Four-Cluster Analysis: A Simple Method to Test Phylogenetic Hypotheses

Andrey Rzhetsky, Sudhir Kumar, and Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

A simple statistical test for comparing three alternative phylogenetic hypotheses for four monophyletic groups is presented. This test is based on the minimum-evolution principle, and it does not require any information regarding the branching order within each monophyletic group. It is computationally efficient and can be easily extended to five or more monophyletic groups.

Introduction

We propose a simple method for inferring the phylogenetic relationships among four monophyletic groups of species by comparing all three alternative phylogenetic hypotheses. In this method, the minimum-evolution tree (Rzhetsky and Nei 1992) can be chosen without knowing the branching pattern of species within each of the four clusters. This is particularly useful for determining the branching pattern of a deep phylogeny using a large number of species. Using the four-cluster analysis, we reexamined the relationships of a group of extinct ratite birds to extant ratite species (Cooper et al. 1992).

Method

Rzhetsky and Nei (1993) recently showed that the expectation of the sum (S) of branch-length estimates is smallest for the true tree when the branch lengths are estimated by the ordinary least-squares method. Let A, B, C, and D be the four monophyletic groups (clusters) and suppose that A, B, C, and D contain n_A, n_B, n_C, and n_D species, respectively. In this case, one of the three possible trees for the four clusters (see fig. 1) must be the correct one, and the expected value of S for this tree should be smallest. That is, if S_I, S_{II}, and S_{III} are the sums of branch lengths for trees I, II, and III in figure 1, we need to compute $S_I - S_{II}$, $S_I - S_{III}$, and $S_{II} - S_{III}$, and test whether one of the S values is significantly smaller than the other two. Using equation (2) in Rzhetsky and Nei (1993) to compute the sums of branch lengths for two trees, say S_X and S_Y and taking the difference between them, we obtain

$$S_X - S_Y = \sum_{i<j} w_{ij} d_{ij}, \tag{1}$$

where d_{ij} is the estimate of evolutionary distance between species i and j, and X and Y refer to the trees compared. The coefficients w_{ij}'s are computed by

$$w_{ij} = \begin{cases} \alpha/(2n_A n_B), & \text{if } i \in \text{A and } j \in \text{B}, \\ \alpha/(2n_C n_D), & \text{if } i \in \text{C and } j \in \text{D}, \\ \beta/(2n_A n_C), & \text{if } i \in \text{A and } j \in \text{C}, \\ \beta/(2n_B n_D), & \text{if } i \in \text{B and } j \in \text{D}, \\ \gamma/(2n_A n_D), & \text{if } i \in \text{A and } j \in \text{D}, \\ \gamma/(2n_B n_C), & \text{if } i \in \text{B and } j \in \text{C}. \end{cases} \tag{2}$$

("$i \in A$" stands for "species i belongs to group A"), where α, β, and γ are computed by equations (3), (4), and (5), respectively.

$$\alpha = \begin{cases} \dfrac{n_A n_B + n_C n_D}{(n_A + n_C)(n_B + n_D)}, \\ \quad \text{if } X = \text{I and } Y = \text{II}, \\ \dfrac{n_A n_B + n_C n_D}{(n_A + n_D)(n_B + n_C)}, \\ \quad \text{if } X = \text{I and } Y = \text{III}, \\ \dfrac{(n_A - n_B)(n_D - n_C)(n_A n_B + n_C n_D)}{(n_A + n_C)(n_B + n_D)(n_A + n_D)(n_B + n_C)}, \\ \quad \text{if } X = \text{II and } Y = \text{III}. \end{cases} \tag{3}$$

Key words: Four-cluster analysis, phylogenetic inference, minimum evolution tree, ratites.

Address for correspondence and reprints: Andrey Rzhetsky, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, 328 Mueller Laboratory, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 12(1):163–167. 1995.

