Permutation Tests for Analysing Cospeciation in Multiple Phylogenies

Mramba, Lazarus
Kenya Medical Research Institute
Department of Statistics-ICT
P.O. BOX 230
Kilifi (80108), Kenya
E-mail:lmramba@gmail.com

Gilks, Wally
University of Leeds, School of Mathematics
Department of Statistics
Leeds (LS2 9JT),United Kingdom
E-mail: W.R.Gilks@leeds.ac.uk

ABSTRACT

The purpose of this study is to develop permutation test statistics that can be used to analyse cospeciation in three phylogenies. The null hypothesis, H_0, is that the three phylogenies are not related, indicating that the host species evolved independently from their parasite species. The alternative hypothesis, H_1, is that there is a close relationship among the three phylogenies which could indicate cospeciation.

Three test statistics have been developed. The first one is the Pearson's partial correlation coefficient, $r_\star$, where $\star$ is either $xy.z$, $xz.y$ or $yz.x$. A triangular association matrix is used when computing the observed, $r_\star^{obs}$, and the permuted partial correlation coefficients, $r_\star^p$. To test the significance of $r_\star^{obs}$, three partial p values: P_x, P_y and P_z (for X, Y and Z held constant respectively) are calculated by summing $r_\star^p \geq r_\star^{obs}$ and dividing by the total number of permutations, N. To conclude on the overall significance of the results, a geometric p value, P_{gm}, is calculated.

The second test statistic is the eigenvalue, λ_r, computed from the correlation matrix based on a principal component analysis (PCA) method. With this statistic, the eigenvalues λ_r are derived from the correlation coefficients of matrix D, where matrix D takes pairs of triangles of relations from the patristic distances to form its rows and the columns are the three trees X, Y and Z. Only the first λ_r corresponding to PC1 is considered since it has the highest variance, is the largest and explains most of the proportion of variation in the data. Similarly, λ_r^{obs} and λ_r^p are calculated as above. To test the significance of the observed eigenvalues, a p value, P_{λ_r}, is computed by summing the number of $\lambda_r^p \geq \lambda_r^{obs}$ and diving by N.

The third test statistic is the eigenvalue, λ_c, computed using the covariance matrix. The procedure is exactly the same as the one discussed above and λ_c^{obs} and λ_c^p are calculated. A p value (P_{λ_c}) is also computed to test the significance of the results.

Computation of type I error suggests that using PCA correlation structure produces uniformly distributed p values as well as for the covariance structure but not when partial correlation statistic is used. These results are irrespective of the size of the phylogenies. The power to reject H_0 drops very fast as more association triangles are added or substituted when the partial statistic is used at 0.01 significance level for small phylogenies. When large phylogenies are used, the power to reject H_0 is found to be consistently high in all the three statistics.

In conclusion, the permutation method on PCA eigenvalues produces reliable test statistics that can be used to test for cospeciation of multiple phylogenies. The partial test statistics can be used but may give biased results for small phylogenies.

BACKGROUND

Phylogenetic trees are diagrammatic representations of the evolutionary relationships (Ewens and Grant,

2001) that occur between taxonomic groups. Trees can be rooted or unrooted. A biological phylogenetic tree must have a root, also called the ancestor of all the leaves. The leaves of a tree represent the species that are most current and are the terminal nodes of a rooted tree. Trees are said to be rooted if their branch lengths represent evolutionary time, with all species in the tree sharing a common ancestor. For cases of unrooted trees, there is no direction to indicate how the evolutionary time might have flowed. The trees are binary implying that an edge that branches will split to two daughter edges. The edges explain the evolutionary divergence that is associated to it and define the measure of distances between the species. Because of the fact that the current species share a common ancestor, they are likely to have similar traits in their DNA or protein sequences inherited from their common ancestor which has changed over evolutionary time through mutational processes such as substitution, deletion and insertion.

Phylogenetic trees convey two types of information. First, the topology defines the branching order of the trees and the way the species are distributed among the leaves. Secondly, the branch lengths represent phylogenetic time, measured by the average amount of mutational change.

The study of cospeciation is limited by the fact that it relies on the accuracy of trees. The assumption is that the host-parasite trees are accurate measures of the evolutionary time. The distance between two or more species from their recent common ancestor may be defined in terms of years if this is known. However, the evolutionary time is not accurately known hence surrogate distances are used instead. However, since all the computations are based on these phylogenetic trees, it is assumed here that the distances are known.

The use of phylogenetic trees and DNA or protein sequences in biology is fundamental to the understanding of the evolutionary relationships and is one of the primary driving tools for describing these associations. New developments in the study of host-parasite phylogenies have given insights into the complexity and necessity of reliable statistical methods that can be used to infer the history of an association between them (Page, 2003). However, reliable statistical test appropriate for assessing cospeciation of more than two parasite-host phylogenies in order to quantify various biological phenomena remain a statistical paradox (Choi and Gomez, 2009).

Cospeciation has been defined by Page (2003) as the joint speciation of lineages that are known to have an ecological association allowing for either of them to have speciated slightly before or after the other. A typical example is that of a host-parasite association. Coevolution may be defined as the joint evolution of any two or more associated organisms and when two phylogenetic trees have similar topologies, they are said to be congruent and therefore incongruent if they have discordance between their topologies (Page, 2003). This idea of congruency has been implicated with cospeciation (Brooks and McLennan, 1991) and that congruent phylogenies signal cospeciation whereas incongruence imply host switching. Also, there is a general consensus derived from Fahrenholz's rule (Fahrenholz, 1913), that parasites diversify together with their hosts. The rule states that parasites' phylogenies reflect hosts' phylogenies. However, complete divergence occurs only when cospeciation is regarded as the only exclusive process (Page, 2003). This is not likely to be the case as there are other biological and environmental processes that occur between host-parasite associations. Thus the incongruence of host-parasite phylogenies would be due to other events besides cospeciation such as parasites switching host lineages, duplication, where the parasites independently speciate from their hosts, extinction, parasites failure to speciate with their hosts, or failing to colonize the descendants of a whole speciating host-lineage (Paterson et al., 1999; Paterson and Gray, 1997; Page, 1990b, 1996b).

In parasitology, parasites have long been used to infer their host phylogeny (Klassen, 1992). If these parasites cospeciated with their hosts then the parasite phylogeny reflects the host phylogeny assuming that the rate at which the parasite evolves is lower than that of the host. Thus the parasite will retain some traits that the host lost. In contrary to this ideology, evidence from studies in molecular biology has shown that parasites evolve at a higher rate than their hosts (Hafner et al., 1994; Moran et al., 1995).

Page (2003) points out that a basic test of cospeciation is one that gives a significant similarity between the topologies of host and parasite phylogenies that is not due to chance alone. Questions about the timing of speciation emerge if host-parasite phylogenies are found to be identical.

Trees are used to analyse host-parasite cospeciation because the method is widely applicable to different types of datasets such as data from molecular and morphological designs. Trees can be represented in other formats that is not necessarily phylogeny. For instance, Becerra (1997) used chemical similarity to compare the phylogeny of host plants with their parasite (blepharida beetles) phylogeny because the insects' evolutionary history is more identical to the host chemistry than the host phylogeny.

