A phylogenetic mixture model for gene family loss in parasitic bacteria

Supplementary Material

Matthew Spencer and Ajanthah Sangaralingam

School of Biological Sciences, University of Liverpool, UK

Appendix

Edge lengths and rate matrix scaling

For a continuous-time Markov process with a rate matrix such as

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

the transition probabilities $\mathbf{P}(t)$ over any non-negative time t (the length of an edge on the tree) are in general given by the matrix exponential

$$P(t) = e^{\mathbf{Q}t} \tag{1}$$

(Norris, 1997, Theorem 2.1.1). For two-state models, it is straightforward to calculate the transition probabilities analytically (Kijima, 1997, pp. 177-178). However, Equation 1 makes it clear that in general, $\mathbf{Q}$ and t are not identifiable. Multiplying $\mathbf{Q}$ by a constant c and dividing t by the same constant does not change the transition probabilities. It is therefore conventional to multiply $\mathbf{Q}$ by a constant so that the expected number of events per unit time at stationarity is 1, in other words

$$-\sum_{i=1}^{s} \pi_Q(i) q_{ii} = 1 \tag{2}$$

where the summation is over all s possible states (in this case, absence and presence). For a two-state model, there is one free parameter, which can be expressed as $\pi_Q(0) = q_{10}/(q_{01} + q_{10})$. Edge lengths are then the expected number of events in a stationary model. In a nonstationary model the edge lengths do not have this interpretation (Yang and Roberts, 1995), but the scaling is convenient for identifiability. We therefore scale all rate matrices using Equation 2, so that there is a common set of edge lengths for all categories. More general models could allow different edge lengths for different categories of genes, at the expense of more parameters.

Likelihood calculation

Consider an edge such as *bc* in Figure 1. Deleting *bc* divides the tree into two subtrees, an upper subtree containing the root, and a lower subtree. We define the lower subtree conditional likelihood $L^{(j)}_{i,lower,bc,v}$ as the likelihood under evolutionary category v for gene

family i on the lower subtree for edge bc, conditional on state j at c. Similarly, the upper subtree conditional likelihood $L^{(j)}_{i,upper,bc,v}$ is the likelihood under evolutionary category v for gene family i on the upper subtree at bc, conditional on state j at b (Boussau and Gouy, 2006).

In the standard pruning algorithm (Felsenstein, 1981), we initialize the lower subtree conditional likelihoods $L^{(j)}_{i,lower,kl,v}$ on terminal edges kl. For gene family i under evolutionary category v, conditional on state j at leaf k,

$$L^{(j)}_{i,lower,kl,v} = \begin{cases} 1 & x_{ik} = j \\ 0 & x_{ik} \neq j \end{cases}$$

where x_{ik} is the presence/absence state of gene family i in the genome at leaf k. We can then traverse the tree in postorder to calculate the lower subtree conditional likelihoods for internal edges. For a vertex b with descendants c and d (Figure 1)

$$L^{(j)}_{i,lower,ab,v} = \sum_{k=1}^{s} p_{jk,v}(t_{bc}) L^{(k)}_{i,lower,bc,v} \sum_{h=1}^{s} p_{jh,v}(t_{bd}) L^{(h)}_{i,lower,bd,v} \tag{3}$$

where the summations are over the s possible states and $p_{ij,v}(t)$ is the transition probability from state i to state j in time t under evolutionary category v.

At the root (for example, vertex a in Figure 1, with immediate descendants b and e), we sum over states weighted by their probabilities at the root to get the full gene family likelihood $L_{i,v}$ for category v:

$$L_{i,v} = \sum_{j=1}^{s} \pi_{j,v} \sum_{k=1}^{s} p_{jk,v}(t_{ab}) L^{(k)}_{i,lower,ab,v} \sum_{h=1}^{s} p_{jh,v}(t_{ae}) L^{(h)}_{i,lower,ae,v} \tag{4}$$

In a stationary model, $\pi_{j,v}$ is the stationary probability of state j in evolutionary category v. We can reroot the tree at any vertex and calculate the likelihood using Equation 4, substituting in the appropriate edge and vertex labels. Because $p_{jk,v}(t_{ab})$ is the only term in Equation 4 that depends on t_{ab}, we do not need to recalculate the lower subtree likelihoods in order to optimize t_{ab} (or any other edge, once we have rerooted at a vertex at one end of the edge).

