# Rooting Gene Trees without Outgroups: EP Rooting

Janet S. Sinsheimer[1,2,3], Roderick J. A. Little[4], and James A. Lake[1,5,*]

[1]Human Genetics Department, University of California, Los Angeles

[2]Biomathematics Department, University of California, Los Angeles

[3]Biostatistics Department, University of California, Los Angeles

[4]Biostatistics Department, University of Michigan School of Public Health

[5]Molecular, Cell and Developmental Biology, University of California, Los Angeles

*Corresponding author: E-mail: lake@mbi.ucla.edu.

**Accepted:** May 1, 2012

## Abstract

Gene sequences are routinely used to determine the topologies of unrooted phylogenetic trees, but many of the most important questions in evolution require knowing both the topologies and the roots of trees. However, general algorithms for calculating rooted trees from gene and genomic sequences in the absence of gene paralogs are few. Using the principles of evolutionary parsimony (EP) (Lake JA. 1987a. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol. 4:167–181) and its extensions (Cavender, J. 1989. Mechanized derivation of linear invariants. Mol Biol Evol. 6:301–316; Nguyen T, Speed TP. 1992. A derivation of all linear invariants for a nonbalanced transversion model. J Mol Evol. 35:60–76), we explicitly enumerate all linear invariants that solely contain rooting information and derive algorithms for rooting gene trees directly from gene and genomic sequences. These new EP linear rooting invariants allow one to determine rooted trees, even in the complete absence of outgroups and gene paralogs. EP rooting invariants are explicitly derived for three taxon trees, and rules for their extension to four or more taxa are provided. The method is demonstrated using 18S ribosomal DNA to illustrate how the new animal phylogeny (Aguinaldo AMA et al. 1997. Evidence for a clade of nematodes, arthropods, and other moulting animals. Nature 387:489–493; Lake JA. 1990. Origin of the metazoa. Proc Natl Acad Sci USA 87:763–766) may be rooted directly from sequences, even when they are short and paralogs are unavailable. These results are consistent with the current root (Philippe H et al. 2011. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. Nature 470:255–260).

**Key words:** rooting, trees, gene sequences, evolutionary parsimony, metazoa, Bayesian statistics, linear invariants.

## Introduction

Fully understanding the evolution of life on Earth requires identifying the ancestor from which all others on the tree evolved, that is, the root. A number of important unresolved questions in evolution, including the origins of modern humans (Reich et al. 2011), the rise of placental mammals (Waddell et al. 1999; Madsen et al. 2001; Murphy et al. 2001a, 2001b; Scally et al. 2001), animal evolution (Aguinaldo et al. 1997; Philippe et al. 2011), prokaryotic evolution (Cox et al. 2008; Lake 2009), and even the beginnings of life (Lake et al. 2009; Ragan et al. 2009), would benefit from phylogenetic methods that can root a tree without having to specify outgroups. And yet, there are no general methods available for determining roots. Trees can be rooted using outgroups and ancient gene duplications (Dayhoff et al. 1972), insertions and deletions (Rivera and Lake 1992), and occasionally gene presences and absences (Lake and Rivera 2004; Simonson et al. 2005). But, these methods cannot be applied to some of the most widely used molecular sequences, including ribosomal RNAs (rRNAs), because useful gene duplications are rare. Furthermore, usable gene duplications and indels are, in general, so infrequent that it is the exception when trees can be rooted using them.

Phylogenetic reconstruction algorithms typically delete rooting information because they assume, either implicitly or explicitly, that evolutionary distances are symmetric, that is, $d_{ij} = d_{ji}$, where $d_{ij}$ is the evolutionary distance between taxon

*i* and taxon *j*. Strictly speaking, the assumption of symmetric distances is only valid if evolutionary processes are time reversible—which they are not (Lake 1997). Assuming symmetric distances thus removes valuable rooting information from many types of sequence analyses. Here we show how to root nucleotide sequences without making this assumption. An alternative for rooting without an outgroup is to use a likelihood-based approach with a nonreversible transition state matrix (Barry and Hartigan 1987a, 1987b; Ferretti et al. 1994; Hendy and Penny 1996; Evans and Zhou 1998). That approach allows for a general model of evolutionary change to be used, however branch lengths, transition rates, and nucleotide distributions at the root must be estimated or integrated out. In contrast, our approach does not need to estimate these parameters, it is fast and simple to implement, and may be easily extended to larger numbers of taxa.

Evolutionary parsimony (EP) (Lake 1987a; 1987b) and its extensions (Cavender 1989; Nguyen and Speed 1992) contain rooting information because they do not assume that evolutionary distances are symmetric. Instead, they are based on the balanced transversion assumption, namely that transversions from purines/pyrimidines produce approximately equal numbers of pyrimidines/purines. Because DNA copying and repair mechanisms can most readily distinguish differences between the larger purines (A and G) and the smaller pyrimidines (C and U/T), those exchanges which substitute one purine for another or one pyrimidine for another (transitions) occur more frequently than those that substitute purines and pyrimidines (transversions), as has been known for nearly 30 years (Brown et al. 1982). In addition because transversions occur less frequently than transitions, they are more desirable for investigating early events in evolution. Thus, methods based on balanced transversions, or its modifications, provide both rooting and topological information, and the slow evolution of transversions makes them better suited for rooting.

Here we explicitly describe how to root trees using EP rooting invariants. The literature on both linear and polynomial phylogenetic invariants and the related topic of Hadamard conjugations, is rich, see as examples (Cavender and Felsenstein 1987; Lake 1987a; Cavender 1989; Nguyen and Speed 1992; Steel et al. 1993, 1998; Ferretti et al. 1994; Sinsheimer 1994; Steel and Fu 1995; Hendy and Penny 1996; Sinsheimer et al. 1996; Waddell et al. 1997; Evans and Zhou 1998; Allman and Rhodes 2003; Yap and Speed 2005), but to our knowledge there has been no work on using linear invariants to root a phylogenetic tree. Extending and simplifying Nguyen and Speed's (1992) results, we show that EP invariants can be classified into three major categories. These classes can be used to: (1) test the EP assumptions (the simplest example is a two-taxon test), (2) distinguish between rooted trees (the simplest example is a three-taxon test), and (3) distinguish between unrooted trees (the simplest example is a four-taxon test). The EP invariants that solely contain root information are explicitly derived, and posterior probabilities are developed to determine three-taxon rooted trees.

The method is illustrated using a simple, but difficult example, rooting the new animal phylogeny directly from short 18S rRNA sequences in the absence of outgroups.

## Results

### Principles of EP

For *s* aligned nucleic acid sequences, there are $4^s$ different nucleotide patterns that can be observed at any nucleotide position, ignoring insertions and deletions. Thus when three nucleic acid sequences are aligned there are $4^3 = 64$ possible patterns at each position of the alignment, for example, (AAA, AAG, . . .). For EP, the set of aligned sequences is represented as the number of times each of these patterns are observed over the entire sequence and are denoted as *xyz* where *x*, *y*, and *z* can take on nucleotide values A, G, C, and T (or U). $N_3$ is a vector with 64 entries containing the observed nucleotide count spectrum for three taxa, $N_3 = (\#AAA, \#AAG, \ldots, \#TTC, \#TTT)$, where # refers to the number of occurrences of each pattern. The components of $N_3$ for a hypothetical 30-mer sequence are shown in figure 1. In this example, the pattern TTT is observed three times so that $\#TTT = 3$.

