HPC for Phylogenetic Tree Inference

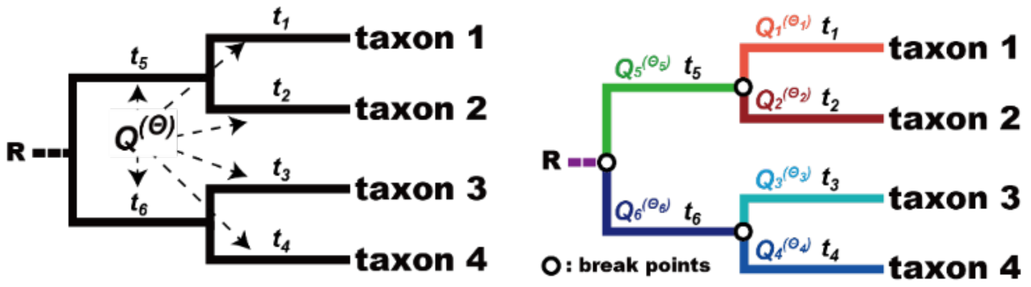
MPI/OpenMP parallelization of the phylogenetic analyses based on non-homogeneous substitution models

Introduction

The non-homogeneous (NH) models, which allocate different model parameters on each node of the tree to evaluate (Fig. 1), are a realistic approach to reconstruct phylogenetic trees appropriately from real-world sequence datasets, since the nucleotide and amino acid sequences in distantly related species certainly evolve under different evolutionary processes. However, the analyses with NH models can be computationally intense as an enormous amount of model parameters need to be optimized.

Fig. 1: Substitution models used in phylogenetic analyses

R: root, t: branch length, Q: rate matrix, Θ: model parameters



Homogeneous model　　Non-homogeneous model

Methods

We applied two parallel computing methods, MPI and OpenMP, to accelerate a phylogenetic program, NHML (Galtier and Gouy, *Mol. Bio. Evol.*, 1998), which implements a NH model for the heterogeneity of guanine and cytosine content across a tree. Two schemes were applied to parallelize the maximum-likelihood (ML) phylogenetic inference by NHML;

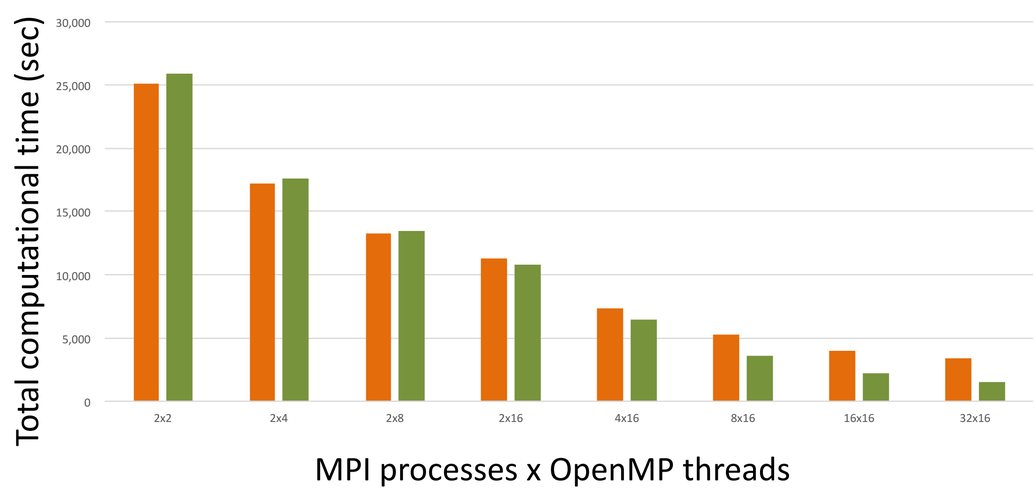
A) Parallel ML estimation of model parameters and branch lengths of a fixed tree

B) Parallel computation of likelihood scores for multiple trees, where each likelihood can be calculated *via* A)

Results

Performances of the two parallel schemes for NHML were evaluated on COMA (PACS-IX) system by analyzing simulated nucleotide sequence data comprising 15 taxa and 10,000 bp. Both two schemes showed good speed-up up to 512 CPU cores (32 computation nodes), especially the hybrid parallel scheme A) + B) achieved ~78 times speed-up compared to the serial code (Fig. 2).

