Multidimensional Vector Space Representation for Convergent Evolution and Molecular Phylogeny

Yasuhiro Kitazoe, Hirohisa Kishino,† Takahisa Okabayashi,* Teruaki Watabe,* Noriaki Nakajima,* Yoshiyasu Okuhara,* and Yukio Kurihara**

*Center of Medical Information Science, Kochi Medical School, Kochi, Japan; and †Graduate School of Agriculture and Life Sciences, University of Tokyo, Tokyo, Japan

With growing amounts of genome data and constant improvement of models of molecular evolution, phylogenetic reconstruction became more reliable. However, our knowledge of the real process of molecular evolution is still limited. When enough large-sized data sets are analyzed, any subtle biases in statistical models can support incorrect topologies significantly because of the high signal-to-noise ratio. We propose a procedure to locate sequences in a multidimensional vector space (MVS), in which the geometry of the space is uniquely determined in such a way that the vectors of sequence evolution are orthogonal among different branches. In this paper, the MVS approach is developed to detect and remove biases in models of molecular evolution caused by unrecognized convergent evolution among lineages or unexpected patterns of substitutions. Biases in the estimated pairwise distances are identified as deviations (outliers) of sequence spatial vectors from the expected orthogonality. Modifications to the estimated distances are made by minimizing an index to quantify the deviations. In this way, it becomes possible to reconstruct the phylogenetic tree, taking account of possible biases in the model of molecular evolution. The efficacy of the modification procedure was verified by simulating evolution on various topologies with rate heterogeneity and convergent change. The phylogeny of placental mammals in previous analyses of large data sets has varied according to the genes being analyzed. Systematic deviations caused by convergent evolution were detected by our procedure in all representative data sets and were found to strongly affect the tree structure. However, the bias correction yielded a consistent topology among data sets. The existence of strong biases was validated by examining the sites of convergent evolution between the hedgehog and other species in mitochondrial data set. This convergent evolution explains why it has been difficult to determine the phylogenetic placement of the hedgehog in previous studies.

Introduction

Morphological cladistics is an important starting point in evolutionary studies, but it is not free from the effects of phenotypic traits acquired by adaptive convergent evolution resulting from similar environments. Molecular phylogenetics has been used successfully to supplement or replace the results obtained by morphological and paleontological approaches (Cann, Stoneking, and Wilson 1987; Kocher et al. 1993). Recently, it has been possible to infer phylogenies using large data sets in whole genomes (Wolf et al. 2002). However, our knowledge regarding the mechanism of molecular evolution is still limited. Almost all current methods of tree building implicitly assume that sequences evolve independently after divergence. When there is convergent evolution among lineages, the estimated pairwise distances tend to be negatively biased. This bias of pairwise distances results in the lineages that underwent convergent evolution being connected in the inferred tree structure. To date, the existence of convergent evolution has been examined by estimating ancestral states after tree reconstruction (Zhang and Kumar 1997), as it has been difficult to infer where and how the tree structure is distorted in multiple convergent evolutions among lineages. However, this two-step approach is reasonable only when convergent evolution has little effect on phylogenetic inference. When strong correlations exist in data sets with long sequences in which stochastic uncertainty becomes negligible, convergence can lead to significant and serious confusion.

Here, we propose a new procedure to concurrently estimate convergent evolution, site heterogeneity of evolutionary rate, and the phylogenetic tree, by developing a recent method of the MVS representation of sequence evolution (Kitazoe et al. 2001). When an estimated pairwise distance reflects the actual substitution process of amino acids or nucleotides, this distance is equal to the sum of branch lengths connecting the two species (this equality is called the "additivity rule"). According to the additivity rule, all the species are arrayed on the orthogonal coordinate axes of the MVS. On the other hand, when strong convergent evolution occurs among lineages, the corresponding pairwise distances are highly underestimated by the statistical model. The biased pairs of species strongly violate the additivity rule and can be detected as outliers by their deviations from the orthogonal coordinate axes of the MVS. Deviations from the additivity rule are also caused by erroneous estimations of the pairwise distances based on the site heterogeneity of evolutionary rate (Yang 1994; Yang and Kumar 1996). In this paper, we introduce a unified index to estimate these two kinds of deviations. The tree is reconstructed by minimizing this index value without assuming any topologies explicitly. The index is useful to evaluate how much the initially estimated pairwise distances distort the tree structure.

We applied our method to the phylogenetic analysis of placental mammals. Relationships among orders in these mammals remain controversial, despite extensive studies over the past decade (O'Brien et al. 1999; Cao et al. 2000; Mouchaty et al. 2000; Murphy et al. 2001; Madsen et al. 2001; Waddel, Kishino, and Ota 2001; Arnason et al. 2002; Madsen et al. 2002; Hudelot et al. 2003; Nikaido et al. 2003), and molecular analyses conflict with morphological

Key words: multidimensional vector space, convergent evolution, bias of pairwise distances, mammalian phylogeny.

E-mail: kitazoey@med.kochi-u.ac.jp.

Mol. Biol. Evol. 22(3):704–715. 2005
doi:10.1093/molbev/msi051
Advance Access publication November 17, 2004

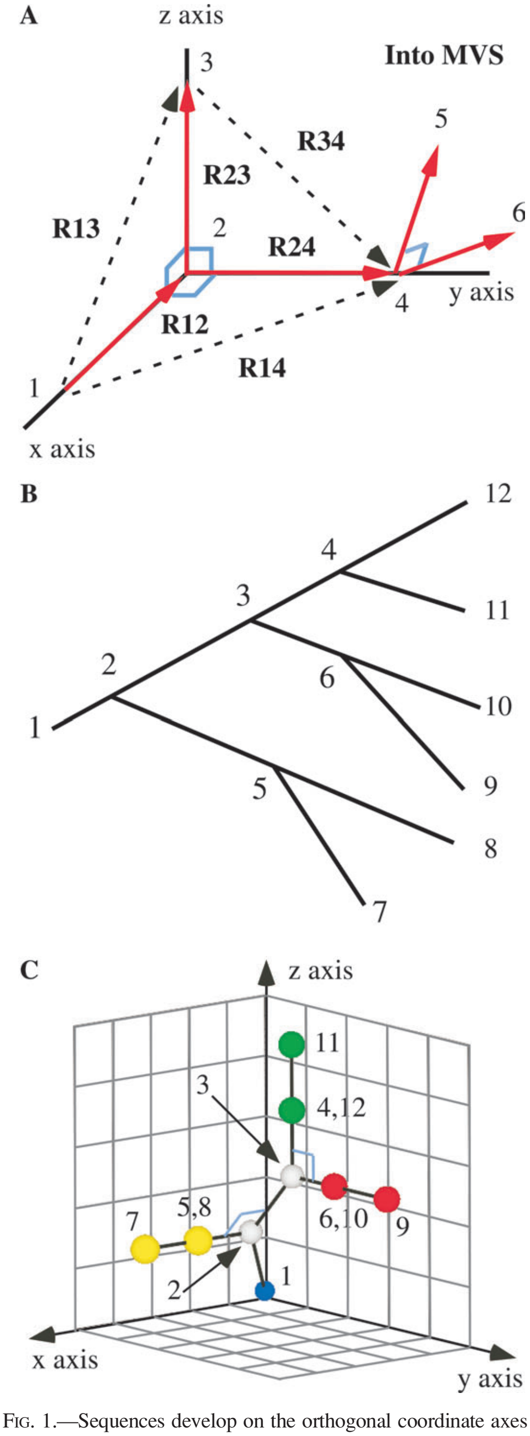

FIG. 1.—Sequences develop on the orthogonal coordinate axes of the MVS (A): a root sequence 1 develops on the *x*-axis over time and splits into two sequences 3 and 4 via the node 2. The sequences 3 and 4 develop on the *z*-axis and *y*-axis, respectively. A further splitting of the sequence 4 into sequences 5 and 6 requires two additional dimensions to make a total of five. Evolution process of many sequences (B) is represented by a three-dimensional MVS map (C) in terms of three quantities, $S_{i,a}(o)$, $S_{i,b}(o)$, and $S_{i,c}(o)$ in equation (1), which are respectively taken as the *x* values, *y* values, and z values. Here, $o=1$, $a=7$, $b=9$, and $c=11$ are assigned as a set of probes.

