By: Scott V. Edwards, Liang Liu, and Dennis K. Pearl
Presented by: Masoud

High Resolution Species Trees without Concatenation

Objectives

- Understanding Concatenation Method
- Understanding ML Method
- Understanding the basic concepts of new method.

Bayesian consensus tree method (BCT)

- Estimates the posterior distribution of trees separately for each gene and the resulting gene trees for each gene are then pooled together as the posterior distribution of species trees
- The consensus tree of the posterior is then used as the point estimate summary of this species tree distribution.
- assumes independent loci

Cont.

$$L^{BCT} = f(D \mid \mathbf{G}) = f(D_1 \ldots D_K \mid G_1 \ldots G_K) = \prod_{i=1}^{K} f_i(D_i \mid G_i)$$

- The prior

$$\text{Prior}^{BCT} = f(\mathbf{G}) = f(G_1 \ldots G_K) = \prod_{i=1}^{K} f_i(G_i).$$

- Attention:
  - For any species tree of five or more taxa, there exist branch lengths in the species tree (invariably short ones) for which gene trees that do not match the species tree are more common than gene trees matching the species tree!

Bayesian concatenation method (BCM)

- Concatenation method using Bayesian approaches to infer gene trees.
- The likelihood is based on the additional assumption that all the genes arise from the same tree G*

$$L^{BCM} = f(D_1 \ldots D_k \mid G^*) = \prod_{i=1}^{k} f_i(D_i \mid G^*)$$

- The prior of gene trees assumes that the gene trees from k genes are all the same.

$$\text{Prior}^{BCM} = f(G^*)$$

Comparition

- BCT has more parameters than BCM, genes can take different trees in the Bayesian consensus tree method, whereas genes are typically assumed to follow the same tree in the Bayesian concatenation method.
- The parameter space is constrained for the Bayesian concatenation method.
- So, BCT will always provide a better fit of model to data but possibly at the expense of introducing extra variability.

But!

- BCT uses independent prior.
  - not ONLY different gene trees are independent, but also that the gene trees and species trees are independent.
  - So, what the [...] are we doing?
  - We are assuming (implicitly) that genes trees are the same as species tree.
- BCM uses a joint prior in which the gene trees across k genes are correlated with correlation = 1
  - Thus, the joint prior appears to be more appropriate than the independent prior if we assume the species tree exists and is distinct from gene trees.
- So, can you guess the idea of this paper, now? ☺

Intermediate Approach (something in the middle)

- The likelihood portion of the method is much like the one in the consensus tree method.
- Use coalescent theory to specify the correlation structure among gene trees.
- Generated samples from the posterior of gene trees for each gene is used to combine those gene trees to infer the species tree.
  - By: coalescent theory
  - coalescent provides a mechanism in which multiple gene trees can be reconciled in a single species history
- By choosing a particular prior of the species tree and the distribution of gene trees given the species tree, the Bayesian hierarchical model can be reduced to the other two basic methods.

Some Details

- **Bayesian Hierarchical Model:**

$\theta = 4N_e\mu$

$$f(S,\theta \mid D) = \frac{1}{f(D)} \int_{\Lambda} \int_{\mathbf{G}} f(D \mid \mathbf{G},\Lambda) * f(\Lambda) * f(\mathbf{G} \mid S,\theta) * f(\theta) * f(S) d\mathbf{G}$$

$f(D|G, \Lambda)$

  - **Likelihood part**

$f(\mathbf{G} \mid S,\theta)$

  - **The distribution of gene trees given species tree is derived from coalescent theory; in which random mating is assumed.**
    - **e.g. Species tree is constrained; because the gene split time predates speciation time.**

$f(\Lambda)$

  - **Includes the parameters in the substitution model and all other parameters in the likelihood function except the gene tree.**

Cont.

f(S)

