HAL open science

On the transformation of MinHash-based uncorrected distances into proper evolutionary distances for phylogenetic inference

Alexis Criscuolo

▶ To cite this version:

Alexis Criscuolo. On the transformation of MinHash-based uncorrected distances into proper evolutionary distances for phylogenetic inference. F1000Research, 2020, 9, pp.1309. 10.12688/f1000research.26930.1 . pasteur-03264539

HAL Id: pasteur-03264539

https://pasteur.hal.science/pasteur-03264539v1

Submitted on 18 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Check for updates

RESEARCH ARTICLE

On the transformation of MinHash-based uncorrected distances into proper evolutionary distances for phylogenetic inference [version 1; peer review: 3 approved]

Alexis Criscuolo

Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, Institut Pasteur, USR 3756, CNRS, 75015 Paris, France

v1 **First published:** 10 Nov 2020, **9**:1309
https://doi.org/10.12688/f1000research.26930.1
Latest published: 10 Nov 2020, **9**:1309
https://doi.org/10.12688/f1000research.26930.1

Abstract

Recently developed MinHash-based techniques were proven successful in quickly estimating the level of similarity between large nucleotide sequences. This article discusses their usage and limitations in practice to approximating uncorrected distances between genomes, and transforming these pairwise dissimilarities into proper evolutionary distances. It is notably shown that complex distance measures can be easily approximated using simple transformation formulae based on few parameters. MinHash-based techniques can therefore be very useful for implementing fast yet accurate alignment-free phylogenetic reconstruction procedures from large sets of genomes. This last point of view is assessed with a simulation study using a dedicated bioinformatics tool.

Keywords

MinHash, p-distance, evolutionary distance, substitution model, phylogenetics, genome, simulation

Open Peer Review

Reviewer Status ✔ ✔ ✔

	Invited Reviewers		
	1	**2**	**3**
version 1 10 Nov 2020	✔ report	✔ report	✔ report

1. **Brian Ondov**, National Human Genome Research Institute, USA, Bethesda, USA
2. **Burkhard Morgenstern**, University of Göttingen, Göttingen, Germany
3. **Guy Perrière**, Université Claude Bernard Lyon 1, Villeurbanne, France

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Alexis Criscuolo (alexis.criscuolo@pasteur.fr)

Author roles: Criscuolo A: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2020 Criscuolo A. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Criscuolo A. **On the transformation of MinHash-based uncorrected distances into proper evolutionary distances for phylogenetic inference [version 1; peer review: 3 approved]** F1000Research 2020, **9**:1309 https://doi.org/10.12688/f1000research.26930.1

First published: 10 Nov 2020, **9**:1309 https://doi.org/10.12688/f1000research.26930.1

Introduction

To estimate the level of proximity between two non-aligned genome sequences x and y, recent methods (e.g. 1–7) have focused on decomposing the two genomes into their respective sets K_x and K_y of non-duplicated nucleotide k-mers (i.e. oligonucleotides of size k). A pairwise similarity may then be easily estimated based on the Jaccard index $j = |K_x \cap K_y| / |K_x \cup K_y|$[8]. The Jaccard index between two sets of k-mers is a useful measure for two main reasons. First, it can be quickly approximated using MinHash-based techniques (MH[9]), as implemented in e.g. Mash[2], sourmash[3], Dashing[4], Kmer-db[6], FastANI[5], or BinDash[7]. Such techniques select a small subset (of size σ) of hashed and sorted k-mers (called sketch) from each K_x and K_y, and approximate j by comparing these two subsets (for more details, see 2,9–12). Second, the proportion p of observed differences between the two aligned genomes (often called uncorrected distance or p-distance) can be approximated from j (therefore without alignment) with the following formula (e.g. 13,14):

$$p = 1 - \left(\frac{2j}{j+1}\right)^{1/k}, \quad (1)$$

provided that both sizes σ and k are large enough, and j is not too low (see below).

As a consequence, a pairwise evolutionary distance d can be derived from the Jaccard index j using transformation formulae of the following form:

$$d = -b_1 \log_e(1 - p / b_2), \quad (2)$$

where p is obtained using Equation (1). Parameters b_1 and b_2 can be defined according to explicit models to estimate the number d of nucleotide substitutions per character that have occurred during the evolution of the sequences x and y, e.g. 15–24. When $b_1 = b_2 = 1$, Equation (2) corresponds to the Poisson correction (PC; e.g. 21) distance. Although it is based on a simplistic model of nucleotide substitution[1,16,25,26], PC is the p-distance transformation implemented in many MH tools (e.g. Mash, Dashing, FastANI, Kmer-db, BinDash). However, more accurate distance estimates may be obtained by using substitution models based on more parameters. Among these models, equal-input (EI, sometimes called F81[18,19,24,27–29]) takes into account the equilibrium frequency π_r of each residue r in $\Sigma = \{A, C, G, T\}$. An EI distance can be estimated using Equation (2) with $b_1 = 1 - \sum_{r \in \Sigma} \pi_r^2$ and $b_2 = 1 - \sum_{r \in \Sigma} \pi_{r_x} \pi_{r_y}$, where π_{r_x} and π_{r_y} are the frequencies of r in the two sequences x and y, respectively[20]. Further assuming that the heterogeneous replacement rates among nucleotide pairs and sites can be modelled with a Γ distribution, an EI distance d can be derived from p using the following formula:

$$d = ab_1[(1 - p / b_2)^{-1/a} - 1], \quad (3)$$

where $a > 0$ is the (unknown) shape parameter of the Γ distribution, e.g. 22,24,30–33. It is worth noticing that when a is high, Equation (2) and Equation (3) yield very similar distance estimates (for any fixed b_1 and b_2).

The aim of this study is to assess the accuracy of Equation (2) and Equation (3) in transforming a MH p-distance $\hat{p}$, where $\hat{p}$ is derived from the MH Jaccard index $\hat{j}$ using Equation (1). In the following, analyses of large sets of simulated nucleotide sequences show three complementary results. First, current MH implementations enable p-distances to be conveniently estimated under several conditions. Second, PC and EI transformations (2) and (3) of MH p-distance estimates can suitably approximate evolutionary distances derived from general time reversible (GTR; e.g. 34) models of nucleotide substitution. Third, PC and EI distances derived from MH estimates enable accurate phylogenetic tree reconstruction from unaligned nucleotide sequences.

Results and discussion

MinHash-based p-distance approximation

Varying d from 0.05 to 1.00 (step = 0.05), a total of 200 nucleotide sequence pairs with d substitution events per character were simulated under the models GTR and GTR+Γ. Each model was adjusted with three different equilibrium frequencies: equal frequencies (f_1; $\pi_A = \pi_C = \pi_G = \pi_T = 25\%$), GC-rich ($f_2$; $\pi_A = 10\%$, $\pi_C = 30\%$, $\pi_G = 40\%$, $\pi_T = 20\%$), and AT-rich (f_3; $\pi_A = \pi_T = 40\%$, $\pi_C = \pi_G = 10\%$). The GTR substitution rates and the Γ shape parameters were obtained based on a maximum likelihood (ML) analysis of 142 real-case phylogenomics datasets. Overall, ML estimates of Γ shape parameters were quite low (i.e. varying from 0.162 to 0.422, with an average of 0.314), confirming that the heterogeneity of the substitution rates across sites is a non-negligible factor when studying evolutionary processes. Every simulation was completed with indel events, resulting in sequences > 3 Mbs with relative lengths (i.e. longer/shorter) varying from 1.0196 ($d = 0.05$) to 1.1117 ($d = 1.00$), on average.

