Structural Analysis of Polarizing Indels Argues the Root of the Tree of Life is Near the Chloroflexi

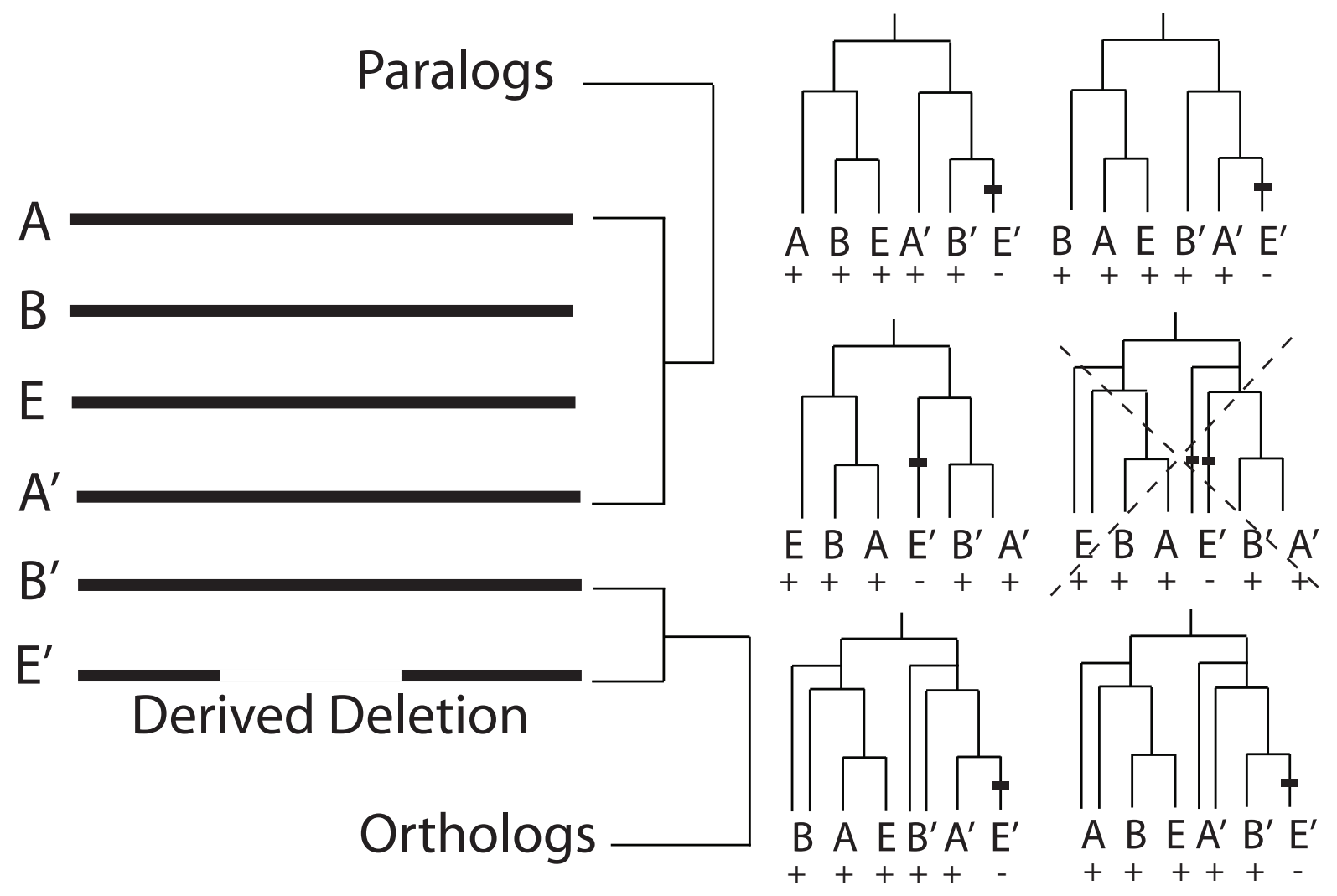
Ruben Valas, Philip Bourne Contact: rvalas@sdsc.edu
Bioinformatics Program, University of California, San Diego

Abstract

Determining which branches of the tree of life have derived features narrows down the possible location of the root. Currently the polarization of indels done by Lake *et al.*[1-5] and the polarizing transitions of Cavalier-Smith[6] arrive at contradictory positions for the root of the tree. We have analyzed the sequence based indel arguments using protein structure wherever possible. Structure strongly supports some of the polarizations, but in other indels it argues for a different conclusion. We conclude that there is no contradiction between Lake *et al.* and Cavalier-Smith; the root of the tree of life must be near the Chloroflexi.

