Inference of Molecular Phylogenetic Tree Based on Minimum Model-based Complexity Method

H.Tanaka, F.Ren, T.Okayama* and T.Gojobori*
Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-Ku, Tokyo 113, Japan
{Tanaka,ren}@mri.tmd.ac.jp
*National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411, Japan
{tokayama,tgojobor}@genes.nig.ac.jp

Abstract

In this study, starting with a newly introduced concept of data complexity ("empirical data complexity"), we specify the concept of complexity more concretely in relation to mathematical modeling and introduce "model-based complexity (MBC) ". Inductive inference based on the minimum model-based complexity method is then applied to the reconstruction of molecular evolutionary tree from DNA sequences. We find that minimum MBC method has good asymptotic property when DNA sequence lengths approach to infinite and compensates the bias of maximum likelihood method due to the difference of tree topology complexity. The efficiency of minimum MBC method for reconstruction of molecular tree is studied by computer simulation, and results suggest that this method is superior to the traditional maximum likelihood method or its modification by Akaike's AIC.

Introduction

The reconstruction of phylogenetic trees from molecular data is one of important problems in evolutionary study and many methods have been proposed so far. These methods are mainly divided into two groups: maximum likelihood methods and distance methods. Both groups of methods, however, have superiorities as well as defects.

Maximum likelihood method (Felsenstein 1981), though it is rigorously based on the probabilistic model of base substitution process along the whole phylogenetic trees, has a defect in that maximum likelihood value itself is conditional on tree topology so that it cannot, at least in principle, determine the goodness of assumed tree topology (Nei 1987, Saitou 1988). On the contrary, the distance methods such as the neighbor-joining method (Saitou & Nei 1987) which are based on the distances (number of base substitutions per site) between the homologous DNA or amino-acid sequences of any pair among the species, have several criteria such as a minimum sum of branch lengths for choosing correct topology, but they are criticized in that they use probabilistic model of base substitution only in calculation of distance between two species, and do not use it any more in the subsequent reconstruction process of the whole evolutionary tree.

With this background, we have been engaged to develop a new method which incorporates both of the superiorities of the ML methods and distance methods for these years(Ren, Tanaka and Gojobori 1995a; Ren et al. 1995b; Tanaka 1996). In our study, the phylogenetic tree reconstruction problem is considered as a kind of inductive inference to extract the minimum complexity model from observed data.

In this study we improve our previous method and give a more exact model of molecular evolution and its complexity. Specially, (1) we investigate the concept of complexity more rigorously in relation to mathematical modeling and define "empirical model-based complexity". (2) This model-based complexity is then applied to the problem of reconstruction of molecular phylogenetic tree from homologous DNA sequence of several species. (3) The efficiency of this minimum model-based complexity estimation method for reconstructing the correct phylogenetic tree is studied by computer simulation.

Model-based Complexity

Concept of minimum complexity is often referred in relation to the inductive inference. 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.

Concept of the complexity of the given data is originated by Solomonoff (Solomonoff 1964), Kolmogrov

(Kolmogorov 1965) and Chaitin (Chaitin 1966) in constructing algorithmic information theory. They measure the complexity of data by the length of program which generates the given data on universal computation machine (Turing machine). As is known, the minimum length program is almost impossible to determine (non-computable).

The subsequent studies such as by Wallace (Wallace 1968), Schwarz (Schwarz 1978) and Rissanen (Rissanen 1978) define the complexity of data in statistical framework, where they use stochastic model instead of Turing machine to measure the complexity of generated data. For example, in Rissanen's MDL principle, data complexity 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 among the assumed model family.

Our minimum complexity estimation is essentially same to the Rissanen's or other similar approach, but different from theirs in starting with defining "absolute stochastic complexity" which in principle need not refer any of statistical model family to describe the complexity of data. Then we formulate the model-based complexity which improves several commonly refered defects of MDL. Anyhow, our starting point is to formulate the definition of model-based complexity of data.

Definition 1(Model-based complexity) *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*

$$K_M(D) \equiv \inf_{M_\lambda}\{K(M_\lambda) + K(D/M_\lambda)\}.$$

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

The details of this concept must be specified in relation to model specification. We start with our definition of absolute stochastic complexity.

Newly Introduced Concept of Empirical Stochastic Complexity

The well-known definition of stochastic complexity is Shannon's one, so that we first start with this definition, that is

$$K_{sh}(x) = \sum_i p(x_i)\log p(x_i).$$

This definition has two problems:

(1)One is the assumption that we have complete knowledge of the probability density function(PDF) of the data $p(x)$. But it is almost impossible to have a complete knowledge of the data generating PDF in the real world.