For each of the 2 (GTR, GTR+Γ) × 3 (f_1, f_2, f_3) × 20 (d = 0.05, 0.10, ..., 1.00) × 200 = 24,000 simulated sequence pairs x and y, the corresponding p-distance was estimated using three MH tools: Mash, BinDash and Dashing. Of note, the accuracy of a MH estimate $\hat{j}$ of the Jaccard index between K_x and K_y is mainly dependent on two parameters: the k-mer size k and the sketch size σ. The size k should be large enough to minimize the probability q of observing a random k-mer shared by x and y by chance alone. Such a value can be obtained from q by $k = \lceil \log_{|\Sigma|}(g(1-q)/q) - 0.5 \rceil$, where g is the length of the largest sequence[2,35,29]. The size σ should be large enough to minimize the error bounds of $\hat{j}$[2], but also to avoid the inconvenient estimate $\hat{j} = 0$. Following 29, σ was set by the proportion s of the average sequence length.

To investigate the impact of both parameters σ and k on the accuracy of the MH estimates, each MH tool was used with s = 0.2, 0.4, 0.6, 0.8 and $q = 10^{-3}, 10^{-6}, 10^{-9}, 10^{-12}$. As in simulated sequences, g ranges from 4.99 Mbs (d = 0.05) to 3.38 Mbs (d = 1.00) on average, s translates into moderately to very large sketch sizes σ, and q into k-mer sizes k = 16, 21, 26, 31.

Two statistics were calculated to assess the linear relationship between the MH estimate $\hat{p}$ (derived from $\hat{j} \neq 0$) and the

'true' p-distance p: the coefficient of determination R^2 and the slope β of the linear least-square regression $\hat{p} = \beta p$. Let $\Psi(p_{max})$ be the subset of pairs $(p, \hat{p})$ such that $p \leq p_{max}$. Varying p_{max} from 0.10 to 0.55, R^2 and β were estimated from $\Psi(p_{max})$ (Figure 1). The cumulative proportions $f_{\hat{j}=0}$ of MH Jaccard index $\hat{j} = 0$ within $[0, p_{max}]$ were also measured (Figure 1). Finally, every value $p_{r>0.99}$ was estimated, where $p_{r>0.99}$ is defined as the highest p-distance such that the subset $\Psi(p_{r>0.99})$ provides a coefficient of correlation $r > 0.99$ (as assessed by a Fisher transformation z-test with p-value < 1%; Figure 1). The highest values $p_{r>0.99}$ were obtained with parameters $k = 26$ ($q \leq 10^{-9}$) and $s = 0.8$ (illustrated in Figure 2).

One important result (Figure 1) is that current MH implementations return suitable estimates of p as long as $p \leq 0.25$, provided that k is sufficiently large. Indeed, when $k \geq 21$ (and any $s \geq 0.2$), the statistics $p_{r>0.99}$ are higher than 0.25 (Figure 1), therefore showing that p and $\hat{p}$ are highly linearly correlated when $p \leq 0.25$ (see e.g. Figure 2). Interestingly, when $p \leq 0.25$, the worthless estimate $\hat{j} = 0$ was almost never observed with the different selected parameters s and q (Figure 1).

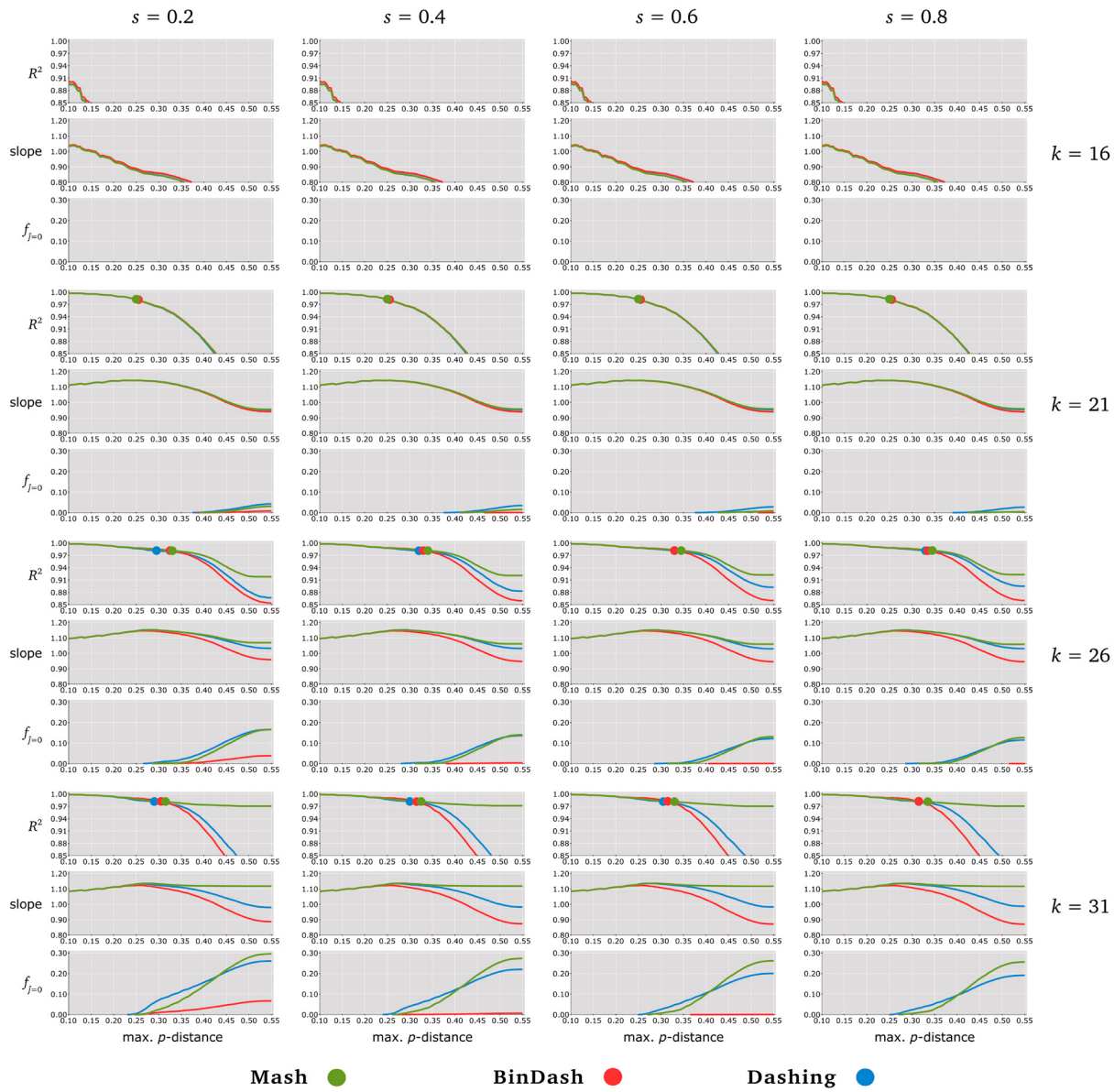

Figure 1. Accuracy of MH tools for estimating *p*-distances from unaligned nucleotide sequences. For each sketch size (columns; set by s = 0.2, 0.4, 0.6, 0.8) and each k-mer size (rows; k = 16, 21, 26, 31), three line charts represent different statistics determined with Mash (green), BinDash (red), and Dashing (blue). For p_{max} ranging from 0.10 to 0.55 (x-axes), represented statistics are (i) the coefficient of determination R^2 (up; y-axis ranging from 0.85 to 1.00) and (ii) the slope of the linear least-square regression through the origin (middle; y-axis ranging from 0.8 to 1.2) computed from estimated $\hat{p}$ and corresponding 'true' p-distances $p \leq p_{max}$, as well as (iii) the cumulative proportion $f_{\hat{j}=0}$ of estimated Jaccard index $\hat{j} = 0$ within $[0, p_{max}]$ (bottom; y-axis ranging from 0.0 to 0.3). Circles in R^2 line charts (up) indicate the largest value $p_{r>0.99}$ such that the subset of pairs $(p, \hat{p})$ defined by $p \leq p_{r>0.99}$ provides a coefficient of correlation $r > 0.99$.