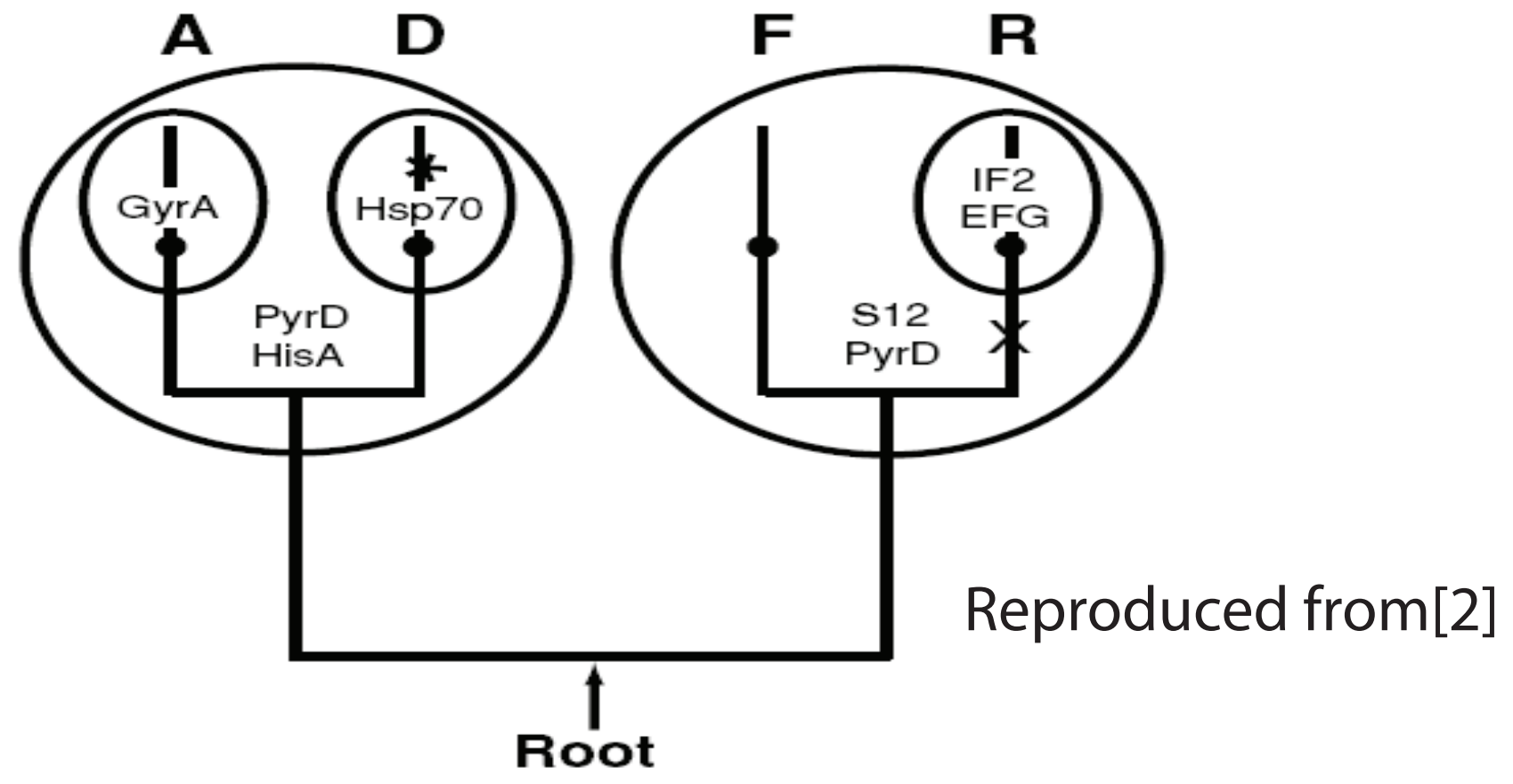
Introduction

Darwin's theory implies that all life arose from a common ancestor. The first split in the branches of the tree corresponds to that Last Universal Common Ancestor (LUCA). The search for LUCA has been composed of arguments of primacy as well as exclusion. However, arguments about which extant species are the most primitive rely on assumptions about what primitive life was like, which leads to circular reasoning when searching for the root. Exclusive methods do not rely on such assumptions.



One exclusive method is indel polarization. As an example, consider the hypothetical set of 2 paralogous genes above. For simplicity lets assume these genes are universally distributed and the duplication that resulted in the paralogs occurred before LUCA. All the Eukaryotes have an indel in E′. It would normally be ambiguous whether this is the result of an insertion or deletion. However, we know all the paralogous and some of the orthologous sequences have that region, which implies the ancestor of both genes had it. So E′ must be a derived form of the gene. There are 6 ways of rooting a tree with 3 taxa. The most parsimonious scenario for explaining the deletion of that region for each of these trees is presented above. A rooting within the Eukaryotes requires at least 2 losses, so it less parsimonious than the other 5 trees, and we can exclude the root from the Eukaryotes in this example.

