Linking phenotypic and genotypic variation: a relaxed phylogenetic approach using the probabilistic programming language Stan

Patrick Gemmell[1]

[1]Department of Organismic and Evolutionary Biology, Harvard University

January 2024

Abstract

PhyloG2P methods link genotype and phenotype by integrating evidence from across a phylogeny. I introduce a Bayesian approach to jointly modelling a continuous trait and a multiple sequence alignment, given a background tree and substitution rate matrix. The aim is to ask whether faster sequence evolution is linked to faster phenotypic evolution. Per-branch substitution rate multipliers (for the alignment) are linked to per-branch variance rates of a Brownian diffusion process (for the trait) via a flexible function. Notably, the method is implemented concisely using the probabilistic programming language Stan. Simulation studies suggest the model can be well estimated using relatively short alignments and reasonably sized trees.

Introduction

The term 'PhyloG2P' was coined in a review article [1] summarizing methodologies that link genotype and phenotype by making comparisons across a tree. PhyloG2P methods treat genetic observations (e.g. accelerated molecular substitution rate) in the appropriate context (e.g. lineages where a trait is lost) as evidence in favour of a biological association between genotypic and phenotypic evolution. Such methods have tended to focus on binary traits, although some methods have been developed for continuous traits too [2, 3, 4, 5].

Here I introduce Halcyon, a new statistically motivated PhlyoG2P method for jointly modelling a continuous trait and a corresponding focal multiple sequence alignment. Like the existing method PhyloAcc-C [5], the Halcyon model makes use of a null/background species tree and substitution rate multipliers, but unlike PhyloAcc-C, these substitution rate multipliers can scale the rate of molecular evolution in an arbitrary way on a per-branch basis and could therefore be described as 'relaxed' (see [6] for a similar use of the term). The conceit of the Halcyon model is that these substitution rate multipliers can then be linked to rate of trait evolution using variance rates that are a function of the aforementioned substitution rate multipliers. A systematic coincidence of deviation from a neutral/background model of sequence evolution and a null/background model of trait evolution may then indicate an interesting biological link between genotype and phenotype.

A notable feature of the implementation of Halcyon is that it makes use of the probabilistic programming language Stan [7], and thus it is both remarkably easy to communicate (see Appendix 1) and also benefits from the infrastructure for computational statistics that has been built up by the Stan community. Accordingly, as well as communicating the model, an additional aim of this document is to demonstrate the utility of general purpose statistical tools for real phylogenetic methods development.

Methodology

Model overview. Consider a vector $\mathbf{y} = (y_1, ..., y_L)$ of continuous trait measurements (e.g. longevity, beak depth, or systolic blood pressure during pregnancy-induced hypertension) placed at the tips of a rooted tree $\mathbf{T}$ having L leaves, $N = 2L - 1$ nodes, and $E = N - 1$ edges of length $\mathbf{t} = (t_1, \ldots, t_E)$. Consider also a matrix $\mathbf{X}_{L\times S}$ representing a DNA multiple sequence alignment made using genetic observations (e.g. enhancers, promoters, concatenated endogenous retrovirus sequences) from the same L species (the rows) at S sites (the columns). One would like to mathematically model all this data together in order to describe an association, if any, between the rate of trait evolution and the rate of sequence evolution. The Halcyon model approaches this objective by scaling branch lengths $\mathbf{t}$ using per-branch substitution rate multipliers $\mathbf{r} = (r_1, \ldots, r_E)$ which are then linked to the rate of change of the trait along any given branch using a functional form $\sigma^2(r_j)$ for the variance rate.