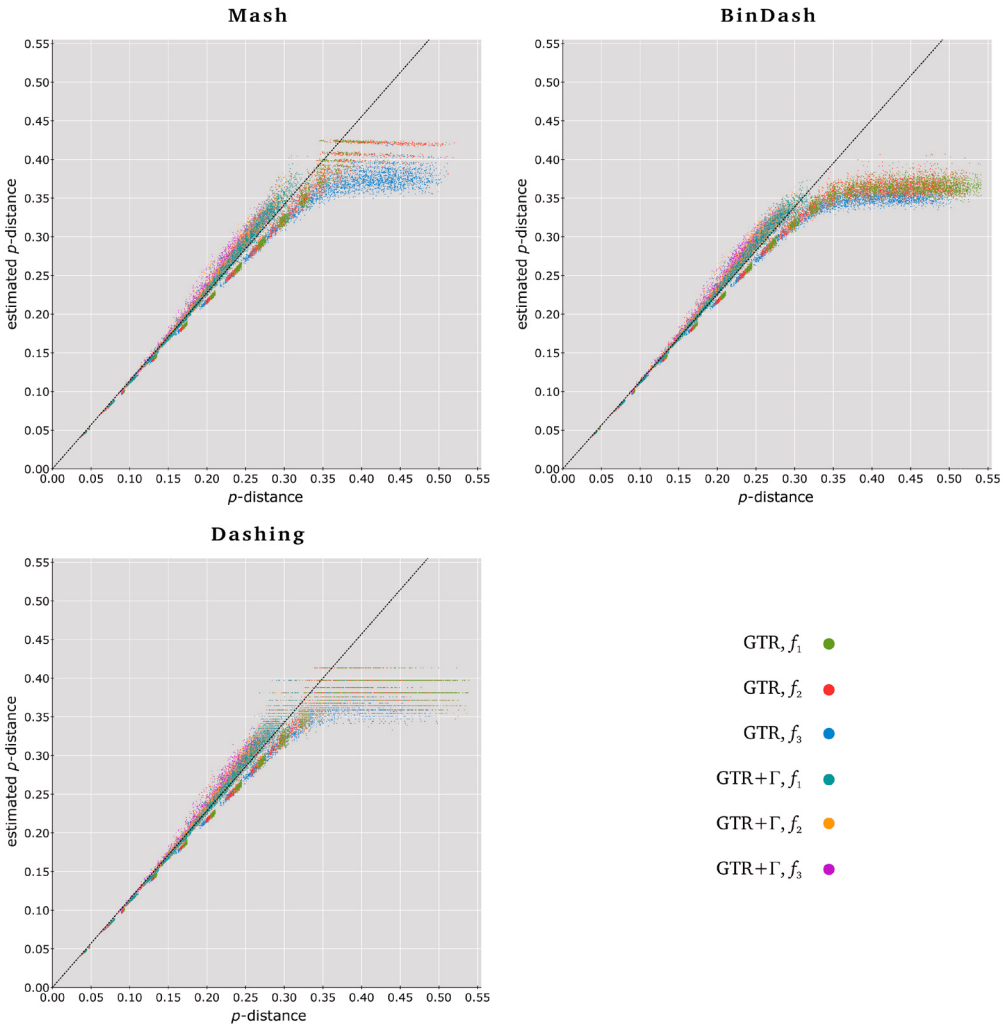

Figure 2. MH *p*-distance estimates between 24,000 pairs of unaligned nucleotide sequences. The *p*-distances $\hat{p}$ estimated by Mash (up left), BinDash (up right) and Dashing (bottom left) with $k = 26$ ($q = 10^{-9}$) and $s = 0.8$ are plotted against the 'true' *p*-distances p between 24,000 pairs of nucleotide sequences simulated under six scenaria of evolution: GTR with equilibrium frequencies $f_1 = (0.25, 0.25, 0.25, 0.25)$ (green points), $f_2 = (0.10, 0.30, 0.40, 0.20)$ (red) and $f_3 = (0.40, 0.10, 0.10, 0.40)$ (blue), and GTR+Γ with f_1 (cyan), f_2 (orange) and f_3 (magenta). Points corresponding to $\hat{j} = 0$ are not represented. Each scatter plot is completed with the least-square regression line through the origin (dashed black line) estimated from the subset of points (p, $\hat{p}$) such that $p \leq p_{r>0.99}$, where $p_{r>0.99} = 0.345$ (Mash), 0.335 (BinDash) and 0.330 (Dashing).

Furthermore, when $p > 0.25$, large *k*-mers are required to obtain satisfactory estimates, i.e. $k > 21$ or $q < 10^{-6}$ (Figure 1). However, dealing with $k > 21$ involves using large sketch sizes to minimize the cases $\hat{j} = 0$ (see $f_{\hat{j}=0}$ in Figure 1). Simulation results suggest that $k = 26$ (i.e. $q = 10^{-9}$) and $s > 0.4$ yield suitable estimates of p, obtained from sequences of lengths > 4 Mbs with pairwise $p < 0.35$ (see Figure 1 and Figure 2). Indeed, when p ranges between 0.25 and 0.35, small sizes k (e.g. $k \leq 21$ or $q \geq 10^{-6}$) always provide underestimated $\hat{p}$ (with any s). Large size k (e.g. $k = 31$ or $q = 10^{-12}$) results in the same trend, but also in high numbers of useless estimates $\hat{j} = 0$ (even with large σ; see $f_{\hat{j}=0}$ in Figure 1).

When $p \geq 0.35$, MH tools always underestimate the *p*-distances between the sequences simulated for this study (Figure 1 and Figure 2). One could suggest that more accurate MH estimates $\hat{p}$ will be expected with larger sketch sizes σ. Nevertheless, results represented in Figure 2 (i.e. $q = 10^{-9}$ and $s = 0.8$, providing the highest $p_{r>0.99}$) are based on average σ varying from ~2.7×10^6 ($d = 1.00$) to ~3.6×10^6 ($d = 0.35$), which are larger than some real genomes.

Transformation of *p*-distances into evolutionary distances

When a pairwise *p*-distance p can be estimated from unaligned nucleotide sequences, it may be transformed into an evolutionary distance d, based on Equation (2) or Equation (3). The relationship between p and d was represented in Figure 3 for different distance estimators: PC transformation (2) ($b_1 = b_2 = 1$), and EI transformations (2) and (3) with equilibrium frequencies f_1 ($b_1 = b_2 = 0.75$ under homogeneous substitution