EP's major assumption, balanced transversions, is a constraint placed on the transversion probabilities (Lake 1987a). It can be represented as four conditional probability statements:

$$P(A|C) = P(G|C), P(A|T) = P(G|T), P(C|G) = P(T|G), P(C|A) = P(T|A) \quad (1)$$

where $P(X|Y)$ denotes the probability of observing nucleotide X at some site in an existing organism's sequence given there was a Y at that site in the ancestral sequence. This constraint implies that for infinitely long sequences, an equal number of both possible transversions will occur along the path from an ancestor to an existing taxon. For example in figure 2, the expected number of A to C transversions will equal the expected number of A to T transversions between the root and taxon 1. In the next section, we illustrate how balanced transversions constrain some invariants to equal zero.

### Illustrating the Balanced Transversions Assumption: Two-Taxon Invariants

We demonstrate the application of EP rooting invariants by considering the simplest possible rooted tree, the two-taxon tree shown in figure 2. For two taxa there are two EP invariants and one possible rooted tree. These are:

$$U_1 = \#AC - \#AT - \#GC + \#GT, \quad (2)$$

and

```
N:          1           11          21          31
Taxon 1:    GTAACAAGGG  GTCTTGCAAG  CGCACTAAGC  --TGC
Taxon 2:    TTACGAAGGG  CTATTGCATA  CTAATTAGTC  CCATG
Taxon 3:    AAAAGCC---  CTAATGAAAT  CCTCGTCTTC  CCTCA

N=30 (Insertions and deletions are ignored.)

                    Observed Patterns:
#AAA = 2  #AAC = 4  #AGT = 1  #ACA = 1  #ATA = 1  #GGG = 1  #GGT = 1
#GCC = 1  #GTA = 1  #GTC = 2  #GTT = 1  #CAA = 1  #CAT = 1  #CGA = 1
#CGG = 1  #CCA = 1  #CCC = 2  #CTG = 1  #TAT = 1  #TTA = 2  #TTT = 3

U2   = R1Y2Y3 = #(A-G)1(C-T)2(C-T)3
     = #ACC - #ACT - #ATC + #ATT - #GCC + #GCT + #GTC - #GTT
     =   0  -   0  -   0  +   0  -   1  +   0  +   2  -   1
     =   0
```

**Fig. 1.**—Hypothetical aligned gene sequences for three taxa. An example illustrating how three aligned sequences can be reduced to a list of the informative EP subpatterns that specifies the number of counts supporting each pattern. These counts can then be used to reconstruct the most likely rooted tree for this hypothetical three-taxon comparison. The data in this example strongly support the G tree (rooted in taxon 3), as discussed in the text.

$$E(U_1^A) = E(\#R_1^A Y_2^A) = 0$$
$$E(U_2^A) = E(\#Y_1^A R_2^A) = 0$$

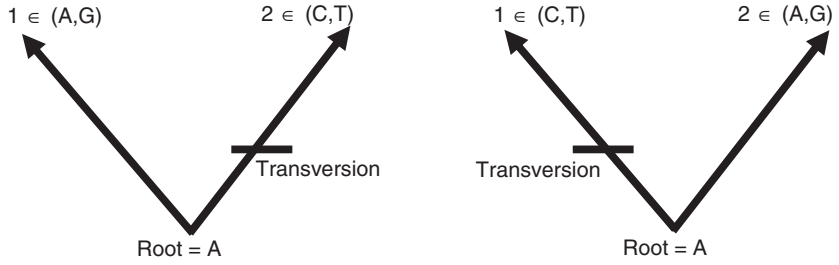


**Fig. 2.**—Two-taxon EP invariants. An example illustrating how balanced transversions require that the two-taxon invariants, $U_1$ and $U_2$, have expected values of zero. Patterns in $U_1 = \#R_1Y_2$ and $U_2 = \#Y_1R_2$ arise from a net transversion in branch 1 or a net transversion in branch 2, but not net transversions in both branches. This is true for all choices of ancestral nucleotide states. Under balanced transversions, the expected number of AC patterns equals the expected number of AT patterns, the expected number of GCs equals the expected number of GTs, the expected number of CAs equals the expected number of TAs, and the expected number of CGs equals the expected number of TGs. Therefore, $U_1$ and $U_2$ have expected values of zero, as is further explained in the text.

$$U_2 = \#CA - \#CG - \#TA + \#TG. \quad (3)$$

We use the shorthand notation $\#(x-y)_1(w-z)_2$ to represent $\#xw - \#xz - \#yw + \#yz$, so that $U_1$ can then be expressed as $\#(A-G)_1(C-T)_2$ and $U_2$ as $\#(C-T)_1(A-G)_2$. This notation can be extended to any number of taxa. For three sequences $\#(u-v)_1(w-x)_2(y-z)_3 = \#uwy - \#uwz - \#uxy + \#uxz - \#vwy + \#vwz + \#vxy - \#vxz$.

In figure 2, we illustrate why $U_1$ and $U_2$ have expected value zero under the balanced transversions assumption. First note that an odd number of transversions at a nucleotide site leads to a net transversion along the path, and an even number of transversions at a nucleotide site leads to either a net transition or no net change along the path. Consider the two-taxon, rooted tree in figure 2. The patterns AC, AT, GT, GC, CA, CG, TA, and TG can only be present if a net transversion has occurred along the path from the root to taxon 2 (branch 2), or along the path from the root to taxon 1 (branch 1), but not both (Lake 1987a).

$U_1$ and $U_2$ can be partitioned into four groups indexed by the unknown nucleotide present in the ancestral sequence. That is, $U_1 = U_1^{(A)} + U_1^{(G)} + U_1^{(C)} + U_1^{(T)}$ and $U_2 = U_2^{(A)} + U_2^{(G)} + U_2^{(C)} + U_2^{(T)}$, where $U^{(X)}$ denotes the counts derived from ancestral sequence positions containing nucleotide X. If one can show that the expected values of $U_1^{(X)}$

and $U_2^{(X)}$ are zero for all nucleotides X, then $U_1$ and $U_1$ have expected value zero.

Using figure 2, we demonstrate that the expected value of the partition $U_1(A)$ is zero. In the first tree in figure 2, the ancestral state is A, so that the nucleotide patterns that comprise $U_1^{(A)} = \#(A-G)_1^{(A)}(C-T)_2^{(A)}$ must have resulted from one net transversion in branch 2 (A to C, or A to T) and a net transition or no net change in branch 1 (A to G, or A to A). To simplify the discussion, we introduce the notation $Y_n = (C-T)_n = C_n - T_n$ for the pyrimidine difference operator, $R_n = (A-G)_n$ for the purine difference operator. Using this notation the two taxon invariants can be written as $U_1 = \#R_1Y_2 = \#(A-G)_1(C-T)_2$ and $U_2 = \#Y_1R_2$. (Note also that multiplication of these operators is commutative.) By the balanced transversions assumption, the number of net A to C transversions equals the number of net A to T transversions in branch 2. Consequently, the expected value of $\#Y_2^{(A)}$ is zero, and the expected value of $\#R_1^{(A)}Y_2^{(A)}$ is zero. By similar reasoning one can show that the expected values of $\#Y_2^{(G)}$, $\#R_1^{(C)}$, $\#R_1^{(T)}$ are zero, and therefore, $U_1^{(G)}$, $U_1^{(C)}$, and $U_1^{(T)}$ have expected value zero. Hence, the sum $U_1 = U_1^{(A)} + U_1^{(G)} + U_1^{(C)} + U_1^{(T)}$ has expected value zero under the balanced transversion assumptions. By symmetry, as shown on the right tree in figure 2, $U_2$ also has expected value zero when the balanced transversion assumptions hold.

