Confidence in evolutionary trees from biological sequence data

M. A. Steel*, P. J. Lockhart† & D. Penny†

* Department of Mathematics, † Molecular Genetics Unit, Massey University, Palmerston North, New Zealand

THE reliable construction of evolutionary trees from nucleotide sequences often depends on randomization tests such as the bootstrap[1] and PTP (cladistic permutation tail probability) tests[2-6]. The genomes of bacteria[7], viruses[8], animals[7,9,10] and plants[11], however, vary widely in their nucleotide frequencies. Where genomes have independently acquired similar G+C base compositions, signals in the data arise that cause methods of evolutionary tree reconstruction to estimate the wrong tree by grouping together sequences with similar G+C content[12-14]. Under these conditions randomization tests can lead to both the rejection of the correct evolutionary hypothesis and acceptance of an incorrect hypothesis (such as with the contradictory inferences from the photosynthetic rbcS and rbcL sequences[14]). We have proposed one approach to testing for the G+C content problem[15]. Here we present a formalization of this method, a frequency-dependent significance test, which has general application.

Suppose we have a collection of four sequences of r-state characters (in practice, r will usually be 2, 4 or 20). If we edit the sites so that only the parsimony sites (phylogenetically informative sites) are present, then c will denote the length of these (edited) sequences. Let T_1, T_2 and T_3 be the three binary trees on the four taxa, where T_i is the tree in which taxon 1 and taxon $i+1$ are grouped together. Let n_i denote the number of sites that favour tree T_i, that is, which fit on T_i with a single mutation. Consider the following null model, in which four sequences are generated purely randomly (each state is independently chosen) with the proviso that the expected frequencies of the states for each sequence equal the actual observed frequencies. Then, by chance alone, among c informative sites generated in this way, a certain random number, which we denote as N_i, will favour tree T_i. We now calculate the probability $Pr[N_i \geqslant n_i]$ that this random contribution is at least the observed value n_i. We also describe a way of normalizing n_i to take account of this random contribution. Let $p_i(j)$ denote the proportion of sites in sequence $i = 1 \ldots 4$ at which state $j = 1 \ldots r$ appears. Evaluate the three sums

$$S_1 = \sum_{j \neq k} p_1(j)p_2(j)p_3(k)p_4(k);$$

$$S_2 = \sum_{j \neq k} p_1(j)p_2(k)p_3(j)p_4(k);$$

$$S_3 = \sum_{j \neq k} p_1(j)p_2(k)p_3(k)p_4(j),$$

and let

$$s_i = \frac{S_i}{S_1 + S_2 + S_3} \quad \text{for } i = 1, 2, 3.$$

Then under the null model, the number, N_i, of sites that fit tree T_i in a random sample of c informative sites has a binomial distribution $B(c, s_i)$. Thus, under the null model, the probability that N_i is at least as large as its observed value, n_i, is given by the cumulative binomial value

$$Pr[N_i \geqslant n_i] = \sum_{k \geqslant n_i} \binom{c}{k} s_i^k (1 - s_i)^{c-k} \qquad (1)$$

