

Are big trees indeed easy?

In a recent *TREE* article, Purvis and Quicke¹ commented on a simulation study performed by Hillis² and concluded that large phylogenies are easier to reconstruct than small ones and that parsimony, which was claimed to have 'democratized systematics', performed better than model-based methods such as neighbour-joining. These conclusions, and Purvis and Quicke's subsequent recommendations, may be misleading and are in part based on misinterpretation of Hillis's simulation results and their relationship with the analysis performed by Kim³.

Hillis used a model tree with 228 species to simulate data, and used parsimony, neighbor-joining and UPGMA to reconstruct the phylogeny. His measure of accuracy, 'percent of tree correct' (p), is calculated using Robinson and Foulds' measure of tree distance⁴, and is equivalent to the percentage of correctly recovered interior branches in the model tree defining species bipartitions (but see below regarding the treatment of estimated polytomies). This is quite different from the frequency with which the entire tree is correctly recovered from simulated replicates (f).

Hillis, and Purvis and Quicke, elaborated on the great number of possible phylogenies for large numbers of taxa and the consequent great difficulty they expected to encounter in attempting to recover a large phylogeny. This argument is valid only if we measure accuracy with f . Using f instead of p in Hillis's study would have given both parsimony and neighbor-joining accuracies near zero for most of the data points Hillis simulated. However, if p is used there is in fact no theory that leads us to expect large trees to be harder to recover than small trees. Although the accuracy (measured by p) is high for the tree simulated by Hillis, the results are just a one-tree example and may not be applicable to other large trees. From simulations of random trees from a model of cladogenesis (a birth-death process with species sampling), we have found that the accuracy can change substantially with the values of parameters in the model. For example, with 1000 sites, we find that parsimony gives $p = 80\%$ for the tree Hillis simulated, but that this proportion can become very low (<20%) or very high (>80%) when parameters in the model of cladogenesis are adjusted within realistic limits – even when the number of sequences and the overall levels of sequence divergence are kept constant.

Furthermore, the performance (p) deteriorates as the number of species increases with the parameters in the cladogenesis model fixed. This has also been corroborated for neighbor-joining⁵. Although more studies are needed to decide whether or not there are more easy large trees than easy small trees, it is apparent that not all large trees are easy, as Purvis and Quicke inferred. Purvis and Quicke's recommendation of adding species in the data as a general approach to avoiding inconsistency may be misleading. The danger of unguarded extrapolation from simulation results obtained using a single tree has been discussed extensively⁶.

The conflicts that Purvis and Quicke perceived between Hillis's and Kim's results are also partly the result of those authors' use of different measures. Kim examined the consistency of the

parsimony estimate of phylogeny; parsimony is consistent if, and only if, the entire tree is always recovered ($f = 100\%$) as the amount of data increases. Kim has found that parsimony is often inconsistent ($f = 0\%$) for a broad range of parameter values. His analyses make few predictions about p .

The conclusion that parsimony performed better for short sequences than the model-based neighbor-joining method appears to us to be an artefact of Hillis's simulation method. With the parsimony method, we were unable to recreate Hillis's results other than by collapsing inferred branches with no changes. In this case, the use of Robinson and Foulds' measure to calculate p effectively counts any inferred polytomy as being 50% correct; this gives parsimony a head start of approximately 50% when there are few data. With neighbor-joining, polytomies are arbitrarily resolved into bifurcating trees and then the method starts from near 0%. As has been suggested previously⁶, if a polytomous tree is compatible with x bifurcating trees, it should be taken as $100\%/x$ correct. When we used this measure for both parsimony and neighbor-joining, we found no difference between them for the tree Hillis simulated.

It is unfortunate that Purvis and Quicke used their interpretation of Hillis's results to argue for a diminished role for statistical models in phylogenetic analyses. Their comment that 'our models are always over simple and could be plain wrong, with inevitable consequences' gives a misleading portrayal of the aims of statistical modelling. Phylogeny reconstruction is well-known to be quite robust to model assumptions⁷, and when the most important factors are considered in the model the results are quite reliable. In addition, it is widely recognized that accurate models lead to better estimates of branch lengths and speciation times⁸. Despite the lack of an explicitly specified evolutionary model in parsimony analysis, there are reasons to believe that parsimony makes very stringent and unrealistic assumptions about substitution processes^{9,10}. In summary, we agree with the dictum 'All models are wrong; some are more useful than others'¹¹.

