A coarse-graining, ultrametric approach to resolve the phylogeny of prokaryotic strains with frequent homologous recombination

CURRENT STATUS: UNDER REVIEW

BMC Evolutionary Biology BMC Series

Tin Yau Pang

pang@hhu.de *Corresponding Author*

DOI:

10.21203/rs.2.18054/v1

SUBJECT AREAS

Evolutionary Biology *Evolutionary Developmental Biology*

KEYWORDS

phylogenetics, homologous recombination, ultrametric tree

Abstract

Background: A frequent event in the evolution of prokaryotic genomes is homologous recombination, where a foreign DNA stretch replaces a genomic region similar in sequence. Recombination can affect the relative position of two genomes in a phylogenetic reconstruction in two different ways: (i) one genome can recombine with a DNA stretch that is similar to the other genome, thereby reducing their pairwise sequence divergence; (ii) one genome can recombine with a DNA stretch from an outgroup genome, increasing the pairwise divergence. While several recombination-aware phylogenetic algorithms exist, many of these cannot account for both types of recombination; some algorithms can, but do so inefficiently. Moreover, many of them reconstruct the ancestral recombination graph (ARG) to help infer the genome tree, and require that a substantial portion of each genome has not been affected by recombination, a sometimes unrealistic assumption.

Results: Here, we propose a coarse-graining approach for phylogenetic reconstruction (CGP), which is recombination-aware but forgoes ARG reconstruction, applicable even if all genomic regions have experienced substantial amounts of recombination, and can be used on both nucleotide and amino acid sequences. CGP considers the local density of substitutions along pairwise genome alignments, fitting a model to the empirical distribution of substitution density to infer the pairwise coalescent time. Given all pairwise coalescent times, CGP reconstructs an ultrametric tree representing vertical inheritance. Based on simulations, we show that the proposed approach can reconstruct ultrametric trees with accurate topology, branch lengths, and root positioning. Applied to a set of *E. coli* strains, the reconstructed trees are most consistent with gene distributions when inferred from amino acid sequences, a data type that cannot be utilized by many alternative approaches.

Conclusions The CGP algorithm is more accurate than alternative recombination-aware methods for ultrametric phylogenetic reconstructions.

Full-text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed.

However, the manuscript can be downloaded and accessed as a PDF.

Figures

Figure 1

Illustration of the procedure of the proposed CGP algorithm. (a) The algorithm takes n aligned sequences as input, which can be nucleotide or amino acid sequences; substitutions on the sequences are represented by coloured markers. (b) Each of the $n(n-1)/2$ genome pairs is divided into equal-sized segments, and the pairwise substitutions on each segment is enumerated to obtain the distribution of local single site polymorphisms (SSPs) density (denoted as $g(x)$). (c) The algorithm aims to infer the distance matrix of the genome sequence pairs from the $n(n-1)/2$ SSP distributions. (d) In particular, the algorithm fits the empirical SSP distributions with a model; the input of this model involves a matrix of $n(n-1)/2$ coalescent times and other model parameters (mutation rate μ , recombination rate ρ , average population divergence θ and transfer efficiency δ_{TE}). (e) In the fitting process, the $n(n-1)/2$ coalescent times are constrained (matrix cells with the same colour have the same value), such that the matrix can be bijectively mapped to a UPGMA tree. (f) the

algorithm 468 explores the model parameter space and tree space to obtain the best fit ultrametric tree.

Figure 2

Histograms showing the distributions of the unrooted symmetric distance (SD) between true trees and trees reconstructed by CGP, RaxML, and Gubbins, from genome sequences derived from Fisher-Wright simulations with low, intermediate, and high recombination levels.

Figure 3

Boxplots showing the distributions of the unrooted branch score distance (BSD) and the distributions of the z-score of unrooted BSD between the true trees and trees reconstructed by CGP, RaxML, and Gubbins, from genome sequences from Fisher-Wright simulations with low, intermediate, and high recombination levels.

Figure 4

Boxplots showing the distributions of the rooted branch score distance (BSD) and the distributions of the z-score of rooted BSD between the true trees and trees reconstructed by CGP, RaxML, and Gubbins, from genome sequences from Fisher-Wright simulations with 480 low, intermediate, and high recombination levels.

Figure 5

Boxplots showing the distributions of the GLOOME posterior likelihood (GPL) and the distributions of the z-scores of GPL. The trees were reconstructed from observed *E. coli* genome sequences, applying CGP, RAXML, ClonalFrameML, and Gubbins to nucleotide sequences, and CGP and RAXML to amino acid sequences.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

[Supplementary_File_S3.tar.xz](#)

[supplementary_text.pdf](#)

[Supplementary_File_S2_results_real_genomes.xlsx](#)

[Supplementary_File_S1_results_simulated_genomes.xlsx](#)