(2)The second is that we have only finite number of samples or a part of data among their possible outputs, so that we can not take summation over the possible data space like in the definition of Shannon complexity.

Hence, more realistic definition of the complexity, or entropy of the data should be provided which only uses finite number of data without any assumptions of perfect knowledge about the distribution. We now introduce a new definition of empirical stochastic complexity or empirical entropy $\tilde{H}_n(\mathbf{x})$ of one dimentional real valued data of $D = \{x_1, x_2, \cdots, x_n\}$ as follows.

Definition 2(Empirical Stochastic Complexity) *Let $D = \{x_1, x_2, \cdots, x_n\}$ be n real-valued data which are supposed to be generated independently from an identical unknown probabilistic distribution, then the empirical complexity of data of n length is defined by,*

$$\tilde{H}_n(\mathbf{x}) \equiv \frac{1}{n}\sum_i^n \log\frac{\Gamma_n}{\gamma_n(x_{(i)})},$$

or, in more succint form,

$$\tilde{H}_n(\mathbf{x}) \equiv -\frac{1}{n}\sum_i^n \log\tilde{\gamma}_n(x_{(i)})$$

where

$$\tilde{\gamma}_n(x_{(i)}) = \frac{\gamma_n(x_{(i)})}{\Gamma_n}.$$

and $\{x_{(i)}\}$ is the permutation of $\{x_i\}$ in the order of increasing magnitude (order statistics). $\tilde{\gamma}_n(x_{(i)})$ is the empirical probability density approximating $\frac{\partial F_n}{\partial x}$ (F_n is empirical distribution function of n real valued data) at $x_{(i)}$where

$$\begin{aligned}
\gamma_n(x_{(1)}) &= \frac{1}{n+1}\{I(x_{(1)}, x_{(2)})\}^{-1}, \\
\gamma_n(x_{(i)}) &= \frac{1}{2(n+1)}[\{I(x_{(i-1)}, x_{(i)})\}^{-1} \\
&\quad + \{I(x_{(i)}, x_{(i+1)})\}^{-1}] \\
&\quad (i = 2, \cdots, n-1), \\
\gamma_n(x_{(n)}) &= \frac{1}{n+1}\{I(x_{(n-1)}, x_{(n)})\}^{-1}.
\end{aligned}$$

and $I(a,b) = b - a$ defines the interval length between a and b. Further,

$$\Gamma_n = \sum_{i=1}^{n-1}\gamma_n(x_i).$$

Remark In the above definition of empirical entropy, the empirical probability density $\tilde{\gamma}(x_{(i)})$ is estimated,

based on the fact that the densely scattered area of data, where the intervals $I(x_{(i-1)}, x_{(i)})$ between the neighboring sample points is small, reflects the high values of empirical probability density, so that we calculate the interval between adjacent points and inverts it to estimate the probability dense. This is original motif of the above definition.

Furthermore, following theorem about the consistency of the above definition of empirical complexity also gives a general idea about why we adopt such kind of definition.

Theorem 1(Consistency of empirical stochastic complexity) *If the empirical distribution function F_n converges to the limiting distribution function F, which is first order differentiable with respect to x, then the empirical stochastic complexity $\tilde{H}_n(\mathbf{x})$ converges to Shannon complexity when $n \to \infty$, so that we have*

$$\tilde{H}_\infty(x) = K_{sh}(x).$$

Proof. On account of the limitation of space, we here only refer to our previous paper where detailed proof is given (Theorem 1 in (Tanaka 1996)).◁

Thus defined empirical complexity can be calculated without any assumption of probabilistic structure. If the data is multidimentional, we can easily extends the above definition of empirical complexity.

Here we assume the model of the data generating mechanism **M**. In defining the model-based stochastic complexity, empirical version of the relative entropy, Kullback-Leibler divergence, must be introduced.

Definition 3 (Empirical KL information) *Let the empirical probability of the data $D = \{x_1, \cdots, x_n\}$ denotes $\tilde{\gamma}_n(x_i)$, the empirical Kullback-Leibler information of D with respect to other probabilistic function $p(x)$ is given by*

$$I^{KL}(D/p(x)) = \frac{1}{n}\sum_i \log \frac{\Gamma_n^p \tilde{\gamma}_n(x_i)}{p(x_i)}.$$

where

$$\Gamma_n^p = \sum_{i=1}^{n} p(x_i)$$

If we model stochastic data generating machinery by some probability density function $p(x)$, then the empirical stochastic complexity can be given by the sum of the empirical Kullback-Leibler information of the data with respect to the model and the complexity of the model, so that we have the following definition.

