Rooting the Tree of Life Using Nonubiquitous Genes

James A. Lake,†‡§ Craig W. Herbold,†§ Maria C. Rivera,*§ Jacqueline A. Servin,†§ and Ryan G. Skophammer*§*

*Department of MCD Biology, University of California; †Molecular Biology Institute, University of California; ‡Department of Human Genetics, University of California; and §UCLA NASA Astrobiology Institute, University of California

Insertion and deletion (indel)–based analyses have great potential for rooting the tree of life, but their use has been limited because they require ubiquitous sequences that have not been horizontally/laterally transferred. Very few such sequences exist. Here we describe and demonstrate a new algorithm that can use nonubiquitous sequences for rooting. This algorithm, top–down indel rooting, uses the traditional logical framework of indel rooting, but by considering gene gains and losses in addition to indel gains and losses, it is able to analyze incomplete data sets. The method is demonstrated using theoretical examples and incomplete gene sets. In particular, it is applied to the well-studied Hsp70/MreB indel, a sequence set thought to have been compromised by gene transfers from Firmicutes to archaebacteria. By sequentially assigning all observable character states, including gene absences, to the questionable archaebacterial Hsp70 and MreB sequences, we demonstrate that this gene set robustly excludes the root of the tree of life from the Gram-negative, double-membrane prokaryotes independently of the archaeal character states. There are very few ubiquitous paralog gene sets, and most of them contain compromised data. The ability of top–down rooting to use incomplete and/or compromised gene sets promises to make rooting analyses more robust and to greatly increase the number of useful indel sets.

Introduction

Indel-based rooting analyses have tremendous potential for answering some of the most fundamental questions in evolutionary biology. One only has to look at the vigorous field of eukaryotic relationships to realize the importance of indel rooting. Among high-level eukaryotic taxa, one of the most trusted rooted sister group relationships consists of animals and fungi, the Opisthokonts, a relationship based on a 12 amino acid–long insert present in protein synthesis elongation factor 1α (Baldauf and Palmer 1993). When indel relationships are strong, they can provide some of the best rooting information available.

Two significant obstacles have prevented the widespread use of indels for rooting in the past. First, very few universally present paralog gene sets are available for indel rooting, and second, many otherwise usable indel sets contain genes that may have been horizontally/laterally transferred, for example, the eukaryotic enolases (Harper and Keeling 2004) and the Hsp70 indel (Gupta 1998; Philippe et al. 1999), making them potentially untrustworthy. These difficulties could have been circumvented if it were possible to use incomplete gene sets for rooting, but it was thought that ubiquitous gene sets were required for indel-based rooting (Rivera and Lake 1992; Gupta and Singh 1994; Philippe et al. 1999) because they were used for sequence-based rooting (Gogarten, Kibak, et al. 1989; Iwabe et al. 1989; Brown and Doolittle 1995; Boucher et al. 2003; Zhaxybayeva et al. 2005). Here we show that ubiquitous genes are not necessarily required for indel rooting. Our algorithm, top–down rooting, mathematically analyzes both indel gains and losses and gene gains and losses. It treats gene losses and gains as an integral part of the evolutionary process, thereby permitting parsimony analyses even when genes are missing from 100% of a particular taxon. It frequently finds that incomplete gene sets contain useful rooting information.

We demonstrate this algorithm through examples, and then apply it to the Hsp70 heat shock protein indel. The interpretation of the Hsp70 indel has been controversial (Gupta 1999; Philippe et al. 1999) particularly because the Hsp70 gene is likely to have been transferred from the Firmicutes to the Archaea. We analyze the Hsp70/MreB paralog gene set and show that this set excludes the root of the tree of life from the double-membrane, Gram-negative prokaryotes for all possible indel, gene, and topological scenarios.

Theory

Indel rooting differs from sequence-based rooting in several ways. For example, in sequence analyses a single sequence set can root the tree of life, whereas indel rooting requires multiple indel–containing sets. In sequence analyses, phylogenetic trees are reconstructed from a pair of universal, paralogous genes, and the root is inferred from the location of the branch connecting the pair of paralogous gene trees (Dayhoff and Schwartz 1980; Gogarten, Kibak, et al. 1989; Iwabe et al. 1989; Brown and Doolittle 1995; Boucher et al. 2003; Zhaxybayeva et al. 2005). In contrast, indel rooting does not provide the location of the root but rather excludes the root from portions of the tree. Thus, multiple indel sets must be analyzed to exclude the root from progressively larger regions of the tree, until ultimately only a single root remains.

