Horizontal Gene Transfer Phylogenetics: A Random Walk Approach

Gur Sevillya,[1] Daniel Doerr,[2] Yael Lerner,[1] Jens Stoye,[2] Mike Steel,[†,3] and Sagi Snir*[†,1]

[1]Department of Evolutionary Biology, University of Haifa, Haifa, Israel

[2]Faculty of Technology, Bielefeld University, Bielefeld, Germany

[3]School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: ssagi@research.haifa.ac.il.

Associate editor: Jeffrey Thorne

Abstract

The dramatic decrease in time and cost for generating genetic sequence data has opened up vast opportunities in molecular systematics, one of which is the ability to decipher the evolutionary history of strains of a species. Under this fine systematic resolution, the standard markers are too crude to provide a phylogenetic signal. Nevertheless, among prokaryotes, genome dynamics in the form of horizontal gene transfer (HGT) between organisms and gene loss seem to provide far richer information by affecting both gene order and gene content. The "synteny index" (SI) between a pair of genomes combines these latter two factors, allowing comparison of genomes with unequal gene content, together with order considerations of their common genes. Although this approach is useful for classifying close relatives, no rigorous statistical modeling for it has been suggested. Such modeling is valuable, as it allows observed measures to be transformed into estimates of time periods during evolution, yielding the "additivity" of the measure. To the best of our knowledge, there is no other additivity proof for other gene order/content measures under HGT. Here, we provide a first statistical model and analysis for the SI measure. We model the "gene neighborhood" as a "birth–death–immigration" process affected by the HGT activity over the genome, and analytically relate the HGT rate and time to the expected SI. This model is asymptotic and thus provides accurate results, assuming infinite size genomes. Therefore, we also developed a heuristic model following an "exponential decay" function, accounting for biologically realistic values, which performed well in simulations. Applying this model to 1,133 prokaryotes partitioned to 39 clusters by the rank of genus yields that the average number of genome dynamics events per gene in the phylogenetic depth of genus is around half with significant variability between genera. This result extends and confirms similar results obtained for individual genera in different manners.

Key words: **gene order, horizontal gene transfer, Markovian processes, phylogenetics.**

Article

Introduction

Building accurate evolutionary trees depicting the history of life on earth is among the most central and important tasks in biology. The leaves of the tree correspond to contemporary extant species and the tree edges (or branches) represent evolutionary relationships. Despite dramatic advances in the extraction of such molecular data, of ever increasing quality, reconstructing an evolutionary tree is still a major challenge requiring reliable approaches for inferring the true evolutionary distances between the species at the tips (leaves) of the tree. This is particularly intensified in prokaryotes due to horizontal gene transfer (HGT), a mechanism by which organisms transfer genetic material not through vertical inheritance (Doolittle 1999; Ochman et al. 2000; Koonin et al. 2001). HGT is pervasive, and links distant lineages in the tree of life, turning it into the "network of life" (Doolittle 1999; Martin 1999; Wolf et al. 2002). Estimates of the fraction of genes that have undergone HGT vary widely with some as high as 99%. Nevertheless, it is widely accepted that even among prokaryotes, the main evolutionary signal, depicting the dominating direction of genetic information flow, is vertical (Beiko et al. 2005; Puigbo et al. 2010) and the evolutionary distances should adhere to this vertical trend.

Evolutionary distances can be inferred from a variety of sources, including morphological, genetic, and other differences between the species at hand. The sought-for tree should preserve the property that the length of the path between any two organisms at its leaves equals the inferred pairwise distance between these organisms. When such a tree exists, these distances are said to be "additive" (Semple et al. 2003).

Over the past few decades, it has become apparent and accepted that statistical modeling, as opposed to parsimony (or combinatorial) approaches, is more accurate and hence is the preferred methodology for phylogenetics. Consequently, vast efforts have been made, first to model data accurately, and then to develop efficient inference methods for the data. In this approach, a fundamental first step is finding and demonstrating provable additivity of a distance measure.

