DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution

Xuhua Xia*[,1,2]

[1]Department of Biology, University of Ottawa, Ottawa, ON, Canada

[2]Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON, Canada

***Corresponding author:** E-mail: xxia@uottawa.ca.

Associate editor: Sudhir Kumar

Abstract

DAMBE is a comprehensive software package for genomic and phylogenetic data analysis on Windows, Linux, and Macintosh computers. New functions include imputing missing distances and phylogeny simultaneously (paving the way to build large phage and transposon trees), new bootstrapping/jackknifing methods for PhyPA (phylogenetics from pairwise alignments), and an improved function for fast and accurate estimation of the shape parameter of the gamma distribution for fitting rate heterogeneity over sites. Previous method corrects multiple hits for each site independently. DAMBE's new method uses all sites simultaneously for correction. DAMBE, featuring a user-friendly graphic interface, is freely available from http://dambe.bio.uottawa.ca (last accessed, April 17, 2018).

Key words: **phylogenetics, bioinformatics, missing distance imputation, rate heterogeneity over sites.**

Brief Communication

DAMBE is for descriptive and comparative sequence analysis (Xia 2013, 2017) featuring a graphic, user-friendly, and intuitive interface, and available free for Windows, Linux, and Macintosh computers at dambe.bio.uottawa.ca. DAMBE7 represents a major upgrade with many new functions including new sets of significance tests for position weight matrix and Gibbs sampler for de novo characterization of sequence motifs. I outline three functions most relevant to molecular evolution and phylogenetics. A supplemental file (Using_New_Functions.docx) is included in Supplementary Material online.

Imputing Missing Distance and Phylogeny Simultaneously

This function is implemented for building large trees of phages which often 1) are too diverged to build a multiple sequence alignment (MSA), and 2) do no share homologous genes/sites (e.g., S3 and S4 in fig. 1*a*). This is also true for many transposons from which one cannot get a meaningful MSA, and researchers are limited to align the sequences against the consensus (Gallus et al. 2015). One can do pairwise alignment among most of the sequences and compute their distances, but some sequence pairs do not share homologous sites and need to have their distances imputed from those computable distances. This allows one to build trees and likely will revolutionize phage taxonomy which is not based on phylogeny.

This distance-imputation function is currently missing. MEGA (Kumar et al. 2016) does not impute missing distances, neither does PHYLIP's DNADIST (Felsenstein 2014). Fitch and Kitsch programs can estimate missing distances if a user tree is provided.

For a distance matrix with N missing distances (parameters), DAMBE searches the tree space and parameter space to find a tree with the N parameters that minimizes

$$RSS = \sum \frac{\left[D_{ij} - E\left(D_{ij}\right)\right]^2}{D_{ij}^m} \quad (1)$$

where D_{ij} and $E(D_{ij})$ are the observed and patristic distance, and m is typically 0, 1, or 2. Figure 1*c* is the phylogenetic tree reconstructed from the distance matrix in figure 1*b* with two shaded distances missing.

For sequences such as that in figure 1*a*, DAMBE will compute all computable distances and impute the missing distances. When bootstrapping/jackknifing is used, distance imputation and phylogeny inference are done for each resampled data. One may also have unaligned sequence data and use PhyPA (Xia 2016) to build phylogenetic trees and obtain bootstrap/jackknife support.

There are cases where a unique solution cannot be obtained. For example, when a missing distance is for two sister taxa (e.g., bonobo and chimpanzee in fig. 1*b* and *c*), we can find minimum RSS but the solution for missing D_{ij} is not unique, with different values for missing D_{ij} resulting in the same minimum RSS. The patristic distances $D_{p.bonobo.i}$ and $D_{p.chimpanzee.i}$, where i stands for other species, do not change when x_1 changes to x_2 (fig. 1*d*), so $D_{p.bonobo.i}$ and $D_{p.chimpanzee.i}$ will remain the same, and so does RSS in equation (1). DAMBE use the midpoint distance in such cases.

