# Estimating Phylogenies from Lacunose Distance Matrices: Additive is Superior to Ultrametric Estimation

*Pierre-Alexandre Landry,* François-Joseph Lapointe,* and John A. W. Kirsch†*

*Département de sciences biologiques, Université de Montréal; and †University of Wisconsin Zoological Museum

Lapointe and Kirsch (1995) have recently explored the possibility of reconstructing phylogenetic trees from lacunose distance matrices. They have shown that missing cells can be estimated using the ultrametric property of distances, and that reliable trees can be derived from such filled matrices. Here, we extend their work by introducing a novel way to estimate distances based on the four-point condition of additive matrices. A simulation study was designed to assess whether the additive procedure is superior to the ultrametric one in recovering distances randomly deleted from complete distance matrices. Our results clearly indicate that the topologies and branch lengths of the trees derived from matrices which were estimated additively are better recovered than those of trees derived from matrices estimated ultrametrically; the original distances are also better recovered with the additive procedure. Except in the case of small matrices with many missing cells for which both methods perform equally well, the additive is generally superior to the ultrametric method for estimating missing cells in distance matrices prior to phylogenetic reconstruction.

## Introduction

There are many limitations in the various existing algorithms for phylogenetic reconstruction based on distances (for a review, see Swofford and Olsen 1990). For example, the generation of a distance matrix suitable for the Fitch-Margoliash method (Fitch and Margoliash 1967) or neighbor-joining analysis (Saitou and Nei 1987) using molecular techniques like DNA-hybridization (Werman, Springer, and Britten 1990) or comparative serology (Maxson and Maxson 1990) requires $n^2$ comparisons among $n$ taxa. Consequently, the majority of such studies are usually limited to few taxa and often lead to lacunose (i.e., incomplete) matrices not directly or immediately amenable to phylogenetic analyses. However, according to a recent article by Lapointe and Kirsch (1995), it is not necessary to have complete matrices to obtain robust phylogenies, since it is possible to estimate missing cells prior to phylogenetic reconstruction by relying on the properties of path-length matrices representing trees. Indeed, there exists a one-to-one correspondence (Hartigan 1967; Johnson 1967) between dendrograms and matrices satisfying the ultrametric inequality, which states that for every triplet of objects {i, j, k}, the two largest distances among the three possible ones are equal:

$$d(i, j) \leq \max[d(i, k); d(j, k)], \text{ for every } i, j, k. \quad (1)$$

As originally proposed by De Soete (1984*a*, 1984*b*), the ultrametric property can be used to fill incomplete path-length matrices, as well as any other types of dissimilarity matrices. De Soete showed that reliable trees could be derived from matrices in which missing distances were estimated ultrametrically, even when adding noise, in some cases with up to 50% of cells missing. Using De Soete's (1984*b*) method, Lapointe and Kirsch (1995) have shown, in another series of simulations, that the ultrametric procedure could be used to fill lacunose DNA-hybridization matrices and still recover correct phylogenies (i.e., the same trees as recovered from complete data). Examples of robust phylogenies were obtained with 35%, 60%, and up to 61% of cells missing, respectively, by Kirsch and Palma (1995), Lapointe and Kirsch (1995), and Bleiwess, Kirsch, and Shafi (1995). In this paper, we show that the ultrametric procedure is not the only method available to fill matrices and we propose an improvement based on the additive property of path-length distances.

Key words: distance matrices, DNA hybridization, missing data, phylogenetic reconstruction, simulation study.

Address for correspondence and reprints: P.-A. Landry, Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale centre-ville, Montréal, Québec, H3C 3J7. E-mail: landryp@ere.umontreal.ca.