Definition 4 (Model-based stochastic complexity) *The stochastic complexity of the data $D = \{x_1, \cdots, x_n\}$, when using model* **M**, *is given*

$$K_M(D) = \inf_M \{K(\mathbf{M}) + I^{KL}(D/\mathbf{M})\},$$

where $K(\mathbf{M})$ is the complexity of the model and

$$I^{KL}(D/\mathbf{M}) = \frac{1}{n}\sum_i \log \frac{\Gamma_n^p \tilde{\gamma}_n(x_i)}{p(x_i/\mathbf{M})}.$$

Remark In relation to the general **Definition 1** in the previous section, the empirical KL complexity corresponds to the complexity of data with respect to the model. Thus, in the empirical context,

$$K(D/\mathbf{M}) \mapsto I^{KL}(D/\mathbf{M}).$$

The empirical KL information $I^{KL}(D/\mathbf{M})$ can be shown to be essentially equal to the log likelihood of model **M** with data D, $p(D|\mathbf{M})$, except for the term not relating to the model **M**. Because,

$$\begin{aligned} I^{KL}(D/\mathbf{M}) &= -\frac{1}{n}\sum_i \log p(x_i/\mathbf{M}) \\ &+ \frac{1}{n}\sum_i \log \frac{\Gamma_n^p \tilde{\gamma}_n(x_i)}{\Gamma_n}. \end{aligned}$$

Thus in our definition of the empirical complexity, $1/n$ of the negative log likelihood function is equal to empirical KL information except for the terms which do not relate to the model to be selected. This shows the naturalness of our definition of empirical entropy.

Complexity of Structured Model

In the induction of the mathematical model from data, the candidate model is not just a simple probabilistic density function but the one that has a structure with certain degree of complexity reflecting our knowledge about the entities generating those data. Hence, when we infer the best model from the data, we should determine the structure as well as the parameter values. But structure also can be represented by values of special parameters or indices. Hence, we treat two kinds of parameters: compositional parameters and inferential parameters when we determine the model. Each of the two kinds of parameters define the complexity of the model.

In this section, we consider the complexity of model in term of its parameters. We introduce the distinction and definitions of compositional and inferential complexity of parameters.

Compositional Complexity

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 such as a model lattice, we can use some index parameters $\xi(M)$ which

define the model classes. We call this kind of parameter as a **compositional parameter** of model space.

The frequent ways to introduce the measure of complexity into these 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 classes, $\log |M^{\xi}|$, as complexity measure of the elements contained in that class if the cardinality is finite. If the cardinality is infinite, we can use ε-entropy for suitably chosen ε-net introduced into the model classes.

Definition 5 (Compositional complexity of model) *If the model space has a sequence of subspaces M^{ξ} which has a strict inclusion relation, that is. $M = M^1 \supset M^2 \supset M^3 \cdots$ which defines the generalization-specification hierarchy into the model space, then complexity of the model m in the subspace M^{ξ} is given by (1) when M is discrete model space,*

$$K_c(m) = \log |M^{\xi}|,$$

(2)when M is continuous model space,

$$K_c(m) = \log N_{\varepsilon}(M^{\xi}),$$

where $N_{\varepsilon}(M^{\xi})$ is the number of elements of Kolmogorov ε-net covering M^{ξ} and ε is given by empirical precision of δm.

Inferential Complexity

Other than the compositional parameters which specifies the model class. there are ordinary parameters which are estimated from data and define a particular model element in the model class. We call this ordinary parameter as an **inferential parameter $\boldsymbol{\theta}$**. There are several approaches to describe the complexity of inferential parameters.

Well-known is Akaike's AIC (Akaike 1977), the half of which is given by

$$\frac{1}{2}AIC = -\log L(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + k,$$

where k is the number of inferential parameters which also describes its complexity and $\boldsymbol{x}$ is data and $\hat{\boldsymbol{\theta}}$ is maximum likelihood estimation of parameter values.