0737-4038/95/1201-0015$02.00

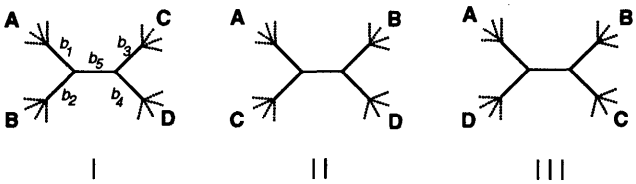

FIG. 1.—Three unrooted trees for four monophyletic clusters of species A, B, C, and D.

$$\beta = \begin{cases} -\dfrac{n_A n_C + n_B n_D}{(n_A + n_B)(n_C + n_D)}, \\ \text{if } X = \text{I and } Y = \text{II}, \\ \dfrac{(n_A - n_C)(n_D - n_B)(n_A n_C + n_B n_D)}{(n_A + n_B)(n_C + n_D)(n_A + n_D)(n_B + n_C)}, \\ \text{if } X = \text{I and } Y = \text{III}, \\ \dfrac{n_A n_C + n_B n_D}{(n_A + n_D)(n_B + n_C)}, \\ \text{if } X = \text{II and } Y = \text{III}. \end{cases} \quad (4)$$

$$\gamma = \begin{cases} \dfrac{(n_A - n_D)(n_C - n_B)(n_A n_D + n_B n_C)}{(n_A + n_B)(n_C + n_D)(n_A + n_C)(n_B + n_D)}, \\ \text{if } X = \text{I and } Y = \text{II}, \\ -\dfrac{n_A n_D + n_B n_C}{(n_A + n_B)(n_C + n_D)}, \\ \text{if } X = \text{I and } Y = \text{III}, \\ -\dfrac{n_A n_D + n_B n_C}{(n_A + n_C)(n_B + n_D)}, \\ \text{if } X = \text{II and } Y = \text{III}. \end{cases} \quad (5)$$

The variance of $S_X - S_Y$ is computed by the standard formula of the variance of a linear combination of random variables:

$$V(S_X - S_Y) = \sum_{i<j} w_{ij}^2 V(d_{ij}) + 2 \sum_{ij<kl} w_{ij} w_{kl} \mathrm{Cov}(d_{ij}, d_{kl}). \quad (6)$$

Once we obtain $S_X - S_Y$ and its variance, we can compute the test statistic

$$Z = (S_X - S_Y)/\sqrt{V(S_X - S_Y)}. \quad (7)$$

Our computer simulation has shown that the distribution of $S_X - S_Y$ is approximately normal for a wide variety of distance measures for nucleotide sequences. Therefore, we can apply a two-tailed normal deviate test to examine the statistical significance of the difference in S between trees X and Y. Note that the level of statistical significance generally increases as the number of species used increases.

One can also test the composite null hypothesis $E(S_{\mathrm{I}}) = E(S_{\mathrm{II}}) = E(S_{\mathrm{III}})$ in the following way, where $E(S_X)$ stands for the expected value of S_X. Under the null hypothesis, random variables $\delta_1 = S_{\mathrm{II}} - S_{\mathrm{I}}$ and $\delta_2 = S_{\mathrm{III}} - S_{\mathrm{I}}$ approximately follow a bivariate normal distribution with expected values $E(\delta_1) = E(\delta_2) = 0$ and covariance matrix

$$V = \begin{bmatrix} V(\delta_1) & \mathrm{Cov}(\delta_1, \delta_2) \\ \mathrm{Cov}(\delta_1, \delta_2) & V(\delta_2) \end{bmatrix}.$$

Then the ellipse containing $100\alpha\%$ of the joint distribution of δ_1 and δ_2 is defined by following equation (Johnson and Kotz 1972, p. 87).

$$\frac{\delta_1^2}{V(\delta_1)} - \frac{2\rho\delta_1\delta_2}{\sqrt{V(\delta_1)V(\delta_2)}} + \frac{\delta_2^2}{V(\delta_2)} = -2(1 - \rho^2)\log(1 - \alpha), \quad (8)$$

where $\rho = \mathrm{Cov}(\delta_1, \delta_2)/\sqrt{V(\delta_1)V(\delta_2)}$, and $\mathrm{Cov}(\delta_1, \delta_2)$ is computed by

$$\mathrm{Cov}(\delta_1, \delta_2) = \sum_{i<j} w_{1,ij} w_{2,ij} V(d_{ij}) + \sum_{ij \neq kl} w_{1,ij} w_{2,kl} \mathrm{Cov}(d_{ij}, d_{kl}). \quad (9)$$

Coefficients $w_{1,ij}$ and $w_{2,ij}$ are calculated according to equations (2)–(5) for $\delta_1 = S_{\mathrm{II}} - S_{\mathrm{I}}$ and $\delta_2 = S_{\mathrm{III}} - S_{\mathrm{I}}$, respectively.

