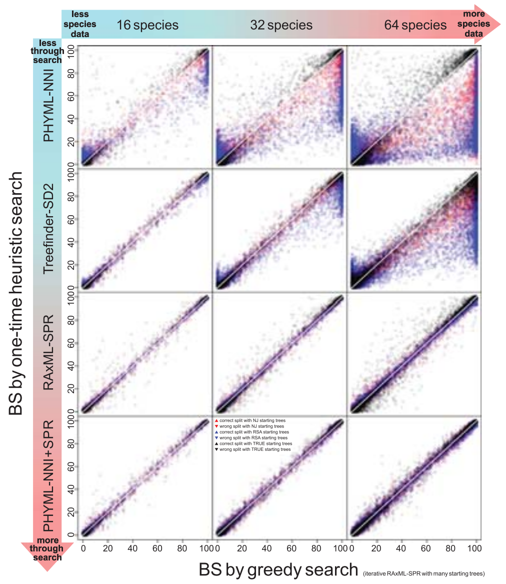
Center for Computational Sciences, University of Tsukuba
www.ccs.tsukuba.ac.jp

HPC for Phylogenetic Tree Inference

Bootstrap Support Values in Maximum Likelihood Phylogenetic Inference May be Strongly Biased by Insufficient Tree Search



Note that all points are drawn with 90% transparency because there are too many points.

Bootstrap support value (BS) is the most commonly used credibility index of phylogenetic tree of living organisms. BS is the percentage of occurrence of "split" that is an internal branch biparting the species on the tree to two groups in the best trees of bootstrap resampled data sets. BS is calculated from 3 steps computation like the following.

1) make 100 or more resampled data sets from original data set based on bootstrap method

2) perform heuristic search for best phylogenetic tree at each of resampled data sets

3) count occurrence of splits in the best trees

Under the maximum likelihood criterion that is the most commonly used criterion, only one-time heuristic search is almost always performed at each of resampled data sets because optimization of likelihood is huge computation. However, best tree vary based on the starting trees. Then, we tested whether BS also vary based on the starting trees or not.

We generated many simulated data sets, calculated BS by several one-time heuristic searches with NJ, RSA, or TRUE starting tree and by very greedy iterated heuristic search with many dispersed starting trees at each data set, and compared BS by one-time heuristic searches with BS by greedy search. NJ starting trees were made by neighbor-joining method. RSA starting trees were made by random sequence addition method. TRUE tree is the model tree used in simulated data generation.

As a result, we found that less through search produced more biased BS, and that more species data raised more biased BS. In addition, TRUE starting trees caused overestimation of BS of correct splits, and NJ and RSA starting trees caused underestimation of BS of correct splits. Because the most through search which produced accurate BS requires too much times, these results suggested that we need a new method for obtaining accurate BS at very huge data sets.

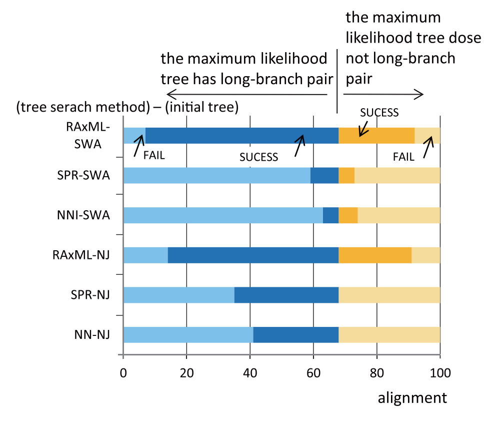
Performance evaluation of heuristic tree optimization methods

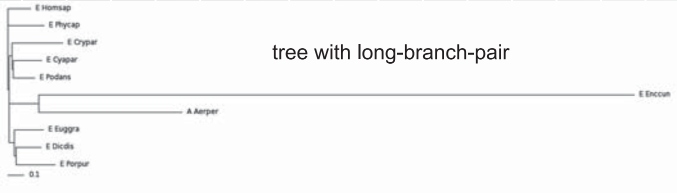
Long branch attraction is the erroneous grouping of two long branches as sister group due to methodogial artifacts.

For each of 100 alignments of 10 OTUs, the likelihoods of all possible trees (more than 2-million trees for an alignment, **200-million trees in total**) are calculated exhaustively by PACS-CS.

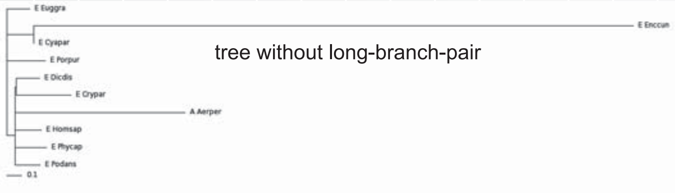
The maximum likelihood trees and the best trees obtained by heuristic tree optimization methods are compared.

The right figure shows that, of 100 alignments, there are 68 maximum likelihood trees which have a pair of long branches and 32 maximum likelihood trees which do not have a pair of long branches. For the latter case, heuristic tree optimization methods fail to find the maximum likelihood trees. Especially, starting with a tree by NJ (Neighbor joining) , SPR and NN cannot find the maximum tree without long-branch pair. NJ always give a tree with long-branch pairs and SPR and NN cannot jump from the tree with long-branch pairs to the one without long-branch pairs over search space.





tree with long-branch-pair



tree without long-branch-pair

New algorithm based on the comparison of results between exhaustive and heuristic methods

A new algorithm is being developed based on the comparison of results between exhaustive and heuristic methods. The algorithm is the extension of genetic algorithm. The genetic algorithm has multiple candidate solutions (trees) through its search. Our algorithm gives a certain ratio of trees with/without pair(s) of long branches.