Points of View

Syst. Biol. 50(6):970–978, 2001

Complexity of the Likelihood Surface for a Large DNA Dataset

LAURA A. SALTER

Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131, USA;
E-mail: salter@stat.unm.edu

The maximum likelihood (ML) criteria for phylogenetic tree estimation is becoming increasingly popular among those wishing to reconstruct the evolutionary history of a collection of DNA sequences. This is happening in part because of the availability of likelihood methods in widely used programs such as PAUP* (Swofford, 1998) and PHYLIP (Felsenstein, 1993), and the recognition that likelihood methods offer several advantages over other reconstruction criteria. These advantages include interpretability of the underlying models, consistency in the statistical sense (Felsenstein, 1981; Hasegawa et al., 1991; Yang, 1994; Chang, 1996; Rogers, 1997), and the possibility of statistical testing of hypotheses by using the likelihood framework (Kishino and Hasegawa, 1989; Goldman, 1993).

The application of likelihood methods to large datasets has been limited, however, by the prohibitive amount of time required by the available algorithms for ML estimation. Consequently, detailed investigation of the likelihood surface has not been undertaken for large datasets (Sanderson and Kim, 2000), and therefore the impact of the complexity of the likelihood surface on estimation procedures has not been studied extensively.

Studies of the complexity of the space of phylogenetic trees have been undertaken for other estimation criteria. For example, Maddison (1991) discussed the existence of islands of most parsimonious trees and the importance of recognizing the existence of such groups of optimal trees. Page (1993) looked at islands of most-parsimonious trees in the context of the various rearrangement procedures that give rise to these islands and discussed the implications for tree-searching algorithms. Both authors agreed that searching for all of the islands of most-parsimonious trees is an important step in any analysis that uses parsimony as the optimality criteria.

Here I examine the likelihood surface for a moderately large, real dataset consisting of 30 papillomavirus sequences: 28 human papillomavirus (HPVs), a rhesus papillomavirus, and a pygmy chimpanzee papillomavirus. The data source is a 1,382-bp segment of the L1 gene from which all sites containing insertions or deletions in one or more of the sequences have been removed. The aligned data were downloaded from the Los Alamos National Database website (http://hpv-web.lanl.gov). For further details about the data, see Chan et al. (1992, 1995), Ong et al. (1997), and Salter (1999).

METHODS

To evaluate the likelihood surface in relation to the impact on possible estimation procedures, I followed the approaches of Maddison (1991) and Page (1993) and demonstrated that several islands of trees exist under both the nearest neighbor interchange (NNI) and tree bisection and reconnection (TBR) strategies of moves for phylogenetic trees. I extended the definition of an island given by Maddison (1991) to consider the ML criteria; that is, an island of trees is a set of n trees that satisfy the following conditions:

1. All of the trees contained in the island have a log likelihood greater than L.
2. Each of the trees in the island is connected to each other tree through a series of trees,

each tree differing from the others by a single rearrangement and each of which has a log likelihood greater than *L*.

3. All of the trees that satisfy the first two criteria are included in the island.

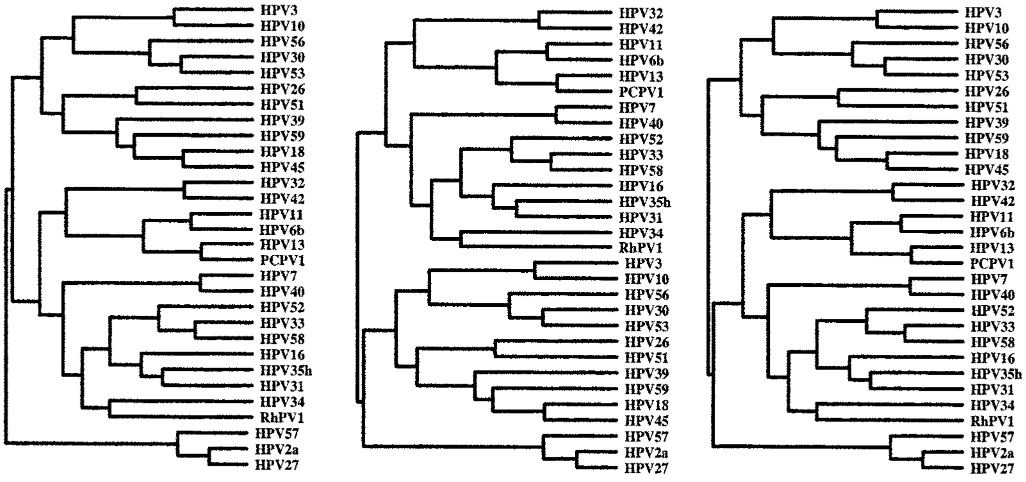
Following the notation of Maddison (1991), I labeled an island of trees with log likelihood greater than *L* as *island-L*. The choice of the cutoff value for *L* is arbitrary in the sense that selecting a smaller value for *L* would also identify an island of trees, and this island would contain all those trees contained in *island-L* as well as any additional trees with a log likelihood greater than the chosen cutoff. Thus, with various selections for *L*, an island may contain only a single tree (that which has the maximum log likelihood in a local area of tree space) or a very large number of trees. In the extreme case, selecting $L = -\infty$ would identify an island that contains all possible trees for *N* taxa. In what follows, the primary criterion in identifying islands was to select a cutoff value *L* that would result in an *island-L* containing a small set of trees with very similar log likelihoods, these trees being separated from other islands of trees by at least two tree rearrangements.