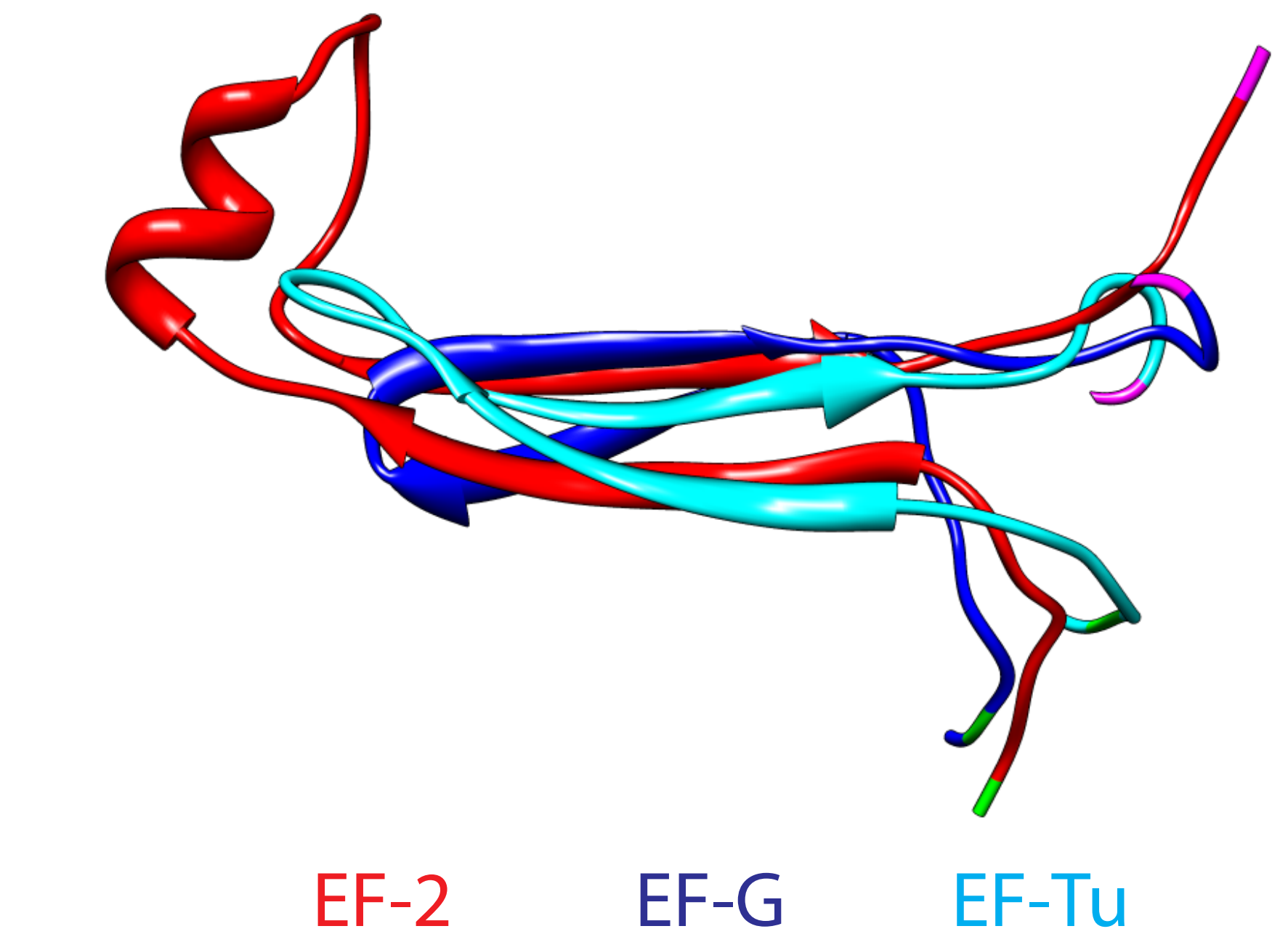
Reality is trickier; horizontal transfer and gene loss muddy the waters. Lake *et al.* have developed a method to polarize indels despite these factors called top-down rooting[1]. They have presented 8 polarized indels that root the tree of life between two clades, shown below. The first is the Actinobacteria (A) and Gram-negatives (D). The second is the Firmicutes (F) and Archaea (R). Cavalier-Smith has presented 13 polarizing transitions that place the root within the Gram-negative bacteria, near the Chloroflexi (marked with an * below). It is important to note that both of these methods agree that the Archaea are derived from Gram-positive Bacteria. However, the major disagreement on the placement within the Bacteria must be resolved. At least one of the polarizations in these methods must be wrong.



