Pacific Symposium on Biocomputing 4: 326-337 (1999)

TOPOLOGY SELECTION IN UNROOTED MOLECULAR PHYLOGENETIC TREE BY MINIMUM MODEL-BASED COMPLEXITY METHOD

H.TANAKA, F.REN
Tokyo Medical and Dental University, 1-5-45 Yusima, Bunkyo, Tokyo 113, Japan

T.OKAYAMA, T.GOJOBORI
National Institute of Genetics, 1111 Yata. Mishima, shizuoka 411, Japan

In reconstruction of phylogenetic trees from molecular data, it has been pointed out that multifurcate phylogenetic trees are difficult to be correctly reconstructed by the conventional methods like maximum likelihood method(ML). In order to resolve this problem, we have been engaged in developing a new phylogenetic tree reconstruction method, based on the minimum complexity principle widely used in the inductive inference. Our method, which we call "minimum model-based complexity(MBC) method", has been proved so far to be efficient in estimating multifurcate branching when the tree is described in the form of rooted one. In this study, we make further investigations about the efficiency of MBC method in estimating the multifurcation in unrooted phylogenetic trees. To do so, we conduct computer simulation in which the estimations by MBC method are compared with those by ML, AIC and statistical test approach. The results show that MBC method also provides good estimations even in the case of multifurcate unrooted trees and suggest that it could be generally used for reconstruction of phylogenetic tree having arbitrary multifurcations.

1 Introduction

In evolutionary studies, many methods have been proposed so far for reconstruction of phylogenetic trees from molecular data. Among these methods, maximum likelihood (ML) method [1] is considered as the most rigorous and widely used. But, as it has been pointed out, even ML method still has some problems for relevant reconstruction of evolutionary trees.

One of such problems lies in the estimation of appropriate tree topology. Like other conventional methods, ML method always reconstructs fully expanded binary phylogenetic trees. But the fully expanded binary tree, in other words, the most complex tree is neither always true nor needed. We think the degree of complexity of the molecular phylogenetic tree should correspond with the information amount contained in the available molecular data. In fact, multifurcate phylogenetic trees would be appropriate when we are not certain about the detailed branching order due to the insufficient amount of information, or when phylogenetic branches are so simultaneous that we could not determine from molecular data alone. In conventional methods, the rel-

evant level of the complexity of phylogenetic trees has been never considered seriously, so that extra-complex trees which do not correspond with the amount of information in the data have been often reconstructed.

To resolve this problem, we have been engaged in developing a new method [2,3] which can deal with this multifurcation problem. Our method is based on the inductive inference theory to extract the model having the relevant complexity level corresponding with information contained in the data.

In our method, the complexity of the data is measured by what we call "model-based complexity (MBC)", which is defined as the sum of the complexity of the model and that of the data with respect to the model. In this MBC method, the model-based complexities of molecular data for the various candidate trees including multifurcations are compared, and the tree which shows the minimum in its model-based complexity is selected to be true one.

In the previous paper [4], we have conducted the numerical experiments to investigate the precision of the MBC method in estimating rooted phylogenetic tree topologies in comparison with those of ML and AIC method. It was found that MBC method provides better results in estimating the tree topology having multifurcation comparing with ML and AIC method.

But, in conventional phylogenetic analyses, most of the trees are estimated in the form of the unrooted tree with branch length described as the number of base substitution per site. Therefore, we develop a new version of MBC method for reconstruction of unrooted phylogenetic trees, which would enable us to perform more straightforward comparison with other methods. We employ the computer simulation to investigate the precision of the new version of our method in estimating unrooted phylogenetic tree topologies, in comparison with those of ML, AIC and the statistical test approach.

2 Method

2.1 Model-based Complexity

As is widely known, in the inductive inference, there would be many theories which can explain the given data to the equal extent, so that we would have to use a certain criterion to select the best one. In this context, so called "principle of parsimony" or "minimum complexity principle " is often used, which states that the theory which has the least complexity and nevertheless explains the data well should be chosen as the first option for true one. There are several ways to make this principle applicable to the real data [5,6,7,8]. For example, in Rissanen's MDL(Minimum Description Length) principle, the complexity of data is measured by the code length of a statistical model M, $l(M)$ plus

code length of data with respect to the model M, $l(D/M)$. Then minimization procedure is taken by varying the model M within the assumed model family. Wallace's MML(Minimum Message Length) is proposed also in the same line of thought except that it considers average length of messages.

