Towards Extracting All Phylogenetic Information from Matrices of Evolutionary Distances

Sebastien Roch,[1*]

[1]Department of Mathematics and Bioinformatics Program
University of California-Los Angeles
Los Angeles, California 90095, USA

[*]To whom correspondence should be addressed; E-mail: roch@math.ucla.edu.

The matrix of evolutionary distances is a model-based statistic, derived from molecular sequences, summarizing the pairwise phylogenetic relationships between a collection of species. Phylogenetic tree reconstruction methods relying on this matrix are relatively fast and thus widely used in molecular systematics. However, because of their intrinsic reliance on summary statistics, distance-matrix methods are assumed to be less accurate than likelihood-based approaches. In this paper, pairwise sequence comparisons are shown to be more powerful than previously hypothesized. A statistical analysis of certain distance-based techniques indicates that their data requirement for large evolutionary trees essentially matches the conjectured performance of maximum likelihood methods—challenging the idea that summary statistics lead to suboptimal analyses. On the basis of a connection between ancestral state reconstruction and distance averaging, the critical role played by the covariances of the distance matrix is identified.

Information about evolutionary trees can be inferred from the fact that species that are close in the tree of life tend to have similar molecular sequences. In its most basic form, the evolutionary distance between two DNA sequences is estimated from the proportion of homologous sites differing between them, typically corrected for back-mutations under common modeling assumptions (*1*). For a collection of sequences, the pairwise evolutionary distances form a matrix—the distance matrix—which underlies a popular class of tree reconstruction methods. Technically, distance-matrix methods include all phylogenetic inference techniques relying solely on pairwise sequence comparisons, including Neighbor-joining (NJ), BIONJ, WEIGHBOR and FastME (*2–5*). Because of their simplicity, such methods are often considerably faster than parsimony and likelihood-based approaches (*6, 7*) and distance-matrix methods are used for large-scale phylogenetic reconstruction and bootstrap analysis, or to produce starting trees for maximum likelihood (ML) heuristics. However, it is unknown if this advantage in speed affects accuracy adversely.

The use of distance-matrix information has been criticized for seemingly ignoring higher-order information, that is, data patterns involving more than two sequences (*1*). Moreover, it has been observed through combinatorial arguments that the conversion from molecular sequences to the distance matrix is far from invertible (*8*). However this hypothesized information loss has not been quantified in a model-based framework. In reality the comparison of distances between different pairs of sequences does involve higher-order signal, albeit in a highly summarized form. But it is unclear how to use such information. In particular the correlation between the entries of the distance matrix has largely been ignored in the design and analysis of distance-matrix methods with a few notable exceptions (*3, 9, 10*).

Formally, phylogenetic data consists of n aligned DNA sequences of length k (without gaps): $s_1^a, \ldots, s_k^a$, where a ranges over the n terminal taxa (Fig. 1). The evolutionary distance between the sequences at a and b is denoted by $\hat{\delta}(a, b)$. In the Jukes-Cantor model, a classical

substitution model which treats all nucleotides symmetrically, the standard distance formula takes the form

$$\hat{\delta}(a,b) = -\frac{3}{4}\log\left(1 - \frac{4}{3}\widehat{\mathrm{p}}_{a,b}\right), \tag{1}$$

where $\widehat{\mathrm{p}}_{a,b}$ is the proportion of homologous sites differing between the sequences at a and b.

A variety of distance-matrix methodologies have been proposed, of which this study focuses on agglomerative methods, including unweighted pair-group method using arithmetic averages (UPGMA) and NJ (*2, 11, 12*). Such methods proceed in two steps that are repeated until termination: a) a selection step where a pair of operational taxonomic units (OTUs) is selected for agglomeration and b) a reduction step where a reduced distance matrix is computed on the remaining OTUs. As an example, in the case of a molecular clock (that is, under the assumption that substitutions occur at the same rate in all branches of the tree), one can simply select the two closest OTUs A' and A'' to form a new composite OTU A and the matrix is reduced with the rule

$$\bar{\delta}(A,B) = \mu\bar{\delta}(A',B) + (1-\mu)\bar{\delta}(A'',B), \tag{2}$$

where $\bar{\delta}$ was used to indicate the reduced distance matrix, with the convention $\bar{\delta}(\{a\},\{b\}) = \hat{\delta}(a,b)$. The choice of the reduction coefficient, μ, is critical and the appropriate value has been debated (*3, 4, 13*). Standard choices are $\mu = |A'|/|A|$ where $|A|$ denotes the number of terminal taxa in A, leading in the clock case to UPGMA, and $\mu = 1/2$, leading in the clock case to the less common weighted version of UPGMA known as weighted pair-group method using arithmetic averages (WPGMA) (*11, 12*).

