*Proceedings of the 28th Annual Hawaii International Conference on System Sciences — 1995*

# Molecular Evolutionary Phylogenetic Trees Based on Minimum Description Length Principle

Fengrong Ren, Hiroshi Tanaka, Norio Fukuda and Takashi Gojobori*

Tokyo Medical and Dental University, Yushima 1-5-45,Bunkyo 113, Japan
*National Institute of Genetics, 1,111 Yata Mishima-city Shizuoka 411,Japan

## Abstract

*Ever since the discovery of a molecular clock, many methods have been developed to reconstruct the molecular evolutionary phylogenetic trees. In this paper, we deal with this problem from the viewpoint of an inductive inference and apply Rissanen's minimum description length principle to extract the minimum complexity phylogenetic tree. Our method describes the complexity of molecular phylogenetic tree by three terms which are related to the tree topology, the sum of branch lengths and the difference between the model and the data measured by logarithmic likelihood. Five mitochondrial DNA sequences from human, common chimpanzee, pygmy chimpanzee, gorilla and orangutan are used for investigating the validity of this method. It is suggested that this method is superior to the traditional method in that it still show the good accuracy even near the root of phylogenetic trees.*

## 1 Introduction

Ever since the discovery of a molecular clock (constancy of molecular evolutionary rate), many methods have been proposed to reconstruct molecular evolutionary phylogenetic trees from the homologous nucleotide sequences of different species. Among these, a maximum likelihood method for inferring evolutionary trees was developed by Felsenstein [5]. In Felsenstein's method, possible topologies of evolutionary trees are first assumed, then the maximum likelihood value for each of these trees to give rise to current difference of homologous nucleotide sequences is calculated, based on the probablistic model of the process of base substitution. The tree topology attaining the highest in its maximum likelihood value is considered as the most likely candidate for the true one.

Although Felsenstein's method is now thought as the most rigorous estimation method for the phylogenetic tree, some have pointed out several problems about this method. For example, Nei [12] and Saitou [19] stated that the likelihood function itself varies its form if the tree topology is different, so that its maximum value is conditional on the tree topology and cannot be compared directly over different tree topologies. This critique implies that likelihood value (fitness of the model to the data) alone is not sufficient for the estimation of phylogenetic tree, but some kind of information about the tree structure (model complexity) must be included in the estimation method.

Along this line, Kishino and Hasegawa [7] proposed to use the minimum Akaike information criterion estimation (MAICE)[1] to solve this problem. In their method, the biases of maximum likelihood values due to the difference of tree topology are compensated by taking into account the difference of the number of parameters contained in the tree model. But, in our view, to include the difference of numbers of the parameters is still insufficient, because, for example, quite different tree topologies can be considered even if the number of the parameters such as the number of branches is same.

In this paper, in order to deal with this problem from more general viewpoint, the phylogenetic tree reconstruction problem is considered as a kind of inductive inference to extract the minimum complexity theory from observed data. Inspired by algorithmic information theory of Kolmogorov [9] and Chaitin [3], Rissanen [17] proposed an inductive inference method based on the concept of the minimum description length (MDL) principle. Rissanen's MDL principle are in some sense an extension of the maximum likelihood method, so as to incorporate qualitative and structural information of the model. Hence, it would be expected that, by MDL principle, we can incorporate structural information in the selection of phylogenetic trees better than MAICE which only con-

siders the number of parameters, such as that of the branches, as the tree complexity.

In the following, we briefly describe MDL principle in inductive inference, followed by the application of MDL principle to estimation of the molecular phylogenetic tree and the results are reported when our method is applied to mitochondrial DNA sequence data of primates to evaluate branching order among Hominoidea. Finally the validity of concept of the minimum complexity in phylogenetic tree problem is discussed.

## 2 Definition of complexity in inductive inference

In inductive inference, there would be many theories (models) which can explain the observed data, so that we need a criterion to select the most suitable one. One of the often used criterion is the principle of parsimony in which the theory that has the smallest complexity and, nevertheless, can explain the observation well should be chosen as the first candidate for true one. In this context, complexity of theory or data must be defined. Along this line, Kolmogorov and Chaitin proposed the algorithmic information theory in which the complexity of data is defined as the shortest length of program which can generate this data. But, to find the minimum length of the program is NP complete and this definition of the complexity is not ready to apply.

