Supplementary Material for "Integrating Sequence Evolution into Probabilistic Orthology Analysis"

Ikram Ullah[1], Joel Sjöstrand[2], Peter Andersson[3], Bengt Sennblad[4] and Jens Lagergren[5]

[1]*School of Computer Science and Communication, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden;*

[2]*Dept. of Numerical Analysis and Computer Science, Science for Life Laboratory, Stockholm University, Stockholm, Sweden;*

[3]*School of Computer Science and Communication, KTH Royal Institute of Technology, Stockholm, Sweden;*

[4]*Atherosclerosis Research Unit, Dept. of Medicine, Science for Life Laboratory, Karolinska Institutet, Solna, Sweden;*

[5]*School of Computer Science and Communication, Science for Life Laboratory, Swedish e-Science Research Center (SeRC), KTH Royal Institute of Technology, Stockholm, Sweden;*

Correspondence to be sent to: Jens Lagergren, School of Computer Science and Communication, KTH Royal Institute of Technology, SE - 100 44 Stockholm, Sweden;
E-mail: jensl@csc.kth.se

1 Analysis details

1.1 Data generation

We used an early in-house version of the PrIME-genphylodata tools, Sjöstrand et al., 2013 to generate 100 synthetic data sets, each from the ABCA and AGP data sets, using parameters sampled from a PrIME-DLRS analysis of the biological ABCA and AGP data sets (Supplementary Table S3). Each generated data set comprised a triplet comprising a gene tree a reconciliation and sequence alignment from biological data sets. In all cases, we used the JTT substitution model for generation of sequence data (no correction for rate variation across sites were used).

1.2 MCMC analyses

We used uninformative, uniform, priors for the gene tree G and parameters θ in all MCMC analyses. Similar MCMC settings were used for all programs (DLRSOrthology, PrIME-GEM and MrBAYES-MPR) for both the fixed-tree and the variable-tree MCMC analyses. For the synthetic data, the MCMC analyses were run for 1 500 000 iterations, while for the biological data, they were run for 2 000 000 iterations. In both cases, every 100th iteration state were recorded, and the 25% initial iterations were discarded as burnin. The posterior means of the parameters in $\theta = (\lambda, \mu, m, \nu)$ are shown in Table S3 (the posterior distribution of gene trees are discussed in the main text).

2 Description of the DLR-ROC and DLRS-ROC procedures

We here describe the comparison techniques and thresholding procedures in detail.

2.1 DLR-ROC

Let D be the biological data. In this paper, DLR-ROC is used for the sake of comparing DLRSOrthology and PrIME-GEM, but the procedure can be applied to compare any two orthology methods that take as input a gene tree – with or without lengths – and a species tree.

1. Until the required number of samples has been obtained, repeat the following:
   (a) Sample parameters θ_i from $P[\theta|D, S]$ according to the DLRS model.
   (b) Generate a synthetic gene tree G_i, with lengths l_i, and a reconciliation γ_i using the DLR process with parameters θ_i and S.
   (c) For each $v \in V(G_i)$, compute the speciation probability using method 1 (*here,* DLRSOrthology using Main Text Equation 4) for all pairs of leaves.
   (d) For each $v \in V(G_i)$, compute the speciation probability using method 2 (*here*, PrIME-GEM) for all pairs of leaves.

(e) For each gene pair $(u, v) \in G_i$ and each of the two methods, do the following: Given a set of threshold values $\Omega \in [0, 1]$, for each threshold $\omega_i \in \Omega$, compute the sensitivity/specificity values based on whether the LCA is a true speciation in γ_i and whether the orthology probability estimate is greater than ω_i.

2. Compute ROC plots based on the sensitivity/specificity data for the two methods.

2.2 DLRS-ROC

Similarly to above, in this paper, we apply DLRS-ROC in order to compare DLRSOrthology and MrBayesMPR, but the procedure can be used to compare any two orthology methods that take as input gene sequences and a species tree.

1. Until the required number of samples has been obtained, repeat the following:

   (a) Sample parameters θ_i from $P[\theta|D, S']$ according to the DLRS model.

   (b) Generate a synthetic gene tree G_i, its reconciliation γ_i, and synthetic sequences D_i using the DLRS model. Let Υ_{D_i} be set of all gene pairs in D_i.

