# DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution

Xuhua Xia*[,1]

[1]Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario, Canada

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

**Associate editor**: Sudhir Kumar

## Abstract

**Since its first release in 2001 as mainly a software package for phylogenetic analysis, data analysis for molecular biology and evolution (DAMBE) has gained many new functions that may be classified into six categories: 1) sequence retrieval, editing, manipulation, and conversion among more than 20 standard sequence formats including MEGA, NEXUS, PHYLIP, GenBank, and the new NeXML format for interoperability, 2) motif characterization and discovery functions such as position weight matrix and Gibbs sampler, 3) descriptive genomic analysis tools with improved versions of codon adaptation index, effective number of codons, protein isoelectric point profiling, RNA and protein secondary structure prediction and calculation of minimum folding energy, and genomic skew plots with optimized window size, 4) molecular phylogenetics including sequence alignment, testing substitution saturation, distance-based, maximum parsimony, and maximum-likelihood methods for tree reconstructions, testing the molecular clock hypothesis with either a phylogeny or with relative-rate tests, dating gene duplication and speciation events, choosing the best-fit substitution models, and estimating rate heterogeneity over sites, 5) phylogeny-based comparative methods for continuous and discrete variables, and 6) graphic functions including secondary structure display, optimized skew plot, hydrophobicity plot, and many other plots of amino acid properties along a protein sequence, tree display and drawing by dragging nodes to each other, and visual searching of the maximum parsimony tree. DAMBE features a graphic, user-friendly, and intuitive interface and is freely available from http://dambe.bio.uottawa.ca (last accessed April 16, 2013).**

***Key words:*** **bioinformatics, phylogenetics, dating, Gibbs sampler, motif discovery, secondary structure, codon usage, hidden Markov model, genomic analysis.**

Article

## Introduction

Data analysis for molecular biology and evolution (DAMBE) is a comprehensive software package for sequence manipulation and descriptive and comparative sequence analysis. Since its first release (Xia 2001; Xia and Xie 2001b), DAMBE is now listed as one of the most widely used software packages in molecular phylogenetics (Salemi and Vandamme 2003; Felsenstein 2004; Lemey et al. 2009). However, with many functions added or improved (table 1), DAMBE now serves as a versatile software workbench for comprehensive data analysis of molecular data. Here, I outline the improvements and additions of DAMBE functions.

## Sequence Retrieval, Manipulation, and Format Conversion

DAMBE reads and writes sequence files in more than 20 standard sequence formats including the annotation-rich GenBank files, MEGA, PHYLIP, NEXUS, FASTA, the binary trace files, and NeXML files (Vos et al. 2012) for descriptive and comparative sequence analysis. NeXML format is an extensible sequence format for richly annotated comparative data to facilitate interoperability (Vos et al. 2012).

The parsing of files in GenBank format has been improved substantially with far more options than those in the first release (supplementary fig. S1, Supplementary Material online). DAMBE can extract annotated sequence features in a GenBank file, such as CDS, exons, introns, rRNA, and tRNA genes, as well as intergenic sequences and sequences upstream/downstream of sequence features. Extraction of sequences between genes or upstream/downstream of genes facilitates the identification of coregulated genes from coexpressed genes derived from functional genomic technologies (Xia and Xie 2001a).

Different exons in the same gene can be under different selection and mutation pressure and consequently may warrant separate sequence analyses (Xia, Xie, Li 2003). With DAMBE, one can extract first exons, middle exons, and last exons from multiexon coding sequences in an annotated GenBank files. The function for extracting introns and for characterizing splicing sites has facilitated the study of the relationship between intron splicing strength and gene expression in the yeast, *Saccharomyces cerevisiae* (Ma and Xia 2011).

## Motif Characterization and Discovery

DAMBE includes the most comprehensive implementation of position weight matrix (PWM) and the Gibbs sampler of which PWM is a component (Xia 2007b, p. 133–147, 2012b), with various ways of specifying the background frequencies,

**Table 1.** Categorization of DAMBE Functions.

| | |
|---|---|
| Sequence retrieval and format conversion | Read/write sequence file in standard formats (PHYLIP, NEXUS, PAML, FASTA, GenBank, Clustal, GCG, NeXML, Trace, etc.) |
| Sequence manipulation | Parsing GenBank files into CDS, exons, introns, exon/intron junctions, rRNA, tRNA, upstream, and downstream of sequence features; concatenate sequences from different files; editing sequences; extract first, second, or third codon sites |
| Motif characterization and discovery | Position weight matrix for characterizing and predicting sequence motifs; perceptron for two-group classification of sequence motifs; Gibbs sampler for characterizing and predicting novel/hidden sequence motifs, etc. |
| Secondary structure | RNA secondary structure prediction, and computation of MFE based on Vienna RNA library; hidden Markov model for predicting protein secondary structure based on training sequences |
| Sequence feature characterization | Codon adaptation index; effective number of codons; RSCU; protein isoelectric point; energetic cost of proteins; skew plots with optimized window size; Z curve |
| Sequence alignment | Global alignment nucleotide and AA sequences by ClustalW, mapping codon sequences to aligned AA sequences for all known genetic codes; local alignment by FASTA algorithms |
| Distances | Robust simultaneous distance estimation by likelihood and LS methods for F84, TN93, and GTR models; codon-based distances for all known genetic codes; patristic distance from input trees; R-F partition distance |
| Distance-based phylogenetic methods | NJ and FastME; statistical test of alternative topologies; use one matrix to evaluate multiple alternative trees, and multiple set of aligned sequences or distance matrices to compute a consensus tree |
| Parsimony method | Searching for MP tree and test for alternative topologies; dragging nodes to each other to find the MP tree |
| Likelihood method | fastDNAML but with estimated s/v ratio; ML analysis based on PAML codes; statistical tests of alternative topologies |
| Substitution models | Finding the best-fitting substitution model by using the likelihood ratio test and information theoretic indices (AIC, BIC); estimate rate heterogeneity over sites and proportion of invariant sites |
| Testing molecular clock | Likelihood-based relative rate test for nucleotide and codon sequences (for all known genetic codes); phylogeny-based test using ML and LS methods |
| Dating speciation or gene duplication events | Regular dating with internal node calibration, with single or multiple soft or hard calibration points, and tip dating frequently used for viruses sampled at different years. |
| Substitution saturation | Test for the presence of phylogenetic information, graphic substitution saturation plot |
| Comparative methods | Independent contrasts for continuous variables and character association for discrete/binary characters |
| Detecting recombination | Simplot; Boot-Scan; compatibility method |
| Other phylogenetic functions | A versatile tree-displaying panel for exporting high-quality trees for publication; tree-drawing by dragging nodes to each other; mapping nucleotide, AA and codon substitutions to tree branches |
| Other graphic functions | In silico 2D gel; plotting AA properties long sequences (e.g., hydrophobicity plot) |