studies (Novacek 1992, 1999). The evolutionary history of placental mammals is also in debate (Easteal 1999; Benton and Ayala 2003): Molecular studies argue that many extant modern orders diversified long before the Cretaceous-Tertiary boundary (Kumar and Hedges 1998; Cao et al. 2000; Waddel, Kishino, and Ota 2001; Springer et al. 2003), in contrast with the paleontological view that most modern orders occurred after this boundary (Alroy 1999; Archibald and Deutschman 2001). To resolve these conflicts, we investigated how much convergent evolution influences the branching patterns and lengths in mammalian phylogeny. We performed bias correction of the initially estimated pairwise distances in three large-sized data sets consisting of complete mitochondrial genomes (CMT), 19 nuclear genes and three mitochondrial genes (22 GENES), and BRCA1. As a result, we observed that systematic deviations of species from the orthogonal coordinate axes exist in the three data sets and strongly reflect the tree building. These systematic deviations were mainly caused by parallel evolution, the existence of which was confirmed by site-by-site analysis of the CMT data comparing the hedgehog with other species (including supraprimates, afrotherians, marsupials, and monotremes). The procedure of bias detection and correction yielded a consistent topology among the three data sets and a remarkable improvement of the branch resolutions. Based on our analysis, convergence explains why unstable or anomalous branching patterns were reported in previous studies, such as those including the hedgehog and rodents in the CMT tree (Waddel, Kishino, and Ota 2001; Arnason et al. 2002).

Methods

MVS Representation of Sequence Evolution

Kitazoe et al. (2001) proposed that the MVS makes it possible to express the sequence evolution precisely because the pairwise distances among the branching nodes (or endpoints) can be reproduced by using the Pythagorean theorem. The MVS is defined as the metric of the space in which the length of a branch vector is equal to the square root of the number of evolutionary changes. The geometry of space is uniquely specified according to the feature that the spatial vectors representing sequence evolution are orthogonal among different branches. This orthogonal feature of the MVS becomes equivalent to the additivity rule. A spatial vector from one sequence to another sequence is expressed as a composition of branch vectors, which describe evolutionary changes in branches connecting the two sequences. An inner product of two vectors represents the number of changes along the shared branch vectors. To explain this MVS feature explicitly, we consider the elementary process in which a species 1 develops on the *x*-axis over time and splits into two species, 3 and 4, via the node 2 in figure 1A. When $D_{i,j}$ denotes the number of amino acid or nucleotide substitutions between *i*-th and *j*-th species (it is normalized by the total site number), we have the additivity equations, $D_{1,3}=D_{1,2}+D_{2,3}$, $D_{1,4}=D_{1,2}+D_{2,4}$, and $D_{3,4}=D_{2,3}+D_{2,4}$. Here, the square of the spatial distance $R_{i,j}$ is equal to the substitution number $D_{i,j}$; that is, ${R_{i,j}}^2=|\mathbf{R}_i-\mathbf{R}_j|^2=D_{i,j}$, where the vector $\mathbf{R}_i$ denotes the spatial position of *i*-th species. Then, the additivity equations can be rewritten as ${R_{1,3}}^2={R_{1,2}}^2+{R_{2,3}}^2$, ${R_{1,4}}^2={R_{1,2}}^2+{R_{2,4}}^2$, and ${R_{3,4}}^2={R_{2,3}}^2+{R_{2,4}}^2$, which express Pythagorean theorem and imply that species 3 and 4 develop on the *z*-axis and y-axis over time, respectively. Furthermore, the splitting of species 4 into species 5 and 6 requires two additional dimensions to make a total of five (fig. 1A). In this way, all the species develop on the orthogonal coordinate axes over time, and the branching events of species correspond to their developments into new orthogonal coordinate axes. The extant species are the endpoints of these developments, and their spatial vectors reflect their individual evolutionary histories.

It is difficult to directly visualize the tree structure of the MVS with a high number of dimensions. However, the orthogonal feature of branch vectors makes it possible to resolve the tree structure by projecting the species vectors onto specific search vectors. First, a species o is placed at the coordinate origin of the MVS, from which the spatial positions, $\mathbf{R}_{o,i}$, of other species i are measured. The inner product, $S_{i,j}(o)$, of two vectors $\mathbf{R}_{o,i}$ and $\mathbf{R}_{o,j}$ gives the branch length from the origin o to the most-recent common ancestor c of species i and j, if the additivity rule is satisfied. This is proved by the relation $S_{i,j}(o) = \mathbf{R}_{o,i} \cdot \mathbf{R}_{o,j} = (\mathbf{R}_{o,c} + \mathbf{R}_{c,i}) \cdot (\mathbf{R}_{o,c} + \mathbf{R}_{c,j}) = R_{o,c}^2 = D_{o,c}$, because $\mathbf{R}_{o,c} \perp \mathbf{R}_{c,j}$, $\mathbf{R}_{o,c} \perp \mathbf{R}_{c,i}$, and $\mathbf{R}_{c,i} \perp \mathbf{R}_{c,j}$. The branch length is rewritten as $S_{i,j}(o) = R_{o,i}R_{o,j}\cos(\theta_{i,j})$, with the angle $\theta_{i,j}$ between the vectors $\mathbf{R}_{o,i}$ and $\mathbf{R}_{o,j}$. Using the cosine theorem $(R_{i,j}^2 = R_{o,i}^2 + R_{o,j}^2 - 2R_{o,i}R_{o,j}\cos(\theta_{i,j}))$ about the triangle (o, i, j) gives the equation,

$$S_{i,j}(o) = (D_{o,i} + D_{o,j} - D_{i,j})/2 \qquad (1)$$

Note that $S_{i,i}(o) = D_{o,i}$. In this way, the branch length $S_{i,j}(o)$ is analytically expressed in terms of $D_{i,j}$, although it was previously obtained using the position vectors $\mathbf{R}_{o,i}$ and $\mathbf{R}_{o,j}$, for which an equation of motion for the many-body system was directly solved (Kitazoe et al. 2001).

Making use of a set of "probe species" (o, a, b, and c), we investigate how the other species i are branched around the probes. The branching pattern can be analyzed by a scatter diagram (hereafter called the "MVS map") in terms of three quantities, $S_{i,a}(o)$, $S_{i,b}(o)$, and $S_{i,c}(o)$, which are respectively taken as the x-values, y-values, and z-values. For instance, the sequence evolution of figure 1B is represented by a three-dimensional display of figure 1C. Here, $o = 1$, $a = 7$, $b = 9$, and $c = 11$ are assigned as a set of probes. The three-dimensional display (fig. 1C) shows that the sequence develops along the diagonal line from the ancestor 1 to the node 2 and splits into two sequences with the right angle at the node 2. One of them develops parallel to the x-axis to the endpoint 7. The other point moves parallel to the y-z plane to the node 3, at which it splits into the two lineages, 9 and 11, with the right angle. These two lineages continue to develop parallel to the y-z plane to their endpoints 9 and 11. The endpoints 8, 10, and, 12 are degenerated into the positions of the nodes 5, 6, and 4 because of the orthogonality of branch vectors, respectively. In this way, the whole tree structure can be resolved using all possible sets of the probes. Therefore, the MVS map implies an alternative representation of phylogeny.

Detection of Convergent Evolution in the MVS Map

In a practical phylogenetic inference, the pairwise distances have to be initially estimated by using the sequences of extant species. Convergent evolution underestimates the corresponding pairwise distances and produces large biases of the distances. Here, we show that such biases can be detected by the MVS map of the two-dimensional display because the MVS map given by taking the biased pairs as the probes represents a systematic deviation of many other species from the additivity rule. To illustrate such a deviation with a simple example, let us assume that convergent evolution occurred between two branches (3→9 and 5→10) in figure 2A, and the distance between points 9 and 10 was highly underestimated. Here, taking the probes o =10, a = 9, and b = 7 yields a systematic deviation of the points 11 to 13 from the diagonal line, as shown by the open circles of figure 2B. These open circles lie on the regression (dotted) line with $y = x + q$. When we increase the distance between the two points 9 and 10 by $2q$, the open circles move to the positions of the solid circles. The solid circles lie on the three lines, **A**, **B**, and **C** (hereafter called the "additivity lines"), and fulfill the following equations for the