All organisms in any ecosystem are linked biochemically (Ahmad et al., 2004). A simple relationship is composed of two or three trophic (eating) levels with feeding relationships. Hosts and parasites form complex food webs that span over several trophic levels. An association is said to be ditrophic when the species of one host are associated with that of one parasite. One common example is the gopher-lice association. The gophers represent the host relating with their parasite which is lice in this case. From this figure, it can be seen that the relationship can be complex in that a single species from the host may be infected by multiple parasite species or multiple host species infected by the same parasite species in addition to the one to one associations.

There are basically three trophic levels in an ecosystem. The first trophic level is composed of primary producers of energy, which are mainly plants. The second trophic level is made up of those that feed directly on the plants, called herbivores or primary consumers and the third level consists of predators. These three levels collectively form a titrophic system. Examples of titrophic relations include plants-herbivores-insects (pollinators), deceit-pollinated plants limited by the pollinators for seed set and the herbivores relying on the plants for fitness, and a parasitoids-plants-herbivores relationship (Micha et al., 2000), where varieties of plants and herbivores confront the parasitoids.

Huelsenbeck et al. (2000, 2001) employ maximum likelihood estimation in a Bayesian framework to analyse the ditrophic host-parasite relations whereas Hommola et al. (2009) use a permutation test. We seek to extend the permutation method to more than two phylogenetic associations.

Permutation methods are useful in that they do not require some parametric assumptions and are therefore reliable since they have no underlying distribution assumptions. One approach suggested by Lapointe and Legendre (1992) is to analyse the statistical significance of the matrix correlation coefficient by comparing independent phylogenetic trees and creating tables of critical values of the Pearson's cross product matrix correlation coefficient. However, our method does not require tables of critical values since the critical values are directly computed from the permuted statistics and compared with the observed statistics.

METHODS

Dataset is made up of simulated phylogenies using the ape package from R Development Core Team (2010) and multivariate statistical techniques employed with a permutation method to test the hypotheses stated below. The test statistics developed are partial correlation coefficients, PCA eigenvalues computed from the correlation matrix and eigenvalues computed from the covariance matrix. The permutation approach is preferred because it makes no assumptions underlying the distribution of the data.

The hypotheses are

H_0 : The host-parasite phylogenies are not related indicating that they have evolved independently;

H_1 : The host-parasite phylogenies are closely related indicating cospeciation.

The dataset consists of phylogenies generated under the null hypothesis H_0 or under the alternative hypothesis H_1 and an association matrix representing the interactions among the phylogenetic trees. A large number of phylogenies are generated under the permutation method and different statistics calculated.

Notation:

Let the trees be denoted as X, Y, Z. Let x denote a tip from tree X, y from tree Y and z from tree Z respectively.

Let (x, y, z) be a triple such that edges xy, xz and yz all exist in the interaction graphs of XY, XZ and YZ respectively such that (x, y, z) picks out a triangle of interactions.

Let T denote the set of all observed triples (x, y, z) and let n denote the number of elements of T where $n > 1$. Let (x_i, y_i, z_i) denote the ith triple of T.

Suppose i and j are distinct elements of T. Let $d(x_i, x_j)$ denote the patristic distance between tips x_i and x_j of tree X. Similarly denoting $d(y_i, y_j)$ and $d(z_i, z_j)$ for trees Y and Z respectively.

Generate a matrix D with $n(n-1)/2$ rows and three columns, where each row represents a distinct pair of triples i and j for $j < i$ and where each column represents a tree. Thus the $(i_1)(i_2)/2 + j$th row of D will contain $d(x_i, x_j), d(y_i, y_j), d(z_i, z_j)$.

Matrix D has $n(n-1)/2$ rows and the number of columns is equivalent to the number of trees. Each row will contain the triples: $d(x_i, x_j)$, $d(y_i, y_j)$, $d(z_i, z_j)$.

Matrix D can be formed from trees that have been generated under H_1 to represent cospeciation. The trees X, Y and Z are set to be very close, by generating them to have the same topologies such as $X = Y = Z$, and to have a triangular association matrix that has associations at corresponding positions of the trees such that $T_1 = (1, 1, 1), T_2 = (2, 2, 2), T_3 = (3, 3, 3), T_4 = (4, 4, 4)$ and $T_5 = (5, 5, 5)$.

Inference for partial correlation

Let the sample partial correlation coefficients from the observed data be denoted by $r^{obs}_{xy.z}$, $r^{obs}_{yz.x}$ and $r^{obs}_{xz.y}$. For each permutation i of the labels of the trees, let the sample partial correlation coefficients obtained after a large number of permutations, say N be denoted by $r^{p}_{xy.z_i}$, $r^{p}_{yz.x_i}$ and $r^{p}_{xz.y_i}$.

To test the significance of the observed partial correlations, the partial p values are computed by summing the number of permuted partial correlation coefficients that are greater than or equal to the observed partial correlation coefficient and divided by the number of permutations (equation 1). The geometric p value of the three partial p values is then computed as the root of the product of the three p values given by equation 4.

The partial p value after controlling for variable Z is calculated as:

$$(1) \qquad P_z = \frac{1}{N}\sum_{i=1}^{N} I(r^{p}_{xy.z_i} \geq r^{obs}_{xy.z}),$$

where N is the total number of permutations, and $r^{p}_{xy.z_i}$ is the partial correlation coefficient between variables X and Y while controlling for variable Z for the ith permutation.

The partial p value after controlling for variable X is calculated as:

$$(2) \qquad P_x = \frac{1}{N}\sum_{i=1}^{N} I(r^{p}_{yz.x_i} \geq r^{obs}_{yz.x}).$$

and likewise, the partial p value after controlling for variable Y is given by

$$(3) \qquad P_y = \frac{1}{N}\sum_{i=1}^{N} I(r^{p}_{xz.y_i} \geq r^{obs}_{xz.y}).$$

Here,

$$I(r^{p}_{\star} \geq r^{obs}_{\star}) = \begin{cases} 1 & \text{if} \quad r^{p}_{\star} \geq r^{obs}_{\star} \\ 0 & \text{otherwise} \end{cases}$$

where $\star$ stands for either $xy.z_i$, or $yz.x_i$ or $xz.y_i$.

A geometric p value may be used as an overall test of significance of the observed partial correlation coefficients and is given by

$$(4) \qquad P_{gm} = \prod_{i=1}^{p} \{P_z P_x P_y\}^{1/p}.$$

If $P_{gm} \leq \alpha$, where α is the significance level, H_0 is rejected and a conclusion made that there is a close relationship among the three host-parasite systems which could indicate cospeciation. Otherwise, there is no evidence to reject H_0.