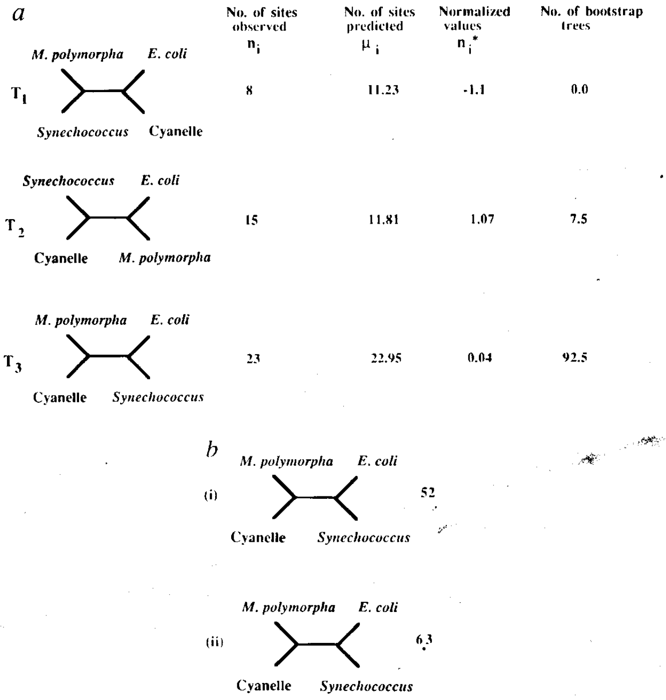


FIG. 1 *a*, The three possible unrooted binary trees for a four-taxon case. Analysis was on the 46 parsimony sites at the first and second codon positions. The cyanelle and chloroplast sequences are used as part of a study on the endosymbiotic origin of organelles and here relationships are difficult to assess[14,15,18] as organelles generally have a high A+T content. The frequencies of the bases are *cyanelle* (0.239a, 0.174c, 0.174g, 0.413t), *E. coli* (0.261a, 0.478c, 0.196g, 0.065t), *Synechococcus* (0.217a, 0.391c, 0.283g, 0.109t) and *Marchantia polymorpha* (0.196a, 0.261c, 0.174g, 0.370t). The number of sites supporting each tree (n_i) is compared with the expected number (μ_i) under the null model. The normalized values n_i^* are given. None of the three values $Pr[N_i \geqslant n_i]$ is significant even at the 0.1 level (the FD test described by equation (2) gives $\chi^2 = 1.79$, d.f. = 2). Applying the PTP test, 8 out of 100 (column) randomized data sets produced trees that under parsimony were shorter or equal to that constructed from the original data, indicating significance for T_3. Bootstrap analysis also gave spurious support to T_3 (92.5/100 recovered). *b*, Majority-rule consensus trees for (row, not column) randomized sequences (i, upper) singleton and parsimony sites at first and second codon positions: 284 sites, and (ii, lower) parsimony sites: 46 sites. Sequences were randomized 100 times and trees reconstructed from these randomized sequences; the values shown are the number of times the given tree was supported. The expected number for the three possible unrooted trees, given a symmetric base composition, would be 33/100. The results are significantly different from 33, showing that the tree-building method grouped the sequences of similar G+C content, even when the sequences were randomized.

Note that N_i has a mean and standard deviation given by $\mu_i = cs_i$, $\sigma_i = \sqrt{(cs_i(1-s_i))}$. Thus, under the null model the normalized value $n_i^* = (n_i - \mu_i)/\sigma_i$ has mean 0 and variance 1. Also, if c is large, n_i^* has, approximately, a standard normal distribution under the null model, which allows the rapid estimation of equation (1) when $c > 50$. Whatever the size of c, it seems sensible to compare trees not on n_i, the number of sites favouring tree T_i, but on the normalized values n_i^*. Also, it is possible to test the null model itself by computing the statistic

$$\chi^2 = \sum_{1 \leqslant i \leqslant 3} \frac{(n_i - \mu_i)^2}{\mu_i} \qquad (2)$$

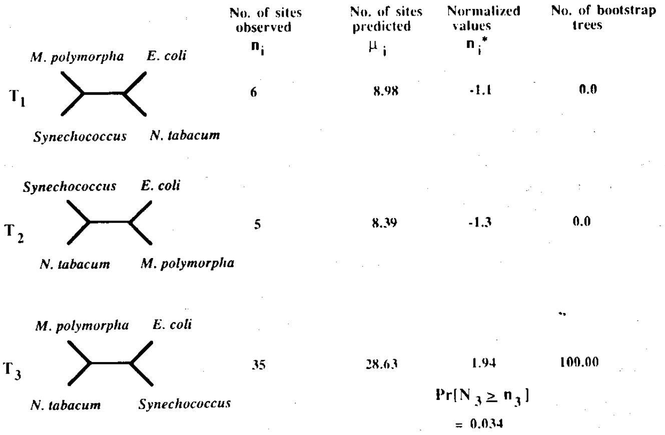
Under the null model, and provided each μ_i is not too small (say $\geqslant 5$) this statistic has a χ^2 distribution with two degrees of freedom. Equations (1) and (2) provide a frequency-dependent significance test (FD test) for the data under the null model.