Rissanen [16] proposed more practical definition of the complexity in 1978, in which models are restricted within a certain family, and the complexity of data is expressed by a sum of the coding length of model and probabilistic residuals that can not be explained by any of this model family. By using this definition of complexity("stochastic complexity"), Rissanen proposes an inductive inference principle for inferring probabilistic models from data. In this principle, the model (theory) which has the minimum description length (MDL) is chosen as the most suitable one. Namely, the MDL principle requests a theory $T$ that minimizes

$$\ell(T) + \ell(x_1, \cdots, x_n \mid T) \cdots (1),$$

where $\ell(T)$ is the length in bits of a machine-readable representation of $T$, $x_1, \cdots, x_n$ are the observations, and $\ell(x_1, \cdots, x_n \mid T)$ is the number of bits needed to encode the observations with respect to $T$. Pendnault [14] showed that the strength of the MDL principle lies in its convergency properties, where it enables the theories (models) to eventually converge on an appropriate one given a sufficient number of observations.

## 3 Relation between MDL and maximum likelihood

In MDL, the residuals from theory, the second term in MDL formula(1), is coded using their estimated probability density function, which gives rise to the close relation to maximum likelihood method.

According to the Shannon's coding theorem of information source, if the coding length of signal $x$ (in this case residuals from the theory) is assigned to $-\log p(x)$ with $p(x)$ true probability density function of $x$, we can describe the information source with minimum average coding length. However, since the true density is usually unknown, $p(x)$ must be inferred from data. In this case we use a family of density function estimate $f(x \mid \theta)$, where $\theta$ is a set of parameters to be adjusted so as to approximate the true probability density function.

If the coding length is assigned to $-\log f(x \mid \theta)$, average coding length of the source is given by

$$-\int p \log f dx = -\int p \log p dx + \int p \log \frac{p}{f} dx \cdots (2).$$

The first term in right-hand side of this formula is the entropy $H(p)$ of the source (residuals) and is not related to choice of distribution model. If residuals from the theory are truly null, this term vanishes. The second term is Kullback's information quantity or divergence $D(p \parallel f)$, which indicates the distance between two probability densities. Therefore the average increase of coding length encoded with $f(x \mid \theta)$, not with the true density $p(x)$, is $D(p \parallel f)$. Because $H(p)$ is not related with model selection, we can obtain the shortest coding length if we minimize $D(p \parallel f)$. Although $-\int p \log f dx$ is not an observable value, but its sample mean $-\frac{1}{n}\Sigma_{j=1}^{n} \log f(x_j \mid \theta)$ is observable where $n$ denotes number of samples. If each sample is independent statistically, it is equal to $-\frac{1}{n}$ of the logarithmic likelihood of the $n$-sampled data $\Sigma_{j=1}^{n} \log f(x_j \mid \theta)$. Therefore, with multiplication by $n$, the negative value of logarithmic likelihood equals to a (true) coding length of $n$-sampled residuals from the theory(entropy) plus its increase due to our ignorance of true density probability function of residuals, so that it can be used as an approximation of the coding length of the $n$ sequenced residuals from the model. This also, reversely, implies that the maximum likelihood estimation can be considered as minimum



Proceedings of the 28th Hawaii International Conference on System Sciences (HICSS '95)
1060-3425/95 $10.00 © 1995 IEEE