An important criterion in designing the selection and reduction steps above is the statistical consistency of the resulting method: as sequence length increases, the reconstructed tree should converge on the true phylogeny. However, because consistency is a coarse statistical property, it is of limited use in comparing different methods (*14*). A case in point is that both reduction coefficients above produce a consistent estimation under a molecular clock. A finer contrast is

obtained with the sample complexity (SC), that is, the sequence length required to guarantee that the reconstruction is correct with a given confidence when the data is generated according to a standard substitution model on the true tree.

Because the sample complexity of a method depends in an intricatre manner on the parameters of the generating model, it is more feasible to evaluate its asymptotic behavior as a designated parameter of interest converges to a limit. For instance, any inference method requires at least an SC proportional to $1/f^2$ as the shortest branch length $f \to 0$ (*15*). In statistical theory, sample complexities are often compared between two inference methods in the form of a ratio called the asymptotic relative efficiency (*16*).

In this study, the regime of interest is large phylogenies and the structural parameters considered are the number of terminal taxa n, the shortest branch length f, and the depth of the tree which we denote by Δ. (All branch and path lengths are expressed in number of substitutions per site.) The depth measures how far the edges are from the terminal taxa (leaves). Precisely, the depth of an edge e is defined to be the length of the shortest path between two leaves that crosses e and the depth of the tree is the largest such quantity over all its edges (*17*). The length of the longest branch, g, also plays a non-asymptotic role. Here it is claimed that the following sample complexity can be achieved with only distance-matrix information under the Jukes-Cantor model (more general models as well as random trees are discussed below):

$$\text{SC} = \begin{cases} C_0 \cdot 1/f^2 \cdot \log n & \text{if } g < g^*_{\text{JC}} \approx 0.26 \\ C_1 \cdot 1/f^2 \cdot e^{C_2 \Delta} \cdot \log n & \text{if } g \geq g^*_{\text{JC}}, \end{cases} \tag{3}$$

where C_0, C_1 and C_2 are positive constants independent of f, Δ, and n, but C_0 may depend on g. (To simplify the formula above, it was assumed that f is small—precisely $f \leq 1$. In general, the factor $1/f^2$ should be replaced with $\max\{1/f^2, 1\}$.) Table 1 shows a comparison with previous results.

The significance of Eq. (3) is that its functional dependence in f, Δ, and n is the best

	$g < g^*_{\rm JC}$	$g \geq g^*_{\rm JC}$
NJ (*18*)	$1/f^2 \cdot e^{\rm Diam} \cdot \log n$	$1/f^2 \cdot e^{\rm Diam} \cdot \log n$
SQM (*17*) (+ UPGMA in MC)	$1/f^2 \cdot e^{\Delta} \cdot \log n$	$1/f^2 \cdot e^{\Delta} \cdot \log n$
Here (+ WPGMA in MC)	$1/f^2 \cdot \log n$	$1/f^2 \cdot e^{\Delta} \cdot \log n$
Conjecture for ML (best possible)	$1/f^2 \cdot \log n$	$1/f^2 \cdot e^{\Delta} \cdot \log n$

Table 1: Summary of previous results on the SC of distance-matrix methods with constants omitted for clarity. MC stands for the molecular clock case. The last row of the table gives the conjectured SC of ML for comparison.

possible among all phylogenetic reconstruction methods except possibly for the values of the constants C_0, C_1, and C_2 (*19*). In particular it is natural to conjecture that the SC of maximum likelihood estimation, currently unknown (*15*), is also given by the same expression. In comparison, the performance of NJ scales exponentially in the diameter (that is, the length of the longest path) for all values of g (*18, 20*). Note that the diameter is larger than the depth, often much larger (think of a tree with a toplogy like that of a caterpillar). Therefore, not all distance-matrix methods achieve the SC in Eq. (3). In fact, the current study produces explicit prescriptions in designing accurate distance-based techniques.

