An approximate maximum likelihood method for phylogenetic tree analysis based on high-temperature Markov Chain Monte Carlo

Ryota Suzuki,Tomoya Taniguchi and Hidetoshi Shimodaira

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology.
email: {ryota.suzuki,taniguc2,shimo}@is.titech.ac.jp

Introduction

Maximum likelihood (ML) method has been widely used because it allows phylogenetic analysis based on probabilistic models of molecular evolution. However, despite its effectiveness and simplicity, ML method does not work properly in analyses of many species - it even fails with only 20-30 species. To overcome this problem, we propose an approximate version of ML method based on Markov Chain Monte Carlo (MCMC). This method has already been used in Shimodaira and Hasegawa (2004) [4]. An example of phylogenetic tree for 32 mammalian species computed by the proposed method will also be shown.

Problems

ML method requires computation of likelihood for all the candidate topologies. For s species, the number of candidates is

$$\frac{(2s-5)!}{2^{s-3}(s-3)!} \tag{1}$$

Since computation of likelihood is time-consuming process, it is virtually impossible to compute for all the candidates when s is large. To avoid this problem, MCMC is often used as an alternative.It generates a set of topologies with frequencies being proportional to posterior probabilities. Theoretically, while we do not have to compute the likelihood values for all topologies, it provides posterior probability of each topology. But MCMC also fails in practice due to the misspecification of probability models and generates too small number of topologies.

Method

Our approach to this problem is as follows. First, we run a MCMC process with "high temperature". Let T denote a topology and $P(T)$ its posterior probability. For $r > 0$, MCMC with inverse temperature r generates topologies to be proportional to the amount

$$P_r(T) \approx \frac{P(T)^r}{\sum_{k=1}^{K} P(T_k)^r} \tag{2}$$

,where K is the number of all the possible topologies. For $r \neq 1$, the generated topologies are not proportional to posterior probabilities. Instead, we can obtain as large number of candidate topologies as r gets smaller. We run a MCMC process with $r < 1$ (high temperature), not to compute posterior probabilities but to obtain a list of candidate topologies. The topologies in the list are expected to include the "true" topology. Then, we compute the likelihood values for all topologies in the list obtained by MCMC procedure. The topology with the highest likelihood value among the list is taken as the approximate maximum likelihood topology. Finally, we perform multiscale bootstrap analysis [3] to assess the uncertainty of the results in terms of hypothesis testings. There are two reasons to higher the temperature of MCMC:

(1) From the viewpoint of ML method, posterior probability is only approximately proportional to likelihood.

(2) Wide range of candidate topologies is needed for (multiscale) bootstrap analysis.

However, regarding the latter it is still not clear how wide range is enough to approximate the p-values of edges computed by (multiscale) bootstrap resampling.

