Paper

Nakhleh L., St John K., Roshan U., Sun J., Warnow T. (2001). Designing fast converging phylogenetic methods. Bioinformatics 17, Suppl 1, S190?8.

First Study - motivation

Neighbor Joining

- polynomial running time
- statistically consistent but there is no provable guarantee regarding the amount of data (sequence length) required to reconstruct the true tree
- suffers from low accuracy on datasets with distantly related taxa, when given sequences of practical length

Goal: To compare NJ to absolute fast converging methods

First study - simulation

- `Seq-Gen` generate sequences of varying length given a tree topology with branch lengths and model of evolution
- Random trees and **biologically-based trees**
- Branch lengths are scaled to create datasets with varying evolutionary diameters

First study - simulation

- Create 50 replicates with sequence lengths of 8000 base pairs
- Jukes-Cantor model of evolution
- Use first 200-8000 base pairs

First study - results

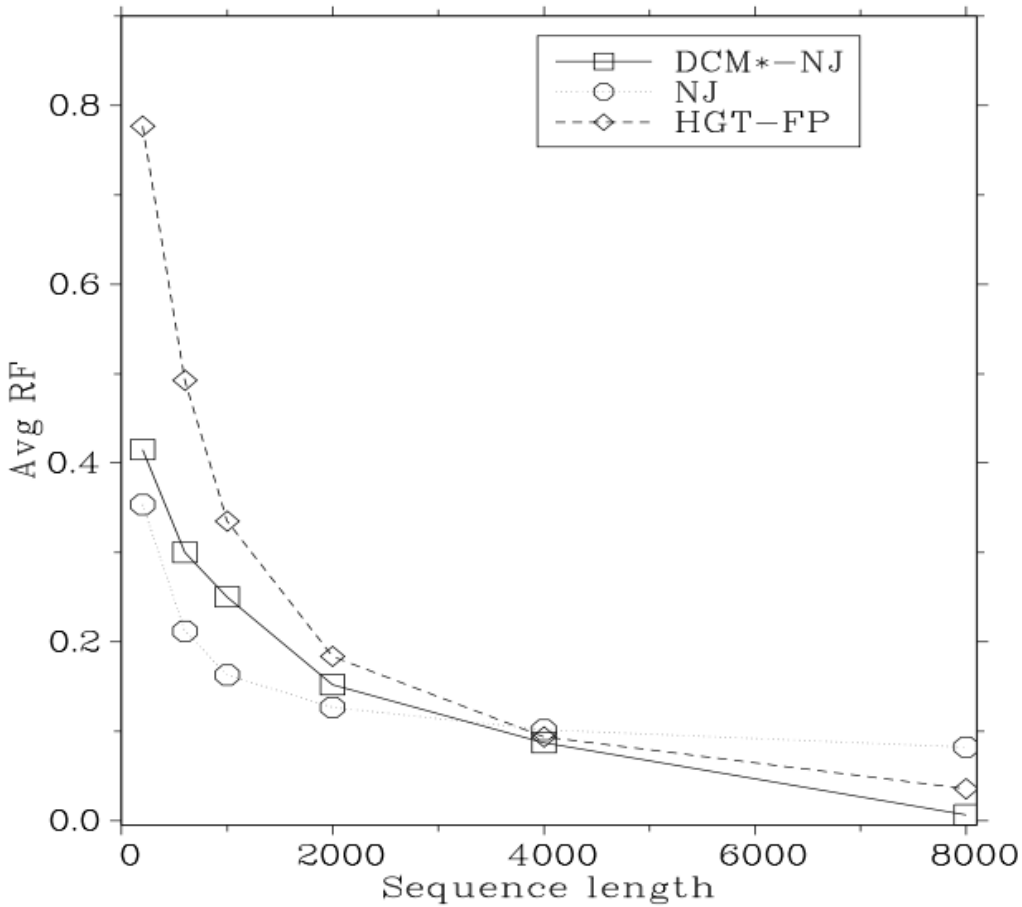

Fig. 3. DCM*-NJ vs. NJ vs. HGT+FP on the rbc*L* 500-taxon tree, under the JC model. Average branch length is 0.264

