A phylogenetic mixture model for gene family loss in parasitic bacteria

Submitted as a research article

Matthew Spencer and Ajanthah Sangaralingam

School of Biological Sciences, University of Liverpool, UK

Correspondence: Matthew Spencer, School of Biological Sciences, University of Liverpool, Liverpool, L69 7ZB, UK.

Phone +44 (0)151 795 4399. Fax +44 (0)151 795 4404.

Email m.spencer@liverpool.ac.uk

Key words: gene families, mixture model, parasites, maximum likelihood,phylogenetics

Running head: gene family gain and loss

Abstract

Gene families are frequently gained and lost from prokaryotic genomes. It is widely believed that the rate of loss was accelerated for some but not all gene families in lineages that became parasites or endosymbionts. This leads to a form of heterotachy which may be responsible for the poor performance of phylogeny estimation based on gene content. We describe a mixture model which accounts for this heterotachy. We show that this model fits data on the distribution of gene families across bacteria from the COG database much better than previous models. However, it still favours an artefactual tree topology in which parasites form a clade over the more plausible 16S topology. In contrast to a previous model of genome dynamics, our model suggests that the ancestral bacterium had a small genome. We suggest that models of gene family gain and loss are likely to be more useful for understanding genome dynamics than for estimating phylogenetic trees.

Introduction

A gene family is the set of members of a group of repeated sequences in a genome (Graur and Li, 2000, p. 264), derived from a common ancestor within some given time (or sequence similarity) threshold. Gene families can be gained by horizontal transfer or by sequence evolution resulting in a change of the gene family into which a gene is classified (it is therefore not impossible, although presumably unlikely, for sequence evolution to result in multiple origins of a gene family). Gene families can be lost by deletion or sequence evolution. Understanding the processes by which gene families are gained and lost is a major part of understanding genome evolution, given the substantial variation in gene content across genomes. Increasingly, this understanding is being gained using progressively more realistic Markov models of gene gain and loss on a phylogenetic tree.

The first phylogenetic model of gene family gain and loss assumed that the rates of gain and loss were equal, and that the same rates applied to all gene families (Hao and Golding, 2006). This is analogous to a Jukes-Cantor model of nucleotide evolution. The first obvious improvement to this model is to allow the gain and loss rates to be different, so that the stationary probabilities of gene family absence and presence need not be equal. Unequal gain and loss rates give much better fits to data on genomes from across the tree of life (Cohen

et al., 2008). Not assuming that the process of gain and loss has reached stationarity further improves the fit. This requires estimating the probability of gene family absence at the root of the tree in addition to the gain and loss rates (Cohen et al., 2008). There is also strong evidence that some gene families are gained and lost at higher rates than others. Cohen et al. (2008) and Hao and Golding (2008) developed mixture models with several categories of gene families, each having the same relative gain and loss rates but different absolute rates. These models are exactly analogous to discretized gamma models of rate variation in sequence evolution (Yang, 1994), and are homogeneous in time. As with sequence evolution, adding this kind of rate variation results in huge improvements in fit (Cohen et al., 2008; Hao and Golding, 2008). However, there are biological arguments for considering heterogeneous models of rate variation among gene families and over time. For example, the number of genes in bacterial genomes varies by approximately an order of magnitude (Mira et al., 2001). In particular, parasitic and endosymbiotic bacteria (parasites for short from now on) often have much smaller genomes than their free-living relatives, probably because they can rely on their host to perform some functions (Mira et al., 2001). The gene families performing these functions in the parasite might then be lost at a higher rate than normal. Gene families involved in functions that cannot be performed by the host are not likely to show accelerated loss rates, although parasites may have reduced opportunities for lateral transfer (Mira et al., 2001). The parasitic lifestyle has arisen multiple times in unrelated lineages (Ochman and Moran, 2001). As a result, each of these lineages may have a common set of gene families with accelerated loss rates (Snel et al., 2005). This common change in evolutionary rates for a subset of data in unrelated lineages is analogous to one form of heterotachy in sequence evolution (Kolaczkowski and Thornton, 2004), and has the potential to mislead phylogenetic methods based on gene content, even those such as conditioned genome reconstruction that do not assume a stationary, homogeneous process (Spencer et al., 2007).

To address the problem of accelerated gene family loss in parasites, we develop a heterogeneous mixture model with two major categories of gene families. One major category of gene families has the same gain and loss rates everywhere (Figure a). A second major category of gene families that are dispensable in parasites has accelerated loss rates in edges leading to these taxa (Figure b). We estimate the probabilities of these major categories. Within each major category, we also allow discrete gamma rate variation of the kind modelled by Cohen

et al. (2008) and Hao and Golding (2008).

We will assume that the tree topology is known, and that we know which genomes belong to parasites. With these assumptions, the pattern of gain and loss rates for each major category on each edge can be specified in advance (although the actual rates must be estimated). In general, a model of this kind will be nonstationary, because we will not force the stationary probabilities to be the same everywhere on the tree. We must therefore work with a rooted tree, and estimate the probabilities of gene family presence and absence at the root in each major category. Previous models have assumed that gain and loss edge lengths are either a constant multiple of substitution edge lengths (Cohen et al., 2008), or that the relationship between gain and loss edge lengths and substitution edge lengths is linear, but differs among edges in ways that can be specified a priori (Hao and Golding, 2006, 2008). Since there is no reason to think either of these assumptions is true, we will also estimate gain and loss edge lengths.