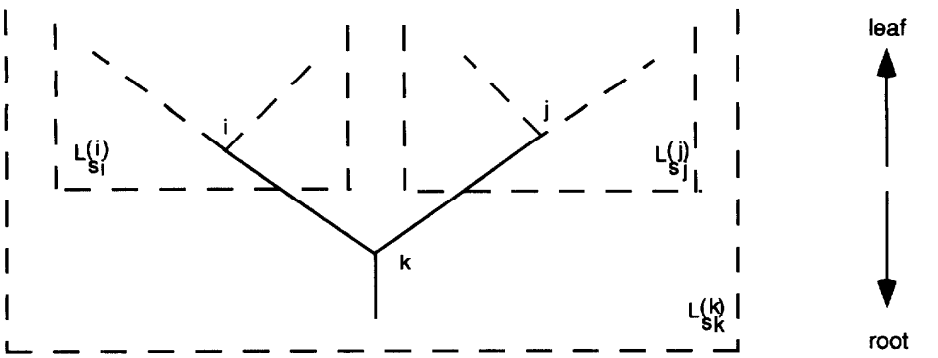


**Figure 1. The tree model used in calculating the likelihood.**

Kullback information estimation or a part of MDL estimation without considering the model complexity.

## 4 A phylogenetic tree based on MDL principle

### 4.1 The MDL formulation

Rissanen gives the following general formula consisting of three terms for total description lengths of the probabilistic data,

$$L(x \mid \theta) = \sum_{j=1}^{n} -\log f(x_j \mid \theta) + L_1(r) + \frac{1}{2}\sum_{i=1}^{m} \log[\theta_i^2 / I^{ii}(\theta)] \cdots (3).$$

where $x = (x_1, \cdots, x_n)$ is the data sequence and $\theta = (\theta_1, \cdots, \theta_m)$ is the parameters to be estimated. Among these, the first is the logarithmic likelihood function, and as stated in the above section, we can consider it to be the coding length of the deviation between theory and data. The second is a term concerned with the model complexity which is related to structural properties of model described by $r$. The third term is the coding length when we describes the values of the parameters with the precision level of their estimation error. This term also can be considered as a quantity describing the complexity of parameter values. The variances of the maximum likelihood estimators are asymptotically given by diagonal elements of inverse of Fisher's information matrix. Fisher's information matrix $I(\theta)$ is given by

$$I_{ij}(\theta) = -E_x[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\ f(x \mid \theta)] \cdots (4),$$

which can be approximated by

$$I_{ij}(\theta) \simeq -\sum_{k=1}^{n} \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\ f(x_k \mid \theta) \cdots (5),$$

where $I^{ii}(\theta)$ is the $ith$ diagonal element of the inverse matrix of $I_{ij}(\theta)$.

### 4.2 The coding length of phylogenetic tree complexity

The complexity of the tree based on the MDL principle is described by two terms like in the general formula(3). One is complexity due to the topology of the tree and the other is that due to the length of tree branches.

The code length to describe the tree topology is given by number of bits needed to specify one topology among possible ones. In the case of encoding a binary tree, since a tree with $u$ leaves has $u-1$ internal nodes, the number of possible topologies of the binary tree $B_u$ would be [2]

$$B_u = u^{-1}\binom{2u-2}{u-1} \sim 4^{u-1}(u-1)^{-1}(\pi(u-1))^{-\frac{1}{2}} \cdots (6).$$

Hence, the coding length to specify one topology among them would be

$$L_1 = 2(u-1)H(\frac{1}{2}) - \frac{1}{2}\log(\pi(u-1)) \cdots (7),$$



Proceedings of the 28th Hawaii International Conference on System Sciences (HICSS '95)
1060-3425/95 $10.00 © 1995 IEEE

where $H(\cdot)$ is an entropy function. This term is only related to the number of leaves $u$, so that it dose not contribute in the selection of binary phylogenetic trees where the number of leaves is fixed by that of current species about which the phylogenetic tree is to be constructed. In the case where it is not restricted to binary trees, Quinlan and Rivest [15] found that the number of possible trees is given by,

$$L_1 = \log(\frac{w+1}{2}) + \log \binom{w}{v} \cdots (8),$$

where $w$ is the number of total nodes and $v$ is the number of internal nodes.