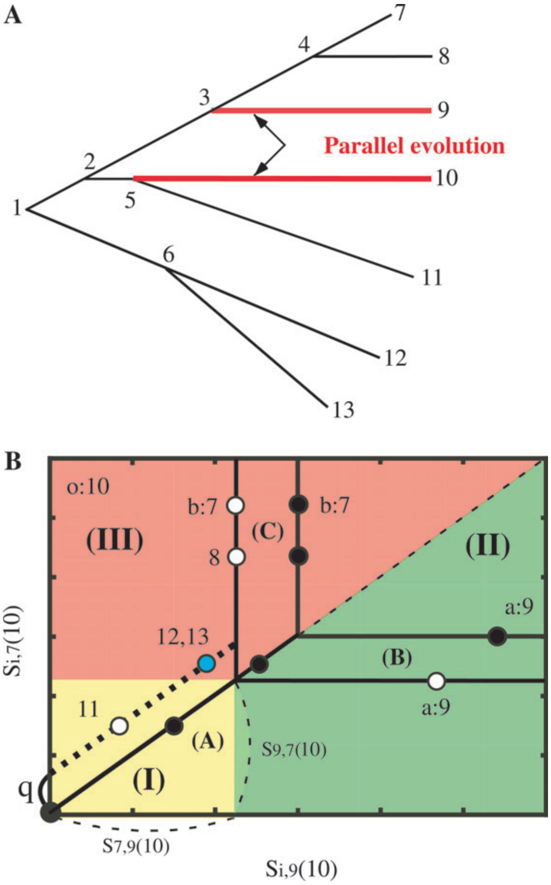

FIG. 2.—Strong convergent evolution between the two branches 3→9 and 5→10 gives rise to underestimation of the distance between the two points 9 and 10 (A) by using any statistical modes. Here, taking o = 10, a = 9, and b = 7 as the probes, a two-dimensional MVS map represents a systematic deviation (the open circles) of the points 11 to 13 from the diagonal line (B). The open circles lie on the regression (dotted) line with $y = x + q$. Increasing the distance between the two points 9 and 10 by $2q$, the open circles move to the positions of the solid circles. The solid circles are arrayed on the three lines, **A**, **B**, and **C**, to fulfill the additivity rule. Consequently, a strong bias of the estimated pairwise distance is removed.

additivity rule: $S_{11,7}(10) = S_{11,9}(10) = D_{10,5}$, $S_{12,7}(10) = S_{12,9}(10) = S_{13,7}(10) = S_{13,9}(10) = D_{10,2}$, $S_{9,7}(10) = S_{9,8}(10) = D_{10,3}$. Then, the solid circles are arrayed in nearest order of branches with the probes. The points 11 to 13, outside the common ancestor 3 of the probes 7 and 9, lie on the diagonal line **A**. The point 9, closest to the probe 8, is located on line **C**. The points 12 and 13 are degenerated into the position of the single node 2. In figure 2B, although only four points exist around the regression line, more points around this line make the minimum modification of distances more reliable. As illustrated in figure 2B, a strong bias between the points 9 and 10 allows the points 12 and 13 to be placed in region III. Then, the four-point condition (Holland et al. 2002) that infers the branch pattern among four species misleadingly connects the points 12 and 13 to the lineage of the probe 7.

A Unified Index to Detect Convergent Evolution

Now, we estimate the deviation from the additivity rule. For a particular probe set (o, a, b), the deviation of a species i is defined as the distance to the nearest additivity line of this species. Because $S_{a,b}(o)$ gives the y value of the line **B** and the x value of the line **C** in figure 2B, the deviation of i in relation to a triplet (o, a, b) is expressed as

$$\begin{aligned} V_{o,a,b,i} &= S_{i,b}(o) - S_{i,a}(o), && \text{if } S_{i,a}(o) < S_{a,b}(o) \text{ and} \\ & && S_{i,b}(o) < S_{a,b}(o)(\text{region I}) \\ &= S_{i,b}(o) - S_{a,b}(o), && \text{if } S_{i,a}(o) > S_{a,b}(o) \text{ and} \\ & && S_{i,a}(o) > S_{i,b}(o)(\text{region II}) \\ &= S_{i,a}(o) - S_{a,b}(o), && \text{if } S_{i,b}(o) > S_{a,b}(o) \text{ and} \\ & && S_{i,b}(o) > S_{i,a}(o)(\text{region II}) \end{aligned} \tag{2}$$

The quantity $V_{o,a,b,i}$ is consistent with an index for the four-point condition (Holland et al. 2002) that infers the branching pattern among four species of o, a, b, and i. This condition indicates that a species i has a sister relation with the probe in the same region. The deviation, $W_{o,a,b}$, for the triplet (o, a, b) and the total deviation W of the system are defined by

$$W_{o,a,b} = \sum_{i} V^2_{o,a,b,i}, \quad W = \sum_{o,a,b} W_{o,a,b} \tag{3}$$

The quantities $W_{o,a} = \sum_b W_{o,a,b}$ and $W_o = \sum_a W_{o,a}$ are used to draw out species pairs (o, a) and species (o) causing large deviations, respectively.

The main function of MVS analysis is to remove the total deviation W by the minimal modification of the initially estimated pairwise distances. We assume that the estimation can be done by means of statistical models (Felsenstein 1996) of amino acid or nucleotide transitions and gives the good initial values for the MVS analysis. Although this estimation using pairs of sequences cannot take account of convergent evolution among lineages, the MVS representation of many-body configurations of species makes it possible to detect the correlated pairs, as systematic deviations of species from the additivity lines appear because of underestimations of these pairwise distances. Therefore, the strength of the present method is to examine a large number of species in which the majority of the estimated pairwise distances approximately satisfy the additivity rule. The systematic deviations can be detected insensitive to models of the transition probability and are removed by modifying the pairwise distances with strong biases so as to satisfy the additivity rule. This procedure implies the minimum modification of distances, which gives rise to a consistent branching pattern, as mentioned below. The usual expression of a tree is given by the neighbor-joining (NJ) method (Saitou and Nei 1987) after the modification.

Phylogenetic Inference Taking Account of Biases in the Evolutionary Model

To obtain a stable and reliable solution in the MVS analysis, it is useful to carry out the tree building from a macroscopic toward a microscopic structure. We assume that species can be decomposed into several groups with very high confidence by selecting sets of probes minimizing the deviations from the additivity rule. Here, we show a useful method to precisely modify the distances among these groups. The method consists of the following two steps.

Distance Modification of Intergroups and Intragroups

The probe o is picked up from one group, and the other probes (a, b) are picked up from the other groups. We consider the case in which species points in the MVS map are systematically deviated from the diagonal line **A**, as seen in figure 2B. These points are fitted by the function $y = x + q$. When the distance between o and a is enlarged by only $2|q|$ (we take b if $q < 0$) and the deviated points are distributed around the diagonal line, the systematic deviation (bias) disappears. This enlargement implies a minimal and unique modification of the pairwise distances. Furthermore, we enlarge the distances between a (or b) and the species i so that they lie on a diagonal line, because it is often difficult to discriminate between noise and small bias. This procedure is repeated in order from smallest to largest $|q|$ value over all possible triplets of o, a, and b. When the orthogonal relation among the groups is fulfilled, the bias correction within a group is continued in a straightforward manner. Any species of the other groups is taken as the probe o, and two species within the group are taken as the probes of a and b. Our task is only to enlarge the distances between species within the group so that they may lie on the additivity lines **B** and **C** over all sets of a, b. We repeat this procedure for all possible sets of a, b and do this in the other groups.

Distance Fluctuation Caused by Stochastic Noise

Our method of bias correction is most efficient when a large data set without stochastic variance is analyzed. When the size of sequence length is not long enough, the estimated pairwise distances do not follow the additivity rule precisely because of random noise, even if the model describes the real evolutionary process well on average. Hence, the distance-modification procedure overcorrects the deviation from the additivity rule. We only consider negative bias of pairwise distances because of difficulty discriminating small bias and random noise. The extent of overcorrection is measured by analyzing unbiased pairwise distances with some noise, $E'_{i,j}$. Because we cannot assume that the original pairwise distances, $D_{i,j}$, are unbiased, we carried out the semiparametric bootstrap to simulate $E'_{i,j}$ using the information of the corrected distances, $E_{i,j}$, and the variance-covariance of the original distances, $D_{i,j}$. Specifically, a multivariate normal random number is added to the set of the corrected pairwise distances, $E_{i,j}$. The variance-covariance matrix is calculated based on the pairwise distances between sequences of resampled sites. By applying the distance-modification procedure to the simulated pairwise distances, $E'_{i,j}$, we obtain modified distances, $F_{i,j}$, and evaluate the overcorrection as the difference, $G_{i,j} = F_{i,j} - E_{i,j}$. Revised distances may be obtained as $H_{i,j} = E_{i,j} - G_{i,j} = 2\,E_{i,j} - F_{i,j}$, but we notice that they do not follow the additivity rule any more. Instead, we calculate the revised distances by application of the distance-modification procedure to the shrinked distances, $D_{i,j}' = D_{i,j} - (F_{i,j} - E_{i,j})$. This whole process is repeated until the convergence is reached.