Using these methods, we show that there is strong evidence for an increase in the loss rate of some but not all gene families in parasitic bacteria. We also show that our model favours a small rather than a large genome in the ancestral bacterium. We compare the fit of our models on a plausible tree topology based on 16S data, and an implausible topology based on gene content data.

Materials and Methods

Modelling gene family gain and loss

We assume that the gain and loss of a gene family can be modelled by a two-state continuous-time Markov process, with states 0 (absence) and 1 (presence). In reality, the process is unlikely to be strictly Markovian. For example, the number of members of a gene family within a genome may also change due to duplication and deletion of segments of DNA. The loss of a gene family with many members may be less likely than the loss of a gene family with only one member in a genome. Models for the number of members of a gene family exist (e.g. Gu and Zhang, 2004; Spencer et al., 2006; Iwasaki and Takagi, 2007) but are computationally complex. We assume that gene families are independent, and that lineages evolve independently. We have discussed the justification for these assumptions

elsewhere (Spencer et al., 2006, 2007). It is important to remember that this model is only an approximation. For example, lateral transfer is an important source of gene gains in bacteria (Doolittle et al., 2003), and violates the strict assumption of lineage independence. However, if we have only sampled a small fraction of the lineages that exist at any time, it is not unreasonable to assume that the rate of gene gains in a given lineage does not depend strongly on other lineages that are present in our data.

For an individual gene family, the Markov process we have assumed is described by a rate matrix

$$\mathbf{Q} = \begin{bmatrix} -q_{01} & q_{01} \\ q_{10} & -q_{10} \end{bmatrix} \tag{1}$$

where q_{ij} is the instantaneous rate of transition from state i to state j. If $\mathbf{x} = [x_0, x_1]$ is a vector of the probabilities of gene family absence (x_0) and presence ($x_1 = 1 - x_0$), and the process is homogeneous in time ($\mathbf{Q}$ does not change) then

$$\frac{d\mathbf{x}}{dt} = \mathbf{x}\mathbf{Q}$$

with stationary probabilities $[\pi_Q(0), \pi_Q(1)]$ of absence and presence at equilibrium.

For a heterogeneous model with a rate shift in some gene families on edges leading only to parasites, we use two different $\mathbf{Q}$ matrices. $\mathbf{Q}_0$ is used on all edges in major category 0 (Figure a) and edges other than those leading only to parasite genomes in major category 1 (solid lines in Figure b). $\mathbf{Q}_1$ is used only on edges that lead only to parasite genomes in major category 1 (dashed lines in Figure b). For rate variation among categories of gene families within major category i, we assume that each rate class v has gain and loss rates $c_v\mathbf{Q}_{(ij)}$, with $\mathbf{Q}_{(ij)}$ being the rate matrix for edge j in major category i. The class-specific rate multiplier c_v is the mean rate for the vth equiprobable partition of a gamma distribution with mean 1 and shape parameter α (Yang, 1994).

The details of our models are in the Supplementary Material. The parameters we need to estimate are as follows. Each rate matrix i is described by its stationary probability of gene family absence, $\pi_Q(0, i)$ (we use this parameterization for compatibility with Cohen et al., 2008, even though our model is non-stationary). Each major category j has a root probability of gene family absence, $\pi_{\mathrm{ROOT}}(0, j)$, because we do not assume that the process of gain and loss is stationary. If we have two major categories, μ_0 is the probability that a gene family belongs to major category 0 (we do not consider models with more than two

major categories). Rate variation among gene families is described by the shape parameter α. To test the hypothesis that accelerated gene family loss affects some but not all gene families in parasites, we fit models with various restrictions on these parameters.

Model selection

We fit the following models:

A. The same relative gain and loss rates for all gene families and all edges. Figure a applies to all gene families. This is equivalent to model M2 in Cohen et al. (2008), except that we estimate edge lengths rather than assuming they are a constant multiple of substitution edge lengths.

B. The relative gain and loss rates are the same for all gene families, but change on edges leading only to parasites. Figure b applies to all gene families.

C. A mixture model in which the relative gain and loss rates change for some but not all gene families on the edges leading only to parasites. Major category 1 is the subset of dispensable gene families in parasites. Major category 0 contains all other gene families. Some gene families evolve according to Figure a, and some according to Figure b.

D. The same relative gain and loss rates for all edges, but some gene families are everywhere gained and lost faster than others. This is model A with gamma rate variation, and is equivalent to model M2+Γ in Cohen et al. (2008), except that we estimate edge lengths.

E. The relative gain and loss rates change for all gene families on edges leading only to parasites, and in addition some gene families are everywhere gained and lost faster than others. This is model B with gamma rate variation.

F. A mixture model in which the relative gain and loss rates change for some (major category 1) but not all gene families on the edges leading only to parasites, and in addition some gene families are everywhere gained and lost faster than others. This is model C with gamma rate variation.

