Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data

Tae-Kun Seo[†‡§], Hirohisa Kishino[‡¶], and Jeffrey L. Thorne[†]

[†]Bioinformatics Research Center, North Carolina State University, Box 7566, Raleigh, NC 27695-7566; and [‡]Professional Programme for Agricultural Bioinformatics and [¶]Laboratory of Biometrics, Graduate School of Agriculture and Life Sciences, University of Tokyo, 1-1-1 Yayoi Bunkyo-Ku, Tokyo 113-8657, Japan

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved February 3, 2005 (received for review November 8, 2004)

Because of the increase of genomic data, multiple genes are often available for the inference of phylogenetic relationships. The simple approach for combining multiple genes from the same taxon is to concatenate the sequences and then ignore the fact that different positions in the concatenated sequence came from different genes. Here, we discuss two criteria for inferring the optimal tree topology from data sets with multiple genes. These criteria are designed for multigene data sets where gene-specific evolutionary features are too important to ignore. One criterion is conventional and is obtained by taking the sum of log-likelihoods over all genes. The other criterion is obtained by dividing the log-likelihood for a gene by its sequence length and then taking the arithmetic mean over genes of these ratios. A similar strategy could be adopted with parsimony scores. The optimal tree is then declared to be the one for which the sum or the arithmetic mean is maximized. These criteria are justified within a two-stage hierarchical framework. The first level of the hierarchy represents gene-specific evolutionary features, and the second represents site-specific features for given genes. For testing significance of the optimal topology, we suggest a two-stage bootstrap procedure that involves resampling genes and then resampling alignment columns within resampled genes. An advantage of this procedure over concatenation is that it can effectively account for gene-specific evolutionary features. We discuss the applicability of the two-stage bootstrap idea to the Kishino–Hasegawa test and the Shimodaira–Hasegawa test.

Phylogenetic relationships can be estimated by a frequentist, Bayesian, or parsimony framework (for a general overview, see ref. 1). Within the frequentist framework, available procedures include maximum likelihood and distance methods. The bootstrap procedure (e.g., refs. 2–4) is widely employed to calculate the significance of a topology that is obtained with the frequentist or parsimony approaches. This procedure assumes that all columns of sequence data are samples from an independent and identical distribution, which we will refer to as the "*iid* assumption." Although site dependency seems to be a more realistic assumption for biological sequence evolution (e.g., refs. 5–8), the *iid* assumption is widely employed because it saves computation time and convenient asymptotic theories of statistics can then be applied.

Because of the increase of genomic data, phylogeny estimation and testing can be based on multiple genes (e.g., refs. 9–12). A simple approach is to concatenate multiple genes, estimate the true tree, and then apply a bootstrap procedure to test whether the estimated topology is significantly better than alternatives. Concatenation may violate the *iid* assumption. Because different genes may have been subject to different evolutionary pressures, it may be appropriate to describe each gene by its own parameter set. The set may include parameters for nucleotide frequencies, branch lengths, etc. This sort of separate analysis of genes may be preferable to concatenation (refs. 9–11 and 13–15, but see ref. 16).

At the extreme, different genes may even support different tree topologies. Permutation tests (17–19) are available to detect departures among genes in "homogeneity" of tree support. However, these tests simply examine whether genes support congruent trees and whether we can ignore gene-specific effects via concatenation. These tests fail to give further direction about combining the separate information when multiple genes support different trees.

To determine the optimal topology even when different genes support different topologies, we should consider what topology selection criterion is most appropriate. The usual sum of parsimony or log-likelihood scores over all columns of separately analyzed multiple genes can be the basis for selecting an optimal tree (e.g., refs. 9–11), and we will refer to this as the "sum criterion." The sum criterion is superior to sequence concatenation because it is affected by gene-specific evolutionary features. When the sum criterion is used to select optimal tree topologies, it is critical to determine how to test the significance of the selected trees. A simple application of the bootstrap procedure merges across genes the pool of sitewise log-likelihood or parsimony scores and then samples sitewise scores from this merged pool with replacement. The simple one-stage procedure may be ill-advised because the variation among genes is not properly considered and because the *iid* assumption among scores within the merged pool is no longer valid. Here, we apply a two-stage bootstrap procedure (20) to avoid problems associated with the one-stage version.

Another criterion to select tree topologies is the "average criterion." If different sequences have different lengths, the sum of scores might fail to be a good criterion because long sequences may have a big impact on the sum of scores and the resulting optimal tree will reflect mainly the tree supported by long sequences and their gene-specific features. This sum of scores criterion therefore might be sensitive to long genes that have experienced unusual evolutionary processes. To remove the effect of sequence length, one can assign each gene an average score per site and then average these averages. In a similar way to the sum criterion, the two-stage bootstrap procedure can then be employed.

For the justification of our two-stage bootstrap procedure, we assume a hierarchical structure of multigene data set generation. First, evolutionary characteristics of each gene are determined. These evolutionary features are summarized by parameters representing nucleotide frequencies, branch lengths, etc. The parameter sets for different genes are independent samples from some common distribution of parameter sets. According to this hierarchical framework, each gene can be visualized as being determined by a common mechanism with perturbation. Because of the perturbation, genes do not all have identical properties. Instead, their properties follow an unknown true distribution. Second, once the properties of each gene are determined, they are then expressed through the sequence columns. These columns are distributed

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: KLD, Kullback–Leibler distance; KH, Kishino–Hasegawa; SH, Shimodaira–Hasegawa; WSH, weighted SH.

[§]To whom correspondence should be sent at the ‡ address. E-mail: seo@iu.a.u-tokyo.ac.jp

around the evolutionary features specific to a particular gene. The degree of dispersion of sequence columns around the features of a gene can vary among different genes.

