Erratum

A Phylogenetic Mixture Model for Gene Family Loss in Parasitic Bacteria

Mol Biol.Evol. 26: 1901-1908. 2009.

Matthew Spencer and Ajanthah Sangaralingam

February 19, 2010

The 16S tree for the COG (Tatusov et al., 1997) data in the original paper was estimated with one category of variation among sites, instead of with four gamma rate categories as reported. Revised versions of Figure 2 (the 16S tree, now estimated with four gamma categories) and Table 1 (based on this revised tree) are given here. The parameter estimates and log likelihoods are similar to those in the original paper.

For six bacterial species (*Neisseria meningitidis* Z2491, *Neisseria meningitidis* MC58, *Sinorhizobium*, *Salmonella typhimurium* LT2, *Helicobacter pylori* J99 and *Mycobacterium tuberculosis* CDC1551), the strain used in the previous 16S tree did not match the strain found in the COG database. These have been corrected in the revised version of Figure 2. Where no strain information was given in COG, or a sequence for a strain could not be found we used an available strain in the RDP database (Cole et al., 2007). The corrected tree files are available at `http://www.liv.ac.uk/~matts/genecontent.html`.

Figure 2: Estimated edge lengths for the 16S topology under model F (Table 1) for the COG data. Edges leading only to parasites/endosymbionts are shown as thicker red lines. Parasites/endosymbionts are indicated by circles and bold genome names. Edge lengths are in expected numbers of gains and losses per gene family at stationarity. Tree drawn using Dendroscope (Huson et al., 2007).

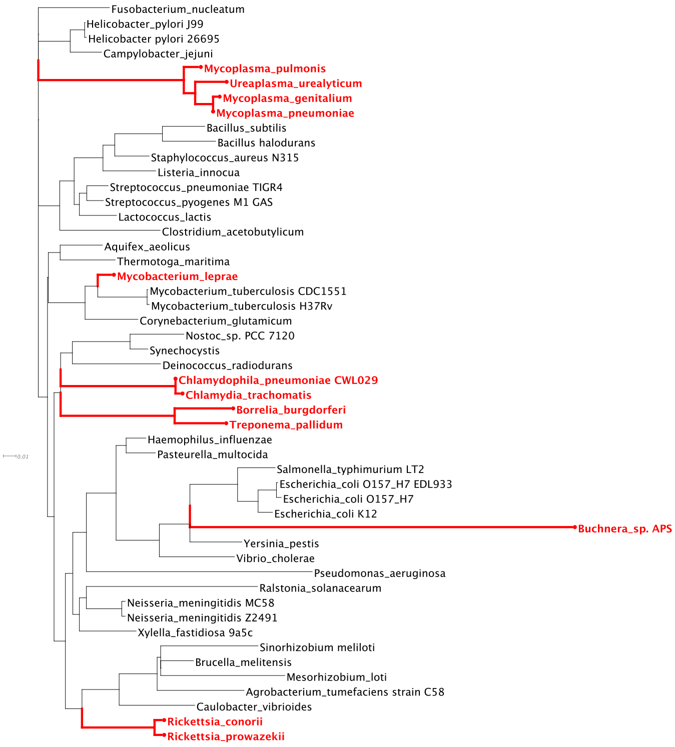

Table 1: Log likelihoods and parameter estimates for six gene family gain and loss models fitted to 50 COG bacterial genomes.

Topology[a]	Model[b]	l[c]	m[d]	$\pi_Q(0,0)$[e]	$\pi_Q(0,1)$	$\pi_{\mathrm{ROOT}}(0,0)$[f]	$\pi_{\mathrm{ROOT}}(0,1)$	μ_0[g]	α[h]
16S	A	-71393	2	0.51 (0.004)	-	0.86 (0.006)	-	-	-
	B	-67253	3	0.34 (0.003)	0.95 (0.0008)	-	0.82 (0.006)	-	-
	C	-66547	5	0.34 (0.003)	0.97 (0.0007)	0.18 (0.03)	0.89 (0.006)	0.09 (0.006)	-
	D	-66091	3	0.02 (0.0004)	-	0.97 (0.003)	-	-	0.50 (0.009)
	E	-64858	4	0.13 (0.003)	0.91 (0.003)	-	0.91 (0.004)	-	0.71 (0.023)
	F	-64767	6	0.13 (0.003)	0.92 (0.003)	0.49 (0.052)	0.93 (0.005)	0.04 (0.005)	0.72 (0.025)
Conditioned logdet	D	-64694	3	0.03 (0.001)	-	0.96 (0.003)	-	-	0.52 (0.011)
	E	-64149	4	0.08 (0.002)	0.84 (0.005)	-	0.93 (0.004)	-	0.67 (0.021)
	F	-64142	6	0.07 (0.002)	0.85 (0.005)	0.03 (0.01)	0.94 (0.004)	0.01 (0.002)	0.67 (0.021)

[a] Models D to F are estimated on two different tree topologies: one derived from 16S sequence data, and one from conditioned logdet distances based on gene content data.

[b] Models as follows. A: no rate shift. B: rate shift in all gene families. C: categories with and without rate shift. D: no rate shift, gamma rate variation. E: rate shift in all gene families, gamma rate variation. F: categories with and without rate shift, gamma rate variation.

[c] Log likelihood.

[d] Number of parameters (excluding the 98 edge lengths on the rooted tree which must be estimated for all models).

[e] $\pi_Q(0,i)$ is the stationary probability of gene family absence in rate matrix i, where $i = 0$ is the rate matrix used throughout major category 0 and on all edges except those leading only to parasites in major category 1. Rate matrix $i = 1$ is used only on edges leading only to parasites in major category 1.

[f] $\pi_{\mathrm{ROOT}}(0,j)$ is the probability of gene family absence at the root in major category j.

[g] Mixing probability for major category 0.

[h] Shape parameter for gamma rate variation.

[i] Approximate standard errors in parentheses for all parameters.

References

Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje. 2007. The ribosomal database project (RDP-II): introducing *myRDP* space and quality controlled public data. Nucleic Acids Research **35**:D169–D172.

Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp. 2007. Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics **8**:460.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science **278**:631–637.