Sequence evolution. Under the proposed model the evolution of the DNA sequences aligned in $\mathbf{X}$ is treated in the usual way using a substitution rate matrix $\mathbf{Q}_{4\times4}$ so that the probability of transition from nucleotide a to b on branch j of length t_j is $\text{expm}\{\mathbf{Q} \cdot r_j \cdot t_j\}_{a,b}$ (see e.g. textbooks [8, 9] for details on the continuous-time discrete-state Markov chain approach to molecular evolution). The tree $\mathbf{T}$ and matrix $\mathbf{Q}$ (as well as its associated stationary distribution π, used to model unconditional nucleotide frequencies at the root of the tree) are treated as fixed, assumed to have been obtained elsewhere using standard methods such as the `phyloFit` command from the PHAST package [10].

Trait evolution. Trait evolution is modelled using a distribution whose variance grows with (potentially) scaled branch length. For concreteness, I present the normal distribution, though other distributions might be equally plausible a priori. Under this scheme, a change from trait value y_i to y_j along branch (i, j) of length t_j is modelled so that $y_j \sim N(y_i, t_j\sigma_j^2)$, where σ_j^2 is a trait variance rate obtained using a function somehow dependent on substitution

rate multiplier r_j.

It is not immediately clear what function should be used to relate the rate of sequence evolution (the substitution rate multipliers $\mathbf{r}$) to the rate of trait evolution (the trait variance multipliers $\boldsymbol{\sigma^2}$). For this reason, a sensible thing to do is to choose a flexible function, allowing positive, negative, and null associations, as well as various degrees of non-linearity, and the possibility of threshold effects. As Figure 1 shows, the logistic function provides such flexibility. Under this scheme, we have $\sigma_j^2 = \sigma^2(r_j) = h/(1+\exp(-a-b\log(r_j)))$ where h, a, and b are scalar parameters shared across all branches of the tree. When $b = 0$ then the trait is following a Brownian motion on the background phylogeny as described by e.g. [11]. When $b \neq 0$ then the trait is following a Brownian motion that has a variance rate that depends on substitution rate multipliers $\mathbf{r}$ (that may arbitrarily accelerate or decelerate the rate of nucleotide evolution on any particular branch in order to better describe the alignment $\mathbf{X}$) as mediated via the shape of logistic function σ^2.

All together then, using the notational convention that R indexes the root node, the likelihood of the proposed model is:

$$P(\mathbf{X}, \mathbf{y}|\mathbf{r}, h, a, b, y_R) = \left[\prod_{s=1}^{S} P(\mathbf{X}_{:,s}|\mathbf{r})\right] P(\mathbf{y}|\mathbf{r}, h, a, b, y_R)\, P(\mathbf{r}, h, a, b, y_R).$$

Note that in order to actually calculate this likelihood one marginalizes over all ancestral nucleotides and trait values.

The following priors were used for simulation studies (see below), though one can easily conceive of different choices based on the nature of the sequence in alignment $\mathbf{X}$ and the tree

