Pacific Symposium on Biocomputing 6:571-582 (2001)

A TREE OBSCURED BY VINES: HORIZONTAL GENE TRANSFER AND THE MEDIAN TREE METHOD OF ESTIMATING SPECIES PHYLOGENY

JUNHYONG KIM[1,2,3], BENJAMIN .A. SALISBURY[1]
[1]*Department of Ecology and Evolutionary Biology,*
[2]*Department of Molecular, Cellular, and Developmental Biology,*
[3]*Department of Statistics,*
Yale University
New Haven, CT 06520, USA
junhyong.kim@yale.edu, ben@aya.yale.edu

A phylogeny is a tree graph representation of genealogical relationships between biological objects. It is of general interest to estimate the phylogeny of whole organisms (species trees) using bio-molecular sequences. When multiple sequences are available for each organism such as with whole genome data, individual phylogenies estimated by each molecule (gene trees) may not be concordant. The lack of concordance may be due to actual biological mechanisms such as horizontal transfer of the molecules. Here, we present a new phylogeny estimation method designed to estimate the species tree despite such horizontal transfer. It uses the idea that horizontal transfer distorts distance relationships between pairs of species but a median estimate of the distances is robust to such distortions. We demonstrate the utility of our method using a simulation study.

1 Introduction

An important view of biological organization is that it is fundamentally based on a bifurcating descent-with-modification process. Whether at the level of whole organisms or at that of protein families, it is thought that bio-diversity is generated through a tree graph where the vertices represent replication and the edges represent modification (for more precise descriptions see [1,2]). Such graphs are called evolutionary trees or phylogenies, and their estimation is crucial to a wide range of basic and applied biological problems[3]. A large volume of literature exists about various tree estimation algorithms (see [4,5,6]). Traditionally, phylogeny estimation involved using data from morphological measurements of the organisms such as presence and absence of specific traits (e.g., wings), counts (e.g., number of appendages), and states (e.g., eye color). However, with the availability of comparative molecular sequence data e.g., RDP[7], phylogenies have been increasingly estimated using bio-molecular sequences that have advantages in terms of quantity of information and putative simpler models of evolution.

When we estimate organismal phylogenies with a particular molecular sequence, we might tentatively assume that the inferred ancestor-descendent relationships of the molecule also represent the ancestor-descendent relationships of the organism. However, when multiple molecules are available for tree inference, it

is not uncommon to observe differences in the estimated trees. This problem has been called the gene tree-species tree problem (e.g., [8,9,10]) where gene tree refers to the inferred genealogical relationships of individual bio-molecules and the species tree refers to the genealogical relationships of the organism. The differences in the gene trees may be due to statistical errors, but it may also be due to actual biological phenomena that confound the genealogical relationships of different molecular components and those of the organisms. There are three broad classes of biological phenomena causing deviation of the gene trees from each other and the species tree: horizontal transfer, ancestral lineage sorting, and gene duplication and loss (see [10]). Figure 1 depicts these phenomena.

It is believed that the discordance between gene trees and species trees becomes exacerbated when the common ancestor to the current taxa is old. This has been seen as a major impediment to estimating deep divergences such as the tree of all life and has led to many controversies including proposals that such a tree does not exist[11-15]. Regardless of whether a unique tree representation exists, it is still reasonable to think that some tree representation is appropriate either as the representation of a presumed unique species phylogeny or as a representation of the "common mode" of genomic evolution. The estimation of such species phylogenies from multiple molecular sequences is expected to become increasingly important as the amount of genomic information increases. Several methods exist in the literature to estimate species trees from gene trees (e.g., [9,16,17,18]). However, these methods are either difficult to apply when a large number of molecules are available or do not address all the conflicts depicted in Figure 1. Here, we suggest a new algorithm designed specifically to estimate species trees from a large collection of gene trees and we evaluate its performance using simulation studies.

