Molecular Phylogenetics and Evolution 102 (2016) 331–343

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

PhyPA: Phylogenetic method with pairwise sequence alignment outperforms likelihood methods in phylogenetics involving highly diverged sequences

Xuhua Xia *

Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa K1N 6N5, Canada
Ottawa Institute of Systems Biology, 451 Smyth Road, Ottawa, ON K1H 8M5, Canada

ARTICLE INFO

Article history:
Received 7 April 2016
Accepted 1 July 2016
Available online 1 July 2016

Keywords:
Sequence alignment
Deep phylogeny
Distance-based methods
Likelihood-based methods
Codon-based alignment

ABSTRACT

While pairwise sequence alignment (PSA) by dynamic programming is guaranteed to generate one of the optimal alignments, multiple sequence alignment (MSA) of highly divergent sequences often results in poorly aligned sequences, plaguing all subsequent phylogenetic analysis. One way to avoid this problem is to use only PSA to reconstruct phylogenetic trees, which can only be done with distance-based methods. I compared the accuracy of this new computational approach (named PhyPA for phylogenetics by pairwise alignment) against the maximum likelihood method using MSA (the ML + MSA approach), based on nucleotide, amino acid and codon sequences simulated with different topologies and tree lengths. I present a surprising discovery that the fast PhyPA method consistently outperforms the slow ML + MSA approach for highly diverged sequences even when all optimization options were turned on for the ML + MSA approach. Only when sequences are not highly diverged (i.e., when a reliable MSA can be obtained) does the ML + MSA approach outperforms PhyPA. The true topologies are always recovered by ML with the true alignment from the simulation. However, with MSA derived from alignment programs such as MAFFT or MUSCLE, the recovered topology consistently has higher likelihood than that for the true topology. Thus, the failure to recover the true topology by the ML + MSA is not because of insufficient search of tree space, but by the distortion of phylogenetic signal by MSA methods. I have implemented in DAMBE PhyPA and two approaches making use of multi-gene data sets to derive phylogenetic support for subtrees equivalent to resampling techniques such as bootstrapping and jackknifing.

© 2016 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Phylogenetic reconstruction becomes difficult with deep phylogenies, mainly due to the difficulty in obtaining reliable MSA (Blackburne and Whelan, 2013; Edgar and Batzoglou, 2006; Herman et al., 2014; Kumar and Filipski, 2007; Lunter et al., 2008; Wong et al., 2008). In contrast to pairwise sequence alignment (PSA) by dynamic programming which is guaranteed to generate, for a given scoring scheme, the optimal alignment or at least one of the equally optimal alignments, MSAs of highly divergent sequences are often poor, especially those obtained from a progressive alignment with a guide tree. Although an iterative approach (Hogeweg and Hesper, 1984; Katoh et al., 2009; Thompson et al., 1994) is typically used for MSA in which a guide tree is used to generate an alignment which is then used to construct a new guide tree to guide the next round of progressive alignment, such approach still often produces poor alignment for deeply diverged sequences, and a poor alignment typically leads to bias and inaccuracy in phylogenetic estimation (Blackburne and Whelan, 2013; Kumar and Filipski, 2007; Wong et al., 2008).

One way to avoid problems associated with poor MSA is simply not to do MSA, i.e., by phylogenetic analysis based on PSA only as pioneered by Thorne and Kishino (1992). Currently, distance-based methods are the only ones that can take advantage of PSA to build phylogenetic trees. Oddly enough, this greatest advantage of distance-based methods has never been realized in practice because no widely used software packages in phylogenetics has implemented any distance-based methods based on PSA.

Thorne and Kishino (1992) did not actually implement the method and evaluate it against the maximum likelihood (ML) method. Consequently it does not change the general conception among molecular phylogeneticists that distance-based methods

* Address: Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa K1N 6N5, Canada.

E-mail address: xxia@uottawa.ca

http://dx.doi.org/10.1016/j.ympev.2016.07.001
1055-7903/© 2016 The Author. Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

are quick and dirty and should be used only when there are too much data to render the ML approach infeasible.

An alternative phylogenetic method (PHYRN) for highly diverged sequences without using MSA has been proposed recently (Bhardwaj et al., 2012). With one set of highly divergent sequences, one first breaks the sequences into short segments and use BLAST to search for local similarities to build a position weight matrix. One then derive a Euclidean distance based on the sharing of such local similarities. In essence, the method is based on a phonetic distance derived from long-word matching. While the web link for the method and the manual (www.ccp.psu.edu/downloads) is no longer accessible, the source codes for the method is still available from G. Bhardwaj who is willing to help with running the protocol (pers. comm.).

I have developed and implemented a phylogenetic reconstruction method based on PSA only (named PhyPA for **Phy**logenetics by **P**airwise **A**lignment) for highly diverged sequences (One should use likelihood-based method whenever sequences are not highly diverged and reliable MSA can be obtained). I evaluated the phylogenetic performance of PhyPA against commonly used likelihood methods based on MSA by using nucleotide, amino acid and codon sequences simulated with different tree topologies and different branch lengths. The combination of likelihood methods and MSA (hereafter referred to as ML + MSA) include MSA by MAFFT (Katoh et al., 2009) and MUSCLE (Edgar, 2004a,b), and phylogenetic reconstruction by PhyML (Guindon and Gascuel, 2003) and PROML/DNAML in the PHYLIP package (Felsenstein, 2014).

PhyPA consistently outperforms the (ML + MSA) approach in phylogenetic reconstruction involving highly diverged sequences. This is true even when all key optimization options are turned on for MAFFT|MUSCLE and PhyML/PROML/DNAML. I have also implemented two approaches making use of multi-gene data sets to derive phylogenetic support for subtrees equivalent to resampling techniques such as bootstrapping and jackknifing. In addition, different candidate topologies can be assessed for relative support. Below I describe the details of PhyPA, the evaluation process for comparing PhyPA against (ML + MSA), and results demonstrating the strength and weaknesses of PhyPA relative to (ML + MSA).

2. Description of PhyPA

PhyPA analysis consists of three steps: (1) pairwise sequence alignment, (2) computing evolutionary distances, and (3) reconstructing phylogenetic tree. These three steps are described below with an emphasis on the new codon-based alignment.

2.1. Pairwise sequence alignment

PhyPA uses the routine dynamic programming approach with affine function gap penalty for PSA. Three scoring matrices were implemented for nucleotide sequence alignment: the standard IUB matrix, and two transition bias matrices with different penalties for transitional and transversional substitutions. All three matrices accommodate ambiguous codes. For example, the substitution score between A and R (which stands for either A or G) is $S_{A,R} = (S_{A,A} + S_{A,G})/2$, and that between R and Y (which stands for either C or T) is $S_{R,Y} = (S_{A,C} + S_{A,T} + S_{G,C} + S_{G,T})/4$, and so on. For amino acid sequences, 25 scoring matrices were implemented including 15 BLOSUM matrices (spanning from BLOSUM30 to BLOSUM100), three PAM matrices and JTT92 matrix. BLOSUM62 represents a good compromise between closely related and highly diverged sequences and is the default.

For codon sequences, the conventional alignment is done by first translating codon sequences into amino acid sequences, aligning the amino acid sequences, and then aligning the codon sequences according to the aligned amino acid sequences (Xia, 2001, Chapter 5). This approach has two shortcomings. First, given the two sequences in Fig. 1a, the approach would yield Alignment 1 (Fig. 1b, where the two sequences differing by a triplet-indel and a transversion), but not Alignment 2 (Fig. 1c, where the two sequences differing by a triple-indel only) which is more parsimonious than Alignment 1. Second, once the codon sequences are translated into amino acid sequences, we lose the information on nucleotide differences between codons which could differ at 1, 2 or 3 codon sites.

I added a new approach with two improvements to solve these two problems. The first accommodates nucleotide differences between codons. The 64 codons are coded with AAA as 0, AAC as 1, ..., TTT as 63. The resulting 64-alphabet sequences are aligned by the affine function gap penalty and a 64-by-64 substitution score matrix. Each entry in the 64-by-64 matrix (score between two paired codons) is $S_{codon1,codon2} = S_{aa1,aa2} - ND_{codon1,codon2}$, where aa1 and aa2 are the amino acids corresponding to codon1 and codon2, $S_{aa1,aa2}$ is the BLOSUM62 score between aa1 and aa2, and $ND_{codon1,codon2}$ is the number of nucleotide sites differing between the two codons, e.g., $ND_{AAA,CAA} = 1$, $ND_{AAG,CAA} = 2$, and so on. One can replace BLOSUM62 with other amino acid score matrices. I have also implemented all 18 known genetic codes documented in www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi.

The second improvement chooses Alignment 2 (Fig. 1c) over Alignment 1 (Fig. 1b) by post-alignment adjustment. To simplify exposition of the rationale for post-alignment adjustment, I use a simple scoring scheme of a match score of 1, a transition substitution score of −1 and a transversion substitution score of −2. The CA dinucleotide (colored red in Fig. 2a) corresponds to a score of

$$S_{CA} = S_{C/C} + S_{A/U} = 1 + (-2) = -1 \qquad (1)$$

Any adjustment would need to have a CA score ($S_{CA.New}$) that is greater than the original S_{CA}.