In Eq. (3) the constants C_0, C_1 and C_2 may be evaluated through simulations or approximations on special cases (Fig. 2). The scaling in $1/f^2$ is necessary and sufficient for all phylogenetic methods, whether distance or likelihood-based (*15*). Likewise, by information-theoretic arguments the factor of $\log n$ is necessary (*19*). Here it arises as a Bonferroni correction, that is, it accounts for the high number of evolutionary distances that must be estimated simultaneously in order to infer a large phylogeny. In particular, to reconstruct a single deep branching in an otherwise known tree, the factor of $\log n$ can be dropped.

The depth, on the other hand, displays an unexpected behavior. The thresholding effect at $g^*_{\rm JC} = 3 \ln 2/8$ mirrors a similar phenomenon in ancestral state reconstruction. Because of the tree-like nature of phylogenies, there exists a fundamental tradeoff between how fast

information is lost through mutations along each path connecting the root to the leaves and, on the other hand, how fast information is replicated through the exponential growth of the tree. Under the Jukes-Cantor model, the critical threshold at which these two effects balance each other out on a bifurcating tree is $g = g^*_{\mathrm{JC}}$ which corresponds roughly to an average of 22 substitutions per 100 sites per branch. When $g < g^*_{\mathrm{JC}}$ (which I refer to as the lower phase of the parameter space), good estimates of ancestral sequences can be obtained no matter how deep the tree is, whereas if $g \geq g^*_{\mathrm{JC}}$ (the upper phase) even the best ancestral estimation methods exhibit a quickly deteriorating accuracy as deeper parts of the tree are analyzed. This phase transition has been studied in statistical physics (*21*).

In addition, this information-theoretic limitation impacts the accuracy of phylogenetic tree reconstruction: the functional dependence in Δ in Eq. (3) is necessary and sufficient for general phylogenetic reconstruction methods, except possibly for the values of the constants C_0, C_1, and C_2 (*19*). In particular, in the lower phase, the sequence length required to infer a deep branching does not grow with the depth.

Regarding distance-matrix methods, however, the best previous results had an SC exponential in Δ for any value of g, using so-called short quartet methods (SQM) (*17*)—which was expected to be best for distance-based techniques (*22*). This study concludes that distance-matrix methods are capable of achieving the asymptotically optimal SC in Eq. (3). This is surprising in that it is not immediately clear how to exploit ancestral state reconstruction through the use of summary statistics alone. This hidden relation between ancestral states and the distance matrix (*22*) forms the basis of the analysis.

The assumption that the largest branch length is bounded, even for large trees, may seem unrealistic. After all, standard random tree models do not satisfy this kind of condition. However classical results on ancestral state reconstruction (*21*) naturally lead to the conjecture that the results discussed here—at least for the reconstruction of a single deep branching—are still true

under the relaxed assumptions that an appropriately defined average branch length $\bar{g}$ is bounded above by g^*_{JC}. The sample complexity for full reconstruction may require an extra factor depending on g, but that is likely for any method. But this is only conjecture as a rigorous analysis of the random tree case is still needed.

Although the molecular clock case is less relevant in practice, it is illustrative of the techniques used. Analysis of the Jukes-Cantor model under a molecular clock was performed abandoning the standard Jukes-Cantor distance of Eq. (1) and, instead, with the following transformation: for each species a we define a new sequence $\sigma^a_1, \ldots, \sigma^a_k$ with $\sigma^a_i = 1$ if $s^a_i =$ `A` or `G` (purines) and $\sigma^a_i = -1$ if $s^a_i =$ `C` or `T` (pyrimidines) letting the evolutionary distance between species a and b be given by $\hat{\delta}_{\mathrm{u}}(a,b) = (1 - \widehat{\Theta}(a,b))/2$ where

$$\widehat{\Theta}(a,b) = \frac{1}{k}\sum_{i=1}^{k} \sigma^a_i \sigma^b_i, \tag{4}$$

is a similarity score between sequences at a and b. A simple calculation shows that $\hat{\delta}_{\mathrm{u}}$ is in fact the standard uncorrected Cavender-Farris-Neyman distance; although an extension to General Time-Reversible (GTR) models reveals some differences with standard distance definitions. Note that, under the molecular clock assumption, correction for back-mutations is not needed (*23*).