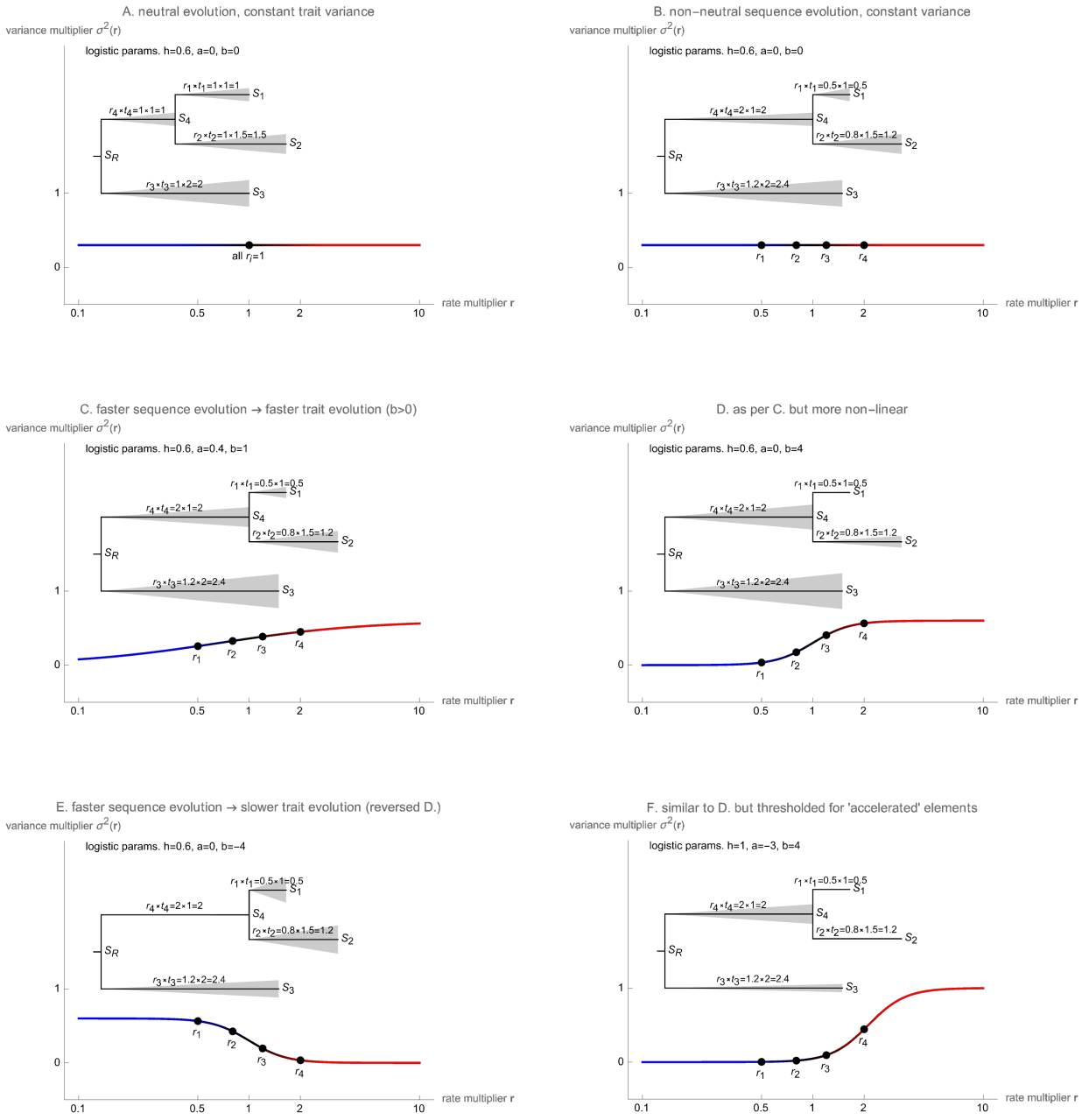

Figure 1: Six example configurations of the Halcyon model based on different parameter values. Outer axes describe variance rate as a function of substitution rate multiplier. Inset in each plot is an example tree, with branch lengths scaled by substitution rate multiplier, and grey triangles showing per-branch trait variance. See the discussion section (below) for an interpretation of the figures assuming the null/background tree $\mathbf{T}$ is based on the neutral substitution rate. Estimating the model involves considering which parameters are more or less likely based on priors and the observed input data.

$\mathbf{T}$:

