

SCRIBE NOTES FOR 03/04/2008

Compiled by: Bakhtiyar Uddin

Content: 'Hobgoblin of Phylogenetics' & 'Inferring Phylogenetic complexities'

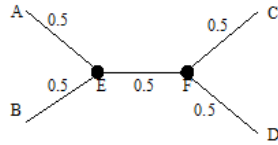


Figure 1: Kimura-2-parameter of evolution with equal probabilities of change in all branches.

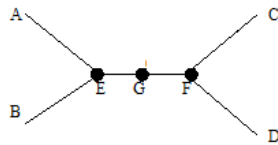


Figure (b)

Figure 2: Tree in figure 1 is ultrametric tree. This is obvious if we create additional node G at the center of edge E-F and then root the tree at node G.

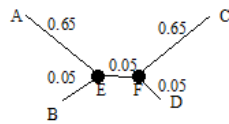


Figure 3: Felsenstein model tree

Figure 1: figures used for the notes

Introduction:

This note describes two articles from Nature that were discussed in the class.

Hobgoblin of Phylogenetics:

The authors (D. Hillis, J. Huelsenbeck) tries to illustrate that statistically consistency does not necessarily lead to preference for a method when only finite amount of data are available. We also have to consider how efficient it is, the amount of data it requires before it converges to the true tree. Attaching too much importance to inefficient method that requires large amount of

data to converge to the true model tree may be misguided.

To illustrate this, they use two computer simulation results.

The first model is a Kimura-2-parameter model of evolution with equal probabilities of change in all the branches (figure 1). As discussed in class, this model is ultra metric. To see how that is true, we can first calculate the edge length of each edge and then we can subdivide the edge at the middle, as shown in figure 2. Then we can root the tree at G. The tree would be ultra-metric because of the symmetry in the probability of change at the edges. So all the leaves will have the same distance from the root G.

The model tree being ultra metric, UPGMA on this model would be consistent. Neighbor Joining method under Kimura distances is also statistically consistent. But, Neighbor joining with Kimura distances requires 5000 nucleotides to reach 90% accuracy and standard parsimony achieves it with only 600 nucleotides. Moreover, parsimony analysis with transversions weighted more heavily than transitions reaches 100% accuracy with only 300 nucleotides.

The second simulation compares the performance of methods when some are known to be inconsistent (figure 2). Neighbor Joining with Jukes Cantor distances, UPGMA and all the parsimony variants are statistically inconsistent under this model. They all converge to the wrong tree. Maximum Likelihood with Kimura distances is statistically consistent because the data generated from the model comes from the model.

Using these simulations they have argued that statistical methods are not always better than Parsimony. Parsimony performs better when frequent events have less weight. Under certain conditions parsimony might fail, but they showed that under those conditions "statistical methods" also fail for sequences of even thousands of bases.

Additional Class Discussion:

Consider a binary sequence data, where each character-edge combination

has mutation probability, $p(e, c)$ which is the probability of change on edge e , given the state of the character at the tail node of the edge.

$$0 < p(e, c) < \frac{1}{2}$$

Suppose we would like to run ML on data generated under a four leaf tree model. Assume we allow 100 characters. There are 5 edges in the tree and we have to consider the probability of change on an edge for every possible initial character.

We would have to estimate 500(100 characters \times 5 edges) probabilities.

ML would return the tree that maximizes the probability.

Inferring complex phylogenies:

In this paper the author(David M. Hillis) looks at one model tree. The tree was modeled under Kimura-2-Parameter. After taking a collection of sequences, the author estimates a phylogeny and various parameters of the model.

He tested three simple algorithms for approximating phylogenetic solutions - stepwise addition under the parsimony criterion, neighbor joining under minimum evolution criterion and UPGMA clustering algorithm. In their parsimony method, they use the stepwise addition technique. Then, they add a taxa incrementally and optimize the MP score.

Parsimony, even though the simplest one, outperformed neighbor joining. Only UPGMA failed to reconstruct more than 99% of the branches in the model tree with 5000 nucleotides. The author also found that, branch swapping on the stepwise addition parsimony tree revealed four equally good solutions. One of those solutions matched the model tree exactly.

The author explains the unexpected ease of reconstruction by dispersal of noise in the data set across many branches in the tree. Because of that the phylogenetic signal are detected above the background noise, even when the phylogenetic signals are weak. The study shows that adding large number of additional taxa to phylogenetic analysis increases the accuracy of estimated trees and reduce the need for complex methods of analysis.