

Materials and Methods

Simple Chained Guide Trees Give Poorer Multiple Sequence Alignments Than Inferred Trees In Simulation and Phylogenetic Benchmarks

Ge Tan, Manuel Gil, Ari P Löytynoja, Nick Goldman, Christophe Dessimoz

Correspondence: c.dessimoz@ucl.ac.uk

In this document, we provide a detailed description of the methods required to reproduce the analysis presented in Figure 1:

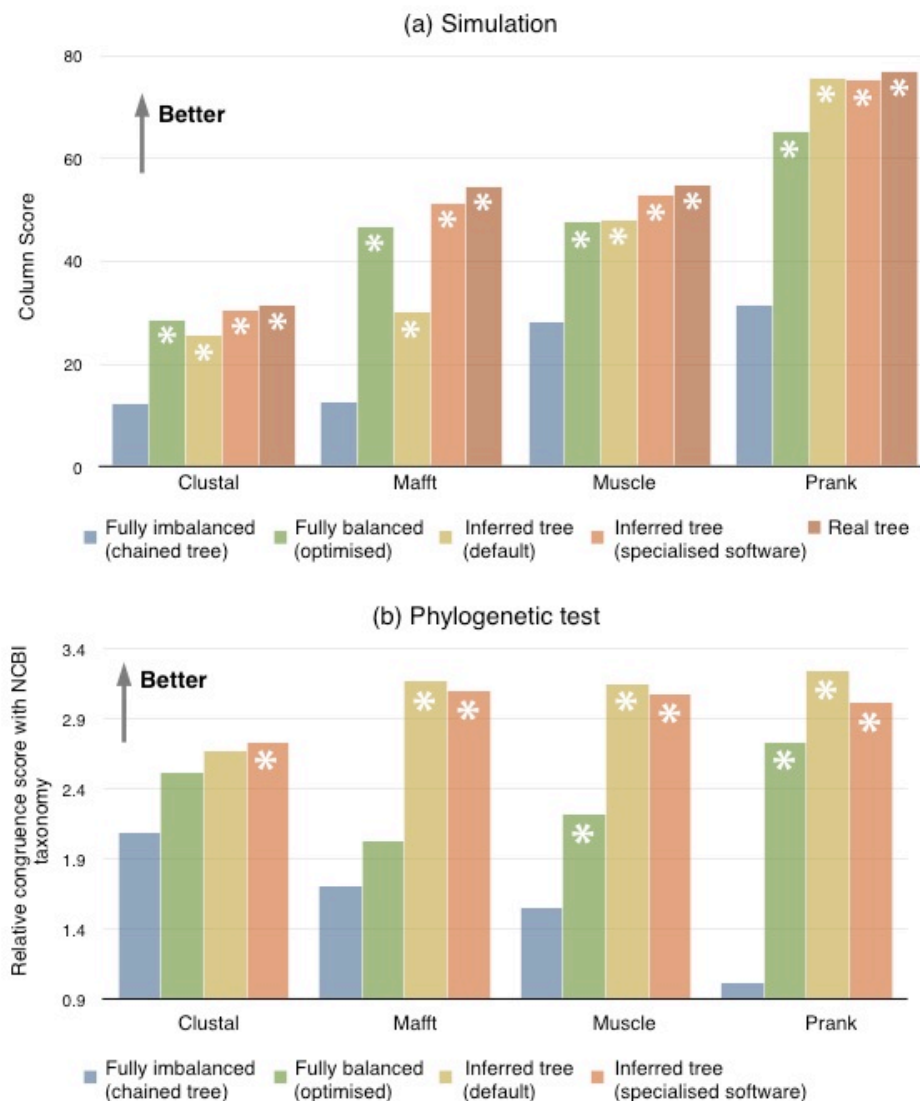


Fig. 1: Evaluation of alignments reconstructed with various aligners and guide tree methods. (a) Average true column score over 113 simulated datasets of 1024 sequences. (b) Average consistency with the NCBI taxonomy over 106 sets of 1024 biological sequences. Note that the real tree is unknown for empirical data. With fully imbalanced trees as input guide tree, Prank failed to reconstruct alignments in 38 empirical data problem instances; results reported in (b) are thus based on the remaining 68 alignments. Significant difference from fully imbalanced guide trees is indicated with a star (Wilcoxon double-sided test, $P < 0.001$). All data available at http://lab.dessimoz.org/14_guidetrees.

Simulation

We used ALF (3) to simulate 113 problem instances. For each instance, a 1024-taxa tree was sampled according to a birth-death process (with parameter $\lambda = 10\mu$) scaled such that the distance from root to deepest branch was 100 PAM units. Sequences were evolved along these trees according to WAG substitution matrices (6), with insertion and deletions introduced following a Poisson distribution with mean=0.0005 event/PAM/site, and length distribution following a Zipfian distribution with exponent 1.821, truncated at 50 characters (default parameters).

