stats

Article

A Phylogenetic Regression Model for Studying Trait Evolution on Network

Dwueng-Chwuan Jhwueng

Department of Statistics, Feng-Chia University, Taichung 40724, Taiwan; dcjhwueng@fcu.edu.tw;
Tel.: +886-975-662-326

Abstract: A phylogenetic regression model that incorporates the network structure allowing the reticulation event to study trait evolution is proposed. The parameter estimation is achieved through the maximum likelihood approach, where an algorithm is developed by taking a phylogenetic network in eNewick format as the input to build up the variance–covariance matrix. The model is applied to study the common sunflower, Helianthus annuus, by investigating its traits used to respond to drought conditions. Results show that our model provides acceptable estimates of the parameters, where most of the traits analyzed were found to have a significant correlation with drought tolerance.

Keywords: regression model; phylogenetic comparative analysis; variance–covariance matrix; reticulate evolution; Brownian motion

check for updates

Citation: Jhwueng, D.-C. A Phylogenetic Regression Model for Studying Trait Evolution on Network. *Stats* **2023**, *6*, 450–467. https://doi.org/10.3390/stats6010028

Academic Editor: Wei Zhu

Received: 21 February 2023
Revised: 13 March 2023
Accepted: 17 March 2023
Published: 18 March 2023

1. Introduction

Hybridizations among closely related species have frequently occurred in nature. Under Mayr's biological species concept, hybrid species can be defined as organisms formed by cross-fertilization between individuals of different species [1,2]. Hybrid speciation occurs in at least two ways: allopolyploid speciation and diploid (homoploid) hybrid speciation. While allopolyploidy is hybrid speciation between two species resulting in a new species that has the complete diploid chromosome complement of both its parents, diploid hybrid speciation results from a normal sexual event in which each gamete has a haploid complement of the nuclear chromosomes from its parent, but gametes that form the zygote come from different species [3]. This means that, in hybrid speciation, the new species may have the same number of chromosomes as its parent (diploid hybridization) or the sum of the number of chromosomes of its parents (polyploid hybridization).

Phylogenetic comparative methods (PCMs) are commonly applied to study correlated trait evolution; most methods were developed by incorporating a phylogenetic tree to represent the affinity among a group of related species [4–6]. However, if evolution involved ancient hybridizations, then we cannot simply use the phylogeny to represent the affinity among species, but instead should use the phylogenetic network (which is a directed acyclic graph, coupled with time constraints). Currently, in the literature, we can observe the development of statistical methods using phylogenetic networks to investigate trait evolution including the hybridization process [7–10]. Note that approaches to phylogenetic analysis typically involve constructing networks using molecular data [11,12], while our approach employs the given phylogenetic network with known topology and branch lengths to study the evolution of traits.

The objective of our research is to examine the evolution of traits in both hybrid and non-hybrid species, specifically through the lens of reticulation evolution. This phenomenon involves the merging of genetic material from different species, resulting in the creation of hybrid offspring that exhibit a unique combination of traits inherited from their parents. Our study aims to investigate the implications of reticulation evolution for correlated trait evolution in a linear regression framework.

The paper is organized as follows. In Section 2, we model the hybrid on the given phylogenetic network and create a phylogenetic regression model to analyze trait data that account for the hybrid information. In Section 3, a heuristic algorithm is proposed to build the variance–covariance matrix given a phylogenetic network and we propose a maximum likelihood framework for parameter estimation. In Section 4, the novel regression model is applied to study the drought tolerance of sunflowers. The discussion for this work is provided in Section 5, and the conclusions are given in Section 6.

2. Model

2.1. Relation between the Hybrid and Its Parents

Figure 1 displays a phylogenetic network that illustrates the connection between three species—X, R, and Y. Species R is a hybrid of species X and Y, and it came into existence at a specific time, $t = t_1$. The root node O served as the ancestor for the three species. The purpose of the network is to show the relationships between the species.

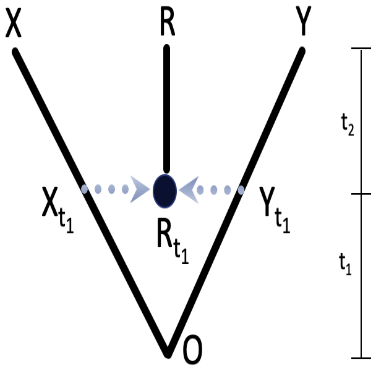

Figure 1. A three-taxa phylogenetic network. The hybrid species R of X and Y on the tips of the network was formed at $t = t_1$.

To model trait evolution with hybridization, we treat the hybrid node on the phylogenetic network by allowing a burst of new variation at the hybridization event. We achieve this by incorporating a hybridization parameter τ. Consider that the trait of the hybrid species is defined in the log scale [7,8] via $\log R = \gamma \log X + (1 - \gamma) \log Y + \log \tau$, where γ is the proportion of the hybrid trait inherited from parent X (i.e., $1 - \gamma$ is the proportion of the hybrid trait inherited from parent Y), and τ is denoted as the hybridization parameter that is designed to model an increase in the variance of the hybrid species.

In raw scale modeling, the relation can be expressed by exponentiation to obtain $R = \tau X^{\gamma} Y^{1-\gamma}$. For a setting, we use $\gamma = 0.5$, where the hybrid was assumed to be inherited equally from both parents. The arithmetic–geometric inequality establishes that $R = \tau\sqrt{XY} \leq \frac{\tau}{2}(X + Y)$. As τ typically ranges between $(0, \infty)$, it follows that $\frac{\tau}{2}$ shares this range with τ. Because the quantitative phenotypic traits are inherently non-negative, the inequality of arithmetic and geometric means condition is met. By incorporating a model that permits variation in the hybrid's variance to be computed from an additive operation on X, Y through τ, we establish the relationship between the hybrid species R and its parent organisms X and Y in Equation (1):

$$R = \tau(X + Y). \tag{1}$$

By incorporating this additive structure, below, we provide an approach to modeling hybrid trait evolution. In Equation (1), the affinity among species at time t in the phylogenetic network can be derived as follows. For any other species Z, the affinity between Z and R is

$$Cov(R, Z) = Cov(\tau(X + Y), Z) = \tau\{Cov(X, Z) + Cov(Y, Z)\} \tag{2}$$

In particular, when $Z = R$, we have

$$Var(R) = Var(\tau(X+Y)) = \tau^2\{Var(X) + Var(Y) + 2Cov(X,Y)\}. \tag{3}$$

Given a phylogenetic network $\mathbb{N}$ of n taxa, one can use Equations (2) and (3) to derive the corresponding similarity matrix $\boldsymbol{G}_{\tau,n} = [g_{\tau,ij}]$, where $g_{\tau,ij}, i, j = 1, 2, \cdots, n$ describe the affinities between taxa i and j, possibly with hybrid species.

Below, we use the Brownian motion (BM) in modeling trait evolution [5,13,14] with the definition in Equation (1) to construct the model and variance–covariance matrix for a group of related species in Section 2.2.

2.2. Covariance Matrix under the Brownian Motion Model

Under the assumption of the BM process for trait evolution[15], we can define $X := X_t, Y := Y_t$ as stochastic variables with $X_t = X_0 + \sigma_x \epsilon_t^X, Y_t = Y_0 + \sigma_x \epsilon_t^Y$, where $X_0 = Y_0$ is the ancestral value at the root of the tree, σ_x and σ_y are parameters of the rate of evolution, and ϵ_t^X and ϵ_t^Y are the Brownian motion variables with $E[\epsilon_t^X] = E[\epsilon_t^Y] = 0$ and $Var[\epsilon_t^X] = Var[\epsilon_t^Y] = t$ for trait X and Y, respectively.

Given the network with a known topology and branch length (times) as shown in Figure 1, we have $Var(X) := Var(X_{t_1+t_2}) = \sigma^2(t_1+t_2)$, $Var(Y) := Var(Y_{t_1+t_2}) = \sigma^2(t_1+t_2)$, and $Cov(X,Y) = 0$ as X and Y are independent. Since the hybrid R is produced at time $t = t_1$, the variation in the hybrid R is decomposed into two parts: one comes from its parent at t_1 and the other comes from its evolution from t_1 to $t_1 + t_2$. Hence, we have $Var(R) := Var(R_{t_1+t_2}) = Var(R_{t_1}) + Var(R_{[t_1,t_1+t_2]}) = Var(\tau(X_{t_1} + Y_{t_1})) + \sigma^2(t_1 + t_2 - t_1) = \tau^2\{Var(X_{t_1}) + Var(Y_{t_1}) + 2Cov(X_{t_1}, Y_{t_1})\} + \sigma^2 t_2 = \tau^2\sigma^2(t_1 + t_1 + 2\cdot 0) + \sigma^2 t_2 = (2\tau^2 t_1 + t_2)\sigma^2$.