Given the alignment in Fig. 2a, there are four possible sub-codon adjustments given the constraint that gap length is three or multiples of three: (1) shifting the dinucleotide CA (colored red[1]) in Fig. 2a to the right of the gap (Fig. 2b), which leads to no gain ($S_{CA.New} = S_{CA} = -1$), (2) shifting the dinucleotide UU (colored blue) in Fig. 2a to the left of the gap (Fig. 2d), which leads to a net loss of −6 ($S_{UU} = 2$ and $S_{UU.New} = -4$), (3) shifting the nucleotide A (colored red) at the left of the gap (Fig. 2a) to the right of the gap (Fig. 2c), which leads to a net gain of 3 ($S_A = -2$ and $S_{A.New} = 1$), and (4) shifting the nucleotide U at the right of the gap (Fig. 2a) to the left of the gap (Fig. 2e), which leads to a net loss of −3 ($S_U = 1$, $S_{U.New} = -2$). Thus, among the four possible adjustments, only the shift in Fig. 2c leads to a net gain, resulting in the alignment in Fig. 2c replacing the original alignment in Fig. 2a. PhyPA automatically checks the two sequences and make these post-alignment adjustments.

Note that the post-alignment adjustment above is particularly useful when one apply nucleotide-based substitution models in phylogenetic reconstruction. For codon-based substitution models, such adjustment should not be used because it would produce partial codons. Nucleotide-based models are generally more robust in phylogenetic reconstruction than codon-based models. However, codon-based sequence alignment is generally better than nucleotide-based alignment.

For very long sequences, PhyPA will search ungapped string matches between the two sequences and used these as anchors so that only sequence segments between anchors need to be

[1] For interpretation of color in Figs. 2, 7, 8, 11, the reader is referred to the web version of this article.

```
(a) Original sequences:
      Arg Ala Gly Lys
Seq1: CGA GCA GGU AAA
Seq2: CGA GCU AAA
      Arg Ala Lys

(b) Alignment 1:
      Arg Ala Gly Lys
Seq1: CGA GCA GGU AAA
Seq2: CGA GCU --- AAA

(c) Alignment 2:
Seq1: CGA GCA GGU AAA
Seq2: CGA GC- --U AAA
```

Fig. 1. Illustrating the problem of codon sequence alignment guided by aligned amino acid sequences. (a) Two codon sequences (Seq1 and Seq2) with their respectively coded amino acid sequences, (b) alignment obtained by translating the codon sequences to amino acid sequences, aligning the amino acid sequences and finally aligning the codon sequences according to aligned amino acid sequences. The two codon sequences in this alignment differ by a transversion and a triplet indel. (c) Alignment 2 in which the two codon sequences differ by only a triplet indel, and is more parsimonious than Alignment 1.

aligned with the computation-intensive dynamic programming algorithm. This allows PhyPA to build trees with long sequences (I used it to build trees with titin coding sequences of which the longest sequence is about 100,000 nt). Anchored alignment function is available only in the Windows version of DAMBE.

2.2. Estimation of evolutionary distances

PhyPA includes a variety of distances for nucleotide and amino acid sequences. For nucleotide sequences, PhyPA implements simultaneously estimated (SE) maximum composite likelihood distances (Tamura et al., 2004) based on the F84 (Felsenstein and Churchill, 1996; Kishino and Hasegawa, 1989) and TN93 (Tamura and Nei, 1993) substitution models. The SE distances have three advantages over the independently estimated (IE) distances (Tamura et al., 2004; Xia, 2009; Xia and Yang, 2011). Two statistical frameworks have been used to derive SE distances, the likelihood framework and the least-squares framework. A numerical illustration of deriving SE distances based on the likelihood and the least-squares framework is in Supplemental file SE.pdf. The SE distances implemented in PhyPA are MLCompositeF84 and MLCompositeTN93 in the likelihood framework. For amino acid sequences, PhyPA uses the Poisson-corrected and Grishin (Grishin, 1995) distances. One may also use distances based on empirical amino acid substitution matrices.

2.3. Distance-based phylogenetic reconstruction

PhyPA uses neighbor-joining (Saitou and Nei, 1987) and FastME (Desper and Gascuel, 2002, 2004) methods for phylogenetic reconstruction. FastME adopts a global optimization criterion with extensive nearest-neighbor interchange (NNI) to search through tree space and, on average, performs better than the neighbor-joining method. One minor extension of FastME is to use neighbor-joining (NJ) to generate an NJ tree, feed the tree to FastME as the initial tree, perform extensive NNI, and compare if the resulting tree is the same as the FastME tree obtained without an initial tree as a check of thoroughness of searching through the tree space.

3. Comparing phylogenetic accuracy

I compared phylogenetic accuracy of PhyPA against likelihood methods represented by PhyML (Guindon and Gascuel, 2003) and PROML/DNAML in the PHYLIP package (Felsenstein, 2014) using MSA generated from MAFFT (Katoh et al., 2009) and MUSCLE (Edgar, 2004a,b). Hereafter I refer this combination of PhyML/PROML/DNAML and MAFFT/MUSCLE as the (ML + MSA) approach.

3.1. Sequence simulation

Two sequence simulation packages, INDELibleV1.03. (Fletcher and Yang, 2009) and indel-Seq-Gen (Strope et al., 2009) have similar functions for sequence simulation. I used INDELible only because I am more familiar with it. Nucleotide, amino acid and codon sequences were simulated with different tree shapes and tree lengths. Two contrasting topologies of 24 leaves (Fig. 3),

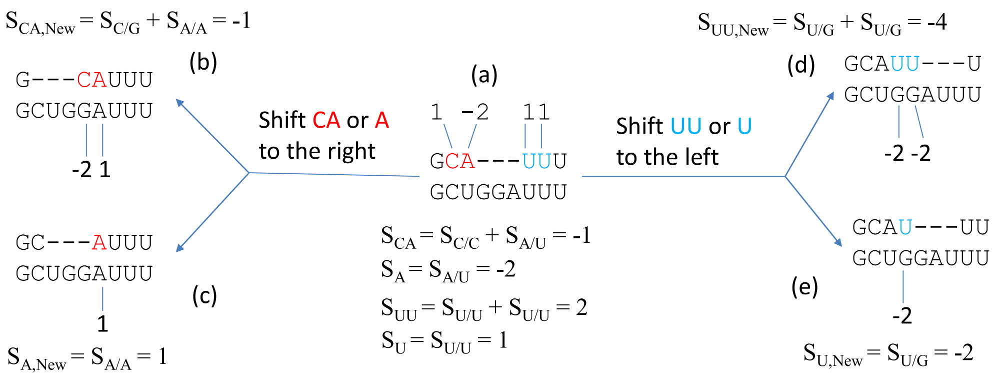

Fig. 2. Post-alignment adjustment for codon sequences with a match score of 1, a transition score of −1 and a transversion score of −2. Shown are the original alignment (a) and four possible adjustments, together with associated scores before (e.g., S_{CA}) and after (e.g., $S_{CA.New}$) the adjustment. Only the adjustment shown in (c) is better than the original.

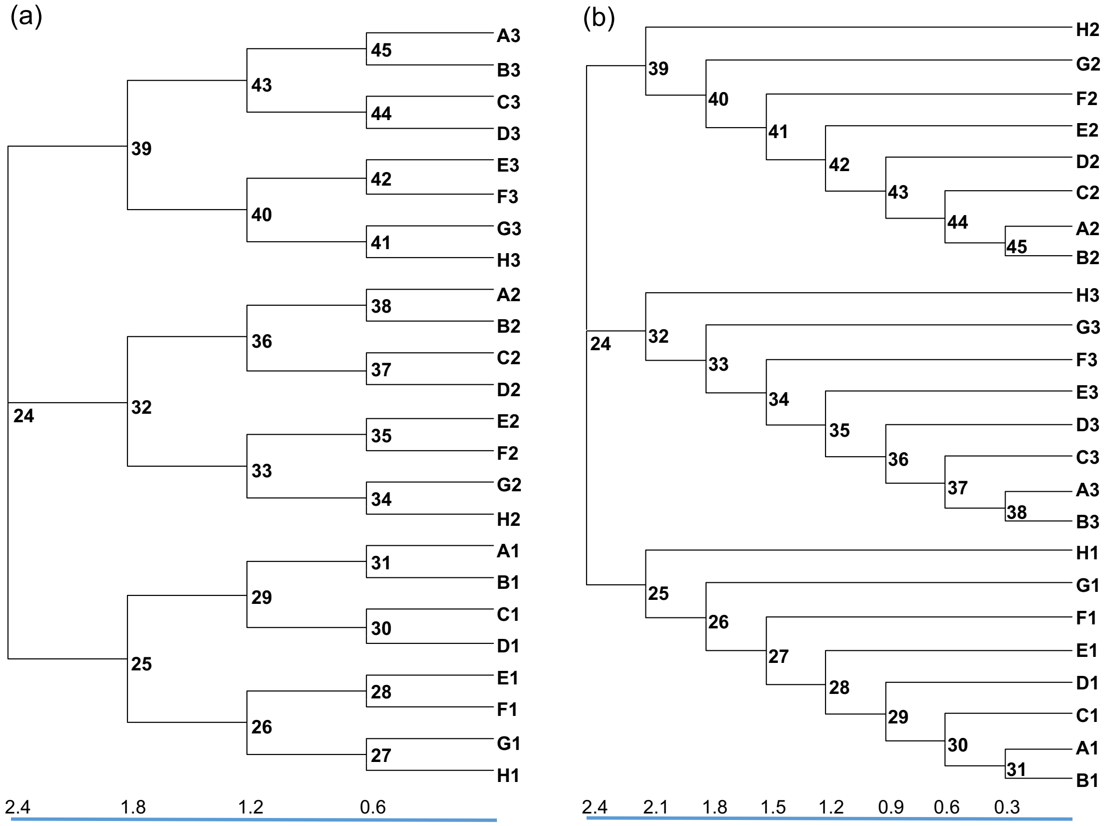

Fig. 3. Symmetric (Sym) and asymmetric (Asym) trees used for sequence simulation, with internal nodes numbered. Each internal branch breaks the 24 OTUs into two groups (a bipartition). Two additional trees with branch lengths half as long, are also used for simulation and designated as SymHalf and AsymHalf, respectively. The branch lengths, shown at the bottom, pertain to simulation of amino acid and codon sequences. The branch lengths of four corresponding trees (Sym, Asym, SymHalf, AsymHalf) for nucleotide sequence simulation are half as long those for amino acid and codon sequence simulation.

designated as Sym (for symmetric tree) and Asym (for asymmetric tree), respectively, were used for sequence simulation. The scales of branch lengths indicated in Fig. 3 are for amino acid and codon sequences. The branch lengths of Sym and Asym trees for simulating nucleotide sequences are half as long as those indicated in Fig. 3. In addition to these Sym/Asym trees, I also used SymHalf/AsymHalf trees with branch lengths half as long as Sym/Asym trees for simulation. The rationale for choosing the branch lengths, indel length and indel frequencies is to cover a range over which true trees are recovered from 100% to nearly 0% by the phylogenetic methods studied. The control files with simulation details such as substitution models, indel length and indel frequencies, starting sequence length, and trees with branch lengths are included in the supplemental file MethodDetails.docx to facilitate reproduction of results in the paper. Three random trees with varying branch lengths are also used but the results from them are very similar to those from the symmetric tree.