The complexity that we proposed is essentially same to the Rissanen or other similar approaches, but, the detailed structure of complexity is investigated in relation to the model inference in our formulation. Suppose we take some family of model set $M = \{M_\lambda / \lambda \in I\}$ (I is some index set of λ) which is supposed to generate data sequence $D = \{x_1, \cdots, x_n\}$. The model-based complexity of data is defined as

$$K_M(D) \equiv \inf_{M_\lambda} \{K(M_\lambda) + K(D/M_\lambda)\}, \tag{1}$$

where $K(M_\lambda)$ is an appropriate measure of complexity defined on the model M_λ, and $K(D/M_\lambda)$ is the complexity of the residuals which the model M_λ cannot explain.

In the model-based complexity, the first term of the equation (1) is formulated in detail. In the ordinary modeling, the model space in which the best model is to be explored has its own structure (composed of classes) exhibiting various degree of complexity. To characterize this structure, we can use some index parameters $\xi(M)$ which define the model classes. We call this kind of parameters as **compositional parameters** of model space. The frequent ways to introduce the measure of complexity into these model classes are: (1) to assign(universal) prior probability $p[\xi(M)]$ to the each element contained in these classes and uses $-\log p[\xi(M)]$ as a measure of complexity for this element, or (2) to assign the logarithm of the size (cardinality) of each j-th class, $\log |M^\xi|$, as complexity measure of the elements contained in that class if the cardinality is finite. In the case that the cardinality is infinite, we can use ε-entropy for suitably chosen ε-net introduced into the model classes.

Other than the compositional parameters which specify the model class, there are ordinary parameters which are estimated from data and define a particular model element in the model class. We call these ordinary parameters as **inferential parameters**. There are several approaches to describe the complexity of inferential parameters. Well-known is Akaike's AIC [9], the half of which is given by $-\log L(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + k$, where k is the number of inferential parameters which also describes its complexity, $\boldsymbol{x}$ is the data, and $\hat{\boldsymbol{\theta}}$ is ML estimator of parameter $\boldsymbol{\theta}$. $L(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ is the maximum likelihood of data x with ML estimator $\hat{\boldsymbol{\theta}}$.

The other approach to the inferential complexity is given by Rissanen. In describing the total code length, he added code length for describing the

precision of data to the ordinary code length of $-\log L(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$: the approximate term of this is $\frac{k}{2}\log n$, where n is the number of data samples. This term is also obtained from the Bayesian viewpoint. In the Bayesian framework, the posterior probability of the model given data $p(\boldsymbol{\theta}|x)$ is proportional to $p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta}$ is a prior probability of $\boldsymbol{\theta}$. If we take its negative logarithm, then corresponding model complexity is given by $-\log\pi(\boldsymbol{\theta})$. We can use non-informative prior of parameters by Jeffrey for $\pi(\boldsymbol{\theta})$, namely, $\frac{1}{2}\log\det I^F(\boldsymbol{\theta})$ where $I^F(\boldsymbol{\theta})$ is Fisher's information matrix. This term asymptotically approaches to $\frac{k}{2}\log n + O(1)$, if n goes to infinite. Thus we have essentially equivalent definition of inferential complexity.

But, in real applications, not all the parameters are independent so that more feasible definition is to introduce the effective dimension of the inferential parameter space by applying eigenvector analysis to Fisher information matrix. Hence, the empirical inferential complexity of data is given by

$$K_{in}(\boldsymbol{\theta}) = \frac{\epsilon-\dim(\boldsymbol{\theta})}{2}\log n, \tag{2}$$

where $\epsilon-\dim(\boldsymbol{\theta})$ is the effective number of empirically independent parameters. Usually the effective number of the components is determined so that the sum of the eigenvalues up to that component amounts to 95% or 90% of the toatal sum of eigenvalues.

Hence, the total model-based complexity is composed of three terms which are (1) compositional complexity of the model, $K_c(m)$, (2)inferential complexity of the model, $K_{in}(m)$ and (3) empirical KL information between the data and model, $I^{KL}(\mathbf{x}/\boldsymbol{\xi},\boldsymbol{\theta})$. Then model-based complexity of data is specified as

$$K_M(D) = \min_{\boldsymbol{\xi},\boldsymbol{\theta}}\{K_c(\boldsymbol{\xi}) + K_{in}(\boldsymbol{\theta}) + I^{KL}(\mathbf{x}/\boldsymbol{\xi},\boldsymbol{\theta})\}. \tag{3}$$

Thus, we have reached the concrete form of the definition of the general model-based complexity. We can use this model-based complexity to extract the model from data by finding the model which minimize the model-based complexity.