Inference for PCA

Computations are done on matrix D, the matrix that takes pairs of the triangular relationships from the patristic distances of the three phylogenetic trees X, Y, and Z. The observed eigenvalues are computed from both the covariance matrix and the correlation matrix before any permutations to the data has been done. Each of these produces three eigenvalues. The first, second and third eigenvalues correspond to the first, second and third principal components respectively. For each permutation i, for say a large number of permutations, N, calculate permuted eigenvalues. Only the first eigenvalues for both the covariance and correlation structures are used for the statistics. This is because they are usually the largest in size, and sometimes the eigenvalues for PC2 and PC3 can be negligibly small, they explain the largest proportion of variation in the data and have the highest variance.

The p values are then calculated as

$$P_{\lambda_r} = \frac{1}{N}\sum_{i=1}^{N} I(\lambda_{r_i}^p \geq \lambda_r^{obs}) \tag{5}$$

$$P_{\lambda_c} = \frac{1}{N}\sum_{i=1}^{N} I(\lambda_{c_i}^p \geq \lambda_c^{obs}), \tag{6}$$

where P_{λ_r} is the p value due to the correlation structure and P_{λ_c} is the p value due to the covariance structure and where

$$I(\lambda_\star^p \geq \lambda_\star^{obs}) = \begin{cases} 1 & \text{if} \quad \lambda_\star^p \geq \lambda_\star^{obs} \\ 0 & \text{otherwise} \end{cases}$$

where $\star$ stands for either r_i, or c_i.

Decision is made based on the p values and the set level of significance, α.

Permutation Algorithm

Step 1: Set the significance level α.

Step 2: For each of the three phylogenetic trees X, Y and Z, calculate their patristic distances.

Step 3: Compute matrix D using both the patristic distances and the association matrix.

Step 4: Using matrix D, calculate the three observed pairwise correlation coefficients: r_{xy}^{obs}, r_{yz}^{obs} and r_{xz}^{obs} and partial correlation coefficients $r_{xy.z}^{obs}$, $r_{yz.x}^{obs}$ and $r_{xz.y}^{obs}$.

Step 5: Compute the observed eigenvalues from both the covariance and the correlation matrices and consider the eigenvalues from the first principal components.

Step 6: Permute the labels of the trees.

step 7: Repeat steps 2 to 6 for a large number of times, say N. For each permutation i, compute the permuted partial correlations coefficients $r_{xy.z_i}^p$, $r_{yz.x_i}^p$ and $r_{xz.y_i}^p$ and the permuted eigenvalues $\lambda_{r_i}^p$ and $\lambda_{c_i}^p$.

Step 8: Compute the three partial p values P_z, P_x and P_y. Compute the geometric mean of these p values, P_{gm}. Compute the p values of the eigenvalues P_{λ_r} and P_{λ_c}.

Step 9: Practically, only one method is applied. This could be either the partial correlation or the PCA using the correlation structure or using the covariance structure. In either case reject H_0 if $P \leq \alpha$, else there is no evidence to reject H_0.

Permutations under the Null hypothesis

The three phylogenetic trees namely X, Y and Z are randomly generated under H_0 to represent species that have evolved independently, given their phylogenetic trees and their association matrix. The triangular association matrix should be randomly generated. All the trees are generated randomly with a number of tips.

A function called $nperm$ has been developed. The parameters required by this function are: the three phylogenetic trees, a triangular association matrix and the number of permutations desired. This function computes the patristic distances of each of the trees and uses the association matrix to build matrix D. It then computes the pairwise correlation coefficients, observed and permuted partial correlation coefficients and observed and permuted eigenvalues from both covariance and correlation matrices. To test the significance of the observed statistics, the partial geometric p value and the p values due to the eigenvalues from both the correlation and covariance structure are computed.

Density plots are drawn with lines of both the observed statistics and the critical value line for the first $95th$ percentile of the permuted coefficients and eigenvalues. If the observed statistics lie below the critical value then it means H_0 is not rejected.

To investigate type I error, data is generated under the H_0 and a random association matrix used. 1000 p values are calculated for each of the statistic and their cumulative distributions checked for uniformity of the p values. The labels of the trees are permuted a large number of times for each p value calculated. For the statistic to be reliable, the p values generated under H_0 should be uniformly distributed.

Power Simulations

To compute the power, the phylogenies are generated under the alternative hypothesis H_1, making them to be exactly the same and having interaction triangles at their corresponding positions. For instance, let the first phylogeny, X, be generated randomly with ten tips. The second and third phylogenies can be made similar to X by either adding say δ and Φ to all branch lengths to get trees Y and Z respectively or by setting trees Y and Z to be exactly the same trees as tree X. For n = 10 triangles, there will be $n(n-1)/2 = 10(9)/2 = 45$ observations in the matrix D and three columns X, Y and Z representing the trees. The association matrix forms triangles T, where $T_1 : (x_1, y_1, z_1), T_2 : (x_2, y_2, z_2), \ldots, T_{10} : (x_{10}, y_{10}, z_{10})$.

To compute the power for the three statistics, two approaches have been implemented, adopted from Hommola et al. (2009) and Legendre et al. (2002): The first approach is to add random triangles as a percentage of the existing number of triangular associations. The phylogenies are generated each with ten tips, giving ten corresponding association triangles. For power simulations, 100 p values can be calculated, while permuting the labels of the trees a larger number of times, say 10000. A sum of the p values greater than α is noted. Plots for rejection rate against the percentage of added triangles are drawn.

The second approach involves randomly substituting corresponding triangles with random ones as a percentage of the existing triangles. This approach has been done for phylogenies with both 10 and 20 tips by substituting $10\%, 20\%, \ldots, 50\%$ of the association triangles randomly.

RESULTS

Simulating the data under the null hypothesis and setting a $seed$ of 100, the observed partial correlation coefficients obtained were: $r^{obs}_{xy.z} = -0.075$, $r^{obs}_{yz.x} = 0.0203$, $r^{obs}_{xz.y} = 0.1505$, and the observed eigenvalues for the three principal components are given in table 1. The observed Pearson's pairwise correlation coefficients

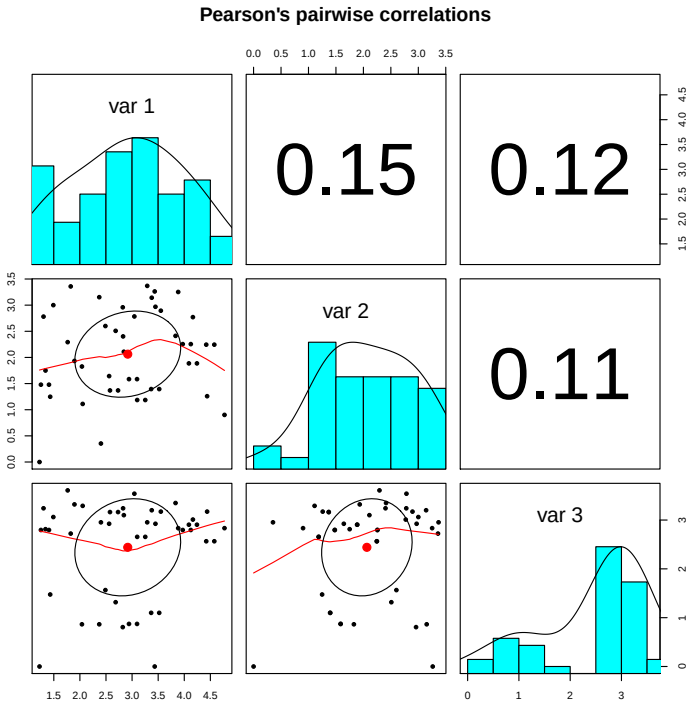

Figure 1: ***Observed pairwise correlation coefficients under*** H_0***. var 1, var 2 and var 3 represents*** X***,*** Y ***and*** Z ***respectively***.

