Phylogenetic Inference Using Parallel Version of MrBayes

Kong Lesheng

(*Department of Biochemistry, NUS*)

1. Introduction

The inference of phylogenies with computational methods is widely used in medical and biological research and has many important applications, such as gene function prediction, drug discovery and conservation biology [1].

Bayesian inference is a powerful method which is implemented in the program MrBayes [2] for estimating phylogenetic trees that are based on the posterior probability distribution of the trees (Figure 1). Comparing to the parsimony and distance methods, Bayesian inference takes full advantage of the information contained in the alignment of DNA sequences (even can make use of morphological data) when estimating phylogenies.

2. The computation problem and the parallel version of MrBayes

Due to the nature of Bayesian inference, the simulation can be prone to entrapment in local optima. To overcome local optima and achieve better estimation, the MrBayes program has to run for millions of iterations (generations) which require a large amount of computation time. For example, the phylogenetic estimation for a medium size dataset (50 sequences, 300 nucleotides for each sequence) typically requires a simulation for 250,000 generations, which normally runs for 50 hours on a PC with an Intel P4 2.8 GHz processor. For multiple sessions with different models or parameters, it will take a very long time before the results can be analyzed and summarized.

With the help of SVU staff, the MPI-enabled parallel version of MaBayes [3] is compiled and installed in one of their Linux cluster. Performance results show there is nearly a linear speed up for the parallel jobs. The same job which originally took two days can be finished within few hours. The significant improvement of simulation speed also allows me to run the application for a longer session to achieve more stable and accurate estimation. Moreover, the job submission through LSF queues is very flexible and convenient for me to manage jobs.

3. Summary

The parallel version of MrBayes (MPI-enabled) greatly improves the simulation speed for phylogenetic inference and dramatically reduces the computation time, which greatly facilitates my research on phylogenetic estimation using Bayesian inference.

Reference:

1. Bader DA, *et al*. Industrial Applications of High-Performance Computing for Phylogeny Reconstruction. *Proceedings of SPIE ITCom*, 2001; 4528:159–168,
2. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003; 19:1572-1574.
3. Altekar G, *et al*. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 2004; 20: 407-415.

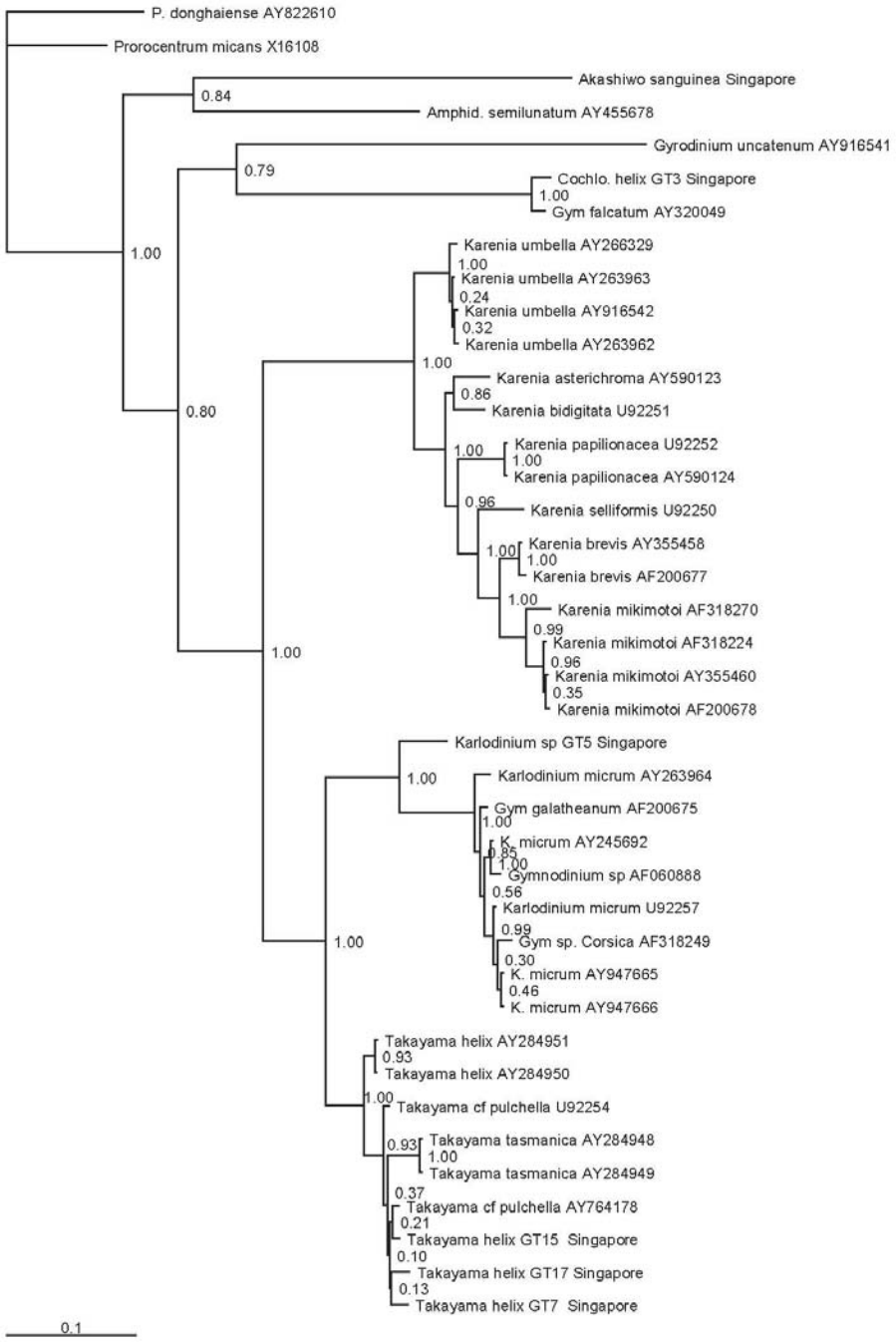

Figure 1, Phylogenetic Tree of Large Sub-unit rDNAs from *Dinophyceae* Species