Reproduced from[2]

A quality alignment is the prerequisite for properly identifying and polarizing an indel. Paralogs that duplicated before LUCA are required to exclude the root from a portion of the tree. That means that polarized indel arguments require alignments between paralogs that diverged over 3.5 Gya. The Achilles' heel of indel polarization is the alignment step. We have used structural alignments and information about quaternary structure to analyze Lake *et al.*'s conclusions. Their exclusion of the root from Actinobacteria based on an insert in GyrA appears incredibly robust based on sequence alone. Our data supports the controversial indel in EF-2 that excludes the root from Archaea and Eukaryotes. However, we find that none of the 3 arguments that actually exclude the root from all Gram-negatives as they claim. Instead we find they actually exclude the root from all Gram-negatives except the Eobacteria (Deinococcus-Thermus and Chloroflexi) which is completely consistent with Cavalier-Smith's rooting near the Chloroflexi.

An insert in EF-2 excludes the root from Eukaryotes and Archaea

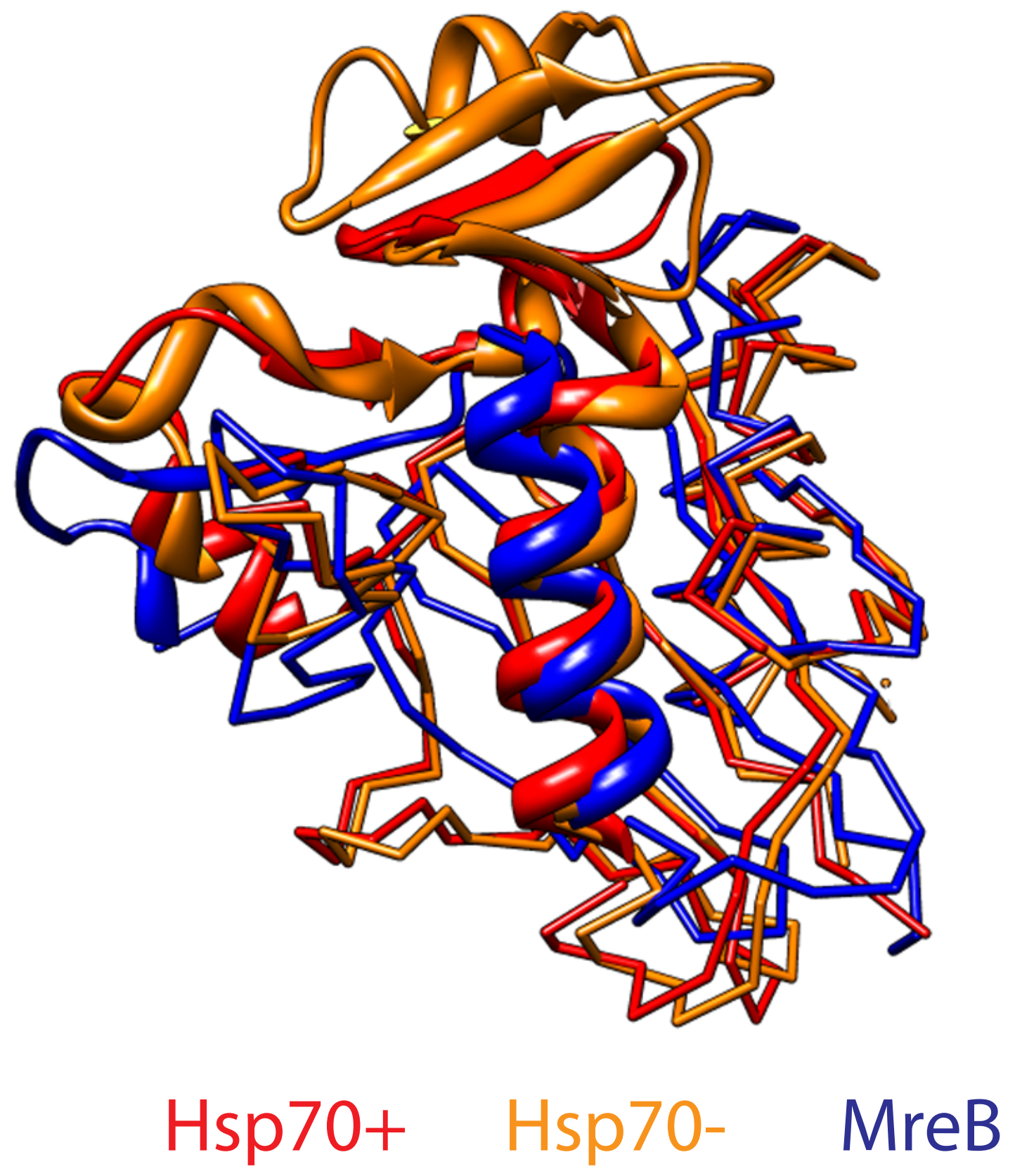


```
EF-2   GITI-KSTAISLYSEMSDEDVKEIKQKTDGNSFLINLIDSPGH--V-D
EF-G   -RGITI-TAAVTTCFWK---------------DHRINIIDAPGH--VD-
EF-Tu  -RGITI-NTSHVEYDTP---------------TRHYAHVDCPGHAD---
```