$$\log(r_1), \ldots, \log(r_E) \overset{iid}{\sim} \text{Normal}(0, 1)$$

$$y_R \sim \text{Normal}(0, 1)$$

$$h \sim \text{Uniform}(0, 4)$$

$$a, b \sim \text{Normal}(0, 2).$$

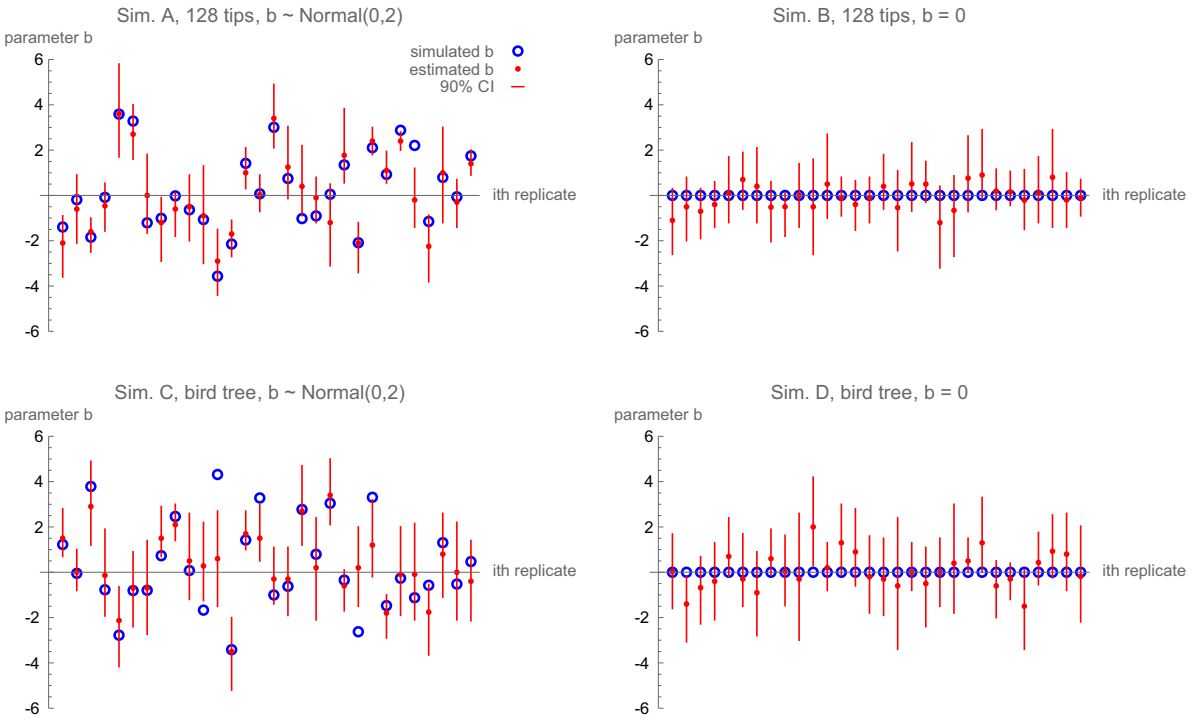

Figure 2: Simulated versus estimated b (mean) under model. Simulations were undertaken on a fully bifurcating ultrametric tree with 128 tips and all branch lengths set to 0.1 (row 1), and a real-world bird tree with 65 tips (row 2). Simulations were performed with an association between the rate of sequence evolution and the rate of trait evolution sampled from the prior (column 1), and with no systematic association (column 2). Simulations used 120 bp alignments.

Simulation studies. To demonstrate the feasibility of the Halcyon model I undertook four simulation studies (Figure 2) focusing on the problem of estimating parameter b, which

encodes the presence of a positive or negative association, if any, between the rate of nucleotide and trait evolution. Each simulation study involved 30 replicates. In Simulation A, I drew parameters from their prior distributions and then simulated alignments and trait values under the Halcyon model by making use of functions from the `phangorn::` [12] and `ape::` [13] packages. I used a bifurcating tree with 128 tips, branch lengths of 0.1, and a fairly short alignment length of 120 bp. I then obtained parameter estimates using the `stan_halcyon_logistic.stan` script described in Appendix 1. This involved running 3 randomly initialized MCMC chains per replicate, with random starting values, a warmup of 1,500 samples, and a additional 2,000 samples for parameter estimation. Figure 2 shows one can recover b to a reasonable accuracy, and the model appears well calibrated as in 27 of 30 replicates the simulated value was inside the 90% CIs (credible intervals), as is expected. The reported Gelman–Rubin convergence diagnostic for the parameters h, a, and b was almost always 1.0 (and < 1.01 otherwise).

