An Integrated System for Phylogenetic Inference using Evolutionary Algorithms

Oclair Prado
Fundação CPqD
Campinas, SP, Brazil 13088-902

Fernando J. Von Zuben
DCA/FEEC/Unicamp
Campinas, SP, Brazil 13083-970

Abstract

This paper presents all the steps to reconstruct phylogenetic trees using an evolutionary algorithm. The fitness criterion is based on maximum likelihood and a toolbox is available. The codification, genetic operators and the optimization procedures involved are clearly described to guarantee reproducibility.

1 INTRODUCTION

Construction of phylogenetic trees is one of the most important problems in evolutionary research. According to Yoshikawa *et al.* (1999), a phylogenetic tree is a tree-structured graph that represents the evolutionary process of genes, and is constructed from sequential data (such as DNA sequences) obtained from several organisms, that will represent the leaves of the tree.

A difficulty here is the lack of information, because we do not have data from the common ancestors, and they must be inferred from the analysis of the current organisms.

Finding a good tree is an NP-Complete problem (Day, 1987), and the number of candidate trees may be calculated using the following formula:

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \qquad (1)$$

2 RESULTS AND DISCUSSION

Using DNA mitochondrial sequences of human, chimpanzee, gorilla, orangutan, and gibbon in the case of 5 leaves (Weir, 1996), Table 1 shows some comparative results obtained from three toolboxes developed to reconstruct trees: our software, called Phylogenetic Tree Project (PTP) (Prado & Von Zuben, 2001), PAML (Yang, 2000) and Phylip (Felsenstein, 1990). The results are close, but not equal, and it is expected to be this way, because each tool has its own features.

Since Maximum Likelihood method tries to maximize the probability of the data given a tree (Nei & Kumar, 2000), the best results in Table 1 are the ones obtained with Phylip. Phylip uses unrooted trees, so that less branches are involved. The final likelihood value is strongly affected by the number of branches. Anyway, the purpose here is not to obtain better results, when compared to other available toolboxes, but to attest the correct implementation of an evolutionary methodology (the reconstructed phylogenetic trees are almost identical), providing all the steps necessary to reproduce the results, what is not possible based on similar approaches in the literature (Matsuda, 1996).

Table 1: Comparative results using PTP, PAML and PHYLIP

TOOL	4 leaves	5 leaves
PHYLIP	-130.74914	-163.87052
PTP	-136.14495	-171.11214
PAML	-136.19721	-174.34676

References

Day, W.H.E, *Computational complexity of inferring phylogenies from dissimilarity matrice*, Bull. Math. Biol, 49:461-467, 1987.

Felsenstein, J. *PHYLIP Manual Version 3.3* University Herbarium, University of California, Berkeley, 1990.

Yoshikawa, T., Tabe, T., Kishinami, R., Matsuda, H. & Hashimoto, A. *On the Implementation of a Phylogenetic Tree Database*, Proc. of IEEE Pacific Conference on Communications, Computers, and Signal Processing, pp.42-45, August, 1999.

Matsuda, H. *Protein phylogenetic inference using maximum likelihood with a genetic algorithm*, Pacific Symposium on Biocomputing. World Scientific, London, pp. 512-523, 1996.

Nei, M. & Kumar, S. *Moledular Evolution and Phylogenetics*, Oxford University press, 2000.

Prado, O. & Von Zuben, F.J. The Phylogenetic Tree Project (PTP). Toolbox available at ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/oclair/. 2001.

Weir, B.S. *Genetic Data Analysis* II, Sinauer, Sunderland, MA, 1996.

Yang, Z. Phylogenetic *analysis by maximum likelihood (PAML)*, version 3.0. University College London, London, England, 2000.