Although it is possible that the evolutionary features of a gene and its length are correlated, we assume that this is not the case. Therefore, our two-stage *iid* assumption leads to a two-stage bootstrap procedure (20) for testing the significance of a tree. In this context, the first stage of the two-stage bootstrap procedure is to sample genes. The second stage samples columns of sequence data with replacement within resampled genes. Although sequence lengths are generally long, the number of available genes may be relatively small. Thus, the expected value of the estimated variance from the two-stage bootstrap procedure may be seriously different from the true variance. This bias should be considered when testing hypotheses. Under the two-stage *iid* assumption, it is possible to analytically calculate the variance of bootstrap resamples. We show how to do this for testing the significance of a topology.

The two-stage hierarchical structure is similar to a random effects model (21). One main purpose of the conventional random effects model is to test for a difference between groups. In our case, groups correspond to genes. A normality assumption is needed for ease of testing. The purpose of our procedure is to test whether the mean of the random effects model is equal to zero. We note that this test can be done without an assumption about the type of distribution by employing a nonparametric bootstrap procedure.

Methods

Here, we show how the hierarchial structure of genes and sequence columns can be used in determining an optimal tree topology and then testing its significance in a likelihood context. Kullback–Leibler distance (KLD) (22) is a distance measure between two models. In our case, these models correspond to tree topologies. We show that our hierarchical assumption of genes and columns can be directly extended to the KLD measure.

KLD. For the ith gene, KLD_i between the true but unknown data-generating mechanism $f(\cdot)$ and the model $g(\cdot|\theta)$ is:

$$KLD_i[f, g] := E_f\left\{\log \frac{f(\mathbf{X})}{g(\mathbf{X}|\theta)}\right\} \approx \frac{1}{n_i}\sum_{j=1}^{n_i}\{\log f(x_{ij}) - \log g(x_{ij}|\theta)\}, \quad [1]$$

where the x_{ij} values are sampled sequence columns from $f(\cdot)$ and n_i is the length of gene i. Each term of $\{\log f(x_{ij}) - \log g(x_{ij}|\theta)\}$ is a sitewise KLD_i. The estimate of the minimum KLD_i can be obtained with the maximum of $\Sigma_{j=1}^{n_i}\log g(x_{ij}|\theta)$, which is denoted $mKLD_i$ here. That is, we obtain $mKLD_i$ at maximum likelihood estimates ($\hat{\theta}$) under the model $g(\cdot|\theta)$,

$$mKLD_i[f, g] = \frac{1}{n_i}\sum_{j=1}^{n_i}\{\log f(x_{ij}) - \log g(x_{ij}|\hat{\theta})\}.$$

The $mKLD_i$ is used to choose the model (topology) which is closest to the unknown true model. Because we do not know the true data-generating mechanism $f(\cdot)$, we cannot directly calculate $mKLD_i$. However, the $\log f(x_{ij})$ terms are common to all competing models and $mKLD_i$ differences can be used in model comparison. Suppose two models (topologies) g_1 and g_2 are compared, then

$$mKLD_i[f, g_1] - mKLD_i[f, g_2] = \frac{1}{n_i}\sum_{j=1}^{n_i}\{\log g_2(x_{ij}|\hat{\theta}_2) - \log g_1(x_{ij}|\hat{\theta}_1)\}, \quad [2]$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are maximum likelihood estimates under the model $g_1(\cdot|\theta_1)$ and $g_2(\cdot|\theta_2)$. The $\{\log g_2(x_{ij}|\hat{\theta}_2) - \log g_1(x_{ij}|\hat{\theta}_1)\}$ term is a sitewise $mKLD_i$ difference and is denoted $mKLD_{ij}^{g_2g_1}$. If the quantity of Eq. **2** is significantly larger (smaller) than zero, model g_2 (g_1) is termed closer to the true model. The variance of $mKLD_{ij}^{g_2g_1}$ is considered in testing significance. The log $g_1(x_{ij}|\hat{\theta}_1)$ and log $g_2(x_{ij}|\hat{\theta}_2)$ terms are not exactly independent among different sites j ($j = 1, \ldots, n_i$) because $\hat{\theta}_1$ and $\hat{\theta}_2$ are functions of all x_{ij} values. However, for large n_i, $\hat{\theta}_1$ and $\hat{\theta}_2$ are close to the true values θ_1 and θ_2, and these true values minimize KLD between $f(\cdot)$ and $g_1(\cdot|\theta_1)$, and between $f(\cdot)$ and $g_2(\cdot|\theta_2)$. Thus, it is almost correct to regard $\log g_1(x_{ij}|\hat{\theta}_1)$ and $\log g_2(x_{ij}|\hat{\theta}_2)$ as functions of solely x_{ij}. Because x_{ij} has a hierarchical *iid* structure, functions of x_{ij} also have a hierarchical *iid* structure.

Testing Phylogeny in a Likelihood Context. Suppose we want to know whether tree topology a or b is closer to the truth. If the estimated difference of $mKLD_i$ between the two tree topologies deviates significantly from zero, it means that one of the two trees is significantly closer to the unknown true distribution than the other [Kishino–Hasegawa (KH) test, ref. 3]. If one of a and b is the optimal tree as estimated by maximum likelihood or parsimony rather than a tree that was of interest *a priori*, the uncertainty of the estimate of the optimal tree should be considered in obtaining a confidence set of trees [Shimodaira–Hasegawa (SH) test, ref. 4]. Here, we show how to apply the hierarchical structure of $mKLD_{ij}^{ab}$ to the KH and SH tests.

For the maximum likelihood estimates of parameters, let the log-likelihood values calculated at the jth column of the ith gene under trees a and b be $l_{a,ij}$ and $l_{b,ij}$ respectively. Let the sitewise log-likelihood difference be $y_{ab,ij}$ ($:= l_{a,ij} - l_{b,ij} = mKLD_{ij}^{ab}$), which is the sitewise difference of mKLD.