   (c) Generate samples from $P[G^{m1}, l^{m1}, \theta^{m1}|D_i, S']$ according to MCMC framework of method 1 (*here*, DLRS). Let $C_{M1} = \{G_i^{m1}, l_i^{m1}, \theta^{m1}\}_{i=1}^{n}$ be the generated samples. For each gene pair $(u, v) \in \Upsilon_{D_i}$, compute its speciation probability across all gene trees in C_{M1} using method 1 (*here*, DLRSOrthology using Main Text Equation 2)

   (d) Generate samples from $P[G^{m2}, l^{m2}, \theta^{m2}|D_i]$ according to MCMC framework of method M2 (*here*, MrBayes). Let $C_{M2} = \{G_i^{m2}, l_i^{m2}, \theta^{m2}\}_{i=1}^{n}$ be the generated samples. For each gene pair $(u, v) \in \Upsilon_{D_i}$, compute its speciation probability across all gene trees in C_{M2} using method 2 (*here*, MrBayesMPR using Main Text Equation 5).

   (e) For each gene pair $(u, v) \in \Upsilon_{D_i}$, do the following. Given a set of threshold values $\Omega \in [0, 1]$, for each threshold $\omega_i \in \Omega$, compute the sensitivity/specificity values based on whether the LCA is a true

speciation in γ_i and whether the orthology probability estimate is greater than ω_i for each of the two methods.

2. Compute ROC plots based on sensitivity/specificity data for the two methods.

In our comparison between DLRSOrthology and MrBayes, we used identical values for MCMC parameters common to DLRS and MrBayes (e.g., number of iterations and thinning). We also used the same model of sequence evolution.

It is worth noting that while comparing DLRSOrthology and PrIME-GEM with respect to DLR-ROC, we only take MPR speciation vertices into account. This is because, by definition, MPR selects the reconciliation with the minimum number of duplications, so any vertex classified as duplication by MPR cannot be a speciation. On the other hand, while comparing DLRSOrthology and MrBayesMPR with respect to DLRS-ROC, we take all gene pairs into account since, in that case, different gene tree are considered.

References

Åkerborg, O., B. Sennblad, and J. Lagergren. 2008. Birth-death prior on phylogeny and speed dating. BMC Evolutionary Biology 8:77.

Hedges, S. B., J. Dudley, and S. Kumar. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Sjöstrand, J., L. Arvestad, J. Lagergren, and B. Sennblad. 2013. GenPhyloData: realistic simulation of gene family evolution. BMC Bioinformatics 14:209.

Table S1: Species, accession numbers and abbreviations for investigated genes of the AGP data

Species	Ensembl/OPTICS accession number	Abbreviation
Canis familiaris	ENSCAFG00000003331	Cfa_agp1
Gallus gallus	ENSGALG00000023820	Gga_agp1
Homo sapiens	ENSG00000187681	Hsa_agp1
Homo sapiens	ENSG00000204154	Hsa_agp2
Monodelhis domestica	ENSMODG00000003862	Mdo_agp1
Mus musculus	ENSMUSG00000028359	Mmu_agp1
Mus musculus	ENSMUSG00000039196	Mmu_agp2
Mus musculus	ENSMUSG00000061540	Mmu_agp3
Ornitorhynchus anatinus	ENSOANG00000013663	Oan_agp1
Taeniopygia guttata	ENSTGUG00000003499	Tgu_agp1

Table S2: Estimated *maxmin* DLRSOrthology thresholds for the investigated biological data sets.

Analysis	$ABCA$	AGP	AGP\HS
Fixed-tree analysis	0.93	0.91	0.96
Variable-tree analysis	0.93	0.75	0.55

Table S3: Posterior mean (standard deviation) of parameters θ, (i.e, duplication λ, and loss μ rates, and substitution rate model mean m and variance ν) for the biological data sets.