## Determining the Roots of Three-Taxon Trees

The principles and reasoning used in the previous section can also be applied to derive the EP statistics for rooting trees. For three taxa, there are three possible rooted trees and three possible operators. The E tree is rooted in taxon 1, the F tree is rooted in taxon 2, and the G tree is rooted in taxon 3. In addition to the $Y_n$ and the $R_n$ operators discussed above, we utilize a third operator, $Z_n = (A+G-C-T)_n$, the transversion difference operator.

The derivation of the expected value of the rooted EP invariant, $U_{F2}$, is illustrated for the F and G trees in figure 3. Invariant $U_{F2}$ evaluates nucleotide patterns, $R_1Y_2R_3$, at positions where the nucleotides in sequences 1 and 3 are purines (Pu) and the nucleotide in sequence 2 is a pyrimidine (Py). When the G tree is correct, then 1 and 2 are sister taxa and $U_{F2}$ has expected value zero as will be demonstrated with reference to figure 3*A*.

We know from the two-taxon EP invariants, that a transversion must occur on either branch 1 or 2 in order to produce the $Pu_1Py_2$ pattern. This is true for all values of the nucleotide present at the node connecting taxa 1 and 2, as in figure 3*A*. Because the two-taxon EP invariant $U_1 = R_1Y_2$ has expected value zero for all four possible values, A, G, C, and T, of the most recent common ancestor of any two-taxon tree, we conclude that the 1, 2 clade of the three-taxon EP invariant, $U_{F1} = (R_1Y_2)\ R_3$ must also have zero expected value. Furthermore, because this value depends only on the value

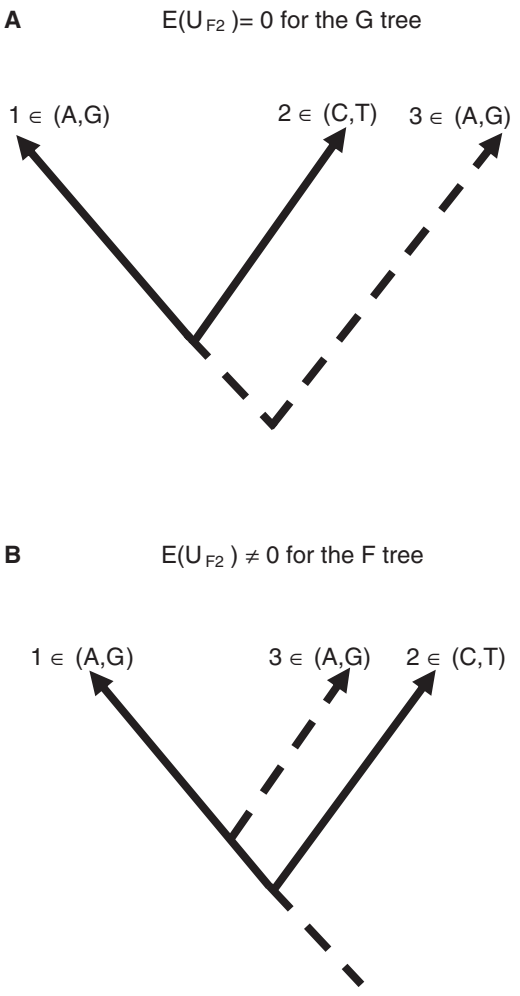


**Fig. 3.**—Three-taxon EP rooting invariants. An illustration of how three-taxon EP rooting invariants provide rooting information. In this example, for the $U_{F2}$ rooting invariant, the unknown nucleotides at the root of the tree and the interior node of the tree, as described in the text, must, be either pyrimidines at the root and purines at the node, Py/Pu, or *vice versa*. When the possible partitions of the roots that result in a transversion are computed, we find that $U_{F2}$ is unconstrained for the F tree, and that it is constrained to 0 expected value for the G tree and, in like manner, for the E tree (not shown). The expected values for all 12 invariants are summarized in table 1.

of the most recent common ancestor, located at the interior node of the F tree, it is therefore independent of any earlier ancestral values present at the root of the three-taxon tree. Thus, $U_{F2}$ has zero expected value for all possible combinations of nucleotides at the interior node and at the root of the G tree (and similarly for the E tree—not shown). Thus for the E and G trees (but not for the F tree):

$$U_{F2} = \sum_{\substack{X \in (Pu,Py) \\ W \in (Pu,Py)}} U_{F2}^{(X)(W)} = 0 \quad \text{(for E and G trees only)} \tag{4}$$

In contrast, when the F tree is correct, see figure 3*B*, the three-taxon rooting invariant, $U_{F2} = R_1Y_2R_3$ is not constrained to zero. This happens because the terminal two-taxon tree relating taxa 1 and 3, corresponds to the operator $R_1R_3$ which is not an EP invariant, and hence is unconstrained. Furthermore, the operator related to the branch leading to taxon 2, $Y_2$, is also unconstrained because transversions have not necessarily occurred in branch 2. Thus their product, $U_{F2} = R_1Y_2R_3$, is unconstrained for the F tree.

The following 12 three-taxon EP statistics contain root information:

$$U_{E1} = \#R_1Y_2Y_3 \quad (5)$$
$$U_{E2} = \#Y_1R_2R_3 \quad (6)$$
$$U_{F1} = \#Y_1R_2Y_3 \quad (7)$$
$$U_{F2} = \#R_1Y_2R_3 \quad (8)$$
$$U_{G1} = \#Y_1Y_2R_3 \quad (9)$$
$$U_{G2} = \#R_1R_2Y_3 \quad (10)$$
$$U_{EF1} = \#R_1Y_2Z_3 \quad (11)$$
$$U_{EF2} = \#Y_1R_2Z_3 \quad (12)$$
$$U_{EG1} = \#R_1Z_2Y_3 \quad (13)$$
$$U_{EG2} = \#Y_1Z_2R_3 \quad (14)$$
$$U_{FG1} = \#Z_1R_2Y_3 \quad (15)$$
$$U_{FG2} = \#Z_1Y_2R_3 \quad (16)$$

There are two types of rooting statistics for three-taxon trees. $U_{E1}$ is a representative of one type and $U_{EF1}$ is a representative of the other type. It is easily seen from the definitions above, that $U_{E1}$, $U_{F1}$, and $U_{G1}$ are permutations of a common pattern, and similarly for $U_{E2}$, $U_{F2}$, and $U_{G2}$. Likewise, $U_{EF1}$, $U_{EF2}$, $U_{EG1}$, $U_{EG2}$, $U_{FG1}$, and $U_{FG2}$ are permutations of the other common pattern.

Unlike the two-taxon EP invariants, the expected values of these three-taxon EP invariants depend on the position of the root. If the E (or F, or G) trees are correct and if the assumptions of EP are met, then the 12 invariants will have the expected values that are summarized in table 1.