I considered several conditions for evaluation of likelihoods. First, one can consider trees that satisfy the molecular clock assumption and either fix the transition/transversion ratio at 2.0 (the default in PAUP* and PHYLIP) or estimate the transition/transversion ratio simultaneously with the tree. For trees that satisfy the molecular clock, I identified two islands under the NNI strategy of moves. Next, considering trees without the assumption of a molecular clock, one can again either fix the transition/transversion ratio at 2.0 or estimate it simultaneously. For this case, I identified three islands under the NNI strategy of moves and two islands under TBR moves. For all of the cases studied, the F84 model (Felsenstein, 1984) was used, and PAUP* was used for computation of likelihoods. Each of these cases is discussed separately below.

RESULTS

Case 1. Trees Estimated Under the Molecular Clock Assumption

Transition/transversion ratio fixed.—In the case in which a molecular clock is assumed and the transition/transversion ratio is assumed to be fixed at 2.0, there are at least two islands of trees in the space defined by NNI moves between trees. The first island contains three trees, shown in Figure 1. Each



(a) ln likelihood = -28,086.23, (R) = -27,871.83

(b) ln likelihood = -28,086.26, (R) = -27,871.83

(c) ln likelihood = -28,086.26, (R) = -27,871.83

FIGURE 1. The three trees in Island 1 when a molecular clock is assumed. The maximized value of the log likelihood is shown below each of the trees for the transition/transversion ratio set to 2.0. The log likelihoods for all of the trees in the case where the transition/transversion ratio is set to 1.1190 (R) are shown below the fixed-ratio log likelihoods.

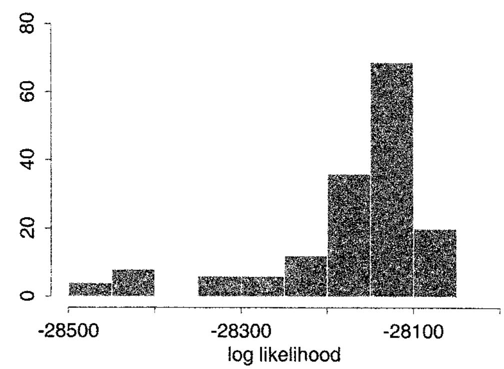
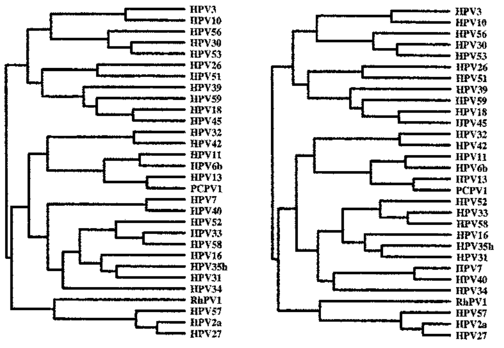


FIGURE 2. Histogram of the log likelihoods of all trees that are a single NNI away from one of the trees in Island 1 when the transition/transversion ratio is fixed at 2.0. The summary statistics are: minimum = −28,495.89, mean = −28,170.77, median = −28,133.51, and maximum = −28,086.56.



(a) ln likelihood = -28,083.67,
(R) = -27,864.73

(b) ln likelihood = -28,084.42,
(R) = -27,864.11

FIGURE 3. The two trees in Island 2 when a molecular clock is assumed. The maximized value of the log likelihood when the transition/transversion ratio is set to 2.0 is shown below each of the trees. Below this is shown the log likelihoods for the trees in the case where the transition/transversion ratio is set to 1.1190 (R).

of these trees has a log likelihood of at least −28,086.50. Interestingly, these trees differ only in the placement of the clade containing HPV2a, HPV27, and HPV57. Examination of the branch lengths that yield the maximized values of the log likelihood shows an extremely short branch length inferred in connecting this clade to the others in all three trees.