The branching lengths of the tree are parameters to be estimated in the tree reconstruction problem. They can be also considered as quantities describing the tree complexity. As a natural unit of the branching length is its estimation error, following term is added to the complexity of the tree in the definition of coding length.

$$L_2 = \frac{1}{2}\sum_{i=1}^{m} \log\{t_i^2 / I^{ii}(t)\} \cdots (9),$$

where $t_i$ denotes the $ith$ branching length. If there are some other parameters, they also can be included in $t$.

### 4.3 The coding length of deviation between tree model and data

The maximum likelihood method requires an explicit probabilistic model of base substitution (mutation) during process of evolution to calculate the likelihood function. Here we construct the likelihood function of the phylogenetic tree following Felsenstein's method. The Markov process is used because the probability of base substitution depends on its current state, but not on its past history. We describe the probability of substitution from one base (base $a$ ) to another (base $b$ ) during an interval of time $t_i$ of the $ith$ branch as follows, where $a$ and $b$ denote one of the four nucleic acids i.e. A, C, T, G,

$$P_{ab}{}^{(t_i)} = exp(-k_i t_i)\delta_{ab} + (1 - exp(-k_i t_i))\pi_b \cdots (10),$$

where parameter $k_i$ related to the occurring rate of substitution during the $ith$ branch of the tree where the substitution (base a to base b) takes place, and $\delta_{ab}$ is 0 if $a$ is unequal to $b$, and 1 if $a$ is equal to $b$(the Kronecker delta function). We distinguish the transversion (a mutation in which a purine is replaced by a pyrimidine or vice verse) from transition(a mutation in which one pyrimidine is substituted by the other pyrimidine or in which one purine is substituted for the other purine), because their substitution rates are much different. We can use $\alpha_i$ for transversion and $\beta_i$ for transition instead of $k_i$. In original Felsenstein's method, $\alpha_i$ and $\beta_i$ are not explicitly used. In equation(10), in the case of $a = b$, the first term together with the second term, give the probability in which base remains same. In the case of $a \neq b$, the second term gives the probability for base substitution and $\pi_b$ is a priori probability of nucleic acid $b$. We assumed the each site of nucleotide sequences is independently changed by point mutation, so that the total likelihood of molecular evolution is the product of likelihoods of evolution of each site. Thus, we begin with describing the likelihood of evolution of one site.

To illustrate how we construct the likelihood function of the total phylogenetic tree, we take a part of the tree shown in Figure 1. For node $k$, whose immediate descendants are $i$ and $j$, we define $L_{s_k}^{(k)}$ as the likelihood for the evolution along a part of the tree stemming from node $k$, branching through $i$ and $j$, reaching the corresponding leaves of total tree, that is a subtree from node $k$ up to corresponding leaves of total tree. We assume the nucleic base of node $k$ is $s_k$. By using $P_{ab}$ we can compute $L_{s_k}^{(k)}$ recursively for all four(A,C,G,T) values of $s_k$ as follows,

$$L_{s_k}^{(k)} = (\sum_{s_i=1}^{4} P_{s_k s_i}^{(t_i)} L_{s_i}^{(i)})(\sum_{s_j=1}^{4} P_{s_k s_j}^{(t_j)} L_{s_j}^{(j)}) \cdots (11),$$

where $L_{s_i}^{(i)}$ is the partial likelihood of the evolution from node $i$ to the corresponding leaves, so with $L_{s_j}^{(j)}$ from node $j$. The first parenthesis in the right-hand side of equation(11) is the likelihood for the evolution from node $k$ to node $i$ ($P_{s_k s_i}{}^{(t_i)}$) and from node $i$ to corresponding leaves ($L_{s_i}^{(i)}$), and the second term is similarly that from node $k$ to node $j$ and node $j$ to corresponding leaves $L_{s_j}^{(j)}$. The full likelihood of the evolution for one site in nucleotide sequence along the considered tree topology is obtained by incremental repetition of this process from top leaves to the bottom fork (root) of the tree. When we eventually reach the bottom fork through this process, we obtain the final form of iteration formula given by

$$L^{one\ site} = \sum_{s_0}^{4} \pi_{s_0} L_{s_0}{}^{(0)} \cdots (12),$$

where we denote root by 0 and $\pi_{s_0}$ is the prior probability of each of four nucleic bases at root 0.