the pseudocounts, and, in particular, a variety of significance tests associated with PWM (Xia 2012b). Discovering such motifs and understand differential selection on them is facilitated by the Gibbs sampler (Neuwald et al. 1995; Xia 2007b, p. 133–147, 2012b), which is for de novo motif discovery. For example, one may have a set of yeast intron sequences and wish to identify the branchpoint site whose nature or location is unknown. This would be a perfect case for applying the Gibbs sampler (fig. 1).

## Descriptive Sequence Analysis

### RNA and Protein Secondary Structure Prediction and Computation of Minimum Folding Energy

DAMBE uses the Vienna RNA Secondary Structure library (Hofacker 2003) to predict secondary structure of RNA sequences and to compute their minimum folding energy (MFE). It features graphic display of secondary structures (supplementary fig. S2, Supplementary Material online). Several studies have used MFE from DAMBE to study the relationship between N-terminal of mRNA and protein translation (e.g., Xia and Holcik 2009; Zid et al. 2009; Xia et al. 2011). DAMBE uses hidden Markov model for predicting protein secondary structure based on training sequences with experimentally determined protein structure (Xia 2007b, p. 109–132).

### Improved Codon Usage Indices

Codon usage bias reflects the joint effect of mutation bias and tRNA-mediated selection (Ikemura 1981; Xia 1996, 1998a, 2005, 2008, 2012c; Xia et al. 1996, 2007; Carullo and Xia 2008; Palidwor et al. 2010; Ran and Higgs 2012). DAMBE implements improved versions of widely used indices of codon usage bias, including the gene-specific codon adaptation index (Sharp and Li 1987; Xia 2007c) and the effective number of codons ($N_c$, Wright 1990; Sun et al. 2012), as well as the codon-specific relative synonymous codon usage (RSCU). These improved codon bias indices have contributed to the discovery of modified tRNA pool for translating HIV-1 late genes (van Weringh et al. 2011), the effect of poly(A) tracts at yeast 5′-untranslated region (5′-UTRs) (Xia et al. 2011), and the elucidation of the function of +4G in the Kozak consensus in mammalian mRNAs (Xia 2007a).

### Nucleotide Skew Plots

The two DNA strands are often subject to different mutation mediated by different DNA replication mechanisms and coding sequence bias. Nucleotide skew plots can often provide hints about mutation and selection operating during the evolutionary process (Lobry 1996; Marin and Xia 2008; Xia 2012a, 2012c). One main problem with the conventional nucleotide

