The Inference of Model-based Minimum Complexity in Reconstructing Molecular Phylogenetic Tree

FR. Ren [1], H. Tanaka [1]
{ren,tanaka}@mri.tmd.ac.jp

T. Okayama [2]
tokayama@genes.nig.ac.jp

[1] Medical Research Institute, Tokyo Medical and Dental Univ.
Yushima 1-5-45, Tokyo 113, Japan

[2] Center for Information Biology, National Institute of Genetics
1111 Yata, Mishima, Shizuoka 411, Japan

Abstract

As a novel application of MDL, the concept of model-based complexity is proposed in this study. This method have good asymptotic property and compensates the bias of maximum likelihood method. The efficiency of this method for reconstructing the correct phylogenetic tree is studied by computer simulation, and results suggest that it is superior to the traditional maximum likelihood method or its modification by Akaike's AIC.

1 Model-based Complexity and Its Application

We have been engaged to develop a new method which can integrates the superiorities of the ML method and distance method for these years[1]. We find that Rissanen's MDL(Minimum Description Length) principle is useful to be applied to this problem, because it is in some sense an extension of the maximum likelihood method and it incorporates the qualitative and structural information of the model in the evaluation of the phylogenetic tree.

In this study, we make the concept of complexity more concretely with respect to its application to induction of mathematical modeling. We call this version of the MDL-complexity as "model-based complexity" in order to distinguish from the ordinary definition of MDL. This model-based complexity of data is defined as

$$K_M(D) = \min_{\boldsymbol{\xi},\boldsymbol{\theta}} \{K(\boldsymbol{\xi}) + K(\boldsymbol{\theta}) + K(D/\boldsymbol{\xi},\boldsymbol{\theta})\}.$$

Where M and D represent the model and data respectively. $\boldsymbol{\xi}$ is the compositional parameters, and $\boldsymbol{\theta}$ is the inferential parameters.

Table 1: Simulation Results

model	estimated tree topology					
	ML		AIC		MDL	
	bi-tree	tri-tree	bi-tree	tri-tree	bi-tree	tri-tree
bi- model	996	4	989	11	983	17
tri- model	867	133	918	82	995	5

In molecular phylogenetic tree, the complexity of tree model include: 1)the complexity of tree topology which is given by $\log^* v + \log \binom{e+v-2}{v}$, where e and v represent the number of leaves and internal nodes respectively. 2)the complexity of branch length which is given by $\log[C_b \|\mathbf{t}\|^b_{I(\mathbf{t})}]$ (b = number of branches, t = branch length, C_b = unit ball volume). Where $I(\mathbf{t})$ is Fisher's information matrix which is given by $I(\mathbf{t}) = -E[\frac{\partial^2}{\partial t^2} \log L(S_1, S_2, \cdots, S_b \mid \mathbf{t})]$, Therefore, we have total complexity of the tree as follows,

$$\begin{aligned} K_{M_T}(S_1, S_2, ..., S_n) &= K(\xi) + K(\boldsymbol{\theta}) + K(D/\xi, \boldsymbol{\theta}) \\ &= -\log L(S_1, S_2, \cdots, S_e/M_T) + [\log^* v + \log \binom{e+v-2}{v}] + \log^*[C_b \|\boldsymbol{\theta}\|^b_{I(\theta)}]. \end{aligned}$$

2 Computer Simulation and Results

The computer simulation used is as follows: First, two topologies (bifurcate and trifurcate) were prepared as the model trees. The both model tree consist of four OTUs. Second, the ancestral sequence of 3024-bp nucleotides was generated by Monte Carlo simulation. This sequence was assumed to evolve along the topologies of the predetermined model trees to produce the sequences of four OTUs at the leaves. Third, the phylogenetic trees were estimated from the generated sequences of four OTUs by the MDL, AIC and ML. This process were repeated 1000 times.

The results of computer simulation is shown in Table 1. From Table 1, we can see that the ML method shows a poor performance in the case that the tree has multifurcations. On the contrary, the MDL method shows good results both in bifurcate and trifurcate model. This is due to that the MDL involve the correction terms to prevent overfitting by extra complexity tree. Therefore we think that this method is superior to the traditional method(ML and AIC).

References

[1] 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(1995) 121-130.