2 The median tree algorithm

A phylogenetic tree is a tree graph $T = \{V, E\}$, where V is the vertex set and E is the edge set. Degree one vertices (leaves) are assumed to represent present day organisms/genes and vertices of higher degree represent unobserved ancestral organisms/genes. More commonly the degree one vertices are called terminal (or current) taxa and higher degree vertices are called ancestral taxa. We assume that the leaves of the tree are labeled from some label set **S**, corresponding to names of terminal taxa. In many cases a special degree two vertex is designated as the "root" of the tree and it is understood that the root is the most ancestral taxon and edges are directed away from the root. We associate with each edge, e, of the tree graph a positive number, $w(e)$. Depending on the context, $w(e)$ may represent time or some expectation under a suitable stochastic model of evolution. For example, it is common to model bio-molecular sequence evolution by a continuous time Markov model of state transitions along each edge. In this case, the edge weights represent

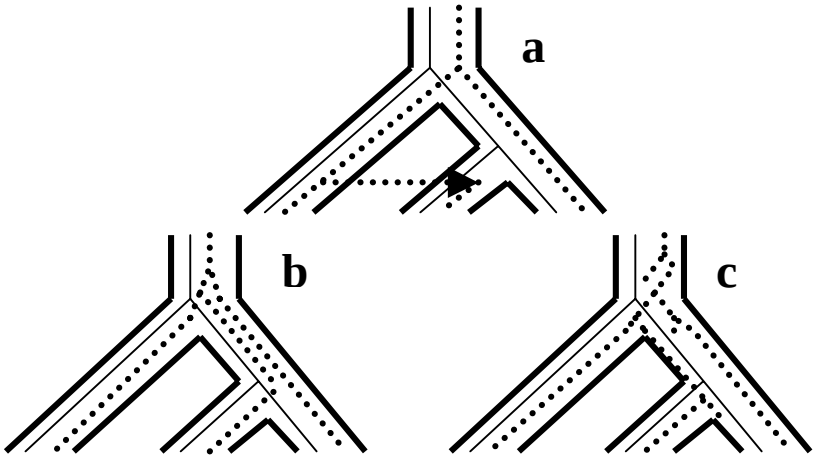

Figure 1. Three different ways in which gene trees may differ from species trees. The bold outlined "tubes" represent organismal species trees. The solid thin line represents a gene lineage that is concordant with the species tree. The bold dashed lines represent another gene whose lineage differs from the species trees by (**a**) a horizontal transfer of the gene, (**b**) ancestral lineage sorting, and (**c**) birth-death process of paralogous genes.

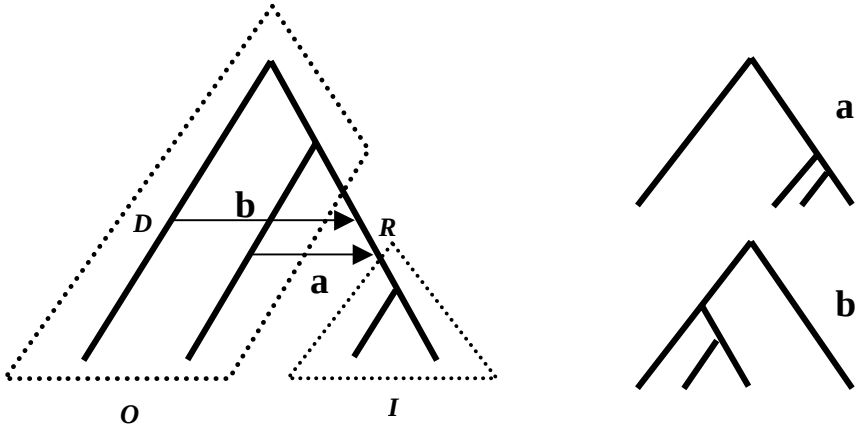

Figure 2. Two kinds of horizontal transfer of gene lineages and the resulting tree topologies**. a.** A transfer between adjacent lineages produces no change in the tree topology but a change in the distance relationships. **b.** A transfer across lineages produces a change in the tree topology and distance relationships.