## Analyzing the Hypothetical Example in Figure 1

If we return to the data generated in the hypothetical example shown in figure 1, we can now estimate the root location from the hypothetical sequences. From these data, we calculate the values of each of the invariants, as shown in table 2. The expected values of these EP root invariants under each of the possible trees are also shown in the table. Clearly, the observed values most closely match the values expected if the G tree is correct, so we predict that the G tree is the most probable tree. In Material and Methods, we formalize this prediction by determining the posterior probability, the probability of a tree given the observed sequence data. We find that the probability of the G tree is 99.68%, the probability of the F tree is 0.20%, and the probability of the E tree is 0.12%, results that overwhelmingly support the G tree.

EP invariants are not restricted to two-, three-, and four-taxon trees (Lake 1987a; Cavender 1989; Nguyen and Speed 1992; Sinsheimer 1994), but can be extended to any number of taxa using the notation developed here (see Sinsheimer 1994). We classify the EP statistics for any number of taxa into three major categories according to their use in statistical inference, namely: (1) for tests of EP assumptions, (2) for distinguishing between rooted trees, and (3) for distinguishing between unrooted trees, topology (see Materials and Methods for details).

## A Second Example: Rooting rRNA Trees

Ribosomal rRNA sequences are particularly difficult to root because the genes are short, paralogous genes are lacking, and nucleotides evolve faster than amino acids often making nucleotide sequences too divergent to be useful. Here we demonstrate the usefulness of EP rooting using a particularly difficult example: rooting a deep, three-taxon metazoan tree using only partial 18S rRNA sequences.

Today the root of the multicellular animals is fairly well known, and even the most challenging, deeper branching parts of the metazoan tree are being reconstructed (Philippe et al. 2011), unlike when the lophophorates were initially shown to be protostomes (Halanych et al. 1995). Here, using short 18S ribosomal DNA sequences and in the complete absence of an outgroup, we show that the lophophorates are protostomes, using three-taxon, EP rooting analyses.

In figure 4, three rooted trees correspond to the hypothesis that the lophophorates are protostomes (G tree), deuterostomes (F tree), or an independent, earlier branching lineage (E tree). Using slowly evolving 18S rRNA sequences of representatives of these three groups, namely the inarticulate brachiopod lophophorate, *Glottidia pyramidata;* the bivalve protostome, *Placopecten magellanicus*; and the echinoderm

**Table 1**
Expected Values of the Three-Taxon EP Rooting Invariants

| TREE | $U_{E1}$ | $U_{E2}$ | $U_{F1}$ | $U_{F2}$ | $U_{G1}$ | $U_{G2}$ | $U_{EF1}$ | $U_{EF2}$ | $U_{EG1}$ | $U_{EG2}$ | $U_{FG1}$ | $U_{FG2}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **E** | **U** | **U** | 0 | 0 | 0 | 0 | **U** | **U** | **U** | **U** | 0 | 0 |
| **F** | 0 | 0 | **U** | **U** | 0 | 0 | **U** | **U** | 0 | 0 | **U** | **U** |
| **G** | 0 | 0 | 0 | 0 | **U** | **U** | 0 | 0 | **U** | **U** | **U** | **U** |

Note.—**U:** unconstrained expected value.

**Table 2**
Expected and Observed EP Rooting Invariants for Hypothetical Sequence

| If E tree is correct | | | If F tree is correct | | | If G tree is Correct | | |
|---|---|---|---|---|---|---|---|---|
| **Invariant** | **Expect** | **Observe** | **Invariant** | **Expect** | **Observe** | **Invariant** | **Expect** | **Observe** |
| $U_{F1}$ | 0 | 0 | $U_{E1}$ | 0 | 0 | $U_{E1}$ | 0 | 0 |
| $U_{F2}$ | 0 | 1 | $U_{E2}$ | 0 | 1 | $U_{E2}$ | 0 | 1 |
| $U_{G1}$ | 0 | 4 | $U_{G1}$ | 0 | 4 | $U_{F1}$ | 0 | 0 |
| $U_{G2}$ | 0 | 4 | $U_{G2}$ | 0 | 4 | $U_{F2}$ | 0 | 1 |
| $U_{FG1}$ | 0 | 8 | $U_{EG1}$ | 0 | 6 | $U_{EF1}$ | 0 | 1 |
| $U_{FG2}$ | 0 | −1 | $U_{EG2}$ | 0 | 3 | $U_{EF2}$ | 0 | −1 |

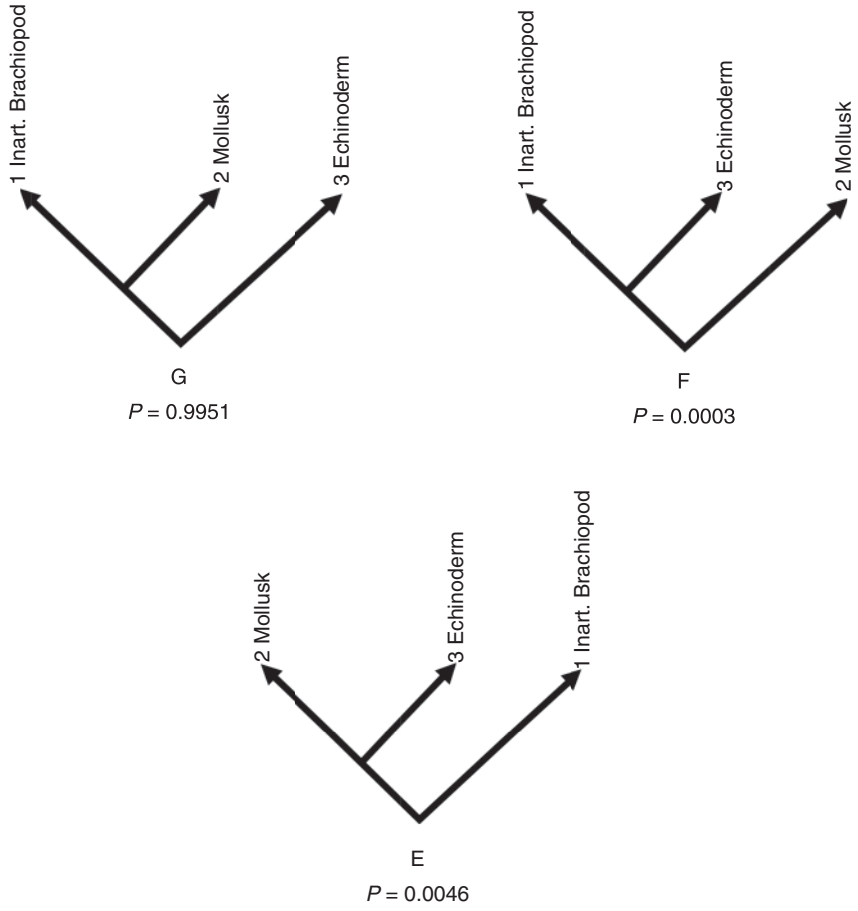


**Fig. 4.**—Rooting a three-taxon metazoan tree. Posterior Bayesian support for the three possible, three-taxon, rooted metazoan trees obtained by analysis of short, nonparalogous, 18S rDNA sequences is listed beneath each of the three rooted trees. In this example, the G tree, rooted in the branch leading to the echinoderm, is strongly supported by the data ($P = 99.51\%$), consistent with the current multicellular animal root, whereas the E and F trees are not significantly supported, $P = 0.46\%$ and 0.03%, respectively.

deuterostome, *Antedon serrata*. We first test the balanced transversion assumption using the six goodness-of-fit invariants for three taxa (see Materials and Methods). The test statistic has a chi-square distribution with 6 degrees of freedom (Mood AM, 1974; Nguyen T, 1992). The null hypothesis of balanced transversions is not rejected at the 5% significance level in these data.