We consider the distribution of sitewise differences of mKLD. That is, for K genes, we assume that $y_{ab,ij}$($i = 1, \ldots, K$; $j = 1, \ldots, n_i$) is an observation of random variable $Y_{ab,i}$ with the properties $E(Y_{ab,i}) = w_{ab,i} < \infty$ and $\mathrm{Var}(Y_{ab,i}) = \sigma_{ab,i}^2 < \infty$. Except for a finite mean and variance, little is assumed about the type of the distribution. The $\sigma_{ab,i}^2$ values need not be the same among genes i, but we do require independence between $w_{ab,i}$ and $\sigma_{ab,i}^2$. The expected value of $\sigma_{ab,i}^2$ is denoted $\sigma_{ab,I}^2$. We consider the hierarchical structure where $w_{ab,i}$ is an observation of random variable W_{ab} with the properties $E(W_{ab}) = \mu_{ab} < \infty$ and $\mathrm{Var}(W_{ab}) = \sigma_{ab,W}^2 < \infty$. We assume that gene lengths n_i are random variables that are independent of $w_{ab,i}$ and $\sigma_{ab,i}^2$ with $E(n_i) = n$ ($0 < n < \infty$) and $\mathrm{Var}(n_i) = \sigma_n^2 < \infty$. The hierarchical structure in our approach is similar to a random effects model. Although a conventional goal with the random effects model would be to test whether $\sigma_{ab,W}^2$ exceeds zero, we are more interested here in testing whether μ_{ab} is zero. For our application, normality assumptions for $Y_{ab,i}$ and W_{ab} are not required.

KH test for multiple genes. Results pertaining to the KH test for multiple genes are below and are justified in the supporting information, which is published on the PNAS web site. We consider the testing problem, $H_0 : \mu_{ab} = 0$ versus $H_1 : \mu_{ab} \neq 0$. If μ_{ab} is greater (less) than zero, then tree a (tree b) can be regarded as more reliable. Define S_{ab} as the sum of sitewise log-likelihood differences between tree a and b over all columns of all genes,

$$S_{ab} := \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ab,ij}. \quad [3]$$

This means

$$E(S_{ab}) = Kn\mu_{ab}$$
$$\mathrm{Var}(S_{ab}) = Kn\sigma^2_{ab,I} + K(n^2 + \sigma_n^2)\sigma^2_{ab,W} + K\sigma_n^2\mu^2_{ab}. \quad [4]$$

If S_{ab} is far from zero, then H_0 is rejected. In general, n^2 and σ_n^2 are big. Because $\sigma^2_{ab,W}$ is multiplied by $K(n^2 + \sigma_n^2)$, the among-gene variation ($\sigma^2_{ab,W}$) can have a big impact on the variance of S_{ab}.

To approximate the sampling distribution of S_{ab} under H_0, we use two-stage bootstrap resampling. First, we resample genes (denoted by $*$) and second, we resample columns of the resampled genes (denoted by $\star$). Because it is computationally intensive to maximize likelihoods for resampled data, we employ the idea of the RELL method (3). That is, we consider two-stage resampling from the set of likelihood values that are already calculated instead of resampling gene and sequence columns. Let the resampled log-likelihood value at the jth column of the ith gene under trees a and b be $l^{*\star}_{a,ij}$ and $l^{*\star}_{b,ij}$. Define

$$S^{*\star}_{ab} := \sum_{i=1}^{K}\sum_{j=1}^{n_i^*} y^{*\star}_{ab,ij},$$

where $y^{*\star}_{ab,ij} := l^{*\star}_{a,ij} - l^{*\star}_{b,ij}$ and n_i^* is the length of resampled ith gene. We have

$$E_{*\star}(S^{*\star}_{ab}|\mathbf{y}_{ab,1}, \ldots, \mathbf{y}_{ab,K}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ab,ij} = S_{ab} \quad [5]$$

and

$$\mathrm{Var}_{*\star}(S^{*\star}_{ab}|\mathbf{y}_{ab,1}, \ldots, \mathbf{y}_{ab,K}) = \sum_{i=1}^{K}\left\{\sum_{j=1}^{n_i} y^2_{ab,ij} + \frac{n_i - 1}{n_i}\left(\sum_{j=1}^{n_i} y_{ab,ij}\right)^2\right\} - \frac{1}{K}\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ab,ij}\right)^2, \quad [6]$$

where $\mathbf{y}_{ab,i}$ is $(y_{ab,i1}, \ldots, y_{ab,in_i})^T$. The expected values of Eq. **5** and **6** are

$$E_{\mathbf{Y}}[E_{*\star}(S^{*\star}_{ab}|\mathbf{y}_{ab,1}, \ldots, \mathbf{y}_{ab,K})] = Kn\mu_{ab},$$

and

$$E_{\mathbf{Y}}[\mathrm{Var}_{*\star}(S^{*\star}_{ab}|\mathbf{y}_{ab,1}, \ldots, \mathbf{y}_{ab,K})] = \mathrm{Var}(S_{ab}) + \{(nK - n - K)\sigma^2_{ab,I} - (\sigma_n^2 + n^2)\sigma^2_{ab,W} - \sigma_n^2\mu^2_{ab}\}. \quad [7]$$

Eq. **7** means that we should consider the bias in the estimation of $\mathrm{Var}(S_{ab})$ by $\mathrm{Var}(S^{*\star}_{ab})$. This bias is represented by the second term of Eq. **7**. Also in testing H_0, this bias should be considered.