Proceedings of the 28th Hawaii International Conference on System Sciences (HICSS '95)
1060-3425/95 $10.00 © 1995 IEEE

$L^{one\ site}$ varies along the site location $(j)$ in the sequences and, as a likelihood function, it depends on the observed data, in this case, the observed $jth$ bases of current nucleotide sequences of all species, given at the leaves of tree. If we denote the $jth$ base in the sequence of $qth$ species($qth$ leaf) by $x_{jq}$, the observed $jth$ bases of all species form vector $x_j = (x_{j1}, x_{j2}, ....., x_{jg})$, where $g$ is the number of species about which the phylogenetic tree is to be constructed. In order to describe these dependence explicitly, $L^{one\ site}$ is rewritten by $L(x_j|t)$, where a set parameters to be estimated is $t = (t_1, ..., t_m, \alpha, \beta)$. Total likelihood is the product of the likelihood of each site of the sequence.

### 4.4 The total description length of a phylogenetic tree

Now we obtain total formula of the coding length of a phylogenetic tree, which is given by the sum of three terms discussed above,

$$L_{total} = \sum_{j=1}^{n} -\log L(x_j \mid t) + \log(\frac{w+1}{2}) + \log \binom{w}{v} + \frac{1}{2}\sum_{i=1}^{m} \log\{t_i^2 / I^{ii}(t)\} \cdots (13),$$

where $n$ is the length of nucleotide sequence, $m$ is total number of branches. This expression states that the tree with the smallest sum of the logarithmic branching lengths is considered as the best tree if the computed trees are equal in the likelihood and the number of nodes. This criterion is considered to be some mixture of the maximum likelihood method and the minimum evolution method as will be discussed in detail.

## 5 An experiment: phylogenetic tree of Hominoidea

### 5.1 Sequence data and computer program

A preliminary investigation about the estimation accuracy of our method was undertaken by an experiment on the reconstruction of phylogenetic tree of Hominoidea. The origin of humans have been a cardinal question in the phylogeny. Although this field has been remarkably advanced since the discovery of the molecular clock, the precise branching pattern and dating in Hominoidea is still controversial. This is partially because nucleotide sequences used to construct the phyogenetic tree of Hominoidea have been not sufficiently long in the sequence or slow in base substitution rate. Horai et al [6] recently sequenced 4938-bp mitochondrion DNA of human, common chimpanzee, pygmy chimpanzee, gorilla and orangutan. This long sequenced mtDNA is sufficiently fast in base substitution rate. Hence we use this Horai's data in this experiment.

These mtDNA sequences contain the complete genes for NADH dehydrogenase subunit 2 ($ND2$), cytochrome oxidase subunit $I$ and $II$ ($COI$ and $COII$), $ATPase$ 8, portions of two genes for $ND1$ and $ATPase6$, and 11 interspersed tRNAs. In this study, since insertion/deletion events are not taken into account in ours model, the noncoding parts in which insertion/deletion is frequent are excluded. In addition, two incompletely determined codons in $ND1$ and $ATPase$ 6, and small overlapping parts are excluded. Consequently, we examined the remaining region (4762-bp) of each sequence. The remaining region is divided into two classes because they show different rates of base substitutions. The data of class 1 include the third codon position of protein coding genes (1336 sites). The data of class 2 include the first and second codon position of protein coding genes and sites in tRNA genes (3426 sites).

### 5.2 Computer program

We develop a computer program for MDL-based reconstruction method. In our method, the downhill simplex method developed by Nelder and Mead[13] is used to obtain the parameter values which attain the minimum code length.

In order to obtain actual evolutionary time, we need some reference ( "outgroup") species with assumed evolutionary time. In this study, we suppose the branching time of orangutan from African apes-human clade is 13 Myr (million years) ago. The base substitution rate is assumed to be constant among all lineages, but different between class1 and class2. The former assumption, constancy of substitution rate is not necessary to our method, but as a preliminary investigation and not to increase the number of parameter $\alpha$ and $\beta$, this assumption is made.

