ESTIMATION OF AN INDEX OF PHYLOGENETIC CORRELATION USING BOOTSTRAP SIMULATION TECHNIQUE

Bahman Shafii and William J. Price
Statistical Programs, University of Idaho

A common objective of bioinformatic analyses is to assess the similarity of species or genotypic variations. Such measures provide a means to evaluate evolutionary models and history as well as having potential application to ecological systems including identification of host preference selection. Phylogenetic correlation, λ, is one index of similarity typically used measuring the deviation of an observed phylogeny relative to a dependent Brownian evolutionary model. Values for λ are estimated through a generalized linear model assuming a variance-covariance structure that has off diagonal elements scaled by λ. A value of λ equal to 1.0 is indicative of the Brownian model, while λ = 0.0 indicates an independent random process. Statistical inference on λ has traditionally been assessed using a likelihood ratio test comparing the estimated value to the theoretical null values, λ=1.0 and λ=0.0. These tests, however, rely on the assumption of a Normal likelihood within the phylogeny. In addition, statistical comparison of estimated λ values between competing phylogenies has not been addressed. The purpose of this paper is to propose an alternative procedure which relies on the re-sampling methodology of the bootstrap. The underlying bootstrap distribution of λ is estimated which in turn provides a means of computing confidence limits and enables hypothesis testing without distributional assumptions. The method will be demonstrated using phylogenetic and metabolomic data related to the host specificity of an insect, Ceutorhynchus cardariae Korotyaev, on a wide range of Brassicaceae species.

Introduction

Phylogenies are used to describe the relationships among species or related organisms. For example, in the simple phylogenetic tree shown in Figure 1, "species" A is more related to "species" B, while both these "species" are less related to "species" C. Relatedness, in this case, is reflected both in the branch lengths, as well as the number of intermediate nodes between species. Often it is of interest to evaluate the association of an ancillary biological trait, not used for the development of the phylogeny, with an existing phylogeny, i.e., do the biological traits correspond to the phylogenetic relationship? Two measures that have been used to quantify this association are Pagel's λ (Pagel 1999) and Blomberg's K (Blomberg et al. 2003). These statistics measure the observed phylogenetic signal of a trait relative to that which is expected under a random Brownian evolutionary model. The Brownian model implies a stochastic "null" condition where traits develop along a path analogous to random Brownian motion.

Pagel's λ may take on values between 0.0 and 1.0, where λ = 0.0 indicates an independent relationship and λ =1.0 a Brownian association. Tests of these conditions are typically carried out through likelihood ratio tests assuming normality of the trait response.

Blomberg's K is a positive value, where $K \leq 1.0$ indicates that the species trait has less association than would be expected from the phylogeny, and $K > 1.0$ is evidence that the association with the phylogeny is strong. Tests of the condition K = 1.0 can be

provided assuming normality, and analysis of variance (ANOVA) procedures have been suggested as a means of comparing K values (Blomberg 2003), although some potential distributional problems with that technique were also noted.

In this study, bootstrap procedures are proposed for addressing inferences on both λ and K. Demonstration will be given for data related to two potential phylogenies of Brassicaceae plant species, and a vector of feeding responses by a potential biological control agent, the weevil Ceutorhynchus cardariae Korotyaev (Coleoptera: Curculionidae). Here, the phylogenies will be assumed as fixed while the random selection process of the bootstrap technique will be made on the feeding trait vector.

Methods

Pagel's λ

A Brownian model may be defined as (Pagel 1999 ; Freckleton, et al. 2002) :

$$\boldsymbol{y_i} = \boldsymbol{\alpha} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{T}} \boldsymbol{\epsilon_{ij}}\ \boldsymbol{t_{ij}} \tag{1}$$

Where, y_i is a trait of interest for species i, α is the ancestral state of the trait, ε_{ij} is a normal random variable of constant variance, σ^2, and the summation is across T branches of length t_{ij}. If Y is a vector of the trait values for n species, then Y has a multivariate normal distribution given by:

$$\boldsymbol{p(Y)} = \frac{\mathbf{1}}{(\mathbf{2\pi\sigma^2})^{\frac{n}{2}}}|\boldsymbol{V}|^{-\frac{\mathbf{1}}{\mathbf{2}}} \times \boldsymbol{exp}\left[-\frac{\mathbf{1}}{(\mathbf{2\sigma^2})}(\boldsymbol{Y}-\boldsymbol{X\alpha})`\boldsymbol{V^{-1}}(\boldsymbol{Y}-\boldsymbol{X\alpha})\right] \tag{2}$$

Here, V is the n × n variance-covariance matrix among species and X is a design matrix. λ is defined as a multiplier of the off-diagonal elements in the variance-covariance matrix, V, such that λ = 1 returns the Brownian model in (2) and λ = 0 returns a model independent of the phylogeny. Intermediate values indicate less than complete Brownian dependency (Pagel 1999). Values for λ may be estimated through maximum likelihood utilizing numerical optimization. While ML estimation provides a point estimate for λ, inferences relative to the null Brownian model are provided by likelihood ratio tests. Further information regarding the statistical properties of λ, e.g. variability, reliability, distribution, etc., have not been addressed.