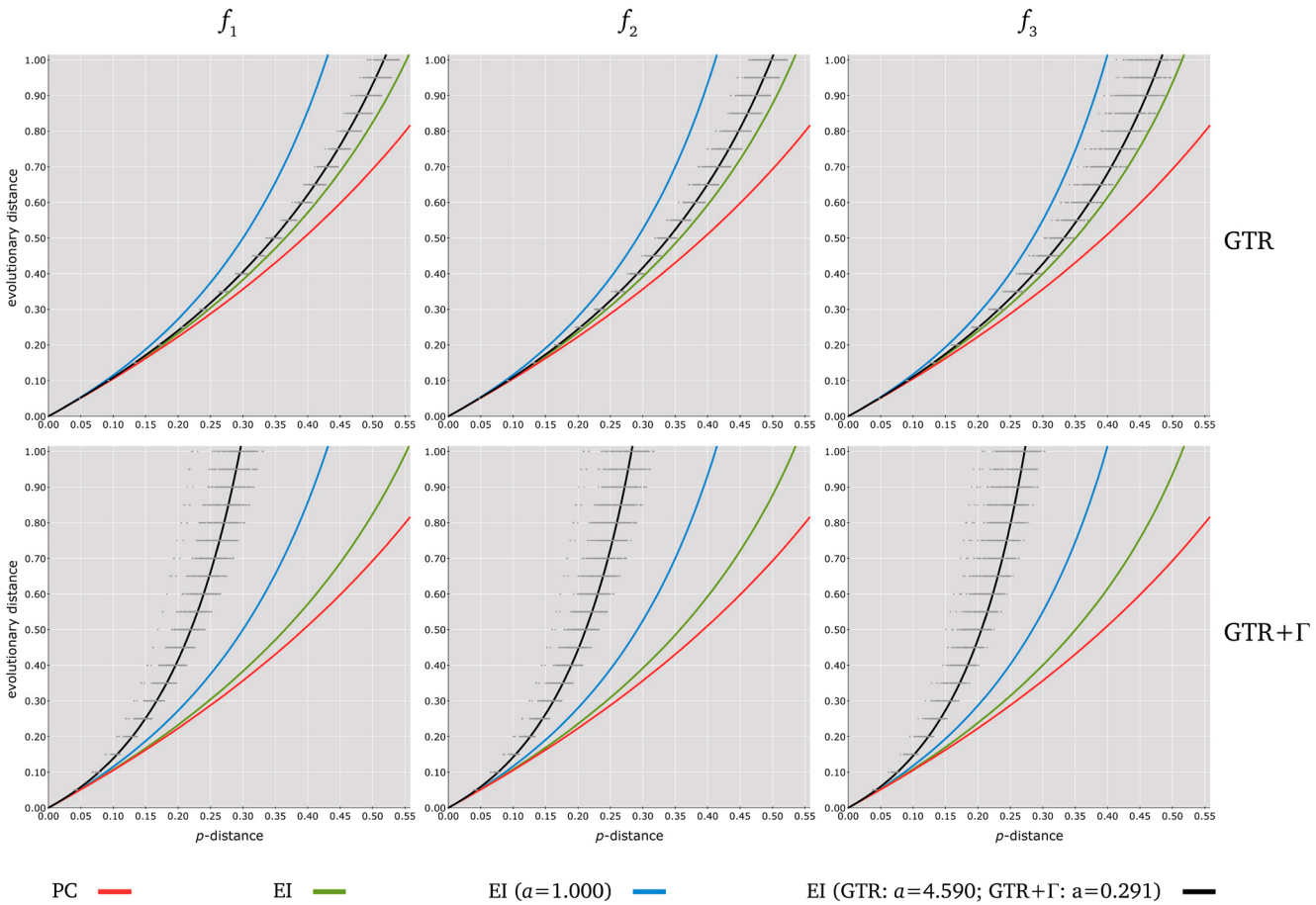

Figure 3. Relationships between *p*-distances and different evolutionary distance estimates under various models of nucleotide substitution. The six charts represent the evolutionary distance d (y-axis ranging from 0.00 to 1.00) against the p-distance p (x-axis ranging from 0 to 0.55). Gray points (p, d) are derived from the simulation of sets of 4,000 sequence pairs, each under six different scenaria of evolution: GTR (top) and GTR+Γ (bottom) with equilibrium frequencies f_1 = (0.25, 0.25, 0.25, 0.25) (left), f_2 = (0.10, 0.30, 0.40, 0.20) (middle) and f_3 = (0.40, 0.10, 0.10, 0.40) (right). PC and EI versions of Equation (2) are represented with red and green curves, respectively. EI version of Equation (3) is represented with blue curves for a = 1, and with black curves for the values a = 4.590 and a = 0.291 determined by least-square regression from the gray points derived from the models GTR (top) and GTR+Γ (bottom), respectively.

pattern[20]), f_2 ($b_1 = b_2 = 0.70$) and f_3 ($b_1 = b_2 = 0.66$). Parameter a in EI Equation (3) was estimated by least-square regression from the pairs (p, d) derived from the sequences simulated under the models GTR and GTR+Γ (see above).

PC and EI p-distance transformations (2) result in improper underestimates as the expected distance d increases. Indeed, when compared with realistic GTR-based distances d, PC and EI transformations (2) give distance estimates that are always lower than d, especially under GTR+Γ and when d is large (e.g. $d > 0.1$; Figure 3). This downward bias is somewhat expected, knowing that PC and EI transformations (2) are based on less parameters than both models GTR and GTR+Γ. However, the additional parameter a in Equation (3) may help dealing with heterogeneous substitution rates among residue pairs (e.g. 36). Hence, the relationship between GTR distances d and the corresponding p-distances p can be approximated by the EI transformation (3) with a = 4.590 (Figure 3). Moreover, as d returned by Equation (3) is inversely proportional to a (for any fixed p), the relationship between d and p under the model GTR+Γ (with Γ shape parameter of 0.314, on average) can also be approximated by the EI transformation (3) with a = 0.291 (Figure 3).

These results show that complex distance measures can be approximated by simple analytical formulae based on few parameters. In practice, nucleotide frequencies (four parameters) can be trivially computed and p-distances (a fifth parameter) can be estimated using MH tools (see above). Therefore, the evolutionary distance d between two sequences that have evolved under the parameter-rich model GTR+Γ can be approximated from these only five parameters using (3) with $a \leq 4.590$ (Figure 3).

At this point, it should be stressed that MH $\hat{p}$ tends to be overestimated. Indeed, MH estimates are of the form $\hat{p} \approx \beta p$ with slope β varying from 1.08 (BinDash, k = 31, s = 0.2) to 1.15 (Dashing, k = 26, s = 0.2) when $p \leq p_{r>0.99}$ (Figure 1). This has a direct impact on the derived distances: using PC and EI transformations (2) on $\hat{p} = \beta p$ with β = 1.15 and $p \leq 0.35$

provide distance estimates that are quite comparable to the ones returned by Equation (3) with a ranging from 1.000 to 4.590 (see Figure 4 for the equilibrium frequencies f_2; similar results were observed with f_1 and f_3 – not shown). The PC transformation (2) on the upward biased MH $\hat{p}$ returns distances that are then comparable to some complex distance measures (e.g. derived from a GTR model), therefore justifying its use by many MH tools. Nevertheless, the EI transformation (3) remains necessary when dealing with distantly related sequences (e.g. $p > 0.2$) and strong heterogeneity of the substitution rate across sites (e.g. often observed Γ shape parameter < 1.000). In such cases, the value of the parameter a should always be slightly increased to compensate the MH upward bias. For instance, EI transformation (3) on p with $a = 0.291$ (i.e. GTR+Γ distance least-square fitting in Figure 3) can be approximated by the same equation on $\hat{p} = \beta p$ with $\beta = 1.15$ and $a = 0.431$.

Phylogenetic reconstruction from MinHash-based evolutionary distances

To assess whether MH p-distance transformations may translate into reliable phylogenetic trees, additional simulations were performed. A total of 142 sets of sequences was simulated under the model GTR+Γ along reference phylogenetic trees. Representative GTR+Γ model parameters (same as above) and reference phylogenetic trees were obtained based on a ML analysis of real-case phylogenomics datasets. Sizes of the reference trees ranged from 10 to 154 taxa (31 on average), with diameters (i.e. maximum distance between any two leaves of a tree) varying from 0.204 to 2.883 (0.975 on average). Sequence lengths and indel events were simulated in the same way as the previous sequence pair simulations.

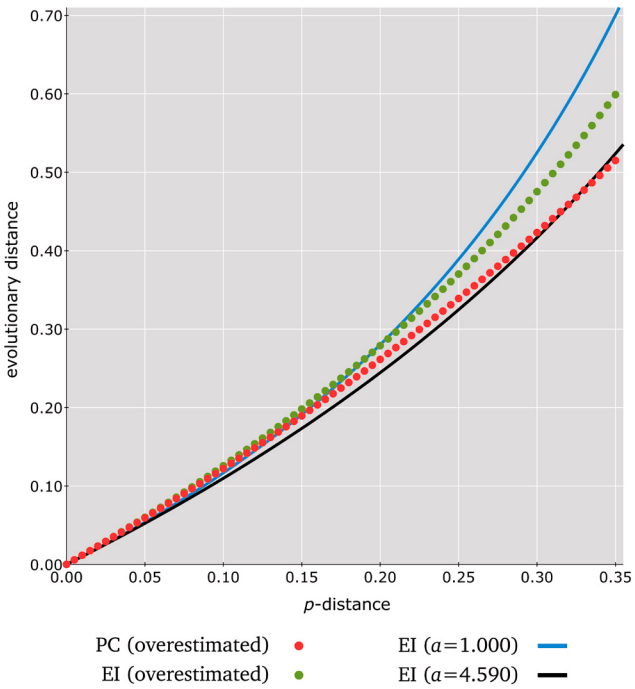

Figure 4. Impact of the MH *p*-distance upward bias on PC and EI transformations. The relationship between the p-distance p (x-axis ranging from 0.00 to 0.35) and the corresponding evolutionary distance d (y-axis ranging from 0.00 to 0.70) is represented when using PC (red dots) and EI (with equilibrium frequencies f_2; green dots) transformations (2) on $\hat{p} = \beta p$ with $\beta = 1.15$. For ease of comparison with Figure 3, EI (f_2) transformation (3) on p are represented with $a = 1.000$ (blue curve) and $a = 4.590$ (black curve).