*Mol. Biol. Evol.* 13(6):818–823. 1996

 ISSN: 0737-4038

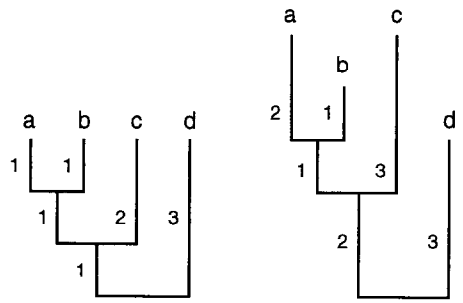
## Methods

A matrix is said to be additive when it meets the following inequality, the so-called four-point condition (Buneman 1971):

$$d(i, j) + d(k, l) \leq \max[d(i, k) + d(j, l); d(i, l) + d(j, k)], \text{ for every quadruplet } i, j, k, l. \quad (2)$$

Then, it is easily shown that an ultrametric matrix represents a particular type of additive matrix, corresponding to a tree for which all terminal nodes are equidistant from the root (Sibson 1972). Because ultrametric distances are more constrained than additive distances, however, one loses accuracy in the estimation of missing additive distances when relying on the ultrametric property. On the other hand, one can always recover ultrametric distances using the four-point condition. Nevertheless, the additive estimation is computationally much more demanding since it requires five known distances among the six possible for any quadruplet {i, j, k, l}; the ultrametric inequality only needs two out of three distances among any triplet {i, j, k}. In the best situation, only $n - 1$ distances are sufficient to fill an ultrametric matrix, whereas $2n - 3$ are needed to estimate a matrix additively (see Lapointe and Legendre 1991, 1992).

To illustrate the difference between the ultrametric and the additive estimation procedures, let us consider




| | a | b | c | d |
|---|---|---|---|---|
| a | 0 | 2 | 4 | 6 |
| b | 2 | 0 | 4 | 6 |
| c | 4 | 4 | 0 | 6 |
| d | 6 | 6 | 6 | 0 |

| | a | b | c | d |
|---|---|---|---|---|
| a | 0 | 3 | 6 | 8 |
| b | 3 | 0 | 5 | 7 |
| c | 6 | 5 | 0 | 8 |
| d | 8 | 7 | 8 | 0 |

FIG. 1.—Examples of hypothetical (*a*) ultrametric and (*b*) additive trees along with their corresponding path-length matrices. In each matrix d(a, c) is assumed to be missing and is filled to illustrate the efficiency of the different estimation procedures. See text for details.

two weighted trees, along with their corresponding path-length matrices (fig. 1). Both matrices satisfy the four-point condition, but the first one (fig. 1*a*) is also ultrametric. Let us pretend that d(a, c) is missing in both matrices. If the ultrametric property were used to recover that missing value, one would need to apply it to every triplet of objects involving the pair {a, c}, taking in each case the maximum of the two known distances as an estimate of d(a, c). Two such triplets are distinguishable here (i.e., {a, b, c} and {a, c, d}) and two estimates of d(a, c) can be computed as (max[d(a, b); d(b, c)] = 4; max[d(a, d); d(c, d)] = 6), the minimum of which is taken as the final estimate of d(a, c). In that case, the estimated distance, d(a, c) = 4, is identical to the real value (fig. 1*a*). However, when the same procedure is applied to the second matrix (fig. 1*b*), which does not satisfy the ultrametric condition, the final estimate of d(a, c), computed as above, is different from the actual distance (i.e., d(a, c) = 5 instead of 6). Therefore, nonultrametric additive matrices will not always be correctly estimated using the ultrametric inequality, even when the data are perfectly clean.

To see whether the reverse case would hold or not, we have used the four-point condition to recover the missing distance d(a, c) in both matrices presented in figure 1. Here, one needs a quadruplet of objects to estimate that missing cell; the only one possible in our case is {a, b, c, d}. The procedure consists of computing the two sums of distances not involving the missing cell (i.e., d(a, b) + d(c, d); d(b, c) + d(a, d)). The maximum of the two sums is returned and the fifth distance (i.e., d(b, d)) is subtracted from it to compute the final estimate of d(a, c). Doing so for both matrices, we get correct estimates in the additive case (fig. 1*b*), as well as for the ultrametric situation (fig. 1*a*). Thus, in accordance with theory, the additive estimation procedure is more general than the ultrametric method used by De Soete (1984*b*) and Lapointe and Kirsch (1995).

To know whether the results obtained in a perfect case (e.g., fig. 1) can be extended to real matrices, we have compared the accuracy of the estimates obtained by the two different procedures using a simulation design. We wanted to assess whether the additive method is always superior, and thus have dealt with DNA-hybridization matrices which, unlike perfect data, are not additive, not symmetrical, and never free of noise (see Lapointe and Kirsch 1995). If the new estimation procedure proves to be more efficient with such messy data, it is likely to be superior with any other type of data.