We should also like to emphasize our and others' interest in understanding processes of sequence evolution. In this regard, statistical models are of critical importance and, when used in the maximum-likelihood framework, provide a powerful and flexible framework for estimating evolutionary parameters and testing hypotheses^{8,12}. In contrast, the lack of a clearly specified model in parsimony analysis makes it difficult to incorporate our knowledge of sequence evolution and renders the method ill-adapted to studying the evolutionary process.

Ziheng Yang

Dept of Integrative Biology,
University of California, Berkeley,
CA 94720-3140, USA
(ziheng@mws4.biol.berkeley.edu)

Nick Goldman

Dept of Genetics,
University of Cambridge,
Downing Street, Cambridge,
UK CB2 3EH
(N.Goldman@gen.cam.ac.uk)

References

- 1 Purvis, A. and Quicke, D.L.J. (1997) *Trends Ecol. Evol.* 12, 49–50
- 2 Hillis, D.M. (1996) *Nature* 383, 130–131
- 3 Kim, J. (1996) *Syst. Biol.* 45, 363–374
- 4 Robinson, D.F. and Foulds, L.R. (1981) *Math. Biosci.* 53, 131–147
- 5 Strimmer, K. and von Haeseler, A. (1996) *Syst. Biol.* 45, 516–523
- 6 Hillis, D.M. (1995) *Syst. Biol.* 44, 3–16
- 7 Gaut, B.S. and Lewis, P.O. (1995) *Mol. Biol. Evol.* 12, 152–162
- 8 Yang, Z., Goldman, N. and Friday, A.E. (1994) *Mol. Biol. Evol.* 11, 316–324
- 9 Goldman, N. (1990) *Syst. Zool.* 39, 345–361
- 10 Yang, Z. (1996) *J. Mol. Evol.* 42, 294–307
- 11 Hunter, W.G. (1982) in *Teaching of Statistics and Statistical Consulting* (Rustagi, J.S. and Wolfe, D.A., eds), pp. 35–69, Academic Press
- 12 Huelsenbeck, J.P. and Rannala, B. (1997) *Science* 276, 227–232

Reply from A. Purvis and D.L.J. Quicke

Ideally, phylogenies would be estimated using a model that accurately captures evolution's complexities. Inaccurate models can fail to reconstruct even very small trees accurately from infinite data¹. We were therefore surprised² at the good performance of a simple method unweighted parsimony in recent simulations of a much larger tree of 228 species³: despite making no allowance for many features of the process used to simulate the sequence data (such as transition bias and rate heterogeneity among sites), parsimony successfully reconstructed the tree from around 5000 simulated bases. This result suggests that large phylogenies may perhaps be easier than previously thought to estimate accurately, even with simple models. Yang and Goldman appear at first sight to disagree with us on many issues. On reading their comments, however, the disagreements seem to be more apparent than real.

Yang and Goldman take issue with two points that they represent as our conclusions, namely, that large trees are easier to construct than small ones, and that parsimony outperformed neighbour-joining and UPGMA. We did not conclude, or even say, that large trees are easier to reconstruct than small ones. Rather, we posed the question, and answered it with, 'What matters isn't the size of the tree *per se*, but the lengths of its branches'. We did say that parsimony outperformed both UPGMA and neighbour-joining, though whether it performed better³ or precisely as well as neighbour-joining has no bearing on what our conclusion actually was. Yang and Goldman's accusations that our conclusions are misleading, and that we misinterpret and misrepresent the results of Hillis³ and Kim⁴, are therefore ill-founded. Their other criticisms have similarly little basis.

We are criticized for extrapolating from simulation results obtained using a single tree. However, we in fact advocated 'a systematic study ... to see how parsimony and other approaches fare with large trees under different combinations of rates, models and other variables like tree topology'.