the expected number of Poisson counting events. Given a tree $T = \{V, E\}$, if we delete an edge that is not adjacent to a leaf, the tree will be split into two connected components, T_1 and T_2. This also results in a bipartition of the leaf-label set, S, into S_1 and S_2. Such a bipartition is called a split[1]. If we delete each edge (non-adjacent to a leaf) in turn, exhaustively, we obtain a set of splits that represents the topology of the tree graph. The topology of a tree is the branching relationship among the labeled leaves of the tree implied by the tree graph. The main objective of most estimation algorithms is to obtain the best estimate of the correct tree topology.

A path distance, P_{ij}, between pairs of vertices v_i and v_j can be defined as the sum of the edge weights in the unique path between v_i and v_j. Given a matrix of all pairwise distances between terminal taxa, an edge-weighted tree graph can be found whose path distances correspond to the distance matrix if the pairwise distances satisfy the so-called four-point condition[19,20]:

$$d_{ij} + d_{kl} \leq \max(d_{ik} + d_{jl}, d_{il} + d_{jk}) \qquad (1)$$

The four-point condition can be simply interpreted as saying that for any quartet of vertices, the pairwise distances behave like path distances on a tree. If a distance matrix satisfies (1) it is said to be an additive distance matrix. The relation between additive distance matrices and tree graphs forms the basis of phylogeny estimation by distance methods. Many distance based phylogeny estimation algorithms[6] will perform well given accurate estimates of additive distances.

Given a bio-molecular sequence, distance based phylogeny algorithms estimate pairwise distances between the terminal taxa. It is standard to assume a continuous time Markov model and attempt to estimate the expected number of Poisson counting events in the path between two terminal taxa. Denote this estimate by $\hat{d}_{ij}$; then if the bio-molecules are *iid* samples of the Markov process, $d_{ij} = E_{\psi}(\hat{d}_{ij})$, the expectation of the estimate under the Markov process ψ satisfies the four-point condition and the resulting distance matrix is an additive distance matrix. If we are given a collection of bio-molecular sequences, m_1—m_k, and they are samples of the same homogeneous stochastic process, it is natural to obtain a weighted average distance estimate $\hat{d}_{ij} = \sum w_k \hat{d}_{ij}^k$ where w_k's are weight coefficients based on the sample size of each individual estimate. This averaged estimate can be used to estimate a single tree representing the combined information from the k molecules and since we have an *iid* process we expect the weighted average to yield a lower variance estimate. However, it is unreasonable to assume that m_1—m_k are samples of a homogeneous process. We expect each molecule to have a unique Markov process associated with it, with different edge weights for each molecule. In general, the four-point condition is not additive and the sum of two additive distance matrices may not be additive. But, if each molecule is a sample from a stochastic process over the same tree topology (with different edge weights possible), the sum of the

expected distance matrices is an additive matrix. Therefore, it is again natural to obtain a weighted average estimate $\hat{d}_{ij} = \sum w_k \hat{d}_{ij}^k$ even when each molecule is a sample from a unique process. (In this case, an efficient estimator would require selecting weights that reflect the variance of individual processes.)

However, when a horizontal transfer type of phenomenon occurs the topology of the tree may change. (The other tree topology changing events such as gene duplication and loss can also be represented as a horizontal transfer event, since each duplication/loss can be seen as generating an alternate tree. Therefore, we will only refer to horizontal transfer from here on.) Figure 2 shows two examples of change in tree topology or distance relationships following a single horizontal transfer event. Different horizontal transfer events will generate a tree with a different topology. Given a collection of molecules that have experienced horizontal events we will have a collection of trees all slightly different from each other. One possible view of this process is to assume that there is a single organismal phylogeny and molecular phylogenies differ from the organismal phylogeny by one or more horizontal transfer events. The goal of our algorithm is to estimate the single organismal phylogeny from a collection of molecular phylogenies.