Table 1 summarizes these models and their parameters. Previous work leads us to expect that models with gamma rate variation will fit better than models without (Cohen et al.,

2008; Hao and Golding, 2008). For comparing models with and without gamma rate variation (A with D, B with E and C with F), we use the likelihood ratio statistic

$$2\Lambda = 2(l_1 - l_0)$$

where l_1 is the log likelihood for the more complicated of a pair of nested models, and l_0 is the log likelihood for the simpler model of the pair. $2\Lambda \sim 0.5\chi_0^2 + 0.5\chi_1^2$ under the hypothesis of no gamma rate variation (Self and Liang, 1987; Ota et al., 2000).

Our main interest is in establishing whether there is a rate shift in edges leading to parasites, and whether this rate shift applies to all genes or only a subset. Standard asymptotic theory does not apply to likelihood ratios between mixture models with different numbers of components (e.g. comparing F with D or E), for two reasons. The null hypothesis (that one of the components has mixing probability zero) is on the boundary of the parameter space, and the parameters of the absent component are not identifiable under the null hypothesis (Chen et al., 2001). Modified likelihood ratio statistics have been developed whose null distribution is known for models with one parameter per component in addition to the mixing probabilities (Chen et al., 2001). These modified likelihood ratio methods have not yet been extended to models with multidimensional parameters in each component. Another possibility is an extension of the $C(\alpha)$ procedure (Lindsay, 1995, pp. 68-73), which tests a one-component model against any nearby alternative with more than one component. These methods would be worth pursuing in future. However, for this application we use the conceptually simple but computationally intensive parametric bootstrap method to estimate the null distribution of the likelihood ratio statistic (McLachlan, 1987). We generate K replicate simulated data sets under the simpler of a pair of nested models, fit both the simpler and the more complex models, and obtain the likelihood ratio statistic for each replicate (we use the true parameters for the simpler model as starting conditions). The p-value can be estimated as the proportion of bootstrap replicates having test statistics at least as large as the observed value.

Data

Our primary dataset is the set of all 50 bacterial genomes in the COG database (Tatusov et al., 1997). This contains 4873 gene families, of which 929 were discarded under the

assumption that observability requires presence in at least three genomes in the subset of data being analyzed (Supplementary Material). For these 50 genomes, we downloaded 16S rRNA sequence alignments from the Ribosomal Database Project release 9.57 (Cole et al., 2007). We then estimated a maximum likelihood tree using PHYML version 2.4.4 (Guindon and Gascuel, 2003) with the general time-reversible model of nucleotide evolution and four gamma rate categories. We rooted the tree between the firmicutes and all other bacteria. Deep branching of firmicutes is supported by analysis of 191 concatenated genes in taxa from across the tree of life (Ciccarelli et al., 2006), and is plausible (Koch, 2003) although not universally accepted.

We used Bergey's Manual of Systematic Bacteriology (Bergey, 1984; Garrity, 2005) to identify the following twelve taxa as parasites/endosymbionts with limited metabolic capabilities: *Mycobacterium leprae*, *Buchnera* sp. APS, *Rickettsia prowazekii*, *Rickettsia conorii*, *Chlamydia trachomatis*, *Chlamydia pneumoniae* CWL029, *Treponema pallidum*, *Borrelai burgdorferi*, *Ureaplasma urealyticum*, *Mycoplasma pulmonis*, *Mycoplasma pneumoniae*, *Mycoplasma genitalium* (Figure). Assignments were based on explicit reference in Bergey's Manual to an obligate intracellular lifestyle, failure or great difficulty to cultivate in tissue culture or artificial medium, or other evidence of reduced metabolic abilities.

Effects of tree topology

To explore the effects of tree topology, we also estimated the three models that included gamma rate variation (D, E, and F) on a tree topology based on a conditioned logdet method (Spencer et al., 2007) applied to the COG gene content data (Supplementary Material). In common with other topologies based on models of gene gain and loss, this topology contains the artefactual parasites clade (Figure). If our mixture model is a good enough representation of the process of gene family gain and loss, we would expect that this incorrect topology would have a worse log likelihood for model F, but a better log likelihood for models D and E, which do not account for heterotachy in parasite gene loss. Using a second topology also lets us see how sensitive our parameter estimates are to tree topology.

Results and Discussion

We first examine the results of fitting all six models on the 16S topology. In all cases, models with gamma rate variation have significantly better likelihoods than models without (Table 1, models A vs. D, B vs. E, C vs. F, likelihood ratio tests with null distribution $0.5\chi_0^2+0.5\chi_1^2$, $p < 1e-16$ in all cases). In conjunction with previous results (Cohen et al., 2008; Hao and Golding, 2008), this is strong evidence that there is variation among gene families in the rate of gain and loss.