Bootstrap/Jackknife Support for PhyPA

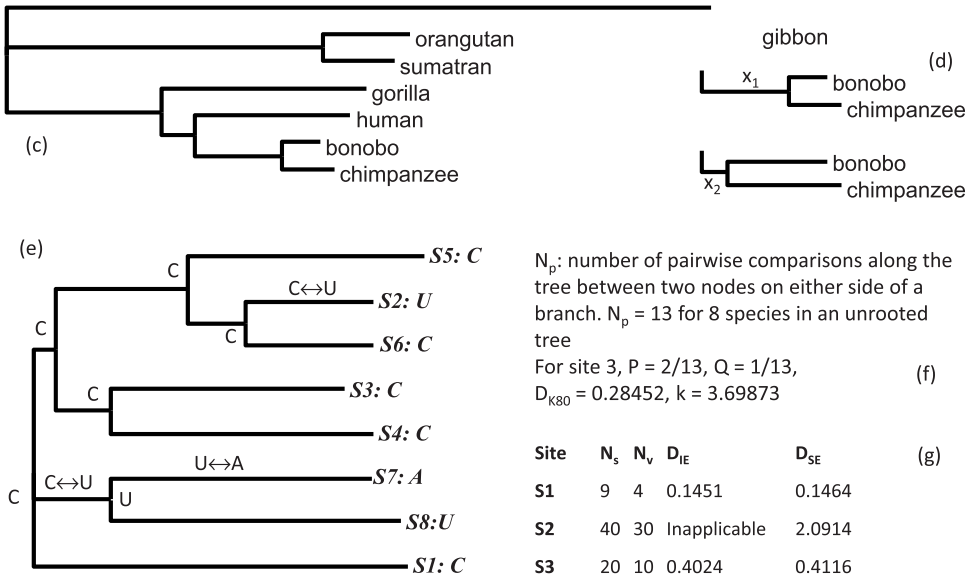
For each pair of sequences, we can obtain a vector **S** of 10 N_{ij} values (number of pairs with nucleotide i in one sequence and j

```
 Gene A                          Gene B
S1 CCGTTA...ACGGCTTTGCCGACGAC ATCAGACGATGCG...AUGACGACTCACGATA
S2 CCGTCA...ACGACTTTGCCGACGAC ACCAGACGATGCA...ACGACAACTTACGATA
S3 CCATTA...ACGGCTTTGCCGACGAC ????????????????????????????????
S4 ???????????????????????????? ATCGGGCGACGCG...ACGACGACTCACGATA
S5 CTGTTA...ACGGCTTTGCCGACGAC ATCAGACGATGCG...ACGGCGACTTACGATA
```

(a)

```
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
```

(b)



N_p: number of pairwise comparisons along the tree between two nodes on either side of a branch. N_p = 13 for 8 species in an unrooted tree

For site 3, P = 2/13, Q = 1/13, D_{K80} = 0.28452, k = 3.69873

(f)

Site	N_s	N_v	D_{IE}	D_{SE}
S1	9	4	0.1451	0.1464
S2	40	30	Inapplicable	2.0914
S3	20	10	0.4024	0.4116

(g)

Fig. 1. Illustration of distance imputation and estimation of the shape parameter in gamma distribution. (*a*) A sequence data set with S3 and S4 sharing no homologous sites to estimate distance. (*b*) Distance matrix with two shaded distance missing. (*c*) Tree reconstructed from the distance matrix in (*b*). (*d*) A case with nonunique solution for a missing distance between bonobo and chimpanzee. (*e*) Tree reconstructed from a multiple alignment with one site mapped to the leaves, together with one of several possible reconstruction of internal nodes. (*f*) Counting changes between neighboring nodes and correction for multiple hits. (*g*) Transitions and transversions at three sites illustrating independently estimated distance (D_{IE}) and simultaneously estimated distance (D_{SE}).

in another). With 10 sequences and 45 pairwise comparisons, there are 45 **S** vectors from which we can compute the 45 pairwise distances. For bootstrapping/jackknifing, we simply resample each pair to generate an **S** vector and use the 45 **S** vectors to produce a new set of 45 pairwise distances from which a tree can be reconstructed. This function complements the function of phylogenetics with imputed missing distances.