The other approach to inferential parameter complexity is given by Rissanen. In describing the total code length. he added code length for describing the precision of data: the approximate term of this is given by

$$K_{MDL} = -\log L(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log n$$

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

Definition 6 (Inferential complexity of the model) *Let $I^F(\boldsymbol{\theta})$ be an empirical Fisher information matrix of the probabilistic model $p(\mathbf{x}/\boldsymbol{\theta})$ which is given by*

$$I_{ij}{}^F(\boldsymbol{\theta}) = -\sum_{l=1}^{n} \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\mathbf{x}_l/\boldsymbol{\theta}),$$

then the inferential complexity of this probabilistic model is given by

$$K_{in}(\boldsymbol{\theta}) = \sup_{M}\{\frac{1}{2}\log\det I(\boldsymbol{\theta})\}.$$

The sup-operation in the above definition is empirically often impossible to execute, and besides, in real application, not all the parameter are independent so that more feasible definition is to introduce the effective dimension of the parameter space by applying eigenvector analysis to Fisher information matrix.

Definition 7 (Empirical inferential complexity of model) *The empirical inferential complexity of data is given by*

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

where $\epsilon - \dim(\boldsymbol{\theta})$ is effective number of empirically independent parameter which is given by the number of parameters whose corresponding eigen value λ_i of empirical Fisher information matrix $I(\boldsymbol{\theta})$ satisfies $\lambda_i \leq \lambda^*$, where λ^* is threshold to reflect inferential precision of the parameters.

Usually the components, the sum of the eigenvalues up to which falls within 95 % of the sum of total eigenvalues are included in the effective components. 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, $I^{KL}(\mathbf{x}/\boldsymbol{\xi}, \boldsymbol{\theta})$.

Definition 8 (Model-based data complexity represented by parameters) *Let $\boldsymbol{\xi}$ be the compositional parameters and $\boldsymbol{\theta}$ be the inferential parameters. then model-based complexity of data is given by*

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

Thus, starting from **Definition 1**, we have reached the final concrete form of the general model-based complexity definition that consists of three different kinds of complexities. We can use this model-based complexity to extract the model from data by finding the model which minimize the model-based complexity(MBC).

Reconstruction of Molecular Phylogenetic Tree

Hereafter we apply the minimum model-based complexity method to the reconstruction of molecular phylogenetic tree. In the evolutionary phylogenetic tree, the model space M_T is decomposed into (1)**tree model** (tree topology T_p and branch lengths $\mathbf{t}$), and (2)**evolution model** (base substitution probability between two of four bases in DNA during time t along the tree).

Complexity of Tree Model

The class of the tree topology is determined by topological parameters: the number of leaves e and that of internal nodes v. Following the graph theory, the number of branches b is related with e and v as

$$e + v = b + 1.$$

The phylogenetic tree is a tree with one root and e leaves, where only v or equivalently b can be varied. We take v as defining class of tree topology space: the compositional parameter of the tree. If v equals 1, we have a star-shaped tree. According to the increase of the v, tree becomes more complex. In this case $v = e - 1$, we have fully expanded binary tree. The complexity of natural number v is given by $\log^* v$. Even if the number of internal nodes v is determined, the tree is not unique. Rissanen (Rissanen 1989) shows approximation of possible number of tree topology of the class defined by v is $\binom{e+v-2}{v}$. If we think each of these trees is equally probable, the resultant complexity of the tree topology is given by

$$K_c(v) = \log^* v + \log \binom{e+v-2}{v}.$$

The branch lengths of the tree are considered as parameters to be estimated in the reconstruction of phylogenetic tree and can be consider as inferential parameters of the tree. Branch lengths are also not arbitrary. If we describe branch lengths in time, for example using Myr as an unit, then each sum of the branch lengths along the pass from the root(common ancestor) to one of the leaves (current species) should be equal, so that in the given topology independent components of branch length is $b' = b - e$. Thus, the complexity of branch lengths is given by

$$K_{in}(\mathbf{t}) = \frac{\epsilon - \dim(\mathbf{t}')}{2} \log n,$$

where $\mathbf{t}' = (t1, ...t_{b'})$ is the independent component of branch lengths, n is the length of nucleotide sequence, and $\epsilon - \dim(\mathbf{t})'$ is the number of efficient eigen vectors of Fisher's information matrix $I^F(\mathbf{t}')$.