Since evolution on different branches occurs independently, the covariation between the hybrid and its parents is $Cov(Y,R) = Cov(X,R) = Cov(X_{t_1}, R_{t_1}) = Cov(X_{t_1}, \tau(X_{t_1} + Y_{t_1})) = \tau[Cov(X_{t_1}, X_{t_1}) + Cov(X_{t_1}, Y_{t_1})] = \tau[Var(X_{t_1}) + 0] = \tau\sigma^2 t_1$. Therefore, with Equations (1)–(3), the corresponding similarity matrix $\boldsymbol{G}_{\tau,3}$ is obtained as in Figure 1.

$$\boldsymbol{G}_{\tau,3} = \begin{array}{c} \\ X \\ R \\ Y \end{array} \begin{array}{c} \begin{array}{ccc} X & R & Y \end{array} \\ \begin{pmatrix} t_1+t_2 & \tau t_1 & 0 \\ \tau t_1 & t_2 + 2\tau^2 t_1 & \tau t_1 \\ 0 & \tau t_1 & t_1 + t_2 \end{pmatrix} \end{array}. \tag{4}$$

Previous work has explained trait evolution in a logarithmic scale, using different parameter notations for the hybrid vigor [7–9], while we use τ. However, it is worth noting that both of these prior methods do account for the hybrid effect. Our proposed approach offers an alternative method of constructing the variance–covariance matrix, which differs from the methods used in the literature. We must acknowledge that our method has a limitation in its ability to handle gene flow, as it can only account for reticulation events. This limitation has been discussed in the literature [8].

2.3. Stepwise Procedure for Constructing the Variance–Covariance Matrix

In this study, we present a novel method for constructing the variance–covariance matrix using a matrix multiplication technique. The proposed approach involves a three-step process, as illustrated in Figure 2.

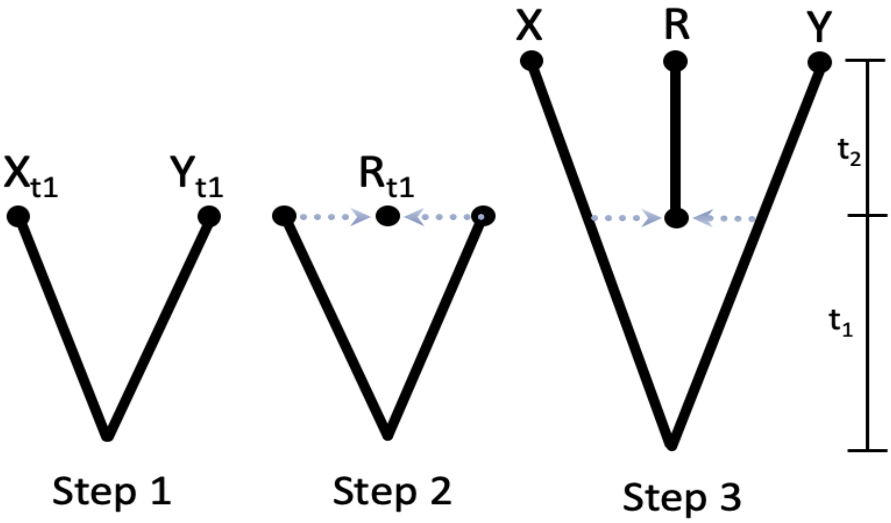

Figure 2. Evolution scenario for 3-taxa phylogenetic network containing a reticular hybridization event.

Figure 2 describes an evolutionary scenario for a phylogenetic network with three taxa, which involves a reticular hybridization event. The scenario consists of three main steps:

1. Step 1: the root O speciates into two distinct taxa denoted as X_{t1} and Y_{t1}.
2. Step 2: a hybrid species, denoted as R, is produced as a result of hybridization between X and Y at a specific time point, denoted as t_1.
3. Step 3: after $t = t_1$, the three species X, R, and Y continue to evolve without undergoing any further speciation or hybridization, ultimately reaching the current time point of $t = t_1 + t_2$.

Note that this calculation of the covariance matrix is a *three-step process* [16], with both steps able to be described using matrix operations.

First, in step 1 in Figure 2, a speciation at the root yields two species X, Y at t_1 with the covariance in Equation (5):

$$\boldsymbol{G}_2 = \begin{array}{cc} & \begin{array}{cc} X & Y \end{array} \\ \begin{array}{c} X \\ Y \end{array} & \begin{pmatrix} t_1 & 0 \\ 0 & t_1 \end{pmatrix} \end{array}. \tag{5}$$

Next, in step 2, there is the instantaneous hybridization event at time t_1. This can be accomplished mathematically by multiplying the previous 2-by-2 matrix describing the variance $\boldsymbol{G}_2$ in Equation (5) in X and Y by a 3×2 path matrix $\boldsymbol{K}_{2,\tau}$ on the left and $\boldsymbol{K}_{2,\tau}^t$ on the right:

$$\boldsymbol{G}_{2,\tau} = \boldsymbol{K}_{2,\tau}\boldsymbol{G}_2\boldsymbol{K}_{2,\tau}^t = \begin{array}{cc} & \begin{array}{ccc} X & R & Y \end{array} \\ \begin{array}{c} X \\ R \\ Y \end{array} & \begin{pmatrix} t_1 + t_2 & \tau t_1 & 0 \\ \tau t_1 & 2\tau^2 t_1 & \tau t_1 \\ 0 & \tau t_1 & t_1 + t_2 \end{pmatrix} \end{array}, \tag{6}$$

where $\boldsymbol{K}_{2,\tau}$ is shown in Equation (7)

$$\boldsymbol{K}_{2,\tau} = \begin{array}{cc} & \begin{array}{cc} X & Y \end{array} \\ \begin{array}{c} X \\ R \\ Y \end{array} & \begin{pmatrix} 1 & 0 \\ \tau & \tau \\ 0 & 1 \end{pmatrix} \end{array}. \tag{7}$$

Finally , the last step is elongation by adding $t_2\mathbf{I}_3$, where $\mathbf{I}_3$ is the 3-by-3 identity matrix. The corresponding covariance structure is shown in Equation (8):

$$G_{3,\tau} = \begin{array}{c} \\ X \\ R \\ Y \end{array} \begin{array}{c} \begin{array}{ccc} X & R & Y \end{array} \\ \begin{pmatrix} t_1 + t_2 & \tau t_1 & 0 \\ \tau t_1 & 2\tau^2 t_1 + t_2 & \tau t_1 \\ 0 & \tau t_1 & t_1 + t_2 \end{pmatrix} \end{array}. \tag{8}$$

Alternatively, standard speciation events, as depicted in Figure 3, can be analyzed using analogous matrix operations.

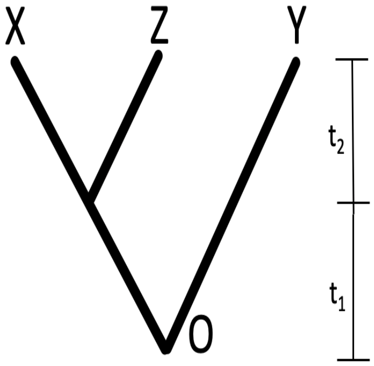

Figure 3. A phylogenetic tree of 3 taxa. X, Z, Y are taxa; O is the root. X and Z share the same branch length on t_1. Y is independent with both X and Z.

The instantaneous speciation event shown in Figure 3 at time t_1 is accomplished by multiplying on the left and on the right by the transpose of the matrix:

$$K_{2,\tau=1} = \begin{array}{c} \\ X \\ Z \\ Y \end{array} \begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \end{array}. \tag{9}$$

For the tree case with only speciation, as shown in Figure 3, one can construct the similarity matrix G_3 in Equation (10):

$$G_3 = K_{2,\tau=1} G_2 K^t_{2,\tau=1} = \begin{array}{c} \\ X \\ Z \\ Y \end{array} \begin{array}{c} \begin{array}{ccc} X & Z & Y \end{array} \\ \begin{pmatrix} t_1 + t_2 & t_1 & 0 \\ t_1 & t_1 + t_2 & 0 \\ 0 & 0 & t_1 + t_2 \end{pmatrix} \end{array}. \tag{10}$$

These operations can be generalized to the k existing species case whenever the $k+1$ taxon arises by hybridization or speciation. Since the form of K changes depending on whether the hybridization of speciation is involved, we adopt the following notation: let K_j denote the $(j+1)$ by j matrix obtained from the j by j identity matrix by inserting a row with a one in column j and zeros elsewhere, where column j denotes the taxon involved in the speciation event. Let $K_{j,\tau}$ denote the $(j+1)$ by j matrix obtained from the j by j identity matrix by inserting a row with τ in columns i and j and zeros elsewhere, where columns i and j denote the taxa involved in the hybridization event. Then, the adjustment from time $t_1 + \cdots + t_{j-1}$ to time t_j is as given in Equation (11):

$$G_{j,\tau} = K_{j-1} G_{j-1,\tau} K^t_{j-1} + t_j \mathbf{I}_j, \tag{11}$$

where, for hybridization, $K_{j-1} = K_{j-1,\tau}$, and for speciation, $K_{j-1} = K_{j-1,\tau=1}$, which sets $\tau = 1$, as is evident from Equation (9) when we compare it with Equation (7).

