes with the three other prokaryotic taxa relatively infrequently.

Indel Analyses

The three genes analyzed in this study, PyrD, HisA, and HisF, are paralogously related members of the eightfold beta alpha barrel gene family, $(\beta\alpha)_8$ (Lang et al., 2000; Hansen et al., 2004). The beta alpha barrel motif is a conserved protein fold consisting of eight alpha helices and eight parallel beta strands that alternate along the peptide backbone. Members of this family perform essential functions in metabolic processes, including nucleotide biosynthesis, amino acid biosynthesis, lipid metabolism, protein synthesis, and glycolysis. Dihydroorotate dehydrogenase, PyrD, is a central enzyme involved in the de novo biosynthesis of pyrimidines. It catalyzes the conversion of dihydroorotate to orotate, a substrate for uridylate synthesis. The histidine biosynthetic proteins, HisA and HisF, catalyze successive reactions in histidine synthesis. HisA converts *N*′-[(5′-phosphoribosyl)-formimino]-5-aminoimidazol-4-carboxamid ribonucleotide (ProFAR) into the 5′-phosphoribulosyl isomer (PRFAR). HisF catalyzes the condensation of PRFAR and ammonia and cleaves the condensation product into 5-aminoimidazole-4-carboxamide ribotide and imidazoleglycerol phosphate.

PyrD contains an indel that separates the clan (Wilkinson et al., 2007) consisting of the Firmicutes and the Archaea from the clan consisting of the Actinobacteria plus the double-membrane prokaryotes (Gupta, 1998). This indel, located between alpha helix 6 and beta strand 7 in the PyrD structure (Hansen et al., 2004), was previously used with another paralogous outgroup, HemE (Skophammer et al., 2007), to exclude the root from within the clan consisting of the Firmicutes and the Archaea.

HisA also contains an indel that separates the Firmicutes and the Archaea from the Actinobacteria and the double-membrane prokaryotes. This previously unknown indel starts just downstream from the PyrD indel and includes beta sheet 7, helix 8, and beta sheet 8. The indel itself is located between alpha helix 7 and beta strand 8 in the HisA structure.

Here we use the PyrD and the HisA indels in combination with two new outgroups, HisA and HisF for PyrD, and HisF and PyrD for HisA, in order to exclude additional roots. These new outgroups were identified through BLAST-based distance analyses that allow one to estimate the statistical significance of potentially paralogous gene sets (see Appendix 2 for details). BLAST analyses employ expect, E, values to determine how frequently alignments are expected by chance (given the size of the search space, the normalized bit score, and a minor constant). They work well for identifying orthologous genes, but they are relatively insensitive to paralogous relationships, especially when the paralogous domains are short. When searching for gene orthologs one would like to know whether a particular subject gene sequence is related by chance to the query sequence. But when searching for paralogous gene sets, one would like to know whether the last common ancestor of the gene set orthologous to the query is related by chance to the last common ancestor of the query gene set. Thus, BLAST E values are not directly relevant to deciding whether two gene sets are paralogous. In Appendix 2 we show how to calculate the BLAST E values between the last common ancestors of the ortholog and paralog gene sets using the modes of the distributions of BLAST scores from each of these sets.

Based on these comparisons we determine that the HisF/PyrD paralogous pair will occur by chance with probability $P < 3 \times 10^{-5}$ (P equals E, for E values $< 10^{-4}$), that the HisA/PyrD paralogous/pair will occur by chance with probability $P < 6 \times 10^{-8}$, and that the HisA/HisF paralogous/pair will occur by chance with probability $P < 10^{-8}$. These are consistent with the close structural similarities observed between the HisA, HisF, and PyrD protein families. Thus, the choice of outgroups is strongly supported.

The approximately 1100 representative PyrD, HisA, and HisF indel-containing sequences used in both analyses are summarized in the alignments shown in Table 1. PyrD contains a two–amino acid insertion in the double-membrane and the actinobacterial sequences, whereas firmicute and archaeal sequences lack this insertion. Similarly, HisA contains a three–amino acid insertion that has the same taxonomic distribution as the PyrD indel. Thus, the distributions of PyrD and HisA indels within double-membrane prokaryotes (D), Actinobacteria (A), Firmicutes (F), and Archaea (R) are coded as (+, +, −, −), where "+" represents insert present, and "−" represents insert absent. The outgroup sequences for the PyrD indel and for the HisA indel lack the insert and are coded as (−, −, −, −) for taxa (D, A, F, R).

The analyses of the PyrD and HisA indels proceed as shown in Figure 1. (The same arguments hold for both indels because the relevant character state patterns are identical in both.) Nine rootings of the most parsimonious unrooted tree, ((D,A), (F,R)) in Newick notation, are shown in Figure 1. Nine roots are possible, rather than the five roots encountered when rooting four taxon trees, because the four groups D, A, F, and R, are clans and not individual sequences. Thus, each leaf must be represented as two separate regions, a crown group (within the taxon) and a stem group (leading to the taxon). Accordingly, roots within crown groups (roots 1, 2, 8, and 9) are shown as two lines, whereas for all other rootings the crown groups are shown as single lines. The proximal portions of the leaves correspond to roots 3, 4, 6, and 7, and the internal branch corresponds to root 5. As shown by large X's in Figure 1, the four least parsimonious rooted trees, roots 1, 2, 3, and 4, require two changes each, thereby excluding the root from the clade consisting of the double-membrane prokaryotes, the Actinobacteria, and their last common ancestor, the (A,D) clade. As shown in Appendix 1, the PyrD indel used in combination with the combined HisA and HisF outgroup provides statistically significant support ($P < .0452$) for excluding the root of the tree of life from clade (A,D). The HisA indel also provides statistically significant support

TABLE 1. A summary of representative PyrD indel– and HisA indel– (insertions or deletions) containing sequences, respectively, showing the distribution of these indels in PyrD, HisA, and HisF genes. The PyrD indel alignments (shown on the left) summarize 1050 sequences, and the HisA indel alignments (shown on the right) summarize 1177 sequences. Sequences are shown for the double-membrane prokaryotes (D), the Actinobacteria (A), the Firmicutes (F), and the Archaea (R). Multiple sequences are shown for firmicutes (Clostridia and Bacilli) and for Archaea (Halobacteria, Thermoplasmata, and Thermoprotei) to illustrate the diversity of these indel containing sequences. PyrD sequences illustrating the PyrD indel (left) and the HisA indel (right) correspond to *Escherichea coli* amino acid positions 271 to 299 and 296 to 323, respectively. HisA sequences illustrating the PyrD indel (left) and the HisA indel (right), correspond to *E. coli* amino acid positions 198 to 224 and 221 to 251, respectively. HisF sequences illustrating the PyrD indel (left) and the HisA indel (right) correspond to *E. coli* amino acid positions 203 to 229 and 226 to 253, respectively. The region labeled "PyrD Indel" spans helix 6, beta strand 7, and helix 8. For reference, beta strand 7 starts at PIIG in sequences from all three proteins. The region labeled "HisA Indel" is adjacent to the PyrD indel; it spans beta sheet 7, helix 8 and beta strand 8; and starts two amino acids downstream of the PIIG sequence. Complete sequence alignments are presented in the Online Supplementary Material, Sections S4 and S5, available online at www.systematicbiology.org.