Model E is significantly worse than model F (Table 1, $2\Lambda = 116$, with only two extra parameters in model F). Of 99 bootstrap replicates, none had test statistics greater than the observed value ($p < 0.01$; the largest bootstrapped likelihood ratio statistic was 7.6). These replicates included four cases in which model F was apparently trapped in a local optimum and had a substantially worse log likelihood than model E (by up to 4.6 log likelihood units), but re-estimating these cases from different starting conditions could not make any difference to our qualitative conclusion. Model D is also significantly worse than model F (Table 1, $2\Lambda = 3121$, with three extra parameters in model F). Of 99 bootstrap replicates, none had test statistics greater than the observed value ($p < 0.01$; the largest bootstrapped likelihood ratio statistic was 17.4). We therefore select model F. We have strong evidence that some but not all gene families have an accelerated loss rate in the edges leading to parasites, and that some gene families also have faster gain and loss rates everywhere than others. Unlike variation in genome size, this form of heterotachy is not handled by methods such as conditioned logdet, and may therefore be responsible for the failure of conditioned logdet methods to give a correct phylogeny (Spencer et al., 2007). One way to deal with this problem would be to partition gene families into categories using empirical Bayes (Supplementary Material), and calculate conditioned logdet distances separately for each category. Simulation studies showed that we can estimate parameters and assign categories with reasonable accuracy under model F, although very rare categories are problematic (Supplementary Material).

In model F, the ratio of gain to loss rates is 5.90 in gene families that are not dispensable in parasites, and on the non-parasite edges (thin edges coloured black in Figure) for dispensable genes. For dispensable genes on parasite edges (thick edges coloured red in Figure), the ratio of gain to loss rates falls to 0.08. This is a substantial change in genome dynamics. It is widely believed that genome reduction in parasites and endosymbionts affects some but

not all kinds of genes (Boussau et al., 2004). However, ours is the first study to test this hypothesis using a probabilistic model of genome dynamics.

The marginal probability of gene family absence at the root is $\sum_{v=1}^{C} \rho_v \pi_{\mathrm{ROOT}}(0, v) = 0.92$ for model F. This is consistent with the idea that the ancestral bacterium had a small rather than a large genome (although we do not know exactly how small, unless we can calculate the total number of gene families to which the marginal probability of gene family absence applies). Furthermore, the probability of absence at the root is much higher for gene families that are dispensable in parasites (93%, Table 1) than for those that are not dispensable in parasites (58%, Table 1). This makes biological sense. We would expect that many of the gene families that are not dispensable in parasites are required by all bacteria.

Our model D is the same as model M2+Γ in Cohen et al. (2008), except that we estimate each edge length, while Cohen et al. (2008) assume that edge lengths are proportional to protein substitution edge lengths, and estimate the constant of proportionality. However, their parameter estimates are quite different from ours. Biologically, their estimates imply a relatively large ancestral genome, with a tendency for the number of gene families to decrease towards the leaves of the tree (Table 2, row M2+Γ). Our estimates (both for model D and for the best-fitting model F) imply the opposite (except for the parasites in model F). Our results are not directly comparable because we used different subsets of the COG database and different tree topologies. To obtain a directly comparable result, we re-fitted model D using the parameters from Cohen et al. (2008) as a starting point. We used both our estimated edge lengths from model D (Table 2, row D^1) and the 16S substitution edge lengths (Table 2, row D^2). We found two local optima, both with the same qualitative pattern as Cohen et al. (2008), but with much worse log likelihoods than our original fit of model D. Non-stationary models for gene content data may be particularly prone to local optima, because loss from a large ancestral genome and gain from a small ancestral genome can generate similar distributions of genome size. However, unless there is another, much better, local optimum with a large ancestral genome which we have not located, there is strong evidence that the ancestral genome was small rather than large.

This agrees with the idea that the last universal common ancestor had a small genome (Koonin, 2003). Previous studies of ancestral genome size have used parsimony, which requires an ad-hoc choice of relative weights for gain and loss events (e.g. Snel et al., 2002;

Kunin and Ouzounis, 2003; Mirkin et al., 2003). Because different weights give different results, parsimony studies cannot provide a definitive answer, and the idea of a large universal common ancestor genome has supporters (Castresana, 2001). It has been suggested that Occam's razor should lead us to assume that ancestral genomes were of similar size to extant genomes (Dagan and Martin, 2007). We can then choose rules for gain and loss events that give ancestral genomes of the right size. Dagan and Martin (2007) used this approach to estimate the proportion of gene families that have experienced lateral gene transfer. Whether this assumption is plausible must depend on how far back in time we look. For example, it seems unlikely that the very first genome would have been of similar size to extant genomes. This is an interesting approach, and it might be possible to do something similar in a model-based method, by using informative priors on ancestral genome sizes. In summary, model-based studies of genome evolution have the potential to resolve the problem of ancestral genome size, but more remains to be done.

More complex models have larger values of the gamma shape parameter α, and thus less rate variation among gene families (Table 1, models D, E, and F). This is probably because rate variation over time in the more complex models can accommodate some of the patterns that would otherwise need to be explained by rate variation among gene families.

For model F, the mean ratio of estimated edge lengths under model F to 16S edge lengths (excluding two zero-length edges in the 16S tree) was 1.37 (standard deviation 4.28). Thus, the rate of gene family gain and loss is of the same order of magnitude as the rate of 16S substitutions. 16S edge lengths are not very strongly associated with gene family gain and loss edge lengths (correlation 0.41). A previous study found weaker correlations between substitution edge lengths and estimated numbers of gene family gains and losses (Boussau et al., 2004). However, their estimates were based on parsimony reconstructions of gains and losses, which will not be reliable if there were multiple gains and losses on some edges. Edges for which the expected ratio of gene family gains and losses to substitutions is very different from the mean may be of biological interest. For example, parsimony reconstructions suggest a high frequency of gene family loss relative to nucleotide substitutions on edges leading to intracellular bacteria (Boussau et al., 2004).