EF-2 is the Archaeal and Eukaryotic homolog of EF-G. EF-2 contains an insert that is not present in EF-G or EF-Tu which implies the Archaea and Eukaryotes are derived. However, the authors' sequence alignment has been called into question. We performed a structural alignment EF-G (2BV3 colored blue), EF-Tu (1EFC colored cyan), EF-2 (1N0U colored red) to investigate whether their conclusion is valid. Our analysis is complicated by a disordered region in the structures that precedes the region of interest. The well conserved RGIT motif does align correctly when using structure alone. This is highlighted by showing the difference in position of the conserved Glycine on the C end of this region (colored green) as well the conserved Aspartic acid in the N end (colored purple). The 4 positions highlighted in red are aligned in the original alignment, which is why the initial result is controversial. This is further complicated by an additional insert in Eukaryotes relative to the Archaea, which is boxed in black. Neither the sequence alignment, nor structural alignment is strictly correct. However, when one combines both pieces of data it becomes clear that there must a derived insertion in the Archaea, which implies the root cannot be in the Archaea or the Eukaryotes.

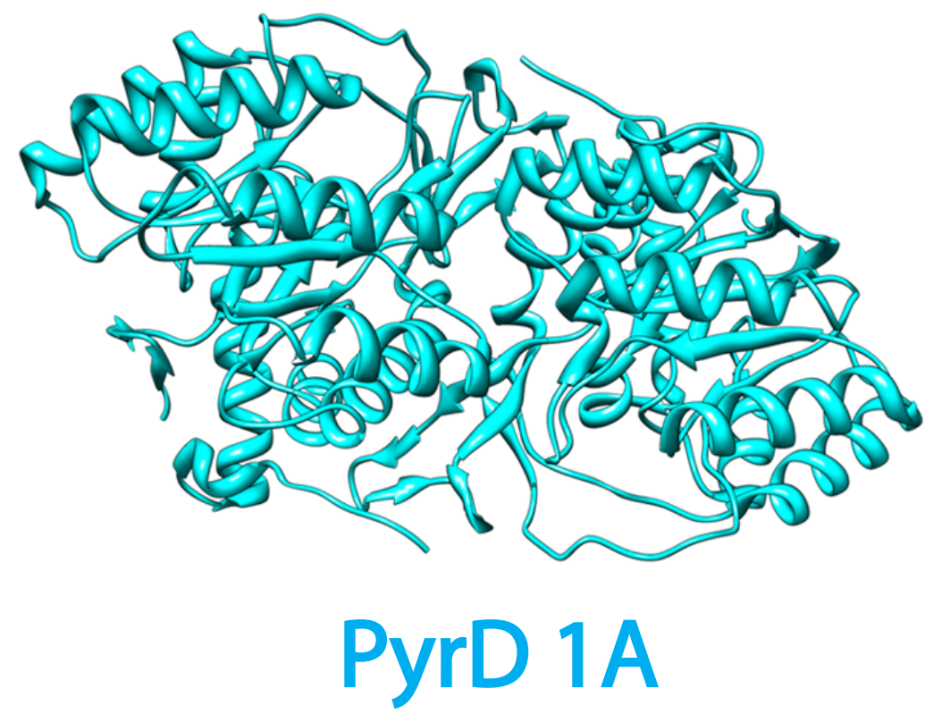
The Mreb Hsp70 indel is inconclusive

The first indel that apparently disagrees with Cavalier-Smith's root is in Hsp70. Hsp70 has a large insert in the Gram-negatives relative to the Gram-positives. Lake *et al.* claim this region is conserved between the Gram-positives and the paralogous MreB. This would imply that the Gram-negative must be derived. However, a structural alignment of representatives of these 3 proteins (MreB 1JCF:A in blue, Hsp70-2V7Y:A in red, and Hsp70+ 1DKG:D in orange) reveals that Hsp70 from Gram-negatives has a significant insert relative to MreB. This makes the ancestral state of the paralogs ambiguous, so the 2 forms of Hsp70 cannot be polarized. This would be insignificant since there are 2 other indels that apparently exclude the root from the Gram-negatives, but we will argue below that neither of these robustly excludes the root from the Eobacteria.