Figure 2 shows a histogram of the maximized log likelihoods for all trees that are a single NNI away from one of the three trees in Island 1. The maximum value of any of these nearby trees is −28,086.56. Thus, the three trees shown in Figure 1 do in fact form an island according to the definition given above. In particular, the tree in Figure 1a is a local maximum.

The second island contains two trees, shown in Figure 3. Each of these trees has a log likelihood of at least −28,084.44. The two trees differ in the placement of the HPV34 sequence and differ from the trees in Island 1 in that the RhPV1 (rhesus papillomavirus) sequence is part of the clade containing HPV2a, HPV27, and HPV57. Figure 4 shows a histogram of the maximized log likelihoods for all trees that are a single NNI away from one of the two trees in this island. The maximum value for any of the nearby trees is −28,084.46. Thus the trees in Figure 3 also form an island in the space of phylogenetic trees for this problem. In particular, the tree in Figure 3A is a local maximum.

The globally optimal solution for this problem is the tree in Figure 3A. The existence of the tree in Island 1 (Fig. 1A), which is only locally optimal, is important in that it demonstrated what difficulties might be encountered during a search for the optimal tree. The existence of local optima is particularly important in studying the performance of heuristic uphill searches, such as those used in PAUP* and PHYLIP, because such searches will terminate once a local optimum is reached. Other methods based on stochastic search procedures (e.g., the genetic algorithm of Lewis [1998] and the stochastic search method of Salter and Pearl [2001]) provide mechanisms to deal with the presence of local optima, but the existence of

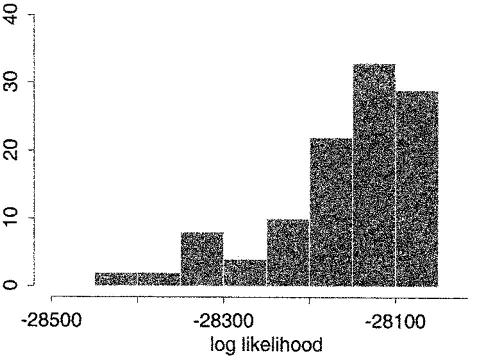


FIGURE 4. Histogram of the log likelihoods of all trees that are a single NNI away from one of the trees in Island 2 when the transition/transversion ratio is fixed at 2.0. The summary statistics are: minimum = −28,438.32, mean = −28,161.55, median = −28,129.93, and maximum = −28,084.46.

multiple islands of trees will complicate the search procedure regardless of the tree-searching method used.

It is of interest to consider the distance between these two islands of trees. Because these islands exist only under NNI swaps, the distance measure of interest is the number of NNI swaps required to move from one of the trees in Island 1 to one of the trees in Island 2. In general, an efficient algorithm for computing the number of NNI rearrangements required to transform a given tree topology into any other tree topology has not been found (Brown and Day, 1984; Page, 1993), but Brown and Day (1984) provide several approximations. Their approximations are implemented in the program COMPONENT (Page, 2001) for unrooted trees and are used here to estimate the number of NNI swaps required to move between the two islands. Table 1 lists the estimated number of NNI swaps required to move between each pair of trees in the two islands when the trees are considered unrooted, using the d_{us} approximation (Brown and Day, 1984). This approximation gives an upper bound on the number of NNI swaps required and indicates that, at most, four swaps are required to move between the two islands. Furthermore, at least two NNI moves are required (otherwise, these two sets of trees would not form distinct islands).

Transition/transversion ratio estimated simultaneously.—As when the transition/transversion ratio was assumed to be fixed, there are also at least two islands of trees when the transition/transversion ratio is simultaneously estimated. The first island contains the same three trees as in the previous case (Fig. 1). Each of these trees has a log likelihood of at least −27,872.00 when the transition/transversion ratio is an estimated 1.1190. Figure 5 shows a histogram of the maximized log likelihoods for all trees that are a single NNI away from one of the trees in this island. The maximum value of the log likelihood of any of these nearby trees is −27,873.10, about 1.3 units of log likelihood less than that of any of the trees in this island. Thus, the trees in Figure 1 represent an island in the space of trees for this problem.

TABLE 1. Estimate of the number of NNI swaps required to move between all pairs of trees in the two islands of trees under the molecular clock assumption. The d_{us} approximation of Brown and Day (1984) as implemented in COMPONENT (Page, 2001) was used to obtain the estimates.

