PhyloAcc-C model description (version 0.1)

Patrick Gemmell

Introduction

The PhyloAcc-C method closely follows the approach taken by PhyloAcc (Hu et al. 2019) and extends it into the realm of continuous phenotypic change. More information on the PhyloAcc family of methods is available at: https://phyloacc.github.io.

Input

The method relies on four inputs: (i) a rooted phylogeny $\mathbf{T}$ having L leaves, $N = 2L - 1$ nodes, and $E = N - 1$ edges, and that encapsulates the relationship between species from which trait data is drawn; (ii) a multiple sequence alignment $\mathbf{X}_{L\times S}$ juxtaposing homologous CNE sequences observed in the L species (rows) at S sites (columns); (iii) a vector $\mathbf{y} = (y_1, y_2, ..., y_L)$ of continuous trait measurements observed in the corresponding species; (iv) a rate matrix $\mathbf{Q}_{4\times 4}$ and stationary distribution π that model the background nucleotide substitution process at neutral sites. The alignment may contain gaps. Both the alignment and the nucleotide substitution parameters can be obtained using standard methods (e.g. see Hu et al. 2019).

Model

Each branch is assigned a **conservation state** z_i that is one of background ($z_i = 1$), conserved ($z_i = 2$), or accelerated ($z_i = 3$). Conservation states are not assigned freely but follow a Markov process from root to tips so that $p(z_j|z_i) = \mathbf{\Phi}_{i,j}$ where

$$\mathbf{\Phi} = \begin{bmatrix} 1-c & c & 0 \\ 0 & 1-a & a \\ 0 & b & 1-b \end{bmatrix}.$$

The structure of this matrix is such that CNEs may become conserved and later accelerated. When a transition from accelerated back to conserved is possible (i.e $b > 0$) then bursts of acceleration can occur on internal branches.

The conservation state of a branch affects both the nucleotide substitution process and the rate at which a trait moves. Conservation states modulate nucleotide substitution rates via **substitution rate multiplier**s $\mathbf{r} = (r_1 = 1, r_2, r_3)$ so that the probability of transition from nucleotide a to b on branch i of length t_i is $\text{expm}\{\mathbf{Q} \cdot r_{z_i} \cdot t_i\}_{a,b}$. This is the well-known and standard approach to molecular evolution.

Similarly, conservation states modulate the magnitude of trait changes along branches via **variance multipliers** $\mathbf{v} = (\sigma^2, \beta_2\sigma^2, \beta_3\sigma^2)$. Under the model traits evolve according to normally distributed displacements along the branches of $\mathbf{T}$ such that cumulative displacements are observed in $y_1, ..., y_L$. Displacements have mean 0 and variance that is proportional to both the branch length t_j and the appropriate variance multiplier:

$$y_j \sim \text{Normal}(y_i, t_j v_{z_j}) = \begin{cases} \text{Normal}(y_i, t_j\sigma^2) & \text{if } z_j = 1 \text{ i.e. background} \\ \text{Normal}(y_i, t_j\beta_2\sigma^2) & \text{if } z_j = 2 \text{ i.e. conserved} \\ \text{Normal}(y_i, t_j\beta_3\sigma^2) & \text{if } z_j = 3 \text{ i.e. accelerated} \end{cases}.$$

Joint likelihood

Using the notational convention that node $pa(i)$ denotes the parent of node i, and that $pa(j) = pa(k) = i$, the joint likelihood of the model is:

$$\begin{aligned} P(\mathbf{X}, \mathbf{y}, \mathbf{z}, \mathbf{r}, \mathbf{v}, \mathbf{\Phi}|\Theta) &= \left[\prod_{s=1}^{S} P(\mathbf{X}_{:,s}|\mathbf{r}, \mathbf{z})\right] P(\mathbf{y}|\mathbf{v}, \mathbf{z})\, P(\mathbf{z}|\mathbf{\Phi})\, P(\mathbf{r}, \mathbf{v}, \mathbf{\Phi}|\Theta) \\ &= \prod_{s=1}^{S}\prod_{j=1}^{E} P(X_{j,s}|\mathbf{r}, \mathbf{z}, X_{i,s}) \prod_{j=1}^{E} P(y_j|\mathbf{v}, \mathbf{z}, y_i) \\ &\cdot \prod_{j=1}^{E} P(z_j|\mathbf{\Phi}, z_i) \\ &\cdot \left[\prod_{s=1}^{S} \pi_{X_{R,s}}\right] P(y_R|\Theta)\, P(z_R|\Theta)\, P(\Theta). \end{aligned}$$

The last line of the above product indicates that at root node R the prior probability of observing a nucleotide is given by the (input) stationary distribution π while the trait and conservation state are specified directly using a prior (below).

Prior specification

At the root, $p(z_R)$ is set using a categorical distribution and y_R is specified using a Normal e.g. $y_R \sim \text{Normal}(0, 1)$. Priors on the entries a, b, c in $\mathbf{\Phi}$ are Beta distributions. Priors on the substitution rate multipliers are Gamma, e.g. in the mammalian case $r_2 \sim \text{Gamma}(5, 0.04)$ and $r_3 \sim \text{Gamma}(10, 0.2)$ would ape the values carefully selected by Hu et al. (2019). Priors on $\log\beta_2$ and $\log\beta_3$ are Normal, e.g. Normal(0, 1), which is mathematically equivalent to setting a Normal(0, 2) on their ratio. The background rate of trait evolution has a log-normal prior e.g. $\log\sigma^2 \sim \text{Normal}(0, 2)$ might be sensible, though PhyloAcc-C can be run with a uniform prior on $\log\sigma^2$ also.