Traditional indel rooting (Rivera and Lake 1992; Baldauf and Palmer 1993; Gupta 1998) uses paralogous aligned regions from 2 ubiquitous genes, referred to as paralog 1 and paralog 2. (We will use plenary to refer to genes present in all taxa in exactly 1 copy per genome; universal to refer to genes present in all taxa, but possibly in multiple copies; and ubiquitous to refer to genes present in almost all taxa, possibly in multiple copies.) In the simplest case, paralog 1 contains both character states of the indel, and paralog 2 contains the indel in only 1 form (Skophammer et al. 2006). Whether the indel under analysis is recent (a synapomorphy) or ancient (a plesiomorphy) can only be decided by knowing whether it is present or absent in

Key words: tree of life, root, indels, cenancestor, double-membrane prokaryotes.

E-mail: lake@mbi.ucla.edu.

Mol. Biol. Evol. 24(1):130–136. 2007
doi:10.1093/molbev/msl140
Advance Access publication October 5, 2006

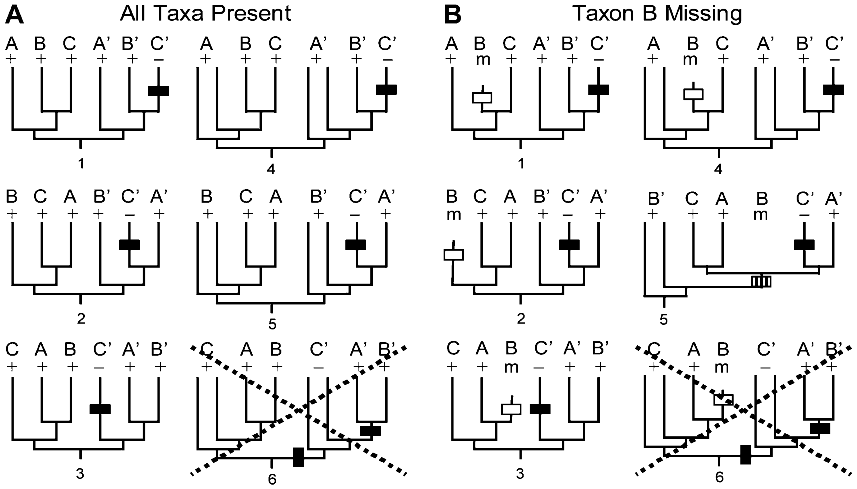

FIG. 1.—An illustration of the calculation of most parsimonious trees for the first example. The character states corresponding to taxa (A, B, C, A′, B′, C′) are (+, +, +, +, +, −) and (+, m, +, +, +, −) for figure 1*A* and *B*, respectively. Solid rectangles represent an indel character–state change, outlined rectangles represent gene deletions, and vertically striped rectangles represent gene duplications. Indel-state changes, gene deletions, and gene duplications are weighted equally. The roots are numbered 1–6, as described in the text. Roots 1–5 are most parsimonious and correspond to 1 and 2 changes in figure 1 (*A*) and 1 (*B*), respectively. Root 5 is the least parsimonious, as indicated by the large Xs across the trees, and corresponds to 2 and 3 changes in figure 1 (*A*) and (*B*), respectively. In some cases alternative, but equally parsimonious, solutions for character-state changes exist (not shown).

paralog 2. If it is absent in paralog 2, then the indel is recent, and the root is excluded from the group containing the indel, and if it is present in paralog 2, then the indel is ancient, and the root is tentatively permitted in the group containing the indel. Thus, analyses of indels present in paralogous gene pairs may be used to eliminate the root from particular regions of the tree.

A Theoretical Example

The logic of rooting differs in traditional indel-rooting analyses and in top–down indel-rooting analyses. Traditional indel-rooting analyses compare all possible rooted trees that relate ubiquitous paralogous taxa and calculate the minimum number of indel-state changes required for each root. Roots corresponding to the fewest indel changes are tentatively accepted, and roots corresponding to the most indel changes (the least parsimonious) are excluded.

Top–down indel analyses follow a similar pattern, with the differences noted in boldface. Top–down analyses compare all possible rooted trees that relate paralogous taxa, **including taxa for which data are missing,** and calculate the minimum number of indel- **and gene**-state changes. Roots corresponding to the fewest indel **and gene** changes are tentatively accepted, and roots corresponding to the most indel **and gene** changes (the least parsimonious) are excluded.

Traditional indel rooting is simpler than top–down rooting because it only examines indel-state changes between 2 states: indel present, "+," and indel absent, "−." In contrast, top–down rooting also analyzes gene-state changes. In this case, the 2 possible gene states, gene present "p" and gene missing "m," must also be considered. Thus, top–down rooting must simultaneously analyze 2 types of changes, indel and gene changes, that produce the observed character states. The observed character states are actually composite states that depend upon the indel state and the gene state. If a gene is present, then the observed composite state is "+" or "−," but if the gene is missing then the observed state can only be "m" because without the gene one cannot know whether the indel is present or absent. This is mathematically analogous to genetic epistasis because a gene presence is required in order to determine the state of an indel. As a result, only 3 states are experimentally observable: "+," "−," and "m."