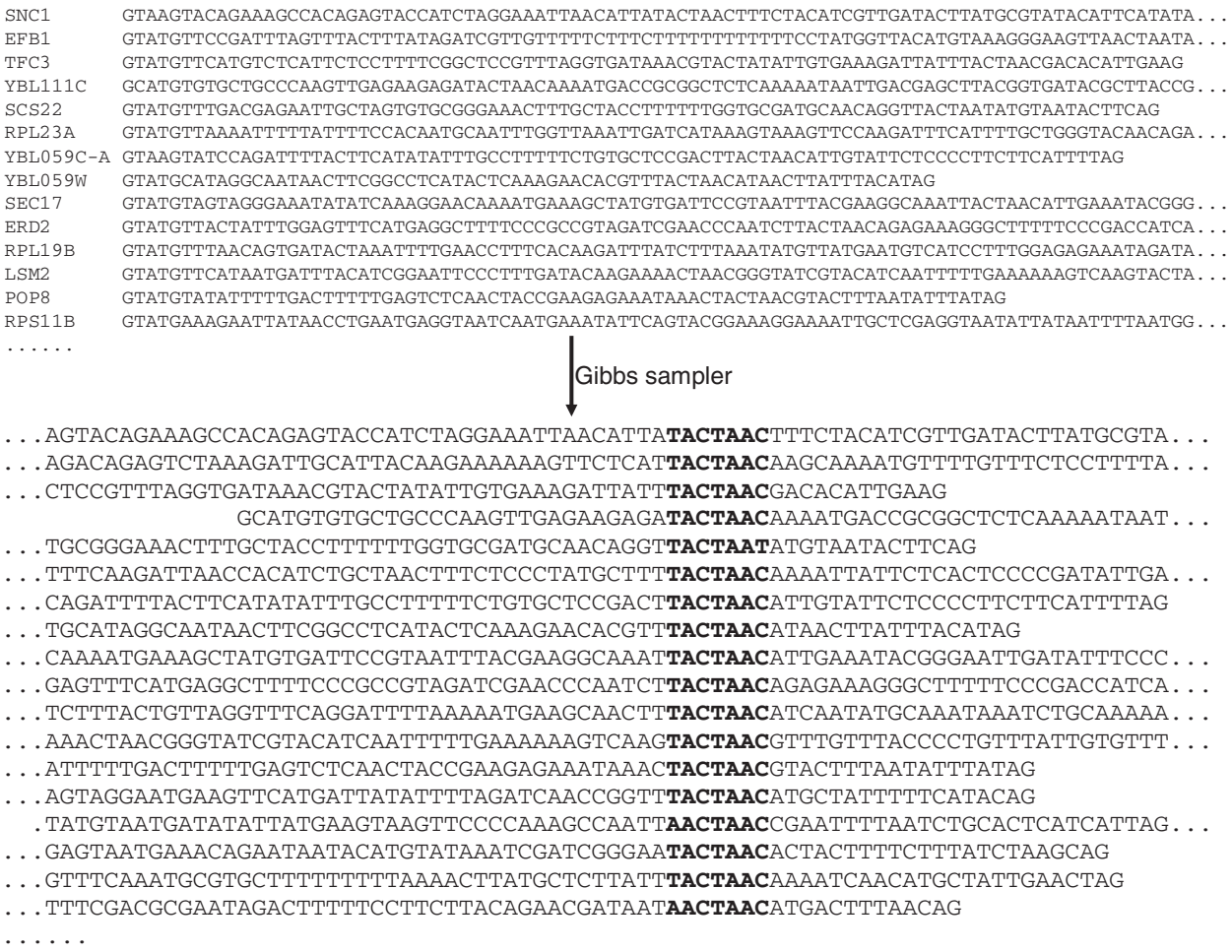


**Fig. 1.** Gibbs sampler in action. The yeast (*Saccharomyces cerevisiae*) intron sequences in the top panel represent the input to the Gibbs sampler. The bottom panel represents part of the output showing the identified motif (i.e., TAATAAC, bolded) shared among the sequences. Output from DAMBE5 also includes the PWM, the significance tests associated with PWM, and the PWM scores for individual motifs as a measure of motif strength, which is correlated with slicing efficiency. The input intron sequence file (YeastAllIntron.fas) is in DAMBE installation directory in FASTA format. Modified from Xia (2012b).

skew plots is the choice of the sliding window size (fig. 2). A window size too small will include too much noise and obscure interesting patterns, and a window size too large will often fail to identify precisely the point where abrupt changes of nucleotide composition occurs (which is typically associated with the origin and termination of DNA replication). DAMBE defines the optimal window size as the one that maximizes the area enclosed by the skew curve and the horizontal line specified by the global skew (fig. 2). The empirical justification of such a definition is that the site where the skew curve changes polarity is always very close to experimentally verified origin and termination of DNA replication in bacterial genomes. Users can specify their own window size and step size.

## Protein Isoelectric Point Profiling

Protein isoelectric point (pI) is important for understanding interactions between proteins and other cellular components because many of such interactions are mediated by electrostatic interactions, for example, a positively charged enzyme is attracted to its negatively charged substrate. DAMBE computes theoretical protein pIs by an iterative algorithm (Xia 2007b, p. 207–219). Empirical data based on protein pI from the acid-resistant gastric pathogen, *Helicobacter pylori*, have been used to test the three key evolutionary hypotheses: the preadaptation hypothesis, the exaptation hypothesis, and the adaptation hypothesis (Xia and Palidwor 2005). pI from DAMBE has also been used to study the adaptive evolution of the matrix extracellular phosphoglycoprotein in mammals and the implication of its change on protein folding (Machado et al. 2011). DAMBE used computed pI in its in silico 2D gel where input protein sequences are displayed on an in silico gel based on their charge and molecular weight (Xia 2007b, p. 207–219). Deviation of the observed protein location on the gel from the in silico prediction indicates posttranslational modification.

## Plot Amino Acid Properties along the Protein Sequence

Amino acids (AAs) are characterized by size, charge, hydrophobicity/polarity, and their tendency to form $\alpha$ helices and $\beta$ sheets. Plotting these properties along the protein sequence