Blomberg's K

An alternative measure, Blomberg's K (Blomberg, 2003), is based on relative measures of variability under specified hypotheses. That is, relative variability, R, is defined as:

$$R = MSE_0 / MSE = (Y - \alpha)^2/(U - \alpha)^2 \qquad (3)$$

Where, MSE_0 is the raw variability of the trait data, MSE is the variability corrected for by the phylogeny covariance, Y is the observed trait response, U is the estimated trait response adjusted for the phylogeny and α is the ancestral trait mean as given above. While R provides a measure of phylogenetic signal, it is influenced by the number of nodes and tips in the phylogeny. This makes it difficult to compare these values across phylogenies. Hence, for comparative purposes, R can be standardized with its expectation under the Brownian model given by:

$$E[R] = (1/(n-1))*(tr(V) - n/\Sigma\Sigma V^{-1}) \qquad (4)$$

Here, n is the number of tips and V is the phylogenetic variance-covariance matrix defined above. K is then defined as:

$$K = R / E[R] = observed[MSE_0/MSE] / expected[MSE_0/MSE] \qquad (5)$$

Bootstrap inference

One means of inference on λ and K is the bootstrap simulation (Efron & Tibshirani 1993). Specifically, in this case, random perturbations are introduced into the trait vector, Y, at each iteration of the bootstrap process. The values of λ and K from (2) and (5) are then re-estimated and the process of perturbation is repeated a large number of times, B, to provide empirical bootstrap distributions of potential λ and K values. While the disturbance values of the trait vector Y can be made either through random sampling of the model error terms (nonparametric bootstrap), or through re-sampling values of the trait, y_i , from a known distribution, only the later is demonstrated here. Earlier investigations involving these data have indicated little difference between the two methods (Price, et al. 2009).

Following bootstrap estimation, interval estimates may be assessed by examining the percentiles of the bootstrap distributions. Subsequent comparison of λ or K values may be carried out either across two or more traits, or within a trait, across different phylogenetic trees. Comparisons are defined on the distribution of differences in the respective bootstrap values of λ or K given by either:

$$\delta_\lambda = \lambda_1 - \lambda_2 ,$$

or (6)

$$\delta_K = K_1 - K_2 ,$$

Where, λ_1 or K_1 and λ_2 or K_2 represent the respective phylogenetic signals from two competing phylogenies or biological traits.

All statistical computations and graphics were carried out using the R statistical system (R Development Core Team 2004). Computations for λ and K were provided by the R packages Geiger (Harmon, et al. 2009) and Picante (Kembel, et al. 2010), respectively.

Demonstration

Feeding Data and Phylogenies

The data set used for this study relates the phylogeny of Brassicaceae species to biological traits of a potential biological control agent as presented by Rapo (2009). The taxonomic group of Brassicaceae covers a large number of economically important crop and weedy species. In this data, the weevil *Ceutorhynchus cardariae* Korotyaev is under assessment for the control of the weedy Brassicaceae species *Lepidium draba* L., which occurs worldwide in many environments. Bio-assays were carried out to assess the potential of *C. cardariae* attack on eleven Brassicaceae species. Such assays provide information regarding host preference and the possibility of attack on non-target Brassicaceae species. In the current study, several measures relevant to attack were recorded, however, only data related to feeding intensity are used for demonstration. Feeding intensity is measured as the number of feeding holes observed in a caged, no choice setting after 48 hours. Ten replications were available for each Brassicaceae species and the average number of feeding holes was the response. The Brassicaceae species used and the corresponding average feeding intensities are given in Table 1. The feeding intensities were further classified into three levels, High (red), Moderate (green) and Low (blue). The species *Lepidium campestre*, *Lepidium draba*, and *Draba nemorosa*, for example, indicated the highest levels of feeding, while *Lepidium latifolium* and *Brassica nigra* showed moderate feeding intensities.

A phylogeny, based on genetic analysis, for the 11 Brassicaceae species is shown in Figure 2a. It might be expected that species closely related to this genetic phylogeny would be equally susceptible to attack (Wapshere 1974). In this case, however, mapping the feeding intensity classes onto this phylogeny indicates that distantly related species, such as *Lepidium draba* and *Draba nemorosa*, are subject to similar levels of attack (Figure 2b). This discontinuity or disjoint host range suggests that another means of assessing species relatedness is required to predict host preference of *C. cardariae*. Such measures could include physical morphological traits such as trichome densities and leaf dry matter content or the chemical profile produced by the plants. Brassicaceae species are well known for their production of glucosinolate compounds. For the species at hand, 34 glucosinolate components were quantified via gas chromography, of which 27 could be reliably identified. Based on these 27 glucosinolate compounds, a separate phylogeny was developed using a neighbor-joining algorithm (Figure 3a). While this phylogeny differs somewhat from the genetic version, the feeding intensity data appears to visually correspond to the phylogeny groups (Figure 3b).