Contrary to our expectations, models D, E, and F all had better log likelihoods on the conditioned logdet tree, which has an artefactual parasites clade (Table 1). This probably

means that none of our models is a good enough representation of the changes in genome dynamics associated with a parasitic lifestyle. Parameter estimates for these three models with the conditioned logdet tree were qualitatively similar to those obtained from the 16S tree (Table 1), with the exception of the exception of a much lower root probability of absence in major category 0 for model F fitted to the conditioned logdet tree. This parameter may be hard to estimate because so few gene families are in this category. In general, it does not seem as though having the correct tree topology is essential for parameter estimation. The improvements in log likelihood when using the conditioned logdet tree over the 16S tree are large, but decrease with increasing model sophistication ($2\Lambda = 3330$ for model D, 1416 for model E, and 1314 for model F). Furthermore, the improvement in log likelihoods between models E and F is much smaller for the conditioned logdet tree ($2\Lambda = 14$) than for the 16S tree ($2\Lambda = 116$). This suggests that when estimated on the more plausible 16S tree, heterotachy is needed to account for the similarities in gene content among widely-separated parasites. On the conditioned logdet tree, heterotachy is much less important because similarities in gene content among parasites can be explained by recent common ancestry. Overall, we do not yet have an adequate model of gene family gain and loss, but our mixture model is a substantial improvement over earlier models.

Conclusion

The variable success of gene content phylogenies so far has been attributed to the lack of realistic models of genome dynamics (McInerney, 2006). Our work has provided strong evidence that changes in the relative rates of gain and loss occur for some but not all gene families in parasitic bacteria. Attempts to estimate phylogenies from gene content data are likely to fail unless they take account of this heterotachy. However, more sophisticated models may be needed in order to get the right topology from gene content data. We may simply need more different categories of gene families. Our software can estimate much more complex models than the ones we considered here, but the structure of these models should be informed by biology, and we need to show that their parameters can be reliably estimated from real data. We might also need to further relax our assumptions about rates of gene gain and loss. Our current models are non-stationary, but all parasite lineages

have one equilibrium genome size, and all non-parasites have another. Conditioned logdet methods were developed to avoid restrictive assumptions about equilibrium genome size (Lake and Rivera, 2004), but are not able to deal with rate variation among gene families. Generalizations of covarion models (Whelan, 2008) with switching between several classes of equilibrium genome size might be useful. There has been some progress in combining models of gene gain and loss with models of sequence evolution (Arvestad et al., 2004). As yet, these models have only a simplistic gain and loss component, but in future, they may be useful for topology estimation.

Alternatively, there may be relatively little information about the ancestry of parasites in gene family presence/absence data (Ed Susko, personal communication). Only a small proportion of gene families are present in parasites, and absences are not very informative, because most gene families are absent from most genomes. Among those that are present in parasites, there are many gene families in common, so that grouping the parasites together will always be a fairly good explanation for the observed pattern.

Models of gene family gain and loss may be more useful for understanding genome dynamics than for topology estimation, especially as most parameter estimates do not seem not very sensitive to tree topology. Potential applications of the models we have described include modelling the evolution of plastid genomes, identifying gene families with unusual gain and loss patterns, and probabilistic reconstructions of ancestral genomes. Probabilistic models are important because existing studies of genome dynamics often depend on parsimony with ad-hoc choices of gain and loss penalties. There are two major problems with the parsimony approach. First, when there are large numbers of gene family gains and losses (which seems likely, given that our estimates suggest rates comparable to those of nucleotide substitutions), parsimony reconstructions are unlikely to be reliable. Second, without a probabilistic model of evolution, it is difficult to test hypotheses about how evolution works. The modelling framework that we and others (e.g. Cohen et al., 2008; Hao and Golding, 2008) have used addresses both of these problems.

Supplementary Material

Details of methods and simulation results.

Acknowledgments

This work was funded by BBSRC grant BB/E019005/1 to MS. Part of this work was carried out during the phylogenetics programme at the Isaac Newton Institute for Mathematical Sciences, Cambridge. We are grateful to David Bryant, Miklós Csűrös, Brian Golding, Weilong Hao, Tal Pupko, Ed Susko and especially Ofir Cohen for ideas, test data sets and help with code development, Ian Smith for computing support, and an anonymous referee for comments that improved the manuscript. Author contributions: MS wrote the software and drafted the manuscript; AS did the data analysis.

Literature Cited

Arvestad, L., A.-C. Berglund, J. Lagergren, and B. Sennblad. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In: Proceedings of the Eighth International Conference on Computational Molecular Biology. ACM Press, New York, 326–335.

Bergey, D. H., editor. 1984. Bergey's Manual of Systematic Bacteriology, volume 1-3. Baltimore: Williams and Watkins, first edition.

Boussau, B., E. O. Karlberg, A. C. Frank, B.-A. Legault, and S. G. E. Andersson. 2004. Computational inference of scenarios for α-proteobacterial genome evolution. Proceedings of the National Academy of Sciences of the United States of America **101**:9722–9727.