The script JolyTree v2.0 was used to reconstruct phylogenetic trees from the simulated sequences. For each pair of unaligned sequences, this script estimates the MH p-distance using Mash, and transforms it into an evolutionary distance. Using these MH-based distances, JolyTree next reconstructs a minimum evolution phylogenetic tree with confidence supports at branches, based on a ratchet-based hill-climbing procedure (for more details, see 29). To obtain accurate MH p-distance estimates, JolyTree was run with parameters $q = 10^{-9}$ and $s = 0.5$ (see above). Evolutionary distances were estimated using the PC and EI transformations (2), as well as the EI transformation (3). To observe the impact of the parameter a, the EI transformation (3) was computed with a varying from 0.05 to 10.0. The accuracy of each p-distance transformation for phylogenetic inference was assessed by the percentage of recovered reference trees, i.e. identical topologies (Figure 5).

Using JolyTree with EI transformations improves the percentage of recovered reference trees (Figure 5). In spite of their limitations, PC distances result in the recovery of 75.3% of the 142 reference trees, but EI transformation (2) increases this percentage to 76.7% (Figure 5). Furthermore, the EI transformation (3) generally provides better results in a large range of a, i.e. up to 83.1% of recovered reference trees (Figure 5). Low a-values (e.g. $a \leq 0.3$) translate into many incorrect tree topologies, whereas high ones (e.g. $a > 6$) tend to provide the same reference tree recovering percentage as the EI transformation (2) (Figure 5). Most suitable values of a (corresponding to the highest reference tree recovering percentages, e.g. 80%) seem to range in the interval [1.0, 2.0] (Figure 5).

These simulation results are consistent with two views which are somehow contradictory. On the one hand, accurate (parameter-rich) distance estimates are required, because biased ones (i.e. corresponding to a concave or convex function of the actual evolutionary distances) may result in incorrect phylogenetic trees[23,37]. On the other hand, simple (underparameterized) distance estimates should often be preferred, because they frequently result in more accurate tree topologies[21,38–42]. Here, the simple PC and EI transformations (2) (one and five parameters, respectively) enable many reference trees to be recovered (Figure 5). However, the EI transformation (3) is able to approximate realistic distance measures (e.g. GTR+Γ) by using only one supplementary parameter a (Figure 3). It therefore enables more reference trees to be recovered (Figure 5).

In line with 43, most suitable values of a (e.g. between 1.0 and 2.0) are all higher than the Γ shape parameter values used for simulating the sequence datasets (i.e. varying from 0.162

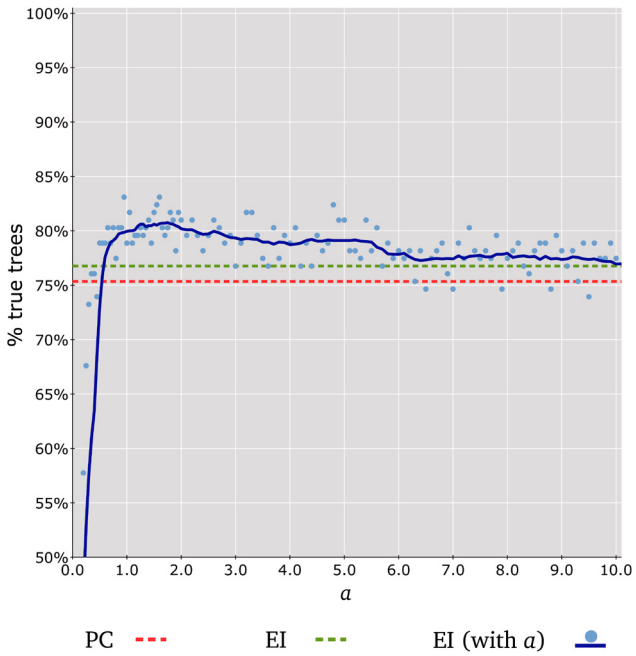

Figure 5. Accuracy of different *p*-distance transformations for phylogenetic inference. The percentage of recovered reference trees (*y*-axis ranging from 50% to 100%) is represented (light blue dots) in function of the parameter a (*x*-axis ranging from 0.0 to 10.0) in EI formula (3). The overall trend of these dots is illustrated using a moving average (dark blue curve). Dashed lines represent the percentages of recovered reference trees obtained with the PC (red) and EI (green) transformations (2).

to 0.422, with an average of 0.314). This can be explained by the MH upward bias (see above), but also by the large variance of the estimate (3) when a becomes low. Reminding that the Γ shape parameters (and the reference trees) used in these simulations were inferred from real-case datasets, these results suggest that using EI transformation (3) with $a \approx 1.5$ may be suitable to infer genus phylogenetic trees. In light of this, it should be stressed that the article[29] describing the first distributed version of JolyTree (v1.1) incorrectly stated that the Mash output is the MH estimate $\hat{p}$ (instead of its PC transformation). As JolyTree v1.1 uses the EI transformation (2), this misinterpretation translates into the odd transformation formula $\delta = -b_1 \log_e (1 + \log_e(1 - \hat{p}) / b_2)$. However, as δ can be approximated on $\hat{p} \leq 0.35$ by the EI transformation (3) with $a = 1.208$, this explains the overall accuracy of JolyTree v1.1 despite its use of δ[29].

Conclusions

Alignment-free phylogenetic inference from pairwise MH-based distance estimates is a promising approach. It enables phylogenetic trees to be quickly reconstructed from a large number of genomes without the burden of multiple sequence alignments (see e.g. 2,29,44–52). This report confirms this view by showing that proper evolutionary distances can be easily derived from MH *p*-distance estimates, therefore enabling accurate phylogenetic inferences.

First, although implemented to approximate nearest neighbors in sequence sets, current MH tools (e.g. Mash, BinDash, Dashing) were shown to be able to conveniently estimate pairwise *p*-distances p up to $p \approx 0.35$. In practice, as p is very similar to the one-complement of the Average Nucleotide Identity (ANI; e.g. 29,53,54), MH estimates of p can then be obtained between genomes gathered from many bacteria, archaea or eukaryota genera, i.e. with pairwise ANI > 65%.

Second, the EI *p*-distance transformation (3) was proven efficient to approximate complex distance measures, e.g. derived from GTR model with heterogeneous substitution rates across sites. Because of an upward bias observed in MH *p*-distance estimates, simpler transformations (based on few parameters, as the commonly used PC) still provide distance measures that are comparable to GTR ones, but with (unrealistic) homogeneous substitution rates across sites. However, thanks to its supplementary parameter a, EI transformation (3) remains necessary to approximate distance measures between distantly related sequences that have arisen from more realistic substitution events.

Third, as proper evolutionary distances can be derived from MH *p*-distance estimates, their efficiency in phylogenetic inference was established using the dedicated tool JolyTree[29]. In particular, the EI transformation (3) with $a \approx 1.5$ enables accurate phylogenetic trees to be inferred.

