Ancient phylogenies: Alignments make a difference

Oliver Deusch[1], Giddy Landan[2], Mayo Röttger[1], Nicole Grünheit[1], Tal Dagan[1] and William Martin[1]

[1] Institute of Botany III, Heinrich Heine University Düsseldorf

[2] Department of Biology & Biochemistry, University of Houston

Abstract

Photosynthetic eukaryotes acquired many of their genes from the cyanobacterial ancestor of plastids through endosymbiotic gene transfer (EGT). Here we evaluate the total fraction of EGT genes by the reconstruction of 11,569 phylogenies for proteins encoded in the genomes of *Arabidopsis thaliana*, *Oryza sativa*, *Chlamydomonas reinhardtii* and *Cyanidioschyzon merolae* and their homologues from 224 prokaryotic and 13 eukaryotic reference genomes. On average, about 14% of the phylogenies examined indicate that the plant nuclear gene was acquired from cyanobacteria during the endosymbiotic origin of plastids. The reconstructed phylogenies span more than 1.2 billion years of evolution, and therefore are prone to reconstruction errors that are common to deep phylogenies. Yet in the absence of a known species tree and a known model of protein evolution, it is impossible to tell how many reconstruction artifacts the trees might contain. One source of such artifacts are errors in the multiple sequence alignments (MSAs) from which the phylogenies are reconstructed. In such cases, analysis of the MSAs' ambiguities might offer opportunities to help separate phylogenetic wheat from chaff. If columns in an alignment differ depending upon the order in which amino acids are introduced into the alignment, then the phylogenetic result is likely to be artifactual. Using this approach we find that the frequency of inferred EGT increases with increasing reliability of alignment columns. Although most current approaches to phylogenomics put their emphasis on the tree reconstruction procedures, the underlying alignments might be the more critical variable.

EGT inference depends upon sequence conservation

On average, about 14% of the proteins branched as nearest neighbor to cyanobacterial homologues. Proteins with higher sequence conservation showed a tendency to display a cyanobacterial origin more often than more divergent proteins did (Fig. 1). For example, 22% of the *Arabidopsis* gene trees with a mean branch length of ≤0.2 substitutions per site indicate cyanobacterial origin of the plant nuclear gene, while only 6% of the gene trees with mean branch length ≥1 do.