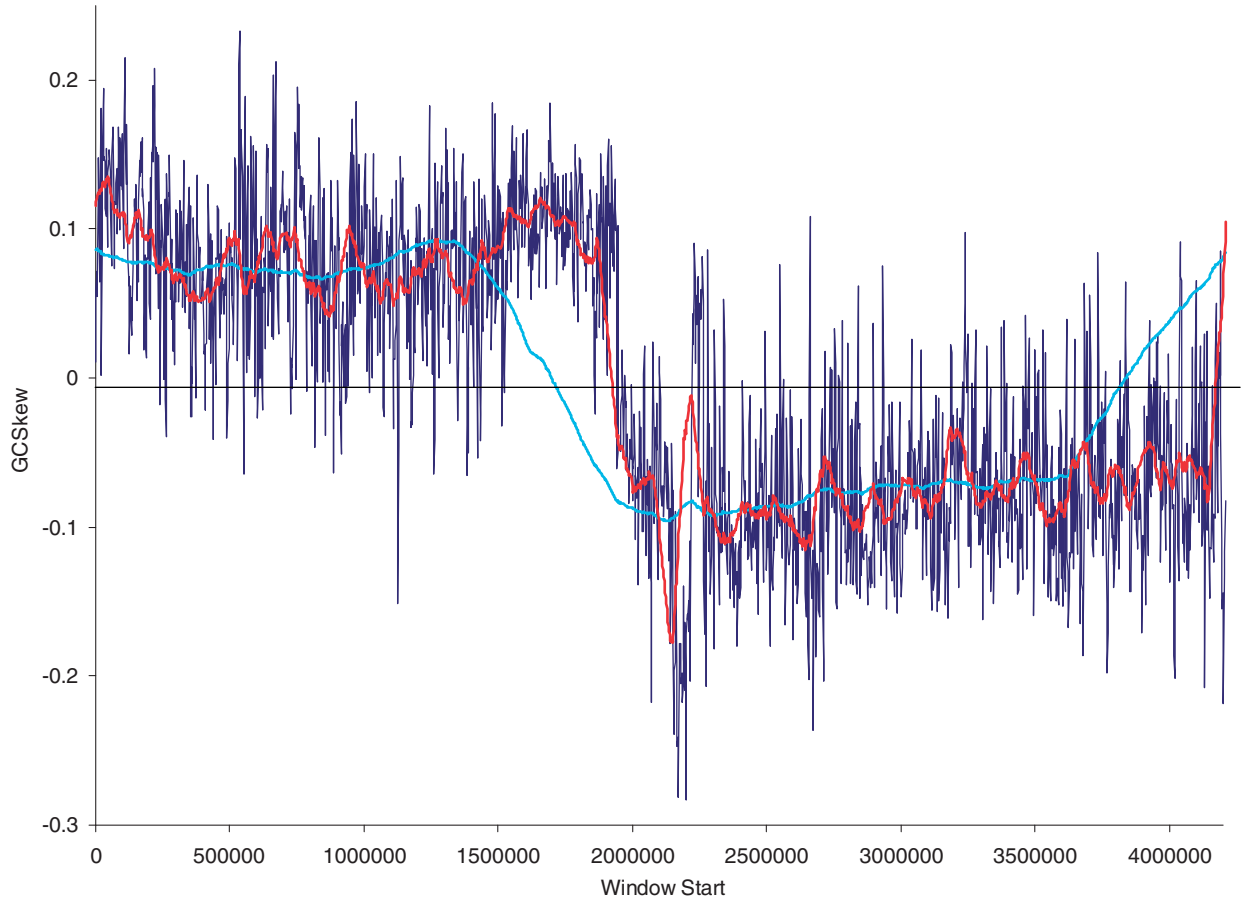


**Fig. 2.** Skew plots of the *Bacillus subtilis* genome at three different window sizes, with the skew curve colored in red having the optimal window size. The horizontal line is the global GC skew computed from the entire genome.

can often shed lights on local structures and functional domains. For example, DNA or RNA binding domains are typically characterized by a stretch of positively charged AAs such as lysine, arginine, and histidine, whereas transmembrane proteins typically contain hydrophobic domains (fig. 3). The presence of these domains creates structural heterogeneity and represents a major source of rate heterogeneity in nonsynonymous substitutions among sites (Xia 1998b; Xia and Li 1998), which can often bias phylogenetic estimation. Several homologous sequences can be plotted jointly for one to visualize how AA substitutions lead to changes in protein phenotype (fig. 3). DAMBE's function for plotting these AA properties along protein sequences is accessed by clicking "Graphics|amino acid properties along sequences."

### Nucleotide, Dinucleotide, AA, and Di-AA Frequencies

These simple frequencies not only serve as excellent entry point for teaching molecular evolution but can also lead to significant biological insights on spontaneous mutation during the evolutionary process (Xia et al. 1996, 2006; Xia 2003, 2012a, 2012c; Xia and Yuen 2005). For example, *Mycoplasma genitalium* has much lower genomic CpG dinucleotide frequencies than *M. pneumoniae*, but differential CpG-specific DNA methylation has been excluded as an explanation because neither species has any CpG-specific methyltransferase. It was found that their sister species, *M. pulmonis*, as well as several other deeper-rooted relatives, have CpG-specific methyltransferases and have even lower CpG dinucleotide frequencies. This restores DNA methylation as an explanation for variation in CpG frequencies between *M. genitalium* and *M. pneumoniae*. That is, the common ancestor of *M. genitalium* and *M. pneumoniae* lost the CpG-specific methyltransferases, and both daughter lineages began to rebound in CpG frequencies. Because *M. pneumoniae* has evolved much faster than *M. genitalium*, its CpG frequency has rebounded to a much higher level than *M. genitalium* (Xia 2003). Similarly, di-AA frequencies among proteomes from diverse array of organisms have revealed constraints of AA by their neighbors (Xia and Xie 2002), and experimental evolution has shown that *Pasteurella multocida* cultured at increasing temperature for over 14,400 generations decreased genomic GC (Xia et al. 2002), contrary to the conventional hypothesis that genomic GC should increase with increasing environmental temperature.

## Local and Global Sequence Alignment

DAMBE implements the FASTA algorithms (Pearson and Lipman 1988; Xia 2007b, p. 4–16) for local sequence alignment and string searching, and the dynamic programming alignment used in ClustalW (Thompson et al. 1994) for global alignment. Since its first release (Xia 2001; Xia and Xie 2001b),

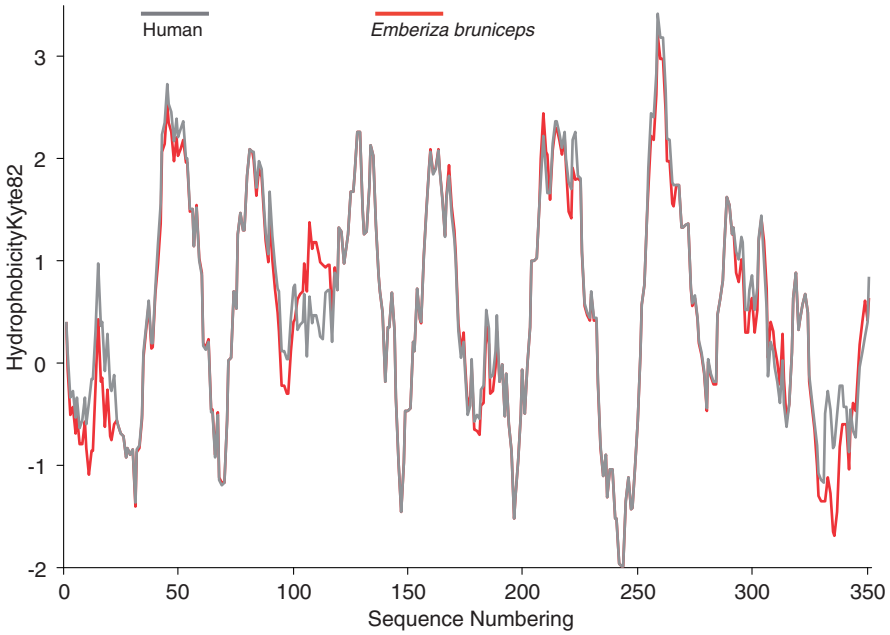


FIG. 3. Hydrophobicity plot for human (NP_000530.1) and avian (*Emberiza bruniceps*: AFK10338) rhodopsin with seven transmembrane domains (peaks). The weak 7th peak is due to a relatively short $\alpha$-helix. Output from DAMBE. A sliding window of 12 AAs is used.

