Submitted to ISBRA 2018 Conference, June 8 - 11, 2018.

Imputing missing distances in molecular phylogenetics

Xuhua Xia

University of Ottawa, Ottawa, Ontario K1N 6N5 Canada
xxia@uottawa.ca

Abstract. Missing data are frequently encountered in molecular phylogenetics and need to be imputed. For a distance matrix with missing distances, the least-squares approach is often used for imputing the missing values. Here I develop a method, similar to the expectation-maximization algorithm, to impute multiple missing distance in a distance matrix. I show that, for inferring the best tree and missing distances, the minimum evolution criterion is not as desirable as the least-squares criterion. I also discuss the problem involving cases where the missing values cannot be uniquely determined, e.g., when a missing distance involve two sister taxa. The new method has the advantage over the existing one in that it does not assume a molecular clock. I have implemented the function in DAMBE software which is freely available at available at http://dambe.bio.uottawa.ca

Keywords: phylogenetics, distance matrix, imputing missing distance, least-squares method.

1 Introduction

The demand to reconstruct supertrees with a large number of taxa from concatenating a large number of genes often results in missing data illustrated in Fig. 1 where a distance between Sp3 and Sp4 cannot be computed because they share no homologous sites. While missing data can be accommodated by the likelihood method with the pruning algorithm [1, 2, 3, pp. 253-255], they can inflate branch lengths and introduce phylogenetic bias [4, 5]. Some popular likelihood-based phylogenetic methods, e.g., PhyML [6], use distance-based methods to build the initial phylogenetic tree, which is then modified in various ways and evaluated in the likelihood framework to find the maximum likelihood tree. Distance-based methods are much faster than other phylogenetic methods such as maximum likelihood, Bayesian inference and maximum parsimony. Consequently, they are used frequently constructing supertrees [7].

```
                Gene A                               Gene B
Sp1 CCGTTA...ACGGCTTTGCCGACGAC | ATCAGACGATGCG...AUGACGACTCACGATA
Sp2 CCGTCA...ACGACTTTGCCGACGAC | ACCAGACGATGCA...ACGACAACTTACGATA
Sp3 CCATTA...ACGGCTTTGCCGACGAC | ????????????????????????????????
Sp4 ?????????????????????????? | ATCGGGCGACGCG...ACGACGACTCACGATA
Sp5 CTGTTA...ACGGCTTTGCCGACGAC | ATCAGACGATGCG...ACGGCGACTTACGATA
```

Fig. 1. A sequence data set from concatenating Gene A and Gene B sequences. A distance cannot be computed between Sp3 and Sp4 because they share no homologous sites.

The least-squares (LS) method is frequently used for imputing the missing distances. The general conceptual framework is to estimate the missing distances given different trees, and then choose the best tree and the best estimates of missing distances with a specific criterion. One of its key advantages of the LS method lies in its optimization with constraints that restrict branch length to be non-negative. I will first outline the general approach, point out problematic cases where unique solution cannot be found, and then develop an efficient computational method similar to the expectation-maximization (EM) algorithm to impute the missing distance. I illustrate the method by applying it to real data.

2 Least-square method for imputing missing distances

Suppose we have four species (S1 to S4 in Fig. 2) with D_{12} = 2, D_{14} = 5, D_{23} = 3, D_{24} = 5, D_{34} =4 but with D_{13} missing. The method developed can also be used to impute multiple missing distances.

2.1 A wrong approach

One may take a wrong approach by thinking that, in this particular case, we have five unknowns and five equations and can solve for D_{13} exactly. For example, given a topology in Fig. 2A, we can write the expected D_{ij} values, i.e., $E(D_{ij})$, as:

$$\begin{aligned}
E(D_{12}) &= x_1 + x_2 \\
E(D_{14}) &= x_1 + x_5 + x_4 \\
E(D_{23}) &= x_2 + x_5 + x_3 \\
E(D_{24}) &= x_2 + x_5 + x_4 \\
E(D_{34}) &= x_3 + x_4
\end{aligned} \tag{1}$$