Our proposed methodology can indeed handle a general ultrametric phylogenetic network with an arbitrary number of hybrid nodes, as later demonstrated in the case study with 13 species and 3 hybrid species in Section 4, as well as the six-taxa network with two hybrids presented in Appendix A.2.

2.4. The Statistical Model and Likelihood Function

Under the regression model framework, let $Y = (y_1, y_2, ..., y_n)^t$ be the trait values for n species, some of which are possibly old hybrids. Let $\boldsymbol{X} = [\mathbf{1}, X_1, X_2, \cdots, X_p]$ be the $n \times k$ design matrix from the covariate trait, where $\mathbf{1} = (1, 1, \cdots, 1)^t \in \mathcal{R}^n$ is the vector of 1s, and we have

$$Y \sim \boldsymbol{X}\beta + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{G}_\tau). \tag{12}$$

Let $\boldsymbol{\theta} = (\tau, \sigma, \beta)$, and the negative log-likelihood function given the traits $Y, \boldsymbol{X}$ and network $\mathbb{N}$ is

$$-\log L(\boldsymbol{\theta}|Y, \boldsymbol{X}, \mathbb{N}) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\sigma^2 + \frac{1}{2}\log|\boldsymbol{G}_\tau| + \frac{1}{2\sigma^2}(Y - \boldsymbol{X}\beta)^t \boldsymbol{G}_\tau^{-1}(Y - \boldsymbol{X}\beta). \tag{13}$$

The least-square estimate is shown in Equation (14):

$$\hat{\beta} = (\boldsymbol{X}^t \boldsymbol{G}_\tau^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{G}_\tau^{-1} Y. \tag{14}$$

As the model assumes a Gaussian process distribution, the estimation of model parameters can be conducted through maximum likelihood inference, utilizing the Nelder–Mead optimization method in the R software [17]. One of the Nelder–Mead optimization's primary benefits is that it can be utilized in a variety of problem settings, without requiring knowledge of the objective function's derivatives. In our specific likelihood function, the covariance matrix contains embedded parameters denoted by τ.

We use maximum likelihood analysis to estimate the hybridized parameter τ by optimizing the negative log-likelihood function, where $|\boldsymbol{G}_\tau|$ is the determinant of $\boldsymbol{G}_\tau$. We set the bound for τ as $[0, 10]$ for the purpose of optimization. We use the golden section method to search for the maximum likelihood estimator (MLE) of the negative log-likelihood function for the Brownian motion model.

Let $J(\tau, \sigma) \equiv -\log L(\tau, \sigma|\beta, Y, \boldsymbol{X}, \mathbb{N})$. By taking the partial differentiation of $J(\tau, \sigma)$ with respect to τ and σ, the Hessian matrix can be obtained:

$$\mathcal{H}(\tau^*, \sigma^*) = \left.\frac{\partial^2 J(\tau, \sigma)}{\partial\tau\partial\sigma}\right|_{(\tau,\sigma)=(\tau^*,\sigma^*)} = \Sigma^{-1}_{\tau^*,\sigma^*}, \tag{15}$$

which is useful to compute the variance of parameters τ, σ for further inference.

For the Gaussian random variable (here, the Brownian motion), the second derivatives of the objective function are constant for (τ, σ) because the objective function is a quadratic function (τ, σ). Therefore, the Hessian matrix can be computed without obtaining the mean vector (τ^*, σ^*). We apply the `R` function `hessian` [18] to compute the Hessian matrix.

It is known that under regularity conditions (smoothness of the likelihood function) [19], the estimator $\hat{\beta}$ (by iterating a finite number of times) is asymptotically distributed as $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{V})$ where n is the taxa size and $\sigma^2 \boldsymbol{X}^t \boldsymbol{G}_\tau^{-1} \boldsymbol{X}$ converge to $\boldsymbol{V}$ in probability. It is assumed that the response variable Y is continuous and that the error terms $\boldsymbol{\epsilon}$ are normally distributed with a mean of 0 and a covariance matrix of $\sigma^2 \boldsymbol{G}_\tau$, which means that the Brownian motion assumption is applied to each tip variable y_i in the response trait vectors Y. The predictor variables X_i are non-stochastic and fixed. Based on these assumptions, the likelihood of the linear regression model is given by an equation, Equation (13), and in order to show that this equation meets the regularity conditions, several properties must be satisfied. The likelihood function must be well-defined, non-negative, continuous in β and $\sigma^2 \boldsymbol{G}_\tau$, and differentiable with respect to β, σ^2, and τ separately. These properties are satisfied because the likelihood function is a product of non-negative terms, the exponential function is always positive, and the sum of continuous functions is continuous. Additionally, the derivatives of the likelihood function with respect to β and σ^2 are continuous. However, we note to the reader that the regularity condition's likelihood function in Equation (13) depends on a certain range of the parameters τ for the network models proposed here and in the literature [7,8]. The derivative of the likelihood function with respect to τ involves the inverse of the covariance matrix $\boldsymbol{G}_{\tau}^{-1}$, which depends on the network structure. First, $\boldsymbol{G}_{\tau}$ is symmetric as a covariance matrix. If $\boldsymbol{G}_{\tau}$ is a positive definite matrix, the derivative of the likelihood with respect to τ will be continuous. The inference can be used to infer the regression effect pending the condition of the τ. In the empirical analysis, we verify the positive definite property of the $\boldsymbol{G}_{\tau}$.

According to Varga and Nabben [20] and Nabben and Varga [21], if the covariance matrix $\boldsymbol{G}_{\tau}$ is an ultrametric matrix, meaning that it satisfies certain mathematical inequalities (i.e., $\boldsymbol{G}_{\tau}[i,i] > \max\{\boldsymbol{G}_{\tau}[i,k], k \neq i\}$ for all $i = 1, 2, \cdots, n$, $\boldsymbol{G}_{\tau}[i,j] \geq \min\{\boldsymbol{G}_{\tau}[i,k], \boldsymbol{G}_{\tau}[k,j]\}$) for all i, j, and k), then the derivative of the likelihood function with respect to τ will be continuous. This is because ultrametricity implies the stronger condition of the triangle inequality, which ensures that the matrix is always positive definite and has no negative eigenvalues. To ensure that all regularity conditions are met, it would be ideal to determine the parameter space for τ that would make $\boldsymbol{G}_{\tau}$ ultrametric before analysis. However, this strict condition depends on the given network and cannot be solved analytically in general. For example, in the case of a three-taxon network, as shown in Equation (8), the parameter space for τ would need to be constrained to $\tau \in \{\tau : (t_1 + t_2) > \tau t_1; 2\tau^2 t_1 + t_2 > \tau t_1\}$ to meet the ultrametric condition.

3. Algorithm and Inference

An extended Newick format (eNewick) uses unique syntax to represent a given phylogenetic network in linear form [22]. A phylogenetic network can be transformed into a phylogenetic tree with some replicated nodes, adequately tagged according to the hybrid nodes, and then traversing the resulting phylogenetic network in postorder to obtain the eNewick description of the phylogenetic network. We modified their representation in the function `newick2phylog` in the `ade4` package [23] in the R software to obtain the eNewick format. The function `Newick2phylog` [23] in the `ade4` package of the R software program was designed to read in phylogenies in Newick format and return an array with three columns, where the first column contains the ancestral nodes and the second and third columns have the two descendants of the corresponding ancestor. Note that the number of rows (ancestors) in this array is $n - 1 + 2k$ as a hybrid node requires two incoming ancestors while a species node only has one ancestor. The root is also included in the count. To provide an example, in a $n = 3$ taxa network with one hybrid ($k = 1$), as in Figure 1 , we have the number of rows equal to 4, which is calculated as $3 - 1 + 2 \times 1$. This is also shown in the following Table 1.

Table 1. Ancestral–descendant relationship corresponding to Figure 1.