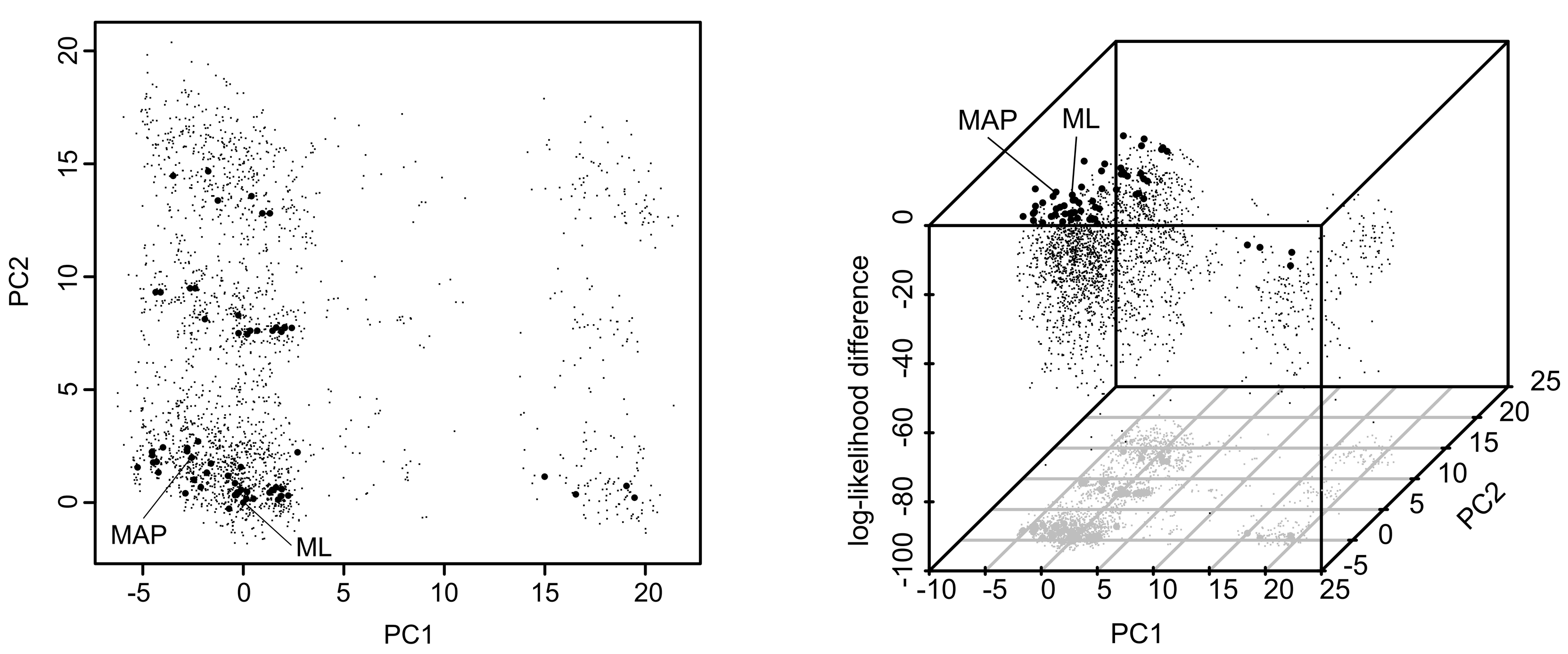
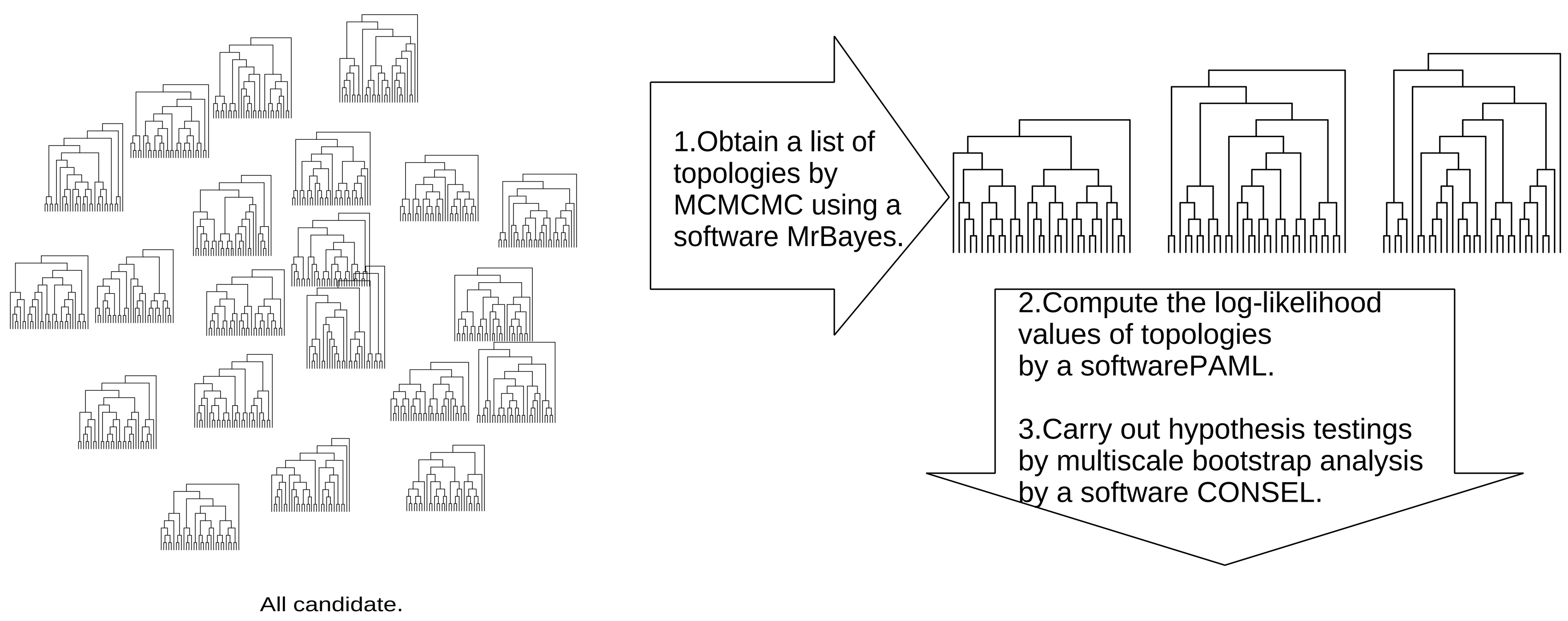


Figure 1: Visualization of the candidate tree topologies.