Castresana, J. 2001. Comparative genomics and bioenergetics. Biochimica et Biophysica Acta **1506**:147–162.

Chen, H., J. Chen, and J. D. Kalbfleisch. 2001. A modified likelihood ratio test for homogeneity in finite mixture models. Journal of the Royal Statistical Society Series B **63**:19–29.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311**:1283–1287.

Cohen, O., N. D. Rubinstein, A. Stern, U. Gophna, and T. Pupko. 2008. A likelihood

framework to analyze phyletic patterns. Philosophical Transactions of the Royal Society of London Series B **363**:3903–3911.

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.

Dagan, T., and W. Martin. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proceedings of the National Academy of Sciences of the United States of America **104**:870–875.

Doolittle, W. F., Y. Boucher, C. L. Nesbø, C. J. Douady, J. O. Andersson, and A. J. Roger. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philosophical Transactions of the Royal Society of London Series B: Biological Sciences **358**:39–58.

Garrity, G. M., editor. 2005. Bergey's Manual of Systematic Bacteriology, volume 2. New York: Springer, second edition.

Graur, D., and W.-H. Li. 2000. Fundamentals of Molecular Evolution. Massachusetts: Sinauer, second edition.

Gu, X., and H. Zhang. 2004. Genome phylogenetic analysis based on extended gene contents. Molecular Biology and Evolution **21**:1401–1408.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology **52**:696–704.

Hao, W., and G. B. Golding. 2006. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. Genome Research **16**:636–643.

———. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. BMC Genomics **9**:235.

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.

Iwasaki, W., and T. Takagi. 2007. Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics **23**:i230–i239.

Koch, A. L. 2003. Were Gram-positive rods the first bacteria? Trends in Microbiology **11**:166–170.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**:980–984.

Koonin, E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nature Reviews Microbiology **1**:127–136.

Kunin, V., and C. A. Ouzounis. 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Research **13**:1589–1594.

Lake, J. A., and M. C. Rivera. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. Molecular Biology and Evolution **21**:681–690.

Lindsay, B. G. 1995. Mixture models: theory, geometry and applications. Hayward, California: Institute of Mathematical Statistics.

McInerney, J. O. 2006. On the desirability of models for inferring genome phylogenies. Trends in Microbiology **14**:1–2.

McLachlan, G. J. 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied Statistics **36**:318–324.

Mira, A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. Trends in Genetics **17**:589–596.

Mirkin, B. G., T. I. Fenner, M. Y. Galperin, and E. V. Koonin. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evolutionary Biology **3**:2.

Ochman, H., and N. A. Moran. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science **292**:1096–1099.

Ota, R., P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Molecular Biology and Evolution **17**:798–803.

Self, S. G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association **82**:605–610.

Snel, B., P. Bork, and M. A. Huynen. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Research **12**:17–25.

Snel, B., M. A. Huynen, and B. E. Dutilh. 2005. Genome trees and the nature of genome evolution. Annual Review of Microbiology **59**:191–209.

Spencer, M., D. Bryant, and E. Susko. 2007. Conditioned genome reconstruction: how to avoid choosing the conditioning genome. Systematic Biology **56**:25–43.

Spencer, M., E. Susko, and A. J. Roger. 2006. Modelling prokaryote gene content. Evolutionary Bioinformatics Online **2**:165–186.

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

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Molecular Biology and Evolution **25**:1683–1694.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution **39**:306–314.

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	-71191	2	0.44 (0.004)[i]	-	0.89 (0.005)	-	-	-
	B	-67238	3	0.34 (0.003)	0.95 (0.001)	-	0.82 (0.006)	-	-
	C	-66560	5	0.35 (0.003)	0.96 (0.001)	0.19 (0.058)	0.89 (0.006)	0.09 (0.007)	-
	D	-66359	3	0.04 (0.001)	-	0.97 (0.003)	-	-	0.51 (0.009)
	E	-64857	4	0.14 (0.003)	0.91 (0.003)	-	0.92 (0.004)	-	0.70 (0.023)
	F	-64799	6	0.14 (0.003)	0.93 (0.003)	0.58 (0.062)	0.93 (0.005)	0.04 (0.005)	0.73 (0.026)
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.

Table 2: Local optima for model D.

Model[a]	l[b]	$\pi_Q(0,0)$[c]	$\pi_{\mathrm{ROOT}}(0,0)$[d]	α[e]
M2+Γ	-	0.87	0.15	0.64
D^1	-70714	0.86	0.27	0.75
D^2	-66594	0.92	0.30	0.76
D	-66359	0.04	0.97	0.68

[a]Each row is a local optimum for model D, obtained from different starting conditions. The row M2+Γ contains the parameter estimates from Cohen et al. (2008). D^1 used the parameters from M2+Γ and the edge lengths we obtained for model D (table 1) as starting conditions. D^2 is the same, except that 16S edge lengths were used as starting conditions. D is our original fit of model D.

[b]Log likelihood.

[c]Stationary probability of absence in rate matrix 0.

[d]Probability of absence at the root in category 0.

[e]Shape parameter for gamma rate variation.