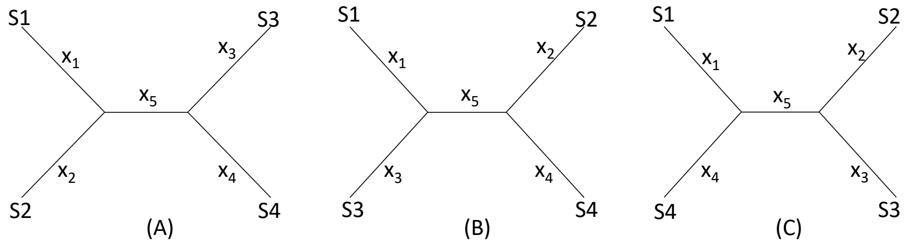


Fig. 2. *Topologies for illustrating the distance-based methods.*

These $E(D_{ij})$ values are termed patristic distances in phylogenetics. If we replace $E(D_{ij})$ by the observed D_{ij} values, we can indeed solve the simultaneous equations in Eq. (1), which give the solution as

$$\begin{aligned}
x_1 &= \frac{D_{12}}{2} + \frac{D_{14}}{2} - \frac{D_{24}}{2} \\
x_2 &= \frac{D_{12}}{2} + \frac{D_{24}}{2} - \frac{D_{14}}{2} \\
x_3 &= \frac{D_{23}}{2} + \frac{D_{34}}{2} - \frac{D_{24}}{2} \\
x_4 &= \frac{D_{34}}{2} + \frac{D_{24}}{2} - \frac{D_{23}}{2} \\
x_5 &= \frac{D_{14}}{2} + \frac{D_{23}}{2} - \frac{D_{12}}{2} - \frac{D_{34}}{2}
\end{aligned} \qquad (2)$$

The missing D_{13} given the tree in Fig. 2A, designated as $D_{13.A}$, can therefore be inferred, as:

$$D_{13.A} = x_1 + x_5 + x_3 = D_{14} + D_{23} - D_{24} \qquad (3)$$

Thus, given the five known D_{ij} values above, we can obtain $x_1 = x_2 = x_3 = x_5 = 1$, $x_4 = 3$, $D_{13.A} = 3$. The tree length (TL), defined as TL = $\sum x_i$, is 7 for the tree in Fig. 2A, i.e., $TL_A = 7$.

One might think of applying the same approach to the other two trees in Fig. 2B,C to obtain $D_{13.B}$ and $D_{13.C}$ as well as TL_B and TL_C, and choose as the best D_{13} and the best tree by using the minimum evolution criterion [8, 9], i.e., the tree with the shortest TL.

This approach has two problems. First, the approach fails with the tree in Fig. 2B where the missing distance, D_{13}, involves two sister species. One can still write down five simultaneous equations, but will find no solutions for x_i, given the D_{ij} values above, because the determinant of the coefficient matrix is 0. For the tree is Fig. 2C, the solution will have $x_5 = -1$. A negative branch length is biologically undesirable and defeats the ME criterion for choosing the best tree and the associated estimate of D_{13}. Second, in most practical cases where missing distances are imputed, there are more equations than unknowns, e.g., if we have five or more species with one missing distance.

2.2 Least-squares approach

Take $E(D_{ij})$ specifications in Eq. (1) for the tree in Fig. 2A, the LS approach find D_{13} and the best tree that minimize the residual sum of squared deviation (RSS) between the observed D_{ij} and the expected $E(D_{ij})$:

$$RSS = \sum \frac{\left[D_{ij} - E(D_{ij})\right]^2}{D_{ij}^m} \qquad (4)$$

where m is typically 0 (ordinary least-squares, OLS), 1, or 2. In the illustration below, we will take the OLS approach with m = 0. It has been shown before that OLS actually exhibits less topological bias than alternatives with m equal to 1 or 2 [10]

Given the three tree topology, the results from the LS estimation are summarized in Table 1. Note that, for the tree in Fig. 2B, there are multiple sets of solutions of x_i that can achieve the minimum RSS of 1.

Table 1. Estimation results from minimizing RSS, with Trees A, B, and C as in Fig. 2, and with the constraint of no negative branch lengths.