2.2 Complexity of Evolutionary Tree

In the molecular phylogenetic tree, the model space M_T is decomposed into two subspaces. One is **tree model** which is determined by tree topology T_p and branch lengths **b** described either in time scale or average number of base substitutions. The other is **evolution model** in which the base substitution probability between two of four bases during time t is given by Markov transition matrix.

First, we describe the **tree model**. In evolutionary tree, the class of tree topology is determined by the number of leaves n_l and that of internal nodes v. From graph theory, we know that the number of branches n_b is related with n_l and v as

$$n_l + v = n_b + 1. \tag{4}$$

This relation holds for both unrooted and rooted trees. In constructing phylogenetic tree, the number of specie (leaves) n_l is fixed, so that only v or equivalently n_b can be varied. We take v as the parameter defining the class of tree topology: the compositional parameter of the tree. If v equals 1, we have a star-shaped unrooted tree. With the increasing of the v, the tree becomes more complex. In the case of unrooted tree, when $v = n_l - 2$, we have fully expanded binary tree. The complexity of natural number v is given by $\log^* v$ by Rissanen, where $\log^* v = \log v + \log\log v + \log\log\log v + \ldots$

Even if the number of internal nodes v is determined, the tree is not unique. Rissanen gives approximation of possible number of the rooted tree topologies defined by v: $\binom{n_l + v - 2}{v}$. In the case of unrooted trees, this approximation is the number of branch times bigger because the root might be possibly located inside any of branches, so that the number of topologies is obtained by dividing it by the number of branches. If we assume each of these trees to be equally probable, the resultant complexity of the tree topology is given by

$$K_c(v) = \log^* v + \log \frac{1}{n_b}\binom{n_e + v - 2}{v}. \tag{5}$$

The complexity of branch lengths which are considered as inferential parameters is given by

$$K_{in}(\mathbf{b}) = \frac{\epsilon - \dim(\mathbf{b})}{2} \log n, \tag{6}$$

where $\mathbf{b} = (b_1, ...b_n)$ is the vector of branch lengths, n is the length of nucleotide sequence, and $\epsilon - \dim(\mathbf{b})$ is the number of efficient eigen vectors of Fisher's information matrix $I^F(\mathbf{b})$.

Second, we describe the **evolution model**. To do it, we first model the base substitution probability with time t, denoted by the transition matrix $\{P_{ij}(t)\}$ where i and j is one of the four nucleotides. The transition matrix in time scale is derived from rate matrix for base substitution $\mathbf{R}$, based on the relation: $P(t) = \exp\{\mathbf{R}t\}$. We use HKY model for rate matrix $\mathbf{R}$ in which the difference of transition (α) and transversion(β) base substitution rate and the frequency of each base(π_i) are taken into account. The details would be referred elsewhere[11]. Then we convert time scale t to relative scale u (average

number of base substitutions per site) to get $\{P_{ij}(u)\}$. The conversion equation from time scale t to relative scale u is given by,

$$u(t) = -t \sum_{k=1}^{4} \pi_k R_{kk} \tag{7}$$

We use this transition matrix $\{P_{ij}(u)\}$ to construct the whole likelihood of the molecular sequences $S_1, S_2, \cdots, S_e$ when the tree topology is assumed to be M_T. The procedure to construct the likelihood function is same as the Felsenstein [1]. Taking the sum of the complexities of the each component of a phylogenetic tree, we have the following total complexity of phylogenetic tree to be minimized:

$$\begin{aligned} K_{M_T}(S_1, S_2, ..., S_{n_l}) &= \min_{v,\mathbf{b}} \{K_c(v) + K_{in}(\mathbf{b}) + I^{KL}(D/v, \mathbf{t})\} \\ &= \min_{v,\mathbf{b}} \Bigg\{ -\log L(S_1, S_2, \cdots, S_e/M_T) \\ &+ \left[\log^* v + \log \frac{1}{n_b} \binom{n_l + v - 2}{v} \right] \\ &+ \frac{\epsilon - \dim(\mathbf{b})}{2} \log n \Bigg\}. \end{aligned} \tag{8}$$

Where $\log L(S_1, S_2, \cdots, S_e|M_T)$ is log likelihood function which approximates empirical KL information of model with respect to sequence data.