We then determine the posterior probability of each of the rooted trees. There are 76 informative positions out of a total of 1457 aligned 18S ribosomal DNA positions, that may be used for determining the rooted tree that relates these three organisms. The rooting statistics are shown table 3 for those invariants that are constrained to zero expected value for the E, F, and G trees, respectively. As described in the Materials

**Table 3**
Expected and Observed Invariant Values for *G. pyramidata, P. magellanicus,* and *A. serrata*

| If E tree is correct | | | If F tree is correct | | | If G tree is correct | | |
|---|---|---|---|---|---|---|---|---|
| **Invariant** | **Expected** | **Observed** | **Invariant** | **Expected** | **Observed** | **Invariant** | **Expected** | **Observed** |
| $U_{F1}$ | 0 | 1 | $U_{E1}$ | 0 | **−4** | $U_{E1}$ | 0 | −4 |
| $U_{F2}$ | 0 | −1 | $U_{E2}$ | 0 | 0 | $U_{E2}$ | 0 | 0 |
| $U_{G1}$ | 0 | −7 | $U_{G1}$ | 0 | −7 | $U_{F1}$ | 0 | 1 |
| $U_{G2}$ | 0 | 5 | $U_{G2}$ | 0 | 5 | $U_{F2}$ | 0 | −1 |
| $U_{FG1}$ | 0 | −2 | $U_{EG1}$ | 0 | 13 | $U_{EF1}$ | 0 | −1 |
| $U_{FG2}$ | 0 | 10 | $U_{EG2}$ | 0 | −3 | $U_{EF2}$ | 0 | 1 |

and Methods, we find that the posterior probability is 99.51% for the G tree, the tree rooted in the branch leading to the echinoderm; 0.03% for the F tree, the tree rooted in the branch leading to the bivalve; and 0.46% for the E tree, the tree rooted in the branch leading to the inarticulate brachiopod. Because EP rooting uses a different type of sequence information than methods designed to test topologies, these analyses provide independent support for the, now well-known result, that the lophophorates are protostomes, and not deuterostomes (Halanych et al. 1995; Philippe et al. 2011).

## Discussion

EP rooting can recover rooted trees directly from nucleotide sequences in the absence of outgroups, even when sequences are relatively short. Given the large amounts of sequence data available, this raises that possibility that outstanding problems in rooted bilateral animal relationships may be resolved using EP rooting. Furthermore, these rooting analyses are quite well suited for Bayesian interpretations.

The three-taxon test, important and useful in its own right, also illustrates the use of EP rooting invariants to target-specific aspects of phylogenetic reconstruction. EP invariants can be classified into three groups based on the phylogenetic information they contain (Sinsheimer 1994). One class is the goodness-of-fit invariants. A second class contains the invariants with topological information but no root information, and a third class contains the statistics with both root and topological information.

As is now well known (Sinsheimer et al. 1996), Bayesian predictions, such as those presented here, provide an attractive alternative to classical hypothesis tests for phylogenetic reconstruction. Multiple hypotheses, as in the example here, where three possible trees exist, are better suited to Bayesian analysis than to classical hypothesis testing (Sinsheimer 1994; Sinsheimer et al. 1996). Further, posterior probabilities estimate the probability of each of the trees given the observed data, a result that is far easier to interpret than the *P* value outcomes of classical hypothesis testing (Burnham KP, 1998).

A major shortcoming of current phylogenetic analyses is that the roots of many major groups are currently unknown due primarily to the difficulty of obtaining useful outgroups. And in some cases, such as the origin of life, outgroups are simply not available. We hope that the EP rooting invariants make even deeper explorations of the origin of life possible.

## Materials and Methods

### General Rules for Defining the Classes of EP Invariants

The same simplifying notation introduced in the previous sections can also be used for EP statistics of any number of taxa. EP statistics are a form of linear invariants. All the work described in this article refers to linear invariants, that is, invariants which allow summing over nucleotide positions. The term linear invariants will therefore be shortened to invariants. In the general *s*-taxon case, EP statistics are linear combinations of counts made up of the building blocks, $U = \#X_1X_2X_3 \ldots X_s$, where for at least one $i \in \{1, \ldots, s\}$ $X_i = (A-G) = R$, for at least one $j \in \{1, \ldots, s\}$ $X_j = (C-T) = Y$, and for all other $k \in \{1, \ldots, s\}$ $X_k \in \{R, Y, S, Z\}$,where $S = A+G+C+T$ and $Z = A+G-C-T$ (Cavender 1989; Nguyen and Speed 1992; Sinsheimer 1994; Sinsheimer et al. 1996). When $X_k = (A+G+C+T) = S$, the nucleotides for the *k*-th taxon are summed over, effectively ignoring that branch. Additional EP assumptions are necessary to preserve balanced transversions when the *k*-th taxon is ignored. These restrictions are $P(A|T) + P(C|T) = P(A|C) + P(T|C)$ and $P(T|A) + P(G|A) = P(T|G) + P(A|G)$ and similar constraints exist for the proportionally balanced transversion model (Cavender 1989; Nguyen and Speed 1992).

In the *s*-taxon case, where $s > 2$, EP statistics can be partitioned into three classes. The first class is comprised of goodness-of-fit statistics that are linear combinations of counts, $U = \#X_1X_2X_3 \ldots X_s$ where for exactly one $i \in \{1, \ldots, s\}$ $X_i = R$, for exactly one $j \in \{1, \ldots, s\}$ $X_j = Y_j$, and for all other $k \in (1, \ldots, s\}$ $X_k = (A+G+C+T)_k$. The second class are statistics containing topological information but no rooting information and are linear combinations of counts, $U = \#X_1X_2X_3 \ldots X_s$ where for exactly two taxa $i, j \in \{1, \ldots, s\}$ $X_i = X_j = R$, for exactly two taxa $m, n \in \{1, \ldots, s\}$ $X_m = X_n = Y$, and for all

other $k \in (1, \ldots, s\}$ $X_k = S$. The statistics containing both root position and topological information comprise the third class and include all EP statistics that are not in Class 1 or 2. Examples of Class 3 statistics include the three-taxon statistics (4)–(15), the four-taxon statistic $\#R_1Y_2S_3Z_4$ and the five-taxon statistic $S_1R_2Y_3R_4S_5$. In the interest of saving space these classifications are not proven here, but can be found elsewhere (Sinsheimer 1994; Sinsheimer et al. 1996).

The hypothesis tests for goodness-of-fit using the Class 1 statistics (Nguyen and Speed 1992), and for topology using Class 2 statistics are easily generalized to $s$ taxa (Sinsheimer 1994). For more than three taxa, the $s$-taxon statistics containing rooting information can also be used to infer the correct rooted tree. In practice, however, inference becomes much more complicated. The expected value patterns of these statistics contain topological information as well as root information. In addition, the number of statistics increases rapidly as the number of taxa increases. For example, there are 92 statistics that contain information to infer the correct four-taxon rooted tree.