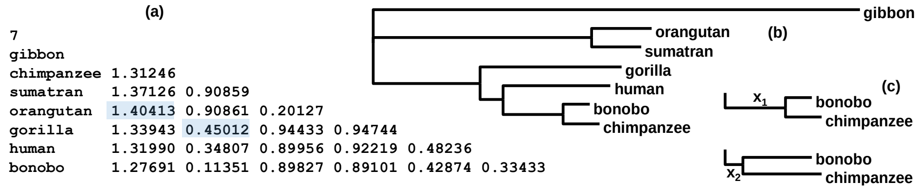
Site	Tree A	Tree B	Tree C
x_1	1	0	1
x_2	1	1.5	1.5
x_3	1	0	1
x_4	3	3.5	3.5
x_5	1	1	0
D_{13}	3	0	2
TL	7	6	7
RSS	0	1	1

We see a conflict between the LS criterion and the ME criterion in choosing the best tree and the best estimate of D_{13}. The ME criterion would have chosen Tree B with TL_B = 6 and D_{13} = 0 because TL_B is the smallest of the three TL values. In contrast, the LS criterion would have chosen Tree A with RSS = 0 and D_{13} = 3. There is no strong statistical rationale for the ME criterion which is based on the assumption that substitutions are typically rare in evolution. The ME criterion is particularly inappropriate for imputing missing distances because it tends to underestimate the missing distances. In contrast, the LS-criterion is well-established. I recommend the LS criterion for the simultaneous imputing of missing distances and inferring phylogenetic trees. Phylogeneticists sometimes think that the ME criterion would be appropriate if the branch lengths are not allowed to take negative values [11-13]. The illustrative example in Table 1 shows that, even when we do not allow branch lengths to become negative, there is still problem with the ME criterion.

3 Implementation in DAMBE

An earlier version of DAMBE implemented the LS approach above by using an iterative approach similar to the EM (expectation-maximization) algorithm as follows. For a given distance matrix with missing values, we simply fill in the missing D_{ij} values by guestimates, e.g., the average of the observed distances. These initial D_{ij} guestimates will be designated as $D_{ij,m0}$ where the subscript “m0” indicates missing distances at step 0. We now build a tree from the distance matrix that minimizes RSS in Eq. (4). From the resulting tree we obtain the patristic distances $E(D_{ij})$ from the tree and replace $D_{ij.m0}$ by the corresponding $E(D_{ij})$ values which are now designated as $D_{ij.m1}$. We now again build a tree, obtain the corresponding $E(D_{ij})$ to replace $D_{ij.m1}$ so now we have $D_{ij.m2}$. We repeat this process until RSS does not decrease any further. This process can quickly arrive at a local minimum. Unfortunately, different topologies have different minimums, and this approach is too often locked in a local minimum with a tree that does not achieve a global minimum RSS.

New version of DAMBE (since version 7) uses a downhill simplex method in multidimensions {Press, 1992 #26508, pp. 408-412} multiple times (with different initial values for the points in the simplex) to increase the chance of finding the global RSS associated with the missing distances and the tree. When there is a single missing distance, then the Brent's method {Press, 1992 #26508, pp. 402-408} is used. Fig. 3 shows an illustrative example. The distance matrix in Fig. 3a is computed from aligned sequence data used before [14]. Fig. 3b is the phylogenetic tree built from this distance matrix. Suppose $D_{\text{gibbon,orangutan}}$ and $D_{\text{gorrila.chimpazee}}$ are missing (shaded in Fig. 3a) and need to be imputed. DAMBE yields $D_{\text{gibbon,orangutan}} = 1.3776$ and $D_{\text{gorrila.chimpazee}} = 0.4600$, which are close to the observed values (Fig. 3a). The final tree built from the distance matrix with the two missing distances is identical to Fig. 3b except for a negligible difference in branch lengths.



```
7
gibbon
chimpanzee 1.31246
sumatran   1.37126 0.90859
orangutan  1.40413 0.90861 0.20127
gorilla    1.33943 0.45012 0.94433 0.94744
human      1.31990 0.34807 0.89956 0.92219 0.48236
bonobo     1.27691 0.11351 0.89827 0.89101 0.42874 0.33433
```

Fig. 3. An example data set for imputing missing distances. (a) A real distance matrix computed from aligned sequences, but we pretend that the two shaded distances are missing. (b) A phylogenetic tree from the distance matrix. (c) A special case illustrating the problem of the estimation.