Methods

Model parameter estimation

To simulate the evolution of nucleotide sequences according to realistic substitution processes, the 187 genus datasets compiled in 29 (available at https://doi.org/10.3897/rio.5.e36178.suppl2) were first considered to infer a representative range of GTR parameter values. For each of the 187 genera, the associated genome assemblies were processed using Gklust v0.1 to obtain one representative genome assembly for each putative species. This analysis provided 142 sets of representative genome assemblies after discarding genera containing < 10 putative species. For each of these 142 sets, coding sequences were clustered using Roary v3.12[55]. Each cluster with at least four coding sequences was used to build a multiple amino acid sequence alignment using MAFFT v7.407[56]. Multiple sequence alignments were back-translated at the codon level and concatenated, leading to 142 supermatrices of nucleotide characters. A phylogenetic tree was inferred from each supermatrix of characters using IQ-TREE v1.6.7.2[57] with evolutionary model GTR+Γ. All data related to these analyses are publicly available as *Extended data* at https://doi.org/10.5281/zenodo.4034244[58].

Sequence simulation

To assess the accuracy of different pairwise distance estimates, a simulation of sequence pairs was performed under both models GTR and GTR+Γ with three different sets (π_A, π_C, π_G, π_T) of equilibrium frequencies: f_1 = (0.25, 0.25, 0.25, 0.25), f_2 = (0.10, 0.30, 0.40, 0.20), and f_3 = (0.40, 0.10, 0.10, 0.40). For each of these six scenaria and for each d varying from 0.05 to 1.00 (step = 0.05), the program INDELible v1.03[59]

was used to simulate the evolution of 200 sequence pairs with d substitution events per character. Initial sequence length was 5 Mbs, and an indel rate of 0.01 was set with indel length drawn from [1, 50000] according to a Zipf distribution with parameter 1.5. For each simulated sequence pair, model parameters (i.e. GTR: six relative rates of nucleotide substitution; GTR+Γ: six rates and one Γ shape parameter) were randomly drawn from the 142 sets of estimated ones (see above). All simulated sequences are publicly available as *Extended data* at https://doi.org/10.5281/zenodo.4034461[60].

To compare the efficiency of p-distance transformations for phylogenetic reconstruction, the program INDELible v1.03 was also used to simulate the evolution of a sequence along each of the 142 phylogenetic trees previously inferred from different genera (see above). For each of the 142 genera, sequence evolution was simulated under the model GTR+Γ with the corresponding parameters (i.e. four nucleotide frequencies, six relative rates, and one Γ shape parameter). Sequence length and indel events were simulated as described above. The 142 simulated sequence sets are publicly available as *Extended data* at https://doi.org/10.5281/zenodo.4034643[61].

Sequence and phylogenetic analyses

MH p-distances were estimated with Mash v2.2, BinDash v1.0, and Dashing v0.3.4-11-gb44a. BinDash and Dashing were used with the MH b-bit flavor with $b = 18$. Of note, as Mash and Bindash directly return the PC distance d, the corresponding p-distance was computed by $p = 1 - e^{-d}$.

Phylogenetic tree reconstructions from simulated sequences were performed with the script JolyTree v2.0. This version implements the PC and EI transformations (2) and (3) of the pairwise p-distances estimated by Mash. If any, missing evolutionary distances $d_{uv} = \varnothing$ (i.e. corresponding to $\hat{j} = 0$ or $p \geq b_2$) between sequences u and v are approximated by JolyTree from the other non-missing evolutionary distances by $d_{uv} = \min_{x \neq u,v; d_{xu}, d_{xv} \neq \varnothing} (d_{xu} + d_{xv})$. This fast approximation is derived from the triangle inequality property $d_{uv} \leq d_{xu} + d_{xv}$ expected from the triplet of evolutionary distances induced by any sequence triplet u, v, x (see e.g. 62).

Data availability

Source data

A list of the 14,244 genome assemblies used to build the 187 genus datasets (Supplementary material of 29). https://doi.org/10.3897/rio.5.e36178.suppl2.

Extended data

Zenodo: Phylogenomic analyses of 142 prokaryotic genera. https://doi.org/10.5281/zenodo.4034244[58].

Zenodo: Simulated pairs of nucleotide sequences for testing (alignment-free) genome distance estimate methods. https://doi.org/10.5281/zenodo.4034461[60].

Zenodo: Model trees and associated simulated nucleotide sequences for testing phylogenetic inference methods. https://doi.org/10.5281/zenodo.4034643[61].

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Acknowledgements

The author thanks Pascal Campagne for his meaningful comments on the manuscript. The author is also obliged to Sylvain Brisse and to the *Hub de Bioinformatique et Biostatistique*, Institut Pasteur, Paris (France), for their support. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris.

References

1. Fan H, Ives AR, Surget-Groba Y, *et al.*: **An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data.** *BMC Genomics.* 2015; **16**(1): 522.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Ondov BD, Treangen TJ, Melsted P, *et al.*: **Mash: fast genome and metagenome distance estimation using MinHash.** *Genome Biol.* 2016; **17**(1): 132.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Titus Brown C, Irber L: **sourmash: a library for MinHash sketching of DNA.** *Journal of Open Source Software.* 2016; **1**(5): 27.
Reference Source
4. Baker D, Langmead B: **Dashing: Fast and accurate genomic distances with HyperLogLog.** *bioRxiv.* 2019.
Publisher Full Text
5. Jain C, Rodriguez LM, Phillippy AM, *et al.*: **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.** *Nat Commun.* 2018; **9**(1): 5114.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Deorowicz S, Gudys A, Dlugosz M, *et al.*: **Kmer-db: instant evolutionary distance estimation.** *Bioinformatics.* 2019; **35**(1): 133–136.
PubMed Abstract | Publisher Full Text
7. Zhao X: **BinDash, software for fast genome distance estimation on a typical personal laptop.** *Bioinformatics.* 2019; **35**(4): 671–673.
PubMed Abstract | Publisher Full Text
8. Jaccard P: **Nouvelles recherches sur la distribution florale.** *Bulletin de la Société vaudoise des sciences naturelles.* 1908; **44**: 223–270.
Publisher Full Text
9. Broder A: **On the resemblance and containment of documents.** In SEQUENCES '97: *Proceedings of the Compression and Complexity of Sequences.* 1997; 21–29.
Publisher Full Text
10. Jain C, Dilthey A, Koren S, *et al.*: **A fast approximate algorithm for mapping long reads to large reference databases.** *J Comput Biol.* 2018; **25**(7): 766–779.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Numanagic I, Gökkaya AS, Zhang L, *et al.*: **Fast characterization of segmental duplications in genome assemblies.** *Bioinformatics.* 2018; **34**(17): i706–i714.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Rowe WPM: **When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data.** *Genome Biol.* 2019; **20**(1): 199.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Mougel C, Thioulouse J, Perrière G, *et al.*: **A mathematical method for**

determining genome divergence and species delineation using AFLP. *Int J Syst Evol Microbiol.* 2002; **52**(Pt 2): 573–586.
PubMed Abstract | Publisher Full Text

14. Sarmashghi S, Bohmann K, Gilbert MTP, *et al.*: **Skmer: assembly-free and alignment-free sample identification using genome skims.** *Genome Biol.* 2019; **20**(1): 34.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Jukes TH, Cantor CR: **Evolution of protein molecules.** In: Munro HN, editor, *Mammalian protein metabolism.* 1969; 21–132.
Publisher Full Text

16. Dickerson RE: **The structure of cytochrome c and the rates of molecular evolution.** *Journal of Molecular Evolution.* 1971; **1**(1): 26–45.
PubMed Abstract | Publisher Full Text

17. Kimura M, Ohta T: **On the stochastic model for estimation of mutational distance between homologous proteins.** *J Mol Evol.* 1972; **2**(1): 87–90.
PubMed Abstract | Publisher Full Text

18. Tajima F, Nei M: **Biases of the estimates of DNA divergence obtained by the restriction enzyme technique.** *J Mol Evol.* 1982; **18**(2): 115–120.
PubMed Abstract | Publisher Full Text

