Efficiency of Model-Based Complexity Method for Estimating Unrooted Multifurcate Phylogenetic Tree

Fengrong Ren **Hiroshi Tanaka**
rencom@mri.tmd.ac.jp tanaka@cim.tmd.ac.jp
Department of Bioinformatics, Tokyo Medical and Dental University
1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

1 Introduction

In evolutionary study, many methods have been proposed so far for reconstructing phylogenetic tree from molecular data. But it has been pointed out that multifurcate phylogenetic trees are difficult to be correctly reconstructed by conventional methods. In order to resolve this problem, we have been engaged in developing a new method for reconstructing molecular phylogenetic tree, based on the minimum complexity principle that is widely used in the inductive inference. The results of our previous studies have proved that this method, which we call "minimum model-based complexity (MBC) method", is quite efficient in estimating rooted multifurcate tree. In this study, we make further investigations about this method in the case that trees are unrooted by using computer simulation and statistical test. It was found that MBC method also shows good performance even in the case of unrooted multifurcate tree and suggest that this method could be generally used for reconstructing phylogenetic tree having arbitrary multifurcation. The details of our MBC method would be referred to our previous paper [1].

2 Computer Simulation of Unrooted Tree

We take a bifurcate and a multifurcate tree with 4 OTUs as the topology models to be compared. The length of branch c in Fig. 1 is varied from 0.0 to 0.1 with interval 0.01 to realize the bifurcate and multifurcate tree. That is, if $c = 0$ then the topology becomes multifurcate tree; if c is relatively small comparing with a and b then the topology will be nearly multifurcate, whereas if c is statistically significantly large then the tree becomes definitely binary. Two groups data which are assumed to be 1000 bp and 3000 bp respectively are generated.

Three criteria, maximum likelihood, Akaike information criterion and model-based complexity are applied to this simulation to select whether the trees are bifurcate or multifurcate. When $c = 0$ or nearly 0, both AIC and MBC are supposed to select the multifurcate tree, but they become more inclined to select the bifurcate tree as the true value of branch length c increases. At some c value, the frequencies to select multifurcate and bifurcate tree become almost equal. We denote this value of c by $c_{AIC}(0.5)$ or $c_{MBC}(0.5)$.

The results by complexity criteria are examined by using statistic test. Since even if the true value of the branch c is 0, the estimated value $\hat{c}$ would distribute around 0. Here the boundary value of 95% confidence interval of the estimation is considered as the threshold $\theta_c(0.05)$ to accept or reject $c = 0$: (1) if the estimation $\hat{c}$ exceeds the threshold $\theta_c(0.05)$, which means the node is bifurcate. (2) if the estimation $\hat{c}$ is below the threshold $\theta_c(0.05)$, which means that the node is multifurcate.

Since the $c_{AIC}(0.5)$ and $c_{MBC}(0.5)$ values in the complexity criterion approach behave the same role as the thresholds in statistical test, we can compare the $\theta_c(0.05)$ with $c_{AIC}(0.5)$ and $c_{MBC}(0.5)$ to clarify the features of these complexity criteria in selecting the tree topology.

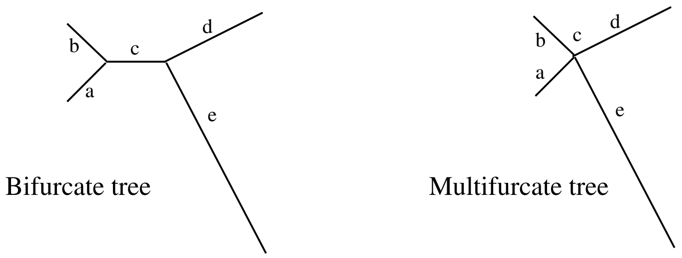

Figure 1: Bifurcate and multifurcate tree models used for computer simulation.

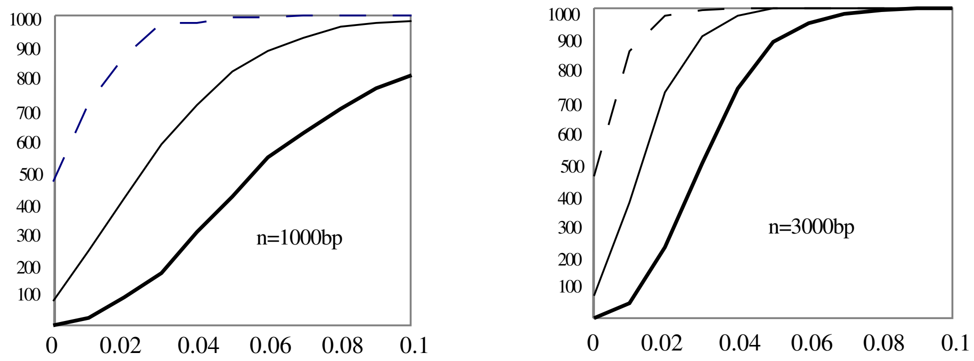

Figure 2: Change of the selection ratio when c is varied. Ordinate denotes the frequency that bifurcate tree is selected among simulated 1000 samples. Abscissa denotes the true c value used in generation of the data. Dotted line indicates ML method, whereas thin line and thick line indicate AIC and MBC method, respectively.

3 Results and Conclusion

The AIC method shows good correspondence with statistical test based on simulated data, but in this simulation, exactly the same evolution model $\{P_{ij}(u)\}$ is used in generation of the data and estimation of branch lengths. In real situations, we should also estimate the evolution model from the data, which certainly produces estimation errors. If we take this effect into account in determining the confidence interval of the estimation, the threshold values of statistical test will becomes larger. Hence we would say that AIC method, needless to say ML method, overestimates the complexity of tree topology to prefer the bifurcate tree. On the contrary, the underestimation in MBC method is not fatal but much more appropriate when the unpredictable sources of estimation variation in the modeling are considered.

References

[1] Tanaka, H., Ren, F., Okayama, T., and Gojobori, T., Inference of Molecular Phylogenetic Tree Based on Minimum Model Based Complexity Method, *Proc. 5th Intl. Conf. on Intelligent Systems for Molecular Biology*, 319–328, 1997.