Simple examples can help explain top–down rooting. In our first example, shown in figure 1*A*, the ancestral character states are ("+," "+," "+") for outgroup taxa (A, B, C), respectively, and ("+," "+," "−") for ingroup taxa (A′, B′, C′), respectively. Because each group in figure 1*A* represents a higher-level phylogenetic taxon, the leaves of the 3-taxon unrooted trees are divided into 2 separate regions, a terminal portion of the leaf representing the diversity of organisms within the clade and an interior portion consisting of the branch leading to the clade. Hence for any 3-taxon tree representing diversified clades, there are 6 possible distinct roots, rather than the 3 roots that would be present if one were simply comparing 3 individual sequences. The 3 roots, labeled 1, 2, and 3 in figure 1, are on the branches leading to groups A, B, and C, respectively, and the 3 roots, labeled 4, 5, and 6, are within groups A, B, and C, respectively. Thus, groups A, B, and C are shown as 2 lines in trees corresponding to roots 4, 5, and 6, representing the branching within their respective clades. The most parsimonious trees, see figure 1*A*, require only a single indel character–state change, shown as a black rectangle, except for root 6, which requires 2 indel character–state changes. Thus, root 6 is eliminated for character state B = "+," as shown by the large "X" in figure 1*A*.

The result in figure 1*A* could have been obtained using traditional indel rooting, but the analysis shown in figure 1*B* can only be performed using top–down rooting. In this case, even though genes from taxon B are missing, denoted by character state "m," root 6 is still eliminated. Three types of

Table 1
Parsimony Analysis of Example

Root B	+	−	m	Root OK
1	1	2	2	Y
2	1	2	2	Y
3	1	2	2	Y
4	1	2	2	Y
5	1	**3**	2	Y
6	**2**	**3**	**3**	**N**

NOTE.—Parsimony scores are calculated for roots 1–6, left column, in the 3 taxon trees discussed in the first example. The character states (+, x, +, +, +, −) correspond to taxa (A, B, C, A′, B′, C′), respectively, and the 6 possible root locations are labeled as in figure 1, where x = "+," "−," or "m." A root location is rejected, Root OK = N, only if all 3 separate analyses reject the root. Maximum scores for each column are printed in bold-faced type.

operations can cause changes in figure 1*B*. These are indel character–state changes, gene deletions (open bars), and gene duplications (vertically striped bars). The 5 most parsimonious roots, 1–5, in figure 1*B* each require 2 changes, whereas root 6 requires 3 changes. Solutions for roots 1–4, and for 6, follow those obtained when taxon B is in character state "+" but use a gene deletion to remove the branch leading to B. Root 5, however, is novel because a second copy of Gene B is never created as an ortholog/paralog gene duplication occurs just prior to the speciation of taxa A and C. As in the previous case, only root 6 is eliminated.

The complete results for all 3 possible character states for Taxon B are summarized in table 1. Note that the parsimony counts shown in different columns do not necessarily exclude the same roots, but all 3 analyses exclude root 6. For example, when taxon B is in character state "−," both roots 5 and 6 are excluded. But because root 5 is allowed by the other 2 character states, "+" and "m," the possibility of a root at this position cannot be excluded, as shown by a Y in the fifth column (Root OK). Root 6, however, is excluded for all 3 possible character states, and hence, this data set rejects root 6 independently of the state of B. For an example of a second, more complex pattern that excludes the root from a larger region of the tree, see Supplementary Analyses and Data, Section S1, Supplementary Material online. A complete listing of all possible 3 taxon patterns and of the taxa they exclude is provided in Supplementary Analyses and Data, Section S2, Supplementary Material online.

Before applying the top–down algorithm to experimental data, we briefly consider how data might be coded in response to experimental uncertainties. In practice, 3 types of uncertainties need to be considered: 1) genes may be genuinely absent from some taxa, 2) genes may be present but have complex interpretations, or (3) gene sequences may be unavailable. Genes may be genuinely absent, as in 1, as a result of gene deletions or gene duplications (as in fig. 1*B*). If so, they should be coded as missing, "m." Data may have complex interpretations, as in 2, when lateral/horizontal gene transfers might have moved genes to taxa originally lacking them or when frequent indel character–state changes within a taxon have made the ancestral state unobservable. If gene transfers are suspected to be responsible for moving genes into a taxon that originally lacked them, data should be coded either as the experimentally observed state, say "+" (assuming no transfer), or as "m" (assuming the gene was transferred). In this case, a root would have to be excluded for both states "+" and "m," in order to be reliably eliminated. Genes may be unavailable, as in 3, if genomes have not yet been sequenced. In this case, data should be coded in all 3 states, "+," "−," and "m," and a particular root can be eliminated only if it is excluded for all 3. Thus, it is possible, in some circumstances, to eliminate a root even if no sequences are available from that taxon.