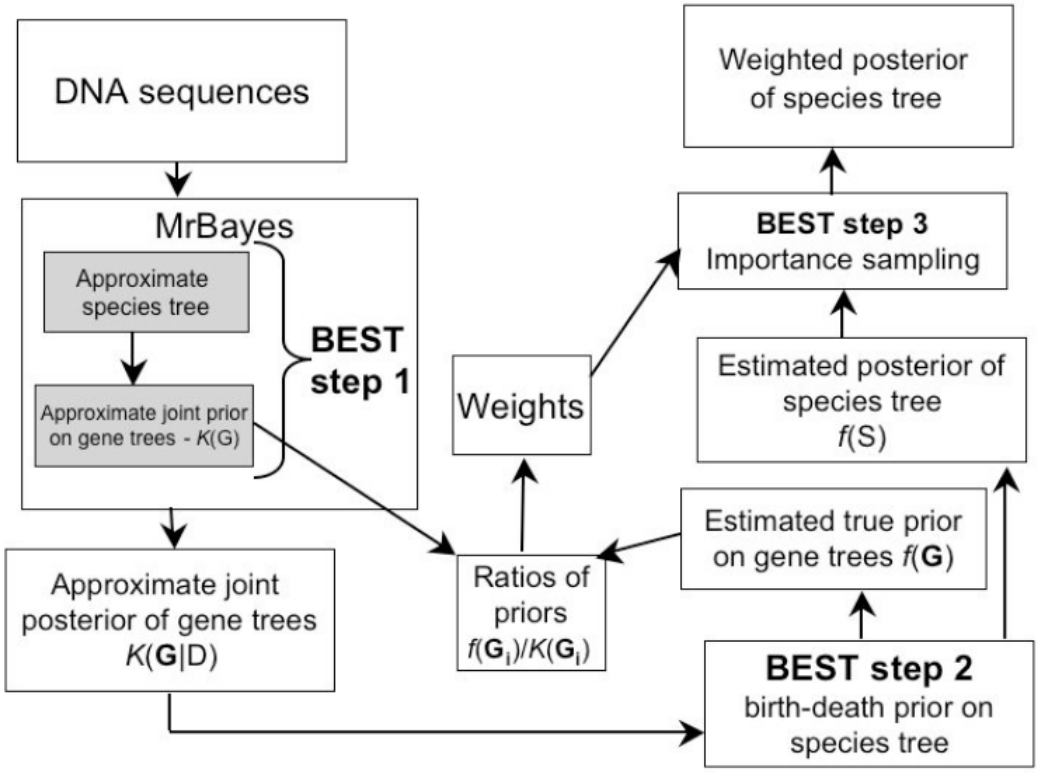
- a birth-and-death process as the prior distribution of the species tree's topology and branch lengths.

f(θ)

- independent gamma distributions as the *prior* of the effective population size $f(x|\alpha, \beta) = \Gamma^{-1}(\alpha)\beta^{-\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}}$

- Two step Markov Chain Monte Carlo (MCMC) is used for implementation.
  - See next picture.