The most common approach to infer evolutionary distance that is also considered accurate is by analyzing point mutations in a "marker gene," commonly a ubiquitous housekeeping gene, shared by all the taxa under study.

Nevertheless, such a gene is highly conserved by definition and hence cannot provide a strong enough signal for sorting the shallow branches of the prokaryotic tree. To cope with the latter, one can consider applying these corrected measures to a concatenation of genes, not necessarily shared by all species (Ciccarelli et al. 2006; Swingley et al. 2008; Rinke et al. 2013). Although this approach provides a richer signal, it may suffer from the problem of where the signal of the concatenated tree comes from, as demonstrated in Thiergart et al. (2014). Specifically, cases where an edge in the inferred tree has very high support, while this edge is not shared by any source (gene) tree, are possible. This violates the premise that the species tree should represent the "main trend of evolution" due to ample evidence that a strong tree-like signal can be extracted despite the presence of extensive HGT (Beiko et al. 2005; Puigbo et al. 2010). All the techniques mentioned above rely on point mutations in the shared gene sequences and hence dubbed "sequence-based" techniques. Approaches taken to cope with the problems described above and provide a complementary information to the point mutation rely on the dynamics of prokaryotes' genomes and are broadly divided into gene-order- and gene-content-based, and hence dubbed "gene-based techniques." Under the order-based approach (Sankoff 1992; Hannenhalli and Pevzner 1999), two genomes are considered as permutations over the gene set and distance is defined as the minimal number of operations needed to transform one genome to the other. The content-based approach (Snel et al. 1999; Tekaia and Dujon 1999) ignores gene order entirely, and similarity is defined as the size of the set of shared genes. The "synteny index" (SI) (Shifman et al. 2014; Adato et al. 2015) was suggested as an alternative method to the above techniques, allowing unequal gene content (GC) on one hand while accounting for the order among the shared genes.

Although a statistical framework has been devised for part of these gene-based models, in the context of corrected distance-based estimations, similarly to the sequence-based correction, to the best of our knowledge, all of them differ from the present work either by the model assumed (gene gain/loss vs. rearrangement; Sankoff and Nadeau 1996; Wang and Warnow 2001; Biller et al. 2015; Serdoz et al. 2017), the model input (GC vs. order; Huson and Steel 2004; Woodhams et al. 2013), or the type of analysis employed (stochastic processes vs. representation theory; Sumner et al. 2017; Terauds and Sumner 2019). All These gene-based models consider the sequence of genome dynamics events that may have led to the observed differences and select the most likely (as opposed to the most parsimonious) explanation. This model-, likelihood-based approach has acquired wide acceptance in the evolutionary community for its robustness and generality (Felsenstein 1981; Hendy et al. 1994).

A related line of research focuses on "reconciling" between a gene tree and a species tree. These works are also divided into parsimony/combinatorial-based approaches (Nakhleh et al. 2005; Doyon et al. 2010), and model/likelihood-based approaches (Szöllősi et al. 2013; Sjöstrand et al. 2014). Under both approaches, a sequence of events, acting on the species tree, and yielding the given gene tree, is sought. These events may contain other than HGT and are commonly denoted "duplication, transfer, loss" (DTL) (Stolzer et al. 2012; Bansal et al. 2018). The combinatorial approach seeks a shortest sequence of DTL events leading to the given gene tree, while the likelihood-based approach assumes a model and attempts to optimize its parameters. Both these approaches do not focus on tree reconstruction let alone based on gene order between multiple genes, and therefore, even the model assumed is different.