Phylogenetic Signal and Bootstrap Estimation

Although the subjective feeding intensity groups are useful for quick visual assessment of the host range, a more objective assessment would be desirable. To quantify the relationship, both Pagel's λ and Blomberg's K were estimated using these

data (Table 2). The values for K showed K > 1.0 (strong correlation) for the glucosinolate phylogeny and a lower value (weaker correlation) for the genetic phylogeny. This concurs with the visual assessment shown above. The corresponding values for λ, however, indicate an opposite pattern than might be expected, showing a near perfect correlation with the genetic structure as well as a lower value for that of the glucosinolate phylogeny. The variability associated with these measures was not directly available, and hence, a further investigation utilizing bootstrap estimation was deemed warranted.

Feeding hole measurements were simulated as Poisson variates on each bootstrap iteration. The Poisson parameter, e.g. the distributional mean of each species, was set equal to the corresponding average number of feeding holes observed in that species. Values for λ and K were then computed for each of B = 1000 bootstrap iterations using both the genetic and glucosinolate phylogenies. Following all bootstrap simulations, the corresponding empirical bootstrap distributions of each metric were developed along with the associated 95^{th} percentile intervals.

Pagel's λ

Figures 4a and 4b display the bootstrap distributions for Pagel's λ based on the genetic and glucosinolate phylogenies, respectively. The glucosinolate distribution follows a reasonable distribution with 95% intervals ranging from 0.38 to 0.76. The genetic distribution, however, is degenerate, centering on a value close to 1.0 with no variability. Further inspection revealed that several bootstrap iterations in both phylogenies had defaulted to either the values $\lambda = 0.0$ or $\lambda = 1.0$. Overall, the estimation of λ was found to be unstable, possibly due to its definition as a multiplicative adjuster in the variance-covariance structure. Small changes relative to the trait data rendered the estimation of λ untenable, thereby reducing its value and reliability as a measure of phylogenetic correlation. For these reasons, λ was not considered for further investigation.

Blomberg's K

The empirical bootstrap distributions of K for the genetic and glucosinolate phylogenies are given in Figures 5a and 5b, respectively. In both cases, the bootstrap process resulted in usable distributions. The genetic phylogeny had a 95^{th} percentile interval of 0.65 to 0.91. This range does not cover K = 1.0, suggesting that the relationship modeled by the genetic phylogeny does not adequately explain the variability present in the feeding data. Alternatively, the percentile range for the glucosinolate data was 0.98 to 1.23, indicating the presence of some correlation between this phylogeny and the feeding data. A comparative plot of the two distributions is shown in Figure 6a. A distinct separation between the two scenarios is evident with little overlap indicated. The distribution of the difference in the two K estimates ($\delta_K = K_{Gluc} - K_{genetic}$) is given in Figure 6b, where the percentile interval is $0.14 < \delta_K < 0.49$, suggesting a significant difference between the measures, where $K_{gluc} > K_{genetic}$. The glucosinolate data, therefore, appear to better predict the host preference of *C. cardariae* as measured by feeding intensity.

Concluding Remarks

Genetic-based phylogenies have been shown to be useful in predicting qualities such as host preference. In some situations, however, they may not work well if other factors are more prevalent in interactions with other organisms. In those cases, other measures of relatedness developed from physical or chemical characteristics may provide more reliable information.

In this study, two measures of phylogenetic correlation, Pagel's λ and Blomberg's K, were proposed for examining the relationship between a phylogenetic structure and a biological trait. Statistical inferences on these metrics were carried out using bootstrap simulation methods. Empirical bootstrap distributions for the feeding data of *C. cardariae* were developed and compared under genetic and glucosinolate phylogeny scenarios. The metric λ was unstable during bootstrap simulations due to its multiplicative nature. The metric K, however, was able to numerically demonstrate the correspondence between feeding data and the glucosinolate phylogeny. Comparison of K for the two phylogenies found glucosinolates to have a better correspondence to the feeding intensity data than the phylogeny developed from genetic information.

These methods will prove useful for future attempts to define the plant-insect relationship utilizing additional chemical profile and plant morphology data. Successful completion of this objective will help predict non-target susceptibility to *C. cardariae*.

References

Blomberg S. P., T. Garland Jr., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution, 57(4), 2003, pp. 717-745.

Efron, B. and R. J. Tibshirani. 1993. An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability). Chapman Hall, 456 pp.

Harmon, L. 2008. geiger: Analysis of evolutionary diversification. http://cran.r-project.org/web/packages/geiger/index.html .

Freckleton R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. The American Naturalist. 160:6, 712-726.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature, 401: 877-884.