	PyrD Indel	HisA Indel
PyrD		
D-Proteobacteria	LQLKSTEIIRRLSLELNGRLPIIGVGGID	GGIDSVIAAREKIAA---GASLVQIYSGFIF
D-Proteobacteria	VREGSTRVIRALCGLLGDAVPIIGVGGIL	GGILAGEHAREKIDA---GAQLVQLYTGLIY
D-Cyanobacteria	LRSRSTEVIRLLHRTTQGQLPIIGVGGIF	GGIFSAEDAWQKIVA---GASLVQVYTGWVY
D-Deinococcus	LTARSTELVRAAYRLTRGRMPIVGVGGIF	GGVFSAEDAYAKLLA---GADLVEVYSALIY
D-Other DM Proks	ILPIAVRMIYQVYEKFGDRIPIIGVGGIT	GGIASFSDAMEFLLV---GASAIQIGTMNFV
A-Actinobacteria	LKARSLEVLSRLYARVGDRITLVGVGGIE	GGIENAEDAWQRILA---GATLVQGYSAFIY
F-Bacilli	IKPVAIRMVHEVSQAV--NIPIIGMGGIE	GGIETAEDVIEFFYA---GASAVAVGTANFI
F-Mollicutes	IKPVAIRMIYQVSQAV--NIPIIGMGGIS	GGISNVQDVIDFISA---GASAVAIGTANFI
F-Clostridia	VKPIALRMVHEVAKTV--DIPVIGLGGIS	GGISTAEDAIEFMMA---GASAIQIGTINFV
R-Halobacteria	IRERATEQVRFVAERT--DQPVVGVGGVA	GGVATAEDAYEKIRA---GASVVQLYTALVY
R-Thermoplasmata	IKPVGIRYVYEVKKET--GKEIIGVGGIS	GGISNYKDAIEYIMA---GASAVQIGTALYK
R-Thermoprotei	LYPIALRIIKDVYEEY--GVDIIGVGGVY	GGVYDWTDVIGMLAA---GAKLVGLGTVLIE
HisA		
D-Proteobacteria	MQGCNVPFTKALAEAT--SIPVIASGGIH	GGIHNLGDIKALLDAKAPGIIGAITGRAIYE
D-Proteobacteria	LQGINIDATVKLAQSV--SIPVIASGGLS	GGLSSLKDIDHLCAVESEGVEGVICGRAIYS
D-Cyanobacteria	LAGPNLAALRSMADAS--TVPVIASGGVG	GGVGCMADLIALLALEPHGVTGVIVGRALYD
D-Deinococcus	LRGLDRDLMRQVRGLW--HGELIVGGGVA	GGVADTNDVR-LLAEE--GIEGAIVGRAIYE
D-Other DM Proks	LEGVDVEPYKEIKKHV--KKPVIASGGVS	GGATTSDDLHKLRSLEKYGVDSVIIGKALYE
A-Actinobacteria	LGGPNLDLLAGVADRT--DAPVIASGGVS	GGVSSLDDLRAIATLTHRGVEGAIVGKALYA
F-Bacilli	LAGPNVEQLELLQKNV--ATRLIASGGVA	GGVASIQDVKKLNDM---NIYGVIIGKALYE
F-Clostridia	LKGPNLQAMKEMADSV--SMDVIASGGVS	GGVSRLKDIIDLKQT---GVSGVIVGKAIYT
R-Halobacteria	LDGVRTDPVRRLVDSV--DIPVIASGGVA	GGVATINDVLALRSA---GAAAVVVGSALYE
R-Thermoplasmata	NSDGTGHISRIEKFWD--EGYFMYAGGVN	GGVNSIDDLKKLENM---GFNGAIIGKALYN
R-Thermoprotei	TKGGIDNNVVEYVKSV--KKIKEYAGGIG	GGVSSDSDITFLKNV---GFDYIIVGMAFYL
HisF		
D-Proteobacteria	KNGFDLGVTRAISDAL--GIPVIASGGVG	GGVGNLQHLADGILE---GHASAVLAASIFH
D-Proteobacteria	KSGFDLELTRAVSDAV--PVPVIASGGVG	GGVGNLQHLADGIKL---GHADAVLAASIFH
D-Cyanobacteria	QAGYDLELTRAVAQAV--PVPVIASGGAG	GGAGCLDHIAAALDQ---GPQGGQASAALLA
D-Deinococcus	RAGFDLEATRAVAREV--DLPVIASGGAG	GGAGKVQDFYDVLTA---GEADAALAASVFH
D-Other DM Proks	KDGYDIELNRAISEAV--NIPVIASGGAG	GGAGKKEHFYEVFSK---TKVEAALAASVFH
A-Actinobacteria	KAGFDLALLRAVRAAV--TVPVIASGGAG	GGAGAVEHFAPAVAA---G-ADAVLAASVFH
F-Bacilli	KNGYDLRLTEEISKSV--SVPVIASGGCG	GGCGHADHIIEVFQK---TAVDAALAASIFH
F-Clostridia	KDGYDIELTRTVSENV--KIPVIASGGAG	GGAGKMEHFKDALVD---GKADAVLAASLFH
R-Halobacteria	KDGYDIPLMKAVCDTV--STPVIASSGCG	SGCGSPEDMEEVFVD---AGADAGLAASIFH
R-Thermoplasmata	KKGFDTDLIRKITGSV--NIPVIASGGAG	GGAGSPEDFLGVFQA---G-ADGALAASIFH
R-Thermoprotei	RLGYDLELTRKIVDSV--NIPVIASGGAG	GGIGSLDDLLKLKSF---GFDYSIIGMSFYN

($P < .0282$) for excluding the root from this region. Together the PyrD- and the HisA-directed indels, assuming independence of sites, provide strong support ($P < 2 \times 10^{-3}$) for excluding the root from the (A,D) clade.