Example

As an example, we estimated an approximate ML tree for 32 mammalian species from a part of the mitochondrial protein sequences of Nikaido et al. (2003) [2] The procedure is as follows:



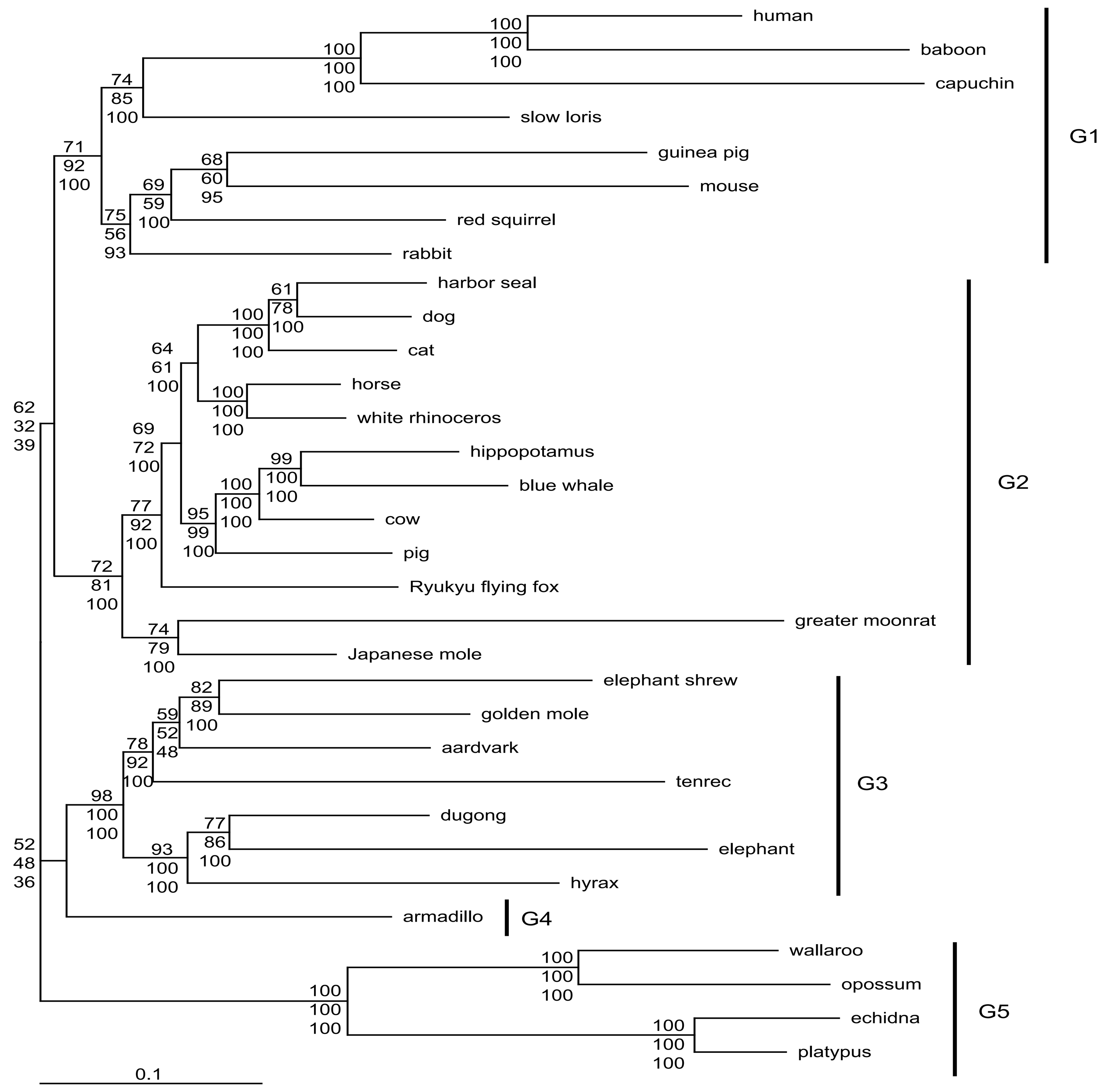


Figure 2: The estimated topology of 32 species.

The estimated approximate ML tree is shown in Figure 2. Three numbers near branches are the approximately unbiased p-value, bootstrap probability, and posterior probability respectively from the top to the bottom, in percentage. The software implementing this method will be available at our website [7].

References

[1] Huelsenbeck, J. P. and Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, pages 17:754–55, 2001.

[2] Nikaido, M. et al.. Mitochondrial phylogeny of hedgehogs and monophyly of eulipotyphla. *Molecular Phylogenetics and Evolution*, page 28(2):276284, 2003.

[3] Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, pages 51:492–508, 2002.

[4] Shimodaira, H. and Hasegawa M. Assessing the uncertainty in phylogenetic inference. *Statistical Methods in Molecular Evolution*(ed. R. Nielsen) Springer-Verlag, 2004.

[5] Yang Z.. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput.Appl. Biosci*, pages 13:555–56, 1997.

[6] Shimodaira, H. and Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, page 17:12461247, 2001.

[7] *http://www.is.titech.ac.jp/˜shimo/prog/*