

Faster Algorithms for Support Value Computation & Parallel Computing for Large-Scale Phylogeny Reconstruction

Alexandros Stamatakis

*Ecole Polytechnique Federale de Lausanne, School of Computer & Communication Sciences,
Laboratory for Computational Biology and Bioinformatics*

Despite the impressive progress that has been achieved with the new generation of Maximum Likelihood (ML) search algorithms for phylogeny reconstruction, the computation of support values based on non-parametric bootstrapping (BS) still represents a major computational challenge.

Initially, I will present novel tree search heuristics that accelerate the search process by factor 2.5 on average while yielding equally good trees.

In addition, I will present a new algorithm to accelerate the BS procedure in RAxML (Randomized Accelerated Maximum Likelihood). In comparison to the standard BS procedure these heuristics yield run time improvements between factor 6 on datasets with 140 sequences up to factor 13 on datasets with several thousands of sequences.

At the same time the support values obtained by the new BS heuristics show correlation coefficients ranging between 0.94 and 0.98 compared to those obtained via the standard method.

In absolute numbers this means that a full phylogenetic analysis including 100 bootstrap replicates and a search for the best-scoring ML tree on single-gene datasets of up to 2,000 taxa can be conducted within less than 48 hours on a single -reasonably fast- processor. This represents an improvement of one order of magnitude over current approaches.

In the second part of my talk I will outline how the computation of large multi-gene datasets with ML can efficiently be parallelized on the IBM BlueGene supercomputer. The parallelization on BlueGene scales well up to 1,024 processors.

We used the IBM BlueGene to compute trees on the largest dataset analyzed under ML to date, which consists of 270 sequences and 500,000 base pairs.