	1a	1b	1c	2a	2b
1a	—	1	1	4	5
1b		—	1	4	5
1c			—	4	5
2a				—	1
2b					—

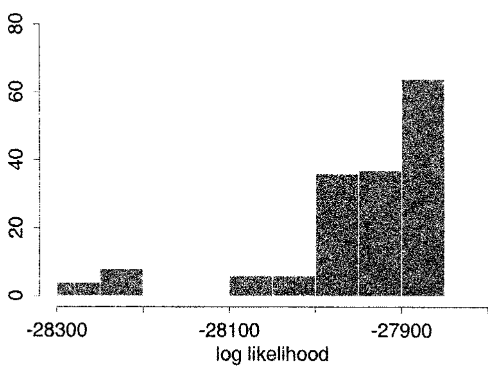


FIGURE 5. Histogram of the log likelihoods of all trees that are a single NNI away from one of the trees in Island 1 when the transition/transversion ratio is fixed at 1.1190. The summary statistics are: minimum = −28,270.82, mean = −27,952.87, median = −27,917.60, and maximum = −27,873.10.

The second island in this case also contains the same two trees as in the case of a fixed transition/transversion ratio (Fig. 3). Each of these trees has a log likelihood that is at least −27,864.80 when the transition/transversion ratio is estimated to be 1.1190. Figure 6 shows a histogram of the maximized log likelihoods

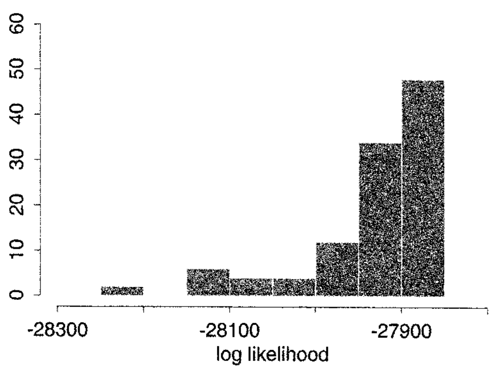


FIGURE 6. Histogram of the log likelihoods of all trees that are a single NNI away from one of the trees in Island 2 when the transition/transversion ratio is fixed at 1.1190. The summary statistics are: minimum = −28,209.11, mean = −27,938.00, median = −27,908.75, and maximum = −27,864.95.

for all trees that are one NNI away from one of the trees in this island. Because the maximum value of the log likelihood for any of these nearby trees is −27,864.95, these trees represent an island. The tree in Figure 3B is both a local maximum and the globally optimal tree in this case. Note that the estimate of the ML tree is different when the transition/transversion ratio is estimated, although the two estimates have very similar log likelihoods and are contained in the same island.

Case 2. Trees Estimated Without the Molecular Clock Assumption

Transition/transversion ratio fixed.—In the case in which a molecular clock is not assumed, but the transition/transversion ratio is assumed to be fixed at 2.0, at least three islands of trees are in the space defined by NNI moves between trees. When TBR moves are considered, two of the three islands are no longer distinct (i.e., a single TBR rearrangement can be used to move from one island to another), but the third remains distinct from the other two (i.e., no single TBR move exists that connects this island to any tree in the other two). The first island contains a single tree, shown in Figure 7. The cutoff value used in defining this island was a log likelihood greater than −28,066.00. Figure 8 shows a histogram of all trees that are a single NNI away from the tree in Island 1. The maximum value of the log likelihood for any of these trees is −28,066.72. Hence, the tree in Figure 7 represents a local maximum.

The second island also contains a single tree, shown in Figure 9. The cutoff value for the log likelihood used for defining this island was −28,052.00, and the tree in Figure 9A has a log likelihood of −28,051.70. This tree differs from the tree in Island 1 in the placement of the clade containing the HPV2a, HPV27, and HPV57 sequences. Figure 10 shows a histogram of the maximized log likelihoods for all trees that are a single NNI away from the tree in Figure 9. The maximum value of any of these trees is −28,054.57, indicating that this tree is a local maximum in the space of trees induced by NNI moves.

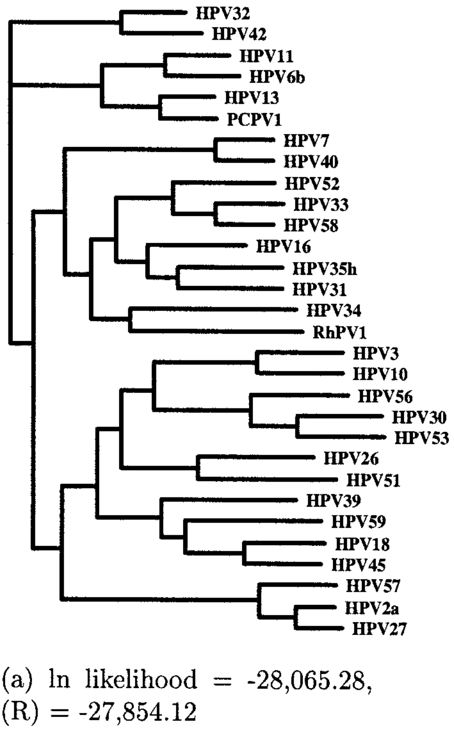


FIGURE 7. The single tree in Island 1 with no molecular clock assumption. The maximized value of the log likelihood is shown below the tree when the transition/transversion ratio is set at 2.0. The log likelihood when the transition/transversion ratio is set to 1.1190 (R) is shown below the fixed-ratio log likelihood.