In this work, we provide for, the first time, such a model for the SI approach in which we model HGT events as a continuous time Markovian process. Based on this, we show that the gene neighborhood in a genome behaves as a birth–death–immigration random process. This allows us to map its SI score to the expected number of "jumps" a gene has undergone from the two genomes' divergence event, and thus makes the SI measure additive. This additivity proof is asymptotic and assumes infinite data in the form of genome size. Therefore, we also devise a heuristic approach taking realistic sizes of input into account, such as the genome size and the gene neighborhood size. The model relies on exponentially decaying functions and provides us with realistic estimates of number of transfers occurring in a genome, which could not be derived by considering the raw, uncorrected SI that was used in Shifman et al. (2014) and Sevillya and Snir (2019). We first demonstrate the accuracy of the model under an extensive simulation study, both pairwise and multigenome setting, demonstrating it indeed attains additivity as required, as opposed to other approaches or the raw, uncorrected, SI. Next, we test the model under a more realistic realm in which insertions and deletions occur, and show that our heuristic copes satisfactorily with this type of events. Applying this heuristic model to 1,133 prokaryotic genomes from the orthology database EggNOG (Powell et al. 2012) yielded a set of 39 clusters of closely related taxa. Having a distance correction function at hand allows an approximate estimation of genome dynamics activity inside these clusters, showing that the average amount of this activity in bacterial genera is around half the genome size, but highly variable.

Results

Asymptotic Estimation of Divergence Times

We now introduce a random process that will play a key role in the analysis of the random variable $\bar{\mathrm{SI}}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$. Consider the location of a gene g_ℓ, not being transferred during time period t, with respect to another gene $g_{\ell'}$. WLG assume $\ell > \ell'$ and let $j = \ell - \ell'$. Now, there are j "slots" between $g_{\ell'}$ and g_ℓ in which a transferred gene can be inserted, but only $j-1$ genes in that interval, that can be transferred. Obviously, a transfer into that interval moves $g_{\ell'}$ one position away from g_ℓ, and a transfer from that interval, moves $g_{\ell'}$ one

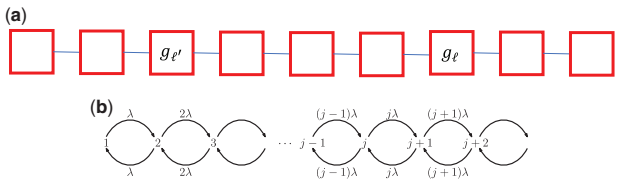

FIG. 1. (*a*) A genome: there are three genes between $g_{\ell'}$ and g_ℓ that can be transferred out of the interval $[\ell', \ell]$, but four slots where new genes can be transferred into the same interval, and hence $X_0 = 4$. (*b*) The continuous time Markov chain defined by X_t: X_t moves to state $j+1$ at rate $j\lambda$ and to state $j-1$ at rate $(j-1)\lambda$.

position closer to g_ℓ (see fig. 1*a* for illustration). The above can be modeled as a continuous-time random walk on state space $1, 2, 3, \ldots$.

Where $g_{\ell'}$ is considered as a reference gene, g_ℓ moves right or left with respect to $g_{\ell'}$, and the state of the process represents the distance between $g_{\ell'}$ and g_ℓ. Transitions from state j to $j+1$ occur due to a transfer into the interval $[\ell', \ell]$ and at rate $j\lambda$ (for all $j \geq 1$) and from j to $j-1$ due to a transfer from the interval $[\ell', \ell]$ and at rate $(j-1)\lambda$ (for all $j \geq 2$), with all other transition rates 0. This is thus a (generalized linear) birth–death process, and the process is illustrated in figure 1*b*. As the process is not affected by the specific values of ℓ and ℓ' (rather by their difference), we can ignore them and let X_t denote the random variable that describes the state of this random walk (i.e., the distance of some g_ℓ from a reference gene $g_{\ell'}$—a number 1, 2, 3, etc.) at time t.