## Statistical Inference Using EP Rooting Invariants

In this section we derive the posterior probabilities for inference of the correct three-taxon rooted tree. The three possible rooted trees correspond to three alternative hypotheses (figs. 3 and 4), namely $H_E$: tree E is the true tree, $H_F$: tree F is the true tree, and $H_G$: tree G is the true tree. Under the principles of EP, the hypotheses can be expressed in terms of the expected values of the components of U (table 1). We assume all three trees are equally probable in the absence of prior sequence data, that is, $P(H_r) = 1/3$ where $r \in \{E,F,G\}$. By Bayes theorem, the probability of a rooted tree given the observed data can be expressed as:

$$P(H_r|U) = \frac{P(U|H_r)}{\sum_{t\in\{E,F,G\}} P(U|H_t)} \tag{17}$$

where $P(U|H_r)$ is the probability of the observed sequence data given tree $r$ is the true tree. Because we assume each of the rooted trees is equally probable a priori, these posterior probabilities are equivalent to Akaike weights (Burnham and Anderson 1998).

Let $\mathbf{M}$ be the vector of expected values, $E(U)$ and let $\mu_k$ the expected value of $k$-th statistic comprising U, $U_k$. One approach to inference is to approximate $P(U|H_r)$ by the multivariate normal density, $f(U|H_r, \hat{M}_r, \Omega_r)$ where $\hat{M}_r$, is the sample estimate of $E(U)$ under $H_r$ and $\Omega_r$ is the sample estimate of the 12 by 12 variance–covariance matrix under $H_r$ (Mood et al. 1974). Equation (17) then yields

$$P(H_r|U) = \frac{|\Omega_r|^{-1/2} \exp\left(-1/2(U - M_r)^T \Omega_r^{-1}(U - M_r)\right)}{\sum_{t\in\{E,F,G\}} |\Omega_t|^{-1/2} \exp\left(-1/2(U - M_t)^T \Omega_t^{-1}(U - M_t)\right)} \tag{18}$$

The matrix $\Omega_r$ is composed of estimates of the variances and covariances of the EP statistics. Expressions for these estimates are derived in the next section.

## Variances and Covariances of the Three-Taxon Rooting Invariants

We first formalize the definition of an EP statistic. For $s$ aligned taxa, there are $4^s$ possible combinations of the four nucleotides at each site. The data consist of the observed nucleotide spectrum $\mathbf{N_s}$, a $4^s$ vector tallying each of these combinations over the observed sites. The sum of the counts, $N = \sum_{i=1}^{4^s} N_i$, is equal to the total number of aligned nucleotides. Ignoring site-to-site variation, $\mathbf{N_s}$ is modeled as a $4^s$ multinomial. We restrict attention to statistics that are linear combinations of these counts, with coefficients $-1$, 0, or 1. Let $U_r$ denote such a statistic. $U_r$ can be written concisely as vector products, $\langle N, V\rangle = \sum_{i=1}^{4^s} N_i V_i$, where $\mathbf{V}$ is a vector of length $4^s$ whose components that are $-1$, 0, or 1. $U_r$ is an EP statistic of tree $\tau$ if its expected value is zero when tree $\tau$ is the true tree. An invariant vector of tree $\tau$, $\mathbf{V_\tau}$, is any non-zero vector that, when multiplied by $\mathbf{N_s}$ generates an EP statistic $U_r$ when tree $\tau$ is the true tree. For example, the two-taxon statistic, $U_1$ is the vector product of $\mathbf{N_2}$ and $\mathbf{V_1} = (0,0,1,-1,0,0,-1,1,0,0,0,0,0,0,0,0)$, and $U_2$ is the product of $\mathbf{N_2}$ and $\mathbf{V_2} = (0,0,0,0,0,0,0,0,1,-1,0,0,-1,1,0,0)$.

For linear invariants, $U_j$ and $U_k$ the expected value of $U_j$, the variance of $U_j$ and the covariance of $U_j$ and $U_k$ can be calculated by recalling that for multinomial counts, $N_m$ and $N_q$, $E(N_m) = Np_m$, $Var(N_m) = Np_m(1-p_m)$, and $Cov(N_m, N_q) = -Np_mp_q$ (Mood et al. 1974).

$$\mu_j = E(U_j) = N\sum_{i=1}^{4^s} V_{j,i}p_i \tag{19}$$

$$Var(U_j) = N\left(\sum_{i=1}^{4^s} V_{j,i}^2 p_i - \left(\sum_{i=1}^{4^s} V_{j,i}p_i\right)^2\right) \tag{20}$$

$$Cov(U_j, U_k) = N\left(\sum_{i=1}^{4^s} V_{j,i}V_{k,i}p_i - \sum_{i=1}^{4^s} V_{j,i}p_i \sum_{i=1}^{4^s} V_{k,i}p_i\right) \tag{21}$$

Estimates of the expected value, variance, and covariance follow from substituting the sample estimate of $p_i$, $p_i = N_i/N$, into equations (19–21):

$$\hat{\mu}_j = \sum_{i=1}^{4^s} V_{j,i}N_i \tag{22}$$

$$\hat{Var}(U_j) = \sum_{i=1}^{4^s} V_{j,i}^2 N_i - \frac{\hat{\mu}_j^2}{N} \tag{23}$$

$$\hat{Cov}(U_j, U_k) = \sum_{i=1}^{4^s} V_{j,i}V_{k,i}N_i - \frac{\hat{\mu}_j\hat{\mu}_k}{N} \tag{24}$$

When $\mu_j$ is constrained to be zero, (22)–(24) reduce to:

$$\hat{\mu}_j = 0 \tag{25}$$

$$\hat{Var}(U_j) = \sum_{i=1}^{4^s} V_{j,i}^2 N_i \tag{26}$$

$$\hat{Cov}(U_j, U_k) = \sum_{i=1}^{4^s} V_{j,i} V_{k,i} N_i \tag{27}$$

Equations (25–27) were used to construct the test statistics (eqs. 5–16, 17) for the three-taxon case. The expressions for the variances and covariances can be easily calculated by introducing rules of multiplication for the simplified notation based on the Kronecker product representation of V and $\sum_{i=1}^{4^s} V_{ji} V_{ki} N_i = U_j \bullet U_k$ where $\bullet$ is multiplication branch by branch,

$$X_1Y_2Z_3 \bullet R_1S_2T_3 = (X * R)_1(Y * S)_2(Z * T)_3 \tag{28}$$

Within any branch *i*, the rules of multiplication, *, are:

$$(A - G) * (A - G) = (A+G) \tag{29}$$

$$(C - T) * (C - T) = (C+T) \tag{30}$$

$$(A+G+C+T) * (A+G+C+T) = (A+G+C+T) \tag{31}$$

$$(A - G) * (C - T) = (C - T) * (A - G) = 0 \tag{32}$$

$$\begin{aligned}(A - G) * (A+G+C+T) \\ = (A+G+C+T) * (A - G) = (A - G)\end{aligned} \tag{33}$$

$$\begin{aligned}(C - T) * (A+G+C+T) \\ = (A+G+C+T) * (C - T) = (C - T)\end{aligned} \tag{34}$$

$$(A+G - C - T) * (A+G - C - T) = (A+G+C+T) \tag{35}$$

$$\begin{aligned}(A - G) * (A+G - C - T) \\ = (A+G - C - T) * (A - G) = (A - G)\end{aligned} \tag{36}$$