To illustrate how top–down rooting can be used to increase the reliability of indel analyses when gene transfers are present, we reanalyze the well-known and controversial heat shock protein Hsp70 indel and its outgroup protein MreB in the next section. A second example, illustrating how top–down rooting can be used when alignments are uncertain, reanalyzes an indel within protein synthesis factors EF-G and EF-Tu and is presented in the Supplementary Analyses and Sequence, Section S4, Supplementary Material online.

Increasing the Number and Reliability of Indel Analyses: A Case of Alternative Interpretations, Heat Shock Response Protein Hsp70

Protein Hsp70 is present in organisms that span the kingdoms of life. It contains 3 primary functional domains. The 1) N-terminal ATPase domain binds ATP and uses energy from this binding to drive conformational changes in a 2) substrate-binding domain, and 3) the C-terminal domain acts as a trap door to close the substrate-binding domain. Among its functions, Hsp70 can protect cells from thermal or oxidative stress and also participate in the disposal of damaged or defective proteins. The well-known Hsp70 indel is found in the N-terminal ATPase domain starting at, approximately, the amino acid position 80 in the *Escherichia coli* sequence. This indel has been used extensively in pioneering studies by Gupta and colleagues to investigate prokaryotic phylogeny (Gupta and Singh 1994; Gupta et al. 1994). These authors also utilized an ancient paralog of Hsp70, MreB (Gupta 1998), making the Hsp70 indel potentially root informative.

However, the usefulness of this indel for rooting purposes has been questioned primarily due to lateral/horizontal gene transfers but also due to an uncertainty about the position and number of gaps (Philippe et al. 1999). These authors point out that "horizontal gene transfers confuse prokaryotic phylogenies based on the Hsp70 protein," and note that "Hsp70 genes have only been characterized in 4 genera of euryarchaeota" and could not be detected in some completely sequenced archaebacterial genomes. They suggest "a recent origin of the known archaebacterial Hsp70 genes," and further suggest "the archaebacterial Hsp70 genes have been acquired through horizontal gene transfer from Gram-positive eubacteria," in accord with the previous suggestion of others (Gogarten et al. 1996). Furthermore, although not mentioned by these authors, we note that the distribution of MreB outgroup sequences in the archaebacteria is very patchy and its origins are also questionable. Thus, the Hsp70/MreB indel represents a challenging problem for top–down analysis.

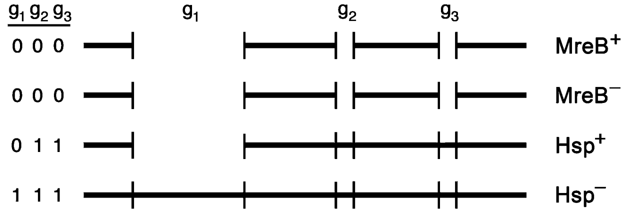

FIG. 2.—A summary of gap locations in the Hsp70/MreB alignment in the vicinity of the Hsp70 indel. For details of the determination of the positions of these indels, see the Supplementary Analyses and Data, Section S3, Supplementary Material online. Regions in which sequences are present are shown as solid lines, and regions in which gaps are present are shown as spaces delimited by vertical lines. MreB sequences are present in single-membrane, Gram-positive eubacteria and in double-membrane, Gram-negative eubacteria. These are labeled MreB$^+$ and MreB$^-$, respectively. Similarly, Hsp70 sequences present in Gram-positive eubacteria and in Gram-negative eubacteria are labeled Hsp$^+$ and Hsp$^-$, respectively. As deduced in the Supplemental Analyses and Data section, Supplementary Material online, the primary gap created by the insert present in Hsp$^-$ sequences, gap 1, is upstream of subsidiary gaps 2 and 3. Hence, for the purposes of this analysis, it is independent of the 2 downstream gaps.

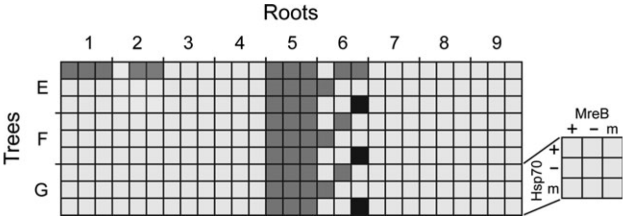

FIG. 3.—A rooting array analysis of the Hsp70/MreB indel. Roots are color coded as follows: roots in red (dark) are excluded, roots in yellow (light) are allowed, and roots in black are allowed, see Supplementary Analyses, Section S3, Supplementary Material online for details. The character states corresponding to the 9 possible combinations of observable character states for archaeal Hsp70 and MreB sequences are labeled on the separate 3 × 3 subarray shown at the lower right of the figure.