The process X_t is slightly different from the much-studied critical linear birth–death process, for which the rates of birth and death from state j are both equal to j (here, the rate of birth is j, but the rate of death is $j-1$), and for which 0 is an absorbing state. However, this stochastic process is essentially a translation of a critical linear birth–death process with immigration rate equal to the birth–death rate λ (the inclusion of immigration has the effect that 0 is no longer an absorbing state). This is the key to establishing both parts of the lemma below. We first define $p_{ij}(t)$ as the transition probability for X_t to be at state j given that at time 0 it was at state i. Note that, here, i and j cannot be ignored as they do not specify absolute locations, rather locations relative to a reference gene, that is, it can be seen that $p_{ij}(t) \neq p_{(i+r)(j+r)}(t)$. Formally,

Definition 1. *For each ordered pair* $i, j \in \{1, 2, 3 \ldots, \}$, let $p_{ij}(t) = \mathbb{P}(X_t = j | X_0 = i)$.

Lemma 1.

a. *The transition probabilities* $p_{ij}(t)$ *satisfy the following tridiagonal differential system:*

$$\frac{1}{\lambda}\frac{\mathrm{d}p_{ij}(t)}{\mathrm{d}t} = -(2j-1)p_{ij}(t) + jp_{i(j+1)}(t) + (j-1)p_{i(j-1)}(t),$$

subject to the initial condition:

$$p_{ij}(0) = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases}$$

b. *The expected value of* X_t *grows as a linear function of* t. *Specifically,*

$$\mathbb{E}[X_t | X_0 = i] = i + t\lambda.$$

Moreover, X_t *has no stationary distribution.*

c. *Conditional on* $X_0 = i$, *and for fixed value of* t *and value* $B > \lambda t$, *the probability that the supremum of* X_s *over the interval* $[0, t]$ *exceeds* B *is at most* $(i-1)/(B-\lambda t)$. *In particular, this probability tends to zero as* $B \to \infty$.

Proof. Consider a critical linear-birth–death process with immigration in which the birth rate and death rate are both equal to λ, and the immigration rate is also equal to λ. Let Y_t denote the random variable counting the number of individuals in the system, and notice that Y_t takes values in $0, 1, 2, \ldots$, in contrast to X_t which takes values from 1 upward.

Then, the process Y_t is stochastically identical to the process $X_t - 1$. To see this, simply note that both processes are Markovian, and the transition probabilities for $Y_t + 1$ correspond precisely to those indicated in figure 1*b*. Thus, if we let $\tilde{p}_{ij} := \mathbb{P}(Y_t = j | Y_0 = i)$, then

$$\mathbb{P}(X_t = j | X_0 = i) = \mathbb{P}(Y_t = j-1 | Y_0 = i-1),$$

and so,

$$p_{ij}(t) = \tilde{p}_{i-1j-1}(t).$$

Now the (tridiagonal) system of differential equations for $\tilde{p}_{ij}$ is the well-known forward Kolmogorov differential equations (see supplementary text, Supplementary Material online and, e.g., Section 6.4.4. of Allen 2010) and by translation, these provide the equations in Part (a).

For Part (b), observe that:

$$\begin{aligned} \mathbb{E}[Y_t - \lambda t] &= \mathbb{E}[Y_t] - \lambda t = \mathbb{E}[X_t - 1] - \lambda t \\ &= \mathbb{E}[X_t] - 1 - \lambda t. \end{aligned} \quad (1)$$

Now, $Y_t - \lambda t$ is a Martingale process, with $\mathbb{E}[Y_t - \lambda t] = \mathbb{E}[Y_0]$ for all $t \geq 0$. Thus, if $X_0 = i$, then

$$\mathbb{E}[Y_t - \lambda t] = \mathbb{E}[Y_0] = i - 1.$$

Combining this with equation (1) gives $\mathbb{E}[X_t] = i + \lambda t$ as claimed.

That X_t has no stationary distribution follows from Theorem 6.1 of Allen (2010).