2.3 Computer Simulation of Unrooted Tree

Model and Method

We take a bifurcate and a multifurcate tree with 4 OTUs (shown in Fig.1) as the topology models to be compared. In the bifurcate tree, the branch lengths of a and b are assumed to be equal to 0.2 in terms of average base substitution number per site, and $d = a + c = b + c$. The branch e is assumed to be connected with an outgroup species. Then the length of branch c is varied from 0.0 to 0.1 with interval 0.01 to realize the bifurcate or 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.

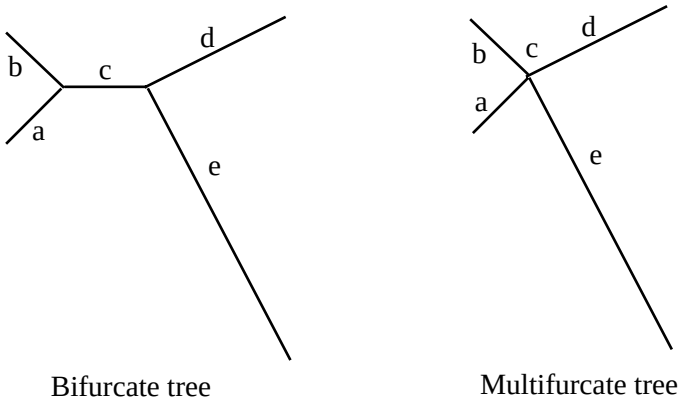

Figure 1: Bifurcate and multifurcate tree models used for computer simulation

Since any point can be a starting point in the unrooted tree, we simulate the evolution process from one of arbitrary root(ancestor sequence) to the each leaf(4 nucleotide sequences) by calculating base transition matrix $\{P_{ij}(u)\}$. In this simulation two groups data are generated: one group consists of 1000 simulated sequences whose lengths are assumed to be 1000bp, and another group consists of 1000 simulated sequences whose lengths are assumed to be 3000bp. Other parameters such as transition rate α and transversion rate β and frequency of the four bases π_i are fixed.

Complexity Criterion Approach

The complexity criteria which are applied to this computer simulation are: (1) negative logarithm of the maximum likelihood (ML) which describes model fitness for the data by $-logL$, (2) Akaike information criterion (AIC) which describes both model fitness for the data and model dimensionality(complexity) by $-logL + k$, and (3) the model-based complexity (MBC) which describes both model fitness for the data and model complexity by $-logL + K_c(v) + \frac{\epsilon - \dim(\mathbf{b})}{2} \log n$. Since all of three methods estimate the parameter values by maximizing the same likelihood function, estimated parameter values are same among them. But, as the additional complexity terms are different, they make different selections for tree topology.

To clarify the global nature of estimation by these complexity criteria, we investigate the change of the frequencies in selecting the bifurcate and

multifurcate tree when the branch length c varies. When $c = 0$ or nearly 0, both AIC and MBC select the multifurcate tree but AIC and MBC become more inclined to select the bifurcate tree as the true value of branch length c increases. Finally 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 $c_{MBC}(0.5)$ and call it equiprobable(EP) value.

Statistical Test Approach

The results by complexity criteria are evaluated by using the statistical test approach. In this approach, we conduct Neyman-Pearson type hypothesis test on whether the length of branch c is 0 or not, using a boundary value of confidence interval of estimated value $\hat{c}$.

Even if the true value of the branch c is 0, the estimated value $\hat{c}$ would distribute around 0. The boundary value of 95% confidence interval of the estimation is ordinarily considered as the threshold $\theta_c(0.05)$ to accept or reject $c = 0$. The hypotheses in this case are: (1)Null hypothesis H_0: the branch length c is 0 and the corresponding node is multifurcate (degenerate). (2)Alternative hypothesis H_A: the branch length c is not 0 and the corresponding node is bifurcate(fully expanded). With the hypotheses of H_0 and H_A, the statistical test framework is as follows: (1)Null hypothesis H_0 is rejected and H_A is accepted if the estimation $\hat{c}$ exceeds the threshold $\theta_c(0.05)$, which means the node is bifurcate. (2)Null hypothesis H_0 is not rejected but accepted if the estimation $\hat{c}$ is below the threshold $\theta_c(0.05)$, which means that the node is multifurcate.