Before analyzing the Hsp70/MreB indel, we first determine the number and distribution of the indels found within proteins Hsp70 and MreB because additional distinct gaps within the MreB sequence that are absent in Hsp70 have been noted (Philippe et al. 1999). We also consider the extent of gene transfer. Detailed analyses of these questions are presented in the Supplementary Analyses and Data section, Supplementary Material online, and summarized below.

Our studies identify 2 separate gaps, g2 and g3, that are present in MreB sequences but absent in Hsp70 sequences, in general agreement with Philippe (1999). Both gaps occur demonstrably downstream from the primary Hsp70 insert, g1, present in Gram-negative eubacterial sequences, labeled Hsp$^-$. The primary Hsp70 insert is absent in Gram-positive eubacterial Hsp70 sequences, labeled Hsp$^+$. It is also absent in all Gram-negative and Gram-positive eubacterial MreB sequences, MreB$^-$ and MreB$^+$, respectively. Thus, for the purpose of these analyses, the primary Hsp70 indel is spatially and phylogenetically independent of the 2 subsidiary indels, g2 and g3 (fig. 2).

The principal obstacle preventing the analysis of the Hsp70 indel is that horizontal/lateral gene transfers from the Firmicutes may have been the source of the archaebacterial Hsp70 sequences, although the patchy distribution of the MreB gene in archaebacteria is also a concern. In order to circumvent these difficulties, we make no assumptions about the history of the archaeal genes and indels. Rather, we consider, in turn, all possible character states for both the archaebacterial Hsp70 and the MreB sequences and reject a particular root only if it is excluded for all possible character states and for all possible tree topologies. The resulting computations are fairly complex, involving several hundred alternative tree topologies, roots, and combinations of character states, but they allow us to test rigorously whether or not the Hsp70 indel excludes the root from within the double-membrane (Gram-negative) prokaryotes.

For this analysis, the relevant 4 prokaryotic groupings are the double-membrane (Gram-negative) eubacteria (D), the Archaea (R), the eubacterial Actinobacteria (A), and the eubacterial Firmicutes (F). Together these 4 groups include all known prokaryotic diversity (Boone and Castenholz 2001). Because the Hsp70 insert is present only in the Gram-negative prokaryotes, the observed character states are (+, −, −, −) for ingroup taxa (D, R, A, F) and (−, −, −, −) for outgroup taxa (D′, R′, A′, F′), respectively. In view of the possible Hsp70 gene transfer to the Archaea, and the uncertainty of the archaeal MreB sequences, we consider each of the 3 possible character states, "+," "−," or "m," for these 2 archaeal taxa. Because the character states are unknown for 2 taxa, nine, 3^2, combinations of character states must be investigated. Because 4 taxa are being analyzed and the topology is unknown, 3 unrooted trees are possible, and each must be excluded for all possible character states. In Newick notation, the unrooted trees represented in figure 3 are the E tree, ((D, R),(A, F)), the F tree, ((D, A),(R, F)), and the G tree, ((D, F),(A, R)). Finally, each of the 9 distinct roots, numbered 1–9, must be evaluated. Roots 1–4 are on the branches leading to taxa D, R, A, and F, respectively. Roots 5–8 are within taxa D, R, A, and F, respectively. And root 9 is within the central branch of the E, F, and G trees. Parsimony scores for each of the 243, 3^5, possible rooted trees are presented in figure 3. Excluded and allowed roots are color coded as follows: roots in red (dark) are excluded, roots in yellow (light) are allowed, and roots in black are allowed, see Supplementary Analyses, Section S3, Supplementary Material online. The 9 possible combinations of character states for archaeal Hsp70 and MreB sequences are displayed in the 3 × 3 subarray at the lower right of the figure.

Indels can exclude a root with more statistical support than is commonly thought because 1 indel set can represent more than a million phylogenetically informative indel quartets. In practice, the effective number of independent quartets is considerably less because indel quartet sequences are correlated by an underlying tree structure. In Section S5 (Supplemental Analyses and Alignments, Supplementary Material online), we calculate the correlations from the sequence variation within the indel-flanking regions and determine that the Hsp70/MreB indel is significant at the $P < 0.005$ level.

Of all 9 roots, only the root within the double-membrane eubacteria, root 5, is excluded in all 27 combinations of character states and unknown tree topologies. Other roots are excluded by some combinations of trees and character states, for example, roots 1, 2, and 6, and