Fig. 2: Performance evaluation on COMA (PACS-IX)



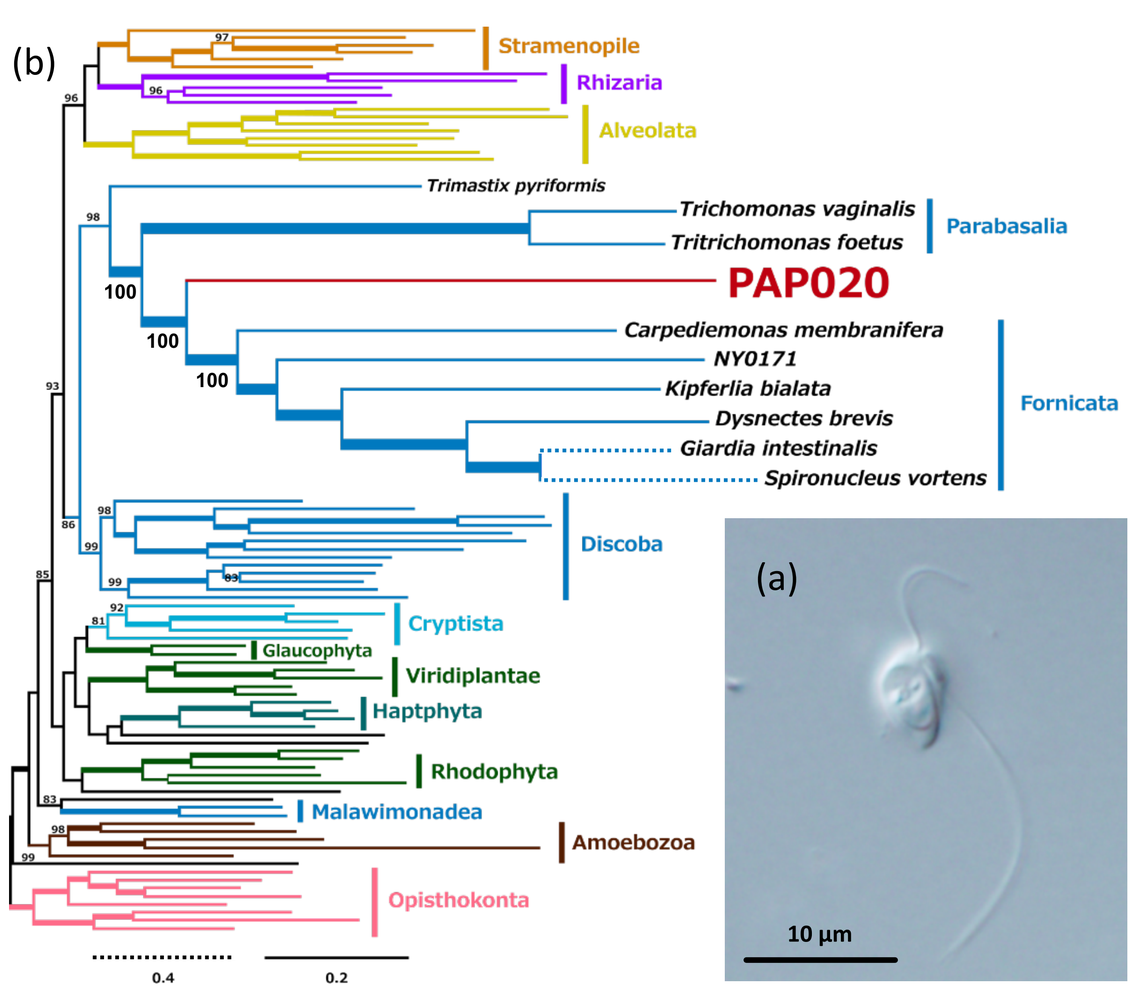
Phylogenetic position of a microaerophilic eukaryotic microbe, strain PAP020, inferred from a large-scale multigene dataset

We isolated a novel eukaryotic flagellate, strain PAP020 (Fig. 3a), from mangrove sediments. This strain has been maintained in the laboratory under the microaerophilic condition. In this study, we conducted RNA-seq analysis on this strain, and assembled a alignment comprising 148 highly conserved protein sequences shared among 83 diverse eukaryotes. The resultant '148-protein' dataset was then subjected to the maximum-likelihood (ML) phylogenetic analyses.

Strain PAP020 branched at the base of an assemblage of anaerobic eukaryotes called Fornicata with high statistical support (Fig. 3b). Thus, this strain most likely represents a previously unnoticed lineage that is related intimately to Fornicata, or is an ancestral member of Fornicata. Either way, PAP020 is critical to understand the anaerobic metabolisms in fornicates including giardiasis pathogen, *Giardia intestinalis*.

Fig. 3: Phylogenetic position of strain PAP020

(a) Strain PAP020. (b) ML tree inferred from a 148-protein dataset. Only bootstrap values greater than 80% are listed on the corresponding bipartitions. Thick branches represent bipartitions with full support in the bootstrap analysis.



contact address: yuji@ccs.tsukuba.ac.jp