Comparison of Information Criteria and Simulation Statistics Method for Topology Selection in Molecular Phylogenetic Tree

Soichi Ogishima ogishima@bioinfo.tmd.ac.jp
Fengrong Ren rencom@mri.tmd.ac.jp
Hiroshi Tanaka tanaka@cim.tmd.ac.jp

Departmentif of bioinformatics, Medical Research Institute,
Tokyo Medical and Dental University

Keywords: model-based complexity, maximum likelihood method, Akaike information criterion

1 Introduction

In evolutionary studies, a number of methods have been proposed so far for reconstruction of phylogenetic tree. But, it has been pointed out, that even maximum likelihood method which is considered as the most rigorous one still has some problems for estimating multifurcate tree. Two kinds of approaches are thought to be appropriate to resolve this problem. One is the approach by information criteria such as Akaike information criterion (AIC) and Rissanen's minimum description length (MDL) criterion by which we can deal with the tree topology selection as a sort of model selection problem. Another is to construct statistical framework to test whether a certain length of branch is 0 or not, and thus multifurcate node is detected. In our previous studies [1, 2], we have investigated the qualitative characteristics of AIC and our model-based complexity (MBC) method, which is a variant of MDL method, in estimating the appropriate tree topology. But we have not rigorously examined the effects of modeling error in base substitution. In this study, we will take this kind of estimation error into account and strictly compare the efficiency of MBC method with ML and AIC methods by computer simulation and the results will be examined by statistical hypothesis tests.

2 Methods and Results

We assume a tree with 4 OTUs (a, b, d, e) as the topology model and the length of internal branch c is varied from 0.0 to 0.1 to distinguish multifurcation from bifurcation. 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 group's data which are assumed to be 1000bp and 3000bp respectively are generated.

1. Information criteria Three criteria, ML, AIC and MBC are applied to this simulation. We investigated the changes of the frequencies in selecting tree topology when the true branch length c varies. As at some c value, the frequencies to select bifurcate or multifurcate tree will become equal, we denote this value of c by $c_{ML}(0.5)$, $c_{AIC}(0.5)$, $c_{MBC}(0.5)$ and call it equiprobable (EP) value.

2. Simulation statistics method The results by three complexity criteria are compared by using statistical hypothesis test. Null hypothesis H_0: the branch length c is 0 and the corresponding node is multifurcate. Alternative hypothesis H_A: the branch length c is not 0 and the corresponding node is bifurcate. We also examined the probability of type II error. It is the probability of error that H_0 is not rejected though H_A is correct. The value $c_{SS}(0.5)$ denotes the true c value in which

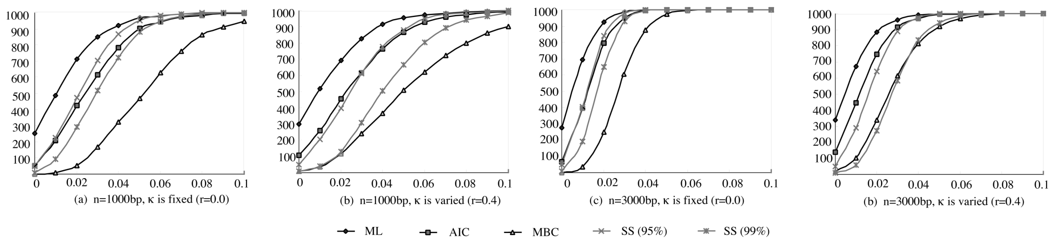

Figure 1: Changes of the selection ratio among the two topologies obtained from different methods. Ordinate denotes the frequencies that the bifurcate trees are selected among1000 samples. Abscissa denotes the true c values used in generation of the data.

Table 1. The results when κ is fixed (r = 0)

	$c_{AIC}(0.5)$	$c_{MBC}(0.5)$	$c_{SS}^{0.05}(0.5)$	$c_{SS}^{0.01}(0.5)$
1000bp	0.024	0.051	0.021	0.029
3000bp	0.012	0.027	0.012	0.017

Table 2. The results when κ is varied (r = 0.4)

	$c_{AIC}(0.5)$	$c_{MBC}(0.5)$	$c_{SS}^{0.05}(0.5)$	$c_{SS}^{0.01}(0.5)$
1000bp	0.023	0.049	0.026	0.041
3000bp	0.011	0.026	0.017	0.029

the estimation of type II error (β) becomes 0.5. This value characterizes the global nature of this statistical test to detect multifurcation and has the same role as the previous EP value.

3. Results The changes of the frequencies in selecting a bifurcate or multifurcate tree when the branch length c varies are shown in Figure 1. We can see that the ML method is almost always to select a bifurcate tree except few cases. On the contrary, AIC, MBC and simulation statistics method show more natural tendencies that fit our intuition. Thus, we only compare the results obtained by AIC, MBC and simulation statistics method in the following.

Table 1 shows the results when κ (κ = transition/transversion rate ratio) is fixed. In the case that simulated sequences are assumed to be 1000bp, the threshold $c_{SS}^{0.05}(0.5)$ is 0.021 and corresponds well with AIC EP value, 0.024, whereas the MBC EP value, 0.051, exceeds even 99% threshold $c_{SS}^{0.01}(0.5)$, 0.029. The same tendency is observed in the case that simulated sequences are assumed to be 3000bp.

However, when κ is varied to take the estimation errors of κ into account, the results are quite different (see Table 2). Although the thresholds $c_{SS}^{0.05}(0.5)$ is 0.026 and still correspond well with AIC EP values, 0.023, but the threshold $c_{SS}^{0.01}(0.5)$, 0.041, corresponds well with MBC EP value, 0.049. Especially, when the sequence length increases to 3000bp, the MBC EP value, 0.026, corresponds very well with threshold $c_{SS}^{0.01}(0.5)$, 0.029.

In realistic world, the exact evolution model of species is usually unknown and thus we have to estimate the model from molecular data. Considering this situation, our MBC method would provide a better criterion than ML and AIC methods for selecting tree topology.

References

[1] Ren, F., Tanaka, H. and Gojobori, T., Construction of molecular evolutionary phylogenetic tree from DNA sequences based on minimum complexity principle. *Computer Methods and Programs in Biomedicine*, 46:121–130, 1995.

[2] 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.