The key to our algorithm is to ask what happens to pairwise distance relationships under horizontal transfer. Suppose D_A is the expected distance matrix representing the original tree in figure 2 and D_B and D_C are the new expected distance matrices for the resulting trees. D_B will be different from D_A in its pairwise elements and D_C will be different as well, but in a *different* subset of the elements. Suppose now we examine distances between *i*th and *j*th taxa and ask what happens to this number when we have *k* different molecules. Let the expected distance matrix for the species tree be D^0 and the expected distance matrix for the *k*th molecule be D^k. D^k will differ from D^0 in some of the elements if horizontal transfer event occurred in the *k*th gene. Denoting the distance estimate between *i*th and *j*th taxa for the *k*th molecule by $\hat{d}_{ij}^k$, we expect $\hat{d}_{ij}^k$ to have either central tendencies $\bar{d}_{ij}^0$ or $\bar{d}_{ij}^k \neq \bar{d}_{ij}^0$ according to whether horizontal transfer distorted the relationship between *i*th and *j*th taxa. Therefore, the distance between *i*th and *j*th taxa is a mixture random variate $d_{ij} = \alpha d_{ij}^0 + (1-\alpha) d_{ij}^H$ with the expectation $\bar{d}_{ij} = \alpha \bar{d}_{ij}^0 + (1-\alpha) E_H(d_{ij})$, where E_H is the expectation under the stochastic model of horizontal transfer and α is the mixture proportion. If $\alpha > 0.5$, it is natural to use the median of $\hat{d}_{ij}^k$ as an estimate of $\bar{d}_{ij}^0$ since it is robust to the mixture. Thus our Median Tree Algorithm (MTA) is:

1. For each pair of taxa *i* and *j*, compute a distance estimate for each of the *k* molecules, $\hat{d}_{ij}^{k}$.
2. Within the distance matrix for each gene, normalize the pairwise distances. (We suggest a normalization procedure in the next section.)
3. Compute a Median Distance Matrix by finding the median of $(d_{ij}^{1}, d_{ij}^{2} \cdots d_{ij}^{k})$ for every *i*th and *j*th taxa pair.
4. Estimate a tree with the median distance matrix and a distance based algorithm.

3 Performance Tests

3.1 Generation of model trees

To explore the efficacy of the Median Tree Algorithm, we applied these methods to data derived from simulations. Each model species tree was accompanied by 25 gene trees based on horizontal transfer of branches of the species tree. A wide range of transfer rates was tested along with distortion of edge weights.

For each replicate, a species tree was created using a Yule process. In this process, a tree begins with a degree two vertex, the root, connected to two leaves. Each leaf bifurcates with a constant rate under a homogeneous pure birth process. For the generation of our species trees, we sampled from the Yule process conditioned on maximum time = 1, terminal taxa (leaves) = 20, and bifurcation rate = $ln10$, the rate at which 20 leaves are expected at time = 1. Gene trees were created by applying a constant rate, *h*, of horizontal transfer to a copy of the species tree. In the first stage, "donor" vertices were added to the tree. For each edge, a waiting time, t_h, was drawn. If $t_h < l(e)$, the time length of the edge, then a donor vertex was added between the edge's vertices. Edges were considered recursively for further addition of donor vertices. In the second stage, a recipient vertex was selected for each donor, processed in order of proximity to the root. Any vertex with an immediately ancestral edge that spanned the donor's time from root was eligible to be a recipient. (This "random choice" model creates more severe estimation problems than a "genetic affinity" model where the choice of recipient lineages is weighted by genealogical distance.) The recipient, chosen equiprobably from the eligible vertices, was then detached from its ancestral vertex and connected to the donor with a branch length that would maintain its distance from the root. The previous ancestor was removed from the tree and its other two neighboring vertices were connected with a distance preserving branch length. By not considering evolutionary divergence in selection of donor-recipient pairs, we presumably made it more difficult to recover the model trees than if closely related lineages were more probable partners.