Price, W. J., Hinz, H.L., Shafii, B., Schwarzlaender, M., Rapo, C.B., Eigenbrode, S. D. 2009. Bootstrap estimation and inference on an index of phylogenetic correlation. Proceedings of the 25th Annual Meeting of the International Society of Chemical Ecology, Neuchatel, Switzerland, August 23-27, 2009.

R Development Core Team (2004). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Rapo C.B., S. D. Eigenbrode, H. L. Hinz, U. Schaffner, W. J. Price, and M. Schwarzlaender. 2009. Metabolic Profiling Analysis: a New Tool in the Prediction of Host-Specificity in Classical Biological Control? Proceedings of the 25th Annual Meeting of the International Society of Chemical Ecology, Neuchatel, Switzerland, August 23-27, 2009.

S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463-1464.

Wapshere, A. J. 1974. A strategy for evaluating the safety of organisms for biological weed control. Annals of Applied Biology . 77:201-211

Plant Species	No. Holes
Barbarea orthoceras	13.4
Brassica nigra	25.25
Camelina microcarpa	0
Draba nemorosa	46.6
Hesperis matronalis	3.4
Lepidium campestre	65.7
Lepidium draba	87.3
Lepidium latifolium	17.53
Lepidium squamatum	9.5
Stanleya pinnata	1.25
Stanleya viridiflora	0

Table 1. Brassicaceae species and the associated average number of feeding holes by *Ceutorhynchus cardariae* Korotyaev recorded over 48 hours.

Phylogeny	λ	K
Genetic	0.99	0.79
Glucosinolate	0.57	1.13

Table 2. Estimated values for Pagel's λ and Blomberg's K using the genetic and glucosinolate based phylogenies.

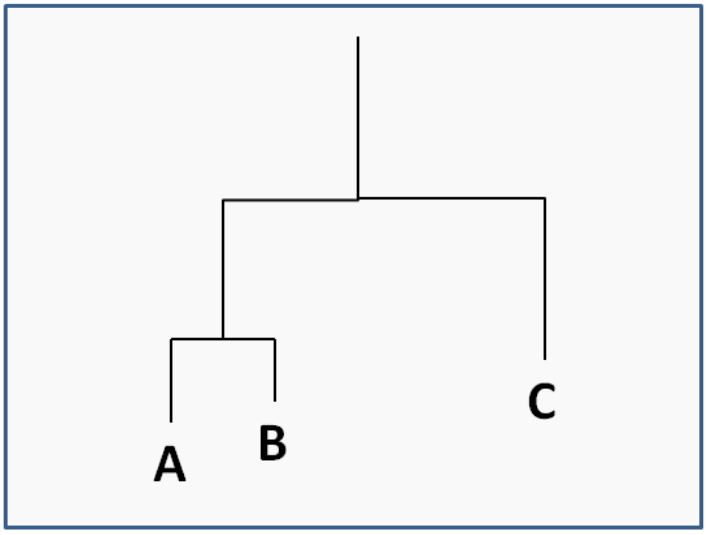


Figure 1. A simple phylogeny representing the relationship between three “species”, A, B, and C. “Species” A and B more closely related to one another than to “species” C.