Following two guidelines of the bootstrap method by Hall and Wilson (23), we approximate the distribution of $(S_{ab} - Kn\mu_{ab})/\hat{\sigma}_{ab}$ with the distribution of $(S^{*\star}_{ab} - S_{ab})/\hat{\sigma}^{*\star}_{ab}$, where $\hat{\sigma}_{ab}$ and $\hat{\sigma}^{*\star}_{ab}$ are the square roots of the unbiased estimators of $\mathrm{Var}(S_{ab})$ and $\mathrm{Var}(S^{*\star}_{ab}|\mathbf{y}_{ab,1}\cdots, \mathbf{y}_{ab,K})$. We have

$$\hat{\sigma}^2_{ab} = \frac{K}{K-1}\sum_{i=1}^{K}\left\{\sum_{j=1}^{n_i} y_{ab,ij} - \frac{1}{K}\sum_{r=1}^{K}\sum_{j=1}^{n_r} y_{ab,rj}\right\}^2, \quad \text{and} \quad [8]$$

$$\hat{\sigma}^{2*\star}_{ab} = \frac{K}{K-1}\sum_{i=1}^{K}\left\{\sum_{j=1}^{n_i^*} y^{*\star}_{ab,ij} - \frac{1}{K}\sum_{r=1}^{K}\sum_{j=1}^{n_r^*} y^{*\star}_{ab,rj}\right\}^2. \quad [9]$$

If $(S_{ab} - 0)/\hat{\sigma}_{ab}$ is outside the 95% interval of $(S^{*\star}_{ab} - S_{ab})/\hat{\sigma}^{*\star}_{ab}$, then we reject the null hypothesis $H_0 : \mu_{ab} = 0$. When gene number is large and sequences are long, $(S_{ab} - 0)/\hat{\sigma}_{ab}$ asymptotically follows a standard normal distribution when H_0 is true and the test can be performed without a bootstrap procedure.

SH test for multiple genes. In many cases, we want to test significance between an optimal and another tree, not between two trees that are both selected prior to the analysis. The SH test is designed for this situation because it considers the uncertainty of choosing the optimal tree (ref. 4; see also ref. 24). Often, the optimal tree is selected from a set of candidates, and the goal is to test whether the optimal tree is significantly better than the others. The candidates that are not significantly worse than the optimal tree are used to construct the confidence set of trees.

The definition of S_{ab} in Eq. **3** can be rewritten

$$S_{ab} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ab,ij} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} l_{a,ij} - \sum_{i=1}^{K}\sum_{j=1}^{n_i} l_{b,ij}. \quad [10]$$

Eq. **10** suggests the sum criterion to determine the optimal tree from multiple genes. Suppose there are p candidate trees, one of which is the optimal tree $\hat{\tau}$. In our method,

$$\hat{\tau} = \underset{\tau\in\{1,\ldots,p\}}{\mathrm{argmax}}\left\{\sum_{i=1}^{K}\sum_{j=1}^{n_i} l_{\tau,ij}\right\}. \quad [11]$$

To apply the two-stage bootstrap to the SH test, let H_{ab} and $H^{*\star}_{ab}$ be $(S_{ab} - Kn\mu_{ab})/\hat{\sigma}_{ab}$ and $(S^{*\star}_{ab} - S_{ab})/\hat{\sigma}^{*\star}_{ab}$, where $a, b = 1, \ldots, p$. If $a = b$, H_{ab} and $H^{*\star}_{ab}$ are zero. Under the least favorable configuration (4) in which the expected log-likelihood sum is the same over all trees, the hypothesized value of μ_{ab} is equal to zero, and this makes $H_{ab} = S_{ab}/\hat{\sigma}_{ab}$. Following the steps below, we can construct a confidence set of trees.

1. For each tree $\tau(\tau \in \{1, \ldots, p\})$, calculate the test statistic $T_\tau = H_{\hat{\tau}\tau}$.
2. Generate q sets of two-stage resampled log-likelihood values $l^{*\star(r)}_{\tau,ij}(r = 1, \ldots, q;\ \tau = 1, \ldots, p)$.
3. Calculate $H^{*\star(r)}_{ab}$ from $l^{*\star(r)}_{a,ij}$ and $l^{*\star(r)}_{b,ij}$, for all $a, b = 1, \ldots, p$.
4. For each tree τ with each resampled data set r $(r = 1, \ldots, q)$, calculate $T^{*\star(r)}_\tau = H^{*\star(r)}_{\hat{\tau}^{**}\tau}$, where $\hat{\tau}^{**}$ of $H^{*\star(r)}_{\hat{\tau}^{**}\tau}$ is estimated with resampled data as follows

$$\hat{\tau}^{**} = \underset{\tau\in\{1,\ldots,p\}}{\mathrm{argmax}}\left\{\sum_{i=1}^{K}\sum_{j=1}^{n_i^*} l^{*\star}_{\tau,ij} - \sum_{i=1}^{K}\sum_{j=1}^{n_i} l_{\tau,ij}\right\}. \quad [12]$$

5. For each τ $(\tau \in \{1, \ldots, p\})$, if T_τ is not in the critical region of the distribution of $T^{*\star}_\tau$ with level α, τ is included in the confidence set of trees.

The Average Criterion to Select Tree Topology. Above, we considered the sum criterion to select the tree topology. As noted, the sum of log-likelihood values over all genes can be sensitive to relatively long genes that happened to experience atypical evolutionary processes. To remove the effect of sequence length, we consider an alternative to the sum criterion that we refer to as the average criterion.

If we define $\hat{\mu}_{ab}$ and $\hat{\mu}^{*\star}_{ab}$ as

$$\hat{\mu}_{ab} := \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ab,ij}, \quad \text{and} \quad \hat{\mu}^{*\star}_{ab} := \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n^*_i}\sum_{j=1}^{n^*_i} y^{*\star}_{ab,ij},$$

then $E(\hat{\mu}_{ab}) = \mu_{ab}$ and $E(\hat{\mu}^{*\star}_{ab}) = \hat{\mu}_{ab}$. Letting $n^{\ominus}$ and $n^{\otimes}$ be the expected values of $1/n_i$ and $1/n_i^2$, the variance of $\hat{\mu}_{ab}$ and $\hat{\mu}^{*\star}_{ab}$ are

$$\mathrm{Var}(\hat{\mu}_{ab}) = \frac{1}{K}\{\sigma^2_{ab,I} n^{\ominus} + \sigma^2_{ab,W}\}, \quad \text{and}$$

$$\mathrm{Var}_{*\star}(\hat{\mu}^{*\star}_{ab}) = \frac{1}{K^2}\sum_{i=1}^{K}\left\{\frac{1}{n_i^2}\left(\sum_{j=1}^{n_i} y^2_{ab,ij} + n_i(n_i-1)\cdot\left(\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ab,ij}\right)^2\right)\right\} - \frac{1}{K}\left\{\frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ab,ij}\right\}^2$$

and the expectation of $\mathrm{Var}_{*\star}(\hat{\mu}^{*\star}_{ab})$ is

$$E_{\mathbf{Y}}[\mathrm{Var}_{*\star}(\hat{\mu}^{*\star}_{ab})] = \mathrm{Var}(\hat{\mu}_{ab}) + \frac{1}{K}\left\{\sigma^2_{ab,I}\left(n^{\ominus} - \frac{n^{\ominus}}{K} - n^{\otimes}\right) - \frac{1}{K}\sigma^2_{ab,W}\right\}.$$