Site-Heterogeneity of Evolutionary Rate

The initial pairwise distances $D_{i,j}$ are estimated by taking into account the site-heterogeneity of evolutionary rate with the shape parameter α (Yang 1994; Yang and Kumar 1996). The value of this parameter can be estimated by multiple sequence comparison. Here, without specifying topology, we estimate this value in a way that the resultant pairwise distances become as consistent as possible with the tree structure. As an analog to the interpretation of the least-squares estimate as a maximum-likelihood estimate with a normal distribution for random noise, we formally consider the following log-likelihood function:

$$\begin{aligned} L(\alpha, \beta) &= -(N/2)\log(\sigma) - \sum_{o,a,b,i} V^2_{o,a,b,i}/(2\sigma) \\ &= -(N/2)\log(\sigma) - W/(2\sigma) \qquad (4) \end{aligned}$$

with $\sigma = D^{\beta}_{av}$, where D_{av} is the average pairwise distance, and N is the number of all possible sets of *o, a, b*, and *i*. Instead of modeling the variance-covariance matrix of $V_{i:o,a,b}$'s for the dispersion σ, we simply take account of the σ dependency on the α value by way of the average pairwise distance. Then, the parameter α is defined as the maximum-value point of the function $L(\alpha, \beta)$, and the parameter β is adjusted so that the function $L(\alpha, \beta)$ may be symmetrically distributed around the maximum value as a function of α.

Results

The Efficacy of the MVS Procedure Through Simulations of Sequence Evolution

By simulating Markov process of sequence evolutions with strong site heterogeneity of evolutionary rate, we examined how precisely the log-likelihood function $L(\alpha, \beta)$ of equation (4) can reproduce the true values. A simulation of the branch pattern of figure 3A was made using the shape parameter value $\alpha = 0.4$ for site heterogeneity and the JTT model (Jones 1992) for amino acid transition probability. Here, the simulation began with an initial long sequence (10,000 sites) to ignore stochastic noise. This sequence was given by a random distribution

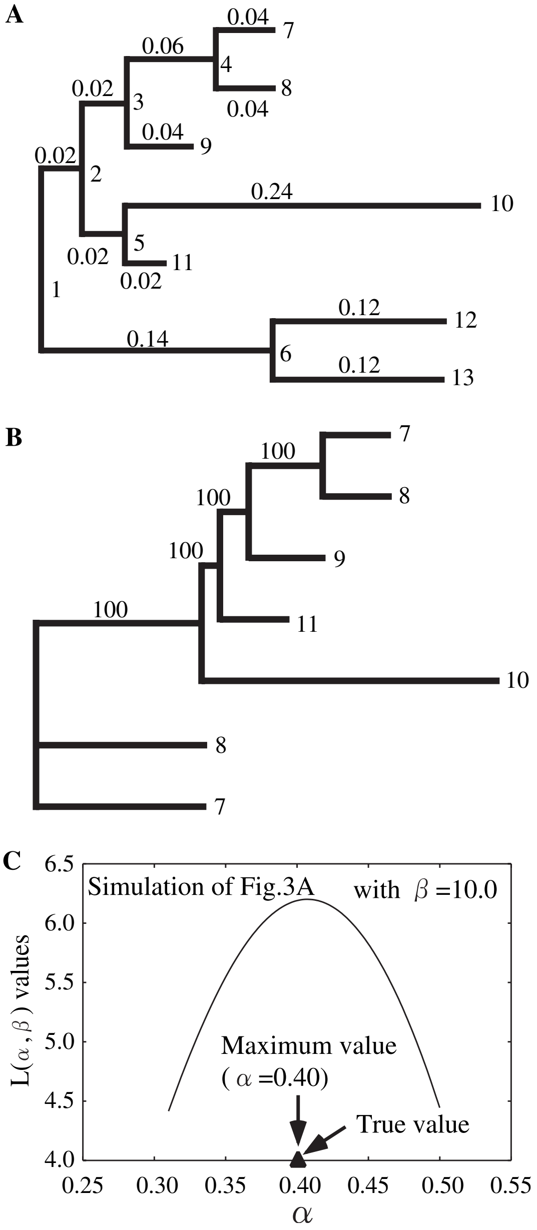

FIG. 3.—Markov process for the branching pattern (A) was simulated with a shape parameter value $\alpha = 0.4$ for site heterogeneity of evolutionary rate (Yang and Kumar 1996) and with the JTT model for amino acid transition probability (Jones, Taylor, and Thornton 1992). With use of the endpoint (7 to 13) sequences, the NJ method yielded an erroneous pattern despite the 100% bootstrap resolution (B), when the pairwise distances were estimated with large α values ($\alpha > 1.2$). In the MVS procedure, the log-likelihood function $L(\alpha, \beta)$ of equation (4) gave a maximum value at $\alpha = 0.4$ without any explicit information regarding the topology (C).

of amino acids. Next, using the obtained sequences of the endpoints (7 to 13), we calculated the pairwise distances with a variety of α values. Using larger values than $\alpha = 1.2$ in the NJ method yielded an erroneous pattern (fig. 3B) with the 100% bootstrap resolution. In the MVS procedure, the function L(α, β) gave a maximum value at $\alpha = 0.4$ ($\beta = 10.0$), without any explicit information regarding the topology (fig. 3C), in contrast to the maximum likelihood (ML) (Felsenstein 1996; Adachi and Hasegawa 1996). A similar value ($\alpha = 0.41$) was given by the ML method, taking account of the site heterogeneity (Yang 1994; Yang and Kumar 1996).

To also demonstrate the detection and distance modification of biased pairs, we simulated molecular evolution of the branching pattern with figure 4A, in a manner similar to the case of heterogeneity. Here, we artificially induced parallel changes only between the two branches 3→9 and 5→10, so that the distance between the points 9 and 10 could be underestimated by 25%. Using the obtained sequences of the endpoints 7 to13, the NJ and ML methods yielded an erroneous pattern (fig. 4B) with the 100% bootstrap resolution, in which the points 12 and 13 were coupled to the points 7 and 8, and, furthermore, the two points 9 and 10 had a sister relation because of a strong attraction between them. On the other hand, in the MVS method, the triplet of probes $o = 10$, $a = 9$, and $b = 7$ gave the maximum deviation. Here, a MVS map quite similar to the open circles of figure 2B was obtained, and the points 11to 13 deviated consistently from the diagonal line, as shown by the regression line with $y = x + q$ of figure 2B. This deviation disappeared by increasing only the distance between the two points 9 and 10 by $2q = 0.07$, and the correct distance ($D_{9,10} = 0.28$) between the two points was obtained. The deviation W of the system gave a minimum value at this correct distance (fig. 4*c*) when it was plotted as a function of the distance between the two points 9 and 10 (fig. 4C).

