From the Department of Applied Economics and Statistics 204, University of Nevada, Reno, NV 89557. Address correspondence to George C. J. Fernandez at the address above.

References

Gabriel KR, 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58:453–467.

Fernandez GCJ, 1991. Analysis of genotype × environment interaction by stability estimates. HortScience 26: 947–950.

Fernandez GCJ, 2000. Quick results from statistical analysis (visited/last modified August 16, 2000). http://www.ag.unr.edu/gf.

Shafii B, Mahler KA, Price WJ, and Auld DL, 1992. Genotype by environment interaction effects on winter rapeseed yield and oil content. Crop Sci 32:922–927.

Shafii B and Price WJ, 1998. Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. J Agric Biol Environ Stat 3:335–345. http://www.uidaho.edu/ag/statprog/ammi/.

Tai GCC, 1971. Genotypic stability analysis and its application to potato regional trials. Crop Sci 11:184–190.

Zobel RW, Wright MJ, and Gauch HG, 1988. Statistical analysis of a yield trial. Agron J 80:388–393.

Received September 7, 2000
Accepted April 30, 2001

Corresponding Editor: Bruce S. Weir

DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution

X. Xia and Z. Xie

DAMBE (data analysis in molecular biology and evolution) is an integrated software package for converting, manipulating, statistically and graphically describing, and analyzing molecular sequence data with a user-friendly Windows 95/98/2000/NT interface. DAMBE is free and can be downloaded from http://web.hku.hk/~xxia/software/software.htm. The current version is 4.0.36.

DAMBE (data analysis in molecular biology and evolution) is an integrated computer program for descriptive and comparative analysis of molecular data (including nucleotide and amino acid sequence data, as well as allele frequency and distance matrix data). It has features either not available or poorly implemented in other programs. These features are grouped into (1) sequence format conversion and manipulation supporting 20 commonly used molecular data formats; (2) descriptive statistics such as nucleotide, amino acid, and codon frequencies, dinucleotide and diamino acid frequencies, analysis of codon usage and amino acid usage bias; and (3) comparative sequence analysis such as phylogenetic reconstruction of trees and ancestral sequences with distance, maximum parsimony and maximum-likelihood methods, bootstrapping and jackknifing, significance tests of molecular clock and alternative phylogenetic hypotheses (i.e., how much better is the best tree compared to alternative trees), fitting statistical distributions to substitution data over sites including Poisson, negative binomial, and gamma distributions. DAMBE features a user-friendly Windows interface with extensive on-line help.

Table 1. Common data file formats used in DAMBE

Sequence format	Read in	Convert to
PHYLIP	+	+
PAUP	+	+
MEGA	+	+
CLUSTAL	+	–
FASTA	+	+
GenBank	+	+
GCG	+	+
MSF	+	+
DNA strider	+	+
PAML	+	+
RST,MP[a]	+	–
PHYLTEST	–	+
IG/Stanford	+	+
NBRF	+	+
EMBL	+	+
FITCH	+	+
PIR/CODATA	+	+
Plain text[b]	+	+
Allele frequency	+	–
Distance matrix	+	+

[a] Sequence formats for storing original sequences and reconstructed ancestral sequences from the original sequences.

[b] The one-sequence-per-file text format from programs such as Sequence Navigator and DNA Star.

Sequence Input and Sequence Format Conversion

DAMBE can read and convert almost all commonly used molecular data formats (Table 1). In particular, DAMBE can take advantage of the rich information contained in the FEATURES table of GenBank files and extract specific segments such as CDS, exons, introns, rRNA, etc., by a few mouse clicks. The user can also use "custom splicing" to extract sequence segments that are not specified in the GenBank sequences.

DAMBE can read sequence files directly from a networked computer, such as a remote UNIX workstation, in the same way as one would read a file from a local hard disk. Extensive network functions have also been implemented for retrieving sequences directly from GenBank, either by LOCUS name or accession number, or by keyword search.

DAMBE features a color-coded sequence editor for either sequence input or visual alignment. Sequences can be as long as 32,768 bp.

Sequence Manipulation

We will only highlight two of the many sequence manipulation features in DAMBE.

Sequence Alignment

DAMBE can align nucleotide and amino acid sequences as most other alignment programs do. However, one particular feature that is not available in most other alignment programs is the ability to align protein-coding nucleotide sequences against aligned amino acid sequences. Other programs often introduce frameshift indels in the aligned protein-coding sequences, even if the protein genes are known to be functional and do not have these frame-shifting indels. In other words, the introduced frame-shifting indels in the aligned sequences are alignment artifacts, and the correctly aligned sequences should have complete codons, not one or two nucleotides, inserted or deleted.