For each of the four trees and each of the three sequence types (amino acid, codon and nucleotide), 100 sets of sequences were generated by INDELible in FASTA format. The simulated data thus contain 12 FASTA files (four trees by three sequence types), each containing 100 sequence data sets, and each data set containing 24 sequences. These simulated sequences are available at http://dambe.bio.uottawa.ca/SimulatedSeq.asp which can be used to replicate the results in this paper. If one uses the control files in the supplemental file MethodDetails.docx to re-simulate the sequences and re-do the analysis, then the result will not be exactly the same as reported here but will exhibit the same pattern and uphold the same conclusion.

For introducing rate heterogeneity into the sequences, I simulated three sets of sequences with different substitution rates and then concatenated them together. The control file in the MethodDetials.docx file gives an example of this approach. Sequences with rate heterogeneity were subsequently analyzed by ML method with estimated alpha parameter for gamma distributed rates.

3.2. Phylogenetic reconstruction by PhyPA

PhyPA is implemented in DAMBE (Xia, 2013). Each sequence file with 100 sets of simulated unaligned sequence data (each set with 24 simulated sequences) was read into DAMBE. One chooses (1) a score matrix (e.g., a transition bias matrix for nucleotide sequences or BLOSUM62 for amino acid sequences) and gap open and gap extension penalties to perform PSA, (2) an evolutionary distance to estimate (e.g., simultaneously estimated TN93 distance for nucleotide sequences or Poisson-corrected distance for amino acid sequences), and (3) a distance-based phylogenetic reconstruction method (e.g., FastME).

3.3. Multiple sequence alignment

The simulated unaligned nucleotide and amino acid sequences were aligned by MAFFT (Katoh et al., 2009) and MUSCLE (Edgar, 2004a,b) which lead to higher phylogenetic accuracy than Clustal (Thompson et al., 1994). The LINSI option that generates the most accurate alignment ('–localpair' and '–maxiterate = 1000') is used for MAFFT. The accuracy of MAFFT with the LINSI option is exemplified by the observation that the MergeAlign approach

(Collingridge and Kelly, 2012) can improve little on the MSA obtained by MAFFT + LINSI. For MUSCLE, the default option is the most accurate. Simulated codon sequences were aligned by (1) translating into amino acid sequences, (2) aligning the amino acid sequences by MAFFT/MUSCLE, and (3) aligning the codon sequences against the aligned amino acid sequences.

I compared performance of MAFFT against MUSCLE by the following approach. The MSAs from MAFFT and MUSCLE were used in phylogenetic reconstruction and the one that recovers more true trees or subtrees is the better alignment program. MAFFT performs slightly (but not consistently) better than MUSCLE. MAFFT tend to produce more 5′ end indels for nucleotide sequences than MUSCLE, suggesting that the two programs penalize 5′ end gaps differently. The results for the ML + MSA approach reported in this paper are based on MSAs derived from MAFFT.

The alignment of multiple sets of sequences in a file is automated in DAMBE (Xia, 2013). To align a file with 100 sets of sequences with each set containing sequences for 24 OTUs, one clicks 'File|Open file with multiple data sets' to read in the file, enter 24 to specify the number of sequences per set, select either 'Built-in ClustalW', 'External MAFFT' or 'External MUSCLE'. If MAFFT is chosen, then browse to the directory containing MAFFT executable (mafft.bat), set alignment options, and click 'Run'. DAMBE will then align all 100 sets of sequences and give the option of writing the multiple sets of alignment sequences in PHYLIP or other sequence format. The same procedure is for aligning with external MUSCLE or built-in CLUSTALW. This function of using external alignment programs is available only in the Windows version of DAMBE.

3.4. Phylogenetic reconstruction by the likelihood method

For PhyML, Blosum62 is used for amino acid substitution model, and HKY85 for nucleotide substitution model with estimated transition/transversion ratio (Blosum62 and HKY85 are the models used in sequence data simulation). The tree improvement option '-s' was set to 'BEST' (best of NNI and SPR search). The '-o' option was set to 'tlr' which optimizes the topology, the branch lengths and rate parameters. For sequences with rate heterogeneity, six categories of rates were used to estimate the shape parameter (α) of the gamma distribution.

For PROML/DNAML, all optimization options such as "not rough" and "Global rearrangement" were turned on and input order was randomized four times. These options are absolutely essential in PROML/DNAML because resulting trees will often be poor without them. The option of "Henikoff/Tillier PMB" was set in PROML which corresponds to the BLOSUM option used in simulating amino acid sequences. PROML is extremely slow and each file with 100 data sets was broken into 20 files each with five sets of sequences and submitted as individual jobs to computer servers. For nucleotide sequences analyzed with DNAML, HKY85 was chosen as the substitution model (which corresponds to the model used in simulation).

PROML/DNAML do not automatically estimate transition/transversion ratio (F84R) or the α parameter of the gamma distribution. I have added a function in the Windows version of DAMBE to automate this process by running PROML/DNAML repeatedly. In short, when both F84R and α need to be estimated, a simplex method is used; when only one of the parameters needs to be estimated a Brent method is used. These methods are taken from the Numerical Recipes (Press et al., 1992). To use these functions in DAMBE, click 'File|Read standard sequence file' to read in a set of aligned sequences, click 'Phylogenetics|Run external program|Phylip' and choose either DNAML or PROML, browse to the directory where PHYLIP executables reside (e.g., C:\phylip-3.69\exe) and click 'OK'. In the next dialog, choose to estimate 'Transition/transversion ratio' and 'Gamma|Estimate'. Click OK and DAMBE will generate a ML tree with estimated F84R and α that maximize the likelihood. To use this function with multiple sets of sequences, click 'File|Open file with multiple data sets'. The rest is the same as above.

3.5. Bipartition analysis

Each internal branch in a tree breaks the 24 OTUs into two mutually exclusive sets of OTUs (bipartitions). A bipartition is typically referred to by the smaller set. For example, the internal branch flanked by node 43 and 45 (Fig. 3a) correspond to the bipartition with one set of OTUs {A3, B3} and the other set with all other OTUs. This bipartition is referred to as bipartition (or partition) {A3, B3}. The topologies with 24 OTUs (Fig. 3) used in simulation each have 21 unique bipartitions. If the same topology is recovered from phylogenetic reconstruction, then all 21 bipartitions will be recovered, otherwise only a fraction of the bipartitions will be recovered. The proportion of true bipartitions recovered can therefore be used as a measure of accuracy for comparing different phylogenetic methods (Robinson and Foulds, 1981). I performed bipartition analysis, also implemented in DAMBE (Xia, 2013), to see which of the three methods (PhyPA, PhyML and DNAML/PROML) produce phylogenetic trees sharing the highest proportion of bipartitions with the true trees (i.e., trees used in sequence simulation). The 21 bipartitions from the true trees will be referred to as true bipartitions hereafter.

4. Results

I have made comparisons of PhyPA against ML + MSA with all optimized options turned on for the latter. PhyPA performs consistently better than ML + MSA when the topology is symmetrical. For an extremely asymmetrical topology, PhyPA performs consistently better than ML + MSA when sequences are highly diverged, but ML + MSA performs better with less diverged sequences (i.e., when reliable MSA can be obtained). The supplemental file MethodDetails2.docx details the procedures for replicating results in the paper.

4.1. Amino acid sequences

For amino acid sequences simulated with the SymHalf tree (Fig. 4a, which have branches half as long as those in the Sym tree in Fig. 4b), true bipartitions are 100% recovered with both PhyPA and ML + MSA methods (Fig. 4a). The two methods, however, differ in recovering deep bipartitions with longer branch lengths. PhyPA recovered substantially more true bipartitions relative to the ML + MSA approach (Fig. 4b). For example, out of 100 sets of sequences, the bipartition including OTUs {A3, B3, C3, D3} were recovered 53 times by PhyPA, but only 33 and 35 times by PhyML and PROML, respectively (Fig. 4b). This pattern is consistent for all the deep nodes (Fig. 4b, numbers highlighted in red).

For the asymmetric topology, the ML + MSA approach recovered more true bipartitions than PhyPA when the sequences are not too diverged (see the AsymHalf tree in Fig. 5a, numbers highlighted in blue). Thus, when sequence divergence does not rule out reliable MSA, ML + MSA is better than PhyPA. However, this advantage disappears quickly when sequences are more diverged. In Fig. 5b where the Asym tree has branches twice as long as those in the AsymHalf tree in Fig. 5a, PhyPA consistently recovered more true bipartitions than ML + MSA (Fig. 5b, numbers highlighted in red).

