An Improved Method For Constructing *Large Phylogenies*

Nomi Hadar Anna Rice and Itay Mayrose

Department of Molecular Biology and Ecology of Plants,
Tel Aviv University, Tel Aviv Israel

Background: Large phylogenies

Inferring the evolutionary relationships among species is one of the ongoing goals of evolutionary biology.

The massive accumulation of sequence data should provide more accurate phylogenies with the hope to resolve the tree of life.

However, recent efforts to reconstruct large phylogenies encompassing thousands of species have revealed that the resulting trees are often of poor quality.

Do large phylogenies really provide accurate description of the true evolutionary history?

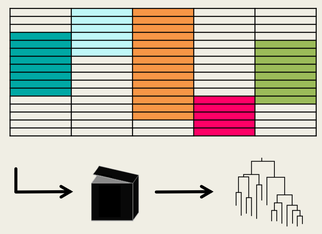
The reconstructed tree can only be as good as its underlying multiple sequence alignment (MSA).

How to build a phylogeny

Phylogeny reconstruction consists of two basic steps:

The supermatrix method

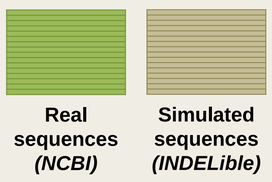
Concatenates sequences from multiple genes into a single, giant phylogenetic matrix. Gaps in place of missing genes.

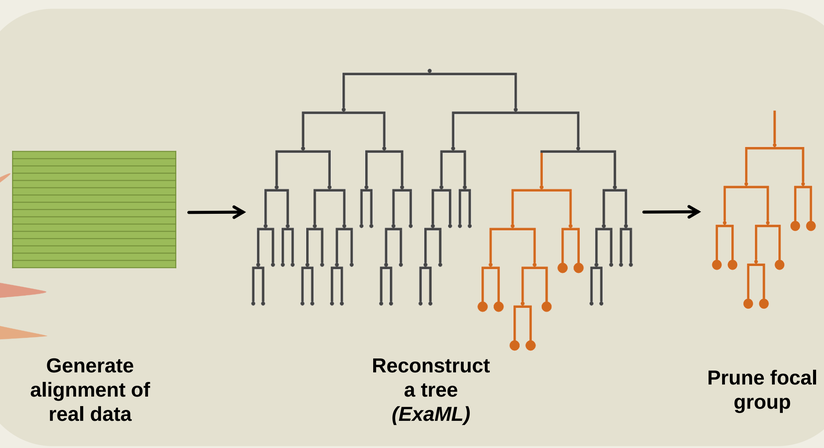


The problem: Aligning rapidly evolving loci for highly divergent taxa leads to poor alignments and inaccurate trees.

Examining the effect of sequence divergence

Is the accuracy of reconstructing a focal group affected when distant species are added to the alignment?



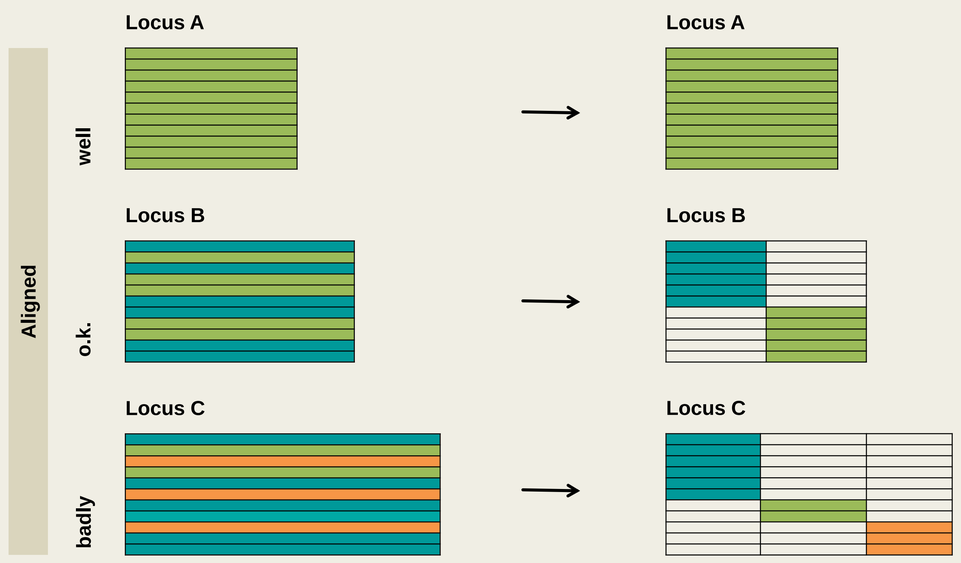


Results: Adding distant species affects the accuracy of the tree.

Average distance between trees		trees from simulated sequences			
	# species	135	816	2412	5113
trees from real sequences	135	0.3854	0.4362	0.4411	0.4436
	816		0.6384	0.6555	0.6667
	2412			0.8909	0.8999
	5113				1.1058

A new approach for reconstructing large phylogenies

Divide the alignment space to well-aligned regions:



Using the resulting alignment to reconstruct a tree with higher accuracy.

This study is supported by the Edmond J. Safra center for bioinformatics at Tel-Aviv University