An Improved Method for Estimating the Shape Parameter of Gamma Distribution

Substitution rate varies over sites and is particularly pronounced in protein-coding genes (Xia 1998). The method by Gu and Zhang (1997) uses the following probability density function (Johnson and Kotz 1969) to estimate α:

$$f(k) = \frac{\Gamma(\alpha + k)}{\Gamma(k+1)\ \Gamma(\alpha)} \left(\frac{\bar{k}}{\bar{k}+\alpha}\right)^k \left(\frac{\alpha}{\bar{k}+\alpha}\right)^\alpha \quad (2)$$

where k, instead of being integers, is replaced by the estimated number of substitutions per site, and k is mean k.

The method's accuracy depends on the accuracy of the estimated k which comes from a multiple alignment in two steps (fig. 1*e* and *f*): 1) construct a phylogenetic tree from the aligned sequences and reconstruct ancestral sequences at internal nodes (fig. 1*e*, showing one of several possible reconstructions for one site with nucleotides mapped to the leaves), and 2) perform pairwise comparisons between two nodes on each side of a branch to obtain observed number of substitutions per site, and apply correction for multiple hits to get k (fig. 1*f*). DAMBE improves this estimation in two ways. First, it uses simultaneous estimation (SE). Take the K80 model for example. At each site,

$$E(P_i) = \frac{1}{4} + \frac{1}{4}e^{-\frac{4D_i}{\kappa+2}} - \frac{1}{2}e^{-\frac{2D_i(\kappa+1)}{\kappa+2}} \quad (3)$$

$$E(Q_i) = \frac{1}{2} - \frac{1}{2}e^{-\frac{4D_i}{\kappa+2}} \quad (4)$$

where D_i is K80 distance and κ is the transition/transversion ratio, not to confuse with k in equation (2) which is the

estimated number of substitution for a site. Applying equations (3) and (4) to data from the three sites (fig. 1g) independently will generate one inapplicable case for site 2 (under D_{IE} in fig. 1g, with IE for independent estimation). We can estimate all D_i and κ simultaneously by maximizing the following log-likelihood:

$$lnL = \sum_{i=1}^{N} \{N_{s.i} \ln[E(P_i)] + N_{v.i} ln[E(Q_i)] + N_{I.i} ln[1 - E(P_i) - E(Q_i)]\} \quad (5)$$

where N is the number of sites, $N_{s.i}$ and $N_{v.i}$ and $N_{I.i}$ are recorded number of transitional, transversional difference and no difference from pairwise comparisons along the tree between nodes on each side of each branch at site i. SE generates no inapplicable cases (D_{SE} in fig. 1g) and leads to the second improvement in using more realistic models such as F84 or TN93 instead of the K80 correction in GZ-gamma (Gu and Zhang 1997). SE distance is used in MEGA (Tamura et al. 2004) and DAMBE (Xia 2009) which includes MLCompositeF84 and MLCompositeTN93 for F84 and TN93 models, respectively, but has never been used in estimating the shape parameter.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by Discovery Grant RGPIN-2018-03878 from Natural Science and Engineering Research Council of Canada.

Literature Cited

Felsenstein J. 2014. PHYLIP 3.695 (phylogeny inference package). Seattle: Department of Genetics, University of Washington.

Gallus S, Hallström BM, Kumar V, Dodt WG, Janke A, Schumann GG, Nilsson MA. 2015. Evolutionary histories of transposable elements in the genome of the largest living marsupial carnivore, the Tasmanian devil. *Mol Biol Evol.* 32(5):1268–1283.

Gu X, Zhang J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Mol Biol Evol.* 14(11):1106–1113.

Johnson NL, Kotz S. 1969. Discrete distributions. Boston: Houghton Mifflin.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol.* 33(7):1870–1874.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A.* 101(30):11030–11035.

Xia X. 1998. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. *Mol Biol Evol.* 15(3):336–344.

Xia X. 2009. Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. *Mol Phylogenet Evol.* 52(3):665–676.

Xia X. 2013. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol.* 30(7):1720–1728.

Xia X. 2016. PhyPA: phylogenetic method with pairwise sequence alignment outperforms likelihood methods in phylogenetics involving highly diverged sequences. *Mol Phylogenet Evol.* 102:331–343.

Xia X. 2017. DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. *J Hered.* 108(4):431–437.