One can access the function in DAMBE [15, 16] by clicking 'File|Open other molecular data|Distance matrix file with missing value', and open a distance file in the format in Fig. 3a with missing distances represented by '.' (a period without quotation marks).

There are cases where missing distances cannot be uniquely determined. For example, when the missing distance is for two sister taxa (e.g., the two chimpanzee species, designated bonobo and chimpanzee in Fig. 3b), we can find a minimum RSS but the solution for the missing distance $D_{\text{bonobo,chimpanzee}}$ is not unique. That is, multiple $D_{\text{bonobo,chimpanzee}}$ values can generate the same RSS. Note that the patristic distances $E(D_{\text{bonobo.i}})$ and $E(D_{\text{chimpanzee.i}})$ where i stands for other species, do not depend on the branch length leading to the common ancestor from bonobo and chimpanzee. This branch can be as short as x_2 or as long as x_1 (Fig. 3c) but RSS in Eq. **Error! Reference source not found.** will remain the same. Thus, in this particular case, a missing $D_{\text{bonobo,chimpanzee}}$ cannot be determined uniquely. The only way to eliminate this problem is to have a more closely related species to break up the sister relationship so that the missing D_{ij} is not between two sister taxa.

The method in the paper has an advantage over a previous method [7] that assumes a rooted tree and a molecular clock for building a tree and for inferring missing distances. This assumption is not needed and is too restrictive in practise.

4 Software availability

DAMBE is available free at http://dambe.bio.uottawa.ca. One can access the function by clicking ‘File|Open other molecular data|Distance matrix file with missing value’, and open a distance matrix file in the format in Fig. 3a, with missing distances represented by ‘.’ (a period without quotation marks).

5 Acknowledgement

This study is funded by the Discovery Grant from Natural Science and Engineering Research Council (NSERC, RGPIN/261252) of Canada.

6 Conflict of interest

I declare no competing interest.

References

1. Felsenstein, J.: Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22, 240-249 (1973)
2. Felsenstein, J.: Inferring phylogenies. Sinauer, Sunderland, Massachusetts (2004)
3. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368-376 (1981)
4. Darriba, D., Weiss, M., Stamatakis, A.: Prediction of missing sequences and branch lengths in phylogenomic data. Bioinformatics 32, 1331-1337 (2016)
5. Xia, X.: Phylogenetic Bias in the Likelihood Method Caused by Missing Data Coupled with Among-Site Rate Variation: An Analytical Approach. In: Basu, M., Pan, Y., Wang, J. (eds.) Bioinformatics Research and Applications., pp. 12-23. Springer (2014)
6. Guindon, S., Gascuel, O.: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696-704. (2003)
7. Kettleborough, G., Dicks, J., Roberts, I.N., Huber, K.T.: Reconstructing (super)trees from data sets with missing distances: not all is lost. Mol Biol Evol 32, 1628-1642 (2015)
8. Rzhetsky, A., Nei, M.: A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9, 945-967 (1992)
9. Rzhetsky, A., Nei, M.: METREE: a program package for inferring and testing minimum-evolution trees. CABIO 10, 409-412 (1994)

10. Xia, X.: Topological bias in distance-based phylogenetic methods: problems with over- and underestimated genetic distances. Evolutionary Bioinformatics 2, 375–387 (2006)
11. Felsenstein, J.: An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol 46, 101-111 (1997)
12. Desper, R., Gascuel, O.: Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9, 687-705. (2002)
13. Desper, R., Gascuel, O.: Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21, 587-598 (2004)
14. Xia, X., Yang, Q.: A distance-based least-square method for dating speciation events. Mol Phylogenet Evol 59, 342-353 (2011)
15. Xia, X.: DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30, 1720-1728 (2013)
16. Xia, X.: DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. J Hered 108, 431-437 (2017)