Complexity of Evolution Model

After the structure of tree is fixed, we need the model for evolution mechanism along the tree from the root to each leave. Elemental base substitution probability with time t, denoted by $\{P_{ij}(t)\}$ where i, j is one of the four nucleotides, must be first modeled. We assume $P_{ij}(t)$ follows Markov process. According to the theory of Markov process, the probability (P_{ij}) that a base which is initially in state i changes to state j after time(t) have elapsed is given by

$$\{P_{ij}(t)\} = exp(\mathbf{Rt}),$$

where i and j represent the one of four bases A,C,G and T. $\mathbf{R}$ is a rate matrix describing the number of bases have changed in a small interval of time of length dt. This rate $\mathbf{R}$ is given by Hasegawa (Hasegawa 1985) as follows,

$$\begin{bmatrix} -\beta\Pi_Y - \pi_G\alpha & \pi_C\beta & \pi_T\beta & \pi_G\alpha \\ \pi_A\beta & -\beta\Pi_R - \pi_T\alpha & \pi_T\beta & \pi_G\beta \\ \pi_A\beta & \pi_C\alpha & -\beta\Pi_R - \pi_C\alpha & \pi_G\beta \\ \pi_A\alpha & \pi_C\beta & \pi_T\beta & -\beta\Pi_Y - \pi_A\alpha \end{bmatrix}$$

where α denotes the transition and β denotes the transversion rate. If we think the rates are varied among branches, we denote them by rate vectors $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_b), \boldsymbol{\beta} = (\beta_1, ...\beta_b)$ which are thought to be the inferential parameters. π_A, π_C, π_T and π_G represent the overall equilibrium frequencies of base A, C, T and G, respectively and $\Pi_Y = \pi_T + \pi_C$ and $\Pi_R = \pi_A + \pi_G$.

This Markov matrix $\mathbf{R}$ satisfies two needs for calculating the probability of base substitution: one is that the transition and transversion rate can be distinguished, and the other is that the frequency of A, C, T and G approaches their equilibrium values when $t = \infty$. In constructing the probability of the observed sequences given the model from the elemental base substitution matrix, we follows Felsenstein's method. We assume that the evolution is independent at different sites in the nucleotide sequences, so that the probability of a given set of data can be computed site by site.

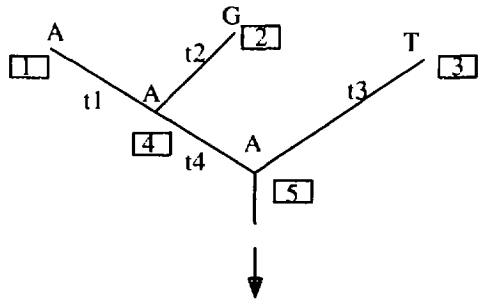

Figure 1: Two model trees used for computer simulation

To illustrate how the probability of current sequences is constructed, we take a particular case, which is shown in Fig.1. In Fig.1, 1.2 and 3 denote the current DNA sequence data which are assumed to be A, G and T respectively, and 4 and 5 are internal nodes. If we assume that node 4 and 5 are base A, the likelihood of the branch t_1(the segment from node 4 to node 5) is $P_{AA}(t)$. From node 4 up to the corresponding node 1 and node 2, the probability of this part will be computed by multiplying the transition probability during t_1 and t_2. We compute the probability along the whole tree in the same manner as the above, and finally obtain the probability of whole tree as to one site of sequences, denoted by:

$$P^{onesite}(leave = A, G, T/root - A)$$

$$= \pi_A P_{AT}^{(t_3)} P_{AA}^{(t_4)} [P_{AA}^{(t_1)} P_{AG}^{(t_2)}].$$

In fact, as the bases at internal node 4 and 5 are unknown, we take the sum of probabilities of the four cases and the probability will be rewritten by:

$$P^{onesite}(A, G, T) = \sum_{s_5=1}^{4} \pi_s P_{s_1 T}^{(t_3)} \sum_{s_4=1}^{4} P_{s_5 s_4}^{(t_4)} [P_{s_4 A}^{t_1} P_{s_4 G}^{(t_2)}],$$

where s_4 and s_5 mean the base states of internal node 4 and 5.

The overall probability for current sequences is a product of the probability observed base of all sites under consideration.

Taking the sum of the above formulation of the complexity of the each component of a phylogenetic tree, we have the total complexity to be minimized.

$$\begin{aligned} K_{M_T}(S_1, \cdots, S_e) &= \min_{v,\mathbf{t}'} \{K_e(v) + K_{in}(\mathbf{t}') \\ &+ I^{KL}(D/v, \mathbf{t}')\} \\ &- \min_{v,\mathbf{t}'} \{ \log L(S_1, \cdots, S_e/M_T) \\ &+ [\log^* v + \log \binom{e+v-2}{v}] \\ &+ \frac{e - \dim(\mathbf{t}')}{2} \log n \}. \end{aligned}$$

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

Computer Simulation

We have used the preliminary version of MBC method to reconstruct the phylogenetic trees of Primate and Mammalian with real DNA sequence data (Ren, Tanaka, & Gojobori 1995). Since the exact evolutionary pathways of the extant species are usually unknown, it is not proper to examine the efficiency of a tree-making method with real molecular data. Hence, in this study we employ a computer simulation to examine the efficiency of the MBC method. The efficiency of MBC method for selecting the correct tree topology is compared with those of ML method and AIC method.