The choice of a reduction coefficient in Eq. (2) can be cast in a more general context by using positive weights $w(a)$, $a \in A$, and $w(b)$, $b \in B$, which sum to 1 on each OTU and letting $\bar{\delta}_{\mathrm{u}}(A,B) = (1 - \overline{\Theta}_{\mathrm{u}}(A,B))/2$ where

$$\overline{\Theta}_{\mathrm{u}}(A,B) = \sum_{a\in A}\sum_{b\in B} w(a)w(b)\widehat{\Theta}(a,b). \tag{5}$$

For instance, the μs corresponding to UPGMA and WPGMA coincide respectively with $w(a) = 1/|A|$ and $w(a) = 2^{-|a|_{a^*}}$, where $|a|_{a^*}$ is the number of branches between a and the most recent common ancestor (MRCA) a^* of A.

It is natural to choose those $w(a)$s that minimize the variance of Eq. (5). If one were to ignore the correlations between the $\widehat{\Theta}(a,b)$s and use the fact that the variances of the $\widehat{\Theta}(a,b)$s are identical under a molecular clock (assuming that A and B correspond to true clades), then the solution is given by the minimizer of $\sum_{a\in A} w(a)^2 \sum_{b\in B} w(b)^2$ which is simply $w(a) = 1/|A|$ and $w(b) = 1/|B|$ showing that UPGMA would be the best choice. However, the $\widehat{\Theta}(a,b)$s are in fact correlated, which complicates the variance minimization. To solve the problem, Eqs. (4) and (5) were combined into

$$\bar{\delta}_{\mathrm{u}}(A,B) = \frac{1}{2}\left(1 - \frac{1}{k}\sum_{i=1}^{k}\left[\sum_{a\in A} w(a)\sigma_i^a\right]\left[\sum_{b\in B} w(b)\sigma_i^b\right]\right). \tag{6}$$

The expressions in square brackets can be interpreted as linear ancestral state estimators: The value of $\sum_{a\in A} w(a)\sigma_i^a$ is a form of weighted majority vote which tends to be positive if the MRCA of A has a purine at site i and negative in the case of a pyrimidine. In other words, the distance average in the reduction step is equivalent to an implicit ancestral sequence estimation. As a result, the task of minimizing the variance of $\bar{\delta}_{\mathrm{u}}(A,B)$ requires the identification of good linear ancestral estimators. Such estimators have been investigated in statistical physics and a formula for the variance of $\sum_{a\in A} w(a)\sigma_i^a$ has been derived (*21, 24*).

With this interpretation of the reduction step, insights about ancestral estimation were used to derive new results about distance-matrix methods. When $g < g^*_{\mathrm{JC}}$ a particularly good choice of weights is $w(a) = 2^{-|a|_{a^*}}$—the WPGMA choice. Although these weights are not strictly speaking variance minimizers, they produce a linear ancestral estimator with a relatively small (bounded) variance no matter how deep the tree. A rigorous analysis of WPGMA on the basis of this observation showed that the asymptotically optimal SC in Eq. (3) is achieved (*24*).

In contrast, although UPGMA is expected to behave similarly to WPGMA on balanced trees, it may perform more poorly on unbalanced ones, for instance, under uneven taxon sampling (*25*). In fact, the performance of UPGMA deteriorates drastically under unbalanced con-

ditions. Consider the idealized example of Fig. 2 where one seeks to infer a deep branching in a phylogeny containing a dense subtree. In this case a uniform weighting of the terminal taxa disproportionately favors the dense subtree and produces, in the lower phase, an ancestral estimator of the MRCA a^*_{d} of the dense subtree rather than of the MRCA a^* of the full clade (i.e., clade A) (E. Mossel, Personal communication). As a result the long path between a^*_{d} and a^* leads to a sample complexity that scales exponentially with the depth (*24*) (Fig. 3). In comparison, WPGMA achieves the lower SC in Eq. (3) for any tree topology. In particular, the asymptotic efficiency of UPGMA relative to WPGMA tends to 0 as $\Delta \to +\infty$ (Fig. 3). In other words, ignoring the correlation structure of the distance matrix may lead to significantly poorer performance in such cases.