To compare the performance of the two methods, we used the same experimental design and matrices as Lapointe and Kirsch (1995); that is, (1) lacunose matrices are generated by removing known distances from complete matrices, (2) missing cells are estimated, (3) a phylogeny is reconstructed, and (4) results of filled matrices are compared with those derived from complete matrices, using either one of the two estimation procedures. Five DNA matrices of varying size (i.e., $n$ = 7: Dasyuridae [Marsupialia], Kirsch et al. 1990; $n$ = 9: hummingbirds [Apodiformes], Bleiweiss, Kirsch, and Matheus 1994; $n$ = 11: Didelphidae [Marsupialia], Kirsch, Dickerman, and Reig 1996; $n$ = 13: Diprotodontia [Marsupialia], Springer and Kirsch 1991; $n$ = 15: Pteropodidae [Chiroptera], Kirsch et al. 1995) were used in our simulations, and increasing percentages of missing cells ($P$ = {0.1, 0.2, 0.3, 0.4, 0.5, 0.6}) were considered by deleting reciprocal distances (i.e., d(i, j) and d(j, i)) from the square matrices. One hundred simulations of lacunose matrices were performed for each $P$ and each matrix. In a first series of simulations, matrices were filled using the ultrametric inequality. Then, in a second series, we repeated the simulations with the same matrices (removing exactly the same cells from the corresponding matrices), this time using the four-point condition to fill the empty cells (computer program available from the authors upon request). Deletion of the same distances in both series was deemed necessary to control for sampling errors due to the randomization. The difference in estimations can thus be explained only with respect to the mathematical property called for in the simulations. As stated above, one major distinction between the ultrametric and additive inequalities is the number of known distances required to estimate missing cells (i.e., 2/3 vs. 5/6). When the number of missing cells in a matrix is large, the additive estimation becomes more difficult to compute than the ultrametric one. In some cases, no additive estimation can be calculated and we have to rely on the ultrametric property to estimate the minimum number of cells required to call the four-point condition. To do so, an iterative procedure is used by calling the ultrametric inequality to fill the first missing cell encountered in the matrix, going back to the additive inequality recursively until the matrix is filled.

Estimates obtained by the two different methods were compared at three distinct levels. First, we correlated the filled matrices with the complete DNA-hybridization tables to quantify distance recovery; this allowed us to evaluate how well experimental values could be