DAMBE can align protein-coding sequences against the aligned AA sequences, thus preventing frameshifting gaps as alignment artifacts, which occur frequently with direct alignment of protein-coding nucleotide sequences. Newer versions have added new genetic codes as they are discovered, allowing researchers to use this function for protein-coding genes in all 18 known genetic codes (Courgnaud et al. 2002; Chen et al. 2011; Du et al. 2011; Reis et al. 2011; Tan et al. 2011; Fu et al. 2012; Glaw et al. 2012).

The inclusion of the Vienna RNA Secondary Structure library (Hofacker 2003) facilitates the identification of secondary structure present in RNA sequences, which can then be used to guide sequence alignment with structural constraints. Such structure-guided alignment has significantly improved the quality of phylogenetic reconstruction (Kjer 1995; Xia 2000; Xia, Xie, Kjer 2003).

## Molecular Phylogenetics

DAMBE is listed as one of the most widely used software packages in molecular phylogenetics (Salemi and Vandamme 2003; Felsenstein 2004; Lemey et al. 2009). It performs routine phylogenetic analysis using the distance-based, maximum parsimony, and maximum-likelihood (ML) methods, with a variety of options (supplementary fig. S3, Supplementary Material online), and can handle multiple distance matrices or multiple sets of sequence alignments in phylogenetic analysis, which arise frequently in simulation studies or in large-scale data compilations. Multiple data sets are read into DAMBE by clicking "File|Read other molecular data file" and choosing relevant options.

### Distance-Based Phylogenetic Methods

The distance-based methods implemented in DAMBE have two advantages over other comparable programs. First, in addition to the popular neighbor-joining (NJ) method (Saitou and Nei 1987), DAMBE implements the fast and accurate FastME method (Desper and Gascuel 2002), which not only has a global criterion for the best tree, in contrast to a local criterion in the NJ method, but also extensive searching through tree space by nearest-neighbor-interchange. Second, DAMBE implements simultaneously estimated (SE) distances for the F84, TN93, and GTR substitution models in both the ML and the least-square (LS) framework (referred as MLCompositeF84, MLCompositeTN93, LSCompositeF84, LSCompositeTN93, and LSCompositeGTR). These SE distances alleviated or eliminated the three major problems associated with the independently estimated distances (Tamura et al. 2004; Xia 2009) and reduced the documented topological biases associated with the over- and underestimated distances (Xia 2006). This offers great flexibility than MEGA, which implemented the maximum composite likelihood distance only for the TN93 model.

One advantage of the distance-based method over other phylogenetic methods is that it theoretically only needs pairwise alignment, which can be done well with dynamic programming. This eliminates two serious problems associated with the progressive multiple sequence alignment (PMSA) with a guide tree, used in Clustal and other practical alignment programs. First, PMSA does not guarantee the multiple alignment to be optimal given the scoring scheme, and extensive manual editing is often necessary (but impractical with a large data set). Second, PMSA has the well-known chicken–egg problem, that is, obtaining a good multiple alignment needs a good guide tree, but a good guide tree needs a good multiple alignment. The advantage of requiring only pairwise alignment implies that distance-based methods can perform phylogenetic analysis with sequences so diverged that it is practically impossible to obtain a biologically meaningful multiple sequence alignment, thus excluding the

application of all other phylogenetic methods that require a multiple alignment.

DAMBE also implements a variety of distances for AA sequences, codon sequences, and allele frequencies including microsatellite data. In particular, because DAMBE implements all 18 known genetic codes, these codon-based distances can be computed for any codon sequences from any species pairs as long as they share the same genetic code. Synonymous and nonsynonymous substitution rates computed from DAMBE have been used to document accelerated evolution in rearranged ape chromosomes (Navarro and Barton 2003) and the evolution and functional divergence in the FoxL2 gene family (Cocquet et al. 2003, 2002; Baron et al. 2004). DAMBE integrates all these distances with phylogenetic functions for phylogenetic reconstruction, testing the molecular clock hypothesis, and dating speciation and population divergence.

### ML Methods

DAMBE has two implementation of the ML method, one based on the source code from fastDNAML (Olsen et al. 1994), but with an estimated transition/transversion ratio (supplementary fig. S3*b*, Supplementary Material online), and the other based on the PAML package (abacus.gene.ucl.ac.uk/software/paml.html) (supplementary fig. S3*c*, Supplementary Material online). The ML method is also used in estimating the shape parameter of the gamma distribution, the proportion of invariant sites, and in finding the best-fitting substitution model (by using information-theoretic indices and by using likelihood ratio test for nested models). The alpha parameter in the gamma distribution is estimated in DAMBE by the rapid divide-and-conquer method (Gu and Zhang 1997), that is, the ancestral states are constructed with a given or reconstructed tree, and the number of substitutions per site is corrected for multiple hits and then fitted to the gamma distribution to find the alpha parameter.

### Testing the Molecular Clock Hypothesis

The molecular clock hypothesis is typically tested by either relative-rate tests or by tree-based tests. For relative-rate tests, DAMBE implements nucleotide-based and codon-based likelihood ratio tests (Muse and Weir 1992; Muse and Gaut 1994). There are four nested models. The general model for nucleotide sequences does not constrain either the transition or transversion rates between the two ingroup lineages, the most restrictive model constrain both, and the two intermediate models with one constraining the transition rate and the other constraining the transversion rate. The analysis for the codon sequences is the same except that transition and transversion rates are replaced by synonymous and nonsynonymous substitution rates. The strength in the codon-based method implemented in DAMBE is that all 18 known genetic codes are implemented. For tree-based tests of the molecular clock hypothesis, DAMBE implements both the ML method (a likelihood ratio test between the clock-constrained model and the unconstrained model) and the LS method (Xia 2009).