$$\begin{aligned}(C - T) * (A+G - C - T) \\ = (A+G - C - T) * (C - T) = -(C - T)\end{aligned} \tag{37}$$

$$\begin{aligned}(A+G - C - T) * (A+G+C+T) = (A+G+C+T) \\ * (A+G - C - T) = (A+G - C - T).\end{aligned} \tag{38}$$

where we have dropped the subscript *i* to reduce the clutter in the notation. If for any branch *i* $(W*V)_i = 0$, then $U_j \bullet U_k = 0$. Using these rules, any of the variance or covariance terms of matrix $\Omega_r$ (eq. 18) can be determined, for example, using the transversion, purine and pyrimidine difference operator notation:

$$\begin{aligned}\hat{Cov}(U_{E1}, U_{EF2}) &= \#R_1Y_2Y_3 \bullet \#Y_1R_2X_3 - \frac{\hat{\mu}_{E1}\hat{\mu}_{EF2}}{N} \\ &= \#0_1 0_2 (T - C)_3 - \frac{\hat{\mu}_{E1}\hat{\mu}_{EF2}}{N} \\ &= -\frac{\hat{\mu}_{E1}\hat{\mu}_{EF2}}{N}\end{aligned}$$

Following the same logic, we can derive the expressions for the covariances and variances for three taxa:

$$\hat{Var}(U_{E1}) = \#(A+G)_1(C+T)_2(C+T)_3 - \frac{\hat{\mu}_{E1}^2}{N}$$

$$\hat{Var}(U_{E2}) = \#(C+T)_1(A+G)_2(A+G)_3 - \frac{\hat{\mu}_{E2}^2}{N}$$

$$\hat{Var}(U_{F1}) = \#(C+T)_1(A+G)_2(C+T)_3 - \frac{\hat{\mu}_{F1}^2}{N}$$

$$\hat{Var}(U_{F2}) = \#(A+G)_1(C+T)_2(A+G)_3 - \frac{\hat{\mu}_{F2}^2}{N}$$

$$\hat{Var}(U_{G1}) = \#(C+T)_1(C+T)_2(A+G)_3 - \frac{\hat{\mu}_{G1}^2}{N}$$

$$\hat{Var}(U_{G2}) = \#(A+G)_1(A+G)_2(C+T)_3 - \frac{\hat{\mu}_{G2}^2}{N}$$

$$\hat{Var}(U_{EF1}) = \#(A+G)_1(C+T)_2(A+G+C+T)_3 - \frac{\hat{\mu}_{EF1}^2}{N}$$

$$\hat{Var}(U_{EF2}) = \#(C+T)_1(A+G)_2(A+G+C+G)_3 - \frac{\hat{\mu}_{EF2}^2}{N}$$

$$\hat{Var}(U_{EG1}) = \#(A+G)_1(A+G+C+T)_2(C+T)_3 - \frac{\hat{\mu}_{EG1}^2}{N}$$

$$\hat{Var}(U_{EG2}) = \#(C+T)_1(A+G+C+T)_2(A+G)_3 - \frac{\hat{\mu}_{EG2}^2}{N}$$

$$\hat{Var}(U_{FG1}) = \#(A+G+C+T)_1(A+G)_2(C+T)_3 - \frac{\hat{\mu}_{FG1}^2}{N}$$

$$\hat{Var}(U_{FG2}) = \#(A+G+C+T)_1(C+T)_2(A+G)_3 - \frac{\hat{\mu}_{FG2}^2}{N}$$

$$\hat{Cov}(U_{E1}, U_{EF1}) = -\#(A+G)_1(C+T)_2(C - T)_3 - \frac{\hat{\mu}_{E1}\hat{\mu}_{EF1}}{N}$$

$$\hat{Cov}(U_{E1}, U_{EG1}) = -\#(A+G)_1(C - T)_2(C+T)_3 - \frac{\hat{\mu}_{E1}\hat{\mu}_{EG1}}{N}$$

$$\hat{Cov}(U_{E2}, U_{EF2}) = \#(C+T)_1(A+G)_2(A - G)_3 - \frac{\hat{\mu}_{E2}\hat{\mu}_{EF2}}{N}$$

$$\hat{Cov}(U_{E2}, U_{EG2}) = \#(C+T)_1(A - G)_2(A+G)_3 - \frac{\hat{\mu}_{E2}\hat{\mu}_{EG2}}{N}$$

$$\hat{Cov}(U_{F1}, U_{EF2}) = -\#(C+T)_1(A+G)_2(C - T)_3 - \frac{\hat{\mu}_{F1}\hat{\mu}_{EF2}}{N}$$

$$\hat{Cov}(U_{F1}, U_{FG1}) = -\#(C - T)_1(A+G)_2(C+T)_3 - \frac{\hat{\mu}_{F1}\hat{\mu}_{FG1}}{N}$$

$$\hat{Cov}(U_{F2}, U_{EF1}) = \#(A+G)_1(C+T)_2(A - G)_3 - \frac{\hat{\mu}_{F2}\hat{\mu}_{EF1}}{N}$$

$$\hat{Cov}(U_{F2}, U_{FG2}) = \#(A - G)_1(C+T)_2(A+G)_3 - \frac{\hat{\mu}_{F2}\hat{\mu}_{FG2}}{N}$$

$$\hat{Cov}(U_{G1}, U_{EG2}) = -\#(C+T)_1(C - T)_2(A+G)_3 - \frac{\hat{\mu}_{G1}\hat{\mu}_{EG2}}{N}$$

$$\hat{Cov}(U_{G1}, U_{FG2}) = -\#(C - T)_1(C+T)_2(A+G)_3 - \frac{\hat{\mu}_{G1}\hat{\mu}_{FG2}}{N}$$

$$\hat{Cov}(U_{G2}, U_{EG1}) = \#(A+G)_1(A - G)_2(C+T)_3 - \frac{\hat{\mu}_{G2}\hat{\mu}_{EG1}}{N}$$

$$\hat{Cov}(U_{G2}, U_{FG1}) = \#(A - G)_1(A+G)_2(C+T)_3 - \frac{\hat{\mu}_{G2}\hat{\mu}_{FG1}}{N}$$

and for all other $U_j$ and $U_k$ combinations, $\hat{Cov}(U_j, U_k) = -\frac{\hat{\mu}_j\hat{\mu}_k}{N}$

### Goodness-of-Fit Invariants and Test Statistics

For three taxa, there are 6 goodness-of-fit invariants, $U_{12A} = \#R_1Y_2S_3$, $U_{12B} = \#Y_1R_2S_3$, $U_{13A} = \#R_1S_2Y_3$, $U_{13B} = \#Y_1S_2R_3$, $U_{23A} = \#S_1R_2Y_3$, and $U_{23B} = \#S_1Y_2R_3$ (Nguyen 1992; Sinsheimer 1994.). Let U be the vector of goodness-of-fit invariants. We approximate the density of U with a multivariate normal density and denote the sample estimate of the six by six variance–covariance matrix as $\hat{\Omega}$. The test statistic has a quadratic form, $U^T\hat{\Omega}^{-1}U$, and has a chi-square density with 6 degrees of freedom (Rao 1973), which provides a test of the null hypothesis that all the goodness-of-fit invariants are zero. (More generally, when there are $r$ taxa there are $r(r-1)$ goodness-of-fit invariants and the corresponding test statistic has a chi-square density with $r(r-1)$ degrees of freedom).