In a similar way to Eq. **7**, we should consider the bias in the estimation of $\mathrm{Var}(\hat{\mu}_{ab})$ when using $\mathrm{Var}_{*\star}(\hat{\mu}^{*\star}_{ab})$. The unbiased estimators of $\mathrm{Var}(\hat{\mu}_{ab})$ and $\mathrm{Var}(\hat{\mu}^{*\star}_{ab})$ are, respectively,

$$\hat{\sigma}^2_{\mu_{ab}} = \frac{1}{K}\cdot\frac{1}{K-1}\sum_{i=1}^{K}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ab,ij} - \frac{1}{K}\sum_{r=1}^{K}\frac{1}{n_r}\sum_{j=1}^{n_r} y_{ab,rj}\right\}^2, \quad \text{and}$$

$$\hat{\sigma}^2_{\mu^{*\star}_{ab}} = \frac{1}{K}\cdot\frac{1}{K-1}\sum_{i=1}^{K}\left\{\frac{1}{n^*_i}\sum_{j=1}^{n^*_i} y^{*\star}_{ab,ij} - \frac{1}{K}\sum_{r=1}^{K}\frac{1}{n^*_r}\sum_{j=1}^{n^*_r} y^{*\star}_{ab,rj}\right\}^2.$$

We can approximate the distribution of $(\hat{\mu}_{ab} - \mu_{ab})/\hat{\sigma}_{\mu_{ab}}$ with the distribution of $(\hat{\mu}^{*\star}_{ab} - \hat{\mu}_{ab})/\hat{\sigma}_{\mu^{*\star}_{ab}}$. The extension of the KH and SH tests for the average criterion is straightforward. With this criterion, optimal trees from original and resampled data, $\hat{\tau}$ and $\hat{\tau}^{*\star}$, are

$$\hat{\tau} = \underset{\tau\in\{1,\ldots,p\}}{\mathrm{argmax}}\left\{\frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i} l_{\tau,ij}\right\}, \quad \text{and}$$

$$\hat{\tau}^{*\star} = \underset{\tau\in\{1,\ldots,p\}}{\mathrm{argmax}}\left\{\frac{1}{K}\sum_{i=1}^{K}\frac{1}{n^*_i}\sum_{j=1}^{n^*_i} l^{*\star}_{\tau,ij} - \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i} l_{\tau,ij}\right\}. \qquad [13]$$

The tree inferred with the average criterion is not necessarily the same as the tree obtained by the sum criterion (see below for examples).

Results

Data. As an example, we analyzed data from Cao *et al.* (10). Sitewise log-likelihood values ($l_{a,ij}$'s) of mitochondrial genes were obtained from the authors. The purpose of Cao *et al.*'s work was to find the position of turtle within the amniotes group (for the 15 tree topologies considered by Cao *et al.*, see Table 1). Rather than placing our focus on the turtle position, here we concentrate on our two-stage resampling procedure and the difference between our method and the weighted SH test after sequence concatenation or after merging separately calculated log-likelihoods. Among 14 mitochondrial genes, we excluded 12S and 16S rRNA data. These two rRNA genes were analyzed by Cao *et al.* (10) with a nucleotide substitution model, whereas the other 12 genes are protein-coding genes, and Cao *et al.* (10) used an amino acid replacement model to analyze them. With our hierarchical approach, it may be unreasonable to assume that the former two genes are observations from the same distribution as the 12 protein-coding genes. Tree support varies among the 12 protein-coding genes. When the 12 protein-coding genes were separately analyzed by Cao *et al.* (10), they found that topologies 2, 3, 4, and 9 were the maximum likelihood topologies for 5, 5, 1, and 1 genes, respectively.

Variability Among Genes. We compared the distribution of sitewise log-likelihood ($l_{a,ij}$) values between gene pairs and tested whether those distributions are significantly different. To do this, we performed the nonparametric Kolmogorov–Smirnov test (25). Many pairwise comparisons show a significant difference between genes (see supporting information). We obtained qualitatively similar results for all 15 topologies. This finding implies that ignoring gene-specific effects may be unwarranted and that neither sequence concatenation nor a conventional one-stage bootstrap procedure are recommended. In a similar way, we considered topology pairs *a* and *b* and tested the among-gene variability of sitewise differences of log-likelihood ($y_{ab,ij}$) values. We observed high variability among genes for most tree pairs that were examined (data not shown).

Comparison with SH Test. Because the second term of Eq. **7** is not negligible in general, we need proper "weighting" by $\hat{\sigma}_{ab}$ and $\hat{\sigma}^{*\star}_{ab}$ as described in *Methods* to approximate the distribution of $(S_{ab} - Kn\mu_{ab})$ with the distribution of $(S^{*\star}_{ab} - S_{ab})$. The weighting scheme also can be adapted to the simple SH test. We refer to this adaptation as the weighted SH (WSH) test (4). Because it considers weighting in its test statistic, our method corresponds more to the WSH test than the unweighted SH-test. We obtained optimal tree topologies with the sum and average criteria and we applied the two-stage resampling idea to the SH test. We compared our two stage procedures to the WSH test after sequence concatenation and the WSH test after merging separately calculated log-likelihoods (Table 1).