### Dating Speciation or Gene Duplication Events

DAMBE used the LS method for dating speciation or gene duplication events (Xia and Yang 2011). It performs both regular dating with internal node calibration, with single or multiple soft or hard calibration points, and tip dating frequently used for viruses sampled at different times (fig. 4). The confidence intervals of the estimated dates are obtained by resampling. There are two sources of the variation in the estimated dates, one from the uncertainty in the calibration points (e.g., the fossil dates) and the other from the sequences used for dating. The resampling method only assesses the variation from the sequences. The only logical way of reducing uncertainty associated with the fossil dates is to have more fossils and more calibration points. The dating method in DAMBE has been used to estimate the time to the most recent common ancestor (Bisconti et al. 2011, 2013; Canestrelli et al. 2012).

### Other Functions Related to Phylogenetic Analysis

The function for assessing substitution saturation (Xia, Xie, Salemi, et al. 2003; Xia and Lemey 2009) has been used in hundreds of publications in phylogenetic research. Other functions include detecting viral recombination by Simplot, Bootscan, and compatibility methods (reviewed in Xia 2011). DAMBE features a versatile tree-displaying panel and a tree-drawing panel (supplementary fig. S3*d*, Supplementary Material online) for exporting high-quality trees for publication and for tree drawing by dragging nodes to each other, and it can print large trees to multiple pages.

## Phylogeny-Based Comparative Methods

Large-scale comparative genomics involves the type of data aimed to understand functional association among genes, between genes and phenotypes, and between genotype/phenotype and the environmental variables (supplementary fig. S4, Supplementary Material online). Detecting functional associations or correlations among the genetic (G), phenotypic (P), and environmental (E) variables are typically done by phylogeny-based comparative methods (Harvey and Pagel 1991; Xia 2011, 2013, Chapter 2). The continuous variables are analyzed by independent contrasts (Felsenstein 1985, 2004, p. 432–459) or by the generalized LS method (Pagel 1999), which has the advantage of not requiring the ancestral value being always somewhere between the values of the two descendent lineages and can therefore detect directional evolution. The method has been used successfully in the test of genome size and ambient temperature in salamanders (Xia 1995) and in the discovery that rRNA genes in thermophilic bacteria have longer stems in their stem-loop structure as well as more GC-rich in their stems than mesophilic bacteria (Wang et al. 2006). The discrete variables (e.g., gene presence/absence data for detecting gene association) are analyzed by the likelihood ratio test (Barker and Pagel 2005; Xia 2011) between two models, one assuming association and the other not. DAMBE implements phylogeny-based comparative methods for both continuous and discrete

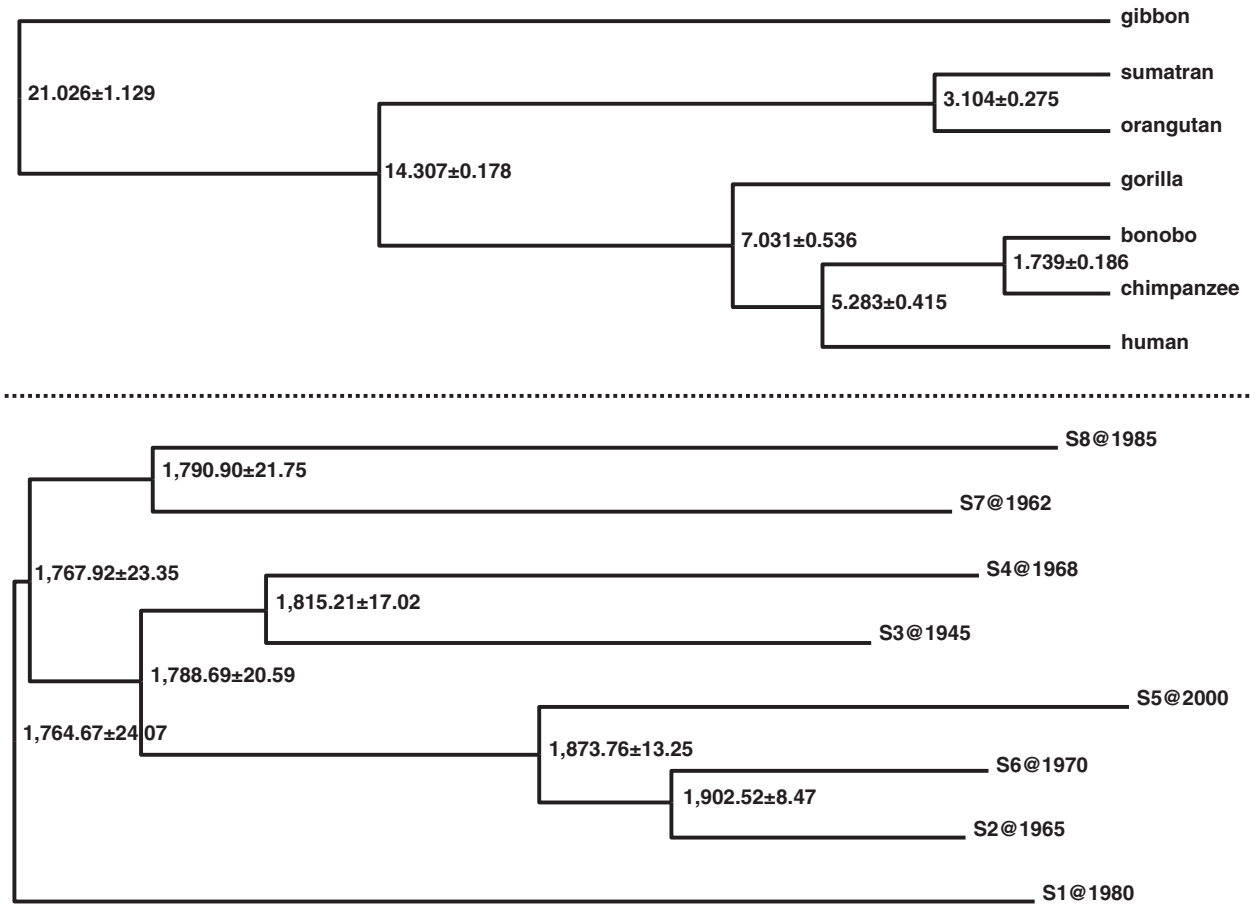


FIG. 4. Regular dating with internal node calibration (top panel) and tip dating with sampling times (specified in OTU name following "@") for calibration, based on the LS method in DAMBE.

variables. The function is accessed by clicking "Phylogenetics| Comparative methods."