Part (c) is established using the Doob Martingale Inequality (Grimmett et al. 2001), which states that for a martingale process Z_t the following inequality holds for any $c > 0$:

$$\mathbb{P}\left(\sup_{0 \leq s \leq t} Z_s \geq c\right) \leq \mathbb{E}[Z_t]/c.$$

Applying this to the martingale $Z_s = Y_s - \lambda s$, and noting that $X_s = Y_s + 1$ gives:

$$\begin{aligned} \mathbb{P}\left(\sup_{0 \leq s \leq t} X_s > B\right) &= \mathbb{P}\left(\sup_{0 \leq s \leq t} Y_s \geq B\right) \\ &\leq \mathbb{P}\left(\sup_{0 \leq s \leq t} Y_s - \lambda t \geq B - \lambda t\right) \\ &\leq \mathbb{E}[Y_t - \lambda t]/(B - \lambda t) \ \text{(by Doob's inequality)} \\ &\leq \mathbb{E}[Y_0 - 0]/(B - \lambda t) \ \text{(by part b above)}, \end{aligned}$$

and the last term on the right is just $(i-1)/(B-\lambda t)$, as claimed. ■

We now set to calculate the probability that a "nonjumping" gene stays in the k-neighborhood of some reference gene. Let $q_{ik}(t)$ be the conditional probability that $X_t \in [k]$ (where $[k]=\{1,2,\ldots,k\}$) given that $X_0=i$. Thus,

$$q_{ik}=\sum_{j=1}^{k} p_{ij}(t). \tag{2}$$

In order to state Theorem 1, we need to define the following quantity. Let,

$$q_k(t) := \frac{1}{k}\sum_{i=1}^{k} q_{ik}(t) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}(t). \tag{3}$$

In words, $q_k(t)$ is the probability that for a gene at an initial state i (i.e., distance from a reference gene) chosen uniformly at random between 1 and k, the process X_* is still between 1 and k after time t (equivalently, $q_k(t)$ is the probability that a birth–death–immigration process with all three rates equal to λ and an initial state chosen uniformly at random between 0 and $k-1$ takes a value at time t that is also at most $k-1$).

Theorem 1. *For any given value of t, and as n grows:*

$$\bar{\mathrm{SI}}(\mathcal{G}_0^{(n)},\mathcal{G}_t^{(n)}) \xrightarrow{p} \exp(-2\lambda t)q_k(t),$$

where $\xrightarrow{p}$ *denotes convergence in probability.*

Corollary 1.

Thus, if the function $t\mapsto \exp(-2\lambda t)q_k(t)$ *has an inverse* φ, *then*

$$\varphi(\bar{\mathrm{SI}}(\mathcal{G}_0^{(n)},\mathcal{G}_t^{(n)})) \xrightarrow{p} t.$$

In particular, for sufficiently large n (including that $\lambda t \ll n$), one can use the expression on the left to estimate (an additive) evolutionary distance and hence construct a tree consistently. The proof of Theorem 1 is fairly technical and hence is given in the Supplementary Material online.

Analysis under Biologically Realistic Values

In the previous section, we have dealt with asymptotic cases, where the size of the genome goes to infinity and therefore, the neighborhood size of gene g_ℓ, has size $2k$, where k is constant, and so of order $o(n)$, and thereby negligible in relation to the length n of the genome. However, real bacterial genomes comprise around 5,000 genes and here, many relaxations used above do not hold. Therefore, in order to analyze real data, we must find a realistic model that imitates real life sizes. Developing analytical results here is substantially harder as the setting is richer than before. Hence we devised the following approach. We first simulate the model and try to learn its behavior. Next, we try to fit the parameters to the model to get the best estimation of the observed behavior. Also and importantly, as the focus here is to develop a "distance measure" rather than a similarity measure as before, we use hereafter the quantity $1-\mathrm{SI}$ that we denote d_{SI} to avoid confusion. Note that, in contrast to SI, d_{SI} starts at zero (identical genomes) and grows in time.

We start with some basic observations that are relevant to this part for the settings different from before.

The next simple lemma gives an upper bound on d_{SI} when $t\to\infty$. We will use it during our simulation study to provide a scaling factor to the inferred function.

Lemma 2. *Under the uniform jump model, when* $t\to\infty$, $d_{\mathrm{SI}} = 1-\frac{2k}{n-1}$.