We applied the WSH test by using the CONSEL software package (26) to sample with replacement from the sitewise log-likelihoods obtained in the analysis of Cao *et al.* (10). The bootstrap probabilities (BPs) of concatenation and merging procedures were calculated with the CONSEL software. The BPs of our two-stage procedure were calculated in similar ways to Eqs. **12** and **13** without centering.

In our two-stage procedures, topologies 2 and 3 are obtained by the sum and average criteria. Topology 2 is also obtained from the procedure of sequence concatenation and merging log-likelihoods. The confidence sets of topologies in Table 1 contain the same trees (trees 2–4) with a 5% significance level, but the *P* values of the topologies vary among sets.

Discussion

In the *Methods*, we explained our method in a likelihood context. If we redefine $l_{a,ij}$ as the parsimony score of the *j*th site of the *i*th gene under tree *a*, then getting and testing the optimal tree by parsimony could be done as described in *Methods.* The idea can also be straightforwardly applied to distance matrix methods. Because the hierarchical structure of gene and sequence columns can be extended to the hierarchical structure of mKLD, it could be also extended to a hierarchical structure of evolutionary distance. From the multiple distance matrices obtained in analyses of separate genes, an average distance matrix could be calculated. The optimal tree would then be reconstructed with this average distance matrix. To assign bootstrap support to subclades, a two-stage bootstrap procedure could be adopted. The idea of an average distance matrix corresponds to the average criterion. The sum criterion also can be extended to the distance method. We can multiply each distance

Table 1. *P* value and bootstrap probability (BP) of each tree calculated with concatenation of sequences, merging separately calculated log-likelihoods, sum criterion, and average criterion

	Concatenation		Merging lnL		Sum criterion		Average criterion	
Tree	WSH	BP	WSH	BP	TS-SH	TS-BP	TS-SH	TS-BP
1	0.0004	0	0.001	0	0.0040	0.0001	0.0076	0.0001
2	0.992*	0.867*	0.931*	0.612*	0.8619*	0.5806*	0.6411	0.3424
3	0.399	0.129	0.784	0.388	0.7823	0.4161	0.8740*	0.6352*
4	0.125	0.004	0.113	0.0002	0.0982	0.0023	0.2000	0.0116
5	0.0004	0	0.002	0.0001	0.0035	0.0001	0.0127	0.0007
6	0.001	0	0.003	0.0003	0.0055	0.0001	0.0304	0.0030
7	0.001	0	0.001	0	0.0019	0	0.0172	0.0003
8	0.001	0	0.001	0	0.0026	0	0.0271	0.0002
9	0.0002	0	0.002	0	0.0251	0.0007	0.0491	0.0063
10	0.001	0	0.001	0	0.0092	0	0.0072	0
11	0	0	0.001	0	0.0013	0	0.0009	0
12	0.0003	0	0.0004	0	0	0	0.0001	0
13	0.0002	0	0.001	0	0.0002	0	0.0001	0
14	0.0001	0	0.001	0	0.0021	0	0.0076	0.0001
15	0.0003	0	0.002	0	0.0051	0	0.0210	0.0001

The WSH and two-stage (TS) SH columns show *P* values, and BP columns show bootstrap probability. Asterisks indicate the optimal topology for each approach. Tree topologies are as follows: 1, (((bird, crocodile), squamate), turtle); 2, (((bird, crocodile), turtle), squamate); 3, ((bird, (turtle, crocodile)), squamate); 4, (((bird, turtle), crocodile), squamate); 5, ((bird, crocodile), (turtle, squamate)); 6, (((bird, turtle), squamate), crocodile); 7, ((bird, turtle), (crocodile, squamate)); 8, (bird, ((turtle, crocodile), squamate)); 9, ((bird, squamate), (turtle, crocodile)); 10, ((bird, (crocodile, squamate)), turtle); 11, (((bird, squamate), crocodile), turtle); 12, (bird, (turtle, (crocodile, squamate))); 13, (bird, ((turtle, squamate), crocodile)); 14, (((bird, squamate), turtle), crocodile); 15, ((bird, (turtle, squamate)), crocodile).

matrix with the length of the gene. The tree can be reconstructed with the sum of these multiple matrices. A two-stage bootstrap procedure can be employed to get the bootstrap support. For long branches on an evolutionary tree of a gene, the estimated distance can be extremely large and have poor precision. Therefore, instead of using the average or sum of distances, it might be better to use the median of multiple estimated distances.

It is possible to estimate $\sigma^2_{ab,W}$. For gene i, $\hat{\sigma}^2_{ab,i}$ can be obtained with samples of $y_{ab,ij}$ ($j = 1, \ldots, n_i$). Using the fact that $\hat{\sigma}^2_{ab}$ of Eq. **8** is an unbiased estimator of $\mathrm{Var}(S_{ab})$ of Eq. **4** together with estimated $\hat{\sigma}^2_{ab,i}$, $\hat{\sigma}^2_n$, $\hat{n}$, and $\hat{\mu}_{ab}$, we can obtain $\hat{\sigma}^2_{ab,W}$. To calculate the precision of $\hat{\sigma}^2_{ab,W}$, we should specify the distributions of $y_{ab,ij}$ and n_i, or at least know their third and fourth moments. Because we make minimal assumptions about the first two moments of $y_{ab,ij}$ and n_i, we do not try to quantify $\hat{\sigma}^2_{ab,W}$. If $y_{ab,ij}$ and $w_{ab,i}$ followed normal distributions with all $\sigma^2_{ab,i}$ values equal, the conventional random effects approach (21) could test whether $\sigma^2_{ab,W} = 0$. Instead, we try to make our method less dependent on the model type and try to make our method less parametric. For this reason, we investigate the heterogeneity of the distribution of $y_{ab,ij}$ with the Kolmogorov–Smirnov test instead of directly quantifying $\hat{\sigma}^2_{ab,W}$. Even when $\hat{\sigma}^2_{ab,W} = 0$, the distribution of $y_{ab,ij}$ can be heterogeneous because the $\sigma^2_{ab,i}$ terms can vary among genes. When all genes have the same $\sigma^2_{ab,i}$, there still may be variation among genes in impact on the sum criterion because of different sequence lengths n_i and because $\sigma^2_{ab,W}$ may not be zero. This means that concatenation of sequences or merging sitewise log-likelihoods followed by the *iid* assumption are not reasonable approaches.