Details of Implementaion

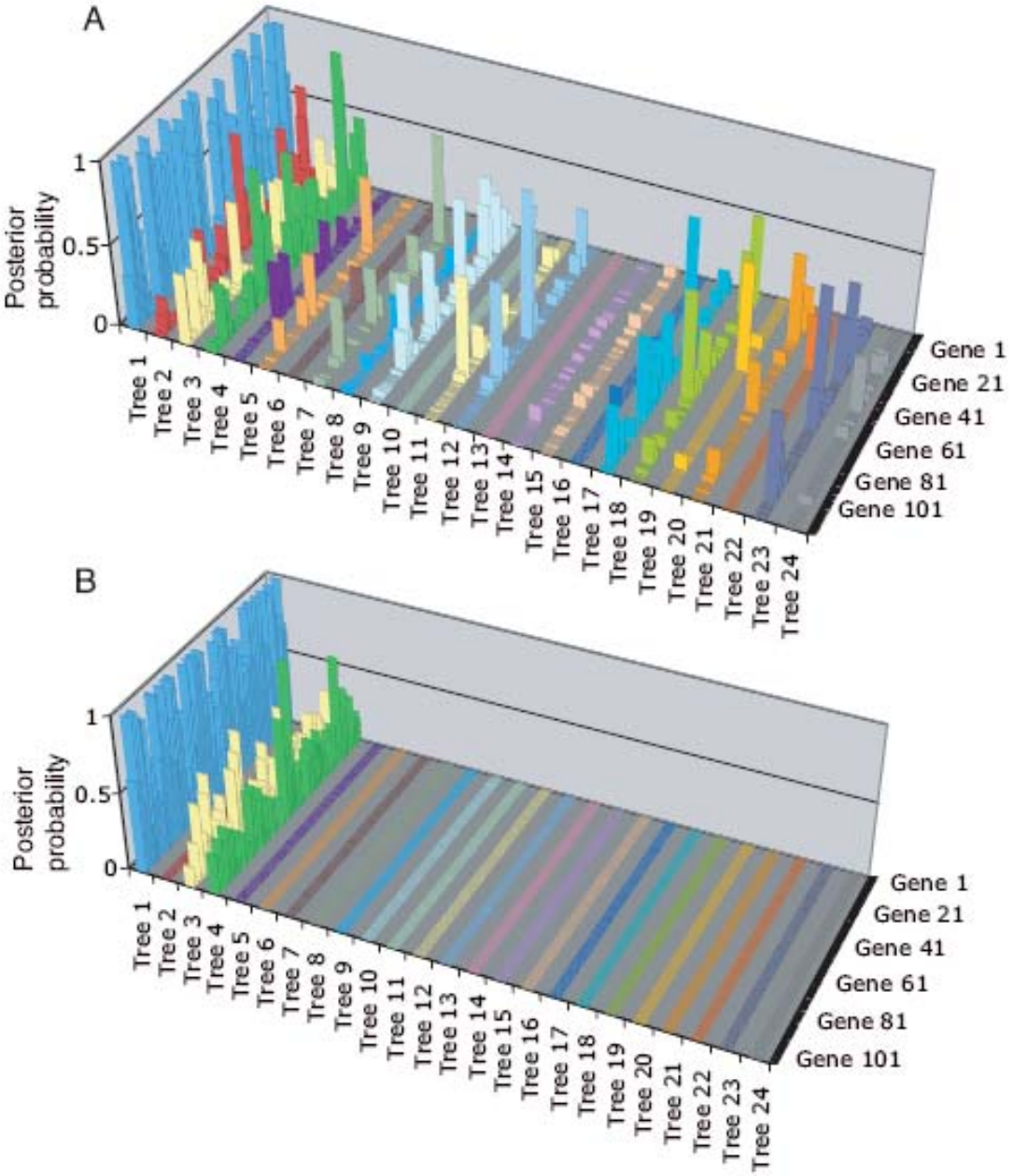


The results

- Assuming a strong positive correlation between trees causes the distribution of expected gene trees NOT to vary wildly!
- The distribution of gene trees *f(G)*, *put more weight on gene trees with similar topologies and branch lengths.*

$$f(G) = f(G_1 \ldots G_K) = \int_S \int_\theta f(G_1|S,\theta) * .. \quad * f(G_K|S,\theta) * f(\theta) * f(S) d\,S d\theta$$

  - Using the proposed priors can result in posterior gene tree distributions that are considerably more concentrated around a few topologies than when loci are analyzed independently of each other or of an overarching species tree
  - See the next picture.
  - See the posterior distribution of Pa-4 in the second next picture, it changes.


A
Posterior probability
1
0.5
0
Tree 1
Tree 2
Tree 3
Tree 4
Tree 5
Tree 6
Tree 7
Tree 8
Tree 9
Tree 10
Tree 11
Tree 12
Tree 13
Tree 14
Tree 15
Tree 16
Tree 17
Tree 18
Tree 19
Tree 20
Tree 21
Tree 22
Tree 23
Tree 24
Gene 1
Gene 21
Gene 41
Gene 61
Gene 81
Gene 101
B
Posterior probability
1
0.5
0
Tree 1
Tree 2
Tree 3
Tree 4
Tree 5
Tree 6
Tree 7
Tree 8
Tree 9
Tree 10
Tree 11
Tree 12
Tree 13
Tree 14
Tree 15
Tree 16
Tree 17
Tree 18
Tree 19
Tree 20
Tree 21
Tree 22
Tree 23
Tree 24
Gene 1
Gene 21
Gene 41
Gene 61
Gene 81
Gene 101