A potential criticism of this work is that the PyrD indel is riddled with indel homoplasies as a result of indels being exchanged between groups extremely frequently. To test this possibility, we compared the observed patterns of indel distribution with the corresponding patterns observed in a 250-taxon gene tree reconstructed from the top BLAST hits to an actinobacterial query (see Online Supplementary Material Section S2, pages 3 to 5 at www.systematicbiology.org). By comparing the concordance between the trees and the indel character states, we were able to determine whether the indels were moving in and out of the clades and thereby estimate the extent of indel homoplasy. Because the actinobacterial and the double-membrane inserts are the source of the rooting information, we looked for any Actinobacteria present within the double-membrane prokaryotes and any double-membrane prokaryotes within the Actinobacteria. There were no actinobacterial sequences present within the double-membrane prokaryotes, and no double-membrane prokaryotes were found within the Actinobacteria. Furthermore, all double-membrane taxa are exclusively contained within their separate clans, except for the Betaproteobacteria, which are present in the expected location in 16S rRNA trees, as a subclan of the Gammaproteobacteria (Ludwig and Klenk, 2001). Thus, there is a high degree of concordance, 99.2%,

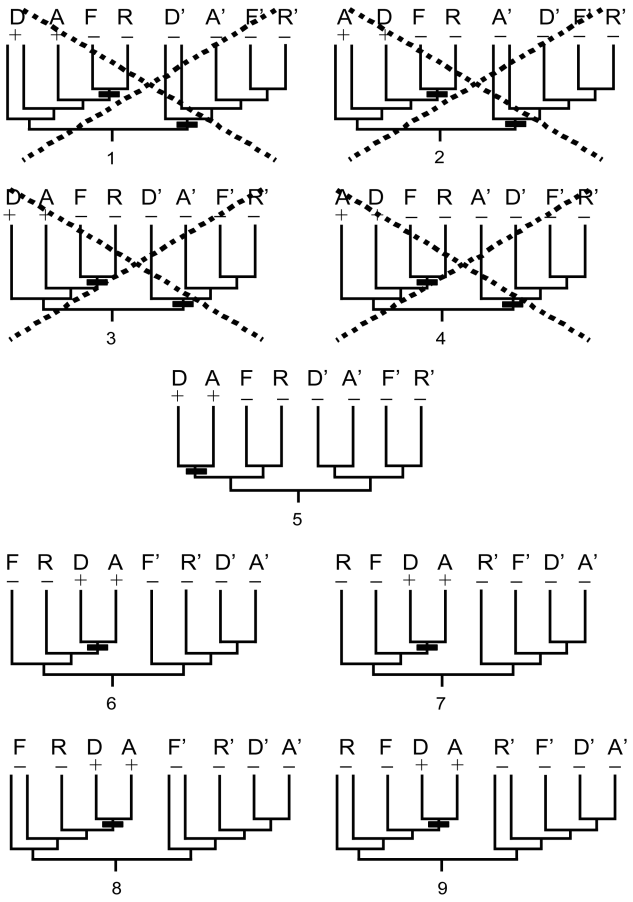

FIGURE 1. A top-down rooting analysis of the excluded roots for the PyrD and the HisA indel (insertion or deletions) sets. The following taxa are analyzed: the double-membrane prokaryotes (D or D′), the Actinobacteria (A or A′), the Firmicutes (F or F′), and the Archaea (R or R′). The character states corresponding to taxa (D, A, F, R) for gene PyrD (or HisA) are (+, +, −, −) and the character states corresponding to taxa (D′, A′, F′, R′) for outgroup genes are (−, −, −, −). Solid bars represent indel character state changes. The roots are numbered 1 to 9 as described in the text. Roots 5 to 9 are most parsimonious and correspond to one change shown by a single solid bar. Roots 1 to 4 are least parsimonious, indicated by large X's across the trees, and correspond to two changes as shown by two solid bars. In some cases alternative, equally parsimonious, locations for character state changes exist (not shown).

between the tree derived from the sequences upstream from the PyrD indel and the observed indel distributions. The observation that the Actinobacteria and the double-membrane prokaryotes are not intermixed in the PyrD tree, as would be expected if extensive gene transfer from either group had been the source of this indel, provides additional confidence in this indel. It is interesting that these analyses provide little evidence of homoplasy despite the potential for long-branch attraction and alignment artifacts (Lake, 1991b). But given that the alignments are obtained from a single query sequence, the resulting star alignments are less likely to be affected by alignment artifacts (Lake, 1991a). Furthermore, restricting the analysis to the 249 highest scoring sequences may reduce long-branch attraction effects. Similar results were also found for the HisA indel, not shown, indicating that indel homoplasy is not a significant concern for either of the indels used in this study.

DISCUSSION

These results, in combination with previous studies (Rivera and Lake, 2004; Skophammer et al., 2006, 2007; Lake et al., 2007; Servin et al., 2008) are summarized in Figure 2, top. They exclude the root from all regions of the tree except for the unique eubacterial root located on the central branch between the actinobacterial–double-membrane clade and the Firmicute-archaeal clade. Recent genomic studies (Rivera and Lake, 2004) based on phylogenetic analyses using gene presences and absences as character states indicate that the eukaryotic genome resulted from a fusion of two diverse prokaryotic genomes as indicated in Figure 2, bottom. Thus, at the deepest levels linking prokaryotes and eukaryotes, the tree of life appears to be a ring of life, with one fusion partner branching from within the double-membrane, photosynthetic clade and the other related to the archaeal prokaryotes. The location of this new root, and the regions excluded by various indel studies, are shown on the ring of life in Figure 2, bottom.

Genome fusions may have important consequences for mapping the root of life, because rings can create alternative pathways for gene flow and thereby introduce new sites for roots. For example, in the ring in Figure 2 there are two pathways for genes to flow into the eukaryotes, K, whereas most trees would connect them only to the Archaea. As a result of the ring, genes may enter eukaryotes either from the Archaea, R, or from the double-membrane prokaryotes, D. The connection between the double-membrane prokaryotes and the eukaryotes is not present in the traditional tree of life but is present in the ring of life. In this case, the new site created for a root is not viable because previous studies of the Hsp70 indel eliminated this possibility (Gupta, 1999; Lake et al., 2007). However, it is theoretically possible that if additional rings were to be found in the "tree/ring of life," these might create new sites for the root that would also need to be eliminated.