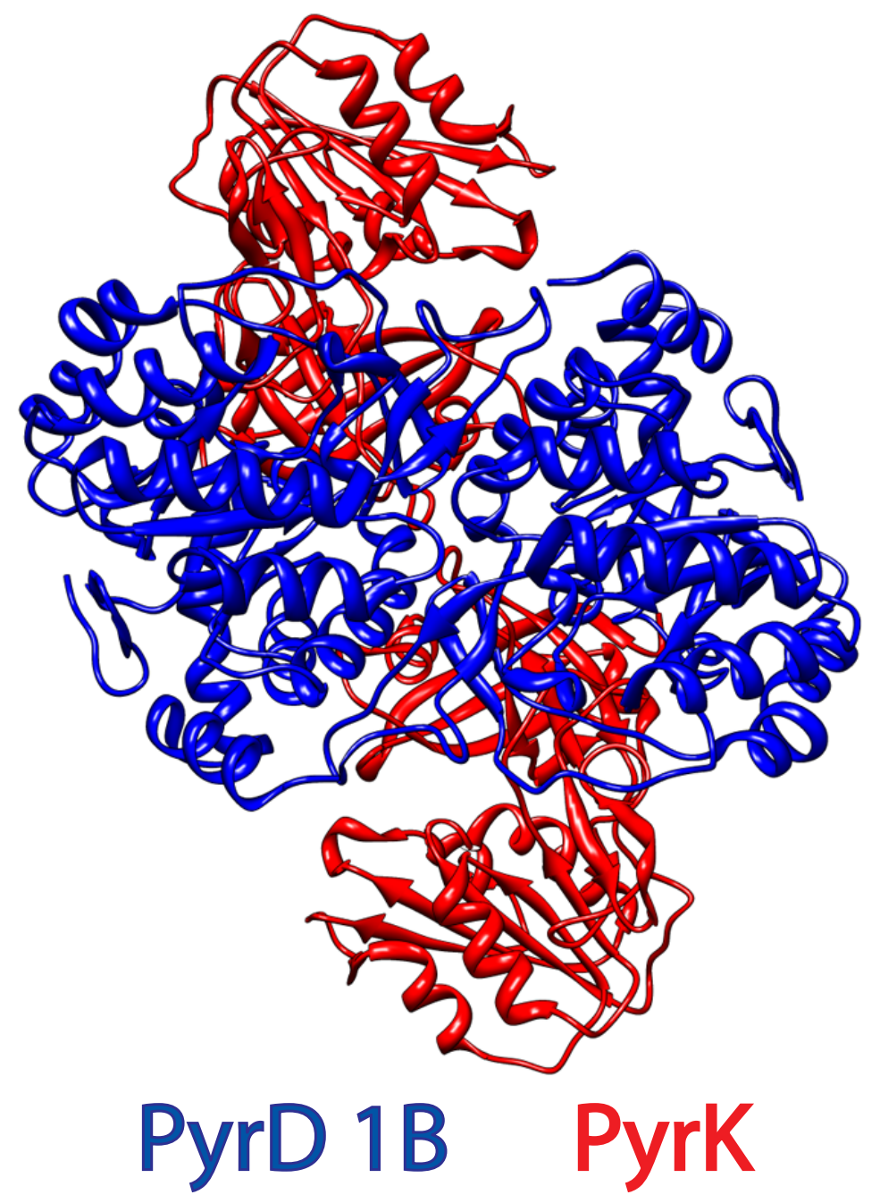


```
Hsp70+  AITN----PNTIISIKRHMGT-----------------------DYKVEIEGKQYTPQEISAIILQYLKSYA
Hsp70-  AVTN---PQNTLFAIKRLIGRRFQDEEVQRDVSIMPFKIIAADNGDAWVEVKGQKMAPPQISAEVLKKMKKTA
MreB    IGKTPATIKAIRPMRDGVI-------------------------------ADYTVALVMLRYFINKA
```

The quaternary structure of PyrD excludes the root from the Firmicutes and Archaea