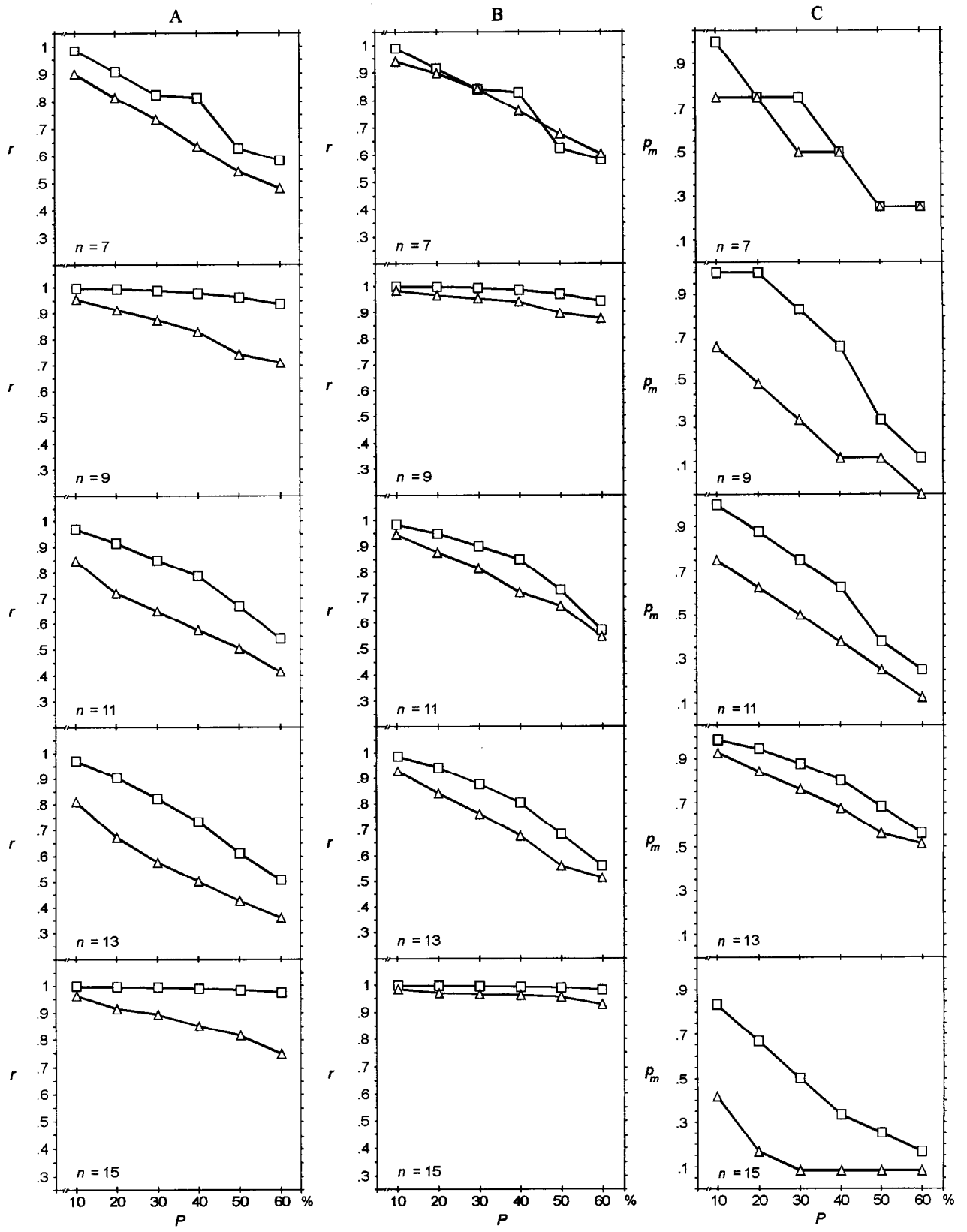



FIG. 2.—Metric and topological recovery values obtained for real matrices of increasing size ($n$ = 7, 9, 11, 13, 15) and for varying percentages of deleted and estimated cells ($P$ = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6), using the ultrametric inequality (open triangles) and the four-point condition (open squares). (*A*) Distance-matrix correlations ($r$). (*B*) Path-length correlations ($r$). (*C*) Standardized partition metric values ($p_m$).

estimated by the two methods. Second, we computed path-length correlations between matrices associated with the trees derived from the original and estimated data; this was done to assess whether the estimation procedure introduces a bias in the path-length distances recovered by the phylogeny reconstruction algorithm. Computations were performed using FITCH (G, S, P = 0 and no negative branches allowed) from the PHYLIP package (Felsenstein 1993). Finally, we confronted the original tree with those derived from the estimated matrices using the standardized partition metric (Robinson and Foulds 1981), this to quantify topological recovery (computations were performed in PAUP, Swofford 1993). In order to compare topological recovery values computed over matrices of different sizes, the partition metric was standardized by its maximal possible value ($2n - 6$) for each matrix (Steel and Penny 1993). The 1-complement of this standardized statistic was then taken as an index of topological similarity, with a maximal value of 1.0 for topologically identical trees, and a value of 0.0 for the most different possible pairs of trees. The average of correlation values and the median of the standardized partition metric values were then recorded for comparison purposes, as in Lapointe and Kirsch (1995).

## Results

The results of the simulations comparing the two estimation methods are presented in figure 2. One of the most important features evident from these comparisons is that the additive procedure provides better results (on average) than the ultrametric method in almost every case. We do observe some variations among matrices, however. For one, the correlations between actual and estimated distance matrices are quite variable (fig. 2*A*); some matrices are well estimated for up to a large proportion of missing cells ($n = 9$ and 15), while others exhibit a steeper decline in correlation values as more cells are deleted ($n = 7$, 11, and 13). There seems to be no direct relation between the size of a matrix and the accuracy of the estimations of missing DNA-hybridization distances, in agreement with the conclusion of Lapointe and Kirsch (1995).