	PC1	PC2	PC3
Importance of components			
Standard deviations	1.078	1.004	0.911
Proportion of Variance	0.388	0.336	0.277
Cumulative Proportion	0.388	0.723	1.000
Eigenvalues: λ_r^{obs}	1.163	1.007	0.830
Eigenvalue: λ_c^{obs}	1.434	1.198	0.999

Table 1: ***Principal components under*** H_0**.**

are displayed in figure 1 whereas figure 2 displays a 3D scatter plot.

The three trees' labels are permuted 10,000 times and for each permutation, the pairwise correlation coefficients, partial correlation coefficients and the eigenvalues for the first principal components (PC1) are used. The geometric mean of the partial p values, $P_{gm} = 0.347$, $P_{\lambda_r} = 0.699$ and $P_{\lambda_c} = 0.860$. The density plots are as shown in figure 3.

In order to calculate the type I error, 1000 p values are calculated for trees with 10 tips and trees with 15 tips. 1000 permutations on the labels of the trees is performed. Thus there will be 1000 p values for each of the statistic. Empirical cumulative distribution function of these results are displayed in figure 4. It is evident from these plots that statistics from using the principal component analysis produces uniformly distributed p values, this being not the case when the partial correlations coefficient statistic is used. These plots suggest the use of PCA technique as being a more reliable statistic than its counterpart.

Results from using a relatively larger number of tips such as 15 showed that it does not change type I error distribution of the p values. The plots are given in figure 5.

Under the perfect alternative hypothesis approach where trees $X = Y = Z$ and and with triangular associations at their corresponding positions, the observed pairwise correlation coefficients are $r_{xy}^{obs} = r_{yz}^{obs} =$

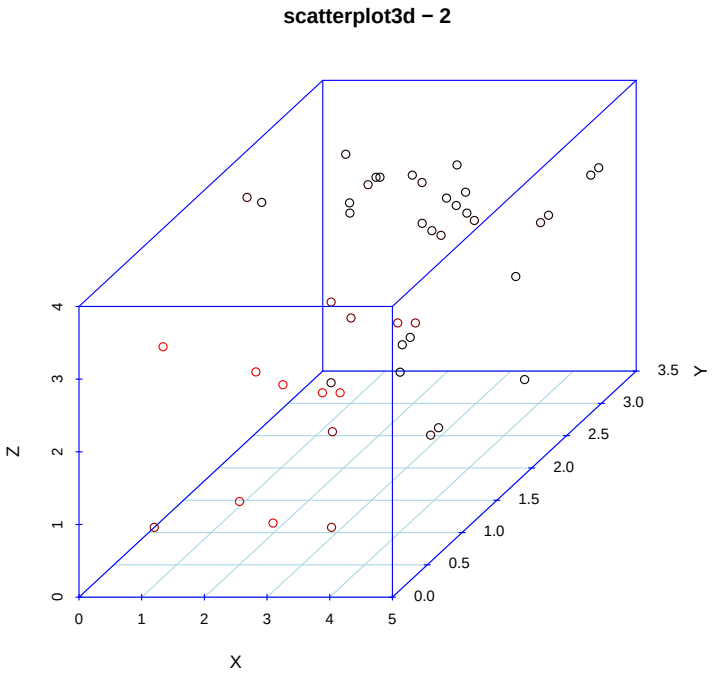

Figure 2: ***Observed Results under*** H_0**.**

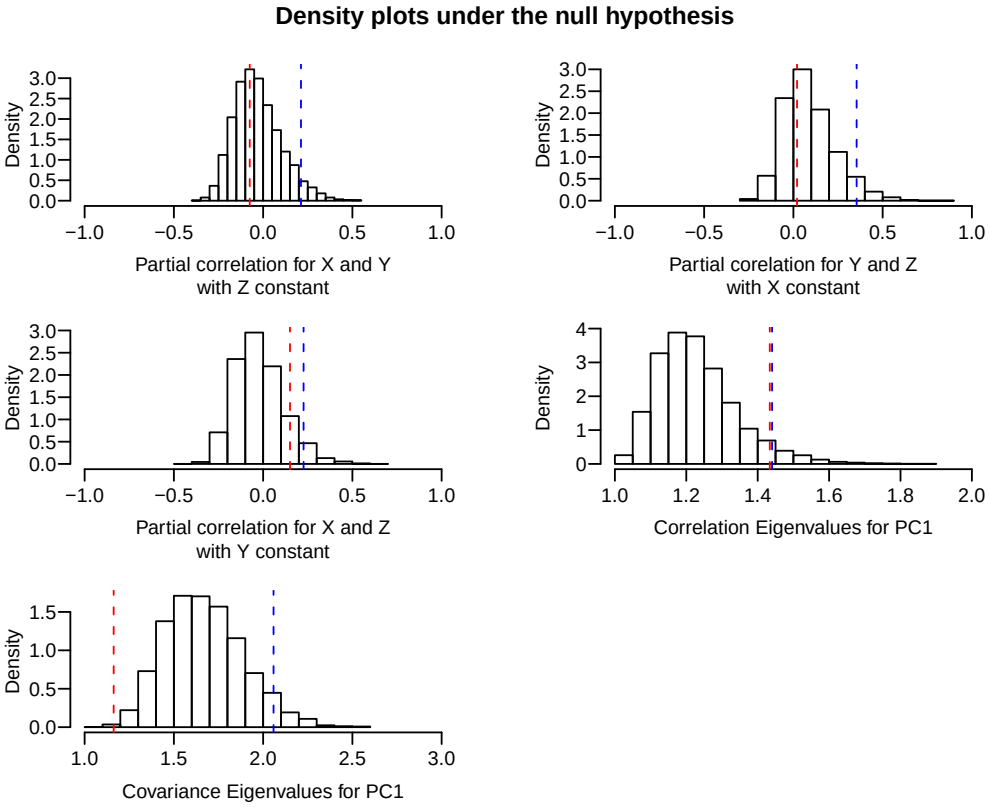

Figure 3: ***Density plots for permuted coefficients and critical values under*** H_0**. The red line represent the observed value and blue represents the critical line at the 95th percentile.**