One alternative way of visualizing the effect of sequence divergence on the efficiency of different methods in recovering

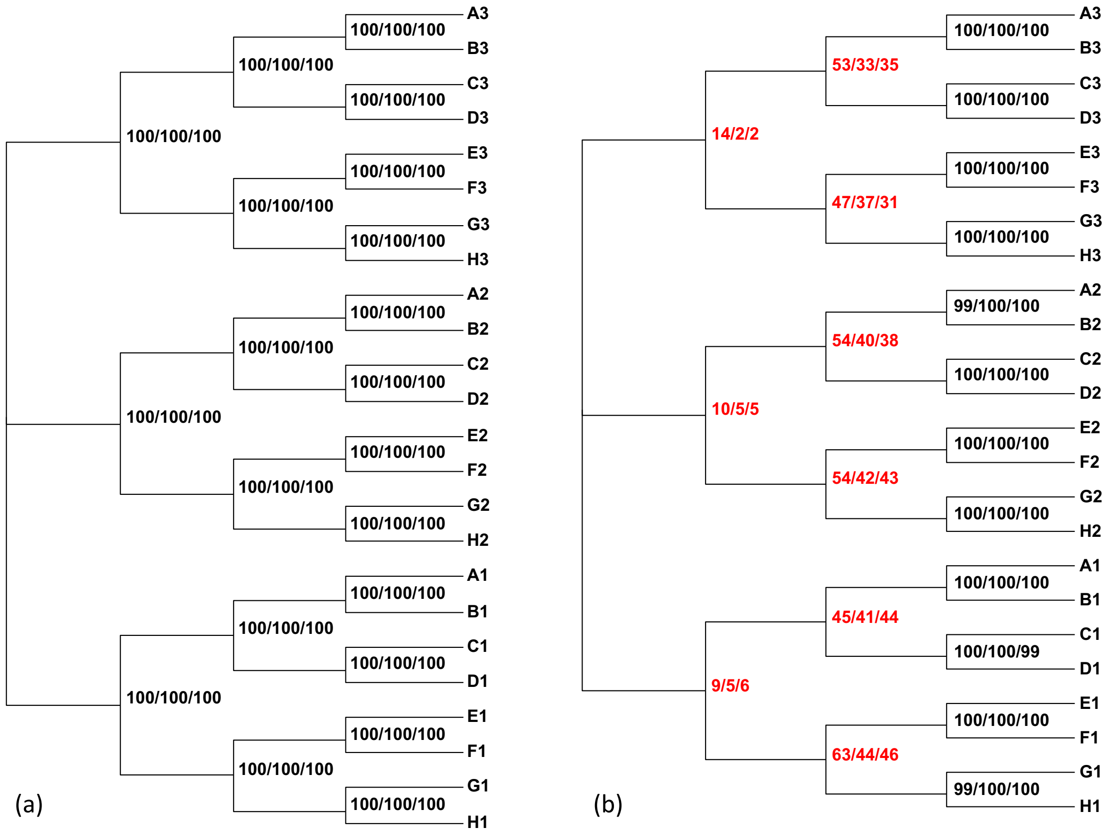

Fig. 4. Contrasting phylogenetic performance between PhyPA and (PhyML/PROML + MAFFT) when optimal options were used in both MAFFT and PhyML/PROML, based on simulated amino acid sequences evolving along the SymHalf (a) and Sym (b) trees. Shown at each bifurcating nodes are the percentage of true bipartitions recovered by the phylogenetic methods, in the format of PhyPA/PhyML/PROML and highlighted red when PhyPA outperforms (PhyML/PROML + MAFFT). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the true bipartitions is to plot the proportion of the true bipartitions recovered ($P_{true.bipartition}$) against the node depth (D_{node}) which is measured by the distance from the node to its descendent terminal leaves. Take the Sym tree in Fig. 3a for example. D_{node} is 1.2 for the internal node 43 containing the bipartition {A3, B3, C3, D3}, 1.8 for the internal node 39 containing bipartitions {A3, B3, C3, D3, E3, F3, G3, H3}, and so on. Similarly for the Asym tree in Fig. 3b, D_{node} is 0.3 for the internal node 45 with bipartition {A2, B2}, and 0.6 for node 44 with bipartition {A2, B2, C2}.

$P_{true.bipartition}$ decreases with increasing D_{node} for both PhyPA and ML + MSA (Fig. 6). However, the decrease is slower with PhyPA than with ML + MSA for both symmetric and asymmetric topologies (Fig. 6a and b, respectively). PhyPA recovers more true bipartitions than ML + MSA when D_{node} is 1.2 or greater for the symmetric and the asymmetric tree (Fig. 6). There is a consistent range of sequence divergence where PhyPA recovers more true bipartitions than ML + MSA. However, ML + MSA outperforms PhyPA when sequences are less diverged. Thus, PhyPA should be used for phylogenetic analysis of highly diverged sequences.

One may find it peculiar that the phylogenetic performance decreases with even moderate D_{node} values. For sequence simulations done without indels, both PhyPA and ML + MSA can recover true trees or subtrees with much higher D_{node} values. However, the introduction of indels aggravates the problem of substitution saturation (Xia and Lemey, 2009; Xia et al., 2003b) by making it more difficult to identify site homology correctly, leading to decrease in phylogenetic performance with even moderate sequence divergence.

4.2. Codon sequences

The phylogenetic results from simulated codon sequences are consistent with those for simulated amino acid sequences. PhyPA recovers more true bipartitions than ML + MSA for symmetric trees (Fig. 7, numbers highlighted in red), or for asymmetric trees with highly diverged sequences (Fig. 8b, numbers highlighted in red). For asymmetric trees with limited sequence divergence, ML + MSA recovered more true bipartitions than PhyPA (numbers highlighted in blue in Fig. 8a where the branch lengths are half as long as those in Fig. 8b).

As in the analysis of amino acid sequences, $P_{true.bipartition}$ decreases with increasing D_{node} (Fig. 9), but the decrease is slower with PhyPA than with ML + MSA for both symmetric and asymmetric topologies (Fig. 9a and b, respectively), consistent with the pattern observed with amino acid sequences (Fig. 6).

The codon-based alignment coupled with the post-alignment adjustment (Fig. 2) increases phylogenetic accuracy dramatically. If we align the codon sequences as nucleotide sequences, not only will the alignment takes longer (as the sequences are three times longer), but the true bipartitions associated with the deep node also become hardly recoverable. For this reason, ribosomal RNA sequences, albeit being claimed as the universal yardstick in phylogenetics, cannot really trace history back very far without elaborate sequence alignment incorporating secondary structure information (Xia et al., 2003a). Coding sequences should be better than non-coding sequences in recovering true phylogeny among highly divergent taxa.

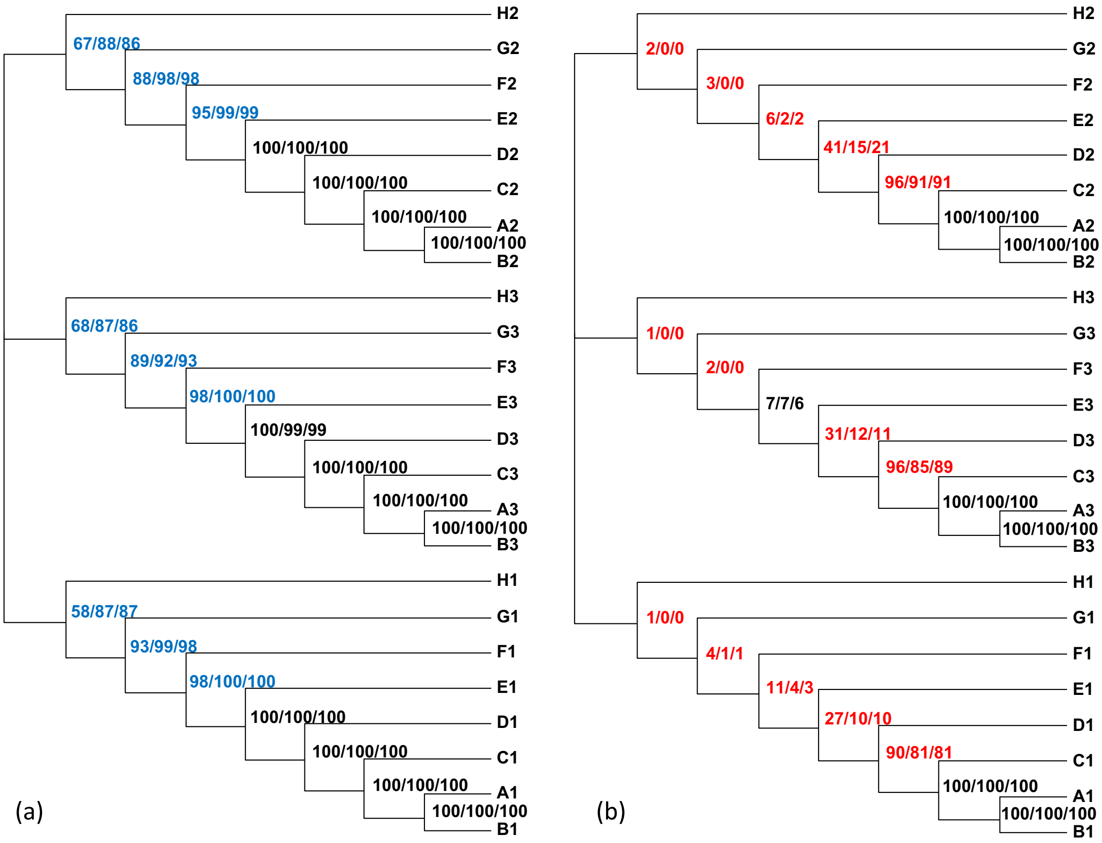

Fig. 5. Contrasting phylogenetic performance between PhyPA and optimized (PhyML/PROML + MAFFT), based on simulated amino acid sequences evolving along the AsymHalf (a) and Asym (b) trees. Shown at each bifurcating nodes are the percentage of true bipartitions recovered by different approaches, in the format of PhyPA/PhyML/PROML. Percentage values are highlighted red when PhyPA outperforms (PhyML/PROML + MAFFT) and blue when opposite. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3. Nucleotide sequences

The results from simulated nucleotide sequences are consistent with those from simulated amino acid or codon sequences in that (1) PhyPA recovers more true bipartitions than ML + MSA for sequences simulated on the symmetric tree (Fig. 10b, numbers highlighted in red), and (2) for sequences simulated on the asymmetric tree, ML + MSA is better than PhyPA for sequences with limited divergence (for Fig. 11a, numbers highlighted in blue), but worse than PhyPA for highly diverged sequences (Fig. 11b, numbers highlighted in red).