Note that, for nucleotide sequences, our null model allows a limited (random) degree of variation in nucleotide frequencies between distinct regions or subsets of sites (for example, coding/non-coding regions or triplet positions). But if these regional nucleotide frequencies differ significantly, then a modified null model, in which sequences are generated randomly but subject to the regional frequency constraints, would predict different numbers of sites supporting each tree from those predicted by our null model (which is based on averages of nucleotide frequencies). For sites from a specified sequence, where varying G+C content within a genome is not important for the sequence, and with constant sites edited out, this factor is not important (this applies for our data sets). In cases where this factor is important, the calculations described below should be carried out separately on the regions and combined using standard statistical techniques.

We illustrate our test for two data sets; one in which the 'historical signal' (patterns in the data that indicate the true species relationship and which have not arisen from convergence) is expected[14,15] to have been lost through multiple substitutions, and a second data set in which we expect some phylogenetic information still to be present.

Figure 1*a* shows three binary unrooted trees for *atpA* sequences from four taxa. This gene encodes the α-subunit of ATP synthetase and is highly conserved in the genomes of mito-

FIG. 2 The three unrooted binary trees for a four-taxon case in which a historical signal is expected between the chlorophyll *a*/*b* containing taxa (replacing the cyanelle sequence with a sequence from tobacco (*Nicotiana tabacum*)). Analysis was carried out on parsimony sites at the first and second codon positions considered jointly (also 46 sites). For tree T_3, we have $n_3 = 35$, $\mu_3 = 28.63$, and from equation (1), $Pr[N_3 \geqslant n_3] = 0.034$, so according to the FD test, T_3 is significant at the 0.05 level. The frequencies of the bases were *Nicotiana tabacum* (0.326a, 0.217c, 0.087g, 0.370t), *E. coli* (0.174a, 0.543c, 0.261g, 0.022t), *Synechococcus* (0.174a, 0.478c, 0.217g, 0.130t) and *Marchantia polymorpha* (0.370a, 0.152c, 0.087g, 0.391t). Applying the PTP test, 0 (out of 100) randomized data sets produced trees that under parsimony were shorter or equal to that constructed from the original data, thus supporting the original tree as significant. The bootstrap test also gave strong support to the expected tree: T_3 (100/100 recovered). Note that the FD test could be extended to more than four taxa, either for parsimony or compatibility, although calculation of the s_i values becomes more difficult as the number of taxa grows (M.A.S. and P.J.L., manuscript in preparation).



chondria, chloroplasts, photosynthetic and non-photosynthetic eubacteria. The sequences are from the cyanelle (a photosynthetic organelle, characterized by a pycobilin light-harvesting system and found in the protist *Cyanophora paradoxa*), *Marchantia polymorpha* chloroplast (characterized by a chlorophyll a/b light-harvesting complex), *Synechococcus* PCC 6301 (a cyanobacterium, characterized by a phycobilin light-harvesting system) and *Escherichia coli* (a non-photosynthetic eubacterium). The edited sequences have a G+C content of 35, 43, 67 and 67%, respectively. The number of parsimony sites n_i that support the three possible trees is given. Greatest support is for tree T_3, and bootstrap support for T_3 is also strong (92.5/100 trees). Similarly, analysis of the data using a PTP (cladistic permutation tail probability) test[5] indicates significance at the 0.1 level, in that only 8 trees out of the 100 (column-randomized) data sets were shorter than or equal to the original tree.