We also compare the MDL-based method with maximum likelihood method in estimation of the phylogenetic tree of Hominoidea. To do so,we use Felsenstein's DNAML in his program package PHYLIP. This program generates the possible topologies of phylogenetic trees and calculates the likelihood of each tree and chooses the best tree having the maximum likeli-



Proceedings of the 28th Hawaii International Conference on System Sciences (HICSS '95)
1060-3425/95 $10.00 © 1995 IEEE

**Table 1. Branching lengths and evolutionary rates from 4762-bp mtDNA sequences**

| Parameters | Branching lengths(*Myr*) and Rates($10^{-3}$/*Myr*) | | |
|---|---|---|---|
| | Horai | MDL | ML |
| t1 (Human) | 5.0 | 4.7 | 6.2 |
| t2 (Common Chimpanzee) | 2.1 | 2.0 | 2.3 |
| t3 (Pygmy chimpanzee) | 2.1 | 2.0 | 2.2 |
| t4 (Gorilla) | 8.0 | 7.0 | 7.2 |
| t5 (Orangutan) | 13.0 | 13.0 | 13.0 |
| t6 | 2.1 | 2.5 | 3.2 |
| t7 | 3.2 | 2.8 | 3.2 |
| t8 | 5.0 | 5.5 | 3.2 |
| sum of branching lengths | 40.5 | 39.5 | 40.5 |
| $\alpha$1 (transition rate of class 1) | | 125.0 | |
| $\beta$1 (transversion rate of class 1) | | 12.5 | |
| $\alpha$2 (transition rate of class 2) | | 20.0 | |
| $\beta$2 (transversion rate of class 2) | | 3.5 | |

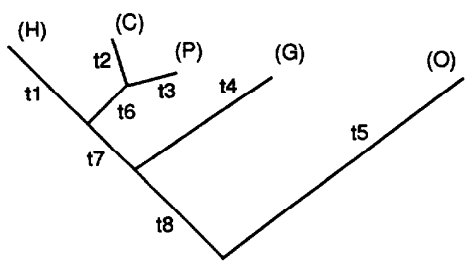


**Figure 2. MDL based tree of mtDNA(4762 sites)**

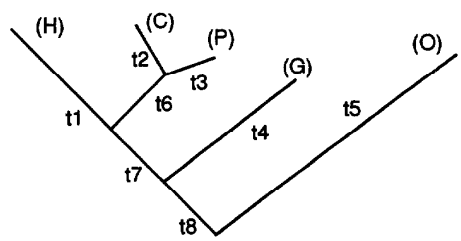


**Figure 3. ML based tree of mtDNA(4762 sites)**



Proceedings of the 28th Hawaii International Conference on System Sciences (HICSS '95)
1060-3425/95 $10.00 © 1995 IEEE

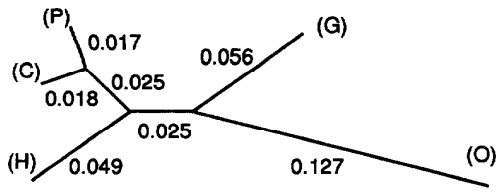


**Figure 4. Unrooted ML based tree of mtDNA(4762 sites). Numbers attached to branches are base substitution number per site.**

hood value. Used data is same as those in MDL-based method.

In Felsenstein's DNAML, the estimated tree is given in the form of unrooted trees, so that it cannot be simply compared with our method. We modified Felsenstein's method according to Hasegawa and Kishino [8], incorporating substitution rate difference and dating of branching of outgroup (in this case, orangutan) to produce the rooted tree.

### 5.3 Results

The estimated tree which attains the minimum coding length is shown in Figure 2. From Figure 2, we can see that common chimpanzee and pygmy chimpanzee have a closer relationship to human rather than to gorilla under the assumption of rate constancy among five species. The results of the branching patterns agree with that obtained by Horai which is thought to be most exact so far.