These first two NNI islands do not form an island in the space of trees induced by TBR moves, because a single TBR move can be used to convert the tree in Island 1 to the tree in Island 2. A third island was identified that is an island in the space of trees

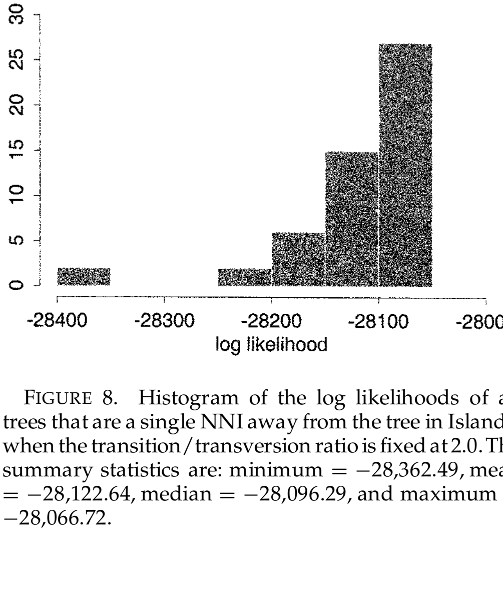


FIGURE 8. Histogram of the log likelihoods of all trees that are a single NNI away from the tree in Island 1 when the transition/transversion ratio is fixed at 2.0. The summary statistics are: minimum = −28,362.49, mean = −28,122.64, median = −28,096.29, and maximum = −28,066.72.

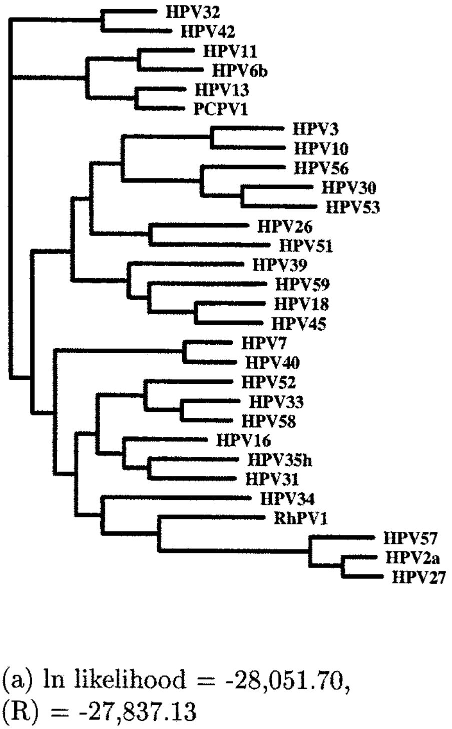


FIGURE 9. The single tree in Island 2 with no molecular clock assumption. The maximized value of the log likelihood is shown below the tree. The log likelihood in the case where the transition/transversion ratio is set to 1.1190 (R) is shown below the fixed-ratio log likelihood.

induced by TBR moves. This island contains three trees, shown in Figure 11, and each tree has a log likelihood of at least −28,053.00. Trees (a) and (b) differ from one another in the placement of the clade containing HPV2a, HPV27, HPV57, and RhPV1. Tree (a) differs from tree (c) in the placement of the clade containing HPV26 and HPV51. The three trees in this island differ from the tree in Island 1 in the placement of the RhPV1 sequence and differ from the tree in Island 2 in the placement of the clade containing HPV2a, HPV27, HPV57, and RhPV1. A histogram of the maximized log likelihoods for all trees that are a single NNI away from one the three trees in this island is shown in Figure 12. The maximum value for any of the nearby trees is −28,053.08, confirming that these trees form an island in the space of NNI moves. That they also form an island in the space of trees induced by TBR moves was verified by running a TBR branch-swapping search in PAUP*. In addition, the tree in Figure 11C is a local maximum.

The globally optimal tree in this problem is the tree in Figure 11C, with the tree in Figure 9 a close second (log likelihoods are −28,051.25 and −28,051.70, respectively). The log likelihoods for the trees in Figures 11a and 11b are also very high (−28,052.10 and −28,051.94, respectively), substantially more than the log likelihood for the tree in Figure 7 (−28,065.28). Hence the likelihood surface is fairly complex, even for a dataset of this size, with at least two islands in the space defined by TBR moves, and at least three islands when NNI moves are used.

