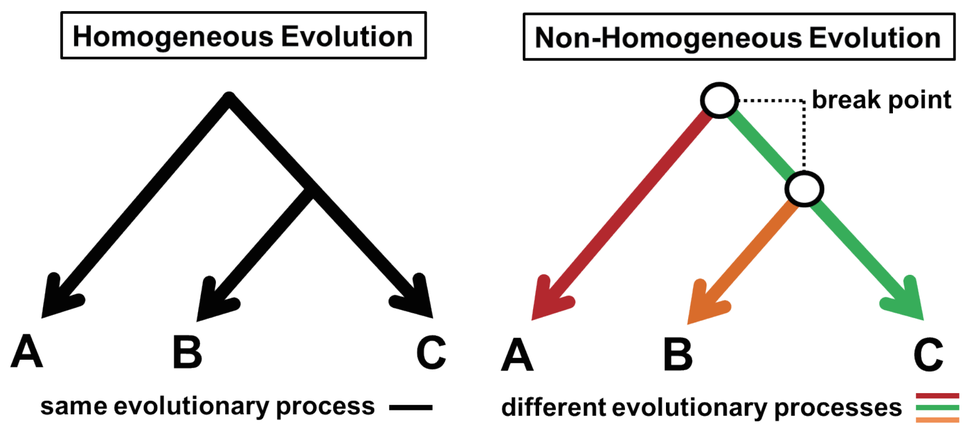
HPC for Phylogenetic Tree Inference

Parallelization of the phylogenetic inference with the non-homogeneous substitution models

Introduction: The nucleotide and amino acid sequences in distantly related species evolve under different evolutionary processes (non-homogeneous evolution; Fig. 1). Non-homogeneous (NH) models, which allocate different model parameters on each node of the tree to evaluate, are the most efficient approach to reconstruct phylogenetic trees appropriately from real-world sequence datasets. NHML, the one of phylogenetic programs implementing NH models, can tolerate the heterogeneity of guanine (G) + cytosine (C) content in nucleotide sequences among lineages and has been generally applied to the analyses of real-world data. However, the analyses with NHML can be computationally intense as enormous amount of model parameters need to be optimized. Toward applying to large-scale sequence data, we parallelized NHML and evaluated its performance on the super cluster in this study.

Fig. 1: Difference of the evolutionary process of molecular sequences



Methods: We simulated two nucleotide sequence data, i) 66 taxa and 10,000 nucleotide positions, and ii) 130 taxa and 2,500 nucleotide positions, based on different model trees. Each sequence dataset was subjected to the likelihood calculation of the corresponding model tree with NHML. Data analyses were run on T2K Tsukuba (~ 16 nodes/256 CPU cores).

Results: Two approaches for parallel computing, OpenMP and MPI, were applied into the algorithm for the maximum likelihood (ML) calculation of a single tree in NHML. From the analyses of two simulated sequence data, regardless of the number of taxa and nucleotide positions, the parallel version of NHML successfully retained good parallel efficiency until using 256 CPU cores (eff >= 0.5; Fig. 2).

Fig. 2: Speed-up of the ML calculation of a tree

Reference: Galtier and Gouy. “Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis.” Molecular Biology and Evolution. 1998. 15:871-9.

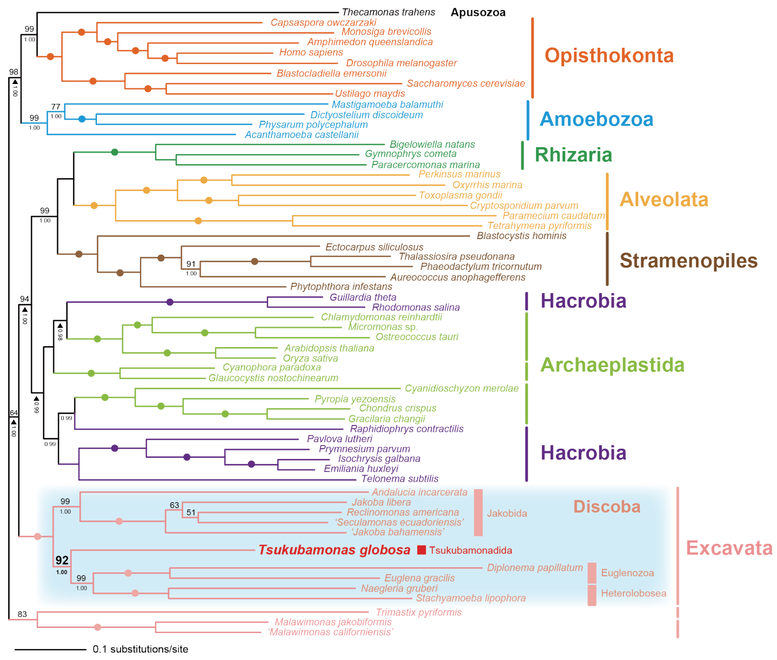
Phylogenetic position of a newly identified protist, *Tsukubamonas globosa*, inferred from a large-scale multigene dataset

Introduction: *Tsukubamonas globosa* is a recently identified single-cell protist and considered to provide key information regarding a potentially early-branching eukaryotic assemblage called Discoba. However, the phylogenetic position of *T. globosa* has yet to be resolved with confident.

Results: We conducted massive expressed sequence tag analyses on *T. globosa* using 454 pyrosequencing technology, and prepared a 157-gene dataset including 41,372 amino acid positions. In the 157-gene analyses, *T. globosa* robustly grouped with previously known discobids (Fig. 3). In particular, *T. globosa* was branched at the base of the clade of euglenozoans and heterolobosean with high statistical support.

Discussion: Our 157-gene analyses successfully pinpointed the phylogenetic position of *T. globosa*. As *T. globosa* was not nested with any of the known discobid subgroups, we can conclude that this flagellate represent a previously unknown group in Discoba.

Fig. 3: Phylogenetic position of *Tsukubamonas globosa*



References: Yabuki et al. “*Tsukubamonas globosa* n. g., n. sp., a novel excavate flagellate possibly holding a key for the early evolution in Discoba.” J. Eukaryot Microbiol. 2011. 58:319-331.

Kamikawa et al. “Gene content evolution in discobid mitochondria deduced from the phylogenetic position and complete mitochondrial genome of *Tsukubamonas globosa*.” Genome Biol Evol. 2014. 6:306-315.