The remained problem is to determine the probabilistic distribution of $\hat{c}$ when the true value of c is 0. In ordinary statistics, the distribution is approximated at normal distribution with estimation error variance. In this simulation, we apply ML method to the simulated data generated assuming $c = 0$ and obtain the distribution of $\hat{c}$. The 95% threshold value $\theta_c(0.05)$ is determined from this distribution. The decision rule for topology selection is whether estimated value $\hat{c}$ is greater than its threshold or not. This threshold also describes the amount of uncertainty in estimated value $\hat{c}$. We can also use the 99% boundary $\theta_c(0.01)$ as a threshold but in this case, the decision is much more conservative to select multifurcate tree.

In comparison with the complexity criterion approaches, the statistical test approach is conventionally more accepted though it needs the calculation of test statistics (thresholds). On the contrary, the complexity criterion approaches are relatively easy to use because what we need is to calculate the magnitudes of the criteria in two models to be compared. Since the EP values

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

3 Results

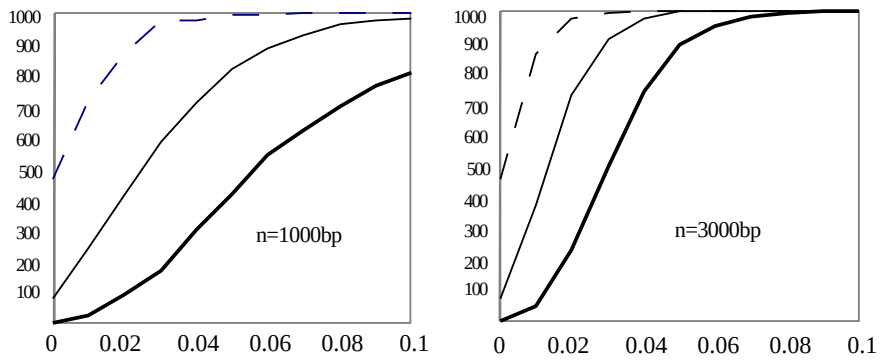

Figure 2: Change of the selection ratio among two topologies by various complexity criteria when c is varied. Ordinate denotes the frequency that bifurcate tree is selected among simulated 1000 samples. Abscissa denotes the true c value used in generation of the data. Dotted line indicates ML method,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

The results of the complexity criterion approach are shown in figure 2. We can see that ML method almost always selects the bifurcate tree with very few exceptions in which the estimations exactly bring $\hat{c} = 0$ within numerical truncation. Even if the true $c = 0$ (exactly multifurcate), the ML criterion selects bifurcation about 50% of the cases. On the contrary, AIC and MBC method show more natural tendencies which fit our intuition. Therefore, in this section we only investigate the features of AIC and MBC method. In the case that the sequence lengths are assumed to be 1000bp, the $c_{AIC}(0.5)$ is 0.025. It means that AIC considers the tree as multifurcate until the true c value is varied up to 0.025 ($0 \leq c \leq 0.025$). But, MBC method is more conservative in selection of bifurcate tree, because the topology is not considered as bifurcate until the c value is varied up to 0.056.

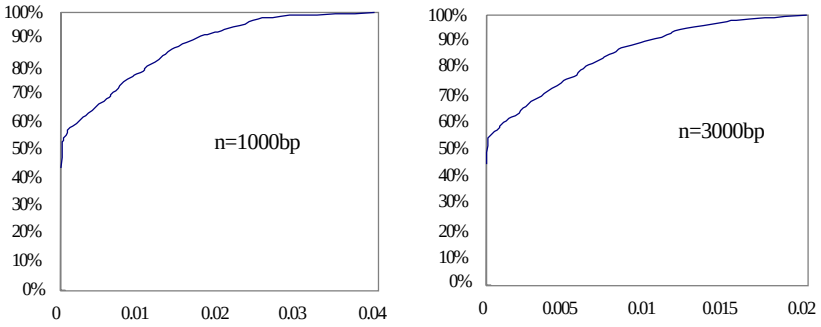

Figure 3: Ogive curve of estimation value $\hat{c}$ when $c = 0$. Ordinate denotes the cumulative probability of $\hat{c}$ and abscissa denotes $\hat{c}$.