The effect of sequence divergence on phylogenetic performance occurs earlier for nucleotide sequences than for amino acid or codon sequences (note the difference in D_{node} value among Fig. 12 for nucleotide sequences and those in Fig. 9 for codon sequences or Fig. 6 for amino acid sequences). Sequence alignment (identification of site homology) is generally easier for codon sequences and amino acid sequences than for nucleotide sequences, which can explain why $P_{true.bipartition}$ decreases faster with D_{node} for nucleotide sequences than for the other two type of sequences.

4.4. Rate heterogeneity and long-branch attraction

For sequences with introduced rate heterogeneity, phylogenetic analysis was done with (1) gamma-distributed rate and (2) assuming the same rate over all sites. The former does not exhibit consistent improvement of phylogenetic accuracy and in fact produced trees that are slightly worse than the latter (as well as worse than PhyPA which assumes constant rate over sites because gamma-distributed rates cannot be estimated from pairwise alignment). It is likely that the highly diverged sequences leads to poor MSA that in turn leads to wrong estimation of site heterogeneity and phylogenetic tree by the ML + MSA approach.

The ML + MSA approach performs much better than PhyPA when sequences are not highly diverged and when non-sister taxa have long branches, with the former being much more robust against long-branch attraction than the latter. However, this advantage of ML + MSA diminishes with increasing divergence, likely because long branches obliterate MSA and phylogenetic signals. Details on rate heterogeneity and long-branch attraction will be presented in a separate paper.

4.5. Why PhyPA performs better than MA + MSA for highly diverged sequences?

One may suspect that the inferior performance of the ML + MSA approach is due to insufficient search of tree space by the ML method. This is not the case. The ML topologies reported by the ML + MSA consistently have higher log-likelihood (lnL) than the true topology evaluated with the same MSA (Fig. 13). Thus, the failure to recover the true topology by the ML + MSA approach involving highly diverged sequences is not because the true topology was not encountered by the search algorithm, but because it has a relatively small lnL and consequently is discarded by the likelihood criterion. When the true alignment from the sequence simulation is used, the ML approach (both DNAML and PhyML) recovered true topologies 100% in ALL simulated sequences, and with much high

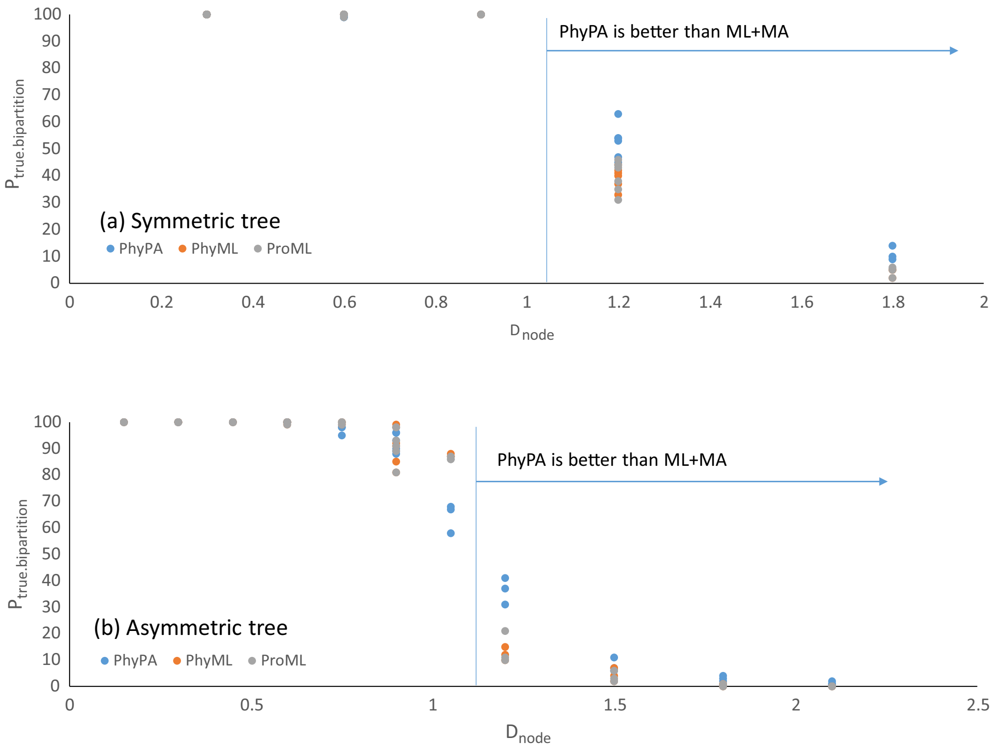

Fig. 6. Percentage of true bipartitions recovered ($P_{true.bipartition}$) by PhyPA and (PhyML/PROML + MAFFT) decreases with increasing sequence divergence (D_{node}), but the decrease is slower for PhyPA than for (PhyML/ProML + MAFFT), based on simulated amino acid sequences on symmetric (a) and asymmetric (b) trees. The range of sequence divergence (D_{node}) over which PhyPA recovers more true bipartitions than (PhyML/ProML + MAFFT) is indicated.

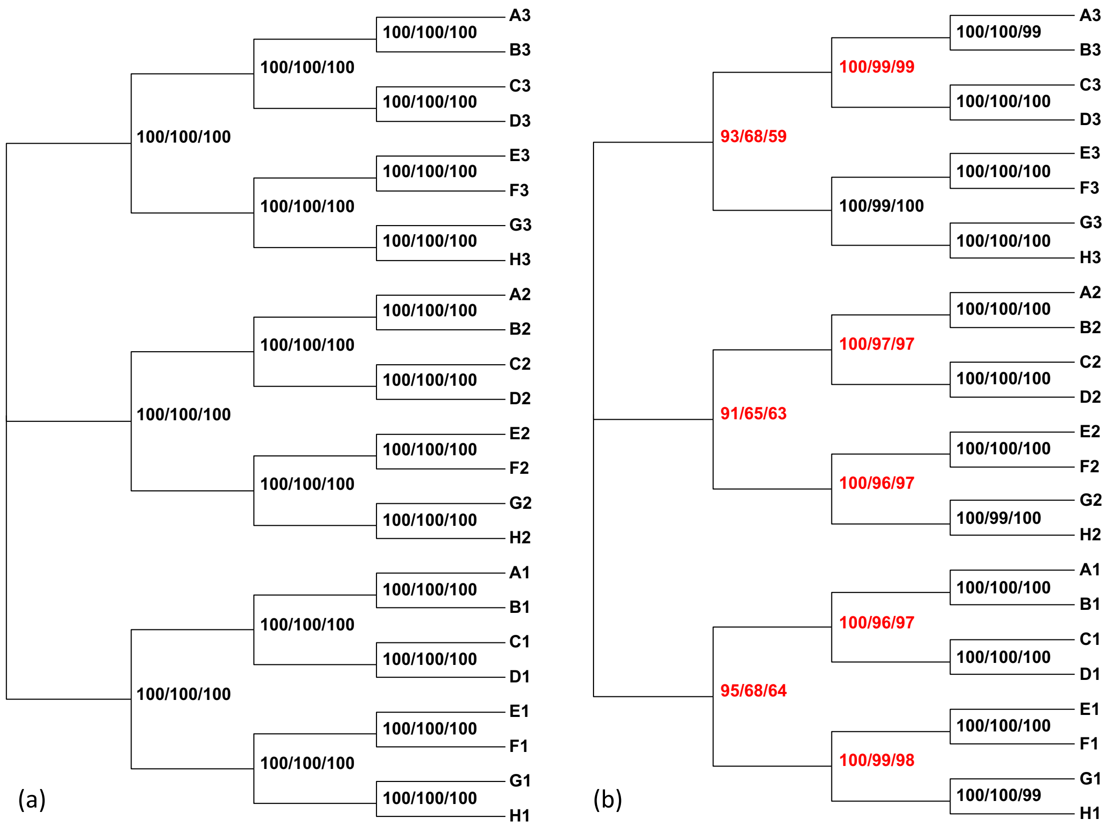

Fig. 7. Contrasting phylogenetic performance between PhyPA and optimized (PhyML/DNAML + MAFFT), based on simulated codon sequences evolving along the SymHalf (a) and Sym (b) trees. Shown at each bifurcating nodes are the percentage of true bipartitions recovered by the three approaches, in the format of PhyPA/PhyML/DNAML.