To extend the analysis to GTR models, a generalized sequence transformation was applied (*22, 24*). A GTR model is specified by a reversible 4×4 substitution rate matrix Q with stationary distribution π. The rate matrix is by convention normalized so that the total rate of change per unit time at stationarity is 1, that is, $\sum_i \pi_i Q_{ii} = -1$. Let ν be an eigenvector of Q corresponding to the largest negative eigenvalue $-\lambda_Q$ (normalized so that $\sum_i \pi_i \nu_i^2 = 1$). For each species a, define a new sequence $\sigma_1^a, \ldots, \sigma_k^a$ with $\sigma_i^a = \nu_{\mathtt{A}}$ (respectively, $\nu_{\mathtt{G}}$, $\nu_{\mathtt{C}}$ and $\nu_{\mathtt{T}}$) if $s_i^a = \mathtt{A}$ (respectively, `G`, `C` and `T`) and let the evolutionary distance between species a and b be given by Eq. (4). This transformation is justified by the fact that, in an asymptotic sense, the best linear ancestral state estimator is a function of ν (*26,27*). In particular, the critical threshold g^*_Q can be expressed as $g^*_Q = \lambda_Q^{-1} \ln \sqrt{2}$. For instance, in the Tamura-Nei model (*28*) the critical branch length is $g^*_Q = \pi_R \pi_Y (1 + R) \ln 2$ where $\pi_R = \pi_{\mathtt{A}} + \pi_{\mathtt{G}}$, $\pi_Y = \pi_{\mathtt{C}} + \pi_{\mathtt{T}}$ and R is the transition to transversion ratio (*24*). For reference, under a uniform stationary distribution with ratios $R = 2$ and $R = 10$ (ratios typical of mammalian nuclear DNA and mitochondrial DNA respectively (*1*)), the critical branch lengths are 0.520 and 1.906.

To remove the molecular clock assumption, the corrected reduced distance matrix

$$\bar{\delta}_{\mathrm{c}}(a^*, b^*) = -\ln\left(\sum_{a\in A}\sum_{b\in B} w(a)w(b)e^{\delta(a^*,a)}e^{\delta(b^*,b)}\widehat{\Theta}(a,b)\right), \tag{7}$$

was used, where a^* and b^* are respectively the roots of clusters A and B. The quantity $\delta(a^*, a)$ is the expected number of changes per site between a^* and a. It is estimated by summing the corresponding branch lengths, which are themselves estimated from previously computed $\bar{\delta}_{\mathrm{c}}$ values progressively as the tree is built (*24*). If the parenthesis above is negative (known as the saturation problem (*29*)), the logarithm is not defined. For the analysis, $\bar{\delta}_{\mathrm{c}}(a^*, b^*)$ was set to $+\infty$ in that case. Eq. (7) can be interpreted as a correction-uncorrection scheme where the reduction step is performed (inside the logarithm) with uncorrected similarity scores, as in Eq. (5); the estimate is then corrected (by taking the logarithm) for the selection step. Beyond guaranteeing consistency, this scheme presents two advantages. By averaging uncorrected distances only, exact formulas can be applied for the variances and covariances needed in designing the reduction step. Secondly, since the reduced distance $\bar{\delta}_{\mathrm{c}}(a^*, b^*)$ is more tightly concentrated (thanks to the variance minimization), the saturation problem is minimized. Similarly to the molecular clock case, a particularly good choice of averaging weight is $w(a) = 2^{-|a|_{a^*}}$. A further issue in removing the molecular clock assumption is that standard agglomerative schemes such as NJ suffer from a basic flaw (*20*): Typical selection criteria involve large distances—which tend to be noisy. Generalized, computationally-efficient agglomerative methods were recently devised that avoid this problem with a local selection criterion, although occasionally backtracking is necessary (*30*). Alternatively, OTUs are merged at internal branches (*31*). By combining a correction-uncorrection scheme with a local agglomerative method, it was established that pairwise sequence comparisons alone achieve the SC in Eq. (3) in this more general setting as well (*24*). (For technical reasons, in the analysis of the nonclock-like case it was assumed that the branch lengths are, roughly speaking, multiples of f (*32*).)