19. Tajima F, Nei M: **Estimation of evolutionary distance between nucleotide sequences.** *Mol Biol Evol.* 1984; **1**(3): 269–285.
PubMed Abstract | Publisher Full Text

20. Tamura K, Kumar S: **Evolutionary distance estimation under heterogeneous substitution pattern among lineages.** *Mol Biol Evol.* 2002; **19**(10): 1727–1736.
PubMed Abstract | Publisher Full Text

21. Nei M, Zhang J: **Evolutionary distance: Estimation.** In *Encyclopaedia of Life Science.* American Cancer Society, 2006.
Publisher Full Text

22. Yang Z: **Models of nucleotide substitution.** In *Computational Molecular Evolution.* 2006; 3–38.

23. McTavish EJ, Steel M, Holder M: **Twisted trees and inconsistency of tree estimation when gaps are treated as missing data – The impact of model misspecification in distance corrections.** *Mol Phylogenet Evol.* 2015; **93**: 289–295.
PubMed Abstract | Publisher Full Text

24. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets.** *Mol Biol Evol.* 2016; **33**(7): 1870–1874.
PubMed Abstract | Publisher Full Text

25. Zuckerkandl E, Pauling L: **Evolutionary divergence and convergence in proteins.** In: Bryson V and Vogel HJ, editors, *Evolving Genes and Proteins.* 1965; 97–166.
Publisher Full Text

26. Jukes TH: **Comparison of polypeptide sequences.** In Le Cam LM, Neyman J and Scott EL, editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Darwinian, Neo-Darwinian, and non-Darwinian Evolution.* 1972; **5**: 101–127.

27. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol.* 1981; **17**(6): 368–376.
PubMed Abstract | Publisher Full Text

28. McGuire G, Prentice MJ, Wright F: **Improved error bounds for genetic distances from DNA sequences.** *Biometrics.* 1999; **55**(4): 1064–1070.
PubMed Abstract | Publisher Full Text

29. Criscuolo A: **A fast alignment-free bioinformatics procedure to infer accurate distance-based phylogenetic trees from genome assemblies.** *Res Ideas Outcomes.* 2019; **5**: e36178.
Publisher Full Text

30. Golding GB: **Estimates of DNA and protein sequence divergence: an examination of some assumptions.** *Mol Biol Evol.* 1983; **1**(1): 125–142.
PubMed Abstract | Publisher Full Text

31. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol.* 1986; **3**(5): 418–426.
PubMed Abstract | Publisher Full Text

32. Rzhetsky A, Nei M: **Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites.** *J Mol Evol.* 1994; **38**(3): 295–299.
PubMed Abstract | Publisher Full Text

33. Gu X: **The age of the common ancestor of eukaryotes and prokaryotes: statistical inferences.** *Mol Biol Evol.* 1997; **14**(8): 861–866.
PubMed Abstract | Publisher Full Text

34. Yang Z: **Estimating the pattern of nucleotide substitution.** *J Mol Evol.* 1994; **39**(1): 105–111.
PubMed Abstract | Publisher Full Text

35. Fofanov Y, Luo Y, Katili C, *et al.*: **How independent are the appearances of n-mers in different genomes?** *Bioinformatics.* 2004; **20**(15): 2421–2428.
PubMed Abstract | Publisher Full Text

36. Bigot T, Guglielmini J, Criscuolo A: **Simulation data for the estimation of numerical constants for approximating pairwise evolutionary distances between amino acid sequences.** *Data in Brief.* 2019; **25**: 104212.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Susko E, Inagaki Y, Roger AJ: **On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled.** *Mol Biol Evol.* 2004; **21**(9): 1629–1642.
PubMed Abstract | Publisher Full Text

38. Zharkikh A, Li WH: **Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock.** *Syst Biol.* 1993; **42**(2): 113–125.
Publisher Full Text

39. Russo CA, Takezaki N, Nei M: **Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny.** *Mol Biol Evol.* 1996; **13**(3): 525–536.
PubMed Abstract | Publisher Full Text

40. Takahashi K, Nei M: **Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used.** *Mol Biol Evol.* 2000; **17**(8): 1251–1258.
PubMed Abstract | Publisher Full Text

41. Rosenberg MS, Kumar S: **Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well.** *Mol Biol Evol.* 2001; **18**(9): 1823–1827.
PubMed Abstract | Publisher Full Text

42. Yoshida R, Nei M: **Efficiencies of the NJp, Maximum Likelihood, and Bayesian Methods of Phylogenetic Construction for Compositional and Noncompositional Genes.** *Mol Biol Evol.* 2016; **33**(6): 1618–1624.
PubMed Abstract | Publisher Full Text

43. Guindon S, Gascuel O: **Efficient biased estimation of evolutionary distances when substitution rates vary across sites.** *Mol Biol Evol.* 2002; **19**(4): 534–543.
PubMed Abstract | Publisher Full Text

44. Dazas M, Badell E, Carmi-Leroy A, *et al.*: **Taxonomic status of *Corynebacterium diphtheriae* biovar Belfanti and proposal of *Corynebacterium belfantii* sp. nov.** *Int J Syst Evol Microbiol.* 2018; **68**(12): 3826–3831.
PubMed Abstract | Publisher Full Text

45. Garcia-Hermoso D, Criscuolo A, Lee SC, *et al.*: **Outbreak of Invasive Wound Mucormycosis in a Burn Unit Due to Multiple Strains of *Mucor circinelloides* f. *circinelloides* Resolved by Whole-Genome Sequencing.** *mBio.* 2018; **9**(2): e00573–18.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Lees J, Kendall M, Parkhill J, *et al.*: **Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study [version 2; peer review: 3 approved].** *Wellcome Open Res.* 2018; **3**: 33.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Petit RA, Hogan JM, Ezewudo MN, *et al.*: **Fine-scale differentiation between *Bacillus anthracis* and *Bacillus cereus* group signatures in metagenome shotgun data.** *PeerJ.* 2018; **6**: e5515.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Bénard AHM, Guenou E, Fookes M, *et al.*: **Whole genome sequence of *Vibrio cholerae* directly from dried spotted filter paper.** *PLoS Neglected Tropical Diseases.* 2019; **13**(5): e0007330.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Halpin JL, Dykes JK, Katz L, *et al.*: **Molecular Characterization of *Clostridium botulinum* Harboring the *bont*/B7 Gene.** *Foodborne Pathog Dis.* 2019; **16**(6): 428–433.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Nadimpalli M, Vuthy Y, Lauzanne A, *et al.*: **Meat and Fish as Sources of Extended-Spectrum β-Lactamase-Producing *Escherichia coli*, Cambodia.** *Emerg Infect Dis.* 2019; **25**(1): 126–131.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Watts SC, Holt KE: **hicap: *In Silico* Serotyping of the *Haemophilus influenzae* Capsule Locus.** *J Clin Microbiol.* 2019; **57**(6): e00190–19.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Zielezinski A, Girgis HZ, Bernard G, *et al.*: **Benchmarking of alignment-free sequence comparison methods.** *Genome Biol.* 2019; **20**(1): 144.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Goris J, Konstantinidis KT, Klappenbach JA, *et al.*: **DNA-DNA hybridization values and their relationship to whole-genome sequence similarities.** *Int J Syst Evol Microbiol.* 2007; **57**(Pt 1): 81–91.
PubMed Abstract | Publisher Full Text

54. Colston S, Fullmer M, Beka L, *et al.*: **Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case.** *mBio.* 2014; **5**(6): e02136.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Page AJ, Cummins CA, Hunt M, *et al.*: **Roary: rapid large-scale prokaryote pan genome analysis.** *Bioinformatics.* 2015; **31**(22): 3691–3693.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol.* 2013; **30**(4): 772–780.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Nguyen LT, Schmidt HA, von Haeseler A, *et al.*: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.** *Mol Biol Evol.* 2015; **32**(1): 268–274.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Criscuolo A: **Phylogenomic analyses of 142 prokaryotic genera**. 2020.
http://www.doi.org/10.5281/zenodo.4034261