Rows	Parent	Descendant 1	Descendant 2
1	O	X_{t_1}	Y_{t_1}
2	X_{t_1}	X	R_{t_1}
3	Y_{t_1}	Y	R_{t_1}
4	R_{t_1}	R	R

The algorithm can generate the covariance matrix $\boldsymbol{G}_{n,\tau}$ by starting from the root, adding a new node in each step, and terminating until the desired matrix of n species is built. For the tree case, each descendant has a unique ancestor. For the node with the reticulated event, the function reads a descendant such as a hybrid species with two ancestors; in one of the ancestral rows, the descendant will be listed by name, and in the other row, the descendant will have a _1 attached to the end of the name. After determining the ancestral–descendant relationships, we find the times from the root at which speciation events or hybridization events occur: t_1, $t_1 + t_2$, $t_1 + t_2 + t_3$, $\cdots$, etc. Note that there are $n-1$ branches, and we build the phylogenetic similarity matrix $\boldsymbol{G}_{n,\tau}$ up from the root. For times $t < t_1$, there are two species present whose evolution is independent given the root. The relationship matrix up until t_1 is thus a 2×2 diagonal matrix with t on the diagonal. For each event, we adjust the similarity matrix according to Equation (11) for the Brownian motion model as follows to generate the variance–covariance matrix $\boldsymbol{G}_{n,\tau}$ for n tips by starting with the root, adding a new node at each *speciation* or *hybridization* event, and terminating when the process reaches the tips. A concrete example with detailed illustration is provided in Appendix A.2.

Our proposed methodology uses a feasible generalized least-squares approach to estimate the model parameters τ and σ, as well as the regression parameter β, through a joint estimation approach. An alternating search procedure is utilized to simultaneously obtain the estimate for $\hat{\beta}$ and the covariance by maximizing the likelihood of the model parameters and minimizing the squared residuals of the regression parameters, as illustrated in Algorithm 1.

Algorithm 1 Procedure for Parameter Estimation.

```
Require: Predictive traits $\boldsymbol{X} = [X_1, X_2, \cdots, X_p]$, and $Y$, network $\mathbb{N}$.
Ensure: Regression estimator $\hat{\beta}$, hybrid vigor estimator $\hat{\tau}$, and rate estimator $\hat{\sigma}$.
1: Get ordinary least-square estimates $\hat{\beta}_0 = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^tY$, $\hat{\sigma}_0 = \sqrt{\frac{n-p}{n}\hat{\epsilon}^t\hat{\epsilon}}$ where
   $\hat{\epsilon} = Y - \boldsymbol{X}\hat{\beta}_0$, $p$ is the number of covariates.
2: Set $\tau_0 = 0.1$.
3: Use the tree traversal algorithm with Equation (11) to construct the variance–covariance
   matrix $\boldsymbol{G}_\tau$.
4: Compute $\ell_0 = -\log L(\tau_0, \hat{\sigma}_0 | \hat{\beta}_0, Y, \boldsymbol{X}, \mathbb{N})$
5: Apply the Nelder–Mead method to search the maximum likelihood $\hat{\tau}$ and $\hat{\sigma}$ and let
   $\ell_1 = -\log L(\hat{\tau}, \hat{\sigma} | \hat{\beta}_0, Y, \boldsymbol{X}, \mathbb{N})$ in Equation (13).
6: Use $\hat{\tau}$ to compute the GLS estimate $\hat{\beta}' = (\boldsymbol{X}^t\boldsymbol{G}_{\hat{\tau}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{G}_{\hat{\tau}}^{-1}Y$.
7:    if $||\hat{\beta}' - \hat{\beta}_0||^2 < 10^{-5}$
8:       return $\hat{\tau}, \hat{\sigma}, \hat{\beta}'$.
9:    else
10:      if $\ell_1 < \ell_0$
11:          set $\tau_0 = \hat{\tau}, \sigma_0 = \hat{\sigma}$.
12:          Set $\ell_{10} = -\log L(\hat{\tau}, \hat{\sigma} | \hat{\beta}_0, Y, \boldsymbol{X}, \mathbb{N})$ and $\ell_{11} = -\log L(\hat{\tau}, \hat{\sigma} | \hat{\beta}', Y, \boldsymbol{X}, \mathbb{N})$
13:             if $\ell_{11} < \ell_{10}$
14:                 Set $\hat{\beta}_0 = \hat{\beta}'$ and go to step 4.
15:             else Go to step 4.
```

4. Empirical Analysis

Hybridization is common in nature, with at least 25% of plant species showing hybridization. Sunflowers are an example of a species that has adapted to a wide range of environmental conditions, including soil types, temperature, and salinity. Studies show that hybridization frequently occurs among sunflowers, resulting in genetically hybrid species. Sunflowers have various uses, including traditional Chinese medicine, edible oil, and soil phytoremediation [24]. The family of Helianthus is the subject of ongoing research on the adaptation of hybrid species to their environment. Sunflowers, in particular, have adapted to tolerate drought and salty conditions in their habitats with lower precipitation levels. Selective sweeps in sunflowers have revealed candidate genes for adaptation to drought and salt tolerance [25]. Studies have also shown that sunflowers vary in their tolerance to drought [26].

The study focused on exploring the correlation between traits and drought tolerance, with soil moisture, precipitation, and rainfall in the area considered as possible factors that affect the response variable, Y. The precipitation data used as the covariates were collected from the `WorldClim` database [27,28]. The geographical data of the longitude and latitude of sunflowers were collected from the Global Biodiversity Information Facility (GBIF) database [29], and the R package `raster` [30,31] was used to download the corresponding data for analysis. To further investigate sunflowers' adaptation to drought tolerance conditions, a phylogenetic regression method was proposed, which can analyze trait data from both hybrid and typical species in the evolutionary mechanism. This method was applied to study a group of common sunflowers, *Helianthus annuus*, using data from the `efloras` database [32]. The collected traits include the plant height, petiole, pedicel, hemispherical bract, bract, stalk, leaf, ray flower, disk, corolla, and calyx achene of sunflowers. The predictor variable used in the study was the annual precipitation amount measured in various locations, which was obtained using the `raster` package from the `WorldClim` database. For example, the precipitation data for uncommon species located at 38.68 latitude degrees and -110.54 longitude degrees were obtained with a setting resolution of 0.5 minutes.

The presented data in Table 2 showcase the response traits of sunflowers, including various characteristics such as annuals, petioles, peduncles, involucres, phyllaries, paleae laminae, ray florets, disc florets, corollas, cypselae, and pappi. The covariate trait in question is the annual precipitation (`AnnPrec`), which represents the yearly precipitation levels at the location of the observed sunflowers.

Table 2. Sunflowers and their traits. Each column represents a sunflower species, while each row records the trait collected from the database.

	Praecox	**Debilis**	**Neglectus**	**Petiolaris**	**Anomalus**	**Deserticol**	**Paradoxus**	**Annuus**	**Argophyllus**	**Bolanderi**	**Exilis**
AnnPrec	1796.71	978.25	148.00	384.80	393.25	154.00	229.00	459.62	695.25	444.67	829.00
Annuals	95.00	7.00	27.50	15.50	34.50	7.25	21.00	13.50	35.00	5.50	2.90
Petioles	1.35	115.00	4.00	29.50	16.00	25.00	7.75	17.50	15.50	30.00	4.75
Peduncles	2.85	1.85	140.00	9.50	25.00	12.00	30.00	9.50	34.00	26.00	150.00
Involucres	6.25	4.50	3.00	120.00	3.00	9.50	17.00	19.50	6.00	17.50	20.00
Phyllaries	75.00	5.25	3.75	2.25	42.50	3.10	6.50	23.50	17.00	7.50	27.50
Paleae	9.50	25.00	7.15	6.80	3.25	25.00	3.50	2.00	19.00	17.00	8.50
Laminae	20.00	10.00	25.00	5.75	4.50	2.05	165.00	3.75	15.00	17.50	20.00
Ray florets	8.50	25.00	16.00	50.00	5.25	3.50	2.70	200.00	11.00	11.00	27.50
Disc florets	25.00	10.00	37.50	23.50	150.00	6.50	4.50	2.75	200.00	6.00	5.00
Corollas	25.00	27.50	10.50	25.00	17.50	150.00	7.00	5.00	2.35	105.00	2.50
Cypselae	8.00	21.00	14.00	10.00	17.00	14.50	75.00	6.00	4.00	2.35	65.00
Pappi	1.60	8.00	17.50	14.50	9.75	17.00	11.50	50.00	5.00	3.25	2.20

This dataset offers valuable insights into the relationship between the response traits of sunflowers and the annual precipitation levels in their growing location. Such findings could have significant implications for plant breeding and cultivation in regions with varying levels of precipitation. As such, a thorough analysis of the presented data can provide critical information that can contribute to the development of more robust and resilient plant species in the future. In light of this, further investigation and exploration of the data presented in Table 2 are warranted, as they may reveal essential correlations and trends that can deepen our understanding of sunflowers and their responses to varying levels of precipitation.

The network in Figure 4 is a modification from [33], where 11 sunflowers species are given at the genus level.