In Simulation B, with the exception of parameter b, I again drew parameters from their prior distributions and then simulated alignments and trait values under the Halcyon model. However, in Simulation B the value of parameter b was fixed at 0 to investigate the false positive behaviour of the model. Figure 2B shows that, if anything, the inference was conservative and favoured false negatives, as 30 out of 30 CIs overlap $b = 0$.

Simulations C and D were analogues of Simulations A and B with the difference that a real-world bird tree (Subir Shakya, personal communication) with 65 tips and heterogeneous branch lengths was used in place of the synthetic bifurcating tree with homogeneous branch lengths. Figure 2C shows that estimation of b is possible on a smaller, unbalanced tree with irregular branch lengths, and Figure 2D shows that the model still behaves well in the absence of a systematic link between genotype and phenotype. In simulation C, 25 of 30 CIs contained the simulated value of parameter b, and in Simulation D 26 of 30 did, again, in both cases, similar to what is expected.

Implementation. The model was investigated using a Metropolis Hastings scheme, a Gibbs sampling scheme, and a Hamiltonian Monte Carlo scheme. The latter approach, implemented using the Stan software, worked well, is described in Appendix 1, and is made available in full at `https://github.com/pgemmell/stan_halcyon`.

Discussion

For presumably practical reasons, popular phylogenetic software has tended to focus on 'hand-crafted' Markov chain Monte Carlo samplers (e.g. [2, 14]) rather than make use of probabilistic programming languages such as BUGS [15], JAGS [16], and Stan [7]. This is contrary to the widespread use of probabilistic programming languages in other areas of biology (e.g. ecology, biomedical science), as well as in the natural and social sciences (for a diversity of examples see textbooks [17, 18, 19]). While the efficiency of hand-crafted samplers can be exceptional, when implementing Halcyon I found the performance of Stan to be 'good enough' for challenging phylogenitic inference, which is to say one can get useful work done, as is demonstrated by the inference done during the simulation studies (above). I was able to reap the benefit of a comprehensive and well tested software tool, with excellent documentation and interfaces (`https://mc-stan.org/`). Further, a Stan script embodies the notion that 'programs must be written for people to read, and only incidentally for machines to execute' (p. xxii) [20], and so is particularly helpful for communicating new models (see Appendix 1).

Turning to potential uses of the model one wonders how to interpret parameter estimates. The Halcyon model was conceived with the idea of treating b as the main parameter of interest. Parameter b describes the association between the rate of genotypic evolution and the rate of phenotypic evolution, while parameter h captures the 'height' of the relationship, and parameter a 'shifts' the relationship, allowing for threshold effects. Under the

model, to the extent the rate of trait evolution is systematically related to a rescaling of the null/background tree via the substitution rate multipliers $\mathbf{r}$ (needed to explain the alignment $\mathbf{X}$) parameter b will differ from 0 (e.g. Figure 1C–F). If branches that must be stretched to explain $\mathbf{X}$ also exhibit faster trait evolution then b will tend to be positive (e.g. Figure 1C, D, and F). If branches that must be stretched to explain $\mathbf{X}$ exhibit slower trait evolution then b will tend to be negative (e.g. Figure 1E).

The Halcyon model is designed under the assumption that departures from any null expectation of evolution will be sporadic, and of varying magnitude and direction, due to correspondingly unpredictable periods of directional or purifying selection that result from rather random factors. In other words, changes to the same trait in different species will generally be due to different events, that occur at different times, in different places, and unfold in different ways. Therefore correlated rate multipliers are not appropriate even though correlated rates (e.g. [21]) might be an important tool for estimating a null/background tree based on neutral substitution rates.