In this case also, it is interesting to estimate the number of NNI swaps required to move between the three islands of trees. Table 2 shows the estimated number of required NNI swaps, using the d_{us} approximation of Brown and Day (1984) as implemented in COMPONENT (Page, 2001). An upper bound on the number of NNI swaps required to move between Islands 1 and 2 is four, between Islands 1 and 3 is five, and between Islands

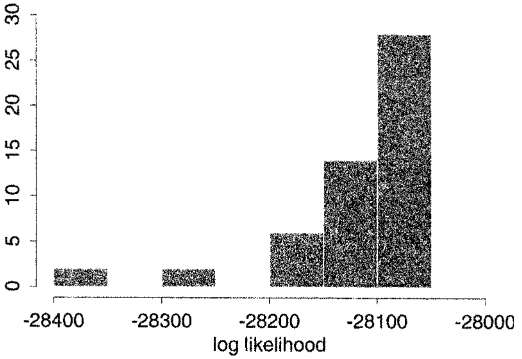
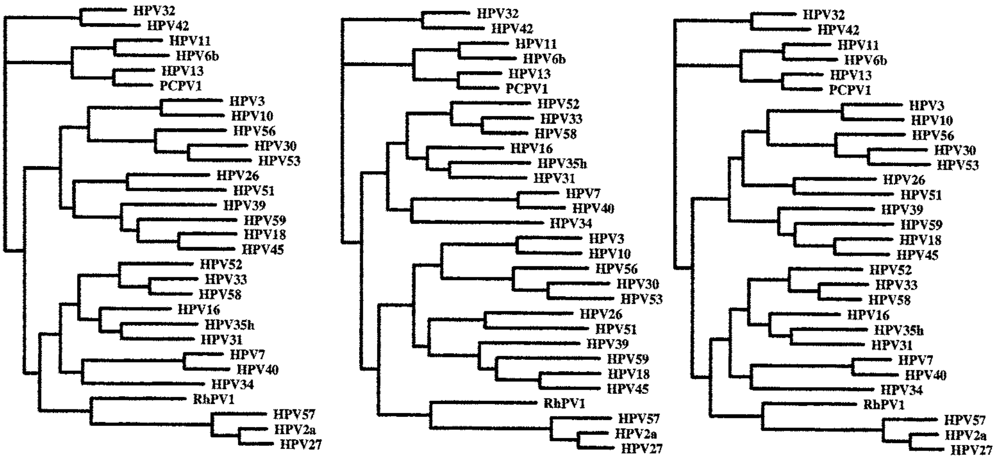


FIGURE 10. Histogram of the log likelihoods of all trees that are a single NNI away from the tree in Island 2 when the transition/transversion ratio is fixed at 2.0. The summary statistics are: minimum = −28,358.48, mean = −28,113.02, median = −28,090.86, and maximum = −28,054.57.

TABLE 2. Estimate of the number of NNI swaps required to move between all pairs of trees in the three islands of trees without the molecular clock assumption. The d_{us} approximation of Brown and Day (1984) as implemented in COMPONENT (Page, 2001) was used to obtain the estimates.

	1a	2a	3a	3b	3c
1a	—	4	6	6	5
2a		—	4	5	3
3a			—	1	1
3b				—	2
3c					—



(a) ln likelihood = -28,052.10
(R) = -27,832.00

(b) ln likelihood = -28,051.94,
(R) = -27,832.63

(c) ln likelihood = -28,051.25,
(R) = -27,833.28

FIGURE 11. The three trees in Island 3 with no molecular clock assumption. The maximized value of the log likelihood is shown below the trees. The log likelihoods in the case where the transition/transversion ratio is set to 1.1190 (R) are shown below the fixed-ratio log likelihoods.

2 and 3 is three. The minimum number of required NNI swaps is two in each of these cases.

Transition/transversion ratio estimated simultaneously.—In the case in which a molecular clock is not assumed and the transition/transversion ratio is estimated simultaneously, three islands of trees were defined as in the previous case. Island 1 contains a single tree (Fig. 7) specified by using a cutoff of −27,854.50. Figure 13 shows a histogram of the maximized values of the log likelihood for all trees that are a single NNI away from this tree. The maximum value was −27,854.92, confirming that this tree is a local maximum for this problem. The single tree in Island 2, obtained with use of a cutoff of −28,837.50, and its corresponding maximized log likelihood in this case is shown in Figure 9. A histogram of the maximized log likelihood for all trees that are a single NNI away from this tree is shown in Figure 14. The maximum log likelihood for any of these trees was −27,837.96. Finally, the