## Conclusion

DAMBE5 is a comprehensive and user-friendly software workbench for data analysis in molecular biology and evolution. It features a variety of analytical functions for descriptive and comparative sequence analysis, as well as functions for allele frequencies and microsatellite data. DAMBE is enhanced by a variety of graphic functions, which makes it ideal for teaching bioinformatics and molecular evolution. DAMBE are available free of charge from http://dambe.bio.uottawa.ca (last accessed April 16, 2013) where a set of tutorials making use of various DAMBE functions can be found. DAMBE is a Windows program but may run on Linux-based systems.

## Supplementary Material

Supplementary figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

The author thanks his students and many colleagues who have used DAMBE and given feedback for improvement, and two anonymous reviewers for their constructive comments. This work was supported by the Discovery Grant of Natural Science and Engineering Research Council of Canada (NSERC).

## References

Barker D, Pagel M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol.* 1:e3.

Baron D, Cocquet J, Xia X, Fellous M, Guiguen Y, Veitia RA. 2004. An evolutionary and functional analysis of *FoxL2* in rainbow trout gonad differentiation. *J Mol Endocrinol.* 33:705–715.

Bisconti R, Canestrelli D, Colangelo P, Nascetti G. 2011. Multiple lines of evidence for demographic and range expansion of a temperate species (*Hyla sarda*) during the last glaciation. *Mol Ecol.* 20: 5313–5327.

Bisconti R, Canestrelli D, Salvi D, Nascetti G. 2013. A geographic mosaic of evolutionary lineages within the insular endemic newt *Euproctus montanus*. *Mol Ecol.* 22:143–156.

Canestrelli D, Salvi D, Maura M, Bologna MA, Nascetti G. 2012. One species, three Pleistocene evolutionary histories: phylogeography of the Italian crested newt, *Triturus carnifex*. *PLoS One* 7:e41754.

Carullo M, Xia X. 2008. An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. *J Mol Evol.* 66:484–493.

Chen WJ, Bu Y, Carapelli A, Dallai R, Li S, Yin WY, Luan YX. 2011. The mitochondrial genome of *Sinentomon erythranum* (Arthropoda: Hexapoda: Protura): an example of highly divergent evolution. *BMC Evol Biol.* 11:246.

Cocquet J, De Baere E, Gareil M, Pannetier M, Xia X, Fellous M, Veitia RA. 2003. Structure, evolution and expression of the *FOXL2* transcription unit. *Cytogenet Genome Res.* 101:206–211.

Cocquet J, Pailhoux E, Jaubert F, et al. (11 co-authors). 2002. Evolution and expression of *FOXL2*. *J Med Genet.* 39:916–921.

Courgnaud V, Salemi M, Pourrut X, et al. (11 co-authors). 2002. Characterization of a novel simian immunodeficiency virus with a *vpu* gene from greater spot-nosed monkeys (*Cercopithecus nictitans*) provides new insights into simian/human immunodeficiency virus phylogeny. *J Virol.* 76:8298–8309.

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol.* 9:687–705.

Du J, Miura E, Robischon M, Martinez C, Groover A. 2011. The *Populus* class III HD ZIP transcription factor POPCORONA affects cell differentiation during secondary growth of woody stems. *PLoS One* 6: e17458.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat.* 125: 1–15.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer.

Fu C, Guo L, Xia R, Li J, Lei G. 2012. A multilocus phylogeny of Asian noodlefishes Salangidae (Teleostei: Osmeriformes) with a revised classification of the family. *Mol Phylogenet Evol.* 62:848–855.

Glaw F, Kohler J, Townsend TM, Vences M. 2012. Rivaling the world's smallest reptiles: discovery of miniaturized and microendemic new species of leaf chameleons (Brookesia) from northern Madagascar. *PLoS One* 7:e31314.

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

Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.

Hofacker IL. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* 31:3429–3431.

Ikemura T. 1981. Correlation between the abundance of *Escheriachia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol.* 146:1–21.

Kjer KM. 1995. Use of ribosomal-RNA secondary structure in phylogenetic studies to identify homologous positions—an example of alignment and data presentation from the frogs. *Mol Phylogenet Evol.* 4:314–330.

Lemey P, Salemi M, Vandamme AM. 2009. The phylogenetic handbook. Cambridge: Cambridge University Press.

Lobry JR. 1996. Origin of replication of *Mycoplasma genitalium*. *Science* 272:745–746.

Ma P, Xia X. 2011. Factors affecting splicing strength of yeast genes. *Comp Funct Genomics.* 2011:212146.

Machado JP, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. 2011. Adaptive evolution of the matrix extracellular phosphoglycoprotein in mammals. *BMC Evol Biol.* 11:342.

Marin A, Xia X. 2008. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. *J Theor Biol.* 253:508–513.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol.* 11:715–724.

Muse SV, Weir BS. 1992. Testing for equality of evolutionary rates. *Genetics* 132:269–276.

Navarro A, Barton NH. 2003. Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science* 300:321–324.

Neuwald AF, Liu JS, Lawrence CE. 1995. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* 4: 1618–1632.

Olsen GJ, Matsuda H, Hagstrom R, Overbeek R. 1994. fastDNAML: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput Appl Biosci.* 10:41–48.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

Palidwor GA, Perkins TJ, Xia X. 2010. A general model of codon bias due to GC mutational bias. *PLoS One* 5:e13431.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A.* 85:2444–2448.

Ran W, Higgs PG. 2012. Contributions of speed and accuracy to translational selection in bacteria. *PLoS One* 7:e51652.

Reis M, Sousa-Guimaraes S, Vieira CP, Sunkel CE, Vieira J. 2011. *Drosophila* genes that affect meiosis duration are among the meiosis related genes that are more often found duplicated. *PLoS One* 6: e17512.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.

Salemi M, Vandamme A-M. 2003. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge: Cambridge University Press.

Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281–1295.