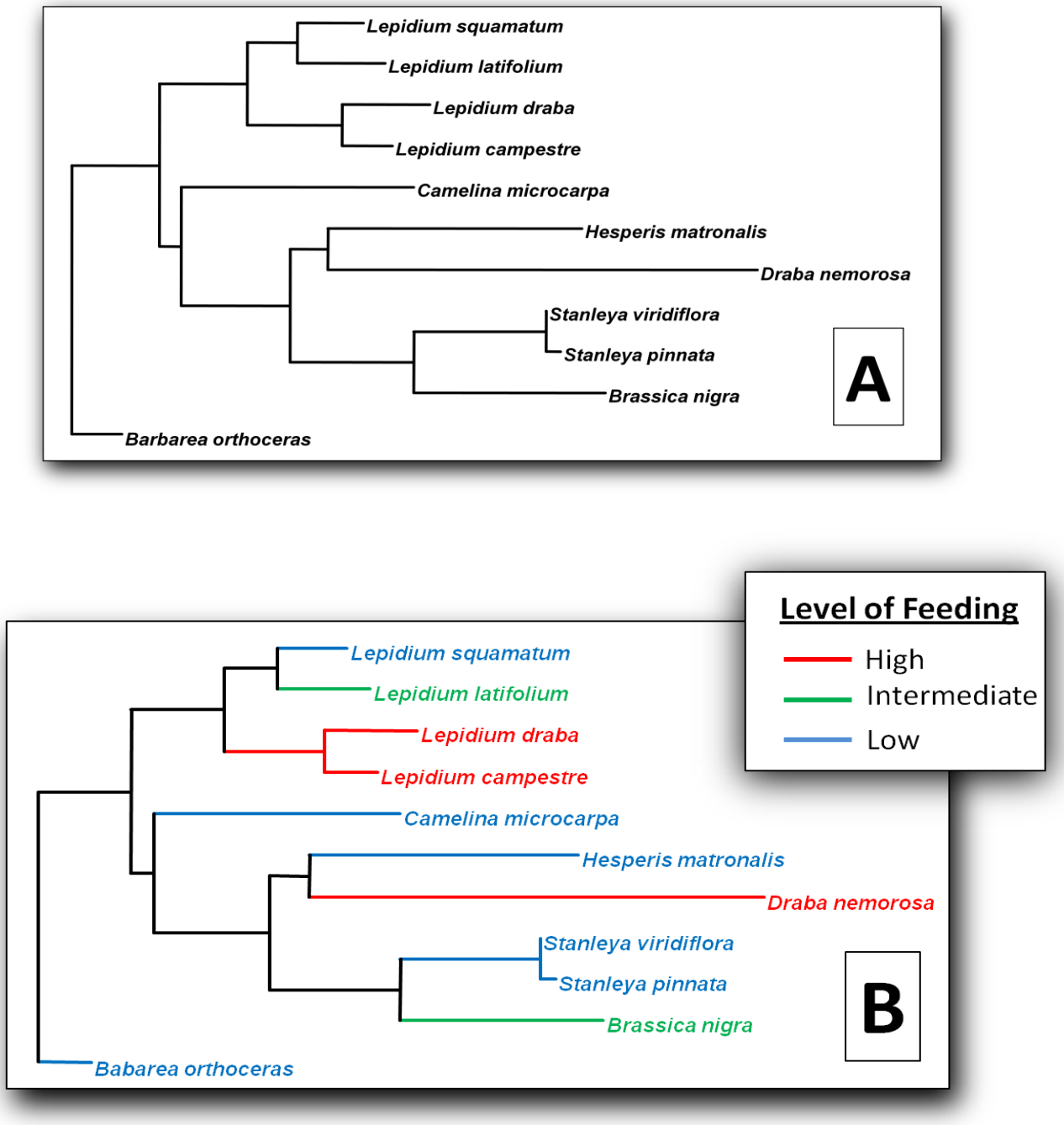


Figure 2. Phylogeny developed from genetic data (A) and the same phylogeny overlaid with *Ceutorhynchus cardariae* Korotyaev feeding intensity classes (B).