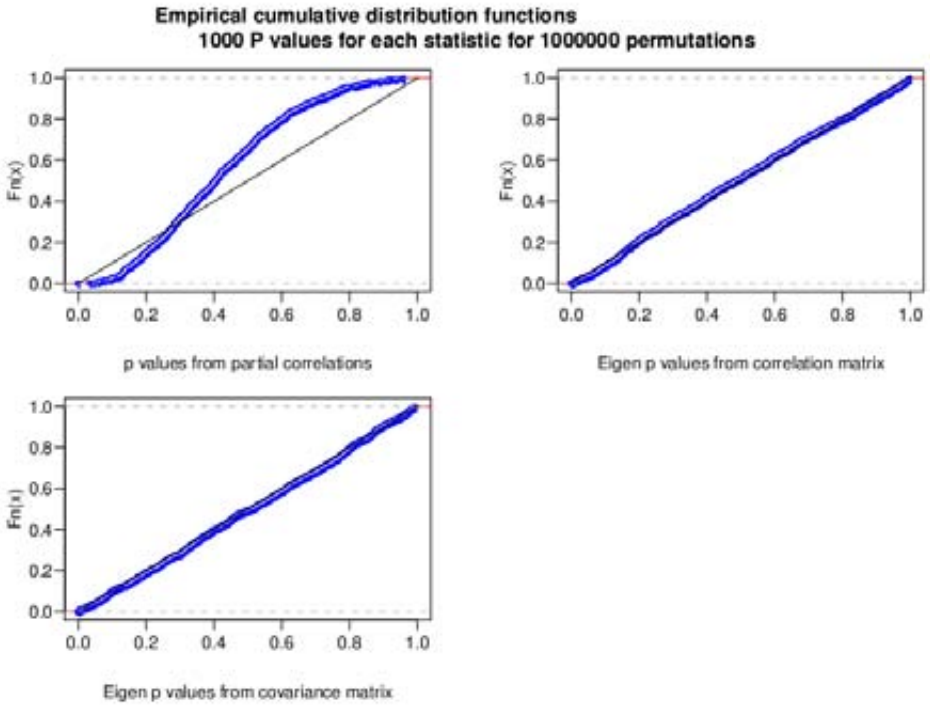

Figure 4: ***Type I error plots for trees with 10 tips performed for 1000*** p ***values permuting the tree labels 10000 times.***

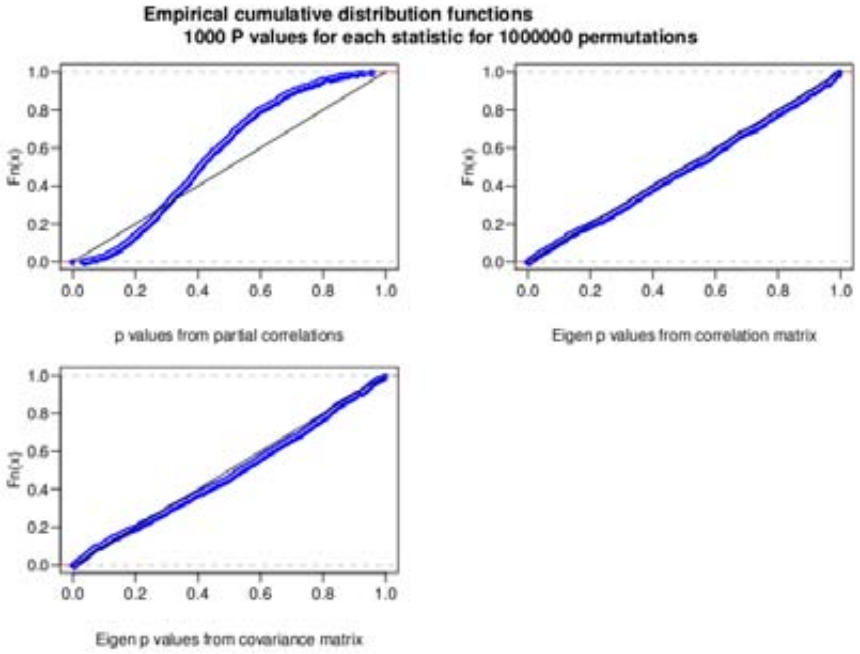

Figure 5: ***Type I error plots for the distribution of 1000*** p ***values for trees with 15 tips, permuting the tree labels 1000 times.***

	PC1	PC2	PC3
Importance of components			
Standard deviations	1.73	0.00	0.00
Proportion of Variance	1.00	0.00	0.00
Cumulative Proportion	1.00	1.00	1.00
Eigenvalues: λ_r^{obs}	3.00	0.00	0.00
Eigenvalue: λ_c^{obs}	3.09	0.00	0.00

Table 2: ***Principal components under perfect conditions of*** H_1**.**

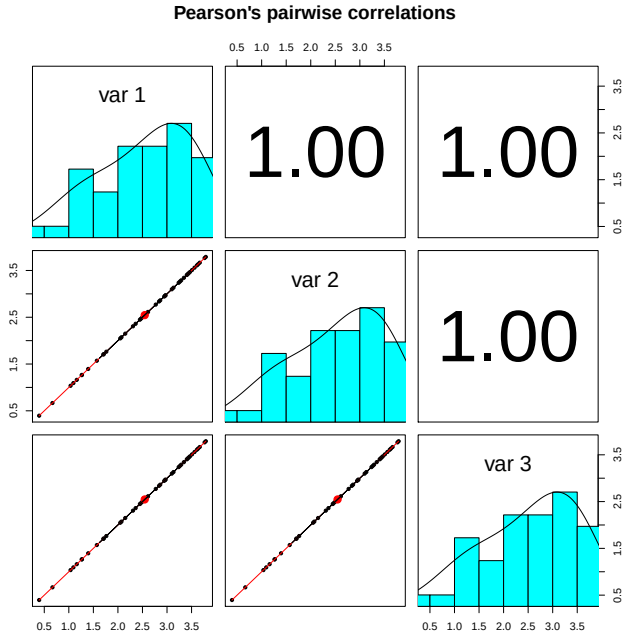

Figure 6: ***Pairwise correlation coefficients produced under a perfect*** H_1**. Var 1, var 2 and var 3 stand for** X**,** Y ***and*** Z ***from the matrix*** D**.**

$1r_{xz}^{obs} = 1$. This makes the partial correlation coefficients $r_{xy.z}^{p}$, $r_{yz.x}^{p}$ and $r_{xz.y}^{p}$ not computable. The pairwise plot is given in figure 6 and table 2 summarizes the PCA results obtained.

Figure 7 shows the density plots from the PCA method. Since the observed eigenvalues lie in the critical region, this means that there is a strong evidence to reject H_0 and conclude that the trees have a close linear relationship which could indicate cospeciation. This is consistent with the p values $P_{\lambda_r} = 0$ and $P_{\lambda_c} = 0$ obtained.

Figure 8 displays plots that are strikingly different from the ones obtained under H_0. The first plot on the first row shows that PC1 explains 100% of the total variation in the data. The standard deviation for PC1 is 1.73 while it is negligibly small for PC2 and PC3. All the box plots are similar and the 3D plot in row three has a straight line of points along the $x = y = z$ line indicating dependency of the observed trees.

Figure 9 displays the power curves for trees with 10 tips. It is evident from these plots that the power to reject H_0 diminishes as more random triangles are added. This is because, adding random triangles weakens the condition for simulation under H_1. The more the random triangles are added, the more H_0 condition is approached and therefore the lower the power to reject H_0.