The SC in Eq. (3) indicates that the difference between distance and likelihood-based methods is more subtle than a cursory analysis would suggest. Note however that, because the bounds examined here are asymptotic, they are mostly relevant for large-scale phylogenies. Moreover the constants C_0, C_1 and C_2 are possibly smaller in the case of likelihood-based methods; although proof is still needed. On the other hand, because the maximum likelihood problem is rarely, if ever, solved to optimality its actual sample complexity may be affected. Another potential advantage of maximum likelihood is that, for some substitution models, good likelihood-based ancestral estimators exist beyond g^* [up to a slightly higher threshold (*33*)], which may translate into a better SC for maximum likelihood. In the presence of noise, however, that advantage disappears (*27*). Finally, under rates-across-sites models, it is possible for topologically different trees to generate identical distributions of pairwise sequence comparisons, making them indistinguishable through distance-matrix computations (*34*); and more general summary statistics are needed. Lastly, practical reconstruction algorithms are needed, a nontrivial task, in order to apply the theoretical insights discussed here.

References and Notes

1. J. Felsenstein, *Inferring Phylogenies* (Sinauer, Sunderland, MA, 2004).

2. N. Saitou, M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).

3. O. Gascuel, *Mol. Biol. Evol.* **14**, 685 (1997).

4. W. J. Bruno, N. D. Socci, A. L. Halpern, *Mol Biol Evol* **17**, 189 (2000).

5. R. Desper, O. Gascuel, *Journal of Computational Biology* **9**, 687 (2002).

6. R. L. Graham., L. R. Foulds, *Math. Biosci.* **60**, 133 (1982).

7. B. Chor, T. Tuller, *J. ACM* **53**, 722 (2006).

8. M. A. Steel, M. D. Hendy, D. Penny, *Nature* **336**, 118 (1988).

9. M. Bulmer, *Mol Biol Evol* **8**, 868 (1991).

10. E. Susko, *Mol Biol Evol* **20**, 862 (2003).

11. R. Sokal, C. Michener, *Univ. Kansas Sci. Bull.* **38**, 1409 (1958).

12. P. H. A. Sneath, R. R. Sokal, *Numerical taxonomy* (W. H. Freeman and Co., San Francisco, Calif., 1973).

13. R. Mihaescu, L. Pachter, *Proceedings of the National Academy of Sciences* **105**, 13206 (2008).

14. J. Kim, *Syst Biol* **47**, 43 (1998).

15. M. A. Steel, L. A. Székely, *SIAM J. Discrete Math.* **15**, 562 (2002).

16. M. J. Schervish, *Theory of statistics* (Springer-Verlag, New York, 1995).

17. P. L. Erdös, M. A. Steel, L. A. Székely, T. A. Warnow, *Random Struct. Algor.* **14**, 153 (1999).

18. K. Atteson, *Algorithmica* **25**, 251 (1999).

19. E. Mossel, *Trans. Amer. Math. Soc.* **356**, 2379 (2004).

20. M. R. Lacey, J. T. Chang, *Math. Biosci.* **199**, 188 (2006).

21. W. S. Evans, C. Kenyon, Y. Peres, L. J. Schulman, *Ann. Appl. Probab.* **10**, 410 (2000).

22. S. Roch, *FOCS'08: Annual IEEE Symposium on Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, CA, USA, 2008), pp. 729–738.

23. A. Rzhetsky, T. Sitnikova, *Mol Biol Evol* **13**, 1255 (1996).

24. Details appear in the Supplementary Materials.

25. R. R. Sokal, P. H. A. Sneath, *Principles of Numerical taxonomy* (W. H. Freeman and Co., San Francisco, Calif., 1963).

26. E. Mossel, Y. Peres, *Ann. Appl. Probab.* **13**, 817 (2003).

27. S. Janson, E. Mossel, *Ann. Probab.* **32**, 2630 (2004).

28. K. Tamura, M. Nei, *Mol Biol Evol* **10**, 512 (1993).

29. D. H. Huson, K. A. Smith, T. Warnow, *WAE '99: Proceedings of the 3rd International Workshop on Algorithm Engineering* (Springer-Verlag, London, UK, 1999), pp. 271–285.

30. C. Daskalakis, E. Mossel, S. Roch, *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing* (ACM, New York, 2006), pp. 159–168.