As in the case of the rooting invariants, we use the within branch multiplication rules (in particular eqs. (29–34)) to derive the entries of $\hat{\Omega}$. The variance of $U_{12A}$ is $\hat{Var}(U_{12A}) = \#(A+G)_1(C+T)_2(A+G+C+T)_3 - \frac{\hat{\mu}^2_{12A}}{N}$ and the variance of $U_{12B}$ is $\hat{Var}(U_{12B}) = \#(C+T)_1(A+G)_2(A+G+C+T)_3 - \frac{\hat{\mu}^2_{12B}}{N}$. The variances for the other goodness-of-fit invariants have the same forms, just with permuted taxon labels. The dependence among invariants is reflected in the covariances, the off diagonal terms of this matrix:

$$\hat{Cov}(U_{12A}, U_{13A}) = \#(A+G)_1(C-T)_2(C-T)_3 - \frac{\hat{\mu}_{12A}\hat{\mu}_{13A}}{N}$$

$$\hat{Cov}(U_{12B}, U_{13B}) = \#(C+T)_1(A-G)_2(A-G)_3 - \frac{\hat{\mu}_{12B}\hat{\mu}_{13B}}{N}$$

$$\hat{Cov}(U_{12A}, U_{23B}) = \#(A-G)_1(C+T)_2(A-G)_3 - \frac{\hat{\mu}_{12A}\hat{\mu}_{23B}}{N}$$

$$\hat{Cov}(U_{12B}, U_{23A}) = \#(C-T)_1(A+G)_2(C-T)_3 - \frac{\hat{\mu}_{12B}\hat{\mu}_{23A}}{N}$$

$$\hat{Cov}(U_{13A}, U_{23A}) = \#(A-G)_1(A-G)_2(C+T)_3 - \frac{\hat{\mu}_{13A}\hat{\mu}_{23A}}{N}$$

$$\hat{Cov}(U_{13B}, U_{23B}) = \#(C-T)_1(C-T)_2(A+G)_3 - \frac{\hat{\mu}_{13B}\hat{\mu}_{23B}}{N}$$

and for all other $U_j$ and $U_k$ combinations, $\hat{Cov}(U_j, U_k) = -\frac{\hat{\mu}_j\hat{\mu}_k}{N}$ (Nguyen T, 1992; Sinsheimer, 1994.) (Rao, 1973).

## Acknowledgments

We thank Ms. Brooke Sarna for discussions and drafting an initial version of figure 3. This work was supported by grants from the NSF (DEB-0719574) and from the NASA Astrobiology Institute Directors Discretionary Fund to J.A.L.

## Literature Cited

Aguinaldo AMA, et al. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493.
Allman ES, Rhodes JA. 2003. Phylogenetic invariants for the general Markov model of sequence mutation. Math Biosci. 186:113–144.
Barry D, Hartigan JA. 1987a. Asynchronous distance between homologous DNA sequences. Biometrics 43:261–276.
Barry D, Hartigan JA. 1987b. Statistical analysis of Hominoid molecular evolution. Stat Sci. 2:191–210.
Brown WM, Prager EM, Wang A, Wilson AC. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol. 18: 225–239.
Burnham KP, Anderson DR. 1998. Model selection and inferences: a practical information theoretic approach. New York: Springer-Verlag.
Cavender J. 1989. Mechanized derivation of linear invariants. Mol Biol Evol. 6:301–316.
Cavender J, Felsenstein J. 1987. Invariants of phylogenies in a simple case with discrete states. J Classif. 4:57–71.
Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A. 105: 20356–20361.
Dayhoff MO, Hunt LT, McLaughlin PJ, Jones DD. 1972. Gene duplications in evolution. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington (DC): National Biomedical Research Foundation, p. 17–30.
Evans SN, Zhou X. 1998. Constructing and counting phylogenetic invariants. J Comput Biol. 5:713–724.
Ferretti V, Lang BF, Sankoff D. 1994. Skewed base compositions, asymmetric transition matrices, and phylogenetic invariants. J Comput Biol. 1:77–92.
Halanych KM, et al. 1995. Evidence from 18s ribosomal DNA that the lophophorates are protostome animals. Science 267: 1641–1643.
Hendy MD, Penny D. 1996. Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. J Comput Biol. 3:19–31.
Lake JA. 1987a. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol. 4:167–181.
Lake JA. 1987b. Determining evolutionary distances from highly diverged nucleic acid sequences—operator metrics. J Mol Evol. 24:59–73.
Lake JA. 1990. Origin of the Metazoa. Proc Natl Acad Sci U S A. 87: 763–766.
Lake JA. 1997. Phylogenetic inference: how much evolutionary history is knowable. Mol Biol Evol. 14:213–219.
Lake JA. 2009. Evidence for an early prokaryotic endosymbiosis. Nature 460:967–971.
Lake JA, Rivera MC. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. Mol Biol Evol. 21:681–690.
Lake JA, Skophammer RG, Herbold CW, Servin JA. 2009. Genome beginnings: rooting the tree of life. Phil Trans R Soc B. 364: 2177–2187.
Madsen O, et al. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.
Mood AM, Graybill FA, Boes DC. 1974. Introduction to the theory of statistics. New York: McGraw-Hill.
Murphy WJ, et al. 2001a. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.
Murphy WJ, et al. 2001b. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2355.
Nguyen T, Speed TP. 1992. A derivation of all linear invariants for a nonbalanced transversion model. J Mol Evol. 35:60–76.
Philippe H, et al. 2011. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. Nature 470:255–260.
Ragan MA, McInerney JO, Lake JA. 2009. The network of life: genome beginnings and evolution. Phil Trans R Soc B. 364: 2169–2175.
Rao CR. 1973. Linear statistic inference and its applications. New York: Wiley.

Reich D, Green RE, Kircher M. 2011. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.
Rivera MC, Lake JA. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76.
Scally M, et al. 2001. Molecular evidence for the major clades of placental mammals. J Mamm Evol. 8:239–277.
Simonson AB, et al. 2005. Decoding the genomic tree of life. Proc Natl Acad Sci U S A. 102:6608–6613.
Sinsheimer JS. 1994. Extensions to evolutionary parsimony [dissertation]. [Los Angeles (CA)]: UCLA.
Sinsheimer JS, Lake JA, Little RJA. 1996. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. Biometrics 52: 193–210.
Steel MA, Fu YX. 1995. Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model. J Comput Biol. 2:39–47.
Steel MA, Hendy MD, Penny D. 1998. Reconstructing probabilities from nucleotide pattern probabilities: a survey and some new results. Discr App Math. 88:367–396.
Steel MA, Szekely L, Erdos PL, Waddell PJ. 1993. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. New Zeal J Botany. 31:289–296.
Waddell PJ, Okada N, Hasegawa M. 1999. Towards resolving the interordinal relationships of placental mammals. Syst Biol. 48:1–5.
Waddell PJ, Penny D, Moore T. 1997. Extending Hadamard conjugations to model sequence evolution with variable rates across sites. Mol Phylo Evol. 8:33–50.
Yap VB, Speed T. 2005. Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol Biol. 5:2.

**Associate editor:** David Bryant