	Independent prior			Joint prior		
	(2,(1,3))	(3,(1,2))	(1,(2,3))	(2,(1,3))	(3,(1,2))	(1,(2,3))
Pa-1	0.184	**0.671**	0.146	0.171	**0.683**	0.146
Pa-2	0.337	**0.353**	0.309	0.299	**0.375**	0.326
Pa-3	0.062	**0.88**	0.058	0.056	**0.915**	0.029
Pa-4	0.331	0.331	0.337	0.221	**0.452**	0.327
Pa-5	0.319	0.319	0.361	0.264	**0.398**	0.338
Pa-6	0.012	**0.966**	0.022	0.047	**0.894**	0.059
Pa-7	0	**1**	0	0	**1**	0
Pa-8	0	**1**	0	0	**1**	0
Pa-9	0.042	**0.912**	0.046	0.026	**0.935**	0.038
Pa-10	0.222	**0.547**	0.232	0.117	**0.699**	0.184
Pa-11	0	**1**	0	0	**1**	0
Pa-12	0.319	**0.353**	0.327	0.293	**0.449**	0.258
Pa-13	0.493	**0.503**	0.004	0.257	**0.743**	0
Pa-14	0.242	**0.503**	0.255	0.254	**0.497**	0.249
Pa-15	0.325	**0.349**	0.325	0.151	**0.578**	0.271
Pa-16	**0.335**	**0.333**	0.331	0.233	**0.496**	0.271
Pa-17	0.042	0.02	**0.938**	0.073	0.156	**0.772**
Pa-18	0	0	**1**	0	0	1
Pa-19	0	0	**1**	0	0	1
Pa-20	0	0.002	**0.998**	0	0	1
Pa-21	0	0	**1**	0	0	1
Pa-22	0.04	0.076	**0.884**	0.045	0.085	**0.87**
Pa-23	0.014	0.064	**0.922**	0.019	0.046	**0.935**
Pa-24	0	1	0	0.002	**0.998**	0
Pa-25	0.311	0.339	**0.349**	0.232	**0.503**	0.265
Pa-26	**0.782**	0.212	0.006	0.482	**0.5**	0.018
Pa-27	**1**	0	0	**1**	0	0
Pa-28	**0.389**	0.305	0.305	0.298	**0.431**	0.271
Pa-29	0.01	**0.653**	0.337	0.001	**0.739**	0.26
Pa-30	0.333	0.327	**0.339**	0.164	**0.68**	0.156
Average support	0.205	**0.434**	0.361	0.157	**0.508**	0.335
Concatenation	0	1	0			
Joint prior (1,139)				0.08	**0.88**	0.04
Joint prior (1,1389)				0.03	**0.95**	0.02
Joint prior (1,10)				0.08	**0.89**	0.03
Joint prior (1,1)				0.01	**0.94**	0.05

For population size, (Finch data)

Table2. Estimates of the ancestral population sizes and divergence times for different priors in the finch data set. The priors for θ are Exponential with means 1/1389, 1/139), ,1/10), or 1. For each prior, the estimates of the divergence times and population sizes of a particular ancestral population are listed in the column2 and column3. (1,2) represents the ancestral population of species1 and species2. (1,2,3) is the ancestral population of species1, species2 and species3.