It will also be important to search for additional indels to further test this new root. In the recent past, new methods for indel analysis have increased the number of directed indels that are suitable for rooting. Initial studies excluded the root from the clan consisting of the archaeal eocytes, the eukaryotes, and their last common ancestor (Rivera and Lake, 1992); from portions of the eukaryotes (Baldauf et al., 1996); and from within the double-membrane (gram-negative) prokaryotes (Gupta et al., 1994). But early examples of high-quality indels were quickly exhausted because it was widely thought that indel-containing genes must be ubiquitous to be useful (Philippe and Forterre, 1999). The introduction of an algorithm for using non-ubiquitous genes for rooting (Lake et al., 2007) increased the number of useable indels for rooting and allowed additional roots to be excluded

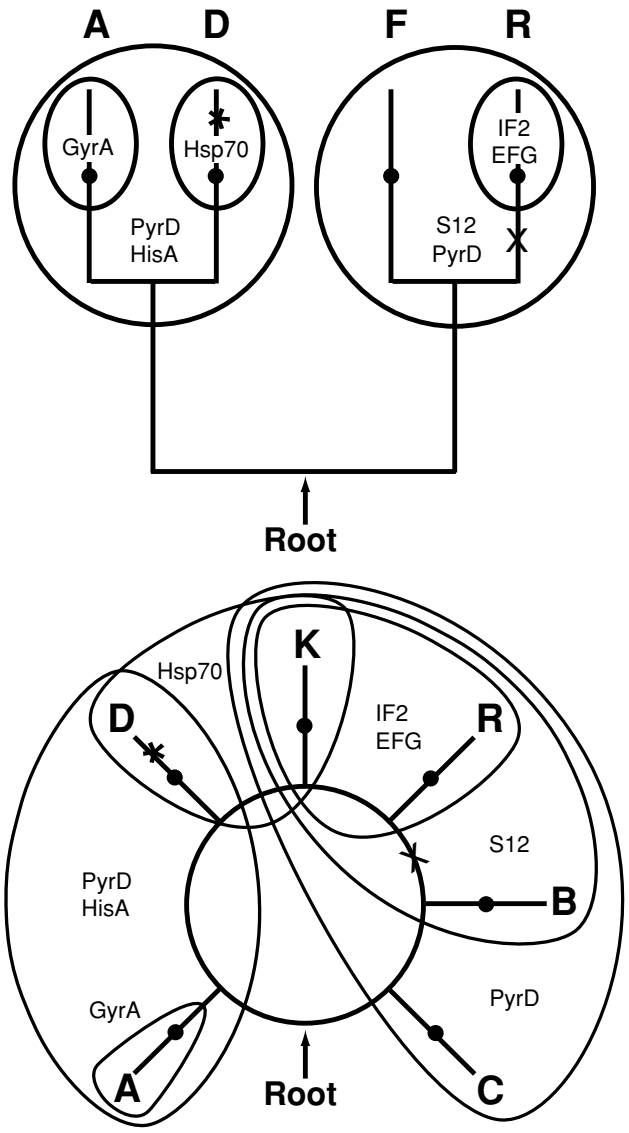

FIGURE 2. A summary of the possible locations for the root of the prokaryotic tree of life (top) and for the root of the ring of life (bottom). The relevant four taxa representing known prokaryotic diversity are the double-membrane eubacteria (D), the eubacterial Firmicutes (F), the eubacterial Actinobacteria (A), and the Archaea (R). The eukaryotes (K) are present in the ring of life (bottom), and the Bacilli (B) and the Clostridia (C) form a paraphyletic grouping within the ring. The regions from which the root is excluded are circled and labeled with the name(s) of the relevant indel(s) that exclude(s) them. The dots present on the distal portions of the leaves represent the last common ancestors of each crown group. For reference, the root based on ground-breaking analyses of anciently duplicated gene paralogs (Gogarten et al., 1989; Iwabe, 1989), marked by an "X," is located between the archaea and the eubacteria, and the root based on the analysis of evolutionary intermediates (transition analyses; Cavalier-Smith, 2006), marked by an "*," is within the double-membrane prokaryotes.

(Skophammer et al., 2007). Now it seems likely that even more indels will become useable if new statistical methods can be developed to allow one to use indels having more homoplasy than those analyzed here. Multiple, imperfect indels may play an increasingly important role in rooting studies.

A rooted tree of life provides a framework for synthesizing paleontological, climatic, geochemical, and biological information about the evolution of life on Earth. The rooted reference tree derived here allows us to infer a few simple facts about the cenancestor. We parsimoniously infer from the distributions of genes and protein families related to the synthesis of ester-linked lipids that members of the cenancestral population were parsimoniously covered by ester-linked lipid membranes. Furthermore, this root reinforces our earlier conclusion that that the unique archaeal lipids are derived from eubacterial genes (Skophammer et al., 2007). Specifically, our tree predicts that the population immediately ancestral to the Bacilli and the Archaea contained a nearly complete archaeal-like mevalonate synthesis pathway (Boucher et al., 2004) including geranylgeranylglyceryl synthase, one of the two enzymes necessary for producing the unique archaeal lipid backbone stereochemistry. Similarly, from the distribution of genes related to peptidoglycan synthesis exclusively in the Actinobacteria, the Firmicutes, and the double-membrane prokaryotes, we infer that the cenancestral population was enclosed by a peptidoglycan layer.

Finally, this root implies that the cenancestral population is not hyperthermophilic because the double-membrane prokaryotes, the Actinobacteria, and the Firmicutes are not primitively hyperthermophilic. A moderately thermophilic cenancestor cannot be excluded for this root because two thermophilic clostridia occupy positions adjacent to the root in concatenated protein sequence trees (Ohno et al., 2000; Wu et al., 2005). More definitive estimations of the growth temperature of the cenancestor will require additional knowledge of the tree topology, perhaps from other sources. These results, together with other evidence and arguments for mesophilic origins (Miller and Lazcano, 1995; Galtier et al., 1999; Philippe and Forterre, 1999), however, make lower temperature, energy-rich, hydrothermal sites such as the Lost City field (Russell and Martin, 2004; Kelly et al., 2005) increasingly attractive for the cenancestral evolution of life. We hope that this new root will form a basis for synthesizing genomic- and Earth science–based information on the evolution of life.

ACKNOWLEDGMENTS

This work was supported by grants from NSF and the UCLA NASA Astrobiology Institute to J.A.L. R.G.S., J.A.S., and C.W.H. were supported by an IGERT training grant from NSF, a Cell and Molecular Biology Training Grant from NIH, and a Genomic Interpretation and Analysis Training Grant from NIH, respectively.

REFERENCES

Baldauf, S. L., J. D. Palmer, and W. F. Doolittle. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA 93:7749–7754.

Boone, D., and R. W. Castenholz. 2001. The *Archaea* and the deep branching and phototrophic bacteria, 2nd edition. Springer, New York, Berlin, Heidelberg.