Mammalian Phylogeny

The method was applied to the phylogenetic inference of placental mammals. We selected three data sets of the CMT (69 species with 3,660 sites) (Nikaido et al. 2003), the BRCA1 (52 species with 5,708 sites) (Madsen et al. 2002), and the 22 GENES (44 species with 17,028 sites) (Murphy et al. 2001) because these data sets have been extensively analyzed in recent years and have low levels of stochastic noise caused by long sequences. Sequence alignments can be obtained from the above-mentioned three works. Here, we analyzed the same sequences as those used in these works. The pairwise distances were calculated using the Jukes-Cantor model (Jukes and Cantor 1969) for nucleotides and the JTT model (Jones, Taylor, and Thornton 1992) for amino acids. Our MVS analysis was fairly robust against transition probability models, such as the JTT-F, Dayhoff, Dayhoff-F (Dayhoff, Schwartz, and Orcutt 1978), and mtREV (Adachi and Hasegawa 1996), because the bias effect based on convergent evolution was frequently much larger than the distance fluctuations caused by stochastic noise. The α values were estimated to be 0.38 for the CMT, 1.5 for the

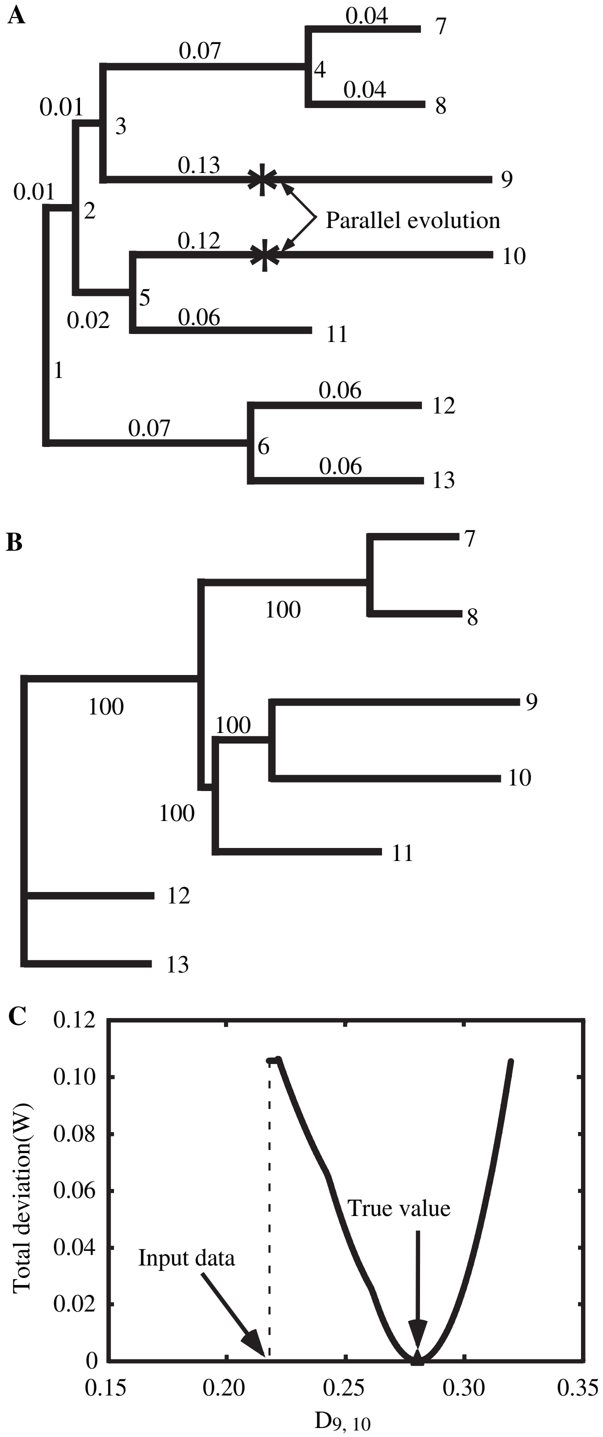

FIG. 4.—Markov process of the branch pattern (A) was simulated, and a strong convergent evolution was added between the two branches 3→9 and 5→10. With use of the endpoint (7 to 13) sequences, the NJ and ML methods gave an erroneous branching pattern, despite the 100% bootstrap resolution (B). In the MVS procedure, the triplet of probes $o = 10$, $a = 9$, and $b = 7$ gave the maximum deviation, and showed a MVS map quite similar to the open circles of figure 2B, in which the points 11 to 13 deviated consistently from the diagonal line. This deviation disappeared by increasing only the distance between the two points 9 and 10 by $2q = 0.07$. When the total deviation W of equation (3) was plotted as a function of the distance between the two points 9 and 10 (fig. 2C), the deviation gave a minimum value at the correct distance ($D_{9,10} = 0.28$) (fig. 2C).

BRCA1, and 0.7 for the 22 GENES, which were similar to those given by the ML method (Yang 1994; Yang and Kumar 1996).

To achieve robustness, we first investigated the global pattern of the phylogenetic relations, and second, we estimated the detailed structure of the tree. We examined separations of superordinal groups by drawing out sets of probes, which are far from each other and minimize the deviations from the additivity rule. As a result, we found that the eutherian trees can be decomposed into four major groups, Laurasiatheria, Supraprimates, Xenarthra, and Afrotheria, in the three data sets of present interest. Figure 5A shows a three-dimensional display of the MVS map in the CMT data set, in which armadillo, aardvark, gray seal, and rabbit were taken as the probes *o*, *a*, *b*, and *c*, respectively. Figure 5A shows a clear separation of Xenarthra (a blue ball), Afrotheria (yellow balls), Laurasiatheria (red balls), and Supraprimates (green balls). Figure 5B shows a three-dimensional display of the MVS map in the 22 GENES, in which sloth, golden mole, tapir, and human were taken as the probes *o*, *a*, *b*, and *c*, respectively. In this figure, the four superordinal groups were also well separated from each other. The separation was robust against the fluctuation caused by random noise. All of the 100 bootstrap samples supported this separation, in which no species were embedded into other groups. Separations were also clear in the 22 GENES and BRCA1 data sets.

Once the four major groups were separated, the bias correction of the pairwise distances was performed automatically according to the phylogenetic-inference method, first among these groups, then within the respective groups, and finally between eutherians and the outgroup (marsupials and monotremes). Consequently, the additivity rule was completely fulfilled in all possible triplets of probes. Thus, we obtained the MVS representation of the eutherian tree, in which the four major groups of placental mammals are clearly separated. Figure 5C presents a global depiction for the CMT data set (corresponding to figure 2A), in which the armadillo, dugon, whale, and mouse were taken as the probes *o*, *a*, *b*, and *c*, respectively. The orthogonal behavior in this MVS map can be reproduced by all sets of the probes, which are taken from the four superordinal groups, respectively.

The branching patterns and lengths were markedly affected by the bias correction of pairwise distances. Figure 6A gives the NJ tree before the bias correction in the CMT data. This was similar to the ML tree of a previous analysis (Arnason et al. 2002) but very different from the NJ tree (fig. 6B) after the bias correction. Almost all branches had very large bootstrap proportions after the bias correction. The low branch resolutions in figure 6A were caused by strong biases spreading to a variety of species pairs. Main differences between figures 6A and 6B were made clear by laying on colors. In figure 6A before the bias correction, the two superordinal groups of the aftrotherians and the xenarthran (armadillo) formed a sister relation in a middle clade of the eutherian tree. The rodents were scattered without comprising a monophyletic lineage, and they also did not form a sister relation with the lagomorphs. The hedgehogs and moonrat formed the first

