Taming the Beast Workshop

Bayesian inference of species tree and *BEAST

Chi Zhang

June 28, 2016
Species tree

- Species tree — the phylogeny representing the relationships among a group of species

![Species tree diagram](image)

Figure adapted from [Rogers and Gibbs, 2014]

- Gene tree — the phylogeny for sequences at a particular gene locus from those species
Gene tree discordance

- Incomplete lineage sorting

![Gene tree discordance diagram with labels τ_{HCG} and τ_{HC} for Human, Chimp, and Gorilla.]

*Figure adapted from Patterson et al., 2006*
Gene tree discordance

- Horizontal gene transfer
- Gene duplication and loss

Figure adapted from [Degnan and Rosenberg, 2009]
Gene tree discordance

- Hybridization

Figure adapted from [Li et al., 2016]
Species tree inference and *BEAST

- A Bayesian method to infer species tree from multilocus sequence data [Heled and Drummond, 2010]
- *BEAST, a functionality of BEAST2

- Gene trees are embedded in the species tree under the multispecies coalescent model [Rannala and Yang, 2003]
  - incomplete lineage sorting
- Gene trees are independent among loci
The prior for species tree $S$ has two parts:

$$P(S) = P(S_T)P(N)$$

- $S_T$ — species time tree
- $N$ — population size functions
- $P(S_T)$ — typically a Yule (pure-birth) or birth-death prior
  - we can assign a hyperprior for the speciation (birth) rate (and extinction (death) rate, if birth-death)
- $P(N)$ — constant or continuous-linear
Species tree prior

- Constant population sizes

\[ N_i \sim \text{gamma}(k, \psi) \]

Figure adapted from [Drummond and Bouckaert, 2015]
Species tree prior

- Continuous-linear population sizes

Figure adapted from [Drummond and Bouckaert, 2015]
Species tree prior

- In *BEAST, the prior type for $N$ is fixed to gamma
- The gamma shape parameter $k$ is fixed to 2, but we can assign a hyperprior for $\psi$, the scale parameter of the gamma
- (This $\psi$ parameter is called ”population mean” in Beauti, but the prior mean is actually $2\psi$ when the population sizes are constant)
Multispecies coalescent model

- The prior for gene tree $g$, given species tree $S$

Figure adapted from [Drummond and Bouckaert, 2015]
The probability distribution of gene time tree $g$ given species tree $S$, is:

$$P(g|S) = \prod_{j=1}^{2s-1} P(L_j(g)|N_j(t))$$

- $s$ — number of extant species ($2s - 1$ branches totally)
- $N_j(t)$ — population size function (linear)
- $L_j(g)$ — coalescent intervals for genealogy $g$ that are contained in the $j$’th branch of species tree $S$
Molecular clock model

- \( P(c) \) — prior for the molecular clock model of genealogy \( g \)
  - strict clock — typically fix to 1.0 for the first locus, and infer the relative clock rates for the rest loci
  - relaxed clock

- \( P(\theta) \) — prior for the substitution model parameters
  - e.g. HKY85,
    - Prior for transition/transversion rate ratio (\( \kappa \)), e.g. \( \text{gamma}(2,1) \)
    - Prior for base frequencies (\( \pi_T, \pi_C, \pi_A, \pi_G \)), e.g. Dirichlet(1,1,1,1)
The probability (likelihood) of data $d_i$ (alignment at locus $i$), given the gene time tree $g_i$, molecular clock $c_i$, and substitution model $\theta_i$, is:

$$P(d_i|g_i, c_i, \theta_i)$$
Priors and likelihood

- $P(S)$ — prior for species tree
- $P(g_i|S)$ — prior for gene tree $i$ (multispecies coalescent)
- $P(c_i)$ — prior for clock rate of locus $i$
- $P(\theta_i)$ — prior for substitution parameters of locus $i$
- $P(d_i|g_i, c_i, \theta_i)$ — likelihood of data at locus $i$
The posterior distribution of the species tree $S$ and other parameters given data $D$ is:

$$P(S, g, c, \Theta|D) \propto P(S) \prod_{i=1}^{n} P(g_i|S)P(c_i)P(\theta_i)P(d_i|g_i, c_i, \theta_i)$$

The data $D = \{d_1, d_2, \ldots, d_n\}$ is composed of $n$ alignments, one per locus.
Assume constant population sizes

Assign i.i.d inverse-gamma(α, β) prior for Nj
  - mean = β/(α − 1)

The population sizes N can be integrated out from P(g|S) [Jones, 2015]

Specify α and β in the invgamma prior (instead of ψ in the gamma prior)
starBEAST2

- A more efficient implementation and an upgrade of *BEAST
  - Population sizes integrated out [Jones, 2015]
  - Relaxed molecular clock per species tree branch (instead of per gene tree branch)
  - More efficient MCMC proposals for the species tree and gene trees (coordinated operators) [Jones, 2015, Rannala and Yang, 2015]
- Available at github.com/genomescale/starbeast2, will be released soon (as a BEAST2 add-on)
References