Exponential mean 0.00072	Divergence times	Population sizes
(1,2)	0.00408(0.00277, 0.00457)	0.00218(0.00072, 0.00553)
(1,2,3)	0.00449(0.00344, 0.00547)	0.00407(0.00252, 0.00604)
Exponential mean 0.0072	Divergence times	Population sizes
(1,2)	0.00297(0.00147,0.00389)	0.00693(0.00069,0.02813)
(1,2,3)	0.00418(0.00325,0.00523)	0.00506(0.00290, 0.00837)
Exponential mean 0.1	Divergence times	Population sizes
(1,2)	0.00235(0.00100,0.00376)	0.0155 (0.00149,0.23625)
(1,2,3)	0.00418(0.00326,0.00493)	0.00481(0.00292, 0.00783)
Exponential mean 1	Divergence times	Population sizes
(1,2)	0.00249(0.00065,0.00262)	0.01237 (0.00310,0.44104)
(1,2,3)	0.00429(0.00343,0.00525)	0.00503(0.00309, 0.00799)

NOT very good!

- **Estimating population size by itself is a very hard problem.**
- **But , the estimation of topology and divergence time seems to be robust in respect to prior of population size.**

Robustness and Efficiency

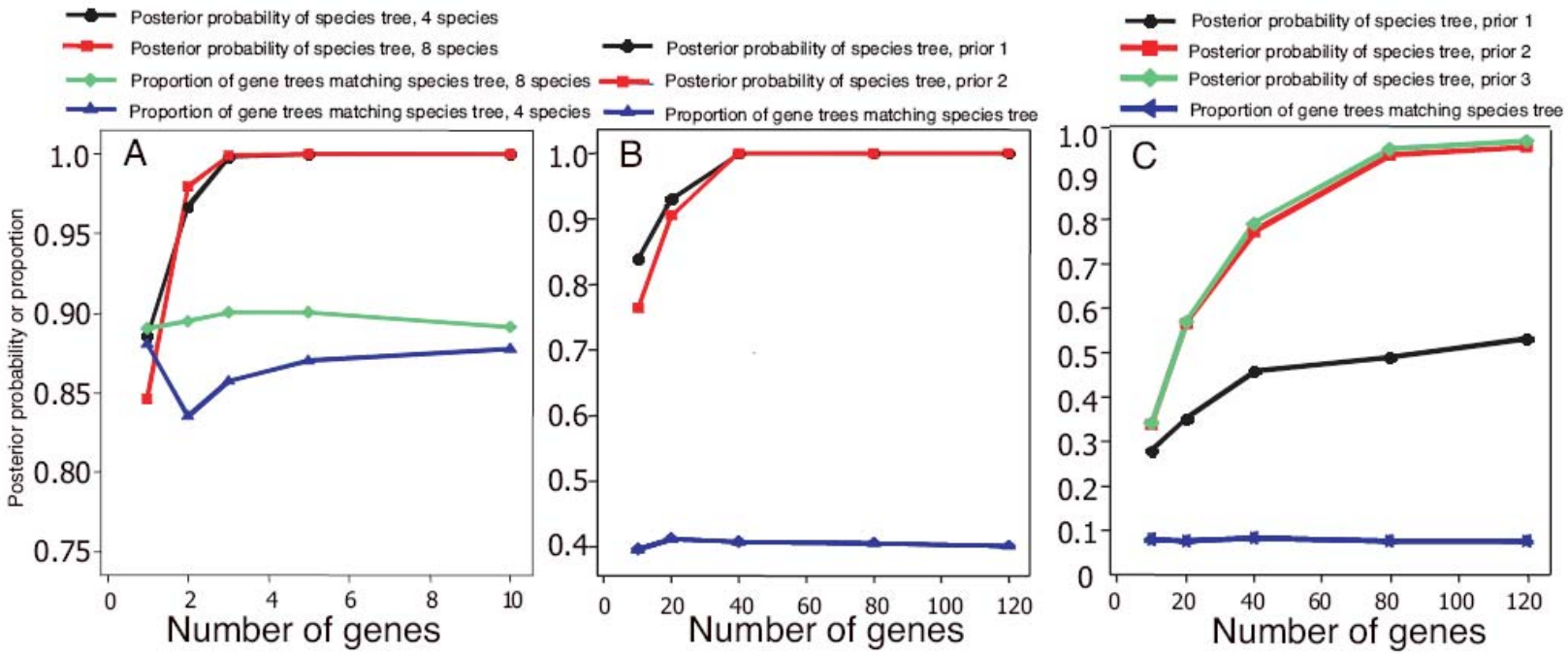


Fig. 3. Robustness and efficiency of the joint model for estimating species trees. (*A*) The number of genes required to resolve the correct species tree with four and eight species when the proportion of gene trees matching the species tree is high. Here this proportion varies between ≈83% and 90% (in blue and green, 100 gene trees per simulation) because the critical internodes in the species tree are relatively long on the scale of the effective population size (θ). The gamma-distributed prior on θ for each node was (1, 200), indicating a mean θ of 1/200 and variance of 1/40,000. A prior mean of 1/200 is consistent with what we know about θ in natural populations of yeast (45, 46). (*B*) The number of genes required to resolve the correct four-species tree when the proportion of gene trees matching the species tree (in blue) is low (≈40%). Prior 1 on θ is (1, 200), and prior 2 is (1, 1,000). (*C*) The number of genes required to resolve the correct eight-species tree when the proportion of gene trees matching the species tree (in blue) is low (<10%). Prior 1 on θ is (1, 100), prior 2 is (1, 500), and prior 3 is (1, 1,000).

Analysis of Macaque Data

- The distance between gene trees and species tree.
- This result suggests that the joint model makes the gene trees closer to the estimated species tree than the independent gene model.