For path-length correlations (fig. 2*B*) we observe an almost uniform convergence of the two curves; the additive procedure is still on top, except for one case where the ultrametric results indicate better estimations ($n = 7$, $P > 0.30$). This convergence in the performance of the competing methods is a direct consequence of the phylogenetic reconstruction algorithm that must return a tree (i.e., a path-length distance matrix). Because path-length distances are more constrained than input distances (i.e., they must satisfy the four-point condition), the correlations between recovered path-length matrices are generally larger than those between the estimated and complete DNA-hybridization matrices (see Lapointe and Kirsch 1995). The difference between trees based on matrices filled using different estimation procedures also tends to vanish for the same reason.

**Table 1**
**Proportion of Individual Cases for Which the Additive Estimation Procedure was More Accurate than the Ultrametric Method**

| | $n = 7$ | $n = 9$ | $n = 11$ | $n = 13$ | $n = 15$ |
|---|---|---|---|---|---|
| Distance correlations..... | 0.765 | 0.992 | 0.908 | 0.927 | 1.000 |
| Path-length correlations... | 0.645 | 0.958 | 0.800 | 0.845 | 0.998 |
| Partition metric ..... | 0.697 | 0.952 | 0.882 | 0.858 | 0.978 |

Interestingly enough, neither the correlation between the original distance matrices (fig. 2*A*) nor that between path-length matrices (fig. 2*B*) is a good predictor of topological recovery (fig. 2*C*). Here again, the additive method is better than, or as good as ($n = 7$), the ultrametric procedure. However, topological-recovery values can be quite different among matrices, including cases of DNA-hybridization tables showing similar correlation values but very distinct topological results between matrices estimated ultrametrically or additively ($n = 9$ and 15). A similar pattern was also observed by Lapointe and Kirsch (1995), who showed that trees with short branches are more vulnerable to topological errors than trees with relatively longer internodes, even though both types may exhibit similar path-length correlations for a given percentage of missing cells.

These results clearly demonstrate that the additive method is by far superior to the ultrametric procedure in estimating missing cells in distance matrices and in recovering reliable phylogenies. However, closer scrutiny reveals that this is not *always* true. In contrast with average recovery values presented in figure 2, table 1 displays the relative performance of the additive method over the ultrametric one for each of the individual matrices estimated. It shows that for some matrices (e.g., $n = 15$) the additive procedure provides better results in almost every instance considered (i.e., relative performance ranging from 97% to 100%, depending on the recovery index computed). In other situations (e.g., $n = 7$), the same approach is superior in only 2/3 (or at most 3/4) of the matrices estimated.

Figure 3 shows the average number of times the ultrametric procedure was called for to fill matrices of different sizes, for varying *P*. One can notice from those results that the ultrametric estimation procedure was never required for up to 40% of cells missing, except for the smaller matrix ($n = 7$). The need for ultrametric estimations is a function of the size of the matrix; indeed, larger matrices have more possible quadruplets, thus increasing the probability of finding one situation suitable for additive estimation (i.e., one quadruplet of objects for which 5/6 of the distances are known). At most, three ultrametric iterations were needed to fill matrices (for $n = \{7$ and $9\}$), whereas only two ($n = 11$) or one iteration was used in the larger cases ($n = \{13$ and $15\}$), and only when 60% of the cells were missing.