Proof. There are $2k$ genes in the original neighborhood $N_{2k}(g,\mathcal{G}^{(n)}(0))$ of g_ℓ. These are scattered uniformly in $\mathcal{G}^{(n)}(\infty)$ and hence also in $N_{2k}(g_\ell,\mathcal{G}^{(n)}(\infty))$. Therefore, in particular, the expected number of these genes in $N_{2k}(g_\ell,\mathcal{G}^{(n)}(\infty))$ is $\frac{2k}{n-1}$ and the result follows. ■

The Linear Model

We start with a simple case that will serve as the basis for the subsequent development. We first define the following.

Definition 2. *The "disjoint events assumption" (DEA) assumes that a transferred gene* g_ℓ *leaves its original, unviolated neighborhood and lands at a new, unviolated neighborhood.*

In other words, under DEA, all neighborhoods associated with transfer events are disjoint. We note that such an assumption violates the randomness of our model as we cannot assume this under a random model. Nevertheless, it holds with high probability for small t, that is, between closely related species.

It is easy to see that under DEA, Lemma 4—the SI local lemma—holds in equality and therefore, the contribution of each event to d_{SI} is approximately $\frac{3}{n}$. Hence, under DEA, for relatively small number of HGT events M, the expected d_{SI} is $\frac{3M}{n}$.

The Expanded Model

As the DEA, and hence linearity of d_{SI}, holds for a relatively short time, we set to develop a more realistic model that also considers nondisjoint events. As discussed earlier, the goal here is not to find an exact model as in the asymptotic model of "Asymptotic Estimation of Divergence Times," rather to find a sound approximation to it. The first approach then is to obtain intuition via simulation study. Figure 2 depicts results of a simulation study between two genomes. Detailed description of this study is provided in the second part of the supplementary text, Supplementary Material online. Figure 2*a* shows d_{SI} as a function of the number events, for various k's. As k is not anymore negligible, and we cannot ignore events at the tips of the tips of the genome, genomes were assumed to be circular. In addition, the theoretical linear model is presented, and we can see that this model (which assumes disjoint events, DEA) departs from the simulation results after about 200 events (20% of the genome size) or less, depending on k. Interestingly, as was shown theoretically

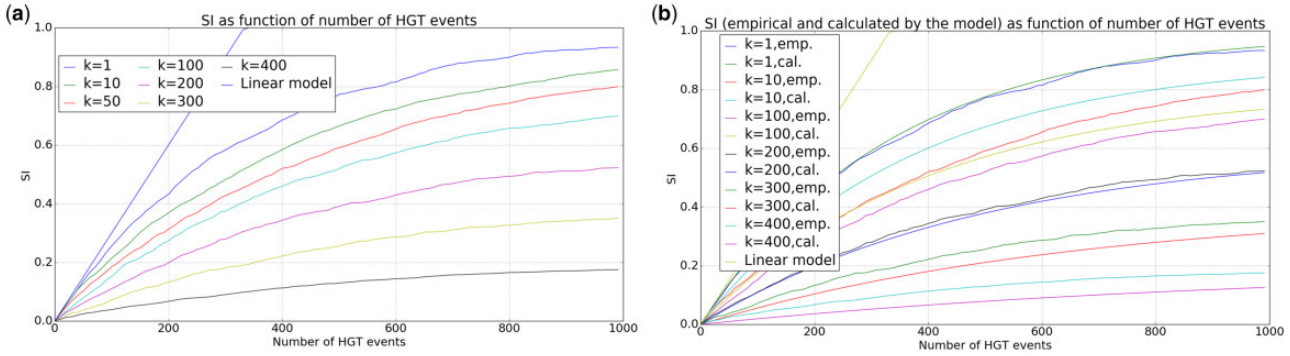

FIG. 2. Results of pairwise simulation between two genomes under realistic values: (*a*) d_{SI} as a function of number of HGT events. Simulations over 1,000-gene genome sizes, under various *k*'s ($k = 1, 10, 50, 100, 200$). (*b*) Actual number of events (as shown in the left) versus predicted values (no. of HGT) as calculated by our suggested model in equation (9).