3.2 Generation of test data

For each gene tree, simulated DNA sequences of length 1500 were generated by the program Seq-Gen[21] using the HKY model[22] and a transition/transversion ratio of 2:1. Given these sequences, pairwise distances were estimated by the program V_MLDIST[23] with the assumption of HKY evolution but an unknown transition/transversion ratio. In one set of experiments, gene trees were further differentiated by distortion of their edge weights simulating non-clock-like evolution because the rate of bio-molecular sequence evolution is known to vary by gene and by lineage. To model this variation, we deformed each edge independently with a coefficient drawn from the uniform distribution [0.5, 1.5].

3.3 Normalization

Step 2 of the MTA, normalization, is important because evolutionary rates may vary among genes regardless of horizontal transfer. Before pairwise distance estimates between terminal taxa can be compared, they must be scaled so that rate differences between genes do not dominate the averaging process. For the MTA to be effective at recovering the species tree, distance variation among genes must reflect primarily the different patterns of horizontal transfer.

The normalization procedure we used begins with an initial pass through the first three steps of the MTA. Let D^M be the median distance matrix computed from unnormalized distances. Then for the kth distance matrix, D^k, we obtain a scaling factor s^k as follows.

$$s^k = \text{median}\{\frac{d_{ij}^k}{d_{ij}^M} \mid \text{for all i, j}\} \qquad (2)$$

That is, we compute the ratio of each distance matrix element to the median distance matrix element. The median of these ratios over all elements is considered an appropriate scaling factor.

3.4 The estimation procedures

Five estimates of the species tree were made for each set of genes. Two were the MTA estimates, one with and one without normalization of the matrices. Estimation of the tree from each matrix (Step 4) was done using the neighbor-joining distance method[24] as implemented in the program NEIGHBOR[25]. Two others were the same as above except that we used mean values rather than median values. The fifth estimate was produced in a different fashion. A neighbor-joining estimate of each of the original gene matrices was made and these trees were processed by the program CONSENSE[25]. This program measures the frequency of each split among a set of

trees. It then constructs an output tree by adding splits in order of decreasing frequency until the next most frequent (or tied) split contradicts an existing split.

3.5 Results

The accuracy of each estimate was measured as dT, the partition metric[26]. This measure is the sum of splits present in the true tree but not in the estimate and splits in the estimate but not in the true tree: false negatives and false positives respectively. Each simulated tree contained 17 splits, thus a maximum error value of dT = 34. Because the neighbor-joining method (as implemented) always produces a single bifurcating tree, the number of false positives and negatives is equal for every estimate produced by the first four procedures in section 3.4. The consensus tree can have fewer false positives than negatives, but the numbers appeared similar in our experiment (not shown). For simplicity, only dT results are reported here.

Samples of 100 species trees were tested for each of nine horizontal transfer rates corresponding to expected numbers of horizontal transfers ranging from 0 to 16 per gene over the tree. Lawrence and Ochman[27] estimated that 17% of genes in *E. coli* originated from horizontal transfer in the last 100 million years, or 1.7×10^{-9} events per gene per year. The total expected amount of time in a Yule tree is

$$2t + (N-2)\frac{1 - e^{-\lambda t} - \lambda t e^{-\lambda t}}{\lambda\left(1 - e^{-\lambda t}\right)}, \qquad (3)$$

where $t = 1$, $N = 20$, and $\lambda t = ln10$ in our simulations. Thus, the expected total time in the trees in our simulation is ~7.8 units and the horizontal transfer rate ranges from approximately 1/7.8 to 16/7.8 events per gene per unit time. Translating this to biological units, our simulation ranges from 76 million to 1.2 billion years of evolution at the horizontal transfer rate of *E. coli*.