If the model is not stationary (as ours is not, because the rate matrices on different edges may not have the same stationary probabilities), we cannot use the approach described above to optimize edge lengths. Instead, Boussau and Gouy (2006) showed that we can calculate

the likelihood using the lower and upper subtree conditional likelihoods on any edge, for example

$$L_i, v = \sum_{j=1}^{s} L^{(j)}_{i,upper,bc,v} \sum_{k=1}^{s} p_{jk,v}(t_{bc}) L^{(k)}_{i,lower,bc,v} \tag{5}$$

The upper subtree conditional likelihoods can be calculated recursively. For example,

$$L^{(j)}_{i,upper,bc,v} = \sum_{h=1}^{s} p_{hj,v}(t_{ab}) L^{(h)}_{i,upper,ab,v} z^{(j)}_{bd} \tag{6}$$

where

$$z^{(j)}_{bd} = \sum_{m=1}^{s} p_{jm,v}(t_{bd}) L^{(m)}_{i,lower,bd,v} \tag{7}$$

The upper subtree conditional likelihoods at the root are needed to initialize the recursion, for example,

$$L^{(j)}_{i,upper,ab,v} = \pi_{\mathrm{ROOT}}(j, v) \sum_{k=1}^{s} p_{jk,v}(t_{ae}) L^{(k)}_{i,lower,ae,v} \tag{8}$$

where the root probabilities $\pi_{\mathrm{ROOT}}(j, v)$ of each state j in each category v are parameters (Boussau and Gouy, 2006). Thus after a postorder traverse down to the root in which we calculate the lower subtree conditional likelihoods, we can do a preorder traverse, during which we calculate the upper subtree conditional likelihoods. Since $p_{jk,v}(t_{bc})$ is the only term in Equation 5 that depends on t_{bc}, we can optimize t_{bc} without recalculating the upper and lower subtree conditional likelihoods. Identifiability of the edge lengths adjacent to the root depends on the arrangement of rate matrices. For example, if in all categories, the rate matrix is the same for the left and right edges at the root, and the state probabilities at the root are the stationary probabilities in each category, then only the sum of these edge lengths will be identifiable.

To obtain the full likelihood L_i for a gene family, we sum up the likelihoods for each category v for that gene family, weighted by the category probabilities ρ_v,

$$L_i = \sum_{v=1}^{C} \rho_v L_{i,v} \tag{9}$$

where we have C categories, and $\sum_{v=1}^{C} \rho_v = 1$. We parameterize the category probabilities as follows. Each category v corresponds to major category u ($u \in \{0, 1\}$ for the models we consider here) and rate class $\nu \in \{0 \ldots k-1\}$. Let μ_u be the probability of major category u (which we estimate, subject to the constraint $\sum \mu_u = 1$). Within each major category, we have k equiprobable rate classes, with rate multipliers obtained from a discrete gamma distribution. Then $\rho_v = \mu_u / k$.

Finally, under the assumption that all gene families are independent, the log likelihood l for all n gene families is

$$l = \sum_{i=1}^{n} \log L_i \tag{10}$$

Conditioning on observability

If some patterns are not observable, we should calculate the gene family likelihood conditional on a pattern being observable, L_i^+.

$$L_i^+ = \frac{L_i}{1 - L_i^-} \tag{11}$$

where L_i^- is the likelihood of unobservable patterns for gene family i (Felsenstein, 1992). If we have multiple categories, substituting Equation 9 into Equation 11 gives

$$\begin{aligned} L_i^+ &= \frac{L_i}{1 - L_i^-} \\ &= \frac{\sum_{v=1}^{C} \rho_v L_{i,v}}{1 - \sum_{v=1}^{C} \rho_v L_{i,v}^-} \end{aligned} \tag{12}$$

where $L_{i,v}^-$ is the likelihood of unobservable patterns for gene family i conditional on category v. Similar conditional likelihood calculations are implemented in the intron gain and loss model in Csűrös et al. (2008) and the restriction site model in MrBayes (Ronquist and Huelsenbeck, 2003). If we have n independent gene families, and the likelihood of unobservable patterns is the same for all gene families, then the conditional log likelihood l^+ is

$$l^+ = \sum_{i=1}^{n} \log L_i - n \log(1 - L_i^-)$$

(Felsenstein, 1992).