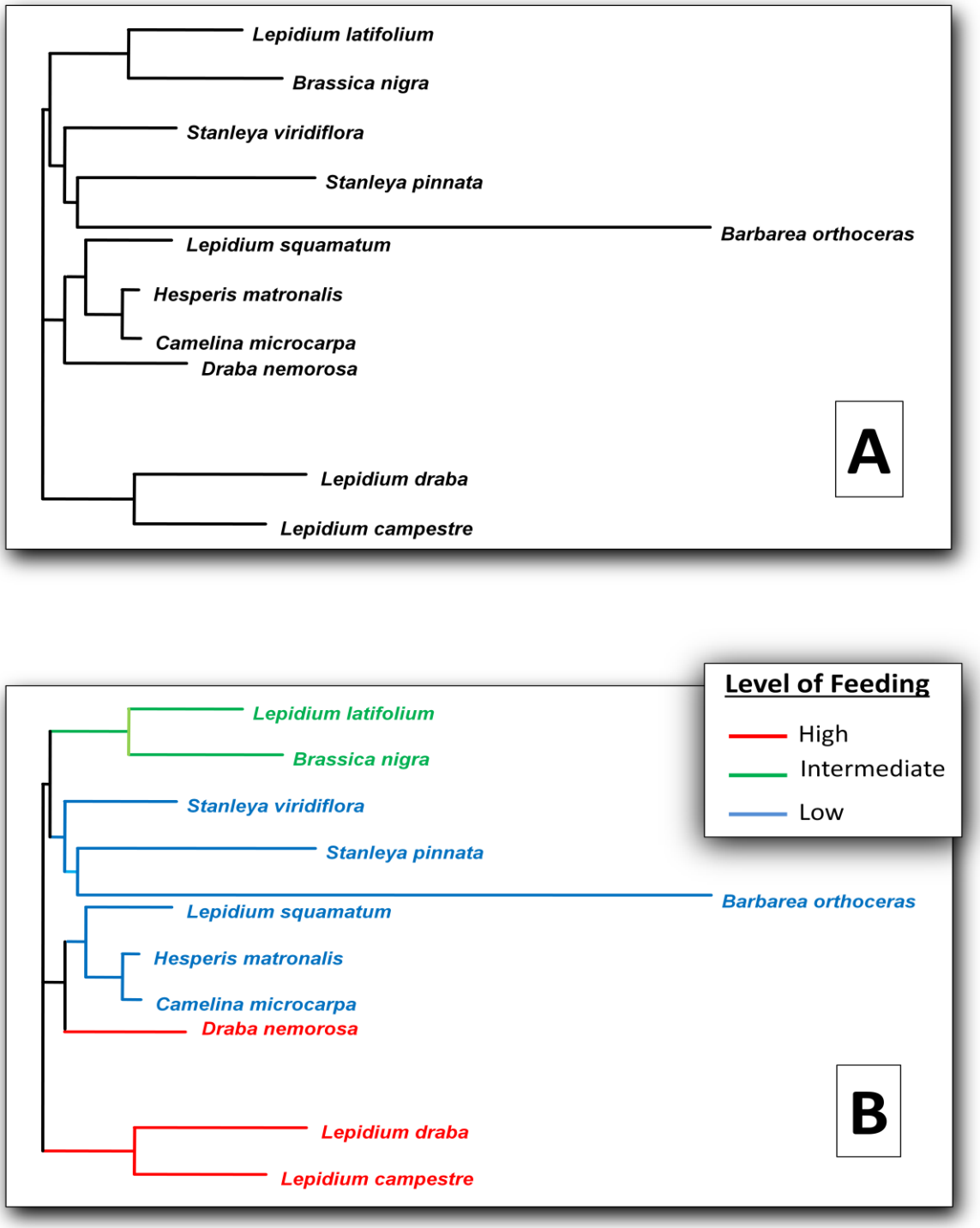
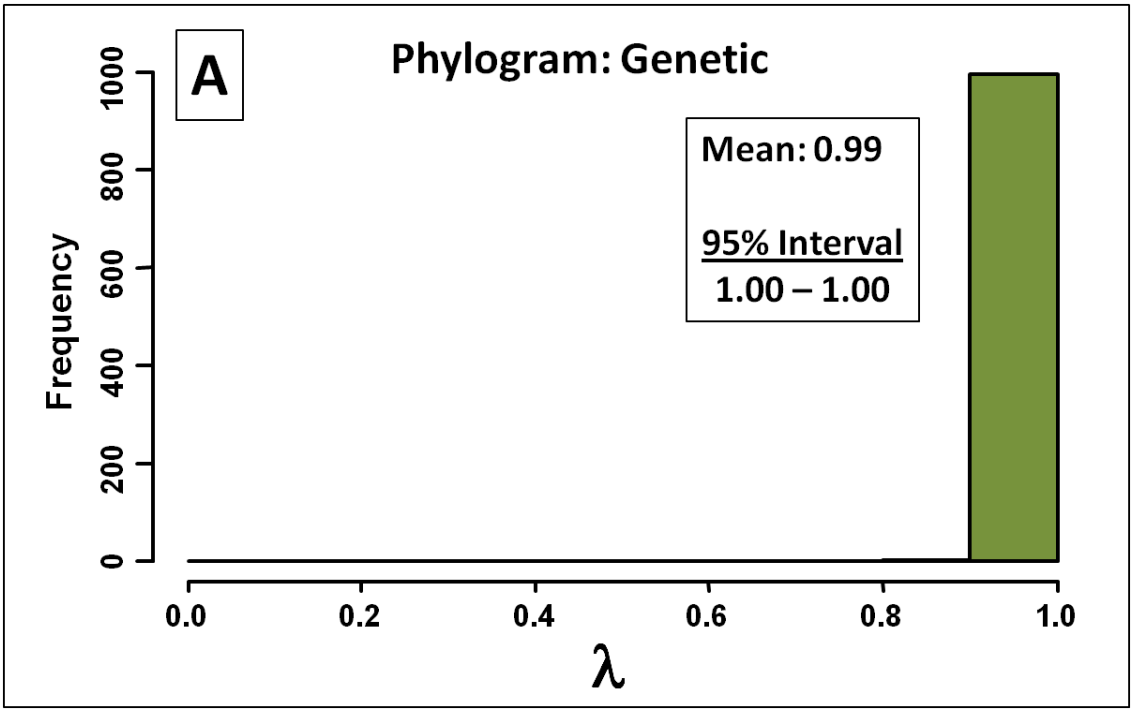


Figure 3. Phylogeny developed from glucosinolate data (A) and the same phylogeny overlaid with *Ceutorhynchus cardariae* Korotyaev feeding intensity classes (B).



