The Accuracy of Fast Phylogenetic Methods for Large Datasets

Pacific Symp. Biocomputing (PSB), 2002

Reviewed by Kajori Banerjee

Motivation

- Compares 4 methods

1) Neighbor-joining
2) Weighbor
3) Greedy parsimony and
4) DCM-NJ+MP

Parameters for comparison :

1) Model of evolution - Jukes-Cantor and Kimura2-Parameter+Gamma
2) Tree diameter
3) Sequence length requirement
4) Taxon sampling

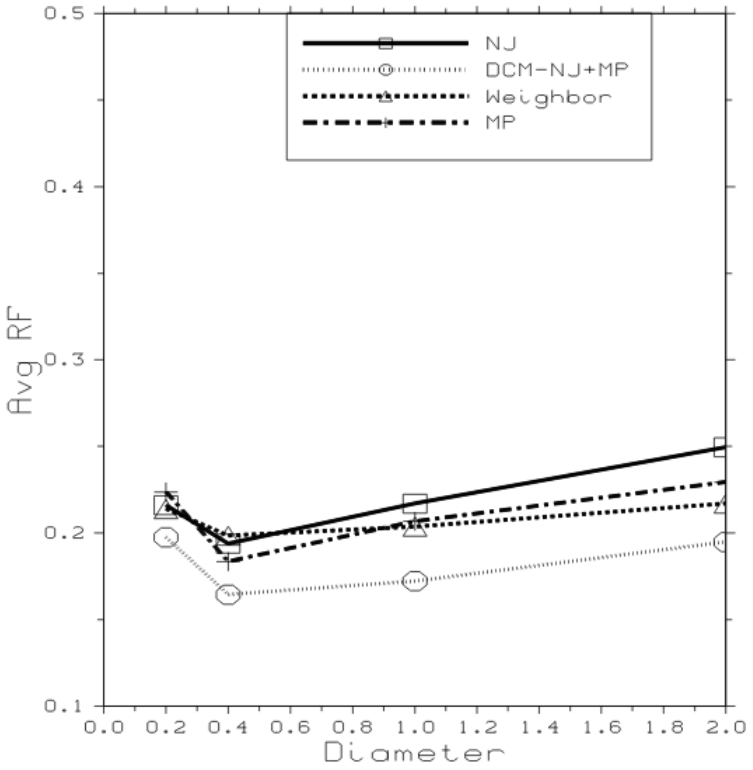
Dataset Generation

1) Generate Model tree : model true tree using Random birth-death process

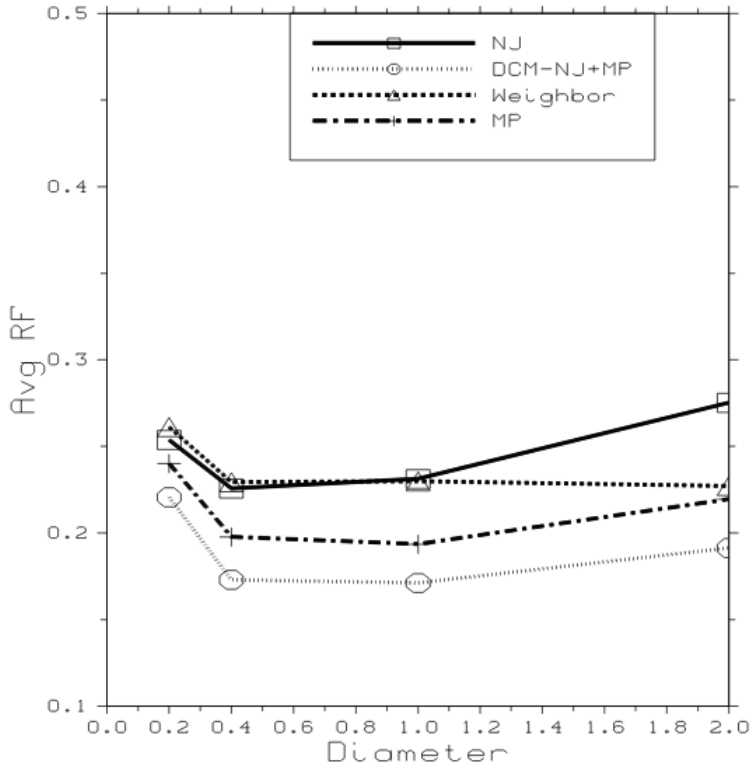
2) Make model trees non-ultrametric - the edges are multiplied with a random number from the interval [1/c, c] is

3) Evolve sequences : Jukes-Cantor or Kimura2-Parameter+Gamma model.

Results



(a) 100 taxa



(a) 400 taxa

Figure 5: Accuracy as a function of the diameter under the K2P+Gamma model for fixed sequence length (500) and two numbers of taxa

Conclusion

1) **Sequence length** :

- Weighbor - better performance for small sequence lengths
- DCM-NJ+MP - more appropriate for data with longer sequences.
- NJ require sequences length to be exponential with respect to the evolutionary diameter of the true tree.

2) Speed :

- Both Neighbor-joining and greedy parsimony are generally faster than Weighbor and DCMNJ+MP .
- Greedy Parsimony is very fast but low topological accuracy.

3) Tree diameter :

DCM-NJ+MP has better topological accuracy then NJ as the evolutionary distance between the taxa in the dataset increases.

4) Model of evolution

5) Taxon sampling

- DCM-NJ+MP has better topological accuracy then NJ with respect to as the number of taxa in the dataset increases.

Comments

- Have not considered boot-strapping.
- Model tree used in these experiments are free from horizontal gene transfer

References

1) Ganapathy, V. Ramachandran, and T. Warnow. Better hill-climbing searches for parsimony. In Algorithms in bioinformatics, pages 245–258. Springer, 2003.

2) T. H. Jukes and C. R. Cantor. Evolution of protein molecules. Mammalian protein metabolism, 3(21):132, 1969.

3) M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotidesequences.Journal of molecular evolution, 16(2):111–120, 1980.

4) W. Li-San. Genome rearrangement phylogeny using weighbor. In Algorithms in Bioinformatics, pages 112–125. Springer, 2002.

5) L. Nakhleh, B. M. Moret, U. Roshan, K. S. John, and T. Warnow. The accuracy of fast phylogenetic methods for large datasets. In Proc. 7th Pacific Symp. Biocomputing PSB 2002 , pages 211–222, 2002.

6) L. Nakhleh, U. Roshan, K. S. John, J. Sun, and T. Warnow. Designing fast converging phylogenetic methods. Bioinformatics,17(suppl 1):S190–S198, 2001.

7) D. Robinson and L. R. Foulds. Comparison of phylogenetic trees.Mathematical biosciences, 53(1):131–147, 1981.

8) N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees.Molecular biology and evolution, 4(4):406–425, 1987