From Eq. **4**, we see that the sum of sitewise log-likelihood values over all columns of all genes is expected to have smaller variance when σ^2_n is small and that variances of the log-likelihoods of genes are significantly affected by the variance of sequence length. On the other hand, the average over all genes of the average log-likelihood per site removes the effect of sequence length and reflects the average structure among genes robustly. The sum and average criteria are two extremes among many possible weighting schemes. Intermediate schemes exist, but it is unclear which intermediate would be best. More experience with genomic data and familiarity with gene-specific variation would help in determining how to select an optimal scheme.

From Cao *et al.*'s data (10), we excluded the two of 14 mitochondrial genes that are not protein-coding. We did this in order not to violate the hierarchical structure of the mKLD differences, $y_{ab,ij}$. We showed that hierarchical structure of genes and sites within genes are straightforwardly extended to the hierarchical structure of mKLD difference. With different amino acid (or nucleotide) substitution models in different genes, or if genes are seriously heterogeneous for reasons such as horizontal gene transfer, the hierarchical *iid* assumption of mKLD is violated. If topological difference has a big impact on mKLD but choice of evolutionary model does not, then different models for different genes or heterogeneous genes would not be a problem.

The 12 protein-coding genes from Cao *et al.* (10) are variable in terms of tree support. Topologies 2, 3, 4, and 9 are the maximum likelihood trees for 5, 5, 1, and 1 individually analyzed genes, respectively. The total concatenated sequence length is 3,235 aa. The sum of lengths of genes that support topology 2 is 1,651 aa, and the sum supporting topology 3 is 1,118 aa. This finding is consistent with the fact that the optimal tree is topology 2 under the concatenation and merging procedures with the sum criterion. The length of the COB gene, which supports topology 9, is 365 aa. This relatively short gene has little effect in the concatenation and merging approaches with the sum criterion. With our two-stage resampling procedure, COB can be resampled more than once and can have a bigger effect. Because the average criterion removes the effect of sequence length, it produces higher *P* values and bootstrap probabilities than the sum criterion for topology 9. Other analyses that we have performed also indicate that the effect of gene sampling on phylogenetic inference can be substantial (e.g., see supporting information).

In our approach, sequence columns are assumed to be *iid* samples from a common distribution. However there may exist heterogeneous partitions within a gene (e.g., ref. 14). For example, the three codon positions of protein-coding genes show significant heterogeneity in evolutionary dynamics (e.g., ref. 28). For protein-coding sequences, partitioning could also be done according to structural environment of the position (e.g., buried or exposed to solvent,

α-helix or β-strand or coil). Here, we do not intensively investigate the sometimes difficult issue of choosing data partitions. More understanding of the evolutionary process would surely assist in defining data partitions.

Although categorizing by codon position or structural environment may lead to substantial variation in evolutionary process among partitions, these categorization schemes do not fit well into our hierarchical framework. We have treated gene-specific perturbations from a common evolutionary mechanism as random effects. Variation among partitions defined by codon position or protein structure would probably be better described with fixed effects. It is not easy to envision randomly sampling partitions defined by codon or protein structure.

Moreover, if each codon position or structural partition has an appropriate evolutionary model, the ratio of lengths between different branches on a tree might not be expected to vary among partitions. There is no obvious reason, for example, that we would expect all α-helix positions in a genome to evolve quickly relative to β-strand positions on one branch of a tree but slowly relative to β-strand positions on another branch of the tree. In contrast, genes are units on which phenotypic selection acts. Natural selection might induce gene-specific effects on branch length that could vary among branches and would yield ratios of branch lengths that vary among genes.

There have been proposals to take into account the variability of evolutionary process among genes (e.g., refs. 28 and 29). One treatment assumes all genes share a topology and the branch lengths for different genes vary only by gene-specific proportionality constants (28). Another assumes all genes evolved via the same topology, but there is no correlation when branch lengths of different genes are examined. Comparisons between the proportional branch lengths, the uncorrelated branch lengths, and the concatenated treatment have been made (30). Recently, the proportionality treatment was adapted to a Bayesian framework (31 and 32), and the uncorrelated treatment could be similarly adapted. In our procedure, all genes (or data partitions) are assumed to be independent samples from the same distribution. However, partitions that seriously violate this assumption (e.g., morphological versus sequence data) can be simultaneously analyzed in a Bayesian framework (refs. 31 and 32, see also ref. 33). Suchard *et al.* (34) recently introduced a promising hierarchical Bayesian approach that can pool information among genes without requiring either uncorrelated or strictly proportional branch lengths.

We have referred to the gene-specific properties that are our focus here as being due to "perturbations" from a "common mechanism." We have intentionally been somewhat vague as to the possible biological sources of these perturbations. The most obvious explanation for gene-specific perturbations is mutation or natural selection. Natural selection operating on a trait coded by the gene could simultaneously affect the branch lengths or other parameters at all gene positions.

Although natural selection and mutation are likely to be important sources of gene-specific perturbations in evolutionary parameters, they should not generate different topologies for different genes. Even if all genes evolved according to the same topology, gene-specific perturbations may be partially responsible for variation among genes in the level of support for the shared topology. Our methods are designed to consider this variation.