Sun XY, Yang Q, Xia X. 2012. An improved implementation of effective number of codons ($N_c$). *Mol Biol Evol.* 30:191–196.

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.* 101:11030–11035.

Tan CY, Ninove L, Gaudart J, Nougairede A, Zandotti C, Thirion-Perrier L, Charrel RN, de Lamballerie X. 2011. A retrospective overview of enterovirus infection diagnosis and molecular epidemiology in the public hospitals of Marseille, France (1985-2005). *PLoS One* 6:e18022.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.

van Weringh A, Ragonnet-Cronin M, Pranckeviciene E, Pavon-Eternod M, Kleiman L, Xia X. 2011. HIV-1 modulates the tRNA pool to improve translation efficiency. *Mol Biol Evol.* 28:1827–1834.

Vos RA, Balhoff JP, Caravas JA, et al. (11 co-authors). 2012. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Syst Biol.* 61:675–689.

Wang HC, Xia X, Hickey DA. 2006. Thermal adaptation of ribosomal RNA genes: a comparative study. *J Mol Evol.* 63:120–126.

Wright F. 1990. The "effective number of codons" used in a gene. *Gene* 87:23–29.

Xia X. 1995. Body temperature, rate of biosynthesis and evolution of genome size. *Mol Biol Evol.* 12:834–842.

Xia X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* 144:1309–1320.

Xia X. 1998a. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics* 149:37–44.

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

Xia X. 2000. Phylogenetic relationship among horseshoe crab species: the effect of substitution models on phylogenetic analyses. *Syst Biol.* 49: 87–100.

Xia X. 2001. Data analysis in molecular biology and evolution. Boston: Kluwer Academic Publishers.

Xia X. 2003. DNA methylation and mycoplasma genomes. *J Mol Evol.* 57: S21–S28.

Xia X. 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. *Gene* 345:13–20.

Xia X. 2006. Topological bias in distance-based phylogenetic methods: problems with over- and underestimated genetic distances. *Evol Bioinform Online.* 2:375–387.

Xia X. 2007a. The + 4G site in Kozak consensus is not related to the efficiency of translation initiation. *PLoS One* 2:e188.

Xia X. 2007b. Bioinformatics and the cell: modern computational approaches in genomics, proteomics and transcriptomics. New York: Springer.

Xia X. 2007c. An Improved Implementation of codon adaptation index. *Evol Bioinform Online.* 3:53–58.

Xia X. 2008. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. *BMC Evol Biol.* 8:211.
Xia X. 2009. Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. *Mol Phylogenet Evol.* 52:665–676.
Xia X. 2011. Comparative genomics. In: Lu HH-S, Scholkopf B, Zhao H, editors. Handbook of computational statistics: statistical bioinformatics. Berlin (Germany): Springer. p. 567–600.
Xia X. 2012a. DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. *Curr Genomics.* 13:16–27.
Xia X. 2012b. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica* 917540:15.
Xia X. 2012c. Rapid evolution of animal mitochondria. In: Singh RS, Xu J, Kulathinal RJ, editors. Evolution in the fast lane: rapidly evolving genes and genetic systems. Oxford: Oxford University Press. p. 73–82.
Xia X. 2013. Comparative genomics. Berlin (Germany): Springer.
Xia X, Hafner MS, Sudman PD. 1996. On transition bias in mitochondrial genes of pocket gophers. *J Mol Evol.* 43:32–40.
Xia X, Holcik M. 2009. Strong eukaryotic IRESs have weak secondary structure. *PLoS One* 4:e4136.
Xia X, Huang H, Carullo M, Betran E, Moriyama EN. 2007. Conflict between translation initiation and elongation in vertebrate mitochondrial genomes. *PLoS One* 2:e227.
Xia X, Lemey P. 2009. Assessing substitution saturation with DAMBE. In: Lemey P, Salemi M, Vandamme AM, editors. The phylogenetic handbook. Cambridge: Cambridge University Press. p. 615–630.
Xia X, Li WH. 1998. What amino acid properties affect protein evolution? *J Mol Evol.* 47:557–564.
Xia X, MacKay V, Yao X, Wu J, Miura F, Ito T, Morris DR. 2011. Translation initiation: a regulatory role for poly(A) tracts in front of the AUG codon in *Saccharomyces cerevisiae*. *Genetics* 189: 469–478.
Xia X, Palidwor G. 2005. Genomic adaptation to acidic environment: evidence from *Helicobacter pylori*. *Am Nat.* 166:776–784.
Xia X, Wang H, Xie Z, Carullo M, Huang H, Hickey D. 2006. Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes. *Mol Biol Evol.* 23:1450–1454.
Xia X, Xie Z. 2001a. AMADA: analysis of microarray data. *Bioinformatics* 17:569–570.
Xia X, Xie Z. 2001b. DAMBE: software package for data analysis in molecular biology and evolution. *J Hered.* 92:371–373.
Xia X, Xie Z. 2002. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. *Mol Biol Evol.* 19:58–67.
Xia X, Xie Z, Li WH. 2003. Effects of GC content and mutational pressure on the lengths of exons and coding sequences. *J Mol Evol.* 56: 362–370.
Xia X, Yang Q. 2011. A distance-based least-square method for dating speciation events. *Mol Phylogenet Evol.* 59:342–353.
Xia X, Yuen KY. 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. *BMC Genet.* 6:20.
Xia XH, Wei T, Xie Z, Danchin A. 2002. Genomic changes in nucleotide and dinucleotide frequencies in *Pasteurella multocida* cultured under high temperature. *Genetics* 161:1385–1394.
Xia XH, Xie Z, Kjer KM. 2003. 18S ribosomal RNA and tetrapod phylogeny. *Syst Biol.* 52:283–295.
Xia XH, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. *Mol Phylogenet Evol.* 26:1–7.
Zid BM, Rogers AN, Katewa SD, Vargas MA, Kolipinski MC, Lu TA, Benzer S, Kapahi P. 2009. 4E-BP extends lifespan upon dietary restriction by enhancing mitochondrial activity in *Drosophila*. *Cell* 139: 149–160.