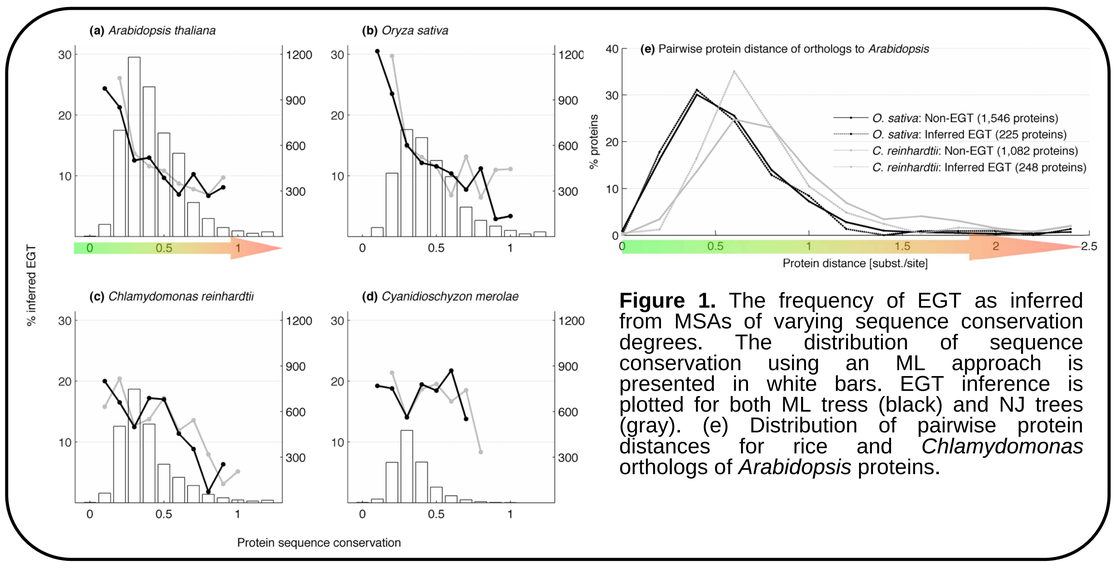


Figure 1. The frequency of EGT as inferred from MSAs of varying sequence conservation degrees. The distribution of sequence conservation using an ML approach is presented in white bars. EGT inference is plotted for both ML tress (black) and NJ trees (gray). (e) Distribution of pairwise protein distances for rice and *Chlamydomonas* orthologs of *Arabidopsis* proteins.

EGT positive proteins are not slow-evolving

No clear tendency was observed for EGTs to be preferentially common in the more slowly evolving proteins (Fig. 1e), suggesting that phylogenetic inference itself might bias the EGT estimate. It has log been known that phylogenetic inference is less accurate with more highly diverged proteins (Nei et al. 1995).

Divergent proteins produce unreliable alignments

We find a strong correlation between sequence conservation as measured by mean branch length in the trees and alignment reliability as estimated by the proportion of identical columns recovered in the heads and tails alignments using the HoT method (Fig. 2). For the most highly divergent sequences in the present data, about 80-90% of the columns generated by multiple alignment are dependent upon the order in which the amino acids are aligned.

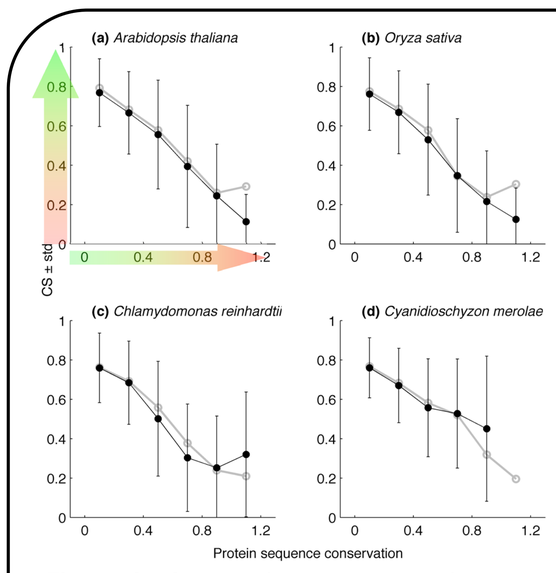


Figure 2. Conserved sequences produce more reliable alignments. The mean *CS* (dots) ± standard deviation is presented for ML trees (black) and NJ trees (gray).

Reliability measures

- Columns Score (**CS**, Thompson et al. 1999)
  - → Proportion of columns in the heads alignment that have a matching column in the tails alignment.
- Sum-of-Pairs Score (**SPS**, Thompson et al. 1999)
  - → Proportion of pairs in the heads alignment that have a matching pair in the tails alignment (less strict than CS).
- Phylogenetic Partitions Score (**PPS**)
  - → Proportion of tree splits (internal branches) that were reconstructed identically from the heads and tails alignments.

	A. thaliana		*O. sativa*		*C. reinhardtii*		*C. merolae*	
	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)
CS	0.60	0.63	0.61	0.66	0.61	0.65	0.42	0.45
SPS	0.71	0.75	0.77	0.79	0.7	0.75	0.48	0.53

Table 1. Correlation coefficient (using Pearson correlation) of PPS and both alignment reliability measures. Both CS and SPS are found in significant correlation (p<0.05) with the proportion of identical phylogenetic partitions (PPS).

