Integration of Minimum Evolution and Maximum Likelihood Method in Constructing Molecular Evolutionary Tree

Hiroshi Tanaka, Fengrong Ren and Norio Fukuda
{tanaka, ren, fukuda}@tmd.ac.jp
Tokyo Medical and Dental University
Yushima 1-5-45, Bunkyo 113, Japan

Abstract

The maximum likelihood method and minimum evolution method are mainly used to reconstruct the molecular evolutionary phylogenetic trees. But, from statistical viewpoints, each of them has its own problem. To improve this, we combine the maximum likelihood method with minimum evolution method, on the framework of Bayesian maximum a posterior probability(MAP) estimation.

1 Introduction

There are main two groups of the methods to reconstruct the molecular evolutionary phylogenetic trees: one is the distance-based methods such as minimum evolution(ME) method, and the other is maximum likelihood(ML) method. From the statistical viewpoint, the ML method is thought as the most rigorous estimation method. However, some have pointed out the validity of the method when applied to the different tree topologies. In ME method, choosing phylogenetic tree with the minimum sum of branch lengths is its superior feature, but the incorrect probabilistic model is assumed for branch length.

In this study, we provide more probabilistic rigorous framework(Bayesian framework) to extend ML method to include the superior feature of ME method for constructing the molecular evolutionary phylogenetic trees.

2 Bayesian Framework for Phylogenetic Tree

We employ the maximum a posterior probability estimation(MAP) for tree estimation. In the MAP estimation, first, a posterior probability of the statistical parameter values is calculated based on the Bayesian theorem,

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \cdots (1),$$

田中博, 任鳳蓉, 福田典夫, 東京医科歯科大学難治疾患研究所医薬情報, 東京都文京区湯島 1-5-45

where θ represents parameters of the model and x is the observed data. The second term means a priori probability of the model. Since the demoninator of the above expression is not related with θ, the quantity which should be maximized is $P(x|\theta)P(\theta)$. With taking logarithm of the above, we get the following maximization criterion

$$\log P(x|\theta) + \log P(\theta) \cdots (2).$$

We select the evolutionary priori on the following assumptions:

(1) Evolution takes costs so that it is more likely the shorter path of the evolution has occurred if it provides the same outcome. (2) Evolutionary time is not linearly scaled. The cost of the evolutionary time increases multiplicatively. (3) The prior follows that of Jeffrey's other than the above assumptions (1) and (2). If we adopt the following assumptions, the prior of the evolutionary tree which have the branching length θ_i is proportional to $\Pi_{i=1}^{p} \frac{1}{exp(\theta_i)}$. If we take a logarithm of the priors, we get the following maximization criterion.

$$\log f(x|\theta) - \sum \lambda|\theta_i| \cdots (4).$$

The above criterion can be considered as a combination of the maximum likelihood method and the minimum evolution method, where λ is the normalized coefficient for a prior.

3 An Experiment

The data computed are from 4938-bp mitochondrion DNA of human, common chimpanzee, pygmy chimpanzee, gorilla and orangutan [1], and we develop a computer program by using Nelder and Mead's downhill simplex method to obtain the MAP estimation of branch lengths.

It is found that ML-ME based method constructs identically same tree with Horai's one, but, the resultant tree by ML method is different in the branching length between gorilla and orangutan near the root. Not only from Horai's study but also from the other related study [2], ML-based tree is slightly inaccurate. Possible accounts for this difference are as follows:

The ML-ME 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 near the leaves likelihood function has a tight relation with observed sequences of present species. But as the tree goes down to the root, likelihood function becomes more inaccurate. On the contrary, the second term preferring the minimum branching lengths becomes more effective. If the principle of parsimony is valid still in molecular evolution, the ML-ME based method well compensates the inaccuracy of ML method.

References

[1] 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.

[2] F. Ren, H. Tanaka, N. Fukuda and T. Gojobori, "Molecular Evolutionary Phylogenetic Tree Based on Minimum Description Length Principle," *to be appeared in Proc. of Hawaii International Conference on System Sciences*, 1995.