In the case the null/background tree is a genetic tree reflecting an appropriate neutral rate of substitution then rate multipliers $\mathbf{r}$ themselves do have an interpretation in terms of population genetics, e.g. $r_j > 1$ can be interpreted as directional selection (p. 262–264) [22], as was done by e.g. [23] in the context of retroviral sequence. Note however that the tree need not be a genetic tree reflecting the neutral substitution rate, but could instead be a genetic tree representing some other expectation (e.g. the evolutionary rate of some known genetic covariate of the trait), or even a non genetic tree. In these cases one would probably need to carefully choose different priors in order to properly capture their intent.

Continuing the discussion under the assumption a tree reflecting neutral substitution rates was used, one could provide the following story about the example configurations in Figure 1. In Figure 1A, the focal sequence is evolving neutrally, and there is no systematic link between the rate of genotypic and phenotypic evolution. In Figure 1B, the focal sequence

may be conserved on some branches (r_1, r_2) and accelerated on others (r_3, r_4), though there is still no systematic link between the rate of genotypic and phenotypic evolution. In Figure 1C, parameters are chosen so there is a link between genotypic and phenotypic evolution, and the link is fairly linear. In Figure 1D the relationship is more non-linear than in Figure 1C (there is a saturation effect for both conserved and accelerated elements), while in Figure 1E the relationship is reversed. Figure 1F shows a shifted version of Figure 1D, whereby only a substantial acceleration of sequence evolution relative to the neutral rate is linked to an increase in the rate of evolution of a trait. This is in contrast to Figure 1C and D, where an acceleration relative to the most conserved elements is sufficient to induce a change in the rate of phenotypic evolution, as is the case for the model [5]. Note that it is the exploration and weighting of configurations like these example ones that is performed when estimating parameter values of the Halcyon model using Markov chain Monte Carlo.

How does Halcyon differ from the other Bayesian PhyloG2P methods that model continuous traits? The Halcyon model differs from [5] in that it is much more flexible: it considers arbitrary evolutionary trajectories at the sequence level and therefore also permits more flexible trajectories for traits. The Halcyon method differs from the established Coevol method [2] in two main ways. First, Coevol treats the rate of sequence evolution itself as a trait undergoing Brownian motion, and therefore it attempts to auto-correlate rates of evolution across the tree. This makes sense if one expects the rate of evolution of e.g. a mouse and a rat to be more similar to each other than to e.g. an elephant, as the former species are both more genetically related and more physically similar to each other than they are to the latter. In contrast, Halcyon uses rate multipliers. Any similarity in the background/neutral rate of evolution between species can be encoded in the guide tree that is provided as input. Halcyon asks how a focal DNA sequence (i.e. the input alignment) behaves relative to this baseline. Where is the focal sequence accelerated and where is it conserved? The second difference between the two models is how they treat traits. Under both models traits un-

dergo a Brownian motion but under Coevol the question is 'are fast (slow) rates of nucleotide substitution associated with high trait values?' Under the Halcyon model the question is 'are fast (slow) rates of nucleotide evolution associated with high rates of change of a trait?' The question asked by Coevol seems appropriate for investigating life history traits but the one asked by Halcyon is arguably more relevant for investigating the genetic correlates of evolutionary innovation.

Acknowledgements

I thank the members of the Jun Liu and Scott Edwards group who asked helpful questions during the Winter 2023 lab meeting. I also thank Timothy Sackton and Subir Shakya for enthusiastic discussions. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Appendix 1

Pre-processing data for the Stan script. Assume null/background tree $\mathbf{T}$, having E edges, L leaves, $N = 2L - 1$ nodes, and a root node index of R. Arrange the edges of $\mathbf{T}$ for post order traversal (e.g. using `ape::postorder`) so that visiting edges in order $E, \ldots, 1$ ensures that ancestral branch $(\_, i)$ is always visited before its decedents (i, j) and (i, k).