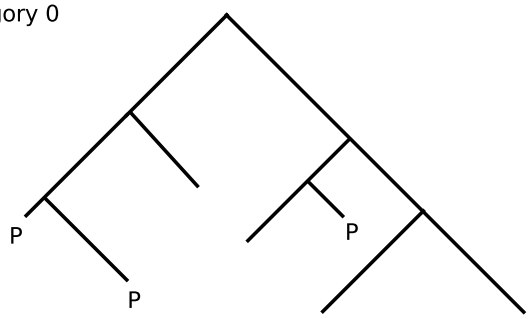
Figure legends

Figure 1. Mixture model for gene loss in parasites. Parasites are leaves labelled P on a rooted tree. We assume that there are two major categories of genes: (a) those that have the same relative rates of gain and loss ($\mathbf{Q}_0$) everywhere; and (b) those that have a different set of gain/loss parameters ($\mathbf{Q}_1$) on the dashed edges leading exclusively to parasitic genomes. The edge lengths are the same for both categories.

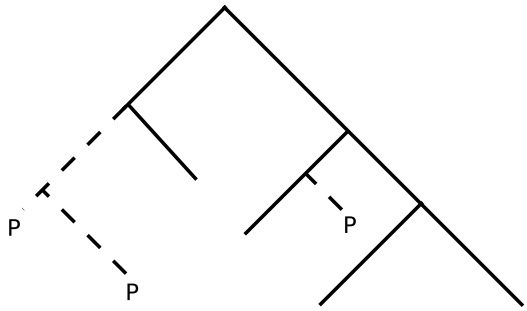
Figure 2. Estimated edge lengths for the 16S topology under model F (Table 1) for the COG data. Edges leading only to parasites/endosymbionts are shown as thicker red lines. Parasites/endosymbionts are indicated by circles and bold genome names. Edge lengths are in expected numbers of gains and losses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

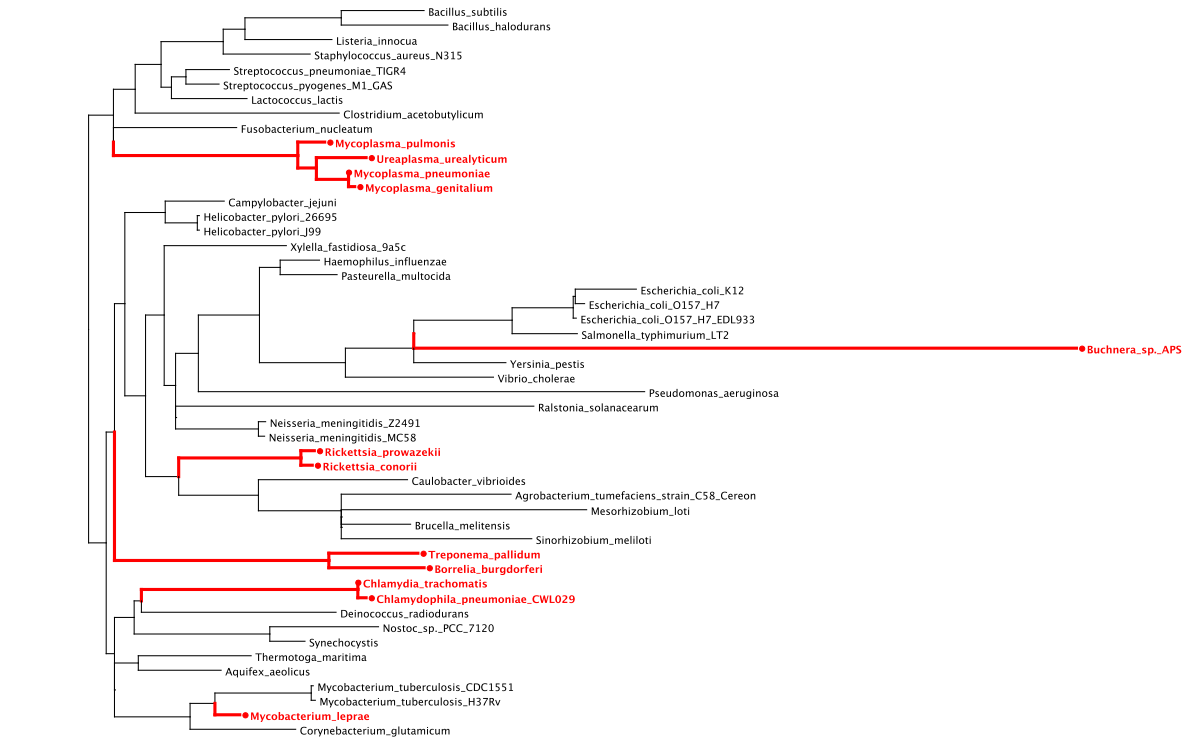
Figure 3. Estimated edge lengths for the conditioned logdet topology under model F (Table 1) for the COG data. Edges leading only to parasites/endosymbionts are shown as thicker red lines. Parasites/endosymbionts are indicated by circles and bold genome names. Edge lengths are in expected numbers of gains and losses per gene family at stationarity. Tree drawn using Dendroscope (Huson et al., 2007).