## Discussion

In the light of the simulation results, one has the right to ask whether the four-point condition should be

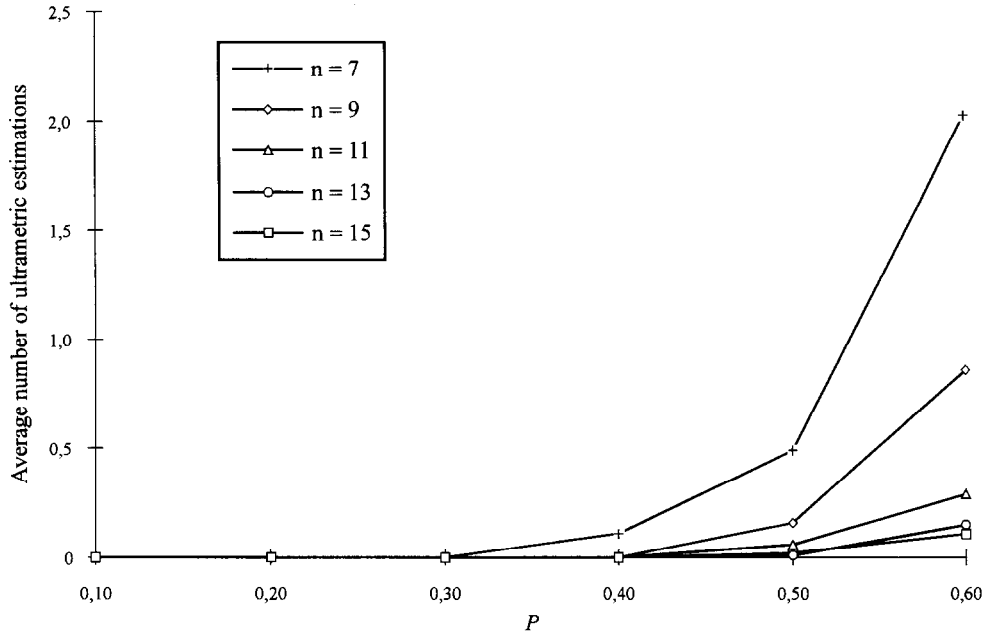



FIG. 3.—Average number of ultrametric estimations performed in the additive procedure for increasing values of $P$ and different matrices.

used all the time. Are there any situations for which the ultrametric method would return better estimates? We already know that for a perfectly ultrametric matrix, it is generally possible to estimate a missing distance with the same accuracy using either the additive or the ultrametric property. Contrariwise, in the case of additive matrices, only the four-point condition gives us correct estimates; the ultrametric method returns an approximation of the actual distance (see fig. 1). However, there exists one notable exception where both techniques fail to recover the actual distance; it involves missing distances between terminal sister taxa. In such cases there is no existing technique to estimate the missing distance correctly. In the best case of an ultrametric matrix without experimental errors, the tree derived from an estimated matrix will bear a trichotomy, lowering the node between the sister taxa down to the next level (see Lapointe and Kirsch 1995). The same problem exists for additive matrices, however. We have observed that the ultrametric procedure can return better estimates (i.e., a closer estimate of a missing distance between terminal sister taxa) in some specific cases (e.g., this happens when the missing distance is much smaller than every other distance in the matrix, including those among other sister taxa). However, when more than one cell is missing and other missing distances are not among terminal sister taxa, the additive estimation is usually better. Moreover, in the absence of outside information, it is impossible to know a priori whether a given missing cell represents a distance between terminal pairs. Therefore, when the "real" tree is not known, it is always wiser to bet on the additive method.

There is another particular case when the ultrametric can return better estimates. Depending on the distances deleted and the quality of the data, the additive estimation of missing cells can result in negative distances because of the subtraction involved. Interestingly, for complete DNA-hybridization matrices, it is possible to obtain negative distances even among real data (Bledsoe and Sheldon 1989). In some very rare instances (11 out of 3,000 simulated matrices, and only for $P > 0.30$ and $n < 13$), negative distances were estimated with the additive procedure. In those 11 special cases, the two methods showed equal capacity to recover the correct topology, neither of them returning better results (the same remark also applies to distance correlations). Thus, even when negative distances are provided by the additive method, the topology recovered can be as good as the one estimated ultrametrically. Let us reiterate, however, that these cases of negative distances are extremely rare and that the additive method should be preferred. For the purpose of experimental estimation, there is no doubt that the additive estimation procedure is in most cases superior to the ultrametric one.

## LITERATURE CITED

BLEDSOE, A. H., and F. H. SHELDON. 1989. The metric properties of DNA–DNA hybridization dissimilarity measures. Syst. Zool. **38**:93–105.