We calculated mean and standard error for each method and sample. To test the significance of differences between the accuracy of the five methods, we performed sign tests, which are more conservative than paired *t*-tests. Normalization had no effect on the accuracy of either the mean or the median estimates when tested across all transfer rates under either the undistorted or distorted branch length model (two-tailed $p = 0.26$, 0.26, 0.15, and 0.42 respectively). In our experiments, all simulated genes evolved at the same rate, except for random variation imposed on edges individually in the distorted model. Therefore, normalization was not expected to improve estimation accuracy. This test confirms that the normalization procedure was not detrimental in the absence of rate variation. In an unreported experiment, we found that normalizing according to the greatest distance estimate in each gene matrix was detrimental and so was abandoned in favor of the procedure proposed here. Because we expect that normalization would be a necessary part of the Median

Tree Algorithm when applied to natural data, all remaining tests below consider only the results for normalized distance estimates.

Figure 3 shows the results for the (a) undistorted and (b) distorted branch length models defined in section 3.2. The mean tree algorithm was the least accurate under most model conditions. The Mean TA outperformed the Median TA only in the case of branch length distortion combined with no horizontal transfer, but did so significantly (two-tailed p = 0.0008). In this case, the only differences in the distance matrices of the gene trees were the symmetric deformations added stochastically under the model, which makes it unsurprising that the means were more accurate estimates than the medians.

The relative success of the Median TA and the consensus method varied more dramatically. For the undistorted simulations, the two methods performed comparably well over the range of 0 to 2 expected horizontal transfers per gene. At higher rates, the MTA had clearly greater accuracy. For the distorted simulations, MTA was again superior at the higher rates of transfer. When less than four transfers were expected per gene, the consensus method outperformed MTA. The consensus method has an advantage in being blind to branch lengths (so long as they do not distort the gene tree topology estimates), which explains its relatively greater success in the distorted experiments. At higher rates this advantage is outweighed by its sensitivity to horizontal transfer; a single transfer can cause a gene tree to share no splits with the species tree.

4 Discussion

Assuming a rooted tree, to every ancestral vertex there is a subtree of all vertices and edges directed away from that vertex. We will call such a subtree a clade. The number of terminal taxa in a clade will be called the size of the clade. A horizontal transfer event involves a donor lineage (marked D in figure 2) and a recipient lineage (marked R in figure 2). The transfer event can be (effectively) thought as picking up the clade subtending from the recipient lineage (we call this the I-clade; marked by a dotted triangle in figure 2) and moving it to the donor lineage D. The total change in the topology and the distance elements depend on the size of the I-clade (recipient lineage), the size of the terminal taxa outside of the I-clade (we call this the O-set; marked as O in figure 2), and the donor position, D.

In terms of distance relationships, suppose the tree contains N terminal taxa and I-clade contains M terminal taxa. The distance relationship within the I-clade will not be affected. The distance relationship within the O-set will also be not affected. Only M(N – M) pairwise distances between the terminal taxa in the I-clade and terminal taxa in the O-set will be affected. More specifically, suppose the transfer event involved moving from lineage D to R (see figure 2), suppose also that Q was the size of the smallest clade containing both D and R (in figure 2, this is the entire

