Molecular Phylogenetics and Evolution 109 (2017) 409–414

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

# A modification of the PHYLIP program: A solution for the redundant cluster problem, and an implementation of an automatic bootstrapping on trees inferred from original data

Makoto K. Shimada *, Tsunetoshi Nishida

*Institute for Comprehensive Medical Science, Fujita Health University, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan*

ARTICLE INFO

*Article history:*
Received 23 December 2016
Accepted 15 February 2017
Available online 20 February 2017

*Keywords:*
PHYLIP
Bootstrap analysis
Originally inferred tree
Hash table
Floating point numbers

ABSTRACT

Felsenstein's PHYLIP package of molecular phylogeny tools has been used globally since 1980. The programs are receiving renewed attention because of their character-based user interface, which has the advantage of being scriptable for use with large-scale data studies based on super-computers or massively parallel computing clusters. However, occasionally we found, the PHYLIP Consense program output text file displays two or more divided bootstrap values for the same cluster in its result table, and when this happens the output Newick tree file incorrectly assigns only the last value to that cluster that disturbs correct estimation of a consensus tree. We ascertained the cause of this aberrant behavior in the bootstrapping calculation. Our rewrite of the Consense program source code outputs bootstrap values, without redundancy, in its result table, and a Newick tree file with appropriate, corresponding bootstrap values. Furthermore, we developed an add-on program and shell script, add_bootstrap.pl and fasta2tre_bs.bsh, to generate a Newick tree containing the topology and branch lengths inferred from the original data along with valid bootstrap values, and to actualize the automated inference of a phylogenetic tree containing the originally inferred topology and branch lengths with bootstrap values, from multiple unaligned sequences, respectively. These programs can be downloaded at: https://github.com/ShimadaMK/PHYLIP_enhance/.



## 1. Introduction

The PHYLIP package is an open source computing toolkit for molecular phylogenetic study that has stood the test of time (Felsenstein, 1989). It has received high evaluations ever since being introduced in October 1980, because of the vast scope of its applications in phylogenetic tree inference. PHYLIP uses a character-based user interface (CUI), which naïve users may find difficult, but is suitable for scripting sequential jobs in queue submission systems for super-computers and massively parallel computing clusters. Because of these advantages, PHYLIP package programs have been implemented in, and modified for, many novel research systems (e.g., Salazar et al., 2015). Regardless, PHYLIP remains with room for improvement, particularly in its bootstrap analysis implementation.

We have improved PHYLIP bootstrap analysis by addressing the following two problems:

(1) The Consense program generates an output text file (outfile by default; hereafter Consense outfile, C-of) that displays calculated bootstrap values in a result table that can be quite confusing (Fig. 1). The problem is this table occasionally displays clusters (combinations of Operational Taxonomic Units, OTUs) involved in the bootstrap calculations redundantly, with each duplicate representing only part of that cluster's total bootstrap value (Redundant Description of Bootstrap Results). This phenomenon occurs more often as the number of clusters increases. When it does happen, users must recognize the aberrant behavior and manually sum the separate bootstrap values of the redundant clusters shown, to know the true bootstrap value for the affected cluster (red numbers in Fig. 1). Worse yet, the other Consense output file, a tree file in Newick format (Felsenstein, 1986) (outtree by default), shows only one of the redundant entries, when this problem occurs. Therefore, in these cases,

* Corresponding author.
*E-mail address:* mshimada@fujita-hu.ac.jp (M.K. Shimada).

http://dx.doi.org/10.1016/j.ympev.2017.02.012
1055-7903/

(1) Right result of 'consense' program (After this rewriting program)

1-a: 'outfile' obtained by rewrote 'consense' program

| | Set of clustering | Bootstrap value |
|---|---|---|
| OTU name | ABHDJGEFIC | (How many times out of 9) |
| Pattern of clustering | .......**. | 9.00 |
| | ..******** | 9.00 |
| | ..****.*** | 6.00 |
| | ..***..... | 6.00 |
| | ..***....* | 6.00 |
| | ..*.*..... | 4.00 |
| | ..***..*** | 2.00 |
| | .********* | 9.00 |