PyrD 1A



PyrD 1B PyrK

The small indel in PyrD has been polarized using several different outgroups to reach contradictory conclusions that are not mutually compatible[2, 5]. The best outgroup is probably HisA, which implies the Gram-negatives and Actinobacteria are derived. However the different forms of the indel correspond to different quaternary structures in this case. PyrD 1A is a homodimer (1JUB colored cyan to the left), and PyrD 1B is a heterotetramer. The homodimer interface at the center of PyrD 1B (1EP3 colored blue to the left) is similar to the interface in PyrD 1A. PyrD 1B has an additional subunit PyrK (colored red to the left). This implies that PyrD 1B is derived from 1A. All the sequences that have the apparently ancestral deletion are in the PyrD 1B family. We argue that in this case the structural polarization trumps the sequence based argument as it a smaller evolutionary event to lose two amino acids than it is to gain an entire protein-protein interface. PyrD 1B is present across the Firmicutes and Archaea, so we can exclude the root from both of these clades.

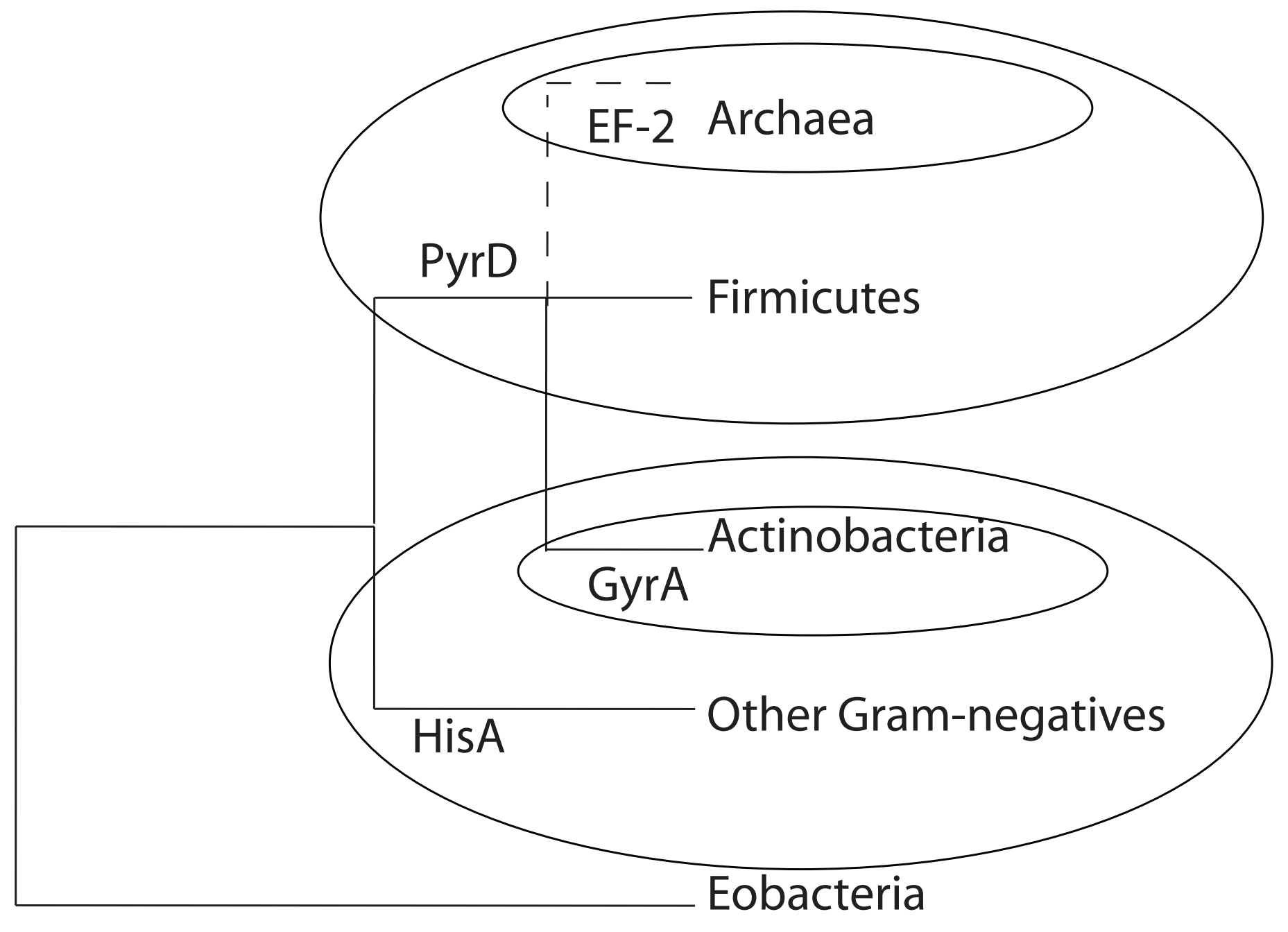
The HisA HisF indel does not exclude the root from the Eobacteria

```
                                                         225       235       245
Streptomyces_coelicolor_A3/1-240              VASGGVSSLDDLRAIAGLVPAGVEGAIVGKALY
Synechocystis_sp._PCC_6803/1-256              IASGGVSQVEDLLNLLSLESLGVNGVIIGKALY
Bacillus_clausii_KSM-K16/1-240                IASGGVSSIADLAQLRAA---SLAGAIVGKALY
Thermus_thermophilus_HB27/1-235               IVGGGIASPEDLHLLQAL---GVDGALIGKALY
Thermus_thermophilus_HB8/1-235                IVGGGIASPEDLHLLQAL---GVDGALVGKALY
Deinococcus_geothermalis_DSM_11300/1-243      IVGGGVATLDDVRLLAEE---GIQGAIVGRAIY
Deinococcus_radiodurans_R1_/1-238             IVGGGVADTNDVRLLAEE---GIEGAIVGRAIY
Herpetosiphon_aurantiacus_ATCC_23779/1-240    IASGGIARVEHLARLAEL---GISGAIIGTALY
Chloroflexus_aurantiacus_J-10-fl/1-241        IASGGIASIDHVRQLAQL---GVEGAIIGKALY
Roseiflexus_castenholzii_DSM_13941/1-237      IASGGISTLDHLRRLADV---GVEGAIVGRALY
Roseiflexus_sp._RS-1/1-237                    IASGGISTIDHLRRLAEV---GVEGAIVGRALY
Dehalococcoides_ethenogenes_195/1-237         IASGGVSSLSHLRLLKDI---GAEGAIVGKAIY
Dehalococcoides_sp._BAV1/1-258                IASGGVSSLSHLRLLKDI---GAEGAIVGKAIY
Dehalococcoides_sp._CBDB1/1-244               IASGGVSSLSHLRLLKDI---GAEGAIVGKAIY
```