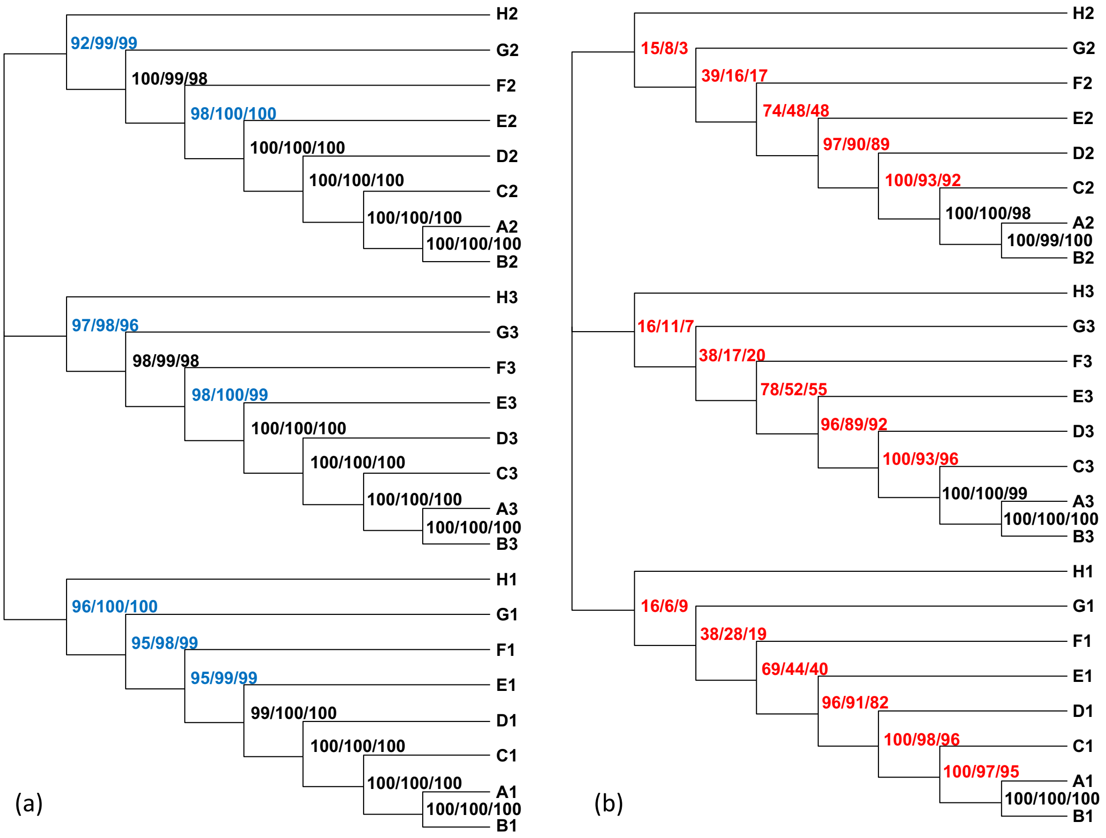

Fig. 8. Contrasting phylogenetic performance between PhyPA and optimized (PhyML/DNAML + MAFFT), based on simulated codon sequences evolving along the AsymHalf (a) and Asym (b) trees. Shown at each bifurcating nodes are the percentage of true bipartitions recovered by the three approaches, in the format of PhyPA/PhyML/DNAML.

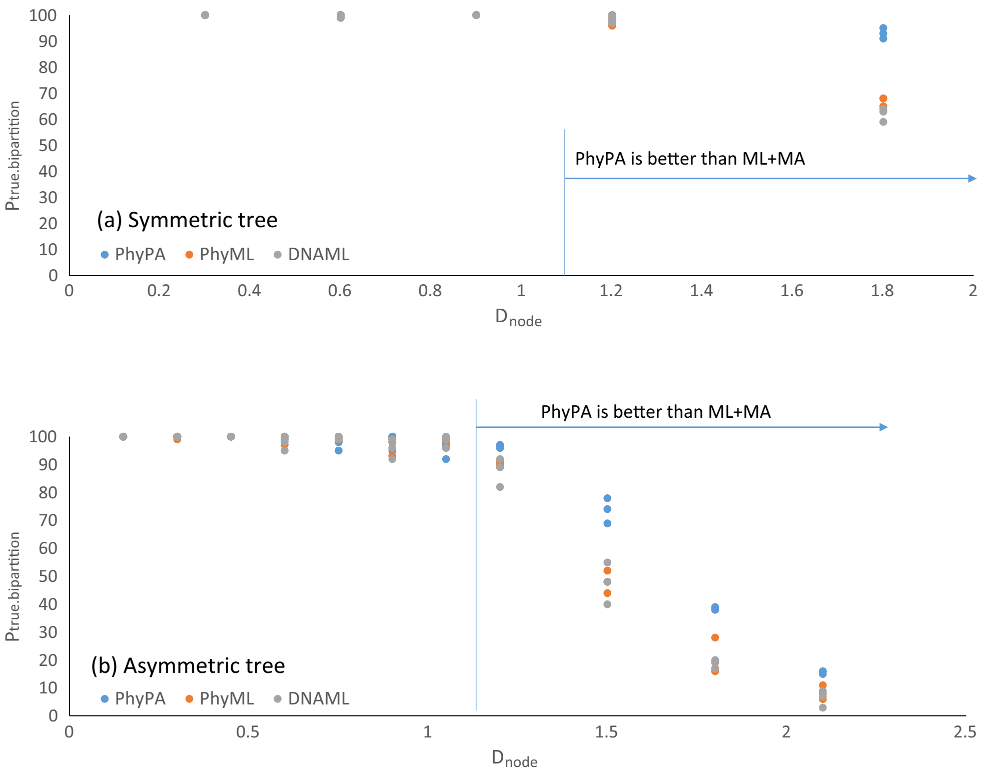

Fig. 9. Percentage of true bipartitions recovered ($P_{true.bipartition}$) through phylogenetic reconstruction decreases with increasing sequence divergence (D_{node}), but the decrease is slower for PhyPA than for (PhyML/DNAML + MAFFT), based on simulated codon sequences on symmetric (a) and asymmetric (b) trees. The range of sequence divergence (D_{node}) over which PhyPA recovers more true bipartitions than (PhyML/DNAML + MAFFT) is indicated.

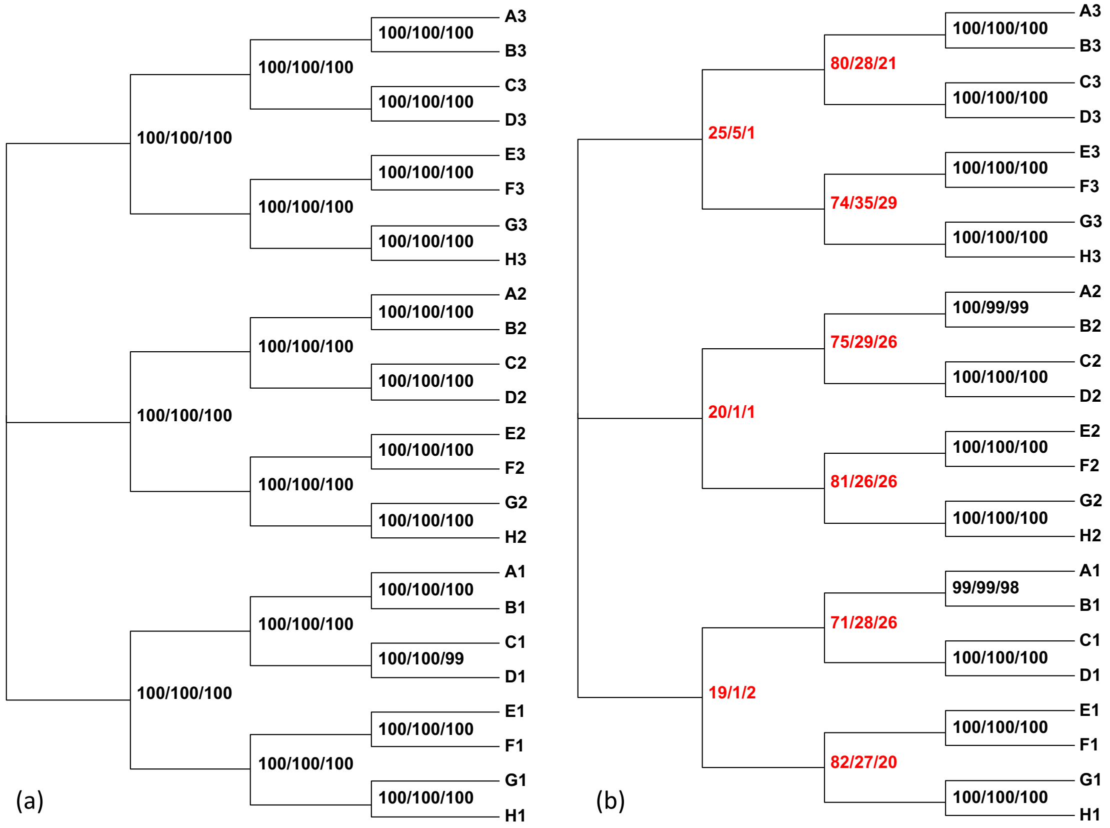

Fig. 10. Contrasting phylogenetic performance between PhyPA and optimized (PhyML/DNAML + MAFFT), based on simulated nucleotide sequences evolving along the SymHalf (a) and Sym (b) trees. Shown at each bifurcating nodes are the percentage of true bipartitions recovered by the three approaches, in the format of PhyPA/PhyML/DNAML.

ln L values (Fig. 13). Unfortunately, true alignment is never known for real sequences.

Another line of evidence confirms that the inferior performance of the ML + MSA approach relative to PhyPA involving highly diverged sequences is due to distortion of phylogenetic signals introduced during MSA, but not by the ML approach. When FastME is applied to the MSA derived from MAFFT, then the recovered trees are in general worse than those from the ML methods. Thus, the outstanding performance of PhyPA is attributable to the pairwise alignment, not to the distance-based phylogenetic reconstruction.