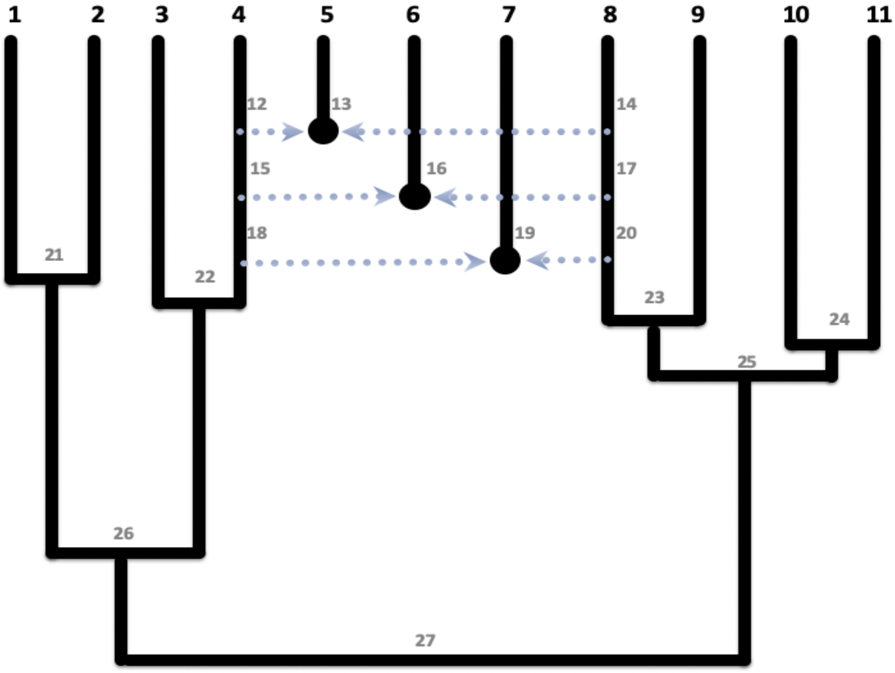

Figure 4. Sunflower network regraphed from [33]. Species on the tip from the leftmost (labeled with the number 1) to the rightmost (labeled with the number 1) are 1. praecox, 2. debilis, 3. neglectus, 4. petiolaris, 5. anomalus, 6. deserticola, 7. paradoxus, 8. annuus, 9. argophyllus, 10. bolanderi, and 11. exilis, where deserticola, anomalus, and paradoxus are hybrids from petiolaris and annuus. The eNewick format is $(1:0,((2:0.84,((3:0.23,4:0.23)16:0.55,(5:0.32,(6:0.15,(7:0.15)13)12:0.18:0.16)17:0.45)20:0.06)21:0.1,(((13,8:0.15:0.35)14:0.2,9:0.35)18,(10:0.16,11:0.16)15)19:0.43)22:0.06)23:0$.

To investigate whether precipitation has a significant impact on traits, it is necessary to check whether the regression slope is zero, represented by the null hypothesis $H_0 : \beta_1 = 0$. The results of the analysis using the phylogenetic regression model are presented in Table 3. The table reports GLS estimates for β, along with its 95% confidence interval, as well as estimates for the rate parameter σ and the hybrid parameter τ.

Table 3. The table provides estimates and corresponding standard errors for the hybrid effect ($\hat{\tau}$), the rate of evolution ($\hat{\sigma}$), and the slope ($\hat{\beta}$) for each of the 12 response traits of sunflowers under a network relationship. The slope estimate represents the effect of precipitation on the particular trait, with a positive value indicating a positive relationship and a negative value indicating a negative relationship. The 95% confidence interval (CI) for the slope estimate provides a range of plausible values for the true effect of precipitation on the trait.

Response Trait	$\hat{\tau}$ ($se_{\hat{\tau}}$)	$\hat{\sigma}$ ($se_{\hat{\sigma}}$)	$\hat{\beta}_1$ **(CI)**	**Significant?**
Annuals	0.841(0.07)	0.133(0.087)	0.243(0.038, 0.449)	Yes
Petioles	0.787(0.157)	0.106(0.069)	−0.091(−0.22, 0.038)	No
Peduncles	0.801(0.176)	0.157(0.103)	0.451(0.165, 0.737)	Yes
Involucres	1.068(0.038)	0.038(0.025)	0.123(0.106, 0.141)	Yes
Phyllaries	0.937(0.042)	0.048(0.031)	0.068(0.041, 0.096)	Yes
Paleae	1.005(0.035)	0.026(0.017)	−0.086(−0.094, −0.079)	Yes
Laminae	1.031(0.055)	0.061(0.04)	−0.019(−0.064, 0.026)	No
Ray florets	0.812(0.046)	0.057(0.037)	−0.038(−0.076, −0.001)	Yes
Disc florets	0.779(0.056)	0.106(0.069)	−0.006(−0.137, 0.124)	No
Corollas	1.065(0.052)	0.032(0.021)	0.048(0.036, 0.06)	Yes
Cypselae	1.221(0.11)	0.053(0.034)	0.071(0.038, 0.105)	Yes
Pappi	1.149(0.159)	0.051(0.033)	−0.019(−0.05, 0.012)	No

Table 3 provides the estimates of hybrid effect $\hat{\tau}$, rate of evolution $\hat{\sigma}$, and slope $\hat{\beta}$, along with their corresponding standard errors for different response traits in a study. The table also provides information about whether the slope estimate is statistically significant or not (significant set to Yes or No) at the 5% significance level.

For example, for the response trait "Annuals", the hybrid effect estimate is $\hat{\tau} = 0.841$ with a standard error of 0.07, indicating that the response of annuals has moderate hybrid weakness among sunflower species. The rate of evolution estimate is $\hat{\sigma} = 0.133$ with a standard error of 0.087, indicating that the evolutionary rate of annuals is relatively slow. The slope estimate is $\hat{\beta} = 0.243$ with a 95% CI of $(0.038, 0.449)$, suggesting that precipitation has a significant positive effect on the trait. The significance of the effect is indicated by the "Significant?" column, which shows "Yes" for a significant effect based on the 95% confidence interval of the slope estimate.

Similarly, the second-to-last row for the response trait Cypselae indicates that the hybrid effect estimate $\hat{\tau}$ is 1.221 (hybrid vigor) with standard error 0.11, the rate of evolution estimate σ is 0.053 with standard error 0.034, and the slope estimate $\hat{\beta}$ is 0.071 with a 95% confidence interval $(0.038, 0.105)$. Additionally, the slope estimate is statistically significant (significance set to Yes) for this trait.

In summary, the table provides estimates and corresponding standard errors for the hybrid effect, rate of evolution, and slope, along with their significance levels for different response traits in a study. These estimates can be used to make inferences about the relationship between the variables being studied and the response traits under consideration.

We further evaluate the correlations among the parameter estimates $\hat{\tau}, \hat{\sigma}$, and $\hat{\beta}_1$ using the 12 sunflower trait datasets; there is a moderate positive correlation (0.73) between the rate of evolution (σ) and the regression slope (b_1), suggesting that an increase in the rate of evolution is associated with an increase in the magnitude of the regression slope. There is a moderate negative correlation (-0.66) between the rate of evolution (σ) and the hybrid effect parameter (τ), suggesting that an increase in the rate of evolution is associated with a decrease in the magnitude of the hybrid effect parameter. There is a weak negative correlation (-0.19) between the regression slope (b_1) and the hybrid effect parameter (τ), suggesting that there is a weak relationship between these variables, and as the hybrid effect parameter increases, the regression slope tends to decrease, but the relationship is not particularly strong.

We performed a benchmark analysis to evaluate the proposed methodology. The baseline model used for comparison is a simple linear regression model. Another model used for comparison is the tree model, which assumes a Brownian motion model [34]. These models were used for the benchmark analysis of our network model. While the existing methodology may not be directly comparable, the analysis still provides insights into baseline estimation and allows us to compare the performance of the proposed methodology with existing baselines. The result is shown in Table 4. The first row of the table compares the performance of the tree model and linear regression model using the "Annuals" trait. The tree model has a benchmark ratio of 1.006, indicating that its RMSE is 0.6% higher than that of the linear regression model. Similarly, the network model has a benchmark ratio of 1.077, which means that its RMSE is 7.7% higher than that of the linear regression model. The results indicate that the tree model has slightly poorer performance compared to the linear regression model, while the network model performs even worse than the linear regression model. This is expected because the network model is more complex. However, despite the larger RMSE values obtained from the network model, the values are still reasonable when compared to the baseline model.

Table 4. The benchmark analysis involves the use of 12 traits, with RMSE1 computed via a simple linear regression baseline model, RMSE2 computed via the tree model [34], and RMSE3 computed via the proposed network model. The fourth and fifth columns of the table present the benchmark ratio for each model.