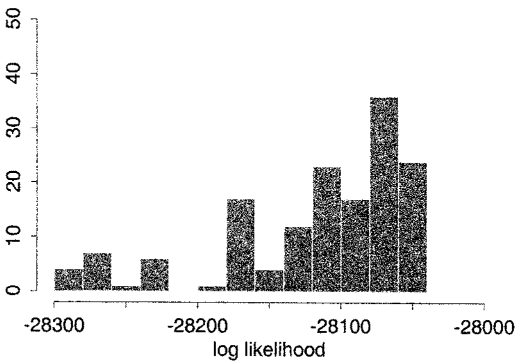


FIGURE 12. Histogram of the log likelihoods of all trees that are a single NNI away from one of the trees in Island 3 when the transition/transversion ratio is fixed at 2.0. The summary statistics are: minimum = −28,283.06, mean = −28,117.07, median = −28,096.03, and maximum = −28,053.08.

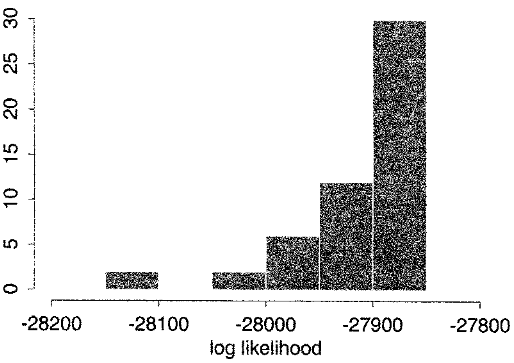


FIGURE 13. Histogram of the log likelihoods of all trees that are a single NNI away from the tree in Island 1 when the transition/transversion ratio is fixed at 1.1190. The summary statistics are: minimum = −28,142.56, mean = −27,908.41, median = −27,883.54, and maximum = −27,854.92.

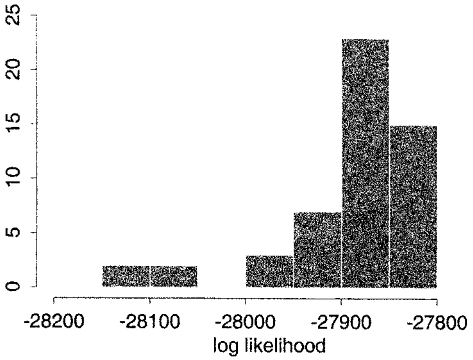


FIGURE 14. Histogram of the log likelihoods of all trees that are a single NNI away from the tree in Island 2 when the transition/transversion ratio is fixed at 1.1190. The summary statistics are: minimum = −28,134.33, mean = −27,895.65, median = −27,876.28, and maximum = −27,837.96.

three trees contained in Island 3 using a cutoff of −27,833.30 and their maximized log likelihoods in this case are shown in Figure 11. The corresponding histogram of maximized log likelihoods for all trees that are a single NNI away from one of these trees is given in Figure 15. The maximum value of the log likelihood for any of these trees was −27,833.38 and thus the tree in Figure 11A is locally optimal.

The globally optimal tree for this problem is the tree in Figure 11A. In this case, the tree selected to be globally optimal is different from the tree chosen (that shown in Fig. 11C) when the transition/transversion ratio is fixed, even though both are in the same island of trees. They differ only in the placement of the HPV26–HPV51 clade. As was the case for trees estimated under the assumption of a molecular clock, the surface of the likelihood is also complex when a molecular clock is not assumed, there being several local optima.

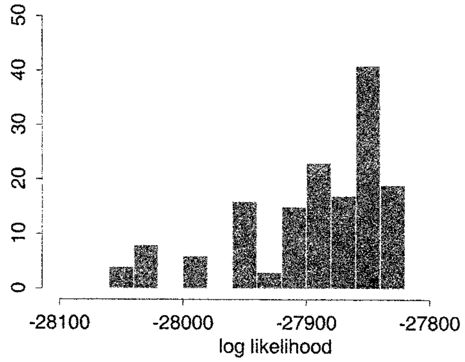


FIGURE 15. Histogram of the log likelihoods of all trees that are a single NNI away from one of the trees in Island 3 when the transition/transversion ratio is fixed at 1.1190. The summary statistics are: minimum = −28,053.99, mean = −27,894.29, median = −27,877.20, and maximum = −27,833.38.

CONCLUSION

The dataset discussed here demonstrates the possibility that numerous local optima may exist when the maximum likelihood criterion is used to infer phylogenetic trees. Local optima were found in several different situations of tree estimation: either with or without the assumption of a molecular clock, with the transition/transversion ratio either fixed or estimated, and with use of either NNI or TBR branch-swapping moves. Several local optima were identified in each of these cases, but those were not necessarily all of the local optima that may exist for this problem. Thus, the likelihood surface for this problem may be even more complex than indicated here.