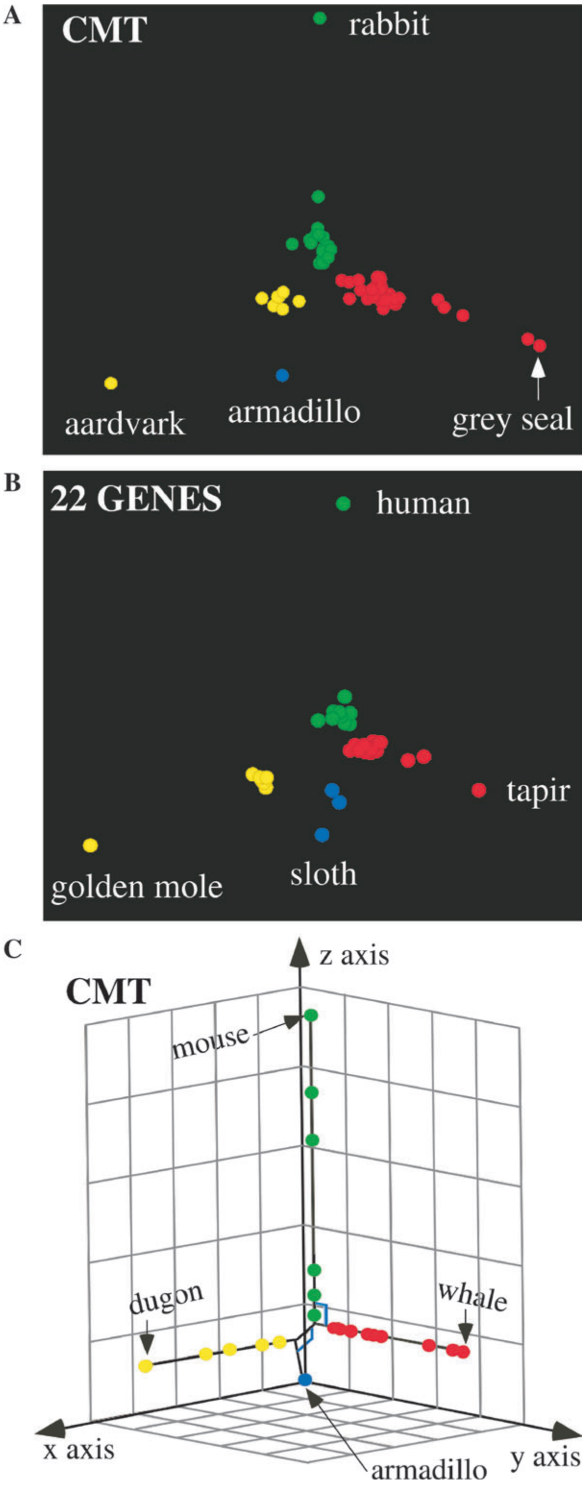

FIG. 5.—A three-dimensional display of the MVS map in the CMT was given by taking armadillo, aardvark, gray seal, and rabbit as the probes *o*, *a*, *b*, and *c*, respectively (A). A three-dimensional display for the 22 GENES was also given by taking the sloth, golden mole, tapir, and human as the probes *o*, *a*, *b*, and *c*, respectively (B). Xenarthra (a blue ball), Afrotheria (yellow balls), Laurasiatheria (red balls), and Supraprimates (green balls) were well separated from each other. All of the 100 bootstrap samples supported this separation, no species were embedded into other groups. A global depiction (C) for the CMT data after the bias correction was given by taking armadillo, dugon, whale, and mouse as the probes *o*, *a*, *b*, and *c*, respectively. Here, the additivity rule was completely satisfied.

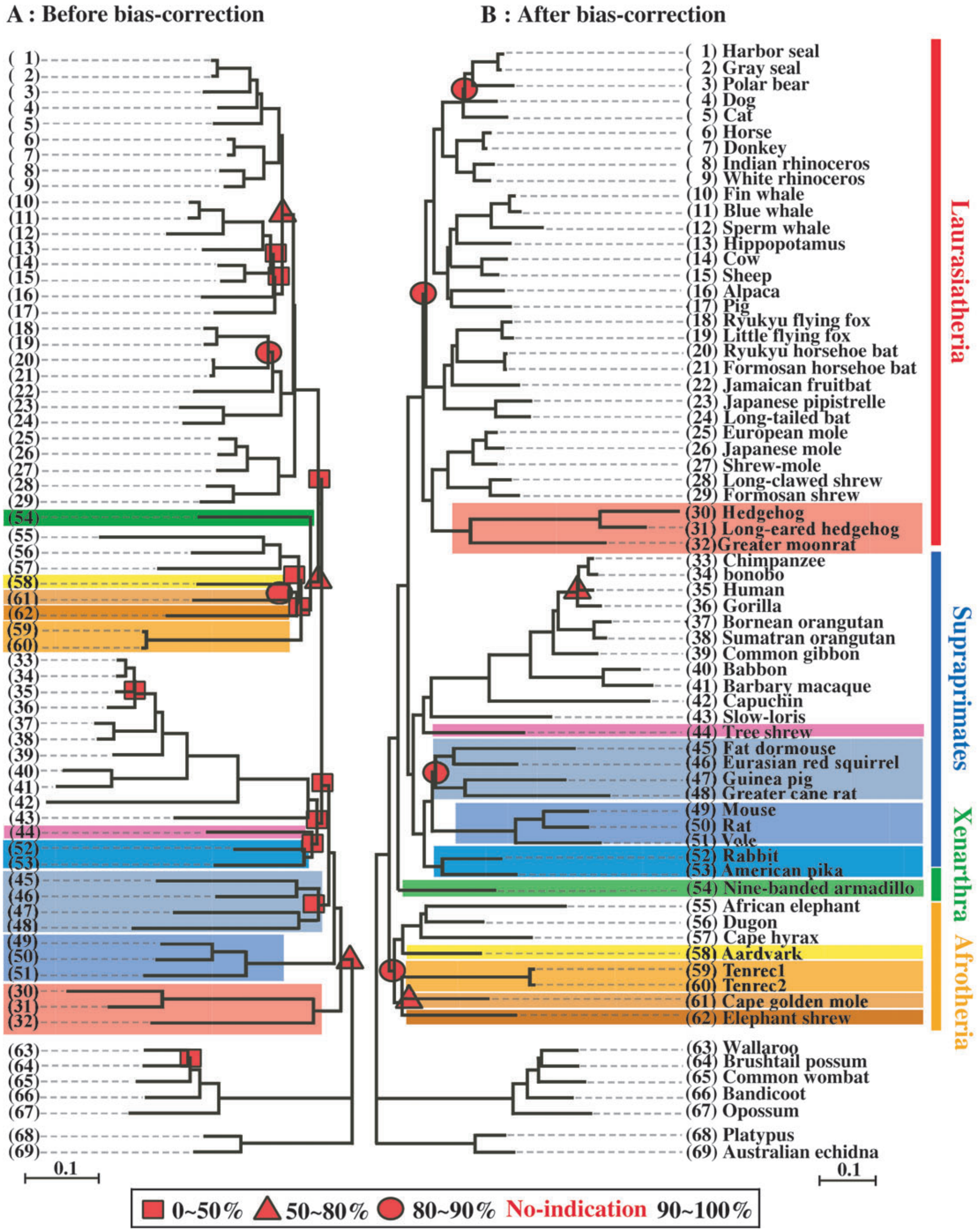

FIG. 6.—The NJ tree before the bias correction (A) of the pairwise distances in the CMT data set is compared with that after the bias correction (B). The bias correction markedly altered the branching pattern and resolution. The symbols ■, ▲, and ● and no-mark stand for the branch resolutions of 0% to 50%, 50% to 80%, 80% to 90%, and 90% to 100%, respectively.

clade of the eutherian tree. On the other hand, in figure 6B after the bias correction, the hedgehogs and moonrat entered into the laurasiatherian group to form a monophyletic lineage with the other laurasiatherian insectivores. The rodents became monophyletic, further made a sister group with the lagomorphs, and became a chief constituent of the supraprimates by finally coupling with the primates. In this way, the three superordinal groups of the afrotherians, the xenarthran, and the supraprimates comprised the first, second, and third clades of the eutherian tree, respectively. Such a superordinal relationship was reproduced by the MVS analysis of the 22 GENES and BRCA1.

This relationship is consistent with a previous ML analysis of the 22 GENES (Murphy et al. 2000) but inconsistent with previous ML analyses of the BRCA1 (Madsen et al. 2001) and CMT (Nikaido et al. 2003), which reported a sister group of the afrotherians and the xenarthrans. Other differences of our results from previous ones in the three data sets are summarized as follows. In the MVS analysis, the aardvark formed a sister group with the paenungulates (the elephant, *Sirenia*, and the hyrax) to