It is common to assume that L_i^- is the likelihood of the pattern with absence at all leaves (e.g. Zhang and Gu, 2004; Hao and Golding, 2006; Cohen et al., 2008; Hao and Golding, 2008). However, this is not always exactly right. For example, in the COG database, orthologs are identified by three-way patterns of sequence similarity among genomes (Tatusov et al., 1997), so a gene family does not appear in the COG database unless it occurs in at least three genomes. For cases where a number of patterns are unobservable, each such pattern is a disjoint event, so we sum over unobservable patterns to get L_i^-:

$$L_i^- = \sum_{j \in \mathcal{U}} L^{j-}$$

where $\mathcal{U}$ is the set of unobservable patterns, and L^{j-} is the likelihood of the jth unobservable pattern. There is an additional complication if we are working with only a subset of the genomes used to construct a database. Some patterns are observable in the subset that would not be observable if extra genomes had not been used to construct the database. We have not attempted to calculated the correct conditional likelihoods for all gene families in these cases. Instead, we discard any gene family not meeting the observability criteria (e.g. presence in at least three genomes for COG) within the subset. The result is correct conditional likelihoods for those gene families that we use, at the cost of throwing away some data. For simplicity we drop the notation indicating whether we are working with likelihoods conditional on observability in the following sections, as this has no effect on the algorithms.

Parameter estimation

We estimate edge lengths one at a time using a golden section method (Press et al., 1992, section 10.1), which is simple and robust. We optimize first the left and then the right descendant edge at each vertex in preorder, using the decomposition in Equation 5 and the algorithm in Figure 2. After changing the descendant edges at a vertex, we need to recalculate the upper subtree likelihoods for the entire tree. This means that optimizing the pair of descendant edges at each vertex requires a full postorder traversal and a partial preorder traversal. There is scope for a more efficient algorithm, but the current implementation is fast enough for our needs.

We use either a Nelder-Mead simplex method (Nelder and Mead, 1965) or a BFGS quasi-Newton method (Nocedal and Wright, 1999, section 8.1) to estimate the other parameters (both implementations from the Gnu Scientific Library version 1.8-2, `http://www.gnu.org/software/gsl/`). The results we report here are from the BFGS method. In test cases, we obtained similar parameter estimates from both methods, but the simplex method gave slightly worse likelihoods. The default starting conditions were the edge lengths from the 16S tree, stationary probabilities of absence 0.5, root probabilities of absence 0.5, mixing probability of major category 0 0.5 (where applicable) and gamma shape parameter 1 (where applicable). In most models there was evidence of local optima, so we used the best result from multiple starting conditions (the default values, starting conditions close to the current estimates from the next most complex model, and starting conditions close to the estimates

from the best current model).

In both cases, we use transformations to turn the constrained optimization into an unconstrained problem. We log-transform the relative gene loss rates $q_{10,v}/q_{01,v}$ and the shape parameter α to ensure that they remain positive. We use a logistic transformation of the root probabilities of state 0, $\pi_{\mathrm{ROOT}}(0, v)$ to keep them within $[0, 1]$. We reduce the dependent, constrained mixing probabilities for the M major categories to a set of $M - 1$ independent, unconstrained parameters $\zeta_v = \log(\mu_v/\mu_M)$, $v = 1 \ldots M - 1$ (Bickel and Doksum, 2001, p. 55). We alternate between edge length and other parameter optimization until the log likelihood converges.

We report approximate standard errors for the parameters other than edge lengths, obtained by treating the edge lengths as fixed. This will underestimate the uncertainty in the other parameters, but including the large number of edge lengths would probably lead to numerical instability. We obtain a numerical estimate $\mathbf{G}$ of the Hessian matrix of second derivatives of the log likelihood with respect to the transformed parameters. Then the Hessian $\mathbf{H}$ for the untransformed parameters is $\mathbf{H} = \mathbf{J}'\mathbf{G}\mathbf{J}$, where $\mathbf{J}$ is the Jacobian matrix of derivatives of transformed parameters with respect to untransformed parameters (Christensen et al., 2008). The covariance matrix $\mathbf{\Sigma}$ for the untransformed parameters is estimated as $\mathbf{\Sigma} = -\mathbf{H}^{-1}$ (Bickel and Doksum, 2001, p. 386), and the standard errors are the square roots of the diagonal elements of $\mathbf{\Sigma}$.