DAMBE solves this problem by aligning the protein-coding nucleotide sequences against aligned amino acid sequences. One can read in the protein-coding nucleotide sequences, translate them into amino acid sequences, align the amino acid sequences, and then align the original nucleotide sequences against the aligned amino acid sequences.

Translation

DAMBE implements all 12 different genetic codes and can therefore translate protein-coding nucleotide sequences from any organism to amino acid sequences. The implementation of these genetic codes greatly facilitates amino acid-based and codon-based analyses.

Descriptive Sequence Analysis

This includes nucleotide, amino acid and codon usage analysis, compositional analysis based on dinucleotide and diamino acid frequencies, quantification of the effect of GC and T_pA frequencies on exon and CDS lengths, and the methylation effect on codon usage bias.

A substitution model used in compara-

tive sequence analysis, such as phylogenetic reconstruction using the maximum-likelihood method, typically has two categories of parameters, the frequency parameters and the rate ratio parameters. The descriptive sequence analysis helps to understand the factors affecting the frequency parameters and to select which substitution model to use in phylogenetic reconstruction.

Comparative Sequence Analysis

Quantification of Substitution Patterns

A substitution model is characterized by frequency parameters and rate ratio parameters, and it is important to know the empirical substitution patterns in order to decide which substitution model to use in analyzing sequences. The quantification of empirical substitution patterns requires pairwise comparisons. However, when the comparison is done between all possible sequence pairs, the resulting substitution pattern may be biased because the comparisons are not independent (Felsenstein 1992; Nee et al. 1996; Xia et al. 1996). For example, if there is one species that has recently experienced a large number of A→G transitions and few other substitutions, then all pairwise comparisons between this species and the other species will each contribute one data point with a large A→G transition bias. One way to avoid such a problem of nonindependence is to reconstruct ancestral states of DNA sequences and estimate the number of substitutions between neighboring nodes along the phylogenetic tree (Tamura and Nei 1993; Xia 1998; Xia and Li 1998). DAMBE automates this process.

Phylogenetic Reconstruction

DAMBE implements most commonly used phylogenetic methods such as distance-based (UPGMA, Fitch-Margoliash, and neighbor-joining methods), maximum parsimony, and maximum-likelihood methods. A variety of genetic distances are implemented. Nucleotide-based distances include the one-parameter (Jukes and Cantor 1969) and two-parameter (Kimura 1980) distances, the paralinear distance (Lake 1994), as well as distances based on the F84 model (Felsenstein 1993) and TN93 model (Tamura and Nei 1993). Codon-based distances include Li's (1993) synonymous and nonsynonymous distances. The UPGMA and neighbor-joining methods can handle tied values in the matrix and generate all possible alternative trees. Most other distance-based programs output only one of the possible trees.

All phylogenetic analyses for protein-coding genes can be performed on individual codon positions or combinations of codon positions, for example, the first and second codon positions when the third codon position experienced substitution saturation. Alternatively, one can also perform analysis on translated amino acid sequences. DAMBE can translate nucleotide sequences from any organism into amino acid sequences because it implements all known genetic codes.

Phylogenetic Analysis Involving Bootstrapping and Jackknifing

Bootstrapping and delete-half jackknifing are implemented in DAMBE in conjunction with the phylogenetic methods mentioned above. Resampling can be nucleotide based, amino acid based, or codon based. The last is necessary for doing bootstrapping and jackknifing with codon-based methods such as Li's (1993) synonymous and nonsynonymous distances. Consensus trees are displayed with bootstrapping values at internal nodes. The branch lengths of a consensus tree can be evaluated.

Testing Alternative Phylogenetic Hypotheses

It is often necessary to evaluate the relative statistical support for alternative phylogenetic hypotheses such as alternative phylogenetic trees. Such hypothesis tests can be carried out in DAMBE with the distance, maximum parsimony, or maximum-likelihood methods. The significance tests make proper multiple comparisons involving multiple trees.

For the distance methods, the test is similar to that detailed in Xia (2000), except that the following equation

$$\text{var } E = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_{ij} - y_{ij})^2}{n(n-1)/2 - m - 1} \quad (1)$$

replaces equation (21.2) in Xia (2000).

For the maximum parsimony method, a rooted tree is required to represent alternative topologies. DNAPARS in PHYLIP has already provided a significance test if you include user trees in the input file. In short, DNAPARS computes the number of steps (changes in character states) for each topology, the difference in the number of steps between the best and each alternative topology, and the associated (large sample) variance of the differences. The z score is computed and declared as significant if it is larger than 1.96 (Felsenstein 1985). The main problem with this test is that the result can be interpreted probabilistically only when you have just two topologies and is not appropriate with multiple comparisons. DAMBE takes the same approach but uses the Newman–Keuls test that is better for multiple comparisons.