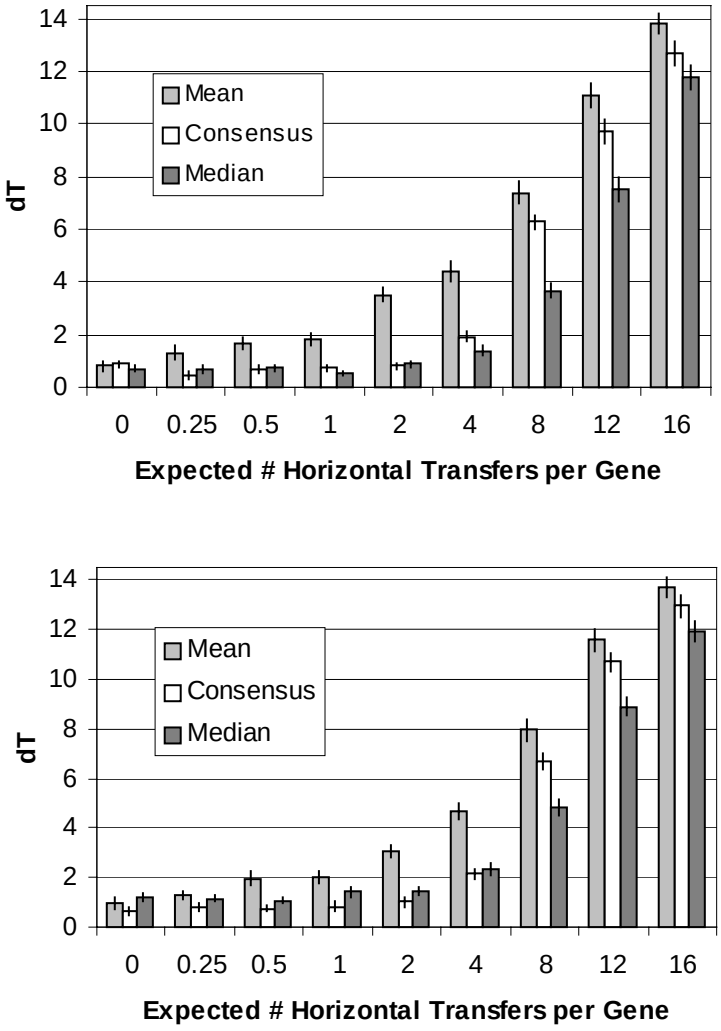

Figure 3. Simulation test results. Species phylogeny was estimated from 25 gene trees three ways: (1) mean estimated distance matrix, (2) consensus, and (3) median distance matrix (the MTA method). The vertical axis is a measure of how different the estimated tree is from the true species tree (a larger value means more different). The category axis shows expected number of horizontal transfer events in each gene. Both the case of molecular-clock evolution (**a**) and non-clock evolution is shown (**b**). The lines above the bars show standard errors.

tree and Q = N), then for each taxon in the I-clade, only (N – M) – (N – Q) = Q – M distances are affected. Therefore, a single horizontal transfer event is expected to change M(Q – M) pairwise distances where M and Q depends on the exact event; at most, ½ of the distance matrix will change by a single horizontal transfer event.

Consensus methods operate on rooted splits (subsets of the label set). If a horizontal transfer event occurs as in figure 2, the subsets representing the I-clade are not affected. The subsets representing all the descendent clades of the donor lineage D are also not affected. However, all clades "above" (towards the root) the points D and R are changed by the horizontal transfer event. As an extreme example, a single horizontal transfer event can change all the subsets of the resulting tree such that no part of the tree topology is preserved.

The simulation study demonstrates that the MTA method is superior to other methods of estimating the species tree. However, all methods performed surprisingly well. In general, the number of distance elements that increase in size versus those that decrease in size will be unequal. Thus, taking arithmetic means of the distance values will not yield unbiased estimates of the species distance matrix. In particular, the bias will tend to increase with the size of the tree. Therefore, using the mean estimate will not be appropriate when horizontal transfer is involved. Similarly, the movement of the donor lineage affects the rooted split relationship of all clades above the donor and recipient lineages, possibly destroying all concordance. The simulation results shown here seem to suggest that when a sufficient number of genes are present, this kind of misleading change may not dominate the data. However, we can expect the consensus method to perform increasingly worse with increasing tree size because the average proportion of splits affected by a single transfer event will increase with the number of terminal taxa whereas the average proportion of distance elements affected will stay the same.