Construct a boolean matrix $\mathbf{B}_{N \times E}$ with $\mathbf{B}_{i,e} = 1$ if edge e is ancestral to node i and $\mathbf{B}_{i,e} = 0$ otherwise. The matrix $\mathbf{B}$ is useful when calculating a variance-covariance matrix later. For example, use the following (pseudocode) procedure:

```
B = zeros(N, E)
for e in E,..,1
  (i,j)  = edge[e]
```

```
  B[j,:] = B[i,:] // all ancestral edges of i are also ancestral to j
  B[j,e] = 1 // edge e leads directly to node j
end
```

The nucleotide alignment $\mathbf{X}$ can be coded `1`–`4` for `a`, `c`, `g`, `t`. Trait data may be normalized and centred. The L rows of $\mathbf{X}$ and $\mathbf{B}$, and L elements of $\mathbf{y}$ must all share the same order as the tips of the tree.

The edge list, edge lengths, substitution rate matrix $\mathbf{Q}$, stationary distribution π, and boolean matrix $\mathbf{B}$ (passed as parameter `path_to_leaf`) are passed to Stan along with the alignment data $\mathbf{X}$, and trait data $\mathbf{y}$.

The Stan Halcyon script. Stan models contain a sequence of program statements that define a probability density function conditioned on observed data. These are converted into a sampler using the Stan tool chain.

Stan does not support discrete parameters (e.g. ancestral nucleotides coded as integers) nor directly support phylogenetic inference. However, one can introduce user defined functions using a familiar procedural syntax. For the Halcyon model I introduce two user defined marginalizations and one convenience function.

The first function computes the likelihood of alignment $\mathbf{X}$ using the Felsenstein algorithm [24], here named `Joe_F()`. The algorithm marginalizes across all possible ancestral nucleotide states. A minor complication to the usual implementation is that under the Halcyon model branch lengths must be rescaled using substitution rate multipliers $\mathbf{r}$.

```
real Joe_F(array[,] int X, vector r, int L, int N, int E, int R, int S,
  matrix Q, row_vector pi_distro, array[,] int edge, vector edge_length) {
  real log_p_X = 0.0;
  array[S] matrix[N,4] lp_Xup; // prob subtree at i if nuc(i) = j, indexed by site
  for (s in 1:S) {
    for (l in 1:L) {
```

```
      for (x in 1:4) {
        if (x == X[1,s]) { lp_Xup[s][1,x] = log(1.0); }
        else             { lp_Xup[s][1,x] = log(0.0); }
      }
    }
    int e = 1;
    while (e <= E) {
      int i = edge[e  ,1];
      int j = edge[e  ,2];
      int k = edge[e+1,2];
      real di = r[e  ] * edge_length[e  ];
      real dj = r[e+1] * edge_length[e+1];
      matrix[4,4] p_ij = matrix_exp(Q * di);
      matrix[4,4] p_ik = matrix_exp(Q * dj);
      for (x in 1:4) {  // x, j, k index rows
        real p_subj = dot_product(p_ij[x], exp(lp_Xup[s][j]));
        real p_subk = dot_product(p_ik[x], exp(lp_Xup[s][k]));
        lp_Xup[s][i,x] = log(p_subj) + log(p_subk);
      }
      e += 2;
    }
    log_p_X += log_sum_exp(lp_Xup[s][R] + log(pi_distro));
  }
  return log_p_X;
}
```

The second user defined function, named `sig_fn()`, is a notational convenience that encapsulates the logistic relationship between the variance rates σ_j^2 and the substitution rate

multipliers $\mathbf{r}$. The function is vectorized and returns a variance rate for every branch.

```
vector sig_fn(real h, real a, real b, vector r) {
  return h ./ (1 + exp(-a - b * log(r)));
}
```