Unreliable alignments produce unreliable trees

Both alignment reliability measures are found in good positive correlation with the tree reliability measure (Table 1). This is sensible because a phylogenetic tree can hardly be more reliable than its corresponding alignment as the tree is reconstructed from the alignment's homology patterns.

The correlation between sequence conservation and alignment reliability (Fig. 2) and its implications for the reliability of the inferred phylogenetic tree (Table 1) suggests that the inferred frequency of EGT depends primarily upon the alignment reliability, and that the tree results are simply a secondary symptom thereof.

EGT inference is dependent upon alignment / tree reliability

The fraction of EGT-positive trees is plotted across the distributions for bin intervals of heads-*vs*.-tails alignment quality as estimated by the criteria *CS*, *PPS*, and *SPS* for each archeaplastidan genome data set (Fig. 3). For each criterion, the most reliable (reproducible) alignments (or trees) give the highest estimate of EGT-positive trees, on the order of 17-25% of all trees examined, with a decreasing trend towards poorer alignments (or trees), notwithstanding variation among intervals containing small sample sizes.

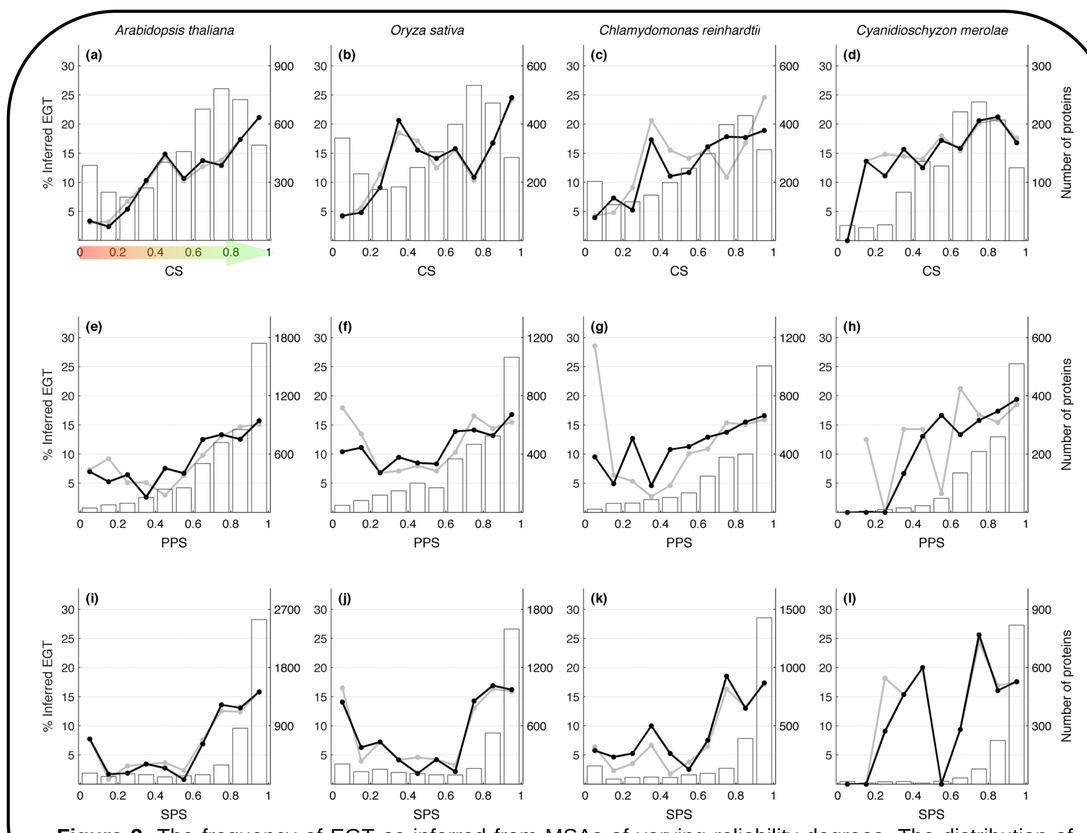


Figure 3. The frequency of EGT as inferred from MSAs of varying reliability degrees. The distribution of MSA reliability as estimated by the three measures is presented in white bars. The frequency of genes inferred as EGT is plotted above for ML trees (black) and NJ trees (gray).