Empirical Bayes estimation of posterior category probabilities

We use an empirical Bayes method (Garthwaite et al., 2002, section 7.8) to estimate the posterior probability $P(c_i = v|\mathbf{D})$ that the category c_i to which a gene family i belongs is v, conditional on the data $\mathbf{D}$. Given the maximum likelihood estimates $\hat{\rho}_j$ of the mixing probabilities for each category j

$$P(c_i = v|\mathbf{D}) = \frac{\hat{\rho}_v L_{i,v}}{\sum_{j=1}^{C} \hat{\rho}_j L_{i,j}}$$

where the likelihoods are maximized with respect to all the model parameters. Our estimate of the category to which a gene family belongs is the category with maximum posterior probability. We are ignoring the uncertainty in the $\hat{\rho}_j$. The same method is used to identify sites under positive selection in codon models (Nielsen and Yang, 1998).

Implementation

We implemented the models described above in C, together with a simulator. The source code is available from `http://www.liv.ac.uk/~matts/genecontent.html` under the Gnu Public License, along with data and examples. The software allows arbitrarily complex specification of major categories and patterns of rate change across the tree within each major category, although we have not investigated the performance of models more complicated than the ones described here.

Our results were obtained on a 3GHz Intel Xeon processor with 4G RAM running 32-bit Debian Etch linux (kernel 2.6.18-6-686). For the COG data, parameter estimation took 2-293 minutes, depending on the model. For parametric bootstrap tests and the simulation study described below, we compiled our code under Cygwin (`http://www.cygwin.com/`) and used a Condor high throughput computing system (`http://www.cs.wisc.edu/condor/`) to distribute replicates among idle PCs running Windows XP at the University of Liverpool.

Simulation study

We also did a simulation study to see whether we could accurately estimate parameters and assign gene families to categories under the most complex model, model F. We generated 100 data sets, each containing 3944 gene families present in at least three genomes (as in the real COG data). We used the estimated parameters and edge lengths for model F fitted to the COG data set (main text, table 1). For each simulated data set, we estimated parameters under model F using the same settings as were used for the real data, except that we only used the default starting conditions.

With the exception of the root probability of absence in major category zero and the gamma shape parameter, the mean estimated parameters from simulated data sets were close to their true values (Table 1). However, the true values of all the parameters other than the stationary probability of absence in rate matrix 0 and the shape parameter lay outside the approximate 95% confidence intervals from the simulation study (mean $\pm$ 2 standard errors), suggesting small-sample biases in the parameter estimates. The only parameter for which this bias may be relatively large is the root probability of absence in major category 0. It may be difficult to estimate this parameter because there are relatively few genes in this category.

For each replicate, we also have an estimate of the standard error for each parameter obtained by inverting the Hessian. We would expect these standard errors to underestimate the true uncertainty, because they ignore uncertainty in edge lengths. The worst case was for the stationary probability of absence in major category zero, where the mean of the standard errors obtained by inverting the Hessian was only 0.07 times the standard error obtained by direct calculation from the distribution of estimates over 100 replicate data sets. In all other cases, the mean of the standard errors obtained by inverting the Hessian was at least 0.46 times the standard error obtained by direct calculation. Overall, standard errors that ignore uncertainty in edge lengths should be interpreted very cautiously.

Using the default starting conditions, the edges leading to *Buchnera sp.* APS (in all replicates) and *Ralstonia solanacearum* (in 10 out of 100 replicates) were estimated at 100 (the largest allowed edge length), suggesting problems with local optima. In the real data, we observed similar problems for some data sets when starting from the 16S edge lengths, but not when starting from other sets of edge lengths (and not at all in the final results reported in the main text, Table 1). Excluding these two edges in replicates for which these problems occurred, the relationship between true and estimated edge lengths had intercept 0.0001 (standard error 0.0001), slope 1.10 (standard error 0.003) and $R^2 = 0.92$. Thus, edge lengths were in general estimated accurately, although there was a tendency for the lengths of long edges to be overestimated. Again, this may be due to finding local rather than global optima. When we ran the same kind of simulation but started the optimization with edge lengths at their true values, the relationship between true and estimated edge lengths had intercept -8.51×10^{-5} (standard error 4.87×10^{-5}) and intercept 1.001 (standard error 9.29×10^{-4}), and there were no estimated edge lengths of 100.