	RMSE1	**RMSE2**	**RMSE3**	$\frac{\textbf{RMSE2}}{\textbf{RMSE1}}$	$\frac{\textbf{RMSE3}}{\textbf{RMSE1}}$
Annuals	0.608	0.612	0.655	1.006	1.077
Petioles	0.558	0.559	0.565	1.002	1.013
Peduncles	0.718	0.727	0.730	1.013	1.017
Involucres	0.227	0.234	0.246	1.033	1.087
Phyllaries	0.276	0.276	0.277	1.001	1.005
Paleae	0.164	0.165	0.171	1.009	1.046
Laminae	0.250	0.267	0.265	1.065	1.059
Ray.florets	0.307	0.308	0.357	1.004	1.162
Disc.florets	0.665	0.676	0.768	1.017	1.154
Corollas	0.125	0.128	0.147	1.020	1.175
Cypselae	0.213	0.219	0.332	1.026	1.555
Pappi	0.175	0.183	0.239	1.046	1.367

5. Discussion

The model utilized to examine trait values in phylogenetic networks through hybridization modeling is of fundamental importance and represents an essential tool in the analysis of this type of data. There is room for improvement by using more appropriate representations for the hybrid R based on its parents X and Y to find suitable functions $R = f(\tau, X, Y)$, which would allow us to model events such as horizontal gene transfers or recombination that are biologically different from hybridization and can affect trait values.

We acknowledge that the covariance structure $\boldsymbol{G}_\tau$ is complex, which creates difficulties in demonstrating the positive definiteness of the Hessian matrix of the likelihood function. This makes it challenging to ensure that the likelihood is jointly convex in all parameters. However, our regression model meets certain conditions, including having a well-defined likelihood function and satisfying the assumption of non-singularity. Our empirical analysis confirms that our method achieves the global maximum within its domain. This is supported by the fact that $\boldsymbol{G}_{\hat{\tau}}$ is positive definite for each dataset, as detailed in Appendix A.1.3.

In order to enhance the current model's capability to analyze phylogenetic network data, several future research avenues could be pursued. Firstly, the model could be extended to include more complex evolutionary processes, such as the Ornstein–Uhlenbeck (OU) model [35] or the early burst model [36]. The OU model could be implemented by introducing a force parameter α to the covariance matrix construction, and the optimization process would require a multidimensional search. For instance, if implementing the OU process [35], one would need to take the non-independent increment condition into account to construct the covariance matrix. One can also consider implementing non-Gaussian processes [37] in the network for trait evolution. Secondly, the algorithm could be generalized to handle the hard polytomy by analyzing multifurcating phylogenetic networks for regression analysis [38].

It is also worthwhile to take into account situations in which characteristics may conform to probability distributions beyond the normal distribution and to evaluate the resilience of our proposed methodology when the assumption of normality is not met. In particular, researchers should examine model misspecification problems [39] and study the consequences of non-normal distributions on the efficacy of the model, as has been done in previous studies [40].

Incorporating more parameters into the model would enable a more functional role of interaction with the hybrid parameters, particularly in the context of richer models such as the OU and early burst models. Furthermore, future work could explore the integration of discrete character evolution or the joint analysis of both discrete and continuous characters [41,42], as well as extend the proposed approach to accommodate diverse types of trait distributions. The development of such extensions would contribute to a better understanding of the evolution of biological traits, and may have practical applications in fields such as conservation biology and agriculture [43].

6. Conclusions

A phylogenetic regression model that incorporates a network structure to examine trait evolution in the context of reticulation events is proposed. Maximum likelihood estimation is utilized to estimate parameters, and an algorithm is developed to build the variance–covariance matrix using a phylogenetic network in eNewick format as input. This model is applied to investigate the response of common sunflower, Helianthus annuus, traits to drought conditions.

Parameter estimation is conducted through maximum likelihood, a widely used method in evolutionary biology, which allows for the estimation of model parameters that maximize the probability of the observed data. Additionally, an algorithm is developed to build the variance–covariance matrix, a crucial component of the model, using a phylogenetic network in eNewick format as input.

Overall, the proposed model and associated methods offer a novel approach to studying trait evolution in the context of reticulation events. By applying the model to the common sunflower and investigating its response to drought conditions, new insights can be gained into the evolutionary patterns of this important species.

Funding: This research and APC were funded by the National Science and Technology Council, Taiwan. MOST 111-2118-M-035-004-.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: I would like to express my sincere gratitude to the reviewers for their valuable feedback, which significantly enhanced the quality and rigor of the analysis. Additionally, I am grateful to Elizabeth Houseworth and Yo-Lun Tsai for their inspiration and support in the early stages of this work.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Appendix A.1. Script and Data files

All files in the manuscript can be accessed at http://tonyjhwueng.info/phyreghyb (accessed on 10 March 2023).

Appendix A.1.1. Model

1. BM: http://tonyjhwueng.info/phyreghyb/bmhydRegV3.r (accessed on 10 March 2023).

Appendix A.1.2. Sunflower Precipitation Dataset

The data for each sunflower can be accessed by executing the R script at the following link:

1. Precipitation data script: http://tonyjhwueng.info/phyreghyb/worldclim (accessed on 10 March 2023).

Appendix A.1.3. Figures and Tables

1 Figure 1: http://tonyjhwueng.info/phyreghyb/3taxanetwork.pptx (accessed on 10 March 2023).
2 Figure 2: http://tonyjhwueng.info/phyreghyb/3taxanetworkstep.pptx (accessed on 10 March 2023).
3 Figure 3: http://tonyjhwueng.info/phyreghyb/3taxatree.pptx (accessed on 10 March 2023).
4 Figure 4: http://tonyjhwueng.info/phyreghyb/sfnet.pptx (accessed on 10 March 2023).
5 Figure A1: http://tonyjhwueng.info/phyreghyb/sixtaxanetwork.pptx (accessed on 10 March 2023).
6 Table 2: http://tonyjhwueng.info/phyreghyb/precdatatolatex.html (accessed on 10 March 2023).
7 Table 3: http://tonyjhwueng.info/phyreghyb/RegSunflower.html (accessed on 10 March 2023).
8 Positive definite of $G_{\hat{\tau}}$: https://tonyjhwueng.info/phyreghyb/pdultracheck.html (accessed on 10 March 2023).
9 Table 4: https://tonyjhwueng.info/phyreghyb/AnnPrecAllResponseTrait.html (accessed on 10 March 2023).

Appendix A.2. Demonstration of Algorithm under Brownian Motion Model

Consider the phylogenetic network given in Figure A1. There are 6 extant taxa, 2 hybridization events, and 9 ancestral nodes in the network.

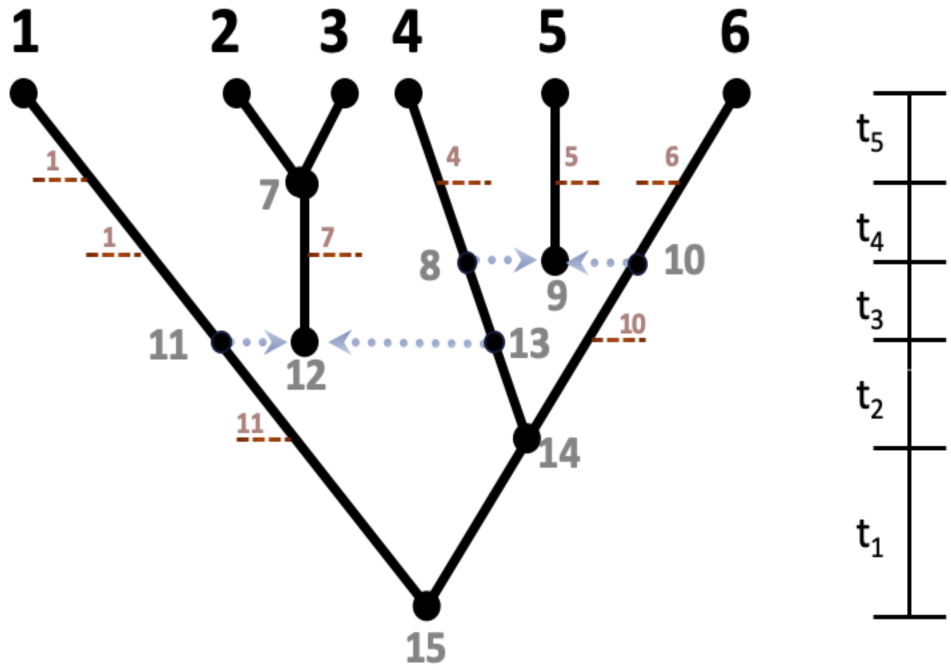

Figure A1. A six-taxa phylogenetic network where 2, 3, and 5 are the hybrid descendants. The eNewick format for the network topology is $((1,((2,3)7)12)11,((12,(4,(5)9)8)13,(9,6)10)14)15$.