5. Discussion

I have shown that the PhyPA approach consistently outperforms the ML + MSA approach when sequences are highly diverged. This pattern is consistently observed for amino acid, codon and nucleotide sequences, and the difference can be quite substantial. For example, for nucleotide sequences, PhyPA recovered about 80% of true bipartitions in contrast to about 25% recovered by the ML + MSA approach (Fig. 12a). As phylogenetics becomes difficult with deep phylogenies, PhyPA should be valuable in such cases.

I should also emphasize that the ML + MSA results reported here are produced with all key optimization options turned on. When I used default options for MAFFT, the ML + MSA approach is much worse in every way than PhyPA even when all the optimization options were turned on for PhyML/PROML/DNAML. Similarly, when I used default options for PhyML/PROML/DNAML, the ML + MSA approach is also much worse in every way than PhyPA even when MAFFT is optimized. Thus, for data analysis involving many highly diverged taxa, one is much better off using PhyPA than using the ML + MSA approach with default options.

It is theoretically not surprising that ML + MSA, while performing better than PhyPA with limited sequence divergence, is worse than PhyPA with highly diverged sequences. For any particular pair of sequences S1 and S2, other sequences may contribute both phylogenetic information and noise to the identification of homologous sites between S1 and S2. With low sequence divergence, phylogenetic information contributed by other sequences overwhelms noise and reduces the problem of inconsistency in homologous site identification often seen in PSA, which is illustrated in Fig. 14 (P. Foster, pers. comm.). Given the three amino acid sequences (S1 to S3) in Fig. 14a, we will get three PSAs shown in Fig. 14b with the scoring scheme defined by Blosum62, gap open equal to 20 and gap extension equal to 2. The residue W in S1 at position 3 (designated by W_{13}, where the first subscript 1 indicates the 1st sequence and the second subscript 3 indicates the 3rd site in S1) is inferred to be homologous to W_{22} based on the PSA between S1 and S2, and homologous to W_{33} based on the PSA between S1 and S3 (Fig. 14b). These two homologous site pairs (W_{13}/W_{22}, W_{13}/W_{33}) imply a homologous site pair W_{22}/W_{33} which however is not true in the PSA between S2 and S3 (Fig. 14b) where site W_{22} pairs with K_{32} instead of W_{33}. Such inconsistency in site homology identification in PSA will be less likely in a MSA (Fig. 14c) obtained with the same scoring scheme. When PSA is derived from the MSA (Fig. 14d), then two inferred homologous pairs (W_{13}/W_{22}, W_{13}/W_{33}) imply a homologous site pair W_{22}/W_{33} which is observed in Fig. 14d. The inconsistency in homology identification in PSA may explain why the ML + MSA approach is generally better than PhyPA when sequences are not highly diverged (the numbers highlight in blue in Figs. 5a, 8a, and 11a).

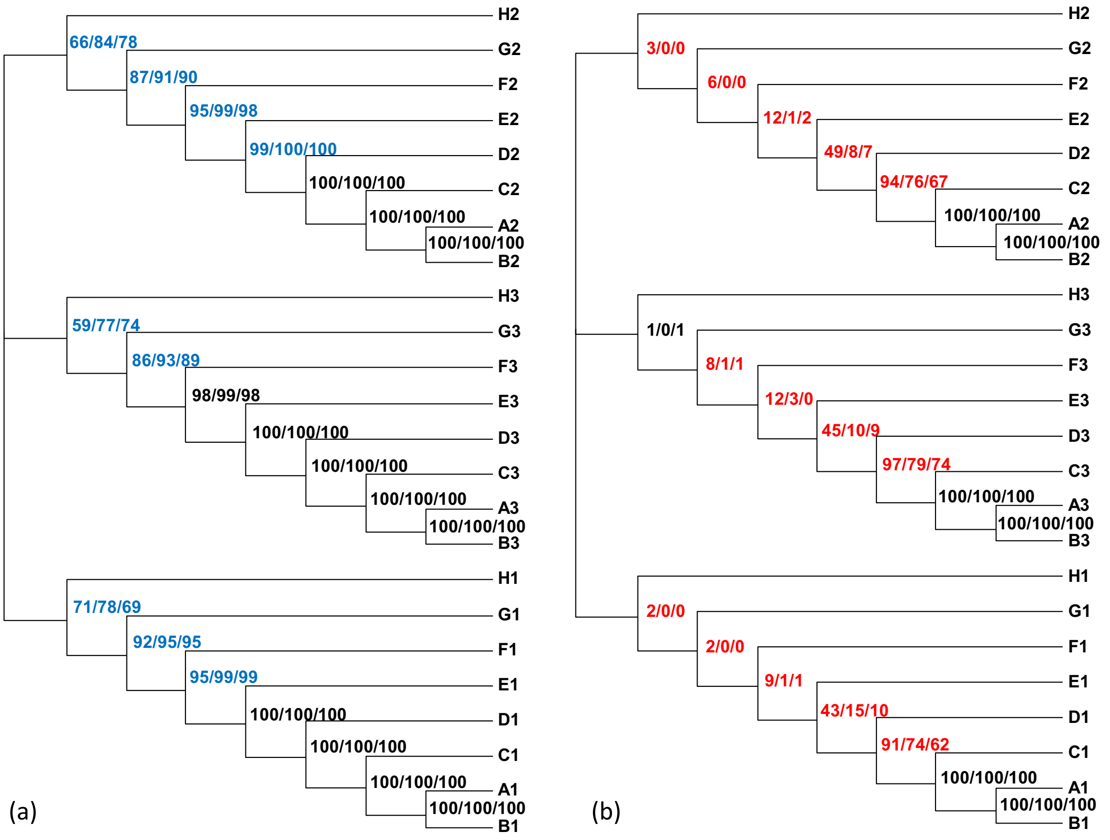

Fig. 11. Contrasting phylogenetic performance between PhyPA and optimized (PhyML/DNAML + MAFFT), based on simulated nucleotide sequences evolving along the AsymHalf (a) and Asym (b) trees. Shown at each bifurcating nodes are the percentage of true bipartitions recovered by the three approaches, in the format of PhyPA/PhyML/DNAML.

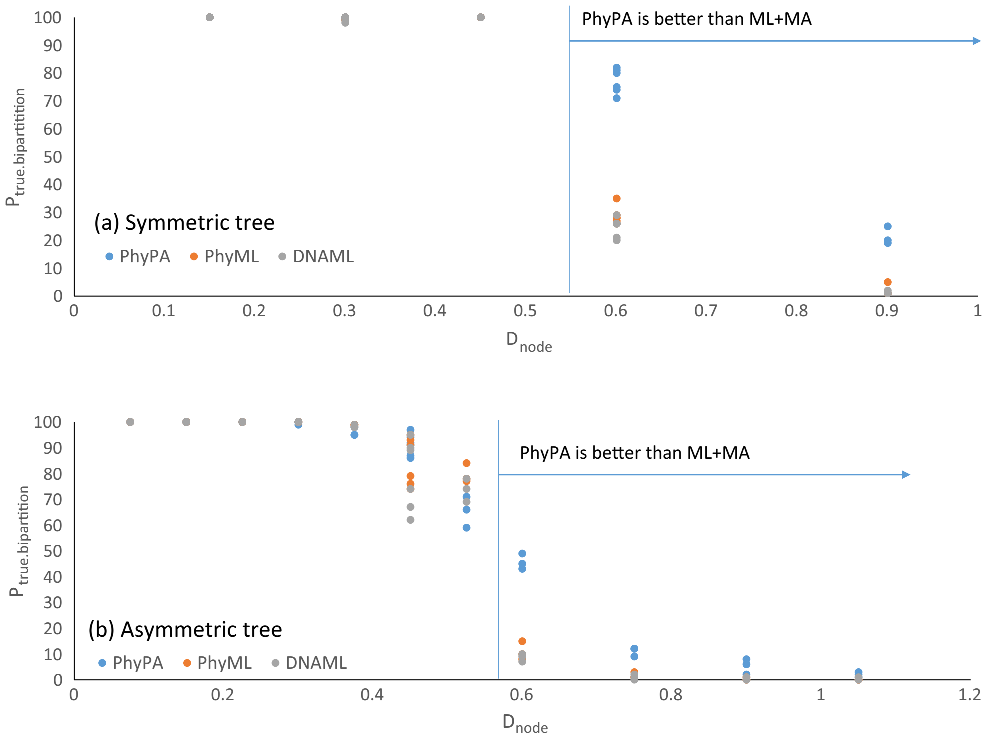

Fig. 12. Percentage of true bipartitions recovered ($P_{true.bipartition}$) decreases with increasing sequence divergence (D_{node}), but the decrease is slower for PhyPA than for (PhyML/DNAML + MAFFT), based on simulated nucleotide sequences on symmetric (a) and asymmetric (b) trees. The range of sequence divergence (D_{node}) over which PhyPA recovers more true bipartitions than (PhyML/DNAML + MAFFT) is indicated.