BLEIWESS, R., J. A. W. KIRSCH, and J. C. MATHEUS. 1994. DNA hybridization evidence for subfamily structure among hummingbirds. Auk **111**:8–19.

BLEIWEISS, R., J. A. W. KIRSCH, and N. SHAFI. 1995. Confirmation of a portion of the Sibley-Ahlquist "tapestry." Auk **112**:87–97.

BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pp. 387–395 *in* F. R. HUDSON, D. G. KENDALL, and P. TAUTU, eds. Mathematics in archeological and historical sciences. Edinburgh University Press, Edinburgh.

DE SOETE, G. 1984*a*. Utrametric tree representations of incomplete dissimilarity data. J. Classif. **1**:235–242.

———. 1984*b*. Additive-tree representations of incomplete dissimilarity data. Qual. Quantity **18**:387–393.

FELSENSTEIN, J. 1993. PHYLIP: phylogeny inference package, version 3.5c. Distributed by the author. University of Washington, Seattle.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic tree. Science **155**:279–284.

HARTIGAN, J. A. 1967. Representation of similarity matrices by trees. J. Am. Stat. Assoc. **62**:1140–1158.

JOHNSON, S. C. 1967. Hierarchical clustering schemes. Psychometrika **32**:241–254.

KIRSCH, J. A. W., A. W. DICKERMAN, and O. A. REIG. 1996. DNA/DNA hybridization studies of carnivorous marsupials. IV. Intergeneric relationships of the opossums (Didelphidae). Marmosiana **1**:(in press).

KIRSCH, J. A. W., T. F. FLANNERY, M. S. SPRINGER, and F.-J. LAPOINTE. 1995. Phylogeny of the Pteropodidae (Mammalia: Chiroptera) based on DNA hybridization, with evidence for bat monophyly. Aust. J. Zool. **43**:557–582.

KIRSCH, J. A. W., C. KRAJEWSKI, M. S. SPRINGER, and M. ARCHER. 1990. DNA/DNA hybridisation studies of carnivorous marsupials. II. Relationships among dasyurids (Marsupialia). Aust. J. Zool. **38**:673–696.

KIRSCH, J. A. W., and R. E. PALMA. 1995. DNA/DNA hybridization studies of carnivorous marsupials. V. A further estimate of relationships among opossums. Mammalia **59**: 403–425.

LAPOINTE, F.-J., and J. A. W. KIRSCH. 1995. Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. Mol. Biol. Evol. **12**:266–284.

LAPOINTE, F.-J., and P. LEGENDRE. 1991. The generation of random ultrametric matrices representing dendrograms. J. Classif. **8**:177–200.

———. 1992. A statistical framework to test the consensus among additive trees (cladograms). Syst. Biol. **41**:158–171.

MAXSON, L. R., and R. D. MAXSON. 1990. Proteins II: immunological techniques. Pp. 127–155 *in* D. M. HILLIS and C. MORITZ, eds. Molecular systematics. Sinauer, Sunderland, Mass.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

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

SIBSON, R. 1972. Order invariant methods for data analysis. J. R. Stat. Soc. Ser. B **34**:311–349.

SPRINGER, M. S., and J. A. W. KIRSCH. 1991. DNA hybridization, the compression effect, and the radiation of diprotodontian marsupials. Syst. Zool. **40**:131–151.

STEEL, M. A., and D. PENNY. 1993. Distributions of tree comparison metrics—some new results. Syst. Biol. **42**:126–141.

SWOFFORD, D. L. 1993. PAUP: phylogenetic analysis using parsimony, version 3.1.1. Distributed by the Illinois Natural History Survey, Champaign, Ill.

SWOFFORD, D. L., and G. J. OLSEN. 1990. Phylogeny reconstruction. Pp. 411–501 *in* D. M. HILLIS and C. MORITZ, eds. Molecular systematics. Sinauer, Sunderland, Mass.

WERMAN, S. D., M. S. SPRINGER, and R. J. BRITTEN 1990. Nucleic acids I: DNA-DNA hybridization. Pp. 204–249 *in* D. M. HILLIS and C. MORITZ, eds. Molecular systematics. Sinauer, Sunderland, Mass.

Stanley Sawyer, reviewing editor

Accepted March 14, 1996