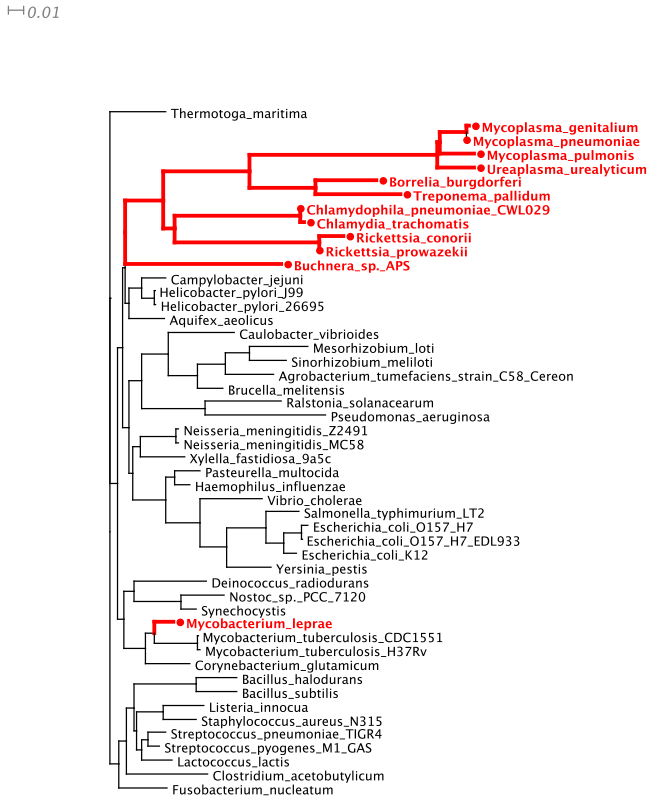
a. Category 0



b. Category 1, dispensable in parasites




0.01
Bacillus_subtilis
Bacillus_halodurans
Listeria_innocua
Staphylococcus_aureus_N315
Streptococcus_pneumoniae_TIGR4
Streptococcus_pyogenes_M1_GAS
Lactococcus_lactis
Clostridium_acetobutylicum
Fusobacterium_nucleatum
Mycoplasma_pulmonis
Ureaplasma_urealyticum
Mycoplasma_pneumoniae
Mycoplasma_genitalium
Campylobacter_jejuni
Helicobacter_pylori_26695
Helicobacter_pylori_J99
Xylella_fastidiosa_9a5c
Haemophilus_influenzae
Pasteurella_multocida
Escherichia_coli_K12
Escherichia_coli_O157_H7
Escherichia_coli_O157_H7_EDL933
Salmonella_typhimurium_LT2
Buchnera_sp._APS
Yersinia_pestis
Vibrio_cholerae
Pseudomonas_aeruginosa
Ralstonia_solanacearum
Neisseria_meningitidis_Z2491
Neisseria_meningitidis_MC58
Rickettsia_prowazekii
Rickettsia_conorii
Caulobacter_vibrioides
Agrobacterium_tumefaciens_strain_C58_Cereon
Mesorhizobium_loti
Brucella_melitensis
Sinorhizobium_meliloti
Treponema_pallidum
Borrelia_burgdorferi
Chlamydia_trachomatis
Chlamydophila_pneumoniae_CWL029
Deinococcus_radiodurans
Nostoc_sp._PCC_7120
Synechocystis
Thermotoga_maritima
Aquifex_aeolicus
Mycobacterium_tuberculosis_CDC1551
Mycobacterium_tuberculosis_H37Rv
Mycobacterium_leprae
Corynebacterium_glutamicum


0.01
Thermotoga_maritima
Mycoplasma_genitalium
Mycoplasma_pneumoniae
Mycoplasma_pulmonis
Ureaplasma_urealyticum
Borrelia_burgdorferi
Treponema_pallidum
Chlamydophila_pneumoniae_CWL029
Chlamydia_trachomatis
Rickettsia_conorii
Rickettsia_prowazekii
Buchnera_sp._APS
Campylobacter_jejuni
Helicobacter_pylori_J99
Helicobacter_pylori_26695
Aquifex_aeolicus
Caulobacter_vibrioides
Mesorhizobium_loti
Sinorhizobium_meliloti
Agrobacterium_tumefaciens_strain_C58_Cereon
Brucella_melitensis
Ralstonia_solanacearum
Pseudomonas_aeruginosa
Neisseria_meningitidis_Z2491
Neisseria_meningitidis_MC58
Xylella_fastidiosa_9a5c
Pasteurella_multocida
Haemophilus_influenzae
Vibrio_cholerae
Salmonella_typhimurium_LT2
Escherichia_coli_O157_H7
Escherichia_coli_O157_H7_EDL933
Escherichia_coli_K12
Yersinia_pestis
Deinococcus_radiodurans
Nostoc_sp._PCC_7120
Synechocystis
Mycobacterium_leprae
Mycobacterium_tuberculosis_CDC1551
Mycobacterium_tuberculosis_H37Rv
Corynebacterium_glutamicum
Bacillus_halodurans
Bacillus_subtilis
Listeria_innocua
Staphylococcus_aureus_N315
Streptococcus_pneumoniae_TIGR4
Streptococcus_pyogenes_M1_GAS
Lactococcus_lactis
Clostridium_acetobutylicum
Fusobacterium_nucleatum