Once $V(\delta_1)$, $V(\delta_2)$, and $\mathrm{Cov}(\delta_1, \delta_2)$ are obtained, we can compute the proportion (p) of the bivariate normal distribution contained by the ellipse for specific values of δ_1 and δ_2 by the following equation.

$$p = 1 - \exp\left\{-\left[\frac{\delta_1^2}{V(\delta_1)} - \frac{2\rho\delta_1\delta_2}{\sqrt{V(\delta_1)V(\delta_2)}} + \frac{\delta_2^2}{V(\delta_2)}\right] \Big/ [2(1 - \rho^2)]\right\}. \quad (10)$$

If p is small, the data do not provide sufficient information to distinguish among trees I, II, and III (fig. 1). If p is large, say >0.95, the null hypothesis $E(S_{\mathrm{I}}) = E(S_{\mathrm{II}}) = E(S_{\mathrm{III}})$ is rejected at the 5% level. Note that value of p in equation (10) does not depend on the choice of δ's. That is, p remains unchanged if we replace $S_{\mathrm{II}} - S_{\mathrm{I}}$ and $S_{\mathrm{III}} - S_{\mathrm{I}}$ (and their corresponding variances and covariances) by either $S_{\mathrm{II}} - S_{\mathrm{I}}$ and $S_{\mathrm{III}} - S_{\mathrm{II}}$ or $S_{\mathrm{III}} - S_{\mathrm{I}}$ and $S_{\mathrm{III}} - S_{\mathrm{II}}$.

It should be emphasized that a high value of p does not necessarily indicate that one of three phylogenies in figure 1 is significantly "better" than the other two. Indeed, one can encounter situations where the values of S_{I} and S_{II} are not significantly different from each other and S_{III} is significantly greater than both S_{I} and S_{II}. In this case p will be large because the null hypothesis $E(S_{\mathrm{I}}) = E(S_{\mathrm{II}}) = E(S_{\mathrm{III}})$ is rejected. However, one can only conclude that tree III is unlikely to be the true tree.

In this four-cluster analysis, the computation of the difference in S between two trees does not require any information regarding the branching order within each cluster. This is because the estimate of a branch length obtained by the ordinary least-squares method depends only on (1) the numbers of species within the four clusters associated with this branch and (2) the intercluster distances among the species (see eq. [2] in Rzhetsky and Nei [1993]) as the estimates of branch lengths within clusters A, B, C, and D give the same contribution to S_{I}, S_{II}, and S_{III}. Therefore, to compute the difference between any pair of S values, we need to estimate the lengths of only five branches of each four-cluster tree (e.g., b_1, b_2, b_3, b_4, and b_5 for tree I in fig. 1).

To simplify the explanation of our method, let us assume that tree I is the true tree. This means that both $S_{\mathrm{I}} - S_{\mathrm{II}}$ and $S_{\mathrm{I}} - S_{\mathrm{III}}$ are expected to be negative, whereas $S_{\mathrm{II}} - S_{\mathrm{III}}$ can be either negative or positive. That is, we have

$$S_{\mathrm{I}} - S_{\mathrm{II}} = -\frac{n_{\mathrm{A}}n_{\mathrm{B}} + n_{\mathrm{C}}n_{\mathrm{D}}}{(n_{\mathrm{A}} + n_{\mathrm{C}})(n_{\mathrm{B}} + n_{\mathrm{D}})}\, b_5 + e_{\mathrm{I,II}},$$

$$S_{\mathrm{I}} - S_{\mathrm{III}} = -\frac{n_{\mathrm{A}}n_{\mathrm{B}} + n_{\mathrm{C}}n_{\mathrm{D}}}{(n_{\mathrm{A}} + n_{\mathrm{D}})(n_{\mathrm{B}} + n_{\mathrm{C}})}\, b_5 + e_{\mathrm{I,II}},$$