Boucher, Y., M. Kamekura, and W. Doolittle. 2004. Origins and evolution of isoprenoid lipid biosynthesis in archaea. Mol. Microbiol. 52:515–527.
Cavalier-Smith, T. 2006. Rooting the tree of life by transition analyses. Biol. Direct 1:1–135.
Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. Science 284:2124–2128.
Esser, C., N. Ahmadinejad, C. Wiegand, C. Rotte, F. Sebastiani, G. Gelius-Dietrich, K. Henze, E. Kretschmann, E. Richly, D. Leister, D. Bryant, M. A. Steel, P. J. Lockhart, D. Penny, and W. Martin. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol. 21:1643–1660.
Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.
Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.
Fitch, W. M., and K. Upper. 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harbor Symp. Quant. Biol. 52:759–767.
Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221.
Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, and M. Yoshida. 1989. Evolution of the vacuolar H^+-ATPase—Implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA 86:6661–6665.
Gupta, R. S. 1998. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. 62:1435–1491.
Gupta, R. S. 1999. Hsp70 sequences and the phylogeny of prokaryotes. Mol. Microbiol. 31:1109–1110.
Gupta, R. S., K. Aitken, M. Falah, and B. Singh. 1994. Cloning of *Giardia-Lamblia* heat-shock protein Hsp70 homologs—Implications regarding origin of eukaryotic cells and of endoplasmic reticulum. Proc. Natl. Acad. Sci. USA 91:2895–2899.
Hansen, M., J. Le Nours, E. Johansson, T. Antal, A. Ullrich, M. Loffler, and S. Larsen. 2004. Inhibitor binding in a class 2 dihydroorotate dehydrogenase causes variations in the membrane-associated N-terminal domain. Protein Sci. 13:1031–1042.
Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA 86:9355–9359.
Jain, R., M. C. Rivera, J. E. Moore, and J. A. Lake. 2003. Horizontal gene transfer accelerates genome innovation and evolution. Mol. Biol. Evol. 20:1598–1602.
Kelly, D., J. Karson, G. Fruh-Green, D. Yoerger, T. Shank, D. Butterfield, J. Hayes, M. Schrenk, E. Olson, G. Proskurowski, M. Jakuba, A. Bradley, B. Larson, K. Ludwig, D. Glickson, K. Buckman, A. Bradley, W. Brazelton, K. Roe, M. Elend, A. Delacour, S. Bernasconi, M. Lilley, J. Baross, R. Summons, and S. Sylva. 2005. A serpentinite-hosted ecosystem: the Lost City hydrothermal field. Science 307:1428–1434.
Korf, I., M. Yandell, and J. Bedell. 2003. BLAST, 1st edition. O'Reilly, Sebastapol, California.
Lake, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. Mol. Biol. Evol. 4:167–191.
Lake, J. A. 1991a. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. 8:378–385.
Lake, J. A. 1991b. Tracing origins with molecular sequences: Metazoan and eukaryotic beginnings. Trends Biochem. Sci. 16:46–50.
Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences—Paralinear distances. Proc. Natl. Acad. Sci. USA 91:1455–1459.
Lake, J. A., C. W. Herbold, M. C. Rivera, J. A. Servin, and R. G. Skophammer. 2007. Rooting the tree of life using non-ubiquitous genes. Mol. Biol. Evol. 24:130–136.
Lang, D., R. Thoma, M. Henn-Sax, R. Sterner, and M. Wilmanns. 2000. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. Science 289:1546–1550.
Ludwig, W., and H. P. Klenk. 2001. Overview: A phylogenetic backbone and taxonomic framework for procaryotic systematics. Pages 49–65 *in* Systematic bacteriology (D. Boone and R. W. Castenholz, eds.). Springer, New York, Berlin, Heidelberg.
Miller, S. L., and A. Lazcano. 1995. The origin of life—Did it occur at high temperatures? J. Mol. Evol. 41:689–692.
Ohno, M., H. Shiratori, M.-J. Park, Y. Saitoh, Y. Kumon, N. Yamashita, A. Hirata, H. Nishida, K. Ueda, and T. Beppu. 2000. *Symbiobacterium thermophilum* gen. nov., sp. nov., a symbiotic thermophile that depends on co-culture with a *Bacillus* strain for growth. Int. J. Syst. Evol. Microbiol. 50:1829–1832.
Philippe, H., and P. Forterre. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49:509–523.
Rivera, M. C., and J. A. Lake. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76.
Rivera, M. C., and J. A. Lake. 2004. The ring of life: Evidence for a genome fusion origin of eukaryotes. Nature 431:152–155.
Russell, M. J., and W. Martin. 2004. The rocky roots of the acetyl-CoA pathway. Trends Biochem. Sci. 29:358–363.
Servin, J. A., C. W. Herbold, R. G. Skophammer, and J. A. Lake. 2008. Evidence excluding the root of the tree of life from the Actinobacteria. Mol. Biol. Evol. 25:1–4.
Skophammer, R. G., C. W. Herbold, M. Rivera, J. A. Servin, and J. A. Lake. 2006. Evidence that the root of the tree of life is not within the Archaea. Mol. Biol. Evol. 23:1648–1651.
Skophammer, R. G., J. A. Servin, C. W. Herbold, and J. A. Lake. 2007. Evidence for a gram positive, eubacterial root of the tree of life. Mol. Biol. Evol. 24:1761–1768.
Ueda, K., A. Yamashita, J. Ishikawa, M. Shimada, T. Watsuji, K. Morimura, H. Ikeda, M. Hattori, and T. Beppu. 2004. Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. Nucleic Acids Res. 32:4937–4944.
Wilkinson, M., J. O. McInerney, R. P. Hirt, P. G. Foster, and T. M. Embley. 2007. Of clades and clans: Terms for phylogenetic relationships in unrooted trees. Trends Ecol. Evol. 22:114–115.
Wu, M., Q. Ren, A. S. Durkin, S. C. Daugherty, L. M. Brinkac, R. J. Dodson, R. Madupu, S. A. Sullivan, J. F. Kolonay, W. C. Nelson, L. J. Tallon, K. M. Jones, L. E. Ulrich, J. M. Gonzalez, I. B. Zhulin, F. T. Robb, and J. A. Eisen. 2005. Life in hot carbon monoxide: The complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. PLoS Genet. 1:563–574.
Zhaxybayeva, O., P. Lapierre, and J. P. Gogarten. 2005. Ancient gene duplications and the root(s) of the tree of life. Protoplasma 227:53–64.

First submitted 21 December 2007; reviews returned 29 February 2008; final acceptance 26 August 2008
Associate Editor: Frank Anderson

APPENDIX 1: SIGNIFICANCE OF THE PYRD AND HISA INDELS

Computing the probability that an indel excludes the root of life from a clan (Wilkinson et al., 2007) is fundamentally like calculating the probability of a four-taxon phylogenetic tree. But it is complicated by having to work with trees based on a single character and by the difficulties associated with computing the relevant statistics under a parsimony model (Felsenstein, 2004). Given the large numbers of sequences that are available for analysis, even a single indel may provide statistically significant evidence for excluding the root from a clan.