The method used in Horai's study is essentially based on the conventional method using the number of different nucleotide among the homologous sequences of species, but a good pieces of domain-specific knowledge, such as detailed division of the nucleotide sequences according to differences of base substitution rate are used to obtain final results. So in this study, in order to investigate the validity and accuracy of the MDL-based method, we use Horai's results only as a most certain criterion for true values instead of investing his method further, but we compare the MDL-based method with other more mathematically well defined method such as maximum likelihood method, assuming the detailed knowledge used in Horai's study is not known.

The results for the maximum likelihood(ML) based tree, modified to have a root, and its original unrooted tree is shown in Figure 3 and 4, respectively. It can be easily seen that the estimated tree by ML method is topologically identical to MDL-based tree, but we can point out the following differences between them.

(1) In the branching length($t_8$) between gorilla and orangutan near the root is longer in MDL-based method than in ML-method.

(2) Nevertheless, the sum of the branch length is minimum in the MDL-based tree.

## 6 Discussion

Here we will discuss about the above two aspects of the experimental results from the intrinsic property of MDL based tree estimation method.

As for the first point of results, MDL-based method agrees with the Horai's result in estimation of $t_8$. Not only Horai's result but other related studies [11] also indicate that the gorilla and orangutan belong to different clades and support the further distance between them than the results of ML-based method. Hence we could say that ML-based method might be slightly inaccurate in estimation of $t_8$. This seems to be caused by the assumption of rate constancy in all lineage, but it might not be so, because the assumption is shared by both ML-based and MDL-based method. Possible accounts for this difference would be as follows.

MDL-based tree tends to become more parsimonious, specially due to the last term which prefers shorter branching lengths. This tendency does not effect so much near the leaves, because in estimation of branching pattern near the leaves likelihood function has a tight relation with observed sequences of present species, and the first term (likelihood term) is more effective . But as the tree goes down to the root, likelihood function becomes less effective (remember that we make summation about the possible bases at unobservable internal nodes to construct the likelihood function), so that the last term preferring the minimum branching lengths becomes more effective than the first term to keep the overall accuracy of estimation. On the contrary, ML-based method does not have such a term, so that as the tree goes down to its root, the estimation becomes inaccurate. So if the principle of parsimony is valid still in molecular evolution, the MDL-based method well compensate the inaccuracy of ML. The further work remains to be done to provide more rigorous proof of this conjecture. We are now preparing this.

As stated in the second point of the results, in spite of longer distance between gorilla and orangutan, MDL-based method attains the minimum in sum of branch lengths. This means MDL-based method reconstructs the minimum complexity tree. The reason



Proceedings of the 28th Hawaii International Conference on System Sciences (HICSS '95)
1060-3425/95 $10.00 © 1995 IEEE

why the minimum complexity tree is most likely candidate for true one is rigorously proved, in different context, by Rzhetsky and Nei [18]. Although they employ "distance-based" least square method in estimating the branch lengths, they discovers that if the topology of the tree is correct, then the sum of estimated branch lengths are minimum. In other words, if our assumed topology is incorrect, the sum of the estimated branch lengths based on this incorrect tree topology becomes larger than that based on the correct topology. Hence they use the sum of branch lengths as the final criterion for the least square estimation of phylogenetic tree, and call their method " the minimum-evolution method". It can be considered that this property is also valid even in case of using maximum likelihood method to fit the model to the data, instead of using least square method. In this context, our method can be considered as an integration of maximum likelihood method and minimum-evolution method.

Alternative applications of MDL to estimate the phylogenetic tree have been studied by Cheeseman [4] and Milosavljević [10]. Although both studies are unique in using MDL to molecular phylogenetic problem, they are different from ours in the formulation of description length in that the former uses the Bayesian framework and the latter uses base-by-base substitution. So our formula is quite new in the use of MDL which combines ML and minimum evolution method.