A possibility that could result in different topological histories among genes is differential lineage sorting. Rannala and Yang (27) explicitly considered lineage sorting by ascribing variability among gene trees to ancestral polymorphism and stochasticity. There are many advantages to the direct and explicit treatment of Rannala and Yang (27). A possible disadvantage is that highly detailed models may lack robustness when assumptions are violated. Our less detailed treatment of gene-specific properties assumes there is a central tree of interest and that random gene-specific perturbations are unbiased relative to this tree. The fact that we do not specify the biological source of these perturbations can be viewed as either an advantage or disadvantage of our procedure.

Regarding the source of gene-specific perturbations, another possibility is they arise from model misspecification. It may be that all genes share a common history but that the model employed to analyze them is not equally appropriate for all genes. If the effects of this model misspecification are unbiased among genes in that the average inference among genes is expected to be the truth, our hierarchical treatment would be appropriate. Undoubtedly, a more desirable solution would be to remedy the misspecification by improving the model.

Here, we use the sum and average criteria to infer the optimal tree from multigene data. We also introduce a two-stage bootstrap procedure to test the significance of the optimal tree. Our method has an advantage over sequence concatenation in that it can effectively consider gene-specific features. It is justified by the assumption of a hierarchical structure of genes and sites within genes. The method has the advantage of being able to combine the information from multiple genes that might support different trees. We expect that the rapid growth of genomic data will lead to a better understanding of the effects of gene sampling on evolutionary inference. For the time being, only a small number of genes are typically available. With these small data sets, it is especially important to consider effects of gene sampling on the bias and power of phylogenetic hypothesis testing.

We thank Y. Cao for providing data. This work was supported by the Institute for Bioinformatics Research and Development of the Japanese Science and Technology Corporation, Japanese Society for the Promotion of Science Grant 16300086, and National Science Foundation Grants INT-99-0934 and DEB-0120635.

1. Felsenstein, J. (2004) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
2. Felsenstein, J. (1985) *Evolution (Lawrence, Kans.)* **39,** 783–791.
3. Kishino, H. & Hasegawa, M. (1989) *J. Mol. Evol.* **29,** 170–179.
4. Shimodaira, H. & Hasegawa, M. (1999) *Mol. Biol. Evol.* **16,** 1114–1116.
5. Pollock, D. D., Taylor, W. R. & Goldman, N. (1999) *J. Mol. Biol.* **287,** 187–198.
6. Jensen, J. L. & Pedersen, A. K. (2000) *Adv. Appl. Prob.* **32,** 499–517.
7. Pedersen, A. K. & Jensen, J. L. (2001) *Mol. Biol. Evol.* **18:** 763–776.
8. Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N. & Thorne, J. L. (2003) *Mol. Biol. Evol.* **20,** 1692–1704.
9. Adachi, J., Waddell, P., Martin, W. & Hasegawa, M. (2000) *J. Mol. Evol.* **50,** 348–358.
10. Cao, Y., Sorenson, M. D., Kumazawa, Y., Mindell, D. P. & Hasegawa, M. (2000) *Gene* **259,** 139–148.
11. Cao, Y., Fujiwara, M., Nikaido, M., Okada, N. & Hasegawa, M. (2000) *Gene* **259,** 149–158.
12. Nikaido, M., Cao, Y., Harada, M., Okada, N. & Hasegawa, M. (2003) *Mol. Phylogenet. Evol.* **28,** 276–284.
13. Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L. & Waddell, P. J. (1993) *Syst. Biol.* **42,** 384–397.
14. Debry, R. W. (1999) *Syst. Biol.* **48,** 286–299.
15. Debry, R. W. (2003) *Syst. Biol.* **52,** 604–617.
16. Gontcharov, A. A., Marin, B. & Melkonian, M. (2004) *Mol. Biol. Evol.* **21,** 612–624.
17. Farris, J. S., Källersjö, M., Kluge, A. G. & Bult, C. (1994) *Cladistics* **10,** 315–319.
18. Swofford, D. L. (1995) PAUP*: *Phylogenetic Analysis Using Parsimony (* and Other Methods)* (Sinauer, Sunderland, MA).
19. Waddell, P. J., Kishino, H. & Ota, R. (2000) *Mol. Biol. Evol.* **17,** 1988–1992.
20. Rao, J. N. K. & Wu, C. F. J. (1988) *J. Am. Stat. Assoc.* **83,** 231–241.
21. Scheffé, H. (1959) in *The Analysis of Variance* (Wiley, New York), pp. 221–260.
22. Kullback, S. & Leibler, R. A. (1951) *Ann. Math. Stat.* **22,** 79–86.
23. Hall, P. & Wilson, S. R. (1991) *Biometrics* **47,** 757–762.
24. Goldman, N., Anderson, J. P. & Rodrigo, A. G. (2000) *Syst. Biol.* **49,** 652–670.
25. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992) in *Numerical Recipes in C* (Cambridge Univ. Press, Cambridge, U.K.), 2nd Ed., pp. 623–628.
26. Shimodaira, H. & Hasegawa, M. (2001) *Bioinformatics* **17,** 1246–1247.
27. Rannala, B. & Yang, Z. (2003) *Genetics* **164,** 1645–1656.
28. Yang, Z. (1996) *J. Mol. Evol.* **42,** 587–596.
29. Yoder, A. D. & Yang, Z. (2000) *Mol. Biol. Evol.* **17,** 1081–1090.
30. Pupko, T., Huchon, D., Cao, Y., Okada, N. & Hasegawa, M. (2002) *Mol. Biol. Evol.* **19,** 2294–2307.
31. Ronquist, F. & Huelsenbeck, J. P. (2003) *Bioinformatics* **19,** 1572–1574.
32. Nylander, J. A. A., Ronquist, R., Huelsenbeck, J. P. & Nieves-Aldrey, J. L. (2004) *Syst. Biol.* **53,** 47–67.
33. Lewis, P. O. (2001) *Syst. Biol.* **50,** 913–925.
34. Suchard, M. A., Kitchen, C. M. R., Sinsheimer, J. S. & Weiss, R. E. (2003) *Syst. Biol.* **52,** 649–664.