Indel rooting is based on testing phylogenetic models. To appreciate the role of the model, consider how the PyrD indel excludes a root from the clans shown in the example in Figure A1.1. As illustrated in Figure A1.1, the four taxa analyzed in this study are $PyrD_D$, $PyrD_A$, $PyrD_{FR}$, and HisA/F. (The first three taxa refer to PyrD indels present in the double-membrane prokaryotes, the Actinobacteria, and the clan consisting of the Firmicutes, the Archaea, and their last common ancestor; and the last refers to the HisA and HisF indels present in all prokaryotes.) We assume that the character state "−," corresponding to insert missing, is the ancestral state because the HisA and HisF genes are universally present in prokaryotes and the indel is missing in 676 of 677 sequences. The rooted tree shown below corresponds to the one that excludes the root from the $PyrD_A \cap PyrD_D$ clan (A,D).

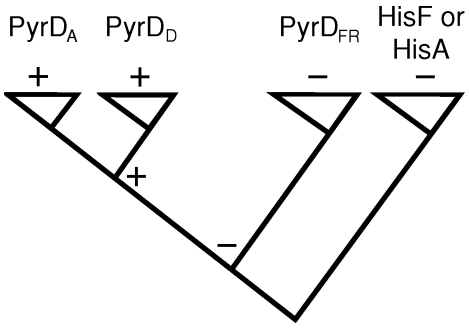

FIGURE A1.1. The principal character states present in ingroup taxa $PyrD_A$, $PyrD_D$, $PyrD_{FR}$ and outgroup taxa, HisA/F. Crown groups are indicated by triangles at the tips of the leaves, although they may possibly extend to the base of their stems. The character states shown above the crown groups parsimoniously exclude the root from the (D,A) clade. Subscripts A, D, and FR refer to sequences within the Actinobacteria, the double-membrane prokaryotes, and the clan consisting of the Firmicutes and the Archaea, respectively.

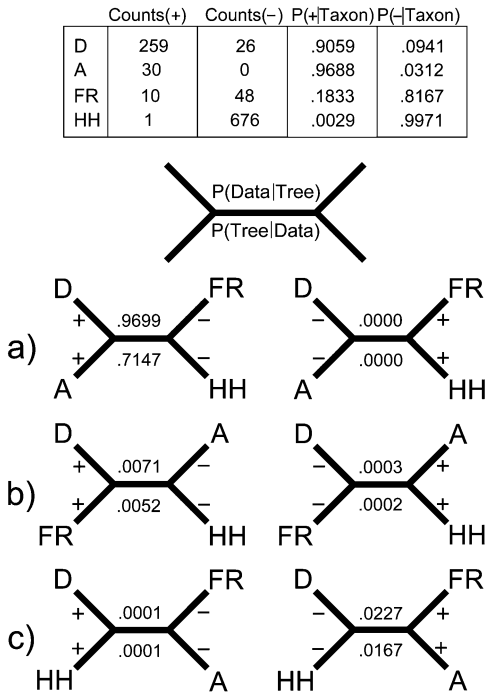

FIGURE A1.2. The data and the model for calculating the probabilities of the possible rooted trees, assuming a multinomial distribution and independent indel counts. The observed indel counts are shown in the table at the top of the figure for the four taxa analyzed in this study, $PyrD_D$, $PyrD_A$, $PyrD_{FR}$, and HisA/F. (The first three taxa refer to PyrD indels present in the double-membrane prokaryotes, the Actinobacteria, and the clan consisting of the Firmicutes, the Archaea, and their last common ancestor; and the last refers to the HisA and HisF indels present in all prokaryotes.) The six possible rooted topologies are shown at the bottom of the figure.

Under this model, six trees may exclude the root from clans. These correspond to the three tree topologies and two labelings of the internal nodes with a plus and a minus. The six possible trees, with internal nodes labeled, are shown in Figure A1.2, the observed number of counts for the PyrD indel are shown in the second and third columns of the table at the top of the figure, and the fractions of pseudocounts (counts + 1) are shown in the fourth and fifth columns. The six trees exclude the root from the (D,A), (FR,HH), (D,FR), (A,HH), (D,HH), and (FR,A) clans, respectively. As illustrated in Figure A1.2, assuming a multinomial model and that the counts in each of the four crown groups are independent, one obtains the probabilities of the data for each of the six rooted trees from the observed frequencies of +'s and −'s within the four taxa. However, the calculation needs to be corrected because it assumes that all indels in any crown group provide independent evidence of the indel character state at the base of the stem.

In order to estimate the support for excluding the root from the (A,D) clan, we calculate a chi-square, consistent with the assumption of a multinomial distribution, and suitably modify it to account for the lack of independence of the observed indel counts within the four taxa. As previously noted, the ((D,A):+,(FR,HH):−) tree excludes the root from the (A,D) clan, consisting of $PyrD_A$ and $PyrD_D$, and the remaining trees exclude the root from the five other clans, (FR,HH), (D,FR), (A,HH), (D,HH), and (FR,A). However, it makes no sense to consider the three clans (FR,HH), (A,HH), and (D,HH), because they exclude the root from the outgroup (an interesting contradiction) and because their inclusion would greatly inflate the chi-square values and thereby lower the P values. Hence, we consider only the (D,A), the (D,FR), and the (FR,A) clans, and assume that each of the three rooted trees supporting these clans are equally likely. We calculate chi-square to test the hypothesis H_0: $\mu_{(A,D)} = 1/2\ \mu_n$, where $\mu_{(A,D)}$ = the mean of the number of counts excluding the root from the (A,D) clan, and μ_n = the mean of the counts excluding the root from the (D,FR) and the (FR,A) clans. Let $\varepsilon = \mu_n/(\mu_n + \mu_{(A,D)})$, the fraction of counts excluding the root from the clans other than (A,D). Given the indel present, +, and absent, −, counts for $PyrD_A$, $PyrD_D$, $PyrD_{FR}$, and HisA (and/or HisF), a_+, d_+, r_+, h_+, and a_-, d_-, r_-, h_-, respectively, we calculate chi-square. The expected number of trees excluding the root from (A,D) under the multinomial model is $(1 - \varepsilon)\ (a_+\ d_+\ r_-\ h_-)/3$, and the expected number of trees excluding the root from the two other clans is $2\ (1 + \varepsilon)\ (a_+\ d_+\ r_-\ h_-)/3$. The observed number of trees excluding the root from (A,D) is $(a_+\ d_+\ r_-\ h_-)$, and the observed number of trees excluding the root from other clans is $\varepsilon(a_+\ d_+\ r_-\ h_-)$. For one degree of freedom we calculate chi-square, where $f = (1 - \varepsilon/2)^2/\ (1 + \varepsilon)$:

$$\chi_1^2 = 2\ f\ a_+\ d_+\ r\_h \qquad \text{(A1.1)}$$

This equation needs to be modified because it overestimates the statistical significance of an indel by assuming that all of the variants in each of the crown groups provide independent support for the character state at the base of the stem group. But in fact not even the most recent common ancestor of the crown groups are likely to extend down to the base of the stem. Thus, the sequences within $PyrD_A$, etc., are not independent and cannot in general provide multiple independent estimates of the character states at the bases of their stems. Here we describe a method to estimate the effective number of variants and substitute these estimates into Equation A1.1 in order to apply it to the PyrD/HisA/HisF- and to the HisA/HisF/PyrD-directed indels.