Conclusion

On average, 14% of the proteins for which trees could be reconstructed were inferred as EGT, regardless of their alignment quality. This estimate should be regarded as a lower-bound. We found a consistent trend that proteins with better sequence conservation produce alignments of higher quality (reliability) and recover a higher proportion of EGT.

In deep phylogenies, however, highly conserved proteins are extremely rare. The HoT method determines when phylogenetic results are solely dependent upon a process no less arbitrary than a coin-toss and thus helps do separate phylogenetic wheat from chaff.

Analyzing alignment artifacts using the HoT method

(Landan and Graur 2007)

The head-tail method (HoT) tests the reliability of an alignment by comparing the original alignment (heads alignment) to the alignment of the same sequences in reversed order (tail alignment).

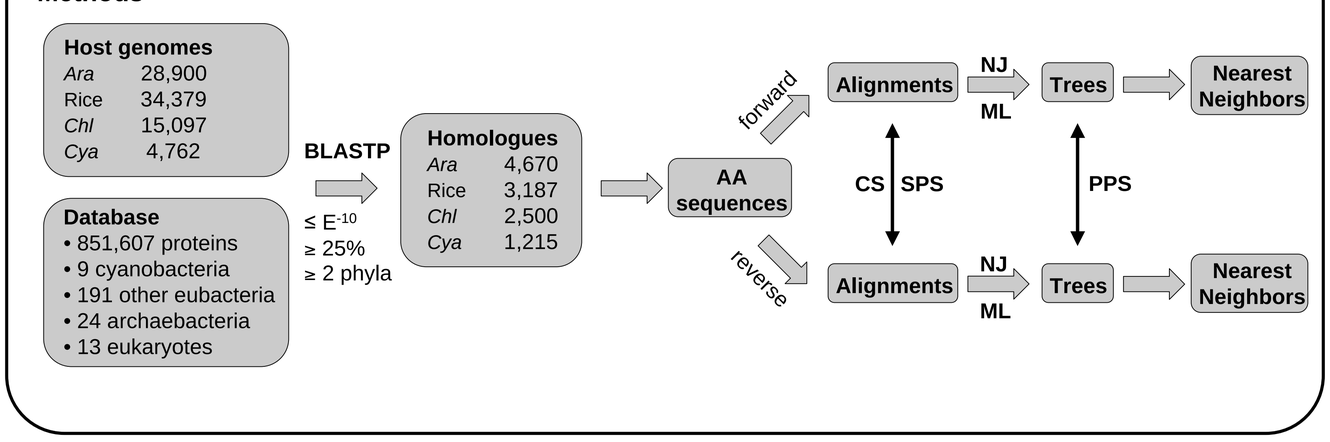
```
Ecoli        AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
Mjannaschii ----------------------ATTCCGGTTGATC-CTGCCGG-AGGCCACTGCTATCGGG
                                 ** * **** * *** *** *****      **

                                vs.

Ecoli        AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAG--GCCTAACACATGC
Mjannaschii --ATTCCGG---TTGATCCTGCCGGAGGCC-ACTGCTATCGGGGTCCGACTAAGCCATGC
              ***   *   ****** ** *  **    *  ***  ***     * ****  *****

                          No identical columns!
```

Methods



References

Landan G and Graur D. Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. Mol. Biol. Evol. 2007. Jun;24(6):1380-3.

Martin W et al.. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A. 2002 Sep 17;99(19):12246-51.

Nei M, Takezaki N and Sitnikova T. Assessing molecular phylogenies. Science. 1995. 267:253-254.

Thompson JD, Plewniak F and Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 1999. 27:2682-2690.