To reiterate, biological events such as horizontal transfer can cause a difference in the gene trees of individual molecules versus that of the whole organisms. The differences cause biased distortions in the pairwise distance relationship of the genes. Such distortions can be seen as mixed central tendency of the distance estimates. Using medians instead of means can provide a robust estimate of the central tendency. Using simulations we demonstrate that such a median procedure is effective at estimating the species tree from a large number of gene trees even when they differ by multiple horizontal transfers. Our method is intuitive and computationally scales linearly with numbers of genes, which we believe will be increasingly important as genome-wide data become increasingly available.

Acknowledgments

This work was supported in part by NSF grant DEB-9806570 to JK. Two anonymous reviewers provided useful comments.

References

1. Dress, A. and Steel, M. A. *Appl. Math. Lett.* **5**, 3 (1992).
2. Steel, M. A. *Appl. Math. Lett.* (1993).
3. Fitch, W. M. *Phil. Trans. Roy. Soc. of Lond. B Biol. Sci.* **349**, 93 (1995).
4. Hillis, D. M. *Curr. Biol.* **7**, R129 (1997).
5. Felsenstein, J. in *Prospects in Systematics* (ed. Hawksworth, D. L.) 112 (Clarendon Press, Oxford, 1988).
6. Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. in *Molecular Systematics* (eds. Hillis, D. M., Moritz, C. & Mable, B. K.) 407 (Sinauer Associates, Sunderland, 1996).
7. Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J. and Woese, C. R. *Nuc. Acid. Res.* **24**, 82 (1996).
8. Doyle, J. J. *Syst. Biol.* **46**, 537 (1997).
9. Page, R. D. M. and Charleston, M. A. *TREE* **13**, 356 (1998).
10. Maddison, W. P. *Syst. Biol.* **46**, 523 (1997).
11. Mayr, E. *Proc. Nat. Acad. Sci., USA* **95**, 9720 (1998).
12. Gupta, R. S. *Microbiol. Mol. Biol. Rev.* **62**, 1435 (1999).
13. Olsen, G. J. and Woese, C. R. *Cell* **89**, 991 (1997).
14. Jain, R., Rivera, M. C. and Lake, J. A. *Proc. Nat. Acad. Sci., USA* **96**, 3801 (1999).
15. Kyrpides, N. C., Olsen, G. J., Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R., Koonin, E. V. *Trends Gen.* **15**, 298 (1999).
16. Page, R. D. M. *Bioinformatics* **14**, 819 (1998).
17. Eulenstein, O., Mirkin, B. and Vingron, M. *J. Comp. Biol.* **5**, 135 (1998).
18. Ragan, M. A. *Mol. Phyl. Evol.* **1**, 53 (1992).
19. Waterman, M. S., Smith, T. F., Singh, M. and Beyer, W. A. *J. Theoret. Biol.* **64**, 199 (1977).
20. Buneman, P. in *Mathematics in the Archeological and Historical Sciences* (eds. Hodson, F. R., Kendall, D. G. & Tautu, P.) 387 (Edinburgh University Press, Edinburgh, 1971).
21. Rambaut, A. and Grassly, N. C. *CABIOS* **13**, 235 (1997).
22. Hasegawa, M., Kishino, H. and Yano, T. *J. Mol. Evol.* **22**, 160 (1985).
23. Drummond, A. and Strimmer, K. *PAL: A Java library for molecular evolution and phylogenetics* (http://www.pal-project.org, 2000).
24. Saitou, N. and Nei, M. *J. Mol. Evol.* **4**, 406 (1987).
25. Felsenstein, J. *PHYLIP: Phylogeny inference package, version 3.57c* (Department of Genetics, Univ. Washington, Seattle, 1995).
26. Robinson, D. F. and Foulds, L. R. *Math. Biosci.* **53**, 131 (1981).
27. Lawrence, J. G. and Ochman, H. *J. Mol. Evol.* **44**, 383 (1997).