1-b: 'outtree' obtained by rewrote 'consense' program

```
((((((C,((H,J):4.00,D):6.00):6.00,(F,I):9.00):2.00,G):6.00,E):9.00,B):9.00,A);
```

(2) Wrong result of 'consense' program (Before this rewriting program)

2-a: 'outfile' obtained by original 'consense' program

| | Set of clustering | Bootstrap value |
|---|---|---|
| OTU name | ABHDJGEFIC | (How many times out of 9) |
| Pattern of clustering | .......**. | 9.00 |
| | ..****.*** | 6.00 |
| | ..***..... | 6.00 |
| | ..***....* | 6.00 |
| | ..*.*..... | 4.00 |
| | ..******** | 4.00 |
| | ..******** | 3.00 |
| | ..******** | 2.00 |
| | ..***..*** | 2.00 |
| | .********* | 9.00 |

2-b: 'outtree' obtained by original 'consense' program

```
((((((C,((H,J):4.00,D):6.00):6.00,(F,I):9.00):2.00,G):6.00,E):2.00,B):9.00,A);
```

**Fig. 1.** Examples of results after (1) and before (2) our conc.c program rewrite using an oversimplified example of nine bootstrap replicates. For example, the cluster containing all operational taxonomic units (OTUs) excluding OTU A and OTU B (..********) appears at the second line with 9.00 out of 9.00 (100%) as a bootstrap value (1-a), which is correctly displayed after our conc.c bug fix (1-b). However, in version 3.696 and earlier, this bootstrap value (9.00) is separated into three lines, the sixth through the eighth (2-a). Although both are equivalent (if users notice the problem and sum the three values), the Consense outtree file incorrectly displays that cluster's bootstrap value by using only one of the three individual bootstrap values (in this case 2.00, i.e. 22%), which disturb the inference of the consensus tree.

not all of the bootstrap values displayed in the consensus tree (outtree) are true. Furthermore, we confirmed that the estimation of the consensus tree in outtree files are disturbed by this problem. Although we noticed this phenomenon in earlier versions of the package released in the 1990's, we only recently discovered the cause of the problem in PHYLIP version 3.695 (http://evolution.gs.washington.edu/phylip.html) and solved it. Our solution is reported below in Section 2.2.

(2) PHYLIP does not have an easy way to generate a tree inferred from the original sequence data (Originally Inferred Tree, OIT) that includes bootstrap values generated from bootstrapped data. We created a program that enables both attributes to be seen with one tree.

When Felsenstein introduced the bootstrap method (Efron, 1979) to phylogenetic study (Felsenstein, 1985), he was interested in how to statistically evaluate an inferred tree, in other words, how close is an inferred tree to the true tree, and how should that 'closeness' be measured. He argued that trees obtained from a series of bootstrapped resampled data were suitable for producing a majority-rule consensus tree that has the advantage of being closer to the true tree than the OIT (Berry and Gascuel, 1996). PHYLIP was initially designed within this context, and provides bootstrap results that can be visualized as a majority-rule consensus tree. Accordingly, PHYLIP cannot generate a Newick tree file that shows a tree inferred from the original sequence data with those estimated branch lengths, along with bootstrap values for those clusters (Lack of OIT File with Bootstrap). However, users may want to know the reliability of interior branches of an OIT using bootstrap values, for example, for Dopazo's (1994) bootstrap interior branch test, which is different from the purpose of development by Felsenstein (Nei and Kumar, 2000). Both types of trees with bootstrap values are available in the MEGA package (Kumar et al., 2016; Kumar et al., 2001) To compensate for this, we created a PHYLIP compatible, add-on program that outputs a Newick tree format file containing topology and branch lengths based on an OIT, as well as those bootstrap values that calculated based on the counts in bootstrap sampling trees (see Section 3.1 for an example of the format).