Data set	$\frac{\lambda}{(10^3\text{Myrs})^{-1}}$	$\frac{\mu}{(10^3\text{Myrs})^{-1}}$	$\frac{m}{(10^3\text{Myrs})^{-1}}$	$\frac{\nu}{(10^3\text{Myrs})^{-1}}$
ABCA	1.963 (0.5220)	2.27 (0.6086)	0.8880 (0.05597)	0.1165 (0.03237)
AGP[1]	6.033 (4.189)	6.555 (5.339)	2.8556 (0.5022)	0.3699 (0.4197)
AGP\HS[1]	5.164 (4.432)	5.663 (5.851)	2.924 (0.7513)	0.6139 (1.048)

[1]NB! The divergence times for the AGP species tree were estimated using MapDP (Åkerborg et al., 2008), which estimates *relative* divergence times with the divergence time of the root being set to 1.0; to enhance comparison, we have here calibrated the rate estimates for AGP and AGP\HS using a species tree root divergence time of 301.7 Myrs, taken from TimeTree (Hedges et al., 2006, http://www.timetree.org).

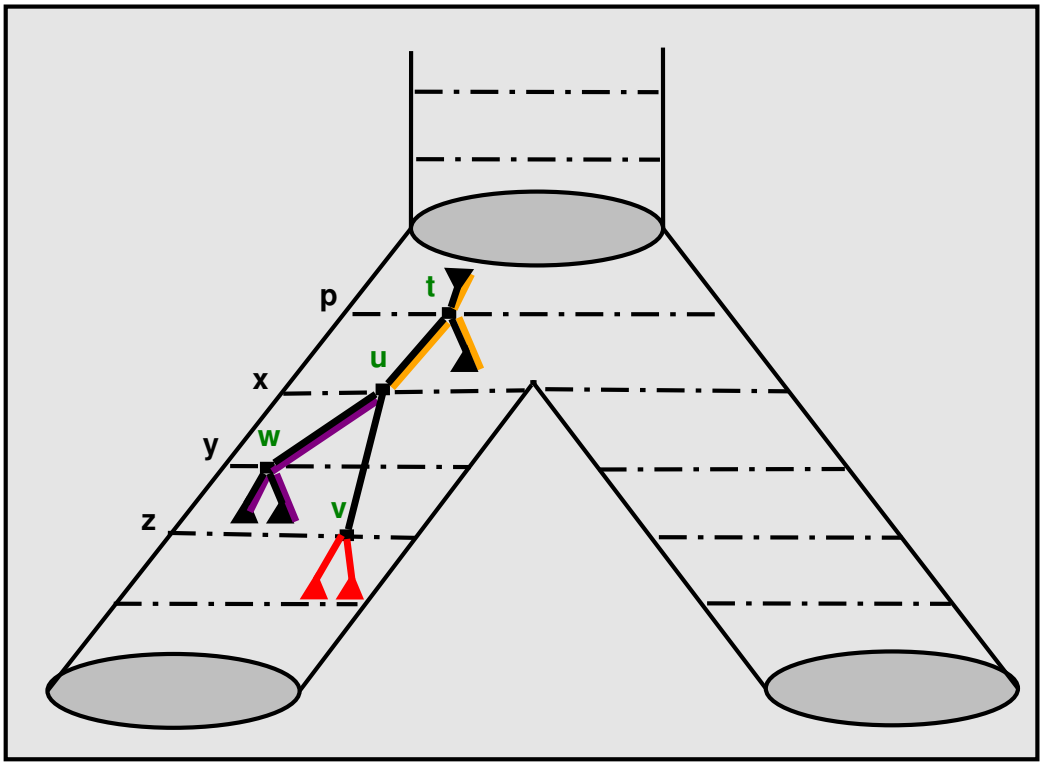

Figure S1: Illustrates the computation of the *above* probability $a(z, v)$, for $z \in V(S')$ and $v \in V(G)$, via a dynamic programming paradigm. Triangles in G refer to collapsed subtrees for clarity. $a(z, v)$ holds the probability density of all discretized realizations of $G \backslash (G_v \backslash v)$, given that v occurs on z. Let w be the sibling of v, and let u be the parent of v. $a(z, v)$ is recursively computed by – summing over each valid placement $x \in V(S')$ for u – the product of the *below* probability $b(e(x), w)$, and the *above* probability of $a(x, u)$. The illustration highlights parts of G corresponding to $a(z, v)$ in black, $b(e(x), w)$ in purple, $a(x, u)$ in orange, and $o(z, v)$ in red, respectively, albeit for a single x. For a speciation vertex x in S, we have $P[v \text{ is a speciation} | x, G, l, \theta, S'] = o(x, v)a(x, v)/P[G]$. Tree layout as in Main Text Figure 1