MDL Method: an Inductive Inference Method for Reconstructing Phylogenetic Trees

Fengrong Ren Hiroshi Tanaka Norio Fukuda

rencom@tmd.ac.jp tanaka@cim.tmd.ac.jp fukucom@tmd.ac.jp

Medical Research Institute, Tokyo Medical and Dental University

Yushima 1-5-45,Bunkyo 113, Japan

Though various computational methods have been proposed to reconstruct the phylogenetic tree from the nucleotide sequence differences so far, there are no decisive reconstruction methods yet so that a new method is expected. The expected method is supposed to be based on the probabilistic model of the base substitution process along the tree, and besides, to incorporate some evaluation of the tree model itself. To do this, we developed a method that based on the general concept of the minimum complexity principle used in the field of inductive theory(model) formation, which states that the model which has the least complexity and nevertheless explains the data well should be chosen as the best one. In applying this principle to the phylogenetic problem, we use Rissanen's definition of stochastic complexity in which the complexity is described by its code lengths [1]. Rissanen called the model estimation by this concept "Minimum Description Length(MDL)"estimation.

In this principle, the model(theory) which has the minimum description length is chosen as the most suitable one. Namely, the MDL principle requests a theory T that minimizes

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

where $\ell(T)$ is the code length needed to encode the theory T. $x_1, \cdots, x_n$ are the observations, and $\ell(x_1, \cdots, x_n \mid T)$ is the code length needed to encode the observations with respect to T.

The complexity of the phylogenetic tree based on the MDL principle is given by

$$L_{total} = -\log(S_1, S_2, \cdots, S_p \mid t) + [\log^* v + \log \binom{w-2}{v}] + \frac{1}{2}\sum_{i=1}^{m} \log\{t_i^2 / I^{ii}(t)\}, \qquad (2)$$

任 鳳蓉、田中 博、福田 典夫：東京医科歯科大学難治疾患研究所情報医学研究部門（医薬情報），

〒 113 東京都文京区湯島 1-5-45

where $S_1, S_2, \cdots, S_p$ is the current nucleic sequences of p species. $t = (t_1, \cdots, t_m)$ in which t_i denotes the ith branch length and m is total number of branches. The first term is a negative log likelihood function, and we can consider it to be the code length of the deviation between the model and the data which is denoted by $\ell(x_1, \cdots, x_n \mid T)$ in the previous formula. The second term is a term concerned with the code length of the model complexity which is related to structural properties of the model, where $\log^* v = \log v + \log\log v + \cdots$, and v and w is the number of the internal nodes and that of the total nodes of the tree, respectively. The third term is the code length when we describe the values of the parameters(the branch length of the tree) with precision level of their estimation error, where I^{ii} is the ith diagonal element of the inverse of Fisher information matrix.

For the description of likelihood function, we use Markov matrix of molecular base substitutions by Felsenstein and Hasegawa. 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 kind of integration of the maximum likelihood method and the minimum evolution method.

A preliminary investigation into the estimation accuracy of our method was executed by an experiment on the reconstruction of phylogenetic tree of the mammalian. The 1008-bp cytochrome b DNA sequences from Human, Bos, Mus and Pelecanus(as an outgroup) are examined. We developed a program for computing MDL method based phylogenetic tree, which calculates the optimal parameter values which attain the minimum code length by employing the downhill simplex method. To compare MDL method with the maximum likelihood method, we used Felsenstein's DNAML in his program package PHYLIP. The sequences data computed are same as those used in MDL method.

The result by MDL method is (((Human: 0.10, Bos: 0.10): 0.055, Mus: 0.125): 0.045, Pelecanus: 0.20) and that by ML method is (((Bos: 0.133, Mus: 0.093): 0.05, Human: 0.16): 0.017, Pelacanus: 0.20). It is clear that both topology and branch length are different to each other. The ML-method tree is obviously contradictory to traditional results which suggest that Rodentia branch off before divergency between Artiodactyla and Primate, whereas the MDL-method tree show that Primate has a closer relationship with Artiodactyla than Rodentia [2]. Moreover, the sum of the branch length by MDL-method is smaller than that by ML-method.These results suggest that the MDL method might be superior to the traditional method because it take both the fitness of the model to data and the tree model itself as a criterion for estimating phylogenetic tree.

References

[1] J. Rissanen, Stochastic complexity and modeling, Ann. Statist. 14(1986) 1080-1100.

[2] Rodney L.Honeycutt, Michael A.Nedbal, Ronald M.Adkins, Laura L.Janecek, Mammalian Mitochondrial DNA Evolution: A Comparison of the Cytochrome b and cytochrome c Oxidase II Genes, J.Mol.Evol. 40(1995) 260-272.

[3] F.Ren, H.tanaka and T.Gojobori, Construction of molecular evolutionary phylogenetic trees from DNA sequences based on minimum complexity principle, Computer Methods and Programs in Biomedicine 46(1995) 121-130.