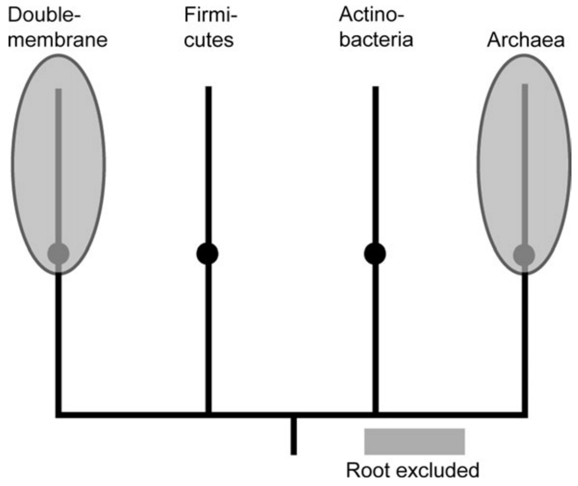

FIG. 4.—A summary of the possible locations for the root of the prokaryotic tree of life. The relevant 4 prokaryotic taxa, representing known prokaryotic diversity, are the double-membrane eubacteria (D), the eubacterial Firmicutes (F), the eubacterial Actinobacteria (A), and the Archaea (R). The 2 regions from which the cenancestral population is excluded are shaded. They correspond to the double-membrane prokaryotes, this study, and the Archaea (Skophammer et al. 2006). The last common ancestors of each group are represented by the dots within the shaded areas. Because the topology relating these 4 groups is unknown, no phylogenetic significance should be attached to the relative placement of these 4 taxa within the figure. Based on these analyses and other considerations related to the molecular architecture of the double-membrane arrangement discussed here, we propose a superphylum, to be known as subdomain Didermataea *subdom. nov.* (sensu stricto Gupta [1998]), consisting of all eubacterial prokaryotes surrounded by a closed system of inner and outer membranes.

allowed by other combinations. But only root 5, located within the double-membrane prokaryotes, is excluded by all combinations. These results are summarized in figure 4, together with results from a previous indel study excluding the root from within the Archaea (Skophammer et al. 2006).

Discussion

It might seem surprising that the Hsp70/MreB analyses overwhelmingly rejected all roots within the double-membrane prokaryotes, even though not a single archaebacterial sequence was used! The reason that archaebacterial sequences were not needed is related to the fact that indel analyses exclude roots from regions, rather than provide positive evidence that a particular root exists. This "top–down" property allows one to exclude roots from recent organisms, even when the data cannot exclude possible roots that are lower down in the tree.

In fact, the top–down property of indel rooting is exactly what is required for finding roots. Imagine, for a moment, that archaebacteria had not yet been discovered, even though they were alive on earth. In that case, in order to root the tree of life, one could only search for a root within the eubacteria because in this example no other prokaryotes were known. Once a unique root was found by exclusion, that Cenacestor (Fitch and Upper 1987) would have been useful for understanding the evolution of known life on earth. Imagine now what would happen if the archaebacteria were discovered! Because this was a top–down rooting, all of the previous work done to exclude the root from the eubacteria would still be useful and valid. But the discovery of archaebacteria would create new potential root locations. Thus, one can see the underlying role of top–down analyses. These new alternatives, and only the new ones, would have to be excluded in order to again obtain a unique root. Because roots are found by exclusion, any root once excluded will remain excluded forever even if new, fundamentally different types of life are found. Returning to the present, but viewed from this perspective, the Hsp70 studies did not, and could not, exclude the root from within the Archaea, or provide phylogenetic information about them because no Hsp70 sequences from them were utilized. But the studies could, and did, exclude the root from within the double-membrane prokaryotes.

Knowing that the root is outside the double-membrane prokaryotes does not greatly constrain the root locations, but it is an interesting start. Remaining potential locations for the root are within the Firmicutes, within the Actinobacteria, on the branches leading to the double-membrane prokaryotes, the Actinobacteria, the Firmicutes, and the Archaea, and within the 3 possible internal branches linking these 4 taxa. The traditional root on the branch leading to the Archaea (Gogarten, Rausch, et al. 1989; Iwabe et al. 1989) is not excluded by the Hsp70 results.

From this perspective, we can now reassess the controversy over the interpretation of the Hsp70 indel. In essence, both research groups were correct because they were focusing on different aspects of the indel. Philippe et al. (1999) correctly asserted that the indel data did not imply that the archaebacteria were more closely related to the Gram-positive eubacteria than they were to any other group. Gupta (1999), on the other hand, interpreted the indel data to imply that the double-membrane organisms were a derived group.

Although the question could not be decided in the absence of top–down analyses, in fact, the exclusion of the root from the double-membrane prokaryotes does not contradict either view. In the unresolved tree shown in figure 4, 9 rooted relationships are possible, and the archaebacteria are the sister taxon to the double-membrane prokaryotes in several of these. In other words, the indel does not demand that the archaebacteria be the sister taxon of the Gram-positive prokaryotes. Gupta on the other hand was equally correct in thinking that the data might exclude the root from within the double-membrane prokaryotes. Thus, both points of view were reasonable.