The third user defined function, named `VCV()`, computes the variance-covariance matrix `Sig` used to calculate the likelihood of seeing trait values $\mathbf{y}$ at the tips of the tree, conditional on the scaled edge lengths obtained via `sig_fn`. Because `Sig` depends on the values of model parameters h, a, b, and $\mathbf{r}$, it needs to be constructed repeatedly during model inference.

```
matrix VCV(array[,] int Path, int L, int E, vector scaled_edge_length) {
  matrix[L,L] Sig;
  for (j in 1:L) {
    for (k in 1:j) {
      array[E] int common_edges;
      for (ii in 1:E) {
        common_edges[ii] =  Path[j,ii] * Path[k,ii];
      }
      real s = 0.0;
      for (ii in 1:E) {
        s = s + common_edges[ii] * scaled_edge_length[ii];
      }
      Sig[j,k] = s;
      Sig[k,j] = s;
    }
  }
  return Sig;
}
```

Given the above three functions only one interesting task remains, which is to specify the likelihood of the Halcyon model in a `model` block:

```
model {
  h  ~ uniform(0, 4);
  a  ~ normal(0, 2);
  b  ~ normal(0, 2);
  yR ~ normal(0, 1);
  r  ~ lognormal(0, 1);
  target += Joe_F(X, r, L, N, E, R, S, Q, pi_distro, edge, edge_length);
  matrix[L,L] Sig = VCV(path_to_leaf, L, E, edge_length .* sig_fn(h, a, b, r));
  y ~ multi_normal(rep_vector(yR, L), Sig);
}
```

References

[1] Smith SD, Pennell MW, Dunn CW, Edwards SV. Phylogenetics is the new genetics (for most of biodiversity). Trends in Ecology & Evolution. 2020;35(5):415-25.

[2] Lartillot N, Poujol R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Molecular Biology and Evolution. 2011;28(1):729-44.

[3] Partha R, Kowalczyk A, Clark NL, Chikina M. Robust method for detecting convergent shifts in evolutionary rates. Molecular Biology and Evolution. 2019;36(8):1817-30.

[4] Yusuf L, Heatley MC, Palmer JP, Barton HJ, Cooney CR, Gossmann TI. Noncoding regions underpin avian bill shape diversification at macroevolutionary scales. Genome Research. 2020;30(4):553-65.

[5] Gemmell P, Sackton TB, Edwards SV, Liu JS. A phylogenetic method linking nucleotide substitution rates to rates of continuous trait evolution. bioRxiv. 2023:2023-10.

[6] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biology. 2006;4(5):e88.

[7] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of statistical software. 2017;76.

[8] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press; 1998.

[9] Yang Z. Computational molecular evolution. OUP Oxford; 2006.

[10] Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in Bioinformatics. 2011;12(1):41-51.

[11] Felsenstein J. Phylogenies and the comparative method. The American Naturalist. 1985;125(1):1-15.

[12] Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592-3.

[13] Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289-90.

[14] Uyeda JC, Harmon LJ. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Systematic Biology. 2014;63(6):902-18.

[15] Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. Journal of the Royal Statistical Society: Series D (The Statistician). 1994;43(1):169-77.

[16] Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing. vol. 124. Vienna, Austria; 2003. p. 1-10.

[17] Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS book: A practical introduction to Bayesian analysis. CRC press; 2012.

[18] McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC; 2018.

[19] Gelman A, Hill J, Vehtari A. Regression and other stories. Cambridge University Press; 2020.

[20] Abelson H, Sussman GJ. Structure and interpretation of computer programs. The MIT Press; 1996.

[21] Thorne JL, Kishino H, Painter IS. Estimating the rate of evolution of the rate of molecular evolution. Molecular biology and evolution. 1998;15(12):1647-57.

[22] Charlesworth B, Charlesworth D. Elements of evolutionary genetics. vol. 42. Springer; 2010.

[23] Gemmell P, Hein J, Katzourakis A. Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split. Retrovirology. 2015;12:1-14.

[24] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution. 1981;17:368-76.