Method and Model

The method of the computer simulation used is as follows: First, two topologies, a bifurcate tree and a trifurcate tree were prepared as the model trees (shown in fig.2). The both model trees consist of four OTUs(species). Second, the ancestral sequence of 1008-bp nucleotides was generated by Monte Carlo simulation. This sequence was assumed to evolve along the topologies of the predetermined model tree to produce the sequences of four OTUs at the leaves. In this simulation, we assume that the nucleotide substitution follows a Markov process and the rate of the nucleotide substitution is allowed to be varied from one lineage to the other lineage. Third, the phylogenetic trees were estimated from the generated sequences of four OTUs

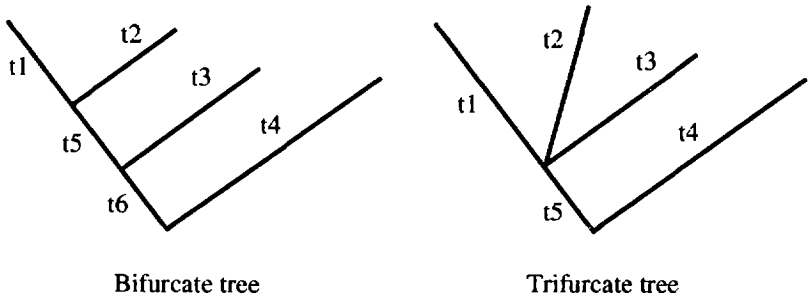

Figure 2: Two model trees used for computer simulation

by MBC method, AIC method and ML method. This process of the simulation of tree-making and reconstruction by 3 methods were repeated 1000 times. We develop the programs for reconstructing trees based on MBC method, AIC method and ML method. In order to compute optimal parameter values which attain the maximum likelihood value or minimum complexity, our programs employ the Downhill Simplex method which is developed by Nelder and Mead (Nelder & Mead 1965). For computing the maximum likelihood method, we developed our own program. Based on this program, we developed the programs for computing AIC method and MBC method in which the term of model complexity is added to ML program.

Our ML program differs from Felsenstein's PHYLIP in several points: First, our ML programs use Hasegawa's model to calculate the nucleotide substitution, and the transition(α) and transversion(β) rate can be estimated from the data by the programs if they are not specified by users. Second, our programs allow the multifurcate tree as the candidates of true tree. Third, in our programs, the evolutionary times can be estimated by incorporating the divergence date of outgroup species(whose divergence time is known). In this study, the sequence d is taken as an outgroup data and its divergence time was assumed to be 200 *Myr* (million years) ago from the other species. The tree models and parameter values in the simulation of this study are taken as examples from the partial mammalian evolution(Human, Bov and mouse).

In this simulation, we only used 4 OTUs to reconstruct the phylogenetic trees considering the calculation time. In this case, if the base substitution rates(α, β) are constant, the difference of parameter number between bifurcate tree and trifurcate tree is very slight (2 and 1). Therefore, the varied base substitution rates are assumed. In addition, in order to examine the significant correlation between the parameter number and each method, the phylogenetic trees are estimated in three different conditions in which the length of sequence n is (1)1008-bp, (2)2016-bp and (3) 3024-bp.

Results

The results under these three conditions are shown from Table 1 to Table 3. Abbreviation is used in the Tables. As for the true model, the "bi-model" means the bifurcate model tree was used to *generate* the data for simulation, and "tri-model" means the trifurcate was used to generate the data for simulation. As for the estimated "bi-tree" and "tri-tree", they mean the number of bifurcate tree or trifurcate tree which is *estimated* as true tree among 1000 data.

Table 1 shows the results with 1008-bp. In the case that the bifurcate tree model is true, ML method reconstructs the correct tree 976 times when simulation is repeated 1000 times and it shows highest performance compared with other methods. The AIC method is a little inferior (953/1000) to the ML method, but it is superior to MBC method(778/1000). On the contrary, in the case when the trifurcate tree model is true, the lowest efficiency in reconstructing the correct tree are observed(598/1000) in the ML method, whereas the MBC method shows very high performance(971/1000). The AIC method still show an intermediate efficiency(771/1000) compared with ML method and MBC method as well as in the case when the bifurcate tree model is true one.