On the other hand, in the statistical test approach, the 95% threshold $\theta_c(0.05)$ in the case of 1000bp is 0.022 (Fig.3). This means that even if true c value equals to 0, $\hat{c}$ varies within 95% interval of estimation ($0 \leq c \leq 0.022$). This threshold 0.022 corresponds well with AIC EP-value 0.025, whereas MBC EP-value 0.056 exceeds even 99% threshold ($\theta_c(0.01) = 0.0303$). Therefore, if we consider the statistical approach is reliable, AIC method appears to be relatively good criterion for selection of multifurcate topology and MBC seems to be biased to multifurcate topology.

In the case that the sequence length is assumed to be 3000bp, $c_{AIC}(0.5)$ is 0.0135 and $c_{MBC}(0.5)$ is 0.0299. On the other hand, the threshold of statistical test is 0.0124 for 95% confidence and 0.0174 for 99% confidence interval. Hence, again, AIC criterion corresponds well with the test threshold and the MBC criterion has a tendency to select multifurcate topology.

4 Discussion

How we interpret above results. If we think the statistical test approach is appropriate, AIC method seems to provide fairly good selection between the bifurcate and multifurcate tree. But in this simulation, exactly 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 (base substitution 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 becomes larger. Here we do

not have the exact value when this uncertainty of the evolutionary model is considered. If this effect is as same as branch length estimation error, we have $\sqrt{2}$ times wider confidence interval. Then 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 are considered. Though MBC still underestimates the complexity of tree to prefer the multifurcate tree, but this tendency will be improved when the number of data is increased. We are now engaged in incorporating other sources of estimation variation into the simulation such as the blurring of the transition matrix to get more actual estimation of tree topology.

Alternative applications of the minimum complexity principle to the reconstruction of phylogenetic tree have been studied by Cheeseman [12] and Allison [13]. Although these studies are same as our study in using the minimum complexity principle. they are different from ours in that they apply the minimum complexity criterion directly to the base by base change of nucleotides along the tree, whereas our method applied the complexity measure along the Felsenstein's way of likelihood modeling.

The results in this study are roughly corresponding to those of rooted tree in our previous results. Thus we can conclude the MBC method provides fairy robust topology selection as compared with ML or AIC method in both rooted and unrooted trees.

5 Conclusion

In this study, by using computer simulation, the efficiency of model-based complexity method in estimation of unrooted phylogenetic tree is investigated in comparison with conventional tree reconstruction methods like maximum likelihood method and Akaike's information criterion methods. The results suggest that, if we consider the real situation in which the evolution model is often inexact, MBC method would provide a better criterion for tree topology selection than the above two conventional methods.

Acknowledgments

We thank three anonymous referees for criticisms and comments which enable us to make many improvements to this paper. This research is partly supported by a grant from National Institute of Genetics(NIG Cooperative Research Program('98-64).

References

1. J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.* 17, 368-376(1981).
2. F. Ren, H. Tanaka and T. Gojobori, Construction of Molecular Evolutionary Phylogenetic Tree from DNA Sequences Based on Minimum Complexity Principle. *Computer Methods and Programs in Biomedicine,* 46, 121-130(1995).
3. H. Tanaka Model-based Complexity and Inductive Inference, *Proc. of 4th Int'l Workshop on Rough Sets, Fuzzy Sets and Machine Discovery,* 144-152(1996).
4. H.Tanaka, F.Ren, T.Okayama and T.Gojobori, 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).
5. R.J. Solomonoff, A formal theory of inductive inference, *Information and Control*, 7:1-22, 224-254(1964)
6. A. N. Kolmogorov, Three approaches to the quantitative definition of information, *Probl. Inform.Transmis*, 1:4-7(1965).
7. C. S. Wallace and D. M. Boulton, An information measure for classification, *Comp. J.*, 11:185-195(1968).
8. J. Rissanen, Modeling by shortest data description, *Automatica.* 14, 465-471(1978).
9. H. Akaike and H. Hirotugu, A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 19(6), 716-723(1974).
10. M. Hasegawa, H. Kishino and T.Yano, Dating of the Human-ape splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol Evol.* 22, 160-174(1985).
11. M. Hasegawa, H. Kishino and T.Yano, Dating of the Human-ape splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol Evol.* 22, 160-174(1985).
12. P. Cheeseman and B. Kanefsky, Evolutionary Tree Reconstruction. *In Working Notes "Minimum message length encoding", AAAI Spring Symposium Series*, Stanford, (1992)
13. L. Allison and C.S. Wallace, The Posterior Probability Distribution of Alignments and Its Application to Parameter Estimation of Evolutionary Trees and to Optimization of Multiple Alignments. *J. Mol Evol.* 34:418-430(1994).