## 7 Conclusion

In this paper, we deal with estimation of the evolutionary phylogenetic tree as a kind of inductive inference and apply Rissanen's minimum description length principle(MDL) to extract the minimum complexity phylogenetic tree from homologous nucleic sequences of different species. Our method describes the complexity of molecular phylogenetic tree by three terms which are related to tree topology, sum of branch lengths and difference between the model and data measured by logarithmic likelihood. This method can compare the goodness of tree having different topology. Five mitochondrion DNA sequences from human, common chimpanzee, pygmy chimpanzee, gorilla and orangutan are used for investigation of the validity of this method. The obtained phylogenetic tree agrees with the most reliable results which have been reported so far, and it is suggested that this method is superior to the traditional method in that it still show the good accuracy even near the root of phylogenetic trees, where likelihood function becomes inaccurate.

## Acknowledgements

We thank Dr.Horai for providing us his mtDNA sequence data in machine readable form, and Dr.Cheeseman for giving us an oppotunity to read his papers on MDL. We also thank the anonymous reviewers for giving helpful comments for us to improve this paper.

## References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle'" *Petrov BN, Csaki F(eds) Second international symposium on information theory*, Akademiai Kiado, Budapest, pp. 267-281, 1973.

[2] Cavalli-Sforz LL, Edwards AWF, "Phylogenetic analysis: models and estimation procedures," *Am J Hum Gen*, Vol. 19, pp. 233-257, 1967.

[3] G. J. Chaitin, "On the lengths of programs for computing finite binary sequences," *J. Ass. Comput. Mach*, Vol. 13, pp. 545-569, 1966.

[4] P. Cheeseman and B. Kanefsky, "Evolutionary tree reconstruction", *In Working Notes,AAAI Spring Symposium Series, Stanford.*

[5] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *J Mol Evol*, Vol. 17, pp. 368-376, 1981.

[6] S. Horai, Y. Satta, K. Hayasaka, R. Kondo, T. Inoue, T. Ishida, S. Hayashi, and N. Takahata, "Man's place in hominoidea revealed by mitochondrial DNA genealogy," *J Mol Evol*, Vol. 35, pp. 32-43, 1992.

[7] H. Kishino and M. Hasegawa, " Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea," *J Mol Evol*, Vol. 29, pp. 170-179, 1989.

[8] H. Kishino and M. Hasegawa, "Converting distance to time: application to human evolution. Methods in Enzymology," *J Mol Evol*, Vol. 183, pp. 550-570, 1990.

[9] A. N. Kolmogorov, "Three approaches to the quantitative definition of information. Probl. Inform.Transmis," *Probl. Inform.Transmis*, Vol. 1, pp. 4-7, 1965.



Proceedings of the 28th Hawaii International Conference on System Sciences (HICSS '95)
1060-3425/95 $10.00 © 1995 IEEE

[10] A. Milosavljević and J. Jurka, "Discovery by Minimal Length Encoding: A Case Study in Molecular Evolution," *Machine Learning*, Vol. 12, pp. 69-87, 1993.

[11] MM. Miyamoto, JL.Slightom and M. Goodman, "Phylogenetic relations of humans and African apes from DNA sequences in the $\psi\eta$-globin region," *Science*, Vol. 238, pp. 369-373, 1987.

[12] M. Nei, "Molecular evolutionary genetics," *Columbia University press, New York*, 1987.

[13] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, Vol. 7, pp. 308-313, 1965.

[14] E. P. D. Pendnault, " Minimal-length encoding and inductive inference," *AAAI press*, pp. 71-92, 1989.

[15] J. R. Quinlan and R. L. Rivest, "Inferring decision trees using the minimum description length principle," *Information and Computation*, Vol. 80, pp. 227-248, 1989.

[16] J. Rissanen, "Modeling by shortest data description," *Automatica*, Vol. 14, pp. 465-471, 1978.

[17] J. Rissanen, "Stochastic complexity and modeling," *Ann.Statist*, Vol. 14, pp. 1080-1100, 1986.

[18] A. Rzhetsky and M. Nei, "Theoretical foundation of the minimum-evolution method of phylogenetic inference," *Mol,Biol.Evol*, Vol. 10(5), pp. 1073-1095, 1993.

[19] N. Saitou, "Property and efficiency of the maximum likelihood method for molecular phylogeny," *J Mol Evol*, Vol. 27, pp. 261-273, 1988.