Over all simulated data sets, the maximum posterior probability category was the true category for 71% of gene families. Category assignments were least reliable for rate classes 1 and 2 in major category 0, and most reliable for major category 1, which made up most of the observable gene families (table 2). Identification of major categories was reliable, with 95% of gene families assigned to the correct major category. This is partly because almost all observable gene families were in major category 1 (dispensable in parasites), and almost all of these were assigned to major category 1 (table 3). Although 66% of the observable gene families that were not dispensable in parasites (major category 0) were correctly assigned, the

specificity of assignment to major category 0 was low. Only 50% of gene families that were assigned to this category truly belonged to it. In summary, we found acceptable accuracy of category assignments, and good accuracy of major category assignments.

Conditioned logdet tree estimation

We estimated conditioned logdet distances from the COG gene content data (Tatusov et al., 1997), for each possible choice of conditioning genome (Spencer et al., 2007). The maximum likelihood estimates of some pairwise distances were non-existent, when the determinant of the pattern probability matrix was zero or negative. For these distances, we used a constrained maximum likelihood and pseudocount method (A. Sangaralingam et al., in preparation). This was implemented as the `-eI` option in the program `cond_logdet 0.3`, available from `http://www.liv.ac.uk/~matts/genecontent.html` under the Gnu Public License. We combined distances from each choice of conditioning genome using a modified version of BIONJ (Gascuel, 1997) and inverse variance weighting (Spencer et al., 2007). This produces an unrooted supertree without edge lengths. We therefore initialized all the edge lengths for this tree at the mean of the edge lengths estimated under model F on the 16S topology, and rooted the tree between firmicutes and all other bacteria, as for the 16S tree. We used the default starting conditions when estimating parameters on this tree.

Literature Cited

Bickel, P. J., and K. A. Doksum. 2001. Mathematical Statistics: Basic Ideas and Selected Topics, volume 1. New Jersey: Prentice-Hall, second edition.

Boussau, B., and M. Gouy. 2006. Efficient likelihood computations with nonreversible models of evolution. Systematic Biology **55**:756–768.

Christensen, T. M., A. S. Hurn, and K. A. Lindsay. 2008. The devil is in the detail: hints for practical optimisation. Economic Analysis and Policy **38**:345–368.

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.

Csűrös, M., I. B. Rogozin, and E. V. Koonin. 2008. Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. Molecular Biology and Evolution **25**:903–911.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution **17**:368–376.

———. 1992. Phylogenies from restriction sites: a maximum-likelihood approach. Evolution **46**:159–173.

Garthwaite, P. H., I. T. Jolliffe, and B. Jones. 2002. Statistical Inference. Oxford: Oxford University Press, second edition.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution **14**:685–695.

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.

Kijima, M. 1997. Markov Processes for Stochastic Modeling. London: Chapman and Hall.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. Computer Journal **7**:308–313.

Nielsen, R., and Z. Yang. 1998. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. Genetics **148**:929–936.

Nocedal, J., and S. J. Wright. 1999. Numerical Optimization. New York: Springer.

Norris, J. R. 1997. Markov Chains. Cambridge, England: Cambridge University Press.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. Numerical Recipes in C: the Art of Scientific Computing. Cambridge, England: Cambridge University Press, second edition.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572–1574.

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

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

Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Molecular Biology and Evolution **12**:451–458.

Zhang, H., and X. Gu. 2004. Maximum likelihood for genome phylogeny on gene content. Statistical Applications in Genetics and Molecular Biology **3**:article 31.

Table 1: Means and approximate 95% confidence intervals for 100 replicate sets of parameters estimated under model F from simulated data sets generated under model F.

Parameter	True value[e]	Mean	95% CI[f]
$\pi_Q(0,0)$[a]	0.14	0.15	$[0.08, 0.22]$
$\pi_Q(0,1)$	0.93	0.92	$[0.90, 0.93]$
$\pi_{\mathrm{ROOT}}(0,0)$[b]	0.58	0.33	$[0.22, 0.44]$
$\pi_{\mathrm{ROOT}}(0,1)$	0.93	0.95	$[0.94, 0.96]$
μ_0[c]	0.04	0.06	$[0.04, 0.07]$
α[d]	0.73	0.83	$[0.72, 0.93]$

[a] $\pi_Q(0,i)$ 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.

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

[c] Mixing probability for major category 0.

[d] Shape parameter for gamma rate variation.

[e] Estimated from the COG data.

[f] Mean $\pm$ two standard errors.