Computations to test the power of the three statistics to reject the null hypothesis for large phylogenetic trees when random triangles are added has been performed on trees with 20 tips. The results are plotted in figure 10. This plot shows that the power to reject H_0 is very high for large phylogenetic trees than for small

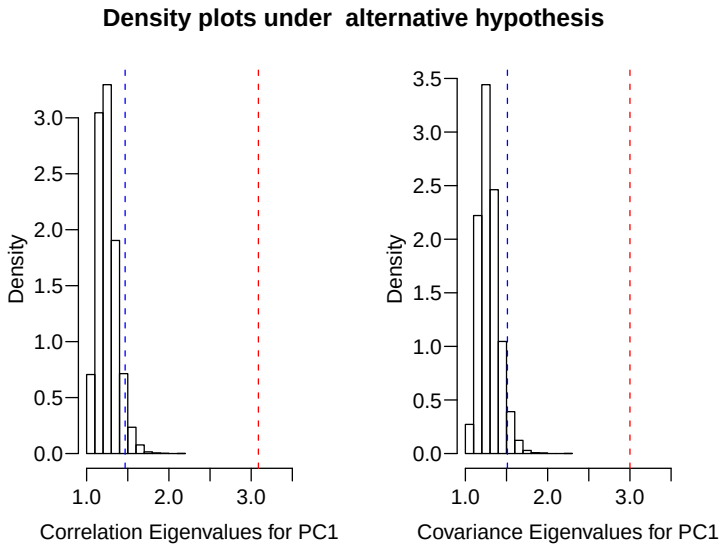

Figure 7: ***Density plots for permuted coefficients and critical values under*** H_1***. The red line represent the observed value and blue represents the critical line at the 95th percentile.***

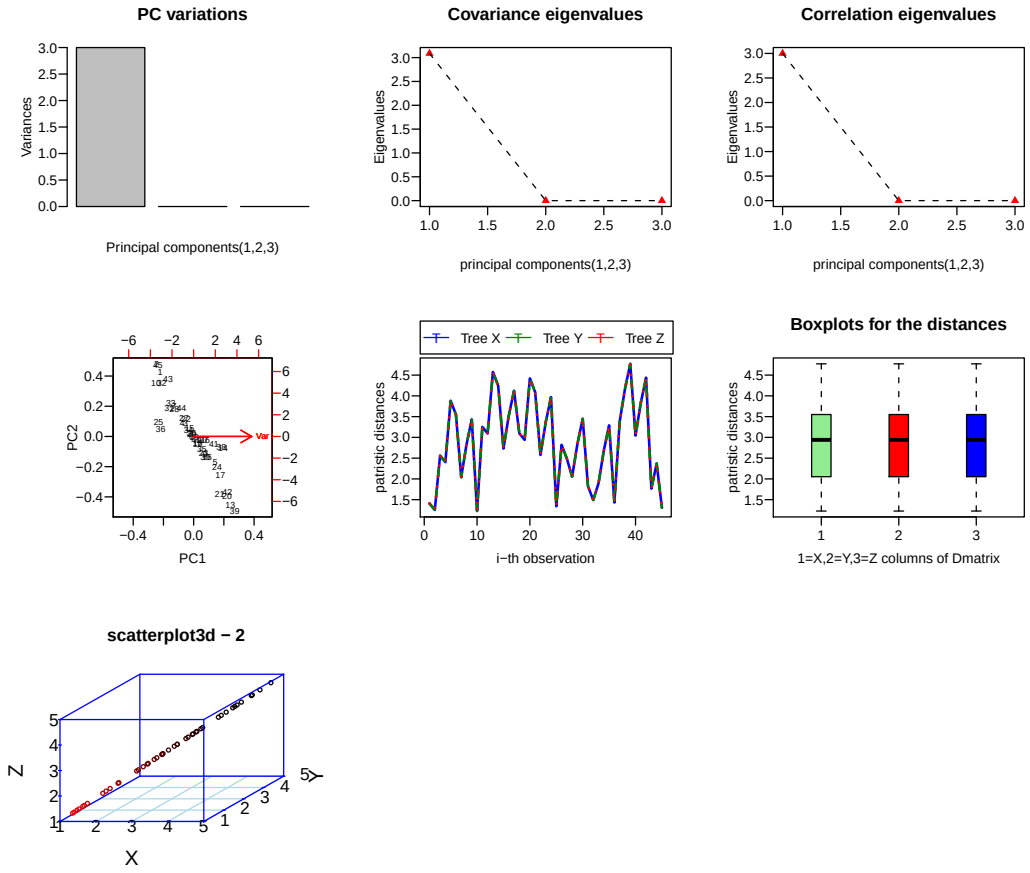

Figure 8: ***Principal components analysis loadings under a perfect*** H_1***. All the plots reveal an overwhelming evidence of correlations among the phylogenies.***

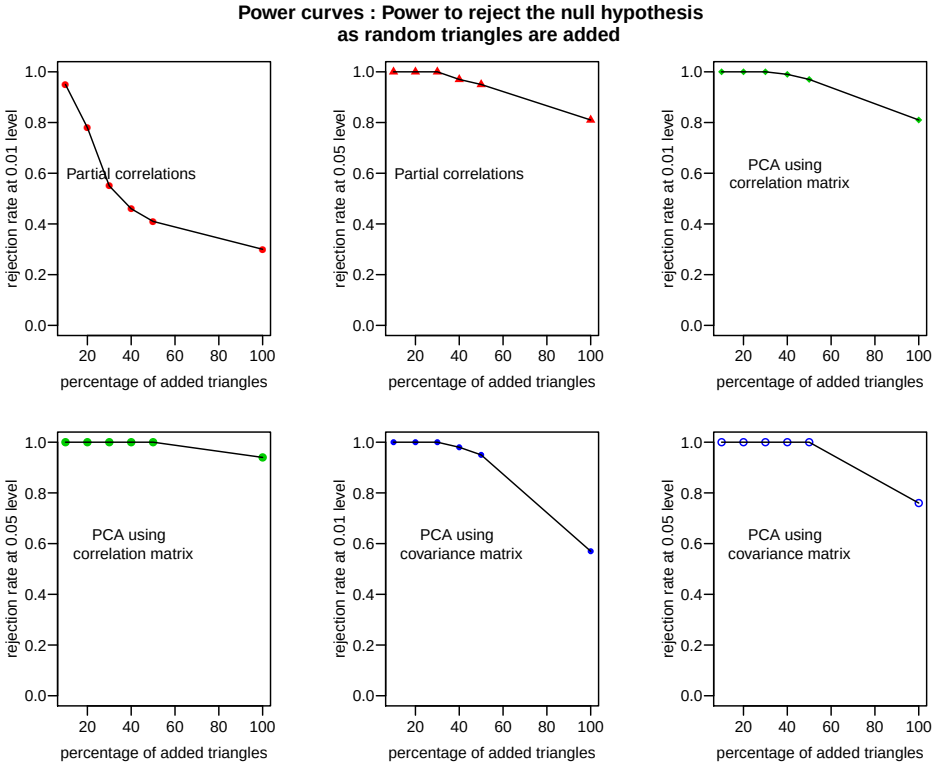

Figure 9: ***Power curves for trees with 10 tips when random triangles are added to the existing association matrix generated under*** H_1.

phylogenetic trees.

