Atlas of Science
another view on science
http://atlasofscience.org

A new way to trace evolutionary history by pairwise alignment only

Genomes and genes change during evolution leading to the observed biodiversity in many colors and shades. This genealogical (phylogenetic) relationship is recorded in DNA and can be reconstructed by molecular phylogenetics and expressed in a tree with branches and leaves.

Sequence alignment is a first step in molecular phylogenetics, aiming to identify site homology (co-ancestry). Site homology is easy to identify for closely related species that differ little in their genomes, but hard for highly diverged species whose genomes may have been partly overwritten several times. In particularly, multiple sequence alignment (MSA) is typically based on tree-guided progressive alignment methods that reduce MSA to pairwise sequence alignment (PSA). This gives rise to a chicken-egg problem. To produce a good MSA, we need a good guide tree, but to have a good guide tree, we need a good MSA. The current approach is to produce an initial guide tree based on pairwise alignment scores, generate an MSA, compute an MSA quality score, use the MSA to generate another, presumably better, guide tree, and use the new tree to generate another MSA and another MSA quality score. This iteration continues until there is no improvement in MSA quality score. Unfortunately, for highly diverged sequences, the quality score changes little over iterations and the MSA remains poor, partly because the MSA quality score has many of its own problems. While PSA is guaranteed to produce an optimal alignment, MSA is not. Many studies have shown poor MSA leads to generate poor phylogenetic results.

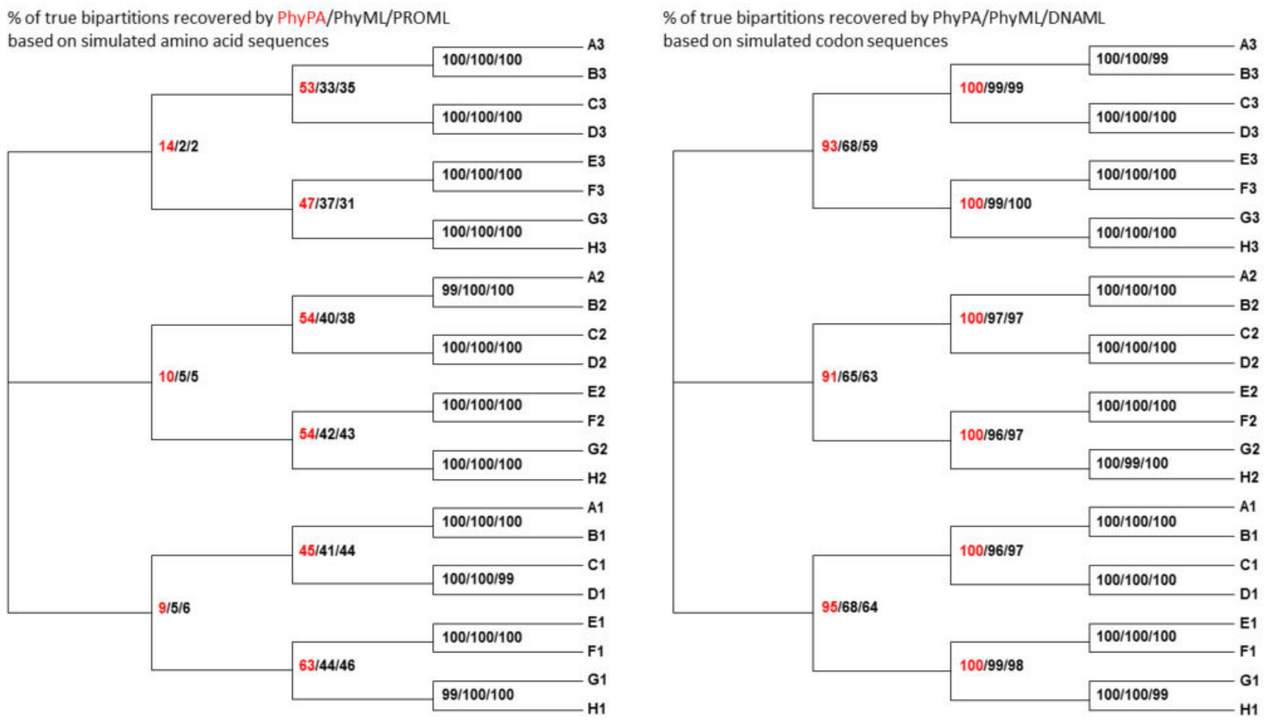

Fig. 1.

One way to avoid the chicken-egg problem is to use PSA only, and this is the approach taken by PhyPA (short for phylogenetics from pairwise alignment). It uses an improved pairwise alignment method to generate PSA, and obtains evolutionary distances with a simultaneous estimation method that uses information from all pairwise alignments instead of just a single sequence pair. This leads to a robust phylogenetic reconstruction from highly diverged sequences with phylogenetic accuracy comparable to that of the maximum likelihood (ML) method using MSA (the ML+MSA approach), a conclusion supported by simulated nucleotide, amino acid and codon sequences. Only when sequences are not highly diverged (i.e., when a reliable MSA can be obtained) does the ML+MSA approach outperforms PhyPA. This result is not due to insufficient searching of tree space by ML+MSA. 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 due to the distorted phylogenetic signals in MSA.

I have implemented PhyPA in DAMBE together with two approaches making use of multi-gene data sets to derive phylogenetic support for subtrees equivalent to resampling techniques such as bootstrapping and jackknifing.

PhyPA itself is evolving. One improvement that should dramatically improve the pairwise alignment is to use different match/mismatch score matrices for pairwise aligning sequences of different divergence. A set of homologous sequences are not equally related to each other, some being closely related and some relatively more diverged. One can perform pairwise alignment, estimating divergence and use an appropriate match/mismatch score matrix for a re-alignment. I was initially uncertain whether this approach will generate any stunning outcome because different match/mismatch score matrices may introduce unexpected inconsistencies. However, preliminary results are highly encouraging. Phylogenetic researchers are likely to be dazzled by a new PhyPA late this year.

Xuhua Xia
Department of Biology, University of Ottawa, Canada

Publication

PhyPA: Phylogenetic method with pairwise sequence alignment outperforms likelihood methods in phylogenetics involving highly diverged sequences.
Xia X
Mol Phylogenet Evol. 2016 Sep