Table 2 shows the results with 2016-bp. Comparing the results of Table 1, all the methods show increase of accuracy. This is because these method are derived from asymptotic considerations. But the increasing rate of the accuracy of estimation is slow both in ML and AIC method, whereas MBC method shows the rapid increase of the accuracy both in the cases of the bifurcate and trifurcate model being used.

Table 3 shows the results with 3024-bp. In the case that the bifurcate model is true, ML method and

Table 1: Simulation results with n=1008-bp

MODEL	ESTIMATED TREE TOPOLOGY					
	ML		AIC		MBC	
	bi-tree	tri-tree	bi-tree	tri-tree	bi-tree	tri-tree
bi- model	976	24	953	47	778	222
tri- model	402	598	229	771	29	971

Table 2: Simulation results with n=2016-bp

MODEL	ESTIMATED TREE TOPOLOGY					
	ML		AIC		MBC	
	bi-tree	tri-tree	bi-tree	tri-tree	bi-tree	tri-tree
bi- model	998	2	993	7	949	51
tri- model	255	745	157	843	15	985

AIC method do not show the increase of accuracy any more(ML: 996/1000, AIC: 989/1000), whereas MBC method still show increase of accuracy(984/1000). In the case that the trifurcate model is true, though all methods show the increase of accuracy, but, as it is same with Table 2, MBC method shows rapid increase of accuracy than ML method and AIC method. Thus we may conclude that the model-based complexity method shows a good asymptotic nature. On the other hand, ML and AIC seem not to have a good asymptotic property and the biases by the difference of parameter number is not compensated so rapid as in MBC method.

Discussion

ML Method Tends to Choose More Complex Model

The MBC method not only has a term for the likelihood of the tree, but also has some terms for estimating the tree complexity, so that it tends to obtain a minimum complexity tree. If we only consider the case of fully expanded binary tree, complexity terms do not effect too much. Because, in this case,the candidate trees are same with each other in the number of parameters of the tree model (such as the number of nodes and branches) even though they have different tree topologies. But, in the case that the tree has a multifurcation, the ML method tends to select an extra complexity tree (fully expanded binary tree) as a true one as shown in Table 1 where ML method estimates binary tree (402/1000) when the trifurcate tree is correct. In the case of multifurcate tree, the number of parameters of the tree model are different, but simple ML method does not involve the correction term to prevent overfitting by extra complexity tree.

On the other hand, MBC method shows relatively poor results when bifurcate tree, which is more complex tree, is correct (778/1000), but not so much poor comparing with the estimation of ML method in the case that the simple tree is correct. For long DNA sequences, MBC method provides good results both in the cases of the bifurcate and trifurcate tree being correct. Exactly, there are tendencies that AIC overestimates and MBC underestimates the number of parameters, but the bias will be compensated much faster in MBC than in AIC. So, which is the recommended method? We think MBC is more appropriate than AIC for the molecular phylogenetic analysis. This is because, in the recent years, the genetic data is rapidly accumulated and it becomes more common to use 10,000 or longer bases of DNA sequences for molecular phylogenetic analysis, so that MBC which have good convergence property is considered as the best method.

Relevancy of Multifurcate Tree

Questions may be raised about whether multifurcate evolutionary trees are biologically relevant. If we rigorously think, there may not exist strictly simultaneous divergence of several species. But molecular phylogenetic tree with multifurcation does not insist on such strictly simultaneous multiple branching of species. Instead, it insists that, with current information available from DNA sequences, we cannot conclude which branch of the multifurcate node separates earlier. If we select one of these by any unreasonable manner, we probably choose a erroneous tree which brings also biologically wrong conclusion. Hence, the multifurcate evolutionary tree, understood in the above sense, may sometimes be the only tree that is methodologically and biologically inerrant.