The ancestral–descendant data gathered from the `eNewick2phylog` function and modified are shown as follows:

Ancestor	[15]	[14]	[12]	[11]	[13]	[9]	[8]	[10]	[7]
Descendants	[11,14]	[13,10]	[7]	[1],[12]	[12,8]	[5]	[4,9]	[9,6]	[2,3]

From this, we can determine the event times and which times lead to which descendants as follows:

Node	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Length	$t_3 + t_4 + t_5$	t_5	t_5	$t_4 + t_5$	$t_4 + t_5$	$t_4 + t_5$	$t_3 + t_4$	t_3

Node	[9]	[10]	[11]	[12]	[13]	[14]	[15]
Length	0	$t_2 + t_3$	$t_1 + t_2$	0	t_2	t_1	0

We also identify the sequence of temporary similarity matrices built up from the root to the tips in terms of the nodes at each event (speciation or hybridization):

$$[15] \to [11,14] \to [11,13,10] \to [11,12,13,10] \to [1,7,8,10]$$
$$\to [1,7,8,9,10] \to [1,7,4,5,6] \to [1,2,3,4,5,6].$$

This sequence contains information for speciation and hybridization events where the speciation replaces the ancestor node with the corresponding two descendants (e.g., for speciation, [15] is replaced by [11, 14]). For hybridization, the hybrid node is inserted between its parents (e.g., [11, 13, 14] → [11, 12, 13, 14] indicates that [12] is hybrid and is inserted between [11] and [13]).

For the first similarity matrix, we obviously have

$$\boldsymbol{G}_2 = \begin{array}{c c} & \begin{array}{cc} 11 & 14 \end{array} \\ \begin{array}{c} 11 \\ 14 \end{array} & \begin{pmatrix} t_1 & 0 \\ - & t_1 \end{pmatrix} \end{array}. \tag{A1}$$

Going from [11,14] → [11,13,10] involves a straightforward speciation event and the new similarity matrix becomes

$$\boldsymbol{G}_3 = \begin{array}{c c} & \begin{array}{ccc} 11 & 13 & 10 \end{array} \\ \begin{array}{c} 11 \\ 13 \\ 10 \end{array} & \begin{pmatrix} t_1 + t_2 & 0 & 0 \\ - & t_1 + t_2 & t_1 \\ - & - & t_1 + t_2 \end{pmatrix} \end{array}. \tag{A2}$$

Going from [11,13,10] → [11,12,13,10] involves a hybridization. The variance for the hybrid [12] can be calculated from $\boldsymbol{G}_3$ with the following formula: $Var([12]) = Var(\tau([11] + [13])) = \tau^2\{Var([11]) + Var([13]) + 2Cov([11],[13])\}$.

Moreover, the covariance between the hybrid species [12] and other species can be obtained by following the formula: $Cov([12], Z) = Cov(\tau([11] + [13]), Z) = \tau\{Cov([11], Z) + Cov([13], Z)\}, \quad Z = 11, 13, 10$. All other elements in $\boldsymbol{G}_{4,\tau}$ can be tracked from $\boldsymbol{G}_3$ because they are identical. Therefore, the covariance for species [11], [12], [13], and [10] at $t = t_1 + t_2$ is

$$\boldsymbol{G}_{4,\tau} = \begin{array}{c c} & \begin{array}{cccc} 11 & 12 & 13 & 10 \end{array} \\ \begin{array}{c} 11 \\ 12 \\ 13 \\ 10 \end{array} & \begin{pmatrix} t_1 + t_2 & \tau(t_1 + t_2) & 0 & 0 \\ - & 2\tau^2(t_1 + t_2) & \tau(t_1 + t_2) & \tau t_1 \\ - & - & t_1 + t_2 & t_1 \\ - & - & - & t_1 + t_2 \end{pmatrix} \end{array}.$$

We elongate from [11,12,13,10]→[1,7,8,10] to obtain

$$\boldsymbol{G}'_{4,\tau} = \begin{array}{c c} & \begin{array}{cccc} 1 & 7 & 8 & 10 \end{array} \\ \begin{array}{c} 1 \\ 7 \\ 8 \\ 10 \end{array} & \begin{pmatrix} t_1 + t_2 + t_3 & \tau(t_1 + t_2) & 0 & 0 \\ - & t_3 + 2\tau^2(t_1 + t_2) & \tau(t_1 + t_2) & \tau t_1 \\ - & - & t_1 + t_2 + t_3 & t_1 \\ - & - & - & t_1 + t_2 + t_3 \end{pmatrix} \end{array}. \tag{A3}$$

The next event from [1,7,8,10] → [1,7,8,9,10] is another hybridization. The 5×5 matrix $\boldsymbol{G}_{5,\tau}$ for species $[1,7,8,9,10]$ is constructed by inserting the hybrid $[9]$ between its parents $[8]$ and $[10]$.

$$\boldsymbol{G}_{5,\tau} = \begin{array}{c} \\ 1 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 7 & 8 & 9 & 10 \end{array} \\ \begin{pmatrix} t_1+t_2+t_3 & \tau(t_1+t_2) & 0 & 0 & 0 \\ - & t_3+2\tau^2(t_1+t_2) & \tau(t_1+t_2) & \tau^2(2t_1+t_2) & \tau t_1 \\ - & - & t_1+t_2+t_3 & \tau(2t_1+t_2+t_3) & t_1 \\ - & - & - & \tau^2(3t_1+2t_2+2t_3) & \tau(2t_1+t_2+t_3) \\ - & - & - & - & t_1+t_2+t_3 \end{pmatrix} \end{array}. \tag{A4}$$

We elongate from [1,7,8,9,10] to [1,7,4,5,6] to obtain

$$\boldsymbol{G}'_{5,\tau} = \begin{array}{c|ccccc} & 1 & 7 & 4 & 5 & 6 \\ \hline 1 & \sum_{k=1}^{4} t_k & \tau\sum_{k=1}^{2} t_k & 0 & 0 & 0 \\ 7 & - & \sum_{k=3}^{4} t_k + 2\tau^2\sum_{k=1}^{2} t_k & \tau\sum_{k=1}^{2} t_k & \tau^2(2t_1+t_2) & \tau t_1 \\ 4 & - & - & \sum_{k=1}^{4} t_k & \tau(2t_1+\sum_{k=2}^{3} t_k) & t_1 \\ 5 & - & - & - & t_4+\tau^2(3t_1+2\sum_{k=2}^{3} t_k) & \tau(2t_1+\sum_{k=2}^{3} t_k) \\ 6 & - & - & - & - & \sum_{k=1}^{4} t_k \end{array}, \tag{A5}$$

where $\sum_{k=1}^{4} t_k = t_1+t_2+t_3+t_4$.

The final step from [1,7,4,5,6]→ [1,2,3,4,5,6] involves a speciation event. The final similarity matrix $\boldsymbol{G}_{6,\tau}$ is given as

$$\boldsymbol{G}_{6,\tau} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & \sum_{k=1}^{5} t_k & \tau\sum_{k=1}^{2} t_k & \tau\sum_{k=1}^{2} t_k & 0 & 0 & 0 \\ 2 & - & 2\tau^2\sum_{k=1}^{2} t_k + \sum_{k=1}^{3} t_k & 2\tau^2\sum_{k=1}^{2} t_k + \sum_{k=3}^{4} t_k & \tau\sum_{k=1}^{2} t_k & \tau^2(2t_1+t_2) & \tau t_1 \\ 3 & - & - & 2\tau^2\sum_{k=1}^{2} t_k + \sum_{k=3}^{5} t_k & \tau\sum_{k=1}^{2} t_k & \tau^2(2t_1+t_2) & \tau t_1 \\ 4 & - & - & - & \sum_{k=1}^{5} t_k & \tau(2t_1+\sum_{k=2}^{3} t_k) & t_1 \\ 5 & - & - & - & - & \tau^2(3t_1+2\sum_{k=2}^{3} t_k)+\sum_{k=4}^{5} t_k & \tau(2t_1+\sum_{k=2}^{3} t_k) \\ 6 & - & - & - & - & - & \sum_{k=1}^{5} t_k \end{array}, \tag{A6}$$

If we assign branch lengths by setting $t_1 = 0.1, t_2 = 0.25, t_3 = 0.15, t_4 = 0.2, t_5 = 0.3$, the eNewick format with branch lengths input into the **R** program will be as follows. `Input`: network $= c(``((1:0.65,((2:0.3,3:0.3)7:0.35)12:0)11:0.35,((12:0,(4:0.5,(5:0.5)9:0)8:0.15)13:0.25,(9:0,6:0.5)10:0.4)14:0.1)15:0")$.