$$S_{\mathrm{II}} - S_{\mathrm{III}} = \frac{(n_{\mathrm{A}} - n_{\mathrm{B}})(n_{\mathrm{D}} - n_{\mathrm{C}})(n_{\mathrm{A}}n_{\mathrm{B}} + n_{\mathrm{C}}n_{\mathrm{D}})}{(n_{\mathrm{A}} + n_{\mathrm{C}})(n_{\mathrm{A}} + n_{\mathrm{D}})(n_{\mathrm{B}} + n_{\mathrm{C}})(n_{\mathrm{B}} + n_{\mathrm{D}})} \times b_5 + e_{\mathrm{II,III}}, \tag{11}$$

where e_{XY} is a linear combination of random errors introduced in the process of estimating evolutionary distances, and b_5 is the expected length of the interior branch of tree I in figure 1. If both $S_{\mathrm{I}} - S_{\mathrm{II}}$ and $S_{\mathrm{I}} - S_{\mathrm{III}}$ are significantly smaller than zero (say, at the 5% level), then tree I is likely to be the true tree.

The four-cluster analysis and the test of interior branch length (Nei et al. 1985; Li 1989; Rzhetsky and Nei 1992, 1993) usually give correlated but not identical results. This is because in the four-cluster analysis we consider differences in sums of estimates for five branches, rather than just one as in the interior-branch test. Further, the interior-branch test is usually applied to a single topology rather than to several alternative topologies. If the topology for interior branch test is not selected a priori but instead chosen to maximize the estimated interior-branch lengths, the interior-branch test may give liberal results (Sitnikova et al. 1995). The four-cluster analysis does not suffer from this problem

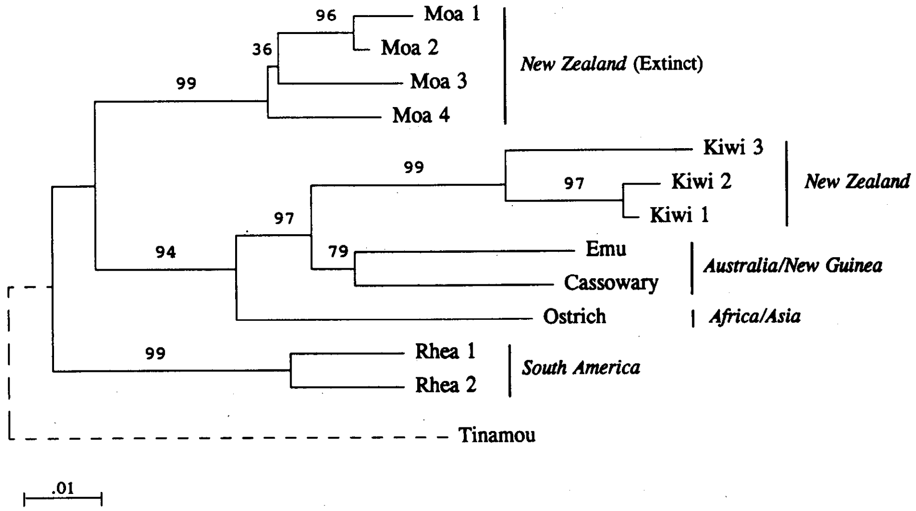

FIG. 2.—The minimum-evolution tree (Rzhetsky and Nei 1993) for 12 ratite birds. Jukes and Cantor's (1969) distances were used. This tree is identical in topology to the one presented by Cooper et al. (1992). The number given to each interior branch is the confidence probability obtained by Rzhetsky and Nei's (1992, 1993) method. The root of the tree was given by using tinamou as the outgroup.

Table 1
Results of Four-Cluster Analyses for Determining the Closest Relative of Moas Using Mitochondrial 12S rRNA Sequences

Best Tree	Statistical Confidence	100(1 − *p*)%
a. [(K, E), (M, R)] . . .	Better than [(K, M), (R, E)] at 0.04% level, Better than [(K, R), (M, E)] at 3.58% level	0.09
b. [(K, C), (M, R)] . . .	Better than [(K, M), (R, C)] at 0.06% level, Better than [(K, R), (M, C)] at 2.04% level	0.22
c. [(K, O), (M, R)] . . .	Better than [(K, M), (R, O)] at 3.44% level, Better than [(K, R), (M, O)] at 28.92% level	12.55
d. [(A, O), (M, R)] . . .	Better than [(A, R), (M, O)] at 10.32% level, Better than [(A, M), (R, O)] at 4.44% level	11.26

NOTE.—In the analysis we used 12S rRNA sequences from four moas (M), three kiwis (K), two rheas (R), one emu (E), one cassowary (C), and one ostrich (O); (A) denotes the kiwis, emu, and cassowary group. Jukes and Cantor's (1969) distance was used in these analyses. The last column shows the value of 100(1 − *p*)% of each of the four-cluster analyses (see eq. [10]).

because all possible arrangements of four groups of species are examined (see example).