Are big trees indeed easy?

In a recent *TREE* article, Purvis and Quicke¹ commented on a simulation study performed by Hillis² and concluded that large phylogenies are easier to reconstruct than small ones and that parsimony, which was claimed to have 'democratized systematics', performed better than model-based methods such as neighbour-joining. These conclusions, and Purvis and Quicke's subsequent recommendations, may be misleading and are in part based on misinterpretation of Hillis's simulation results and their relationship with the analysis performed by Kim³.

Hillis used a model tree with 228 species to simulate data, and used parsimony, neighbor-joining and UPGMA to reconstruct the phylogeny. His measure of accuracy, 'percent of tree correct' (p), is calculated using Robinson and Foulds' measure of tree distance⁴, and is equivalent to the percentage of correctly recovered interior branches in the model tree defining species bipartitions (but see below regarding the treatment of estimated polytomies). This is quite different from the frequency with which the entire tree is correctly recovered from simulated replicates (f).

Hillis, and Purvis and Quicke, elaborated on the great number of possible phylogenies for large numbers of taxa and the consequent great difficulty they expected to encounter in attempting to recover a large phylogeny. This argument is valid only if we measure accuracy with f . Using f instead of p in Hillis's study would have given both parsimony and neighbor-joining accuracies near zero for most of the data points Hillis simulated. However, if p is used there is in fact no theory that leads us to expect large trees to be harder to recover than small trees. Although the accuracy (measured by p) is high for the tree simulated by Hillis, the results are just a one-tree example and may not be applicable to other large trees. From simulations of random trees from a model of cladogenesis (a birth-death process with species sampling), we have found that the accuracy can change substantially with the values of parameters in the model. For example, with 1000 sites, we find that parsimony gives $p = 80\%$ for the tree Hillis simulated, but that this proportion can become very low (<20%) or very high (>80%) when parameters in the model of cladogenesis are adjusted within realistic limits – even when the number of sequences and the overall levels of sequence divergence are kept constant.

Furthermore, the performance (p) deteriorates as the number of species increases with the parameters in the cladogenesis model fixed. This has also been corroborated for neighbor-joining⁵. Although more studies are needed to decide whether or not there are more easy large trees than easy small trees, it is apparent that not all large trees are easy, as Purvis and Quicke inferred. Purvis and Quicke's recommendation of adding species in the data as a general approach to avoiding inconsistency may be misleading. The danger of unguarded extrapolation from simulation results obtained using a single tree has been discussed extensively⁶.

The conflicts that Purvis and Quicke perceived between Hillis's and Kim's results are also partly the result of those authors' use of different measures. Kim examined the consistency of the

parsimony estimate of phylogeny; parsimony is consistent if, and only if, the entire tree is always recovered ($f = 100\%$) as the amount of data increases. Kim has found that parsimony is often inconsistent ($f = 0\%$) for a broad range of parameter values. His analyses make few predictions about p .

The conclusion that parsimony performed better for short sequences than the model-based neighbor-joining method appears to us to be an artefact of Hillis's simulation method. With the parsimony method, we were unable to recreate Hillis's results other than by collapsing inferred branches with no changes. In this case, the use of Robinson and Foulds' measure to calculate p effectively counts any inferred polytomy as being 50% correct; this gives parsimony a head start of approximately 50% when there are few data. With neighbor-joining, polytomies are arbitrarily resolved into bifurcating trees and then the method starts from near 0%. As has been suggested previously⁶, if a polytomous tree is compatible with x bifurcating trees, it should be taken as $100\%/x$ correct. When we used this measure for both parsimony and neighbor-joining, we found no difference between them for the tree Hillis simulated.

It is unfortunate that Purvis and Quicke used their interpretation of Hillis's results to argue for a diminished role for statistical models in phylogenetic analyses. Their comment that 'our models are always over simple and could be plain wrong, with inevitable consequences' gives a misleading portrayal of the aims of statistical modelling. Phylogeny reconstruction is well-known to be quite robust to model assumptions⁷, and when the most important factors are considered in the model the results are quite reliable. In addition, it is widely recognized that accurate models lead to better estimates of branch lengths and speciation times⁸. Despite the lack of an explicitly specified evolutionary model in parsimony analysis, there are reasons to believe that parsimony makes very stringent and unrealistic assumptions about substitution processes^{9,10}. In summary, we agree with the dictum 'All models are wrong; some are more useful than others'¹¹.