`Output`: The similarity matrix for the species $[1,2,3,4,5,6]$ on the tips of the tree is

$$\boldsymbol{G}_{6,\tau} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 1 & 0.175 & 0.175 & 0 & 0 & 0 \\ 2 & - & 0.825 & 0.525 & 0.175 & 0.1125 & 0.05 \\ 3 & - & - & 0.825 & 0.175 & 0.1125 & 0.05 \\ 4 & - & - & - & 1 & 0.3 & 0.1 \\ 5 & - & - & - & - & 0.8 & 0.3 \\ 6 & - & - & - & - & - & 1 \end{array}. \tag{A7}$$

It can be seen that the covariance matrix is a 6 by 6 matrix where the upper diagonal is shown due to its symmetry.

References

1. Rieseberg, L.H.; Carney, S.E. Plant hybridization. *New Phytol.* **1998**, *140*, 599–624. [CrossRef] [PubMed]
2. Mitchell, N.; Owens, G.L.; Hovick, S.M.; Rieseberg, L.H.; Whitney, K.D. Hybridization speeds adaptive evolution in an eight-year field experiment. *Sci. Rep.* **2019**, *9*, 6746 . [CrossRef] [PubMed]

3. Bock, D.G.; Kantar, M.B.; Rieseberg, L.H. *Population Genomics of Speciation and Adaptation in Sunflowers*; Springer: Berlin/Heidelberg, Germany, 2020.
4. Harmon, L.J.; Weir, J.T.; Schulte, L.A. *Phylogenies and Comparative Methods in Ecology and Evolution*; University of California Press: Berkeley, CA, USA, 2005.
5. Harvey, P.H.; Pagel, M.D. Comparative methods for explaining adaptations. *Nature* **1991**, *351*, 619–624. [CrossRef] [PubMed]
6. Clutton-Brock, T.H. *Phylogenetic Perspectives on the Evolution of Mammalian Social Behavior*; University of Chicago Press: Chicago, IL, USA, 2010.
7. Bastide, P.; Solis-Lemus, C.; Kriebel, R.; Sparks, K.W.; Ané, C. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Syst. Biol.* **2018**, *67*, 800–820. [CrossRef]
8. Jhwueng, D.C.; O'Meara, B. Trait evolution on phylogenetic networks. *bioRxiv* **2015**, 023986. [CrossRef]
9. Teo, B.; Rose, J.P.; Bastide, P.; Ané, C. Accounting for within-species variation in continuous trait evolution on a phylogenetic network. *bioRxiv* **2022**, 490814. [CrossRef]
10. Jacquemyn, H.; Merckx, V.; Brys, R.; Tyteca, D.; Cammue, B.P.; Honnay, O.; Lievens, B. Analysis of network architecture reveals phylogenetic constraints on mycorrhizal specificity in the genus Orchis (Orchidaceae). *New Phytol.* **2011**, *192*, 518–528. [CrossRef]
11. Solís-Lemus, C.; Bastide, P.; Ané, C. PhyloNetworks: A package for phylogenetic networks. *Mol. Biol. Evol.* **2017**, *34*, 3292–3298. [CrossRef]
12. Solís-Lemus, C.; Ané, C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* **2016**, *12*, e1005896. [CrossRef]
13. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **1985**, *125*, 1–15. [CrossRef]
14. Revell, L.J. Phylogenetic signal and linear regression on species data. *Methods Ecol. Evol.* **2010**, *1*, 319–329. [CrossRef]
15. Ané, C. Analysis of comparative data with hierarchical autocorrelation. *Ann. Appl. Stat.* **2008**, *2*, 1078–1102. [CrossRef]
16. Jhwueng, D.C. Some Problems in Phylogenetic Comparative Methods. Ph.D. Thesis, Indiana University, Bloomington, IN, USA, 2010.
17. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
18. Gilbert, P.; Varadhan, R. numDeriv: Accurate Numerical Derivatives; R Package Version 2016.8-1.1, CRAN Repository. 2019. Available online: https://cran.r-project.org/web/packages/numDeriv/index.html (accessed on 21 February 2023).
19. Wald, A. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* **1949**, *20*, 595–601. [CrossRef]
20. Varga, R.S.; Nabben, R. On symmetric ultrametric matrices. In *Numerical Linear Algebra;* De Gruyter: Berlin, Germany, 1993; pp. 193–199.
21. Nabben, R.; Varga, R.S. A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant Stieltjes matrix. *SIAM J. Matrix Anal. Appl.* **1994**, *15*, 107–113. [CrossRef]
22. Cardona, G.; Rosselló, F.; Valiente, G. Extended Newick: It is time for a standard representation of phylogenetic networks. *BMC Bioinform.* **2008**, *9*, 532. [CrossRef]
23. Dray, S.; Dufour, A.B. The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Softw.* **2007**, *22*, 1–20. [CrossRef]
24. Tsai, Y.L. Regression Analysis of Hybrid Species's Trait Data. Master's Thesis, Feng-Chia University, Taichung, Taiwan, 2016.
25. Kane, N.C.; Rieseberg, L.H. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, Helianthus annuus. *Genetics* **2007**, 175, 1823–1834. [CrossRef]
26. Koziol, L.; Rieseberg, L.H.; Kane, N.; Bever, J.D. Reduced drought tolerance during domestication and the evolution of weediness results from tolerance—Growth trade-offs. *Evol. Int. J. Org. Evol.* **2012**, *66*, 3803–3814. [CrossRef]
27. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
28. Cerasoli, F.; D'Alessandro, P.; Biondi, M. Worldclim 2.1 versus Worldclim 1.4: Climatic niche and grid resolution affect between-version mismatches in Habitat Suitability Models predictions across Europe. *Ecol. Evol.* **2022**, *12*, e8430. [CrossRef]
29. GBIF. Global Biodiversity Information Facility Database. *Rumex acetosella* L. 2010. Available online: https://www.gbif.org/ (accessed on 31 January 2023).
30. Hijmans, R.J.; Van Etten, J.; Mattiuzzi, M.; Sumner, M.; Greenberg, J.; Lamigueiro, O.; Bevan, A.; Racine, E.; Shortridge, A. Raster Package in R Version. 2023. Available online: https://cran.r-project.org/web/packages/raster/raster.pdf (accessed on 10 March 2023).
31. van Etten, R.J.H.J. Raster: Geographic Analysis and Modeling with Raster Data; R Package Version 2.0-12, CRAN Repository, 2012. Available online: https://cran.r-project.org/web/packages/raster/index.html (accessed on 21 February 2023).
32. Brach, A.R.; Song, H. eFloras: New directions for online floras exemplified by the Flora of China Project. *Taxon* **2006**, *55*, 188–192. [CrossRef]
33. Gross, B.; Rieseberg, L. The ecological genetics of homoploid hybrid speciation. *J. Hered.* **2004**, *96*, 241–252. [CrossRef] [PubMed]
34. Ho, L.S.T.; Ane, C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **2014**, *63*, 397–408. [PubMed]
35. Hansen, T.F. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **1997**, *51*, 1341–1351. [CrossRef]
36. Harmon, L.J.; Losos, J.B.; Jonathan Davies, T.; Gillespie, R.G.; Gittleman, J.L.; Bryan Jennings, W.; Kozak, K.H.; McPeek, M.A.; Moreno-Roark, F.; Near, T.J.; et al. Early bursts of body size and shape evolution are rare in comparative data. *Evol. Int. J. Org. Evol.* **2010**, *64*, 2385–2396. [CrossRef] [PubMed]

37. Blomberg, S.P.; Rathnayake, S.I.; Moreau, C.M. Beyond Brownian motion and the Ornstein-Uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. *Am. Nat.* **2020**, *195*, 145–165. [CrossRef]
38. Jhwueng, D.C.; Liu, F.C. Effect of Polytomy on the Parameter Estimation and Goodness of Fit of Phylogenetic Linear Regression Models for Trait Evolution. *Diversity* **2022**, *14*, 942. [CrossRef]
39. McCulloch, C.E.; Neuhaus, J.M. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics* **2011**, *67*, 270–279. [CrossRef]
40. Sheng, Y.; Yang, C.; Curhan, S.; Curhan, G.; Wang, M. Analytical methods for correlated data arising from multicenter hearing studies. *Stat. Med.* **2022**, *41*, 5335–5348. [CrossRef]
41. Caetano, D.S.; O'Meara, B.C.; Beaulieu, J.M. Hidden state models improve state-dependent diversification approaches, including biogeographical models. *Evolution* **2018**, *72*, 2308–2324. [CrossRef]
42. Grundler, M.C.; Rabosky, D.L. Complex ecological phenotypes on phylogenetic trees: A hidden Markov model for comparative analysis of multivariate count data. *Syst. Biol.* **2019**, *69*, 1200–1211. [CrossRef] [PubMed]
43. Boyko, J.D.; O'Meara, B.C.; Beaulieu, J.M. A Novel Method for Jointly Modeling the Evolution of Discrete and Continuous Traits. *Evolution* **2023**, *77*, 836–851. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.