Table 2: True (rows) and estimated (columns, estimated by maximum posterior probability) categories for 100 replicate simulations, each of 3944 observable gene families generated on the tree in the main text, figure 2, with parameters from model F fitted to the COG data (main text, table 1).

True category			Estimated category									
Major category[a]	Rate class[b]	Category	0	1	2	3	4	5	6	7	Proportion in true category[c]	Proportion correct[d]
0	0	0	2060	147	124	2	321	329	2	0	0.007	0.69
0	1	1	1228	510	573	121	251	1729	423	4	0.01	0.11
0	2	2	523	498	1522	553	124	445	1550	146	0.01	0.28
0	3	3	86	122	650	3485	40	72	162	738	0.01	0.65
1	0	4	5106	1146	973	70	9861	9214	84	0	0.07	0.37
1	1	5	724	1128	1209	551	7482	67281	15967	191	0.24	0.71
1	2	6	20	183	508	508	778	30682	83147	10641	0.32	0.66
1	3	7	0	0	40	206	4	188	14012	113956	0.33	0.89

[a]Major category 0 is gene families that are not dispensable in parasites, and major category 1 is gene families that are dispensable in parasites.

[b]Rate classes are in increasing order of gain and loss rate.

[c]Proportion of observable gene families that were in this true category.

[d]Proportion of observable gene families in this true category for which the estimated category was correct.

Table 3: True (rows) and estimated (columns, estimated by maximum posterior probability) major categories for 100 replicate simulations, each of 3944 observable gene families generated on the tree in the main text, figure 2, with parameters from model F fitted to the COG data (main text, table 1).

	Estimated major			
True major[a]	0	1	Proportion in true major[b]	Proportion correct[c]
0	12204	6336	0.05	0.66
1	12372	363488	0.95	0.97

[a]Major category 0 is gene families that are not dispensable in parasites, and major category 1 is gene families that are dispensable in parasites.

[b]Proportion of observable gene families that were in this true major category.

[c]Proportion of observable gene families in this true major category for which the estimated major category was correct.

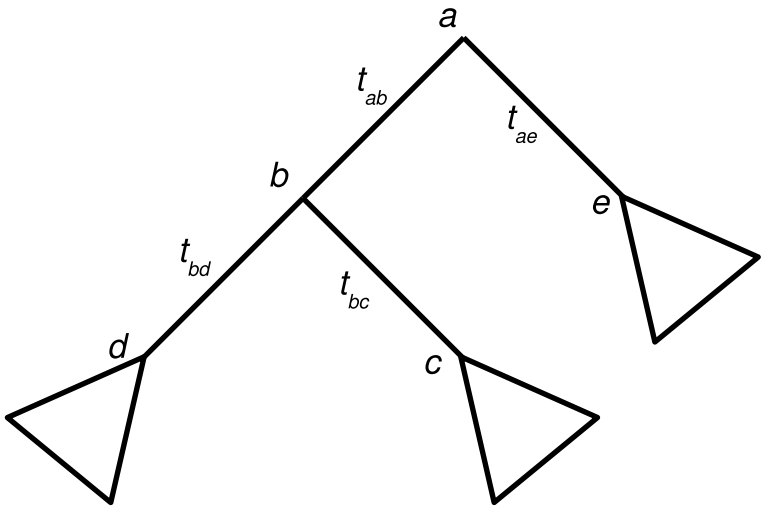

Figure 1: A tree with vertices $a \ldots e$ and edges with weights $t_{ab} \ldots t_{ae}$. The root is at a and the triangles represent descendant subtrees.

Figure 2: Algorithm for edge length optimization. We apply this to each vertex in turn, in preorder.

1. Traverse the tree in postorder, storing $z_{bd}^{(j)}$ (Equation 7) for each edge bd and each state j at the vertex b closest to the root. Also store lower subtree conditional likelihoods (Equation 3).

2. Initialize the upper subtree conditional likelihoods at the root using Equation 8.

3. Traverse the tree in preorder up to the vertex of interest (for example, b in Figure 1), calculating upper subtree conditional likelihoods at each vertex using Equation 6.

4. Optimize the log likelihood as a function of the left descendant edge (e.g. t_{bc} in Figure 1), using Equations 5, 9 and 10 and a golden section method.

5. Update the right upper subtree likelihood at the vertex of interest to account for the optimized left edge.

6. Optimize the log likelihood as a function of the right descendant edge (e.g. t_{bd} in Figure 1) in a similar way to the left edge.