We should also like to emphasize our and others' interest in understanding processes of sequence evolution. In this regard, statistical models are of critical importance and, when used in the maximum-likelihood framework, provide a powerful and flexible framework for estimating evolutionary parameters and testing hypotheses^{8,12}. In contrast, the lack of a clearly specified model in parsimony analysis makes it difficult to incorporate our knowledge of sequence evolution and renders the method ill-adapted to studying the evolutionary process.

Ziheng Yang

Dept of Integrative Biology,
University of California, Berkeley,
CA 94720-3140, USA
(ziheng@mws4.biol.berkeley.edu)

Nick Goldman

Dept of Genetics,
University of Cambridge,
Downing Street, Cambridge,
UK CB2 3EH
(N.Goldman@gen.cam.ac.uk)

References

- 1 Purvis, A. and Quicke, D.L.J. (1997) *Trends Ecol. Evol.* 12, 49–50
- 2 Hillis, D.M. (1996) *Nature* 383, 130–131
- 3 Kim, J. (1996) *Syst. Biol.* 45, 363–374
- 4 Robinson, D.F. and Foulds, L.R. (1981) *Math. Biosci.* 53, 131–147
- 5 Strimmer, K. and von Haeseler, A. (1996) *Syst. Biol.* 45, 516–523
- 6 Hillis, D.M. (1995) *Syst. Biol.* 44, 3–16
- 7 Gaut, B.S. and Lewis, P.O. (1995) *Mol. Biol. Evol.* 12, 152–162
- 8 Yang, Z., Goldman, N. and Friday, A.E. (1994) *Mol. Biol. Evol.* 11, 316–324
- 9 Goldman, N. (1990) *Syst. Zool.* 39, 345–361
- 10 Yang, Z. (1996) *J. Mol. Evol.* 42, 294–307
- 11 Hunter, W.G. (1982) in *Teaching of Statistics and Statistical Consulting* (Rustagi, J.S. and Wolfe, D.A., eds), pp. 35–69, Academic Press
- 12 Huelsenbeck, J.P. and Rannala, B. (1997) *Science* 276, 227–232

Reply from A. Purvis and D.L.J. Quicke

Ideally, phylogenies would be estimated using a model that accurately captures evolution's complexities. Inaccurate models can fail to reconstruct even very small trees accurately from infinite data¹. We were therefore surprised² at the good performance of a simple method unweighted parsimony in recent simulations of a much larger tree of 228 species³: despite making no allowance for many features of the process used to simulate the sequence data (such as transition bias and rate heterogeneity among sites), parsimony successfully reconstructed the tree from around 5000 simulated bases. This result suggests that large phylogenies may perhaps be easier than previously thought to estimate accurately, even with simple models. Yang and Goldman appear at first sight to disagree with us on many issues. On reading their comments, however, the disagreements seem to be more apparent than real.

Yang and Goldman take issue with two points that they represent as our conclusions, namely, that large trees are easier to construct than small ones, and that parsimony outperformed neighbour-joining and UPGMA. We did not conclude, or even say, that large trees are easier to reconstruct than small ones. Rather, we posed the question, and answered it with, 'What matters isn't the size of the tree *per se*, but the lengths of its branches'. We did say that parsimony outperformed both UPGMA and neighbour-joining, though whether it performed better³ or precisely as well as neighbour-joining has no bearing on what our conclusion actually was. Yang and Goldman's accusations that our conclusions are misleading, and that we misinterpret and misrepresent the results of Hillis³ and Kim⁴, are therefore ill-founded. Their other criticisms have similarly little basis.

We are criticized for extrapolating from simulation results obtained using a single tree. However, we in fact advocated 'a systematic study ... to see how parsimony and other approaches fare with large trees under different combinations of rates, models and other variables like tree topology'.