Final study - motivation

Modifications to absolute converging method, DCM*-NJ, to improve its accuracy on shorter sequences

Final study - simulations

- Modifications to absolute converging method, DCM*-NJ, to improve its accuracy on shorter sequences
- Kimura 2-parameter model of evolution with variable site rates drawn from the gamma distribution
- Varied number of taxa with fixed branch lengths and sequence lengths

Final study - results

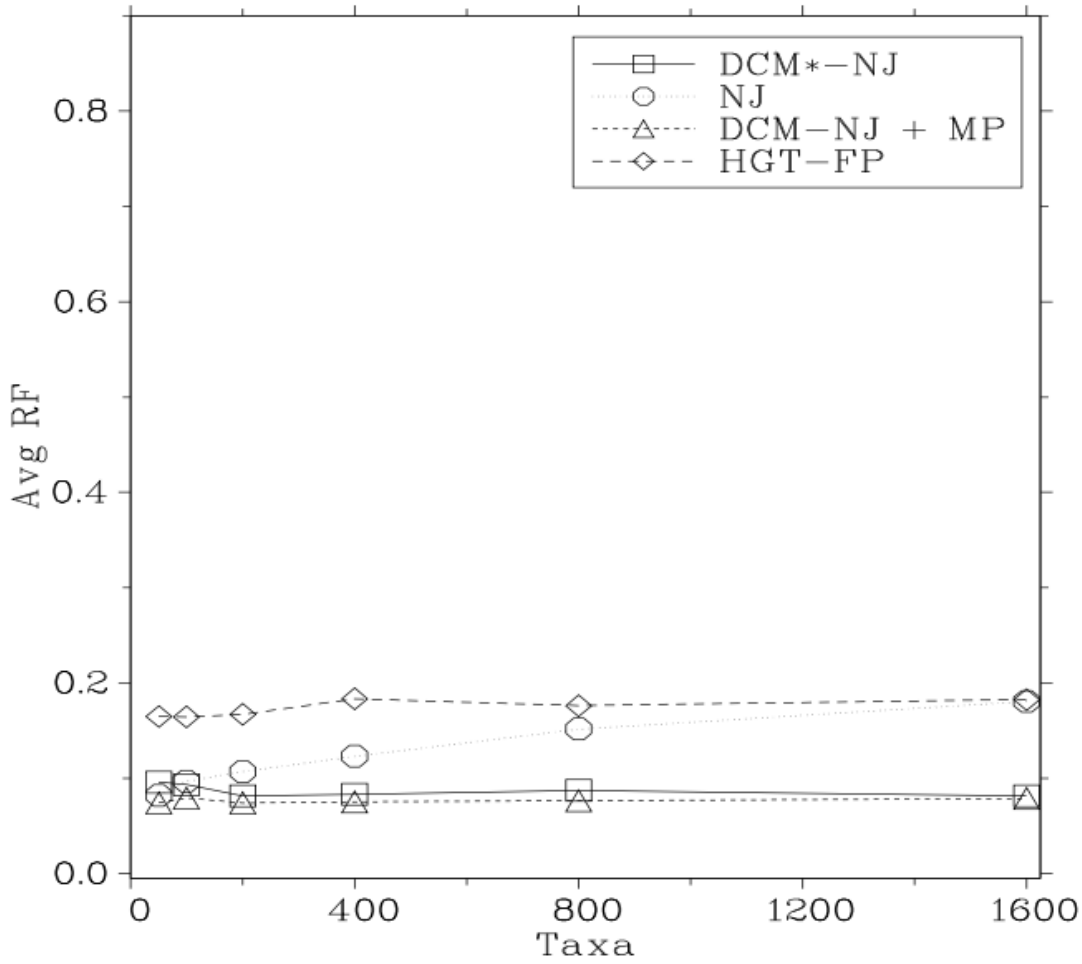

Fig. 8. DCM-NJ+MP vs. DCM*-NJ vs. NJ vs. HGT+FP on random trees under the K2P+Gamma model. Sequence length is 1000. Average branch length is 0.005.

Simulation Improvements

- `Seq-Gen` models all evolutionary events as substitutions
  - Effect of missing data or gaps
  - Impact of alignment error
- Biological sequence data