Table3. The average ± st.dev. of distances between two posterior distributions for each gene in the macaques data set. There are three posterior distributions for each gene, the posterior of species trees, and the posterior of gene trees with the independent gene model and the posterior of gene trees with the joint coalescent-based model. The average distances between the posterior of specie trees and the posterior with the independent prior (denoting by independent-species) as well as the posterior of species trees and the posterior with the joint prior (denoting by joint-species) are calculated by Phylip (Felsenstein, 2004) using the symmetric distance measure (Robinson.and Foulds, 1981).

	independent-species	joint-species
Y-Chromosome	0.781 ± 0.089	0.656 ± 0.085
mtDNA	0.739 ± 0.092	0.646 ± 0.084
C4 Intron 9	0.779 ± 0.054	0.659 ± 0.078
IRBP Intron 3	0.838 ± 0.052	0.659 ± 0.070

Sensitivity Analysis

- One gene is omitted and the algorithm is performed.
- The distances between each Si and S (the posterior of species trees using all four genes) can be found in the following table.
  - The average distance of the proposed method is **lower**.

- This suggests that the estimation of the species tree is not overly subject to the strong influence of a single outlier gene in this method.

Table4. The average ± st. dev. of distances between two posterior distributions for each gene in the macaques data set. There are two posterior distributions for each gene, the posterior of gene trees assuming independent genes and the posterior of gene trees with the joint coalescent-based model. The average distances between each posterior and itself (denoted by independent-independent or joint-joint) are calculated in Phylip. The average distance between the two different posterior (denoted by independent-joint) is also calculated in Phylip.

	independent-independent	joint-joint	independent-joint
Y-Chromosome	0.309 ± 0.067	0.164 ± 0.066	0.356 ± 0.070
mtDNA	0.215 ± 0.077	0.070 ± 0.062	0.272 ± 0.087
C4 Intron 9	0.319 ± 0.067	0.237 ± 0.062	0.501 ±0.047
IRBP Intron 3	0.257 ± 0.068	0.101 ± 0.068	0.362 ±0.052

Final Notes

- Incorporation of multiple sequences sampled per species, as well as hybridization, gene flow, and lateral gene transfer into a more general model of phylogenetic inference is an important goal for future work.

- Also, the software is available free from:
  - http://www.stat.osu.edu/~dkp/BEST

- Objectives accomplished ?
- Thanks anyway! ☺
- A very "gentle" approach to coalescent theory.
  - N. A. Rosenberg and, M. Nordborg "Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms", Volume 3, May 2002, Nature.

- Questions?

- "Any altruistic system is inherently unstable, because it is open to abuse by selfish individuals, ready to exploit it." Selfish Genes, Richard Dawkins