The four-cluster analysis has advantages over the currently available methods in simplicity, computational speed (there is no need to examine all possible tree topologies), and universality (it is easy to use any estimator of evolutionary distance). It also can easily be extended to five or more groups, though the number of comparisons of S values required increases.

Example

As an example, let us consider the evolutionary relationships of a group of extinct flightless birds (moas) from New Zealand with other ratite birds that currently live in New Zealand (kiwis). Because of their geographic proximity and morphological similarity moas were originally thought to share a most recent common ancestor with kiwis rather than with ratite birds from other parts of the world. However, analyzing mitochondrial 12S rRNA gene sequences, Cooper et al. (1992) concluded that kiwis are more closely related to other ratites emu and cassowary (Australia and New Guinea) and ostrich (Africa and Asia) than to extinct moas. Their conclusion is based on the confidence interval test of the interior branch of the maximum likelihood tree obtained. However, since the confidence interval test is known to be too liberal under certain circumstances (Tateno et al. 1994), it is desirable to test this hypothesis by another statistical method. We therefore constructed the minimum-evolution tree (fig. 2) and applied our four-cluster analysis to test the null hypothesis that extinct New Zealand moas are closer to kiwis than to emu, cassowary, ostrich, or rhea (South America). In this test, we assumed that the moas, kiwis, and rheas each constitute a monophyletic group and used emu, cassowary, or ostrich as the fourth group. These results (table 1*a*–*c*) show that in all three tests the null hypothesis is rejected because the cluster of kiwis and moas is significantly worse than the cluster of kiwis with emu, cassowary, or ostrich ($P < 5\%$). This is consistent with the distant relationship of moas and kiwis in our minimum-evolution tree and the conclusion reached by Cooper et al. (1992).

Furthermore, Cooper et al. suggested that kiwis, emus, and cassowaries share a most recent common ancestor with ostrich, not with moas. This hypothesis can be tested by using moas, rheas, the kiwi-emu-cassowary group, and ostrich as the four monophyletic groups and conducting the four cluster analysis (table 1*d*). Our test does not clearly establish Cooper et al.'s view that kiwi, emu, cassowary, and ostrich shared a most recent common ancestor, even though the interior branch leading to this group seems to be long (interior branch test confidence = 94%). The cluster of ostrich with the kiwi-emu-cassowary group is significantly better than the ostrich-rhea cluster ($P < 5\%$), but the hypothesis that ostrich is closer to the kiwi-emu-cassowary group is not significantly better than the second alternative hypothesis in which ostrich is closer to moas ($P > 10\%$). In general, this dataset does not seem to resolve the position of ostrich ($1 - p > 0.11$).

Acknowledgments

We are grateful to two anonymous referees for helpful comments and checking equations. We also thank Naoyuki Takahata for helpful suggestions. This work was supported by grants from the National Science Foundation and the National Institutes of Health to M.N.

LITERATURE CITED

COOPER, A., C. MOURER-CHAUVIRÉ, G. K. CHAMBERS, A. VON HAESELER, A. C. WILSON, and S. PÄÄBO. 1992. Independent origins of New Zealand moas and kiwis. Proc. Natl. Acad. Sci. USA **89**:8741–8744.

JOHNSON, N. L., and S. KOTZ. 1972. Distributions in statistics: continuous multivariate distributions. Wiley, New York.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. M. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

LI, W.-H. 1989. A statistical test of phylogenies estimated from sequence data. Mol. Biol. Evol. **6**:424–435.

NEI, M., C. STEPHENS, and N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

RZHETSKY, A., and M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9**:945–967.

———. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. **10**:1073–1095.

SITNIKOVA, T., A. RZHETSKY, and M. NEI. 1995. Interior-branch and bootstrap tests of phylogenetic trees. Mol. Biol. Evol. (accepted).

TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

NAOYUKI TAKAHATA, reviewing editor

Received April 19, 1994

Accepted July 28, 1994