Although this study is primarily focused on developing a new method of indel analysis, the results obtained here are not without import. The double-membrane prokaryotes are an enormously successful group of eubacteria broadly distributed across the face of earth. They are characterized by an outer membrane that surrounds an inner peptidoglycan layer and an inner cytoplasmic membrane (not to be confused with the nonhomologous double-membrane arrangement surrounding the archaebacterium, *Ignicoccus* [Rachel et al. 2002]). It has been proposed by others (Cavalier-Smith 2002) that the root is within the double-membrane prokaryotes because it would be evolutionarily impossible to evolve a double-membrane prokaryote from a single-membrane prokaryote. However, an analysis of Hsp70 indel variants based on correlations provides strong statistical support, $P < 0.005$, for excluding the root from the double-membrane prokaryotes, clearly indicating that it is not within the double-membrane prokaryotes, see section S5 in Supplemental Analyses and Alignments, Supplementary Material online for statistical details.

Derivation of the double-membrane arrangement from single-membrane prokaryotes fits well with the general knowledge that the outer membrane greatly complicates many processes that are much simpler in single-membrane prokaryotes. For example, the process of flagellar assembly is considerably more complex in double-membrane prokaryotes than in single-membrane prokaryotes (Macnab 2003). In double-membrane prokaryotes, the process requires the construction of novel flagellar rings, the L and P ring assemblies, in addition to the M and S rings associated with the cytoplasmic membrane in both single- and double-membrane prokaryotes. The L and P rings, associated with the outer membrane and the peptidoglycan layer, respectively, permit flagella to pass through the outer membrane. In addition, other design changes are required to accommodate transport across the double-membrane arrangement. Special ATP-binding casette transporters differing considerably from those present in single-membrane prokaryotes, for example, are used to facilitate the uptake of vitamin B_{12} in double-membrane prokaryotes (Locher et al. 2002). Numerous molecular synapomorphies in addition to the Hsp70 indel correlate with the presence of the double membrane, including an indel present in the beta subunit of DNA-dependent RNA polymerases (Morse et al. 2002) as well as many others directly related to the double membrane like those noted above. Although they cannot confirm the direction of the root, they independently validate the phylogenetic significance of the Hsp70 indel. These results emphasize the importance of the novel evolutionary events that resulted in the double-membrane prokaryotes. Clearly, this innovation has produced one of the largest and most broadly distributed groups of prokaryotic life found on the face of earth.

Top–down rooting also makes it possible to analyze data from incomplete gene sets. We estimate that this algorithm may considerably increase the number of root informative data sets. Because indel sets require 2 paralogous gene sets, the probability of finding a useful indel pair increases as the square of the number of useful sets. In an analysis of 8 genomes spanning the prokaryotic tree of life with a median genome size of 2,500 genes per taxon, there were about 400 genes present in all taxa, approximately 350 genes present in all but 1 taxon, and about 300 genes present in all but 2 taxa (Rivera and Lake 2004). Using these numbers as a guide, if one were to accept an average of 1 missing taxon in a 7-taxon ortholog/paralog set, then the number of useful ortholog sets would nearly double and the total number of available ortholog pairs would increase by 350%. If an average of 2 missing taxa were acceptable, as in this Hsp70 study, then the total number of available ortholog/paralog sets would increase by 700%. Because many of the 400 ubiquitous genes appear to be compromised by gene transfers, the increase in the number of usable gene sets is probably considerably larger.

One of the principal advantages of top–down indel analyses is that they allow one to analyze with considerable certainty indel data that, even though present in all taxa, are so ambiguous that they previously would have been thought to be useless. Thus, top–down analyses have the potential to increase both the robustness of indel-rooting studies and the number of useful data sets.

Finally, we need to ask what data might argue against these analyses. Certainly, gene transfers between phylogenetic groups can present significant difficulties (Doolittle 1999). In fact, the proposed Hsp70 gene transfers between the Gram-positive eubacteria and archaebacteria (Philippe et al. 1999) stimulated this paper. Because Hsp70 indel transfers between double-membrane eubacteria and single-membrane eubacteria may confound these analyses, we estimated gene transfer rates by counting the numbers of exceptions within the double- and single-membrane prokaryotes; for details, see Supplementary Analyses and Data, Table S2, Supplementary Material online. The distribution of indels was consistent with relatively few gene transfers between these 2 groups. The Hsp70 insert is present in 98.0% of the double-membrane, Gram-negative sequences and in 2.6% of the single-membrane, Gram-positive sequences. Conversely, the Hsp70 gap is present in 2.0% of the double-membrane sequences and in 97.4% of the single-membrane sequences. We interpret these figures to indicate that gene transfers between Gram-negative and Gram-positive eubacteria have affected 2–3% of the sequences. This level of gene transfer seems to be tolerable and probably would not have significantly affected the conclusion that the root of the tree of life is not within the double-membrane prokaryotes.

We are optimistic that top–down indel rooting has great potential to improve the reliability with which existing indel gene sets can be analyzed. They may also be useful for studying eukaryotic relationships, even when the relevant genes from some taxa may not be available. And finally, we anticipate that they may increase the number of available indel sets considerably beyond those presently available.