Our statistical approach implements the correlation analysis developed by Roderick Little (see Lake, 1987) for use with evolutionary parsimony. Because his method is based on sampling sequence variation and because the variation within an indel cannot be sampled for those sequences that lack an insert, we obtain the necessary sequence information by analyzing sequences from both margins of the indel having a combined sequence length equal to that of the insert (Lake et al., 2007; Skophammer et al., 2006). This allows one to estimate the effective number of sequences that are within the four relevant crown groups in order to utilize these in Equation A1.2, shown below.

$$\chi_1^2 = 2\ f\ E_a\ E_d\ E_r\ E_h \qquad \text{(A1.2)}$$

where E_a, E_d, E_r, and E_h are the effective numbers of sequences in each of the groups. The effective numbers are calculated from the correlation coefficients, the ρ's:

$$E_a = m/(1 + [m - 1]\rho_a);$$

$$E_d = n/(1 + [n - 1]\rho_d);$$

$$E_r = x/(1 + [x - 1]\rho_r); \quad \text{and}$$

$$E_h = y/(1 + [y - 1]\rho_h).$$

To estimate the correlation coefficients corresponding to the $PyrD_A$, $PyrD_D$, $PyrD_{FR}$, and His_{HH} variants, ρ_a, ρ_d, ρ_r, and ρ_h, respectively, let

L = the length of the indel,

m, n, x, and y = the number of sequence pairs in taxa $PyrD_A$, $PyrD_D$, $PyrD_{FR}$, and $His_{A/F}$,

D^j_m = the distance (Hamming) between the jth pair of sequences in taxon $PyrD_A$ (likewise for $PyrD_D$, $PyrD_{FR}$, and $His_{A/F}$), and

V^j_m = the variance corresponding to the D^j_m, etc.

Compute:

D^j_m = the number of positions that are different between the jth pair of sequences,

$V_m = L$, (Poisson assumption)

$D_m = \Sigma_j D^j_m/m$, (similarly for $PyrD_D$, $PyrD_{FR}$, and $His_{A/F}$)

$D^2_m = \Sigma_j (D^j_m - \underline{D}_m)^2/(m-1)$, and(similarly for $PyrD_D$, $PyrD_{FR}$, and $His_{A/F}$)

$\rho_m = 1 - D^2_m/V_m$ (similarly for $PyrD_D$, $PyrD_{FR}$, and $His_{A/F}$)

In applying these equations to the $PyrD_A$, $PyrD_D$, $PyrD_{FR}$, and $His_{A/F}$ sequences, we use the sequence positions immediately flanking the indel. Of these, approximately one half of the sequences are upstream from the indel and the other half downstream (if the length of the indel is an even number, then it is exactly one half). Because the correlation coefficients are dependent on the length of the flanking sequences being analyzed, we set the combined length of the flanking regions equal to the length of the indel.

For the PyrD insert we obtain:

$$\chi^2 = 1.88356 \times (1.2273 \times 1.2029 \times 1.1790 \times 1.1925) = 3.910, \text{ HisA outgroup, } P < .0479 \text{ (1 d.f.)},$$

$$\chi^2 = 1.88356 \times (1.2273 \times 1.2029 \times 1.1790 \times 1.2161) = 3.987, \text{ HisF outgroup, } P < .0458 \text{ (1 d.f.)},$$

and

$$\chi^2 = 1.88356 \times (1.2273 \times 1.2029 \times 1.1790 \times 1.2233) = 4.010, \text{ combined outgroup, } P < .0452 \text{ (1 d.f.)}.$$

For the HisA insert we obtain:

$$\chi^2 = 1.79823 \times (1.3613 \times 1.2611 \times 1.2463 \times 1.2549) = 4.676, \text{ HisF outgroup, } P < .0305 \text{ (1 d.f.)},$$

$$\chi^2 = 1.79823 \times (1.3613 \times 1.2611 \times 1.2463 \times 1.0939) = 4.076, \text{ PyrD outgroup, } P < .0434 \text{ (1 d.f.)},$$

and

$$\chi^2 = 1.79823 \times (1.3613 \times 1.2611 \times 1.2463 \times 1.2916) = 4.814, \text{ combined outgroup, } P < .0282 \text{ (1 d.f.)}.$$

APPENDIX 2: IDENTIFYING ORTHOLOG/PARALOG SETS USING BLAST ANALYSES

When searching for paralogous gene sets, one wants to know whether the common ancestor of the ortholog gene set is related by chance to the common ancestor of the paralog gene set. Thus, BLAST E values are not directly relevant to deciding whether two gene sets are paralogous.

A paralogous relationship is suggested, however, when BLAST analyses present a bimodal distribution of gene sets within the listing of "ortholog" hits. A tree representing an ancient ortholog/paralog relationship and the maxima in the bimodal LogBit distance distribution it generates is diagrammed in Figure A2.1. This tree differs from the standard tree because distances are measured from a reference taxon at the upper left of the tree and mapped along the line at the top of the tree.

As one moves away from the reference taxon at distance D = 0.0 and continues down the tree, successive nodes are encountered. At any node one can either move up in time toward the tips of the tree or continue moving down toward the last common ancestor of the ortholog, O, and ultimately to the branch that leads to the last common ancestor of the paralog, P. When one moves up in time, from any node, the speciation that occurs is indicated by the step-like broadening of the lines.

Assuming an approximately constant rate of speciation, increasing numbers of species will be generated as one moves toward the cenancestral, orthologous node, "O." The final round of ortholog speciation occurs when the cenancestral node of the ortholog is reached. On average, one half of the ortholog BLAST hits will come from the cenancestral node, corresponding to the maximum frequency of ortholog hits. Once beyond the cenancestral ortholog, the last node is the last common ancestor of the ortholog and paralog shown at the bottom of the tree. As one moves up the right side of the tree, the final node is the cenancestral paralog, marked by a "P." At this point the paralog branch begins to speciate, so that in contrast to the continuous generation of ortholog species, all of the paralogs will appear at approximately the same distance, D_p, from the query sequence.

Under the assumption that the rate of speciation is uniform, the largest peak for the ortholog set will appear at distance D_o, corresponding to twice the distance from the query to the cenancestral ortholog. The largest peak for the paralog set appears at distance D_p, corresponding to twice the distance from the query to the common ancestor of the ortholog and the paralog. Thus, the frequency of ortholog BLAST hits will increase gradually to a maximum value at D_o and then decrease rapidly to zero. In contrast, the frequency of paralog BLAST hits will cluster relatively closely around D_p. Variations in evolutionary rates are expected to broaden the maximal peaks, D_o and D_p, but are assumed not to move their positions.