59. Fletcher W, Yang Z: **INDELible: a flexible simulator of biological sequence evolution.** *Mol Biol Evol.* 2009; **26**(8): 1879–1888.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Criscuolo A: **Simulated pairs of nucleotide sequences for testing (alignment-free) genome distance estimate methods**. 2020.
http://www.doi.org/10.5281/zenodo.4034462

61. Criscuolo A: **Model trees and associated simulated nucleotide sequences for testing phylogenetic inference methods**. 2020.
http://www.doi.org/10.5281/zenodo.4034644

62. Guénoche A, Grandcolas S: **Approximations par arbre d'une distance partielle.** *Mathématiques et Sciences humaines.* 1999; **146**: 51–64.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 1

Reviewer Report 03 December 2020

https://doi.org/10.5256/f1000research.29746.r74631

© **2020 Perrière G.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Guy Perrière**
Laboratoire de Biométrie et Biologie Evolutive, CNRS, UMR5558, Université Claude Bernard Lyon 1, Villeurbanne, France

I have no special criticism on the methodology and my comments on that part are really minor. First, what is the justification to the choice of a value equal to 1.5 for the shape of the Zipf distribution used to simulate the indels. As a user of INDELible myself I know very well how this value can be of importance so I would like to know how this value was chosen (arbitrary choice?). Second, I am not sure that the right reference for the computation of missing distances from the triangle inequality property is Guénoche and Grancolas (1999). The first paper I have seen on that topic is Lapointe and Kirsch (1995[1]). Also, I have no concerns on the results presented and everything seems ok for me.

The conclusion of the manuscript presents the approach of phylogenetic inference from pairwise MH-based distance estimates as a possible alternative to classical methods of phylogenetic reconstruction (especially in bacteria). Indeed, a classical phylogenomic approach involves the chaining of a complex set of procedures: the identification of orthologous genes in multiple species; the alignment of these genes; and finally, the building of a tree from the concatenation of those alignments. I think that the inclusion in the paper of a comparison of two phylogenies obtained with: i) the MH-based distance and ii) a concatenation would strengthen this claim. A lot of concatenation datasets extracted from complete bacterial genomes are available and it would be not too difficult to do the experiment. I think this addition is of importance because I have no idea on the influence horizontal gene transfers can have on MH-based reconstructions when classical approaches try to minimize this influence.

References

1. Lapointe FJ, Kirsch JAW: Estimating Phylogenies from Lacunose Distance Matrices, with Special Reference to DNA Hybridization Data. *Molecular Biology and Evolution*. 1995; **12** (2): 266-284
Publisher Full Text | Reference Source

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Phylogenetics, comparative genomics and bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 01 December 2020

https://doi.org/10.5256/f1000research.29746.r74629

© 2020 Morgenstern B. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Burkhard Morgenstern
Department of Bioinformatics, Göttingen Center for Molecular Biosciences, Institute for Microbiology and Genetics, University of Göttingen, Göttingen, Germany

The author investigates how phylogenetic distances between genomic sequences, calculated by MinHash methods, can be transformed into more meaningful distances, based on complex models of nucleotide substitutions.

In the last few years, a number of alignment-free methods have been developed to calculate distances between genomic sequences. While earlier alignment-free methods used rough measures of sequence (dis)similarity, some methods have been proposed recently that accurately estimate distances - the number of substitutions per position since two sequences evolved from their last common ancestor - based on stochastic models of evolution (see below). These methods are restricted, however, to the simplest possible model, the Jukes-Cantor model.

In recent papers, MinHash techniques were proposed as an attractive way of estimating the p-distance between DNA sequences, i.e. the number of mismatches per position in an (unknown) alignment of the compared sequences.

The present paper shows how distances based on complex models of evolution (GTR, GTR + \Gamma) can be approximated using MinHash-based p-distances. The approach is carefully evaluated based on simulated sequences.

The proposed method is a welcome and useful addition to existing alignment-free methods, extending them to more realistic models of evolution. The approach is novel, the paper is very well written and clear, and suitable references are given to the literature for more details. Therefore, I support indexing of the manuscript.

A certain limitation is that the manuscript is restricted to MinHash distances. According to the author, these methods are accurate for p-distances roughly < 0.25 (for reasonable k-mer length), but are less accurate for larger distances. However, a number of other alignment-free methods have been proposed that accurately estimate distances for much larger distances, e.g. Kr (Haubold *et al.*, 2009[1]), FSWM (Leimeister *et al.*, 2017[2]), Phylonium (Klötzl and Haubold, 2019[3]), Slope-SpaM (Röhling *et al.*, 2020[4]). All these methods estimate distances based on the simple Jukes-Cantor model; the program Co-phylog estimates non-corrected p-distances (Yi and Jin, 2013[5]). It should be straight-forward to transform these distances to more complex models, as done in the present paper, and to compare them to the MinHash methods evaluated in the paper.

References

1. Haubold B, Pfaffelhuber P, Domazet-Loso M, Wiehe T: Estimating mutation distances from unaligned genomes.*J Comput Biol*. 2009; **16** (10): 1487-500 PubMed Abstract | Publisher Full Text
2. Leimeister C, Sohrabi-Jahromi S, Morgenstern B: Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*. 2017. Publisher Full Text
3. Klötzl F, Haubold B: Phylonium: fast estimation of evolutionary distances from large samples of similar genomes. *Bioinformatics*. 2020; **36** (7): 2040-2046 Publisher Full Text
4. Röhling S, Linne A, Schellhorn J, Hosseini M, et al.: The number of k-mer matches between two DNA sequences as a function of k and applications to estimate phylogenetic distances.*PLoS One*. 2020; **15** (2): e0228070 PubMed Abstract | Publisher Full Text
5. Yi H, Jin L: Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res*. 2013; **41** (7): e75 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Software development for sequence comparison and phylogeny. Alignment-free sequence comparison.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 26 November 2020

https://doi.org/10.5256/f1000research.29746.r74633

© 2020 Ondov B. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

Brian Ondov
Genome Informatics section, National Human Genome Research Institute, USA, Bethesda, MD, USA

The manuscript describes a series of simulations assessing various parametric transformations of MinHash-based genomic distance estimates into more robust evolutionary distances. It concludes that (1) MinHash methods accurately estimate nucleotide distances, (2) that accurate estimates of corrected evolutionary distances can be derived from these, with the more parameterized of two models performing better, and (3) that accurate phylogenies can be inferred from those estimates.

The manuscript is well written and generally clear, although some of the methods are described so technically as to be hard to follow (e.g. p_{max} and $p_{r>0.99}$ in "MinHash-based p-distance approximation").

My main concern, which is not large, is the choice of the sketch size, *s* (or σ). The MinHash family of algorithms assumes *s* << *G* (genome size). The smallest sketch size tested here, however, is 1/5 of the average genome size. This size, described as "moderately large," is actually two or three orders of magnitude larger than Mash's default, for example. The justification for this is to lower the error bounds. As a consequence, however, what is being tested here is, in a sense, the viability

of pure k-mer counting a la Fan *et al.*, rather than the MinHash approaches, which trade some accuracy of the Jaccard estimation for speed. This is still useful, but could be somewhat misleading for those that need the speed benefits of running MinHash with more typical parameters. Additionally, since the sweep of *s* is linear, but MinHash error bounds relate to an exponential of *s*, the sweep ends up being less informative than it could be, which is likely why the columns of Figure 1 are nearly identical.

As a simple remedy, I would suggest sweeping the parameter *s* exponentially rather than linearly, starting closer to the default sketch size for the tools. It could still end closer to the genome size to get a sense of the behavior of MinHash as it degenerates to the actual k-mer Jaccard score. Along these lines, it would also make sense for the later phylogenetic experiments to either use a smaller *s* or do another sweep of values.

As a minor point when listing relevant software in the introduction: Dashing is based on HyperLogLog sketching, which is similar in concept to MinHash but algorithmically distinct.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: High-Performance Computing, Locality Sensitive Hashing, Sequence Alignment

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research