ThreeTree: A New Method to Reconstruct Phylogenetic Trees

Satoshi OOta
oota@thinker.lab.nig.ac.jp
Laboratory of Evolutionary Genetics, National Institute of Genetics, Mishima, 411-8540 Japan

1 Introduction

Numerous methods have been invented to reconstruct phylogenetic trees. However, criteria used here may not necessarily suitable in actual biological data. On the other hand, if we obtain the same result from methods based on different criteria, the inference can be considered to be robust and reliable. Therefore, some complementary method to existing methods will improve accuracy at inference of phylogenetic trees when they are applied together. The objectives in this paper are as follows:

- Development of a simple method which is not based on the minimum evolution criterion.
- Development of a method which has complementary characteristics to both the neighbor-joining method [2] and the maximum likelihood method [3].
- Development of a robust method in different evolutionary rates among lineages.

2 Algorithm

2.1 The ThreeTree method

Every unrooted tree contains trees having three OTUs (Operational Taxonomic Units). In other words, every unrooted tree can be divided into trees having only three OTUs, whose number is ${}_nC_3$ ($n \geq 3$, n is the number of the OTUs). Here for simplicity, we assume that pure additiveness is hold in a given distance matrix. According to Fitch and Margoliash [1], we can easily infer such ${}_nC_3$ trees. If we can find true neighbor, it is obvious that we can always find the true topology. On the other hand, the true tree, which can be spanned on the distance matrix D, has at least one neighbor. This "true" neighbor is necessarily contained in the ${}_nC_2$ neighbors (The ${}_nC_3$ trees have ${}_nC_2$ neighbors). Once we find the "true neighbor", the true tree can be reconstructed in a recursive way such that the neighbor-joining method does.

Theorem 1 *Consider an unrooted tree having n leaves (OTUs). Let S be the set of the leaves. We can obtain ${}_nC_3$ trees having three leaves i, j, and k ($i, j, k \in S$). Leaves i and j are neighbor to each other, if and only if lengths of all branches ip_x are the same and lengths of all branches jp_x are the same, where p_x is an internal node connected a certain pair of OTUs i and j and the third OTU k (x is the number of possible leaves as the third OTU, so $1 \leq x \leq n-2$). We call such trees "ThreeTrees".*

We should note that more than one neighbors can be found during one iteration if pure additiveness is hold. Unfortunately, practical data satisfying pure additiveness are rare. We may not find any neighbor in the above theorem. It is necessary to apply Theorem 1 to the practical problem in an approximate way, not the exact way. We can obtain a set of variance pairs of branch lengths $l(ip_x)$ and $l(jp_x)$ for all possible trees having three OTUs. Such variance pairs for leaves i and j in ${}_nC_3$ trees are computed for all possible pairs of given OTUs. We can consider that the variance pair having the smallest "value" indicates the true neighbor.

Furthermore, in the maximum likelihood optimization for a branch length, the number of substitutions on the branch is inferred. In other words, the branch length is purely additive if the optimization is perfect. Therefore, by applying the maximum likelihood optimization to the ${}_nC_3$ trees, it is expected that the additiveness is recovered in some degree. The algorithm is as follows:

1. A distance matrix is generated from a given set of sequences.
2. All possible three sequences are chosen from the sequences.
3. All possible trees having three OTUs are generated from the sequences. Branch lengths are inferred by Fitch and Margoliash's method.
4. Using the above branch lengths as initial values, ${}_nC_3$ trees are optimized according to the maximum likelihood method.
5. The most probable neighbor is found according to Theorem 1.
6. Ancestral sequence of the approximate ThreeTree is inferred by the maximum likelihood method.
7. A set of sequences is revised replacing neighbor sequences by their ancestral sequence.
8. The distance matrix is revised with composite OTUs derived above.
9. Steps 2-8 are iterated until three (composite) OTUs are remained.

3 Conclusion

According to simulations, ThreeTree has two interesting characteristics. Firstly, this method gave complementary results to the neighbor-joining method and the maximum likelihood method. Secondly, when ThreeTree algorithm fails to cluster the correct pair of OTUs in a step finding neighbors, multifurcated trees are often generated such that part of the shortest branches is not resolved. Especially, when extreme parallel substitutions do not occur in a given sequence set, ThreeTree method almost always generates such multifurcated trees at worst. When we compare results of the ThreeTree method and those of the other methods to examine reliability of obtained topologies, this characteristic will be very useful.

On the other hand, this method obviously leads to wrong results when pure additiveness is not recovered in a given distance matrix because of extreme parallel substitutions. The reconstruction of additiveness in this case is the remained issue in future.

Acknowledgements

I thank Dr. Naruya Saitou for his insightful advice to develop the ThreeTree method. This work was supported by grants of Research Institute For Advanced Information Technology, Japan and grants-in-aid of Ministry of Education, Science, Sports and Culture, Japan to N. S.

References

[1] Fitch, W. M. and Margoliash, E., Construction of phylogenetic trees, *Science*, 155:279–284, 1967.

[2] Saitou, N. and Nei, M., The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 4:406–425, 1987.

[3] Felsenstein, J., Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.*, 17:368–376, 1981.