31. R. Mihaescu, Distance methods for phylogeny reconstruction, Ph.D. thesis, University of California, Berkeley (2008).

32. C. Daskalakis, E. Mossel, S. Roch, Evolutionary trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture (2009). *Probab Theor Relat Field (In Press).*

33. E. Mossel, *Ann. Appl. Probab.* **11**, 285 (2001).

34. M. Steel, *Journal of Theoretical Biology* **256**, 467 (2009).

35. This work was triggered by a discussion with E. Mossel about lower bounds for distance methods in which he pointed out that the distance matrix has a potentially useful correlation structure. I am also indebted to Y. Peres for his assistance with technical work which made

the results presented here possible. I thank J. Chayes and C. Borgs from Microsoft Research, where part of this work was performed.

Fig. 1. Example of a DNA sequence dataset. The alignment is typically obtained using a multiple sequence alignment heuristic applied to the collected sequences. The character $-$ is a gap. The columns are homologous sites, that is, they are derived from a common ancestor through substitutions. Those columns that include gaps are ignored. For instance, using the notation introduced in the text, we have $k = 10$ and the sequences at $a = \text{Homo sapiens}$ and $b = \text{Pan}$ are $s_1^a, \ldots, s_{10}^a = \mathtt{A}, \mathtt{C}, \mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{A}, \mathtt{G}, \mathtt{A}, \mathtt{A}, \mathtt{A}$ and $s_1^b, \ldots, s_{10}^b = \mathtt{A}, \mathtt{T}, \mathtt{A}, \mathtt{T}, \mathtt{A}, \mathtt{A}, \mathtt{G}, \mathtt{A}, \mathtt{A}, \mathtt{A}$. In particular, the preceding sequences agree on 8 out of 10 sites, so that $\widehat{\mathrm{p}}_{a,b} = 0.2$ and $\hat{\delta}(a, b) = 0.457$ using the Jukes-Cantor formula.

Fig. 2. A. Illustrative example (not to scale) where f_0, g_0, and g_1 values are branch lengths. The H_0, h_0, and h_1 values indicate numbers of levels. The subtrees B and C are complete binary trees. The subtree A is a complete binary tree with one subtree at height H_0 replaced by a complete binary tree with h_1 levels. The top dashed subtree is referred to as the deep triplet in the text. The dark subtree is denser, that is, $g_1 < g_0$. To satisfy the molecular clock assumption $h_1 g_1 = h_0 g_0$. B. Sample complexity for correct reconstruction of the deep triplet in A with $h_0 = h_1 = 0$ (no dense subtree) with WPGMA at 99% confidence level. Results are shown for various tree depths H_0: from 5 (lowest curve; corresponding to $n = 3 \cdot 2^5 = 96$ terminal taxa) to 23 (highest curve; corresponding to $n = 3 \cdot 2^{23} \approx 25 \cdot 10^6$). Although f_0 and g_0 may not be the shortest and longest branch lengths, they play the roles of f and g in the analysis (*24*). A normalized branch length of 1 corresponds to $g_0 = g^*$. We factor out the effect of f by fixing the value of f_0 to 2 (which explains the absence of the standard U-shaped form that is typical for such curves). The bold dashed line indicates the predicted sample complexity for an infinitely deep tree. The results were obtained from a Gaussian approximation with full variance-covariance matrix (*24*).

Fig. 3. On the example of Fig. 2A, a correct reconstruction of the deep triplet using UPGMA or WPGMA requires that the estimated distance between the clades A and C exceeds the estimated

distance between the clades A and B, or equivalently, that $\delta = \bar{\delta}_{\mathrm{u}}(A, C) - \bar{\delta}_{\mathrm{u}}(A, B) > 0$. Shown is the performance of the test $\{\bar{\delta}_{\mathrm{u}}(A, C) - \bar{\delta}_{\mathrm{u}}(A, B) > 0?\}$ over 3000 Jukes-Cantor simulations with $h_1 = H_0$, $h_0 = 1$, $g_0 = 0.7$, $f_0 = 2$, and sequence length $k = 500$ (where the subtrees corresponding to A, B and C are known). The top figure shows the variance of the estimator δ. The bottom figure shows the success rate. Confidence intervals at 95% are shown (using an approximation by the central limit theorem).