Table 3: Simulation results with n=3024-bp

MODEL	ESTIMATED TREE TOPOLOGY					
	ML		AIC		MBC	
	bi-tree	tri-tree	bi-tree	tri-tree	bi-tree	tri-tree
bi- model	996	4	989	11	984	16
tri- model	133	867	82	918	2	998

Other Studies of Evolutionary Tree Based on Minimum Complexity Principle

Alternative applications of minimum complexity principle method to estimate the phylogenetic tree have been studied by Cheeseman (Cheeseman & Kanefsky 1992) and Milosavljević (Milosavljević & Jurka 1993). Although both studies are unique in using MDL method to molecular phylogenetic problem, they are different from ours in definition of evolutionary complexity in that their reconstruction method is primarily based on the conventional parsimony method (Cavalli-Sforza & Edwards 1967). In the parsimony method, minimization procedure is taken with base-by-base comparison among the sequences of nodes,so that the global mathematical nature is not so clear. Moreover, several comparative studies about the conventional molecular tree reconstruction methods show that parsimony method is not so effective as other methods such as maximum likelihood method and neighbor-joining method (Tateno, Nei and Tajima 1982). Hence, we could think that MBC-based tree reconstruction method might be more proper to be based on the maximum likelihood method and to improve it.

Conclusion

In this study, the concept of complexity in inductive inference is investigated more closely in relation to mathematical modeling and model-based complexity. Then we apply this concept to develop a new method to extract the minimum complexity phylogenetic tree from homologous DNA sequences of different species. This method describes the molecular phylogenetic tree by three terms, which are related to tree topology, the estimated parameters and fitness between the model and data measured by likelihood function. The resultant method has the good asymptotic properties and compensate the bias of the maximum likelihood method when model has a structural variability. The computer simulation is used for investigation of the efficiency of this method. The results suggest that this method is superior to the traditional method because it avoids excess-complexity of the tree model in relation to the amount of the information available from DNA sequences of current species. [1]

Acknowledgments

This research is partly supported by the National Institute of Genetics under a grant for collaboration studies.

References

Akaike, H. 1977. On Entropy Maximization Principle. in Application of Statistics(P. R. Krishnaiah, ed.): 27-41. Amsterdam: North-Holland.

Cavalli-Sforza, LL. and Edwards, AWF. 1967. Phylogenetic analysis:models and estimation procedures. *Am.J.Hum.Gen.*19: 233-257.

Chaitin, G. J. 1966. On the lengths of programs for computing finite binary sequences. *J. Ass. Comput. Mach.*13: 545-569.

Cheeseman, P. and Kanefsky, B. 1992. Evolutionary Tree Reconstruction. In Working Notes "Minimum message length encoding", AAAI Spring Symposium Series, Stanford.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.*17: 368-376.

Hasegawa, M., Kishino, H. and Yano, T. 1985. Dating of the Human-ape splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol. Evol.*22: 160-174.

Kolmogorov, A. N. 1965. Three approaches to the quantitative definition of information. *Probl. Inform.Transmis*1: 4-7.

Milosacljević, A. and Jurka, J. 1993. A Case Study in Molecular Evolution. *Machine Learing*12: 69-87.

Nei, M., 1987. Molecular evolutionary genetics. *Columbia University press, New York.*

Nelder, J. A. and Mead, R. 1965. A Simplex Method for Function Minimization. *The Computer Journal*7: 308-313.

Ren, F., Tanaka, H., Fukuda, N. and Gojobori,

T. 1995a. Molecular Evolutionary Phylogenetic Tree Based on Minimum Description Length Principle, In Proceedings of the 28th Hawaii International Conference on System Sciences, 165-173.

Ren, F., Tanaka, H. and Gojobori, T. 1995b. Construction of Molecular Evolutionary Phylogenetic Tree from DNA Sequences Based on Minimum Complexity Principle. *Computer Methods and Programs in Biomedicine*.46: 121-130.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica*14: 465-471.

Rissanen, J. 1989. Stochastic Complexity in Statistical Inquiry. *World Scientific.*

Saitou, N., 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny, *J. Mol. Evol.*27: 261-273.

Saitou, N., and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees, *J. Mol. Biol. Evol.*4(4): 406-425.

Schwarz, G. 1978. Estimating the Dimension of a Model. it Ann.Statist.6: 461-464.

Solomonoff, R. J. 1964. A Formal Theory of Inductive Inference. *Part I, Information and Control* 7: 1-22; *Part II, Information and Control* 7: 224-254.

Tanaka, H. 1996. Model-based Complexity and Inductive Inference, In Proceedings of the 4th International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, 144-152.

Tateno, Y., Nei, M. and Tajima, F. 1982 Accuracy of Estimated Phylogenetic Trees from Molecular Data. *J. Mol. Evol.*18: 387-404.

Wallace, C.S. and Boulton. D.M. 1968. An Information Measure for Classification. *Computer J.*11: 185-194.