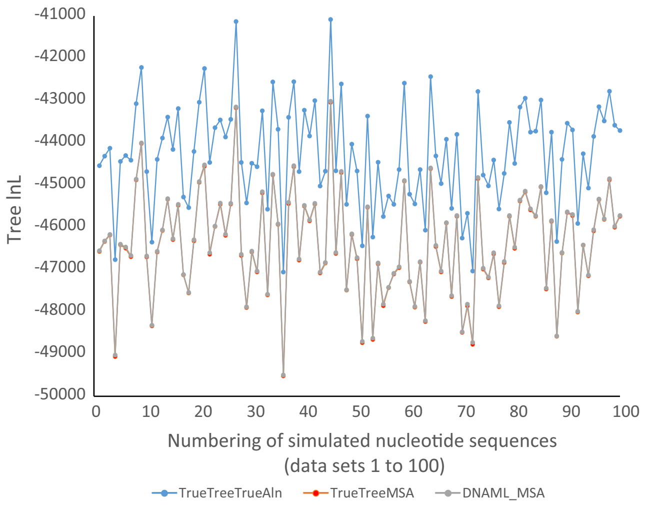

Fig. 13. Phylogenetic information from maximum likelihood analysis of 100 simulated sets of nucleotide sequences based on the HKY85 model and the Sym tree (Fig. 3a). The true topology is 100% recovered by DNAML with the true multiple alignment and their log-likelihood (ln L) is indicated by the blue line (TrueTree-TrueAln). When multiple sequence alignment (MSA) from MAFFT is used, the resulting trees exhibit much reduced ln L values (DNAML_MSA) which, however, are consistently greater than the ln L from evaluating the true topology with the same MSA (TrueTreeMSA). This implies that the failure to recover the true topology is not because of insufficient search of tree space, i.e., the true topology is discarded because of its relatively small ln L, not because it is not encountered by the search algorithm. The pattern is consistent with the simulated amino acid and codon sequences or with PhyML. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

```
(a)                   (c)
S1  DKWWGAAP          S1  DKWWGA--AP
S2  DWWGARRAP         S2  D-WWGARRAP
S3  DKWARRAP          S3  DKW--ARRAP

(b)                   (d)
S1  DKWWGA--AP        S1  DKWWGA--AP
S2  D-WWGARRAP        S2  D-WWGARRAP

S1  DKWWGAAP          S1  DKWWGA--AP
S3  DKWARRAP          S3  DKW--ARRAP

S2  DWWGARRAP         S2  D-WWGARRAP
S3  DKW-ARRAP         S3  DKW--ARRAP
```

Fig. 14. Identification of homologous sites in pairwise and MSA, illustrating the problem of inconsistency in site homology identification associated with PSA (P. Foster, pers. comm.). (a) Three amino acid sequences S1 to S3; (b) three PSAs from the three sequences in (a), with scoring scheme of Blosum62, gap open equal to 20 and gap extension equal to 2; (c) MSA of three sequences in (a) with the same scoring scheme; (d) three PSAs derived from the MSA in (c).

However, when sequence divergence has reached a level when other sequences can contribute mostly noise instead of phylogenetic information to the identification of homologous site between two sequences, PhyPA consistently performs better than ML + MSA (Figs. 4–12).

PhyPA should contribute to resolving controversies in evolutionary biology. For example, coliphages in general exhibit codon usage similar to their *Escherichia coli* host (Chithambaram et al., 2014a, 2014b; Prabhakaran et al., 2015). However, codon usage in phage PRD1 deviates much from that of *E. coli*. One hypothesis is that phage PRD1 may have switched to parasitize *E. coli* recently from a host with a different codon usage. A phylogenetic tree with mapped hosts would facilitate the test of this hypothesis. However, this is not done by the researchers presumably because sequence homology is often weak to obtain reliable MSA. PhyPA would be suitable in such cases.

One implication of the results reported here is that the limiting factor for phylogenetic accuracy involving highly diverged sequences may depend mainly on sequence alignment rather than on tree-building algorithms. Thus, there is a strong need for phylogenetic researchers to redirect their effort from refining tree-building algorithms to refining sequence alignment methods, possibly by incorporating other relevant information such as secondary structure. Improved sequence alignment of rRNA sequences with the aid of secondary structure has been shown to significantly improve phylogenetic accuracy (Xia, 2000; Xia et al., 2003a).

One disadvantage of PhyPA is that it cannot employ site-specific resampling techniques such as bootstrapping or jackknifing to generate a measure of node support on a tree. The approach of bootstrapped pseudo-replicates has been proposed before by Thorne and Kishino (1992) but this is different from the site-based bootstrapping that phylogeneticists are familiar with. Here I propose two alternatives based on the observation that more and more phylogenetic analyses involve multiple genes, e.g., 62 genes was used to elucidate arthropod phylogeny (Regier et al., 2010) and 2320 coding genes were used to study bat phylogeny (Tsagkogeorga et al., 2013). PhyPA can be applied to N genes generating N trees which can then be summarized into a consensus tree to produce node-support statistics equivalent to bootstrapping or jackknifing. Alternatively, if one has M alternative topologies and wish to know which one receives greater support from the N genes, one can analyze these N trees to see which candidate topology has more bipartitions recovered by the N trees. Both of these approaches have been implemented in DAMBE.

One shortcoming of the current study is insufficient exploration of the effect of rate-heterogeneity over sites. While it is a common practice to model rate heterogeneity by gamma distribution, partition-based approach (Zoller et al., 2015) is probably more appropriate than a blind modeling of rate heterogeneity by gamma distribution, in particular because rate heterogeneity in our sequences is introduced by concatenating sequences simulated with different rates. However, as I have emphasized, PhyPA is intended only for highly divergent sequences where reliable MSA is difficult to obtain. It is not intended to replace existing phylogenetic methods operating on reliable MSA.

Availability

I have integrated PhyPA as a function within DAMBE (Xia, 2013) to take advantage of the large number of sequence manipulation functions in DAMBE. To access the function, click 'File|Open standard sequence file' (DAMBE understands about 20 different sequence formats) to read in a set of unaligned sequences. Click 'Phylogenetics|Distance-based method|Phylogenetics by pairwise alignment' to build the tree based on PSA only. The supplemental file MethodDetails.docx contains instructions on how to analyze multiple sets of sequences. DAMBE is freely available at http://dambe.bio.uottawa.ca/dambe.asp. I offer full support for the installation and executable programs as well as unmodified source code.

Acknowledgment

This study is supported by Discovery Grant from Natural Science and Engineering Research Council (NSERC) of Canada. I thank P. Foster for sharing the problem of homologous site identification, L. Jermiin for Fig. 13, N. Rodrigue, and J. Thorne for discussion and comments, and N. Benabbou for assistance with using High Performance Computing Virtual Laboratory.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.07.001.

References

Bhardwaj, G., Ko, K.D., Hong, Y., Zhang, Z., Ho, N.L., Chintapalli, S.V., Kline, L.A., Gotlin, M., Hartranft, D.N., Patterson, M.E., Dave, F., Smith, E.J., Holmes, E.C., Patterson, R.L., van Rossum, D.B., 2012. PHYRN: a robust method for phylogenetic analysis of highly divergent sequences. PLoS ONE 7, e34261.

Blackburne, B.P., Whelan, S., 2013. Class of multiple sequence alignment algorithm affects genomic analysis. Mol Biol Evol 30, 642–653.

Chithambaram, S., Prabhakaran, R., Xia, X., 2014a. Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*. Mol Biol Evol 31, 1606–1617.

Chithambaram, S., Prabhakaran, R., Xia, X., 2014b. The effect of mutation and selection on codon adaptation in *Escherichia coli* bacteriophage. Genetics 197, 301–315.

Collingridge, P.W., Kelly, S., 2012. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinformatics 13, 117.

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

Desper, R., Gascuel, O., 2004. 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.

Edgar, R.C., 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Edgar, R.C., 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797.

Edgar, R.C., Batzoglou, S., 2006. Multiple sequence alignment. Curr Opin Struct Biol 16, 368–373.

Felsenstein, J., 2014. PHYLIP 3.695 (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle.

Felsenstein, J., Churchill, G.A., 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13, 93–104.

Fletcher, W., Yang, Z., 2009. INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26, 1879–1888.

Grishin, N.V., 1995. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 41, 675–679.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696–704.

Herman, J.L., Challis, C.J., Novak, A., Hein, J., Schmidler, S.C., 2014. Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. Mol Biol Evol 31, 2251–2266.

Hogeweg, P., Hesper, a.B., 1984. The alignment of sets of sequences and the construction of phylogenetic trees: an integrated method. J. Mol. Evol. 20, 175–186.

Katoh, K., Asimenos, G., Toh, H., 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol. 537, 39–64.

Kishino, H., 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.

Kumar, S., Filipski, A., 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res. 17, 127–135.

Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., Hein, J., 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. Genome Res. 18, 298–309.

Prabhakaran, R., Chithambaram, S., Xia, X., 2015. *E. coli* and *Staphylococcus* phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. Virology. http://dx.doi.org/10.1099/vir.0.000050.

Press, W.H., Teukolsky, S.A., Tetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

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

Strope, C.L., Abel, K., Scott, S.D., Moriyama, E.N., 2009. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Mol. Biol. Evol. 26, 2581–2593.

Tamura, K., 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.

Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA 101, 11030–11035.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucl. Acids Res. 22, 4673–4680.

Thorne, J.L., Kishino, H., 1992. Freeing phylogenies from artifacts of alignment. Mol. Biol. Evol. 9, 1148–1162.

Tsagkogeorga, G., Parker, J., Stupka, E., Cotton, J.A., Rossiter, S.J., 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr. Biol. 23, 2262–2267.

Wong, K.M., Suchard, M.A., Huelsenbeck, J.P., 2008. Alignment uncertainty and genomic analysis. Science 319, 473–476.

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. Kluwer Academic Publishers, Boston.

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., 2013. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol. Biol. Evol. 30, 1720–1728.

Xia, X., Lemey, P., 2009. Assessing substitution saturation with DAMBE. In: Lemey, P., Salemi, M., Vandamme, A.M. (Eds.), The Phylogenetic Handbook. Cambridge University Press, Cambridge, UK, pp. 615–630.

Xia, X., Xie, Z., Kjer, K.M., 2003a. 18S ribosomal RNA and tetrapod phylogeny. Syst. Biol. 52, 283–295.

Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003b. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26, 1–7.

Xia, X., Yang, Q., 2011. A distance-based least-square method for dating speciation events. Mol. Phylogenet Evol. 59, 342–353.

Zoller, S., Boskova, V., Anisimova, M., 2015. Maximum-likelihood tree estimation using codon substitution models with multiple partitions. Mol. Biol. Evol. 32, 2208–2216.