In contrast, application of our FD test shows that the patterns observed in the data are close to random. The reason for this disparity is that similar G + C contents may suggest phylogenetic structure even when there is none. To emphasize this, we can randomize each sequence across the variable or parsimony sites and construct trees from these (row) randomized sequences. Usually, the original tree will be recovered (Fig. 1*b*), even though the sequences are random, apart from their G + C content.

This example illustrates an important but poorly appreciated distinction in taxonomy between convergence (the degree to which longer sequences (less sampling error) would result in the same tree) and consistency (the increasing certainty that the tree generated from the sequences is the correct tree—the tree that generated the sequences). In the present example both the PTP and bootstrap results suggest a high level of convergence, but because the sequences are effectively random, no tree building method can be consistent.

Figure 2 gives a second example in which the cyanelle sequence has been replaced with a chloroplast sequence from tobacco (*Nicotiana tabacum*). In this case, greatest parsimony support is also for T_3. The tree is strongly supported by the bootstrap (100/100 trees) and the PTP test. With this data set our FD test indicates non-random patterns (presumably historical patterns) shared between tobacco and liverwort (*Marchantia polymorpha*), thus giving confidence that the tree is not an artefact of G+C composition.

For some data sets, there could be many determinants of pattern that result in contradictory signals. Irregular base composition is one such determinant which might converge sequences[12-14,17] both at silent and replacement sites in coding genes, as well as in ribosomal RNAs[7,8,10,13,14,16]. In view of this, our test should help to evaluate whether near or closest neighbours in phylogenetic trees share 'historically informative patterns', or whether they group together solely because of similar base composition. □

Received 24 February; accepted 14 May 1993.

1. Felsenstein, J. *Evolution* **39,** 783–791 (1985).
2. Archie, J. W. *Syst. Zool.* **38,** 239–252 (1989).
3. Henderson, I. M., Penny, D. & Hendy, M. D. *Nature* **326,** 22 (1987).
4. Henderson, I. M., Hendy, M. D. & Penny, D. *J. theor. Biol.* **140,** 289–303 (1989).
5. Faith, D. & Cranston, P. S. *Cladistics* **7,** 1–28 (1991).
6. Steel, M. A., Hendy, M. D. & Penny, D. *J. Class.* **9,** 71–90 (1992).
7. Jukes, T. H. & Bhushan, V. *J. molec. Evol.* **24,** 39–44 (1986).
8. Keese, P., MacKenzie, A. & Gibbs, A. *Virology* **172,** 536–546 (1989).
9. Bernardi, G. & Bernardi, G. *J. molec. Evol.* **24,** 1–11 (1986).
10. Crozier, R. H. & Crozier, Y. C. *Genetics* **133,** 97–117 (1993).
11. Oliver, J. L., Marin, A. & Martinez-Zapater, J. M. *Nucleic Acids Res.* **18,** 65–73 (1990).
12. Penny, D., Hendy, M. A., Zimmer, E. A. & Hanby, R. K. *Aust. Syst. Bot.* **3,** 21–38 (1990).
13. Lockhart, P. J., Howe, C. J., Bryant, D. A., Beanland, T. J. & Larkum, A. W. D. *J. molec. Evol.* **34,** 153–162 (1992).
14. Lockhart, P. J. *et al. FEBS Lett.* **301,** 127–131 (1992).
15. Lockhart, P. J. & Penny, D. *Research in Photosynthesis* Vol. III, 499–505 (Kluwer, Dordrecht, 1992).
16. Hasegawa, M. & Hashimoto, T. *Nature* **361,** 23 (1993).
17. Saccone, C., Pesole, G. & Preparata, G. *J. molec. Evol.* **29,** 407–411 (1989).
18. Lockhart, P. J., Penny, D., Hendy, M. D. & Larkum, A. W. D. *Photosyn. Res.* (in the press).

ACKNOWLEDGEMENTS. Supported by a grant from the New Zealand Lotteries Commission (M.A.S.) and the New Zealand Health Research Council (P.J.L.).