Data for phylogenetic test

A total of 3,038 sets of six fungal orthologous sequences were sampled from the OMA database (Sep 2008 release; 7). For each set, additional homologs were automatically collected from the Mar 2014 OMA release via NCBI BLAST using the script Mafft-Homologs (8), with a threshold E-value of 10^{-10} . Sets for which Mafft-Homologs returned fewer than 1,018 matches were discarded. For each remaining set, exactly 1,018 matches were randomly selected, and the corresponding full sequences were retrieved from SwissProt, ending up with 1,129 sets of 6+1,018=1,024 homologous sequences. Our analysis was performed on a random selection of 106 such gene families.

Construction of the guide trees

Balanced trees with leaf assignment optimised using the Travelling Salesman Problem heuristic were computed by first computing a circular tour over the sequences using the ComputeTSP() function in the programming environment “Darwin” (9) and breaking the tour at its longest edge. Guide trees inferred with specialised phylogenetic software were computed based on pairwise alignments with PAM distance estimation using the *Align()* function in Darwin followed by the least-squares tree optimisation using the function *MinSquareTree()* in Darwin.

Aligners

The aligners used were Clustal Omega v.1.2.1 with command line options “--max-guidetree-iterations=0” (10), Mafft v.7.58, with command line options “--anysymbol --retree 2 --maxiterate 0 --unweight” (11), Muscle v.3.8.31 with command line option “-maxiters 2” (12), and Prank v.140603 with command line options “-once -nobppa -uselogs” (13).

Description of taxonomy congruence score

The taxonomy congruence score adapts the tests introduced in Dessimoz and Gil (5) to the large gene trees at hand. If we assume that genes evolve along a species tree with occasional gene duplication and loss events, the resulting gene trees can be expected in many parts to be congruent with the species tree. For instance, in the absence of gene loss, the species represented in the left and right subtrees of duplication splits should be identical. Likewise, the species contained in the left and right subtrees of speciation nodes should be related clades of the species tree. Our measure attempts to capture this by traversing each gene tree bottom-up and by counting the NCBI lineage terms in common in the subtrees of each internal node. Formally, the definition of the consistency score is as follows. Let T be a rooted gene tree, where the leaves are labeled with the species to which the corresponding gene belongs. $T.Left$ ($T.Right$) denotes the left (right) subtree of T . For a leaf l , $L(l)$ lists the NCBI taxonomic lineage of the corresponding species¹. We define the consistency $C(T)$ of a tree T with L recursively as

¹ For example, if l is *Homo sapiens* then $\{L(l)\} = \{\text{cellular organisms, Eukaryota, Opisthokonta, Metazoa, Eumetazoa, Bilateria, Deuterostomia, Chordata, Craniata, Vertebrata, Gnathostomata, Teleostomi, Euteleostomi, Sarcopterygii, Dipnotetrapodomorpha, Tetrapoda, Amniota, Mammalia, Theria, Eutheria, Boreoeutheria, Euarchontoglires, Primates, Haplorrhini, Simiiformes, Catarrhini, Hominoidea, Hominidae, Homininae, Homo}\}$ and $|L(l)| = 30$.

$$C(x) = \begin{cases} 0, & \text{if } x \text{ is a leaf} \\ |s(x)| + |s(x.Left)| + |s(x.Right)|, & \text{if } x \text{ is an internal node} \end{cases}$$

where

$$s(x) = \begin{cases} \{L(x)\}, & \text{if } x \text{ is a leaf} \\ s(x.Left) \cap s(x.Right), & \text{if } x \text{ is an internal node} \end{cases}$$

The guide tree methods produce trees that are unrooted, or not necessarily rooted at the biological root. To compute the consistency score, the trees were (re-)rooted with the midpoint rooting method (14). For each gene family, the congruence scores obtained by the different guide tree methods were converted to fractional ranks, such that more congruent methods obtain higher ranks. Specifically, the scores were sorted in ascending order and the rank of each method was determined; ties were assigned the same rank, defined as the mean of their ordinal ranking.

References:

6. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699.
7. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res* 39:D289–94.
8. Katoh K, Kuma K-I, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518.
9. Gonnet GH, Hallett MT, Korostensky C, Bernardin L (2000) Darwin v. 2.0: an interpreted computer language for the biosciences Cited by me. *Bioinformatics*.
10. Sievers F et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.
11. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780.
12. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
13. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* 102:10557–10562.
14. Felsenstein J (2004) *Inferring phylogenies* (Sinauer Associates, Sunderland, MA).