Identification of local optima is important for understanding the limitations of the algorithms commonly used to estimate phylogenetic trees under optimality criteria such as ML. Algorithms based on uphill searches such as PAUP* and PHYLIP will find a particular local optimum on any single run. Hence, the identification of the globally optimal phylogenetic tree is generally obtained only after such searches are run numerous times, with random starting points for each run. Such a strategy does not guarantee that the globally optimal tree will be found, but it does help identify islands of trees that often contain trees of high likelihood, as in some of the cases identified here. The feasibility of this approach is questionable for very large datasets, for which the likelihood surface is expected to be even more complex. Results of phylogenetic analyses for large datasets must therefore be interpreted with caution and with reference to the limitations of the methods used to produce them.

This study also brings up other interesting questions relating to the complexity of the likelihood surface. For example, how is the depth of the valleys separating local maxima affected by increasing the number of sites in the sequences under consideration? How will the situation be affected by combining data from two or more regions of the genome? To what extent do the particular assumptions made in the analysis (e.g., independence of sites, evolutionary model, same substitution rate across sites) affect the number and type of local maxima? Such questions provide a rich source for investigation into the efficiencies of phylogenetic reconstruction methods.

Acknowledgments

I thank Dennis Pearl, Richard Olmstead, and two anonymous reviewers for helpful comments on an earlier draft of this manuscript.

References

Brown, E., and W. Day. 1984. A computationally efficient approximation to the nearest neighbor interchange metric. J. Classif. 1:93–124.

Chan, S., H. Bernard, S. Ong, S. Chan, B. Hofmann, and H. Delius. 1992. Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: A showcase for the molecular evolution of DNA viruses. J. Virol. 66:5714–5725.

Chan, S., H. Delius, A. Halpern, and H. Bernard. 1995. Analysis of genomic sequences of 95 papillomavirus types: Uniting typing, phylogeny, and taxonomy. J. Virol. 69:3074–3083.

Chang, J. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Bio. 137:51–73.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1984. Distance methods for inferring phylogenies: A justification. Evolution 38:16–24.

Felsenstein, J. 1993. PHYLIP (Phylogenetic inference package), version 3.5p. Univ. of Washington, Seattle.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Hasegawa, M., H. Kishino, and N. Saitou. 1991. On the maximum likelihood method in molecular phylogenetics. J. Mol. Evol. 32:443–445.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

Lewis, P. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol. Biol. Evol. 15:277–283.

Maddison, D. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40:315–328.

Ong, C., S. Nee, A. Rambaut, H. Bernard, and P. Harvey. 1997. Elucidating the population histories and transmission dynamics of papillomaviruses using phylogenetic trees. J. Mol. Evol. 44:199–206.

Page, R. 1993. On islands of trees and the efficacy of different methods of branch swapping in finding most-parsimonious trees. Syst. Biol. 42:200–210.

Page, R. 2001. COMPONENT, version 2.0. The Natural History Museum, London.

Rogers, J. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Syst. Biol. 46:354–357.

Salter, L. 1999. Simulation-based estimation of phylogenetic trees. Ph.D.Dissertation, Ohio State Univ., Columbus.

Salter, L., and D. Pearl. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst. Biol. 50:7–17.

Sanderson, M., and J. Kim. 2000. Parametric phylogenetics? Syst. Biol. 49:817–829.

Swofford, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Yang, Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.

Received 8 March 2000; accepted 9 October 2000
Associate Editor: R. Olmstead

Syst. Biol. 50(6):978–982, 2001

Difficulties in Detecting Hybridization

Mark T. Holder,[1,2] Jennifer A. Anderson,[1] and Alisha K. Holloway[1]

[1] *Section of Integrative Biology, School of Biological Sciences, University of Texas, Austin, Texas 78712, USA;*
E-mail: mtholder@mail.utexas.edu, janders@mail.utexas.edu, aholloway@mail.utexas.edu
[2] *Institute for Cellular and Molecular Biology, University of Texas, Austin, Texas 78712, USA*

Sang and Zhong (2000) proposed a test to distinguish between hybridization and lineage sorting, two of the many evolutionary processes that can produce discordant gene trees. Consider two gene trees for three taxa A, B, and C (Fig. 1A,B). If B and C are sister taxa on gene tree 1, and A and B are sister taxa on gene tree 2, then either taxon B is a hybrid species (Fig. 1C), or one of the gene trees is incorrect because of lineage sorting (Fig. 1D,E). When faced with these two discordant gene trees, Sang and Zhong