(Lemma 4), this line is independent of *k*. In addition, we can see that the maximum value of d_{SI} behaves according to Lemma 2.

Guided by the results depicted in figure 2*a*, we can now approach the task of developing a heuristic model tracking this behavior. As can be perceived from the figure, all curves follow a diminishing increase in the measured quantity—d_{SI}, alluding to an "exponential decay" trend. Now, a quantity *B* is subjected to exponential decay (or growth) if its increase rate is proportional to its current value, that is, $\frac{dB}{dt} = \Lambda B$, where *B* is the quantity measured, *t* is time, and $\Lambda < 0$ is the decay constant. Such a function behaves as $B_t = B_0 e^{\Lambda t}$ (Durrett 2008). Note that, *B* here is not d_{SI} the rate of change (increase) in the expected value of d_{SI} with respect to the expected number of HGT events—λt. By integrating this expression, we get the expected value of d_{SI}, $\mathbb{E}[d_{SI}]$, resulting from HGT events. Our goal is to develop an expression for the expected change in d_{SI} after time *t*, that under our Poisson model, is linear in the HGT events. Conforming with the derivation of the asymptotic part above, we develop the model with respect to time, and relate it to number of HGT events, only at the end. We will denote this target expression as $\frac{d}{dt}\mathbb{E}[d_{SI}]$, since it is the derivative of $\mathbb{E}[d_{SI}]$. That is, the model we develop, calculates the expected change in d_{SI} per time, which by integration yields the expected number of events. As our approach is simulation-based (in contrast to analytic), the model (function) sought needs to approximate best the jump model that we simulate (see more details below and in the Supplementary Material online). Finally, we determine how to set the actual parameters (constants) in the developed function in order to approximate best the simulation results.

Recall that, HGT events are distributed uniformly throughout the genome. Considering this, we start by finding the number of events required for each gene g_ℓ to get an d_{SI} score of $\frac{1}{2k}$ (and hence total d_{SI} of $\frac{1}{2kn}$). To tackle this question, we assume that most of the events conform with DEA, as the events are uniformly distributed and genomes are relatively similar—implying small d_{SI}, the simulation process is at its beginning (i.e., next to the origin, see fig. 2*a*).

Observation 1. *After* $\frac{n}{6k}$ *events, the expected* d_{SI} *at every gene* g_ℓ*, and hence the total* d_{SI}*, is* $\frac{1}{2k}$.

Proof. Consider the genome as a sequence of $\frac{n}{2k}$ adjacent (nonoverlapping) 2*k*-neighborhoods. By Lemma 2, under DEA, each event contributes $\frac{6k}{2kn} = 3/n$ to the total d_{SI}. Hence, under uniform distribution, we get that after $\frac{n}{6k}$ events the expected number of events occurring at a neighborhood is 1/3, yielding contribution $\frac{n}{6k}\frac{3}{n} = \frac{1}{2k}$ to the total d_{SI}, and the observation follows. ■

Recall from Definition 3 that a violation at a neighborhood is a gene not originally from that neighborhood.

Lemma 3. *If the neighborhood of each gene* g_ℓ *contains m violations, the (expected) addition to the* d_{SI} *score resulting from the next event is* $\frac{6k-3m}{2kn}$.

Proof. We will show that this holds for any $m < 2k$. Having *m* violations in each gene's neighborhood, means that each gene has SI (i.e., $1 - d_{SI}$) score of $\frac{2k-m}{2k}$, that is, each gene is missing *m* of its original neighbors. Let us consider the next event, as some gene g_ℓ is jumping to a new neighborhood. First, gene g_ℓ, as all other genes, is missing *m* old neighbors. That is, when making this jump, it loses its remaining $2k - m$ original neighbors, and also, these $2k - m$ genes are losing gene g_ℓ as their old neighbor. That is, its contribution to the d_{SI} score is $\frac{2(