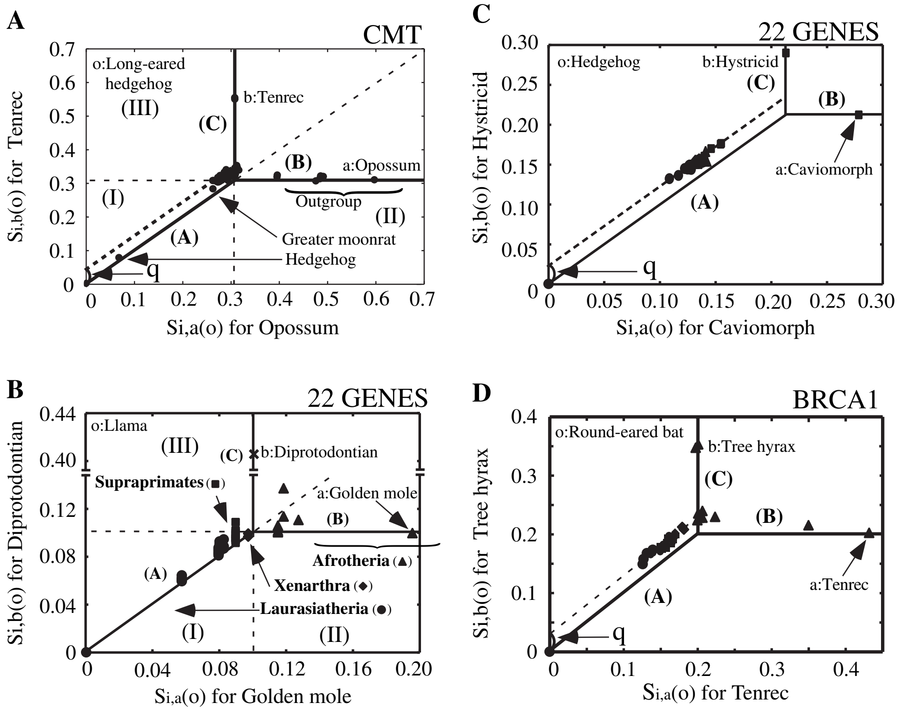

FIG. 7.—The MVS map (A) before the bias correction of the pairwise distances in the CMT showed a systematic deviation pattern similar to figure 2B and was given by taking the long-eared hedgehog (*o*), opossum (*a*), and tenrec (*b*) as the probes. Here, the majority of eutherians were located in the region III of the probe *b*. The deviations were greatly decreased by increasing only the distances between the three insectivores and outgroup by 2*q*. As a result, all laurasiatherians shifted to the region I and became distributed around the diagonal line **A** to be connected with the three insectivores. Attractions between the eutherians and outgroup were also strong in the 22 GENES, and became explicit after the bias correction within the eutherians (B). Here, some of supraprimates and Afrotherians entered into region III of the outgroup. Figure C and D show a strong attraction between the hedgehog (*o*) and caviomorph (*a*) in the 22 GENES (C), and that between round-eared bat (*o*) and tenrec (*a*) in BRCA1 (D), respectively.

reform the African ungulates, although it belonged to the African insectivorous lineage in previous analyses of the 22 GENES (Mouchaty et al. 2000) and CMT (Nikaido et al. 2003). The tree shrew and flying lemur formed the first and second clades in the base of the supraprimate lineage, whereas they were located outside the rodent and primate lineages in a previous BRCA1 analysis (Madsen et al. 2001), and they did not form a sister relation at the base of the primates lineage in a previous 22-GENES analysis (Murphy et al. 2000). The hedgehog, mole, and shrew became monophyletic within the laurasiatherian group, although it was difficult to treat them in previous CMT analyses (Waddel, Kishino, and Ota 2001; Nikaido et al. 2003). In this way, the MVS analysis led to a consistent tree reconstruction in the three representative data sets of placental mammals through the bias correction.

Discussion

We discuss the efficacy of the present bias correction by examining pairs of species with large distance modifications. First, we explain why the hedgehogs and moonrat in the CMT form the first clade of the eutherian in the NJ tree (fig. 6A) before the bias correction and a previous ML tree (Arnason et al. 2002). The MVS map (fig. 7A) before the bias correction showed a systematic deviation pattern similar to figure 2B and was given by the probes of the long-eared hedgehog (*o*), opossum (*a*), and tenrec (*b*). Here, the majority of eutherians were located in region III of probe *b*. The four-point condition combines the species of the same region into one group, whereas the three insectivores (two hedgehogs and moonrat) remained in the region I and became closer to the outgroup (marsupials and monotremes) than to the other eutherians. This observation is the main reason of the strange pattern of the hedgehogs and moonrat in the standard methods. We similarly could explain why the mouse, rat, and vole formed the second clade of the eutherians in the NJ tree (fig. 6A) and the previous ML tree (Arnason et al. 2002). On the other hand, in the MVS analysis, the major portions of the deviations were removed by increasing only the distances between the three insectivores and the outgroup in figure 7A by $2|q|$, in a similar way to the case of figure 2B. As a result, all laurasiatherians shifted to region I and became distributed around the diagonal line **A** to be connected with the three insectivores.

Attractions between the outgroup and the eutherians were also strong in the 22 GENES and became explicit after the bias correction for the eutherians (fig. 7B). Here, the attractions disappeared by enlarging only the distances between the diprotodontian (marsupial) and the eutherians so that the eutherians might be located on the additivity lines **A** and **B** of figure 7B, because the distance modification among the eutherians was already finished at this stage. Figure 7B shows that some of supraprimates and Afrotherians enter into region III of the outgroup in the four-point condition, because of the very strong attractions. We also observed many strong biases among

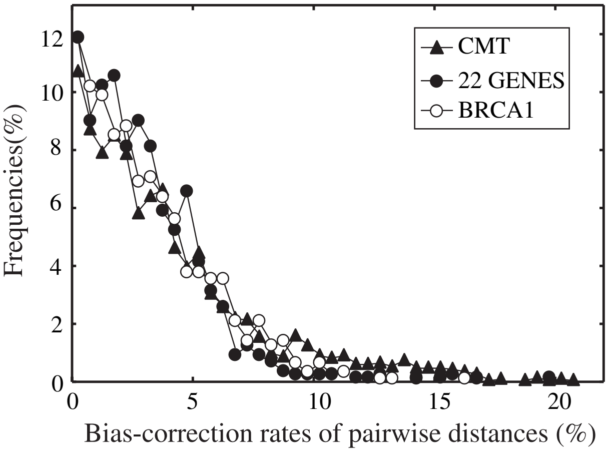

FIG. 8.—The frequencies of bias-correction rates of the pairwise distances were calculated in the CMT, 22 GENES, and BRCA1. Here, the effect of random noise was removed. The correction rates did not diminish in the 22 GENES, the sequence of which is threefold longer than that of the BRCA1.

eutherians that cause systematic deviations in the three data sets. Figures 7C and D show a strong attraction between the hedgehog (*o*) and the caviomorph (*a*) in the 22 GENES and between the round-eared bat (*o*) and the tenrec (*a*) in BRCA1, respectively. The deviations were largely removed by increasing the distance between the two species by $2q$, according to the phylogenetic-inference method.

We survey the bias correction of the initially estimated pairwise distances in the CMT, 22 GENES, and BRCA1 data sets. We calculated the modification rates (%) of the initial distances over all possible pairs of species. Figure 8 gives the frequency distribution (%) of these modification rates relative to the total pair number and show that the bias effect was spread over a wide range of species pairs in the three data sets and was very strong in specific pairs. The numbers of pairs with modification rates exceeding 10% were 201 (2,346) in the CMT, 14 (903) in the 22 GENES, and 20 (1,326) in the BRCA1 (table 1). Here, the numbers in parentheses denote the total numbers of pairs. Figure 8 and table 1 demonstrate that the bias effect does not diminish in the 22 GENES data set with small distance fluctuations caused by the very long sequence size, which is threefold longer than that of the BRCA1 data set. In the estimation of the distance modification rate, the magnitude of distance fluctuations caused by stochastic noise was excluded to estimate only the bias correction by the bootstrap resampling procedure in the phylogenetic-inference method. The average values of distance fluctuations relative to the initial distances were 4.6% for the CMT, 2.5% for the 22 GENES, and 6.7% for the BRCA1, which were much smaller than the distance-modification rates in the case of strong biases.

To confirm how directly the strong bias detected by the MVS analysis can be related to the convergent evolution, we made an alignment analysis of two sequences between the long-eared hedgehog and other species (except for the laurasiatherian insectivores) in the CMT data, because the hedgehog caused strong attractions with other species. We calculated the number of sites at which two sequences had same amino acids but were different from the consensus sequence of the eutherians. We here regarded this number as that of parallel changes between the two sequences. Then, the number of parallel changes was strongly correlated with the modified portion (given by subtracting the initial distance from the modified one) of the corresponding pairwise distance (fig. 9). The correlation coefficient gave a quite high value of $R = 0.95$, although the bias correction included other abnormal multiple substitutions.