2 of the indel polarizations that exclude the root from the Gram-negatives are flawed. This would be a moot point if the HisA HisF indel robustly excluded the root from all Gram-negatives. HisA and HisF are ideal paralogs for indel polarization because they are widely distributed and highly conserved at the sequence level. That reduces the chance of a conclusions being based on an alignment artifact. Our analysis of this data is mostly in agreement with the authors. There is a derived insertion in many HisA sequences that can be used to exclude the root. However, we noticed that this insert does not appear to be present in any Eobacteria. A sequence alignment between the Eobacteria with a representative sequence from Actinobacteria (that have the insert), other Gram-negatives (that have the insert), and from Firmicutes (that lack the insert) shows that all the Eobacterial sequences have the same form of the indel as the Firmicutes. The polarization of this indel excludes the root from all Actinobacteria and Gram-negatives except the Eobacteria.

Conclusion

Exclusive rooting methods offer great promise in determining the nature of LUCA. However, any single argument could be wrong due to the difficulty of correctly polarizing a transition. The insert in EF-2 excludes the root from the Archaea and Eukaryotes. The insert in GyrA robustly excludes the root from the Actinobacteria. Our analysis of the indel data finds no evidence that excludes the root from all Gram-negatives. If any of these 3 arguments held up it would prove Cavalier-Smith's rooting wrong. We take the fact that none of these arguments can robustly exclude the root from Eobacteria as evidence supporting Cavalier-Smith's rooting. When the indel data is combined with other the polarizations a consensus mapping of the major evolutionary events emerges: 1) the root of the tree of life is in the Eobacteria, 2) the Gram-positives are derived from the Gram-negatives, and 3) the Archaea are derived from the Gram-positives. At this point we will not argue whether the ancestor of the Archaea is more like a Firmicute or an Actinobacteria so that line is drawn ambiguously.



References

1. Lake, J.A., et al., Rooting the tree of life using nonubiquitous genes. Mol Biol Evol, 2007. 24(1): p. 130-6.
2. Lake, J.A., et al., Evidence for a new root of the tree of life. Syst Biol, 2008. 57(6): p. 835-43.
3. Servin, J.A., et al., Evidence excluding the root of the tree of life from the actinobacteria. Mol Biol Evol, 2008. 25(1): p. 1-4.
4. Skophammer, R.G., et al., Evidence that the root of the tree of life is not within the Archaea. Mol Biol Evol, 2006. 23(9): p. 1648-51.
5. Skophammer, R.G., et al., Evidence for a Gram-positive, eubacterial root of the tree of life. Mol Biol Evol, 2007. 24(8): p. 1761-8.
6. Cavalier-Smith, T., Rooting the tree of life by transition analyses. Biol Direct, 2006. 1: p. 19.