For the maximum-likelihood method, the Kishino–Hasegawa test (Kishino and Hasegawa 1989), which is also called the RELL test, is implemented as in PAML (Yang 2000) from which I have taken part of the code. The Kishino–Hasegawa test, as is practiced in literature, is analogous to the test in DNAPARS mentioned above, except that the test is based on the likelihood values rather than on the number of steps. In short, one calculates the log-likelihood for each topology, the difference in log-likelihood between the best tree and each of the alternative topologies, and the variance of the differences estimated by resampling methods such as bootstrapping. The z score is then calculated and declared as significant if it is larger than 1.96. Again, such interpretation is heuristic and is not appropriate probabilistically if there are more than two topologies being compared. DAMBE does the same computation but uses the Newman–Keuls test which is more appropriate for multiple comparisons.

Phylogenetic Tree Viewing and Manipulation

DAMBE can graph and print publication-quality trees. The tree-displaying window is scrollable and therefore can accommodate very large trees. The displayed tree can also be copied to presentation programs such as Microsoft PowerPoint.

Fitting Statistical Distributions to Substitutions Over Sites

It is important to know if the substitution rates vary among sites, because such rate heterogeneity, according to a comparative study based on simulated data (Kuhner and Felsenstein 1994), results in failure to recover the true phylogenetic relationships in virtually all commonly used phylogenetic programs (or algorithms), including the maximum-likelihood method (e.g., PHYLIP), maximum parsimony (e.g., PAUP), or neighbor-joining (e.g., MEGA) methods. DAMBE can fit the Poisson, negative binomial, and gamma distributions to substitution data over sites. The maximum-likelihood estimator for k in the neg-

ative binomial distribution is from Johnson et al. (1992), and that for the shape parameter in the gamma distribution is from Evans et al. (1993).

Graphics

In addition to graphically displaying and printing trees, DAMBE also produces a variety of graphic outputs including plotting one or more amino acid properties along amino acid sequences (e.g., polarity plot), saturation plots (i.e., transitions and transversions over divergence), variability-over-site plots, substitution-over-site plots, etc.

In short, DAMBE is a user-friendly program for the Windows platform that features a suite of unique features as well as the capability of performing most routine data analyses in molecular biology, ecology, and evolution.

From the Bioinformatics Laboratory, Department of Ecology and Biodiversity, University of Hong Kong, Pokfulam Road, Hong Kong. DAMBE has incorporated codes from PHYLIP with permission from J. Felsenstein, and from the BASEML program in PAML with permission from Z. Yang. Part of the codes for sequence alignment in DAMBE are taken from the program CLUSTAL by D. Higgins, J. Thompson, and T. Gibson. DAMBE development is supported by a CRCG grant from the University of Hong Kong (10203043/27662, 10203435/27662) and an RGC grant from Hong Kong Research Grant Council (HKU7265/00M; to X.X.). We thank W. H. Li for using DAMBE in his bioinformatics course and H. Kong for assistance. Address correspondence to Xuhua Xia at the address above or e-mail: xxia@hkusua.hku.hk.

References

Evans M, Hastings N, and Peacock B, 1993. Statistical distributions. New York: John Wiley & Sons.

Felsenstein J, 1985. Confidence limits on phylogenies with a molecular clock. Syst Zool 34:152–161.

Felsenstein J, 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet Res 59:139–147.

Felsenstein J, 1993. PHYLIP 3.5 (phylogeny inference package). Seattle: Department of Genetics, University of Washington.

Johnson NL, Kotz S, and Kemp AW, 1992. Univariate discrete distributions. New York: John Wiley & Sons.

Jukes TH and Cantor CR, 1969. Evolution of protein molecules. In: Mammalian protein metabolism (Munro HN, ed). New York: Academic Press; 21–123.

Kimura M, 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111–120.

Kishino H and Hasegawa M, 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179.

Kuhner MK and Felsenstein J, 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468 [published erratum appears in Mol Biol Evol 1995;12: 525].

Lake JA, 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91:1455–1459.

Li W-H, 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99.

Nee S, Holmes EC, Rambaut A, and Harvey PH, 1996. Inferring population history from molecular phylogenies. In: New uses for new phylogenies (Harvey PH, Brown AJL, Maynard Smith J, Nee S, eds). Oxford: Oxford University Press; 66–80.

Tamura K and Nei M, 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526.

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

Xia X, 2000. Data analysis in molecular biology and evolution. Boston: Kluwer Academic.

Xia X, Hafner MS, and Sudman PD, 1996. On transition bias in mitochondrial genes of pocket gophers. J Mol Evol 43:32–40.

Xia X and Li W-H, 1998. What amino acid properties affect protein evolution? J Mol Evol 47:557–564.

Yang Z, 2000. Phylogenetic analysis by maximum likelihood (PAML). London: University College.

Received July 30. 2000
Accepted February 14, 2001

Corresponding Editor: Sudhir Kumar