A strength of the MVS approach is bias detection and correction without explicit assumption of the topology. The initial pairwise distances are estimated as the expected numbers of substitutions assuming some model of evolutionary process. When systematic deviations are observed in the MVS map, this is evidence that some of the estimated pairwise distances are seriously biased, most likely because of convergent evolution. The biased pairwise distances are then corrected via the minimum-modification criterion. Our procedure detected significant systematic deviations in the three large data sets of placental mammals and eliminated inconsistency among data sets in the estimated relationships among eutherian orders. This suggests the validity of our method of bias correction. When a large data set is analyzed, bias detection and correction are of primary importance in phylogeny inference because random noise becomes negligible and biases can lead to significant inconsistency (Phillips, Delsuc, and Penny 2004). The large biases detected here should be interpreted as a warning. Conventional methods of phylogeny inference strongly rely upon probabilistic models of sequence evolution. When the actual amount of sequence convergence between some lineages far exceeds the amount predicted by these models,

Table 1
Bias Correction Rates in the Pairwise Distances

Bias Correction Rates	22 GENES	BRCA1	CMT
The upper 1% point (%)	10.6 ± 2.3 (9)	10.4 ± 3.6 (13)	18.0 ± 4.1 (23)
The upper 5% point (%)	5.6 ± 2.5 (45)	8.1 ± 3.3 (66)	13.9 ± 5.2 (117)
The median point (%)	1.8 ± 2.5 (452)	2.6 ± 2.9 (663)	5.6 ± 4.7 (1,173)
The total number of pairs	903	1,326	2,346

NOTE.—The average values and standard deviations of bias correction rates in the input pairwise distances were estimated by using data sets in a number of bootstrap resamplings. The upper 1% value, the upper 5% value, and the median of these rates are given for 22 GENES, BRCA1, and CMT. The parentheses denote the numbers of species pairs that have larger values than the correction rate at the upper 1%, upper 5%, or median point. The rates become important in larger values than 5% because then the NJ method is likely to give different topologies.

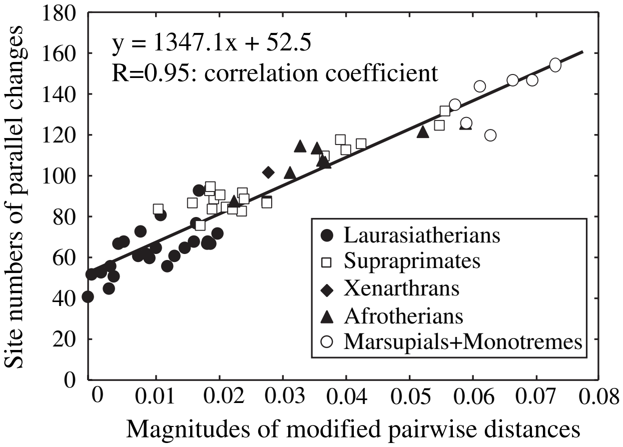

FIG. 9.—The site numbers of parallel changes between the long-eared hedgehog and other species in the CMT sequences were strongly correlated with the modified portions of the corresponding pairwise distances. These numbers were counted by assuming the consensus sequence of the eutherians as the ancestor.

phylogenetic inference via conventional methods could be positively misleading. Collecting an entire genome of sequence data for each taxon of interest and then analyzing the data via a conventional model-based approach could yield topologies that are strongly supported but biologically incorrect. An advantage of the MVS approach is that the cause and nature of convergence between lineages need not be known and explicitly incorporated into an evolutionary model before data analysis. Instead, our technique detects and adjusts for convergence between lineages by relying upon the geometry of the MVS map. A forthcoming application of the MVS approach is to reexamine the divergence times (Kumar and Hedges 1998; Springer et al. 2003) of modern orders and the multiple origins (Springer et al. 1997; Stanhope et al. 1998) of insectivores, which have given discrepancies between molecular and paleontological studies.

Acknowledgments

We thank Jeffrey L. Thorne for his helpful comments and insights regarding the method. This work was supported by BIRD of the Japan Science and Technology Agency (JST) and Japan Society for the Promotion of Science (JSPS).

Literature Cited

Adachi, J., and M. Hasegawa. 1996. MOLPHY: programs for molecular phylogenetics. Version 2.3. Institute of Statistical Mathematics, Tokyo.

Alroy, J. 1999. The fossil record of North American mammals: evidence for a Paleocene evolutionary radiation. Syst. Biol. **46**:107–118.

Archibald, J. D., and D. H. Deutschman. 2001. Quantitative analysis of the timing of the origin and diversification of extant placental orders. J. Mammal. Evol. **8**:107–124.

Arnason, U., J. A. Adegoke, K. Bodin, E. W. Born, Y. B. Esa, A. Gullberg, M. Nilsson, R. V. Short, X. Xu, and A. Janke. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. Proc. Natl. Acad. Sci. USA **99**:8151–8156.

Benton, M. J., and F. J. Ayala. 2003. Dating the tree of life. Science **300**:1698–1700.

Cann, R. L., M. Stoneking, and A. C. Wilson 1987. Mitochondrial DNA and human evolution. Nature **325**: 31–36.

Cao, Y., M. Fujiwara, M. Nikaido, N. Okada, and M. Hasegawa. 2000. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene **259**:149–158.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. Dayhoff, ed. Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, DC.

Easteal, S. 1999. Molecular evidence for the early divergence of placental mammals. Bioessays **21**:1052–1058.

Felsenstein, J. 1996. PHYLIP (phylogeny inference package). Version 3.572. University of Washington, Seattle.

Holland, B. R., K. T. Huber, A. Dress, and V. Moutton. 2002. δPlots: a tool for analyzing phylogenetic distance data. Mol. Biol. Evol. **19**:2051–2059.

Hudelot, C., V. Gowri-Shankar, H. Jow, M. Rattray, and P. G. Higgs. 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol. Phylogenet. Evol. **28**:241–252.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. in Biosci. **8**:275–282.

Jukes, T. H., and C. R. Cantor. 1969. Pp. 21–132 *in* H. N. Munro, ed. Evolution of protein molecules: mammalian protein metabolism. Academic Press, New York.

Kitazoe, Y., Y. Kurihara, Y. Narita, Y. Okuhara, A. Tominaga, and T. Suzuki. 2001. A new theory of phylogeny inference through construction of multidimensional vector space. Mol. Biol. Evol. **18**:812–828.

Kocher, T. D., J. A. Conroy, K. R. McKaye, and J. R. Stauffer. 1993. Similar morphologies of cichlid fish in lakes Tanganyika and Malawi are due to convergence. Mol. Phyl. Evol. **2**:158–165.

Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917–920.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature **409**:610–614.

Madsen, O., D. Willemsen, B. M. Ursing, U. Arnason, and W. W. de Jong. 2002. Molecular evolution of the mammalian alpha 2B adrenergic receptor. Mol. Bio. Evol. **19**: 2150–2160.

Mouchaty, S., A. Gullberg, A. Janke, and U. Arnason. 2000. The phylogenetic position of the Talpidae within Eutheria based on analysis of complete mitochondrial sequences. Mol. Bio. Evol. **17**:60–67.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature **409**:614–618.

Nikaido, M., Y. Cao, M. Harada, N. Okada, and M. Hasegawa. 2003. Mitochondrial phylogeny of hedgehogs and monophyly of Eulipotyphla. Mol. Phylogenet. Evol. **28**:276–284.

Novacek, M. J. 1992. Mammalian phylogeny: shaking the tree. Nature **356**:121–125.

———. 1999. 100 million years of land vertebrate evolution: the Cretaceous-Early Tertiary Transition. Ann. MO Bot. Gard. **86**:230–258.

O'Brien, S. J., M. Menotti-Raymond, W. J. Murphy, W. G. Nash, J. Wienberg, R. Stanyon, N. G. Copeland, N. A. Jenkins,

J. E. Womack, and J. A. Marshall Graves. 1999. The promise of comparative genomics in mammals. Science **286**:458–462.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. **21**:1455–1458.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Springer, M. S., G. C. Cleven, O. Madsen, W. W. de Jong, V. G. Waddell, H. M. Amrine, and M. J. Stanhope. 1997. Endemic African mammals shake the phylogenetic tree. Nature **388**:61–64.

Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc. Natl. Acad. Sci. USA **100**:1056–1061.

Stanhope, M. J., V. G. Waddell, O. Madsen, W. de Jong, S. B. Hedges, G. C. Cleven, D. Kao, and M. S. Springer. 1998. Molecular evidence for multiple origins of Insectivora and a new order of endemic African insectivore mammals. Proc. Natl. Acad. Sci. USA **95**:9967–9972.

Waddel, P. J., H. Kishino, and R. Ota. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Informat **12**:141–154.

Wolf, Y. I., I. B. Rogozin, N. V. Grishin, and E. V. Koonin. 2002. Genomes trees and the tree of life. Trends Genet. **18**:472–479.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable sites: approximate methods. J. Mol. Evol. **39**:306–314.

Yang, Z., and S. Kumar. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13**:650–659.

Zhang, J., and S. Kumar. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol. Biol. Evol. **14**:527–536.

Arndt von Haeseler, Associate Editor

Accepted November 9, 2004