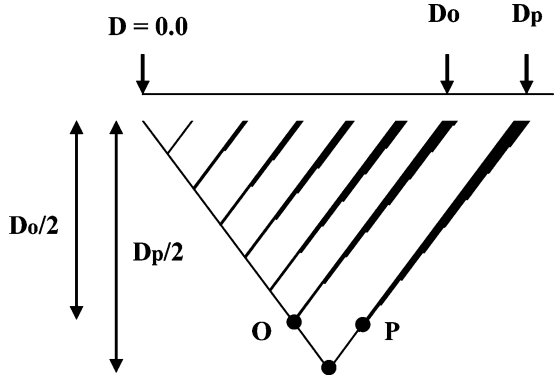

FIGURE A2.1. A tree representing an ancient ortholog/paralog relationship and the maxima in the bimodal LogBit distance distribution it generates. Speciation under a clock-like model is shown for illustration purposes. Hence, the speciation that occurs when one moves up and to the right is indicated by the step-like broadening of the lines.

Thus, if we can convert BLAST bit scores to distances, and vice versa, it should be possible to determine D_o, the distance between the cenancestral ortholog and the query, and D_p, the distance between the cenancestral paralog and the query. Subtracting these two provides the distance that separated these two genes at the time of the cenancestor, D_{po}. Converting this distance into a BLAST bit score allows us to determine the BLAST score that would have been obtained at the time of the cenancestor if the sequence of D_o were used as the query and if the sequence D_p were used as the subject. In other words, the problem of identifying paralogous genes, which BLAST analyses are not designed to do, has been transformed to the problem that BLAST is designed for, that of identifying two recently diverged, formerly orthologous sequences.

The following formula is very useful for converting between BLAST bit scores and evolutionary distances.

$$d_{qx} = -\ln(B_{qx}/B_{qq}). \quad \text{(A2.1)}$$

where d_{qx} is the distance between the query q and sequence x; B_{qx} is the BLAST bit score between the query q and sequence x; and B_{qq} is the bit score between the query and itself. The formula was derived through an analysis of a BLAST two-character state model using symmetric scoring matrices and Paralinear/LogDet distances (Lake, 1994). This model shows that when the subject is within the range of target frequencies (Korf et al., 2003) for the BLAST matrices employed—i.e., BLOSUM45 for the studies here—the resulting distances have the desired properties, including (1) the monotonically increasing property; (2) the zero property; (3) the symmetry property, $d_{qx} = d_{xq}$; and (4) the nonnegative property, $d_{qx} \geq 0$, for all q and x (see Online Supplementary Material; Skophammer et al., 2007).

Equation A2.1 has the advantage that it allows one to obtain all the required information from a single BLAST report. The BLAST bit scores used to calculate the distance between the query and the last common ancestor of the ortholog set, d_{op}, are obtained from the mode of the orthologous BLAST scores, B_{qmo}, and from the mode of the paralogous BLAST scores, B_{qmp}. The BLAST score between the query and itself, B_{qq}, is not needed to calculate the distance d_{op} but is needed to calculate the expect value corresponding to d_{op}.

The distance between the ortholog and paralog sets, d_{op}, is the distance between the paralogs and the query, d_{qmp}, minus the distance between the orthologs and the query, d_{qmo}. Thus,

$$\begin{aligned} d_{op} &= d_{qmp} - d_{qmo} \\ &= -\ln(B_{qmp}/B_{qq}) + \ln(B_{qmo}/B_{qq}) \\ &= -\ln(B_{qmp}/B_{qmo}). \end{aligned}$$

The distance between common ancestor O and common ancestor P is also given by:

$$d_{op} = -\ln(B_{op}/B_{oo}) = -\ln(B_{op}/B_{pp}),$$

where the first step follows from the definition of d_{op}, and the second step follows from the relationship, $B_{oo} = B_{pp} = B_{qq}$, assuming that the ancestral O and P sequences used as queries have similar amino acid compositions. Combining the two expressions for d_{op}, we find:

$$\ln(B_{qmp}/B_{qmo}) = \ln(B_{op}/B_{qq}).$$

From this we obtain the following useful estimate of B_{ab}:

$$B_{op} = B_{qq}(B_{qmp}/B_{qmo}) \quad \text{(A2.2)}$$

Example 1: Calculating the Expect Value That HisF and PyrD Are an Ortholog/Paralog Set

For the HisF indel query (for alignments see Online Supplementary Material, Section S1B), $B_{qq} = 76.1$, $B_{qmp} = 29.8$, and $B_{qmo} = 49.1$. From these values we compute that $B_{op} = 46.2$. From the BLAST listing, part of which is shown below,

gi |21284321| ref |NP_647409.1| cyclase-likeprotein hisF [Staph 46.5 2e=05
gi |57865752| ref |YP_189851.1| imidazoleglycerolphosphate synt 46.2 3e − 05
gi |27467194| ref |NP_763831.1| cyclase-like protein hisF [Staph 45.9 4e−05

We estimate that a bit score of 46.2 corresponds to $E = 3 \times 10^{-5}$. Thus, HisF/PyrD ortholog/paralog sets are expected to occur by chance with probability $P < 3 \times 10^{-5}$ (using the standard conversion, $P = 1 - e^{-E}$; Korf et al., 2003), and we infer that the PyrD and HisF sequences are an ortholog/paralog set.

Example 2: Calculating the Expect Value That HisA and PyrD Are an Ortholog/Paralog Set

For the HisA indel query, $B_{qq} = 107$, $B_{qmp} = 34.5$, and $B_{qmo} = 63.5$ (for alignments see Online Supplementary Material, Section S1A). From these values we compute that $B_{op} = 58.1$. From the BLAST listing, part of which is shown below,

gi |15616141| ref |NP_244446.1 | 1−(5−phosphoribosyl)−5 −[(5 − phosp 64.1 4*e* − 11
gi |56964809| ref |YP_176540.1 | 1−(5−phosphoribosyl)−5 −[(5−phosp 63.5 6e − 11
gi|89202117|ref|ZP_01180847.1| Phosphoribosylformi min o−5−*a* min 44.4 3e−05

We interpolate that a bit score of 58.1 corresponds to an expect value of $E = 6 \times 10^{-8}$, implying that HisA/PyrD ortholog/paralog sets will occur by chance for this data set with a probability $P < 6 \times 10^{-8}$ (using the standard conversion, $P = 1 - e^{-E}$; Korf et al., 2003). We infer that the PyrD and HisA sequences are an ortholog/paralog set.