Inference

Inference is performed using collapsed Gibbs sampling with some Metropolis within Gibbs steps. (For clarity, $\mathbf{Q}$ and t are implicit.) The following steps are repeated:

Step 1: sample ancestral trait values y and trait variance multipliers v

Perform Metropolis steps (default is 500) to propose/update σ^2, β_2, β_3, and latent $\mathbf{y}$. On 60% of iterations proposals to modify σ^2, β_2, and β_3 are made; on the remaining occasions proposals to perturb latent $y_{L+1}, ..., y_N$ are made.

Step 2: sample ancestral nucleotides X

First use the familiar pruning algorithm to calculate the likelihood of subalignment $\{\mathbf{X}_{i,s}\}$ rooted at node i for all sites $s = 1...S$ using the recurrence:

$$\begin{aligned} P(\{\mathbf{X}_{i,s}\}|X_{i,s}, \mathbf{z}, \mathbf{r}) = &\sum_{X_{j,s}} P(X_{i,s} \to X_{j,s}|r_{z_j}) P(\{\mathbf{X}_{j,s}\}|X_{j,s}, \mathbf{z}, \mathbf{r}) \\ &\cdot \sum_{X_{k,s}} P(X_{i,s} \to X_{k,s}|r_{z_k}) P(\{\mathbf{X}_{k,s}\}|X_{k,s}, \mathbf{z}, \mathbf{r}). \end{aligned}$$

Next, forward sample ancestral nucleotides $\mathbf{X}_{L+1...R,1...S}$. For sites at the root node $P(X_{R,s}|\{\mathbf{X}_{R,s}\}, \mathbf{z}, \mathbf{r}) \propto P(\{\mathbf{X}_{R,s}\}|X_{R,s}, \mathbf{z}, \mathbf{r})P(X_{R,s})$ and $P(X_{R,s}) = \pi_{X_{R,s}}$ by assumption. For the remaining internal nodes we work from root to tips on a per-site basis using:

$$P(X_{j,s}|X_{i,s}, \mathbf{z}, \mathbf{r}) \propto P(\{\mathbf{X}_{j,s}\}|X_{j,s}, \mathbf{z}, \mathbf{r}) P(X_{i,s} \to X_{j,s}|r_{z_j}). \tag{1}$$

Step 3: sample per-branch latent rate categories z

First, from tips to root, calculate the joint likelihood of trait values and nucleotide emissions $\{\mathbf{XY}_i\}$ occurring on the subtree rooted at node i using the recurrence:

$$\begin{aligned} P(\{\mathbf{XY}_i\}|\mathbf{X}, \mathbf{y}, z_i, \mathbf{r}, \mathbf{v}, \mathbf{\Phi}) = &\prod_{s=1}^{S} P(X_{pa(i),s} \to X_{i,s}|r_{z_i}) \\ &\cdot P(y_{pa(i)} \to y_i|v_{z_i}) \\ &\cdot \sum_{z_j} P(z_i \to z_j|\mathbf{\Phi}) P(\{\mathbf{XY}_j\}|\mathbf{X}, \mathbf{y}, z_j, \mathbf{r}, \mathbf{v}) \\ &\cdot \sum_{z_k} P(z_i \to z_k|\mathbf{\Phi}) P(\{\mathbf{XY}_k\}|\mathbf{X}, \mathbf{y}, z_k, \mathbf{r}, \mathbf{v}). \end{aligned}$$

Next, sample $\mathbf{z}$ from root to tip. At the root our (domain specific) prior is $P(z_R) = (0.5, 0.5, 0.0)$. For descendent nodes the appropriate probabilities are:

$$P(z_j|\mathbf{X}, \mathbf{y}, z_i, \mathbf{r}, \mathbf{v}, \mathbf{\Phi}) \propto P(\{\mathbf{XY}_j\}|\mathbf{X}, \mathbf{y}, z_j, \mathbf{r}, \mathbf{v}, \mathbf{\Phi}) P(z_i \rightarrow z_j|\mathbf{\Phi}). \tag{2}$$

Step 4: sample per-category nucleotide substitution rate modifiers r

Perform Metropolis step to propose/update substitution rate multipliers r_2 and r_3.

Step 5: sample latent rate category transition probabilities $\mathbf{\Phi}$

The Beta prior on entries a, b and c of $\mathbf{\Phi}$ leads to a Beta posterior. For example, the posterior on a is directly sampled based on transitions from 1 → 2 as follows:

$$a \sim \text{Beta}\left(a_\alpha + \sum_j^E I(z_i = 1, z_j = 2), a_\beta + \sum_j^E I(z_i = 1, z_j = 1)\right) \tag{3}$$

The posteriors of b and c are calculated similarly using the count of transitions 2 → 3 and 3 → 2 respectively.

Model selection/ranking CNEs

A collection of candidate elements can be ranked with respect to a trait of interest based on the Bayes factor in favour of the 'full model' described above. In the full model variance multiplier $\mathbf{v}$ is free to vary whereas in the restricted model $\beta_2 = \beta_3 = 1$, and no systematic relationship between the rate of trait evolution and relative substitution rates is specified. As the null hypothesis is nested and the priors on β_2 and β_3 are common to all candidate elements, the Bayes factor is estimated using the posterior density of $(\log\beta_2, \log\beta_3)$ at $(0, 0)$ (see e.g. Wagenmakers et al. 2010).

References

Hu, Z., Sackton, T. B., Edwards, S. V., and Liu, J. S. (2019). Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees, Molecular biology and evolution, 36(5), 1086–1100.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. Cognitive psychology, 60(3), 158-189.