Power curves generated using the second approach for the trees with 10 tips are given in figure 11. Power to reject H_0 remain high for all the three statistics for the first 30% of substituting triangles but drops significantly as 50% of these association triangles become substituted. The results are displayed in figure 12.

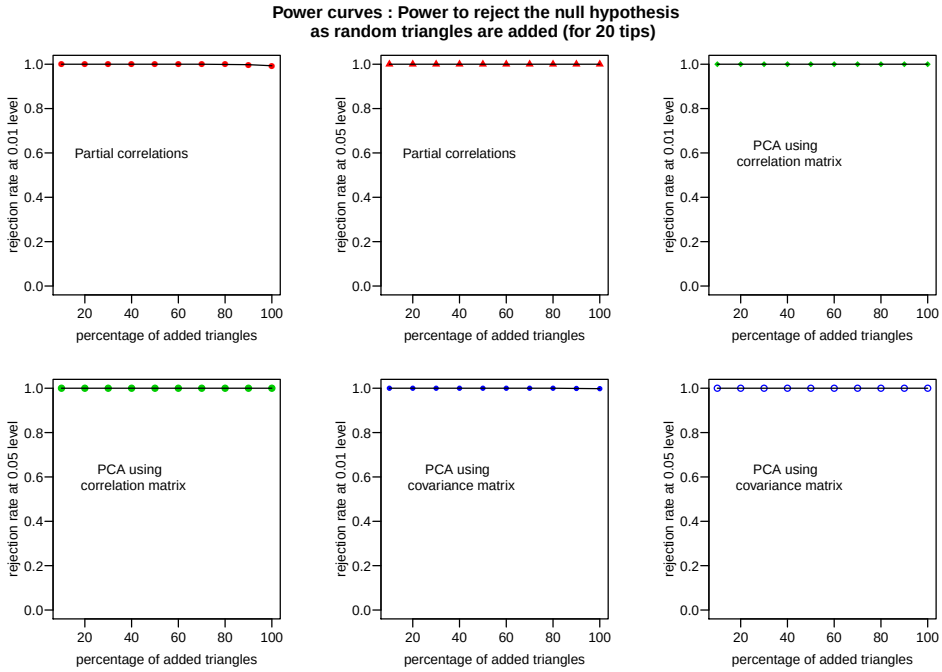

Figure 10: ***Power curves for trees with 20 tips as a result of adding random triangles to the existing association matrix.***

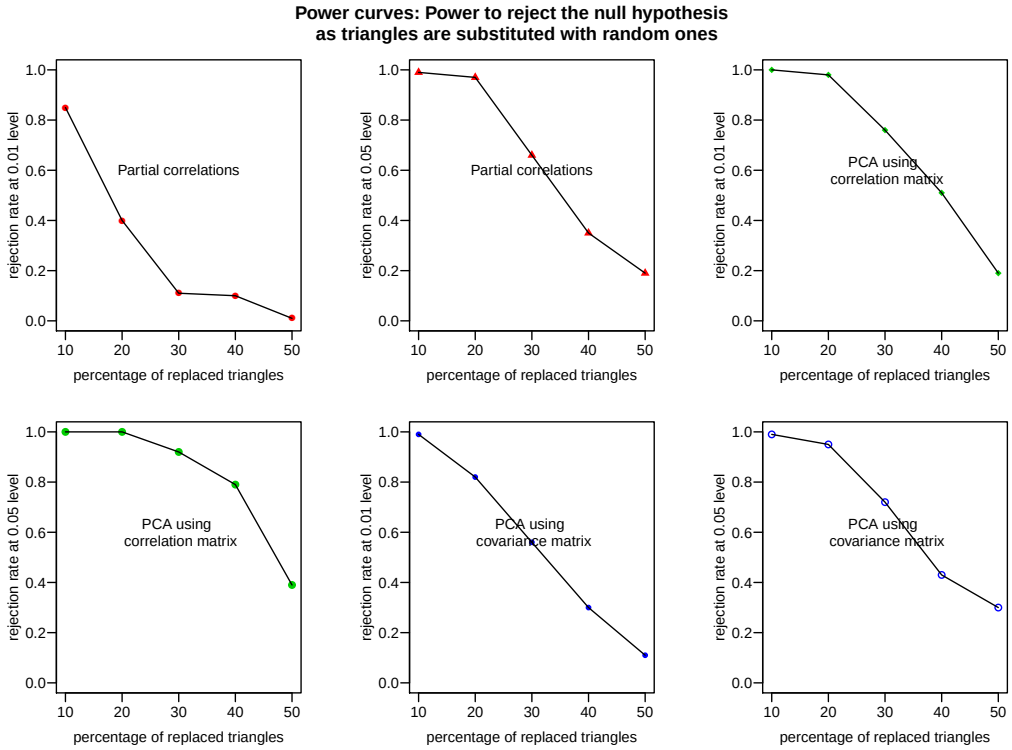

Figure 11: ***Power curves for trees with 10 tips as a result of substituting random triangles into the association matrix.***

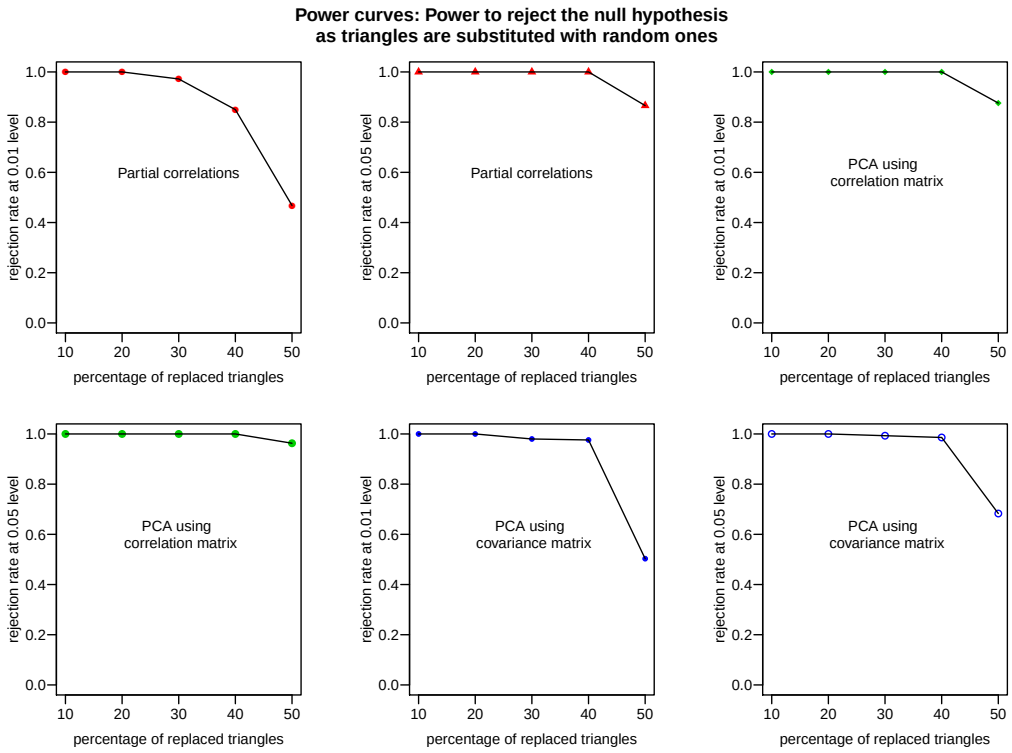

Figure 12: ***Power curves for trees with 20 tips, with random substitution of triangles.***