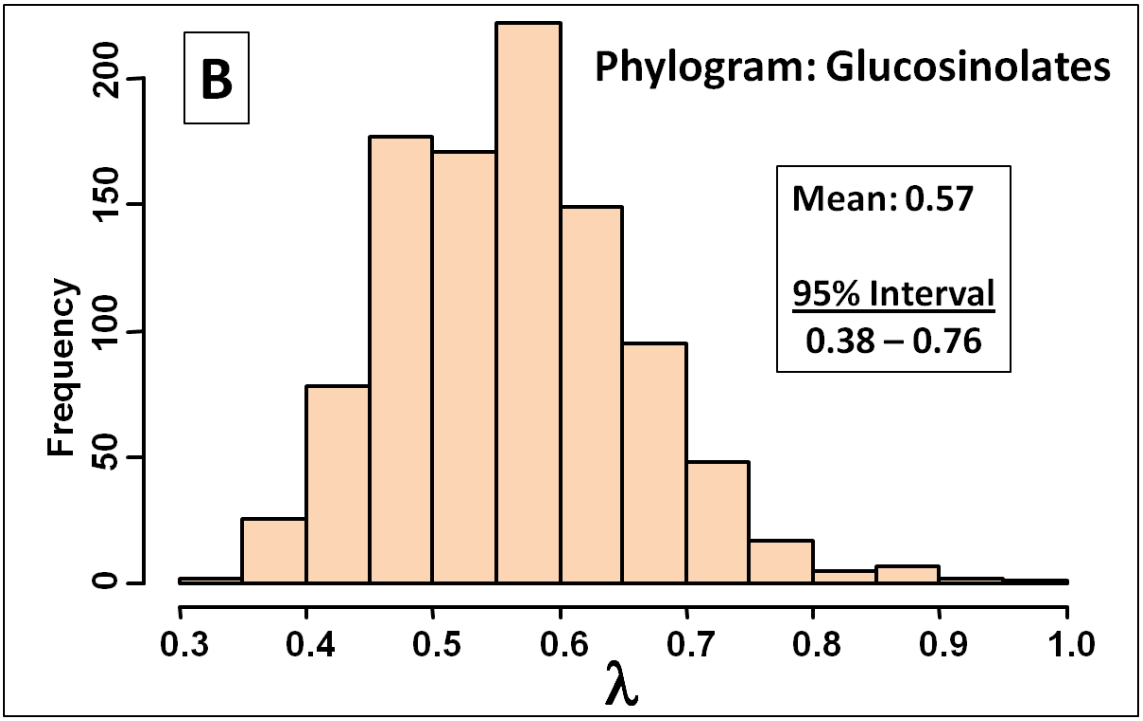


Figure 4. Empirical bootstrap distributions for Pagel's λ based on the genetic phylogeny (A) and the glucosinolate phylogeny (B).

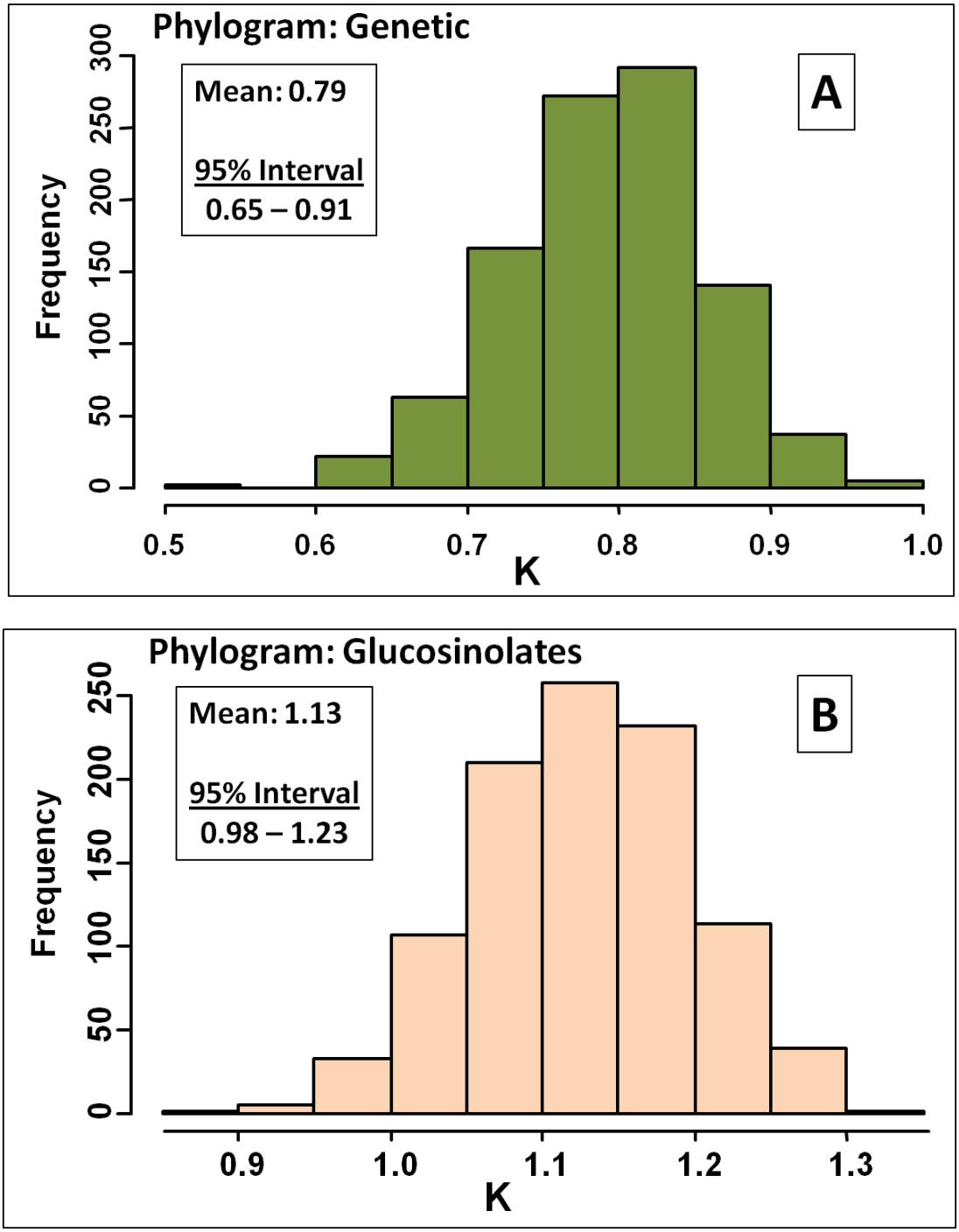


Figure 5. Empirical bootstrap distributions for Blomberg's K based on the genetic phylogeny (A) and the glucosinolate phylogeny (B).

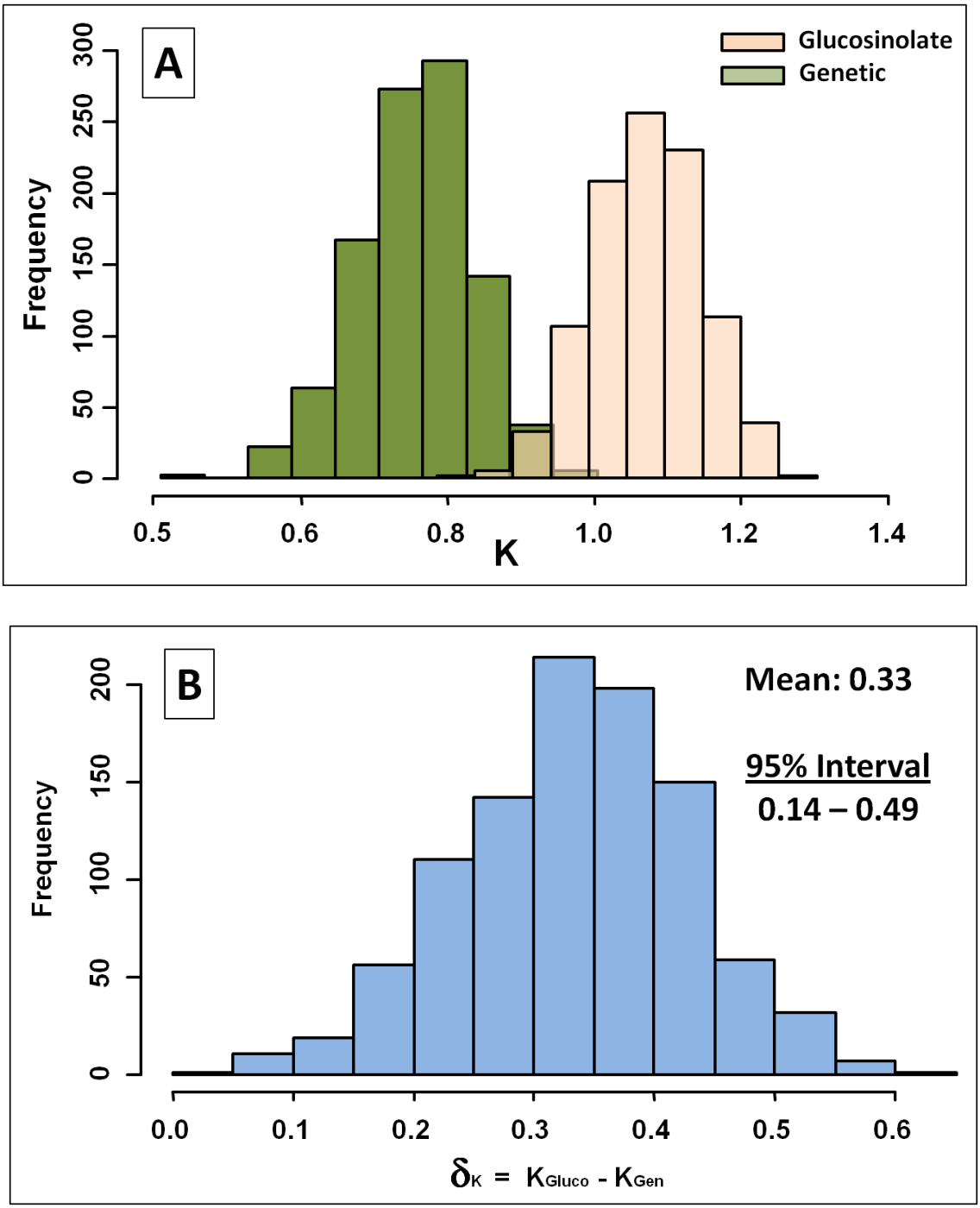


Figure 6. Comparative plots of the empirical bootstrap distributions for Blomberg's K based on genetic (K_{Gen}) or glucosinolate (K_{Gluco}) data (A) and the distribution of the difference, δ_K (B).