Supplementary Material

All supplementary materials are indexed and contained in the Supplementary Analysis and Sequences Section, which is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org).

Acknowledgments

Dedicated to Walter Fitch who first taught me (J.A.L.) parsimony. Supported by grants from National Science Foundation (NSF) and the University of California, Los Angeles–National Aeronautics and Space Administration Astrobiology Institute to J.A.L. R.G.S., J.A.S., and C.W.H. were supported by an Integrated Graduate Education and Research Training Grant from NSF, a Cell and Molecular Biology Training Grant from National Institutes of Health (NIH), and a Genomic Interpretation and Analysis Training Grant from NIH, respectively.

Literature Cited

Baldauf S, Palmer J. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA. 90:11558–11562.

Boone D, Castenholz RW. 2001. The *Archaea* and the deep branching and phototrophic *Bacteria*. New York: Springer.

Boucher Y, Douady C, Papke R, Walsch D, Boudreau M, Nesbo C, Case R, Doolittle W. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet. 37:283–328.

Brown JR, Doolittle WF. 1995. Root of the universal tree of life based on ancient aminoacyl-transfer RNA synthetase gene duplications. Proc Natl Acad Sci USA. 92:2441–2445.

Cavalier-Smith T. 2002. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol. 52:7–76.

Dayhoff MO, Schwartz RM. 1980. Prokaryote evolution and the symbiotic origin of eukaryotes. In: Schwemmler W, Schenk HEA, editors. Endocytobiology: endosymbiosis and cell biology: a synthesis of recent research, Vol. 1. Proceedings of the International Colloquium on Endosymbiosis and Cell Research, Tuebingen, West Germany, April, 1980. Berlin (West Germany): Walter De Gruyter and Co. p. P63–P84.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284:2124–2128.

Fitch WM, Upper K. 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harb Symp Quant Biol. 52:759–767.

Gogarten JP, Kibak H, Dittrich P, et al. (13 co-authors). 1989. Evolution of the vacuolar H+-ATPase—implications for the origin of eukaryotes. Proc Natl Acad Sci USA. 86: 6661–6665.

Gogarten JP, Olendzenski L, Hilario E, Simon C, Holsinger KE. 1996. Dating the cenancestor of organisms. Science. 274:1750–1751.

Gogarten JP, Rausch T, Bernasconi P, Kibak H, Taiz L. 1989. Molecular evolution of H+-ATPases. I. Methanococcus and Sulfolobus are monophyletic with respect to eukaryotes and eubacteria. Z Naturforsch Sect C J Biosci. 44:641–650.

Gupta RS. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev. 62:1435–1491.

Gupta RS. 1999. Hsp70 sequences and the phylogeny of prokaryotes. Mol Microbiol. 31:1109–1110.

Gupta RS, Aitken K, Falah M, Singh B. 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.

Gupta RS, Singh B. 1994. Phylogenetic analysis of 70-kD heat-shock protein sequences suggests a chimeric origin for the eukaryotic cell-nucleus. Curr Biol. 4:1104–1114.

Harper JT, Keeling PJ. 2004. Lateral gene transfer and the complex distribution of insertions in eukaryotic enolase. Gene. 340:227–235.

Iwabe N, Kuma K, Hasegawa M, Osawa S, 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.

Locher KP, Allen TL, Rees DC. 2002. The *E. coli* BtuCD structure: a framework for ABC transporter architecture and mechanism. Science. 296:1091–1097.

Macnab RM. 2003. How bacteria assemble flagella. Annu Rev Microbiol. 57:77–100.

Morse R, O'Hanlon K, Collins MA. 2002. Phylogenetic, amino acid content and indel analyses of the beta subunit of DNA-dependent RNA polymerase of Gram-positive and Gram-negative bacteria. Int J Syst Evol Microbiol. 52:1477–1484.

Philippe H, Budin K, Moreira D. 1999. Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family. Mol Microbiol. 31:1007–1012.

Rachel R, Wyschkony I, Riehl S, Huber H. 2002. The ultrastructure of *Ignicoccus*: evidence for a novel outer membrane andfor intracellular vesicle budding in an archaeon. Archaea. 1:9–18.

Rivera MC, Lake JA. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science. 257:74–76.

Rivera MC, Lake JA. 2004. The ring of life: evidence for a genome fusion origin of eukaryotes. Nature. 431:152–155.

Skophammer RG, Herbold CW, Rivera M, Servin JA, Lake JA. 2006. Evidence that the root of the tree of life is not within the Archaea. Mol Biol Evol. 23:1–4.

Zhaxybayeva O, Lapierre P, Gogarten JP. 2005. Ancient gene duplications and the root(s) of the tree of life. Protoplasma. 227:53–64.

Martin Embley, Associate Editor

Accepted September 29, 2006