DISCUSSION

The three test statistics developed are useful in testing the cospeciation of the three phylogenies. All the three test statistics give a similar conclusion when applied on the data under the null hypothesis. Taking a significance level, α = 0.05, the observed partial correlations are insignificantly small as all the p are above the significance level. A similar conclusion is made at 0.01 significance level. However, when the type I error is performed on the data, the p values from the partial correlation coefficients do not portray a uniform distribution. This implies that the test statistic is unreliable. The type I error results are not any different when trees with 10 tips are used and when trees with 15 tips are used.

The empirical distribution density functions for both eigenvalues using correlation and using covariance matrix show that their p values have uniform distribution for both trees with 10 tips and 15 tips, which is a positive result.

When simulations are performed under the alternative hypothesis with trees $X = Y = Z$, and with triangular associations at their corresponding positions, the partial correlation coefficient statistic fails to be useful because the coefficients can no longer be estimated. This is because of the partial correlation formula which uses the pairwise correlation coefficients. Thus in this situation, no test of significance can be performed using this test statistic. The weakness of the partial test statistic makes the use of the eigenvalues preferred since it is not limited by this fact. Hence, using the PCA technique to compute the two test statistics: eigenvalues under covariance and under correlation matrices is preferred. The proportion of variance that is explained by the first principal component is 100% and the p values from both of these statistics is zero. So, there is overwhelming evidence to reject H_0 of no relationship and conclude that these species have cospeciated.

Random triangles are added in one approach and replaced in another, to the existing association matrix whose relation is at corresponding positions. This is done in order to test the power of the three test statistics. The results obtained when 10 tips are used for each tree and when random triangles are added show a relatively poor performance of the partial correlation coefficient statistic performed at 0.01 significance level because the

power drops very fast as these trees are added. The gradient is not very sharp for all the other statistics at the 0.01 and 0.05 significance level. However, when large phylogenies are used such as trees with 20 tips, the power improved in all the three statistics.

When the original triangles are substituted with random triangles for trees with 10 tips, the same characteristic of the partial statistic is observed. The gradient is very steep for the partial statistic at 0.01 significance level. However, when large phylogenies with 20 tips are used, the power improves a great deal and remains high for at least the first 30% of the substitutions in all the three statistics. The power curves obtained when correlation matrix is used is very high and only starts to drop from 50% substitutions. This is case for both 0.01 and 0.05 significance levels.

It is worthy noting that the geometric p value, P_{gm} computed as a summary of the three partial p values, P_x, P_y and P_z, is affected by any of these values being zero. If it happens that any of the partial p values is zero, then the final P_{gm} will also be zero indicating an overwhelming evidence to reject H_0.

CONCLUSION

The three test statistics can be used to test for cospeciation of multiple phylogenies, with the PCA technique giving the most reliable results. The partial correlation coefficient test statistic is less efficient for small phylogenies.

References

Ahmad, F., Aslam, M. and Razaq, M. (2004). Chemical ecology of insects and tritrophic interactions, *Journal of Research (Science)* **15**: 181–190.

Becerra, J. X. (1997). Insects on Plants: Macroevolutionary chemical trends in host use, *Journal of Science* **276**: 253–56.

Brooks, D. R. and McLennan, D. A. (1991). *Phylogeny, ecology and behaviour*, The university of Chicago Press, Chicago and London.

Choi, K. and Gomez, S. M. (2009). Comparison of phylogenetic trees through alignment of embedded evolutionary distances, *Journal of BMC Bioinformatics* **10**: 423.

Ewens, W. J. and Grant, G. R. (2001). *Statistical Methods in Bioinformatics*, Statistics for Biology and Health, Springer.

Fahrenholz, H. (1913). Ectoparasiten und Abstmmungslehre, *Journal of Zoology* **41**: 371–374.

Hafner, M. S., Sudman, P. D., Villablanca, F. X., Spradling, T. A., Demastates, J. W. and Nadler, S. A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites, *Journal of Science* **265**: 1087–90.

Hommola, K., Smith, J. E., Qiu, Y. and Gilks, W. R. (2009). A Permutation Test of Host-Parasite Cospeciation, *Journal of Molecular Biology Evolution* **26**: 1457–1468.

Huelsenbeck, J. P., Rannala, B. and Larget, B. (2000). A Bayesian framework for the analysis of cospeciation, *Journal of Evolution* **54**: 352–364.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001). Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology, *Science Journals: Review* **294**: 2310–2314.

Klassen, G. J. (1992). A history of the macroevolutionary approach to studying host-parasite associations, *Journal of parasitology* **78**: 573–87.

Lapointe, F.-J. and Legendre, P. (1992). Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees, *Journal of Systematic Biology* **41**: 378–384.

Legendre, P., Desdevies, Y. and Bazin, E. (2002). A statistical test for host-parasite coevolution, *Journal of Systematic Biology* **51**: 217–234.

Micha, S. G., Kistenmacher, S., Mölck, G. and Wyss, U. (2000). Tritrophic interactions between cereals, aphids and parasitoids: Discrimination of different plant-host complexes by Aphidius Rhopalosiphi, *European Journal of Entomology* **97**: 539–543.

Moran, N. A., VanDohlen, C. D. and Baumann, P. (1995). Faster evolutionary rates in endosymbiotic bacteria than in cospeciating hosts, *Journal of Molecular Evolution* **41**: 727–31.

Page, R. D. M. (1990b). Temporal congruence and cladistic analysis of biogeography and cospeciation, *Journal of Systematic Zoology* **39**: 205–26.

Page, R. D. M. (1996b). Temporal congruence revisted: Comparison of mitochondrial DNA sequence divergence in cospeciating pocket gophers and their chewing lice, *Journal of Systematic Biology* **45**: 151–67.

Page, R. D. M. (2003). *Tangled trees: Phylogeny, cospeciation, and coevolution*, The University of Chicago Press, Chicago and London.

Paterson, A. M. and Gray, R. D. (1997). *Host-Parasite cospeciation, host switching and missing the boat*, Vol. Probability and Mathematical Statistics of *Host-Parasite Evolution: General Principles and Avian Models*, Oxford: Oxford Press, University of Leeds,UK.

Paterson, A. M., Palma, R. L. and Gray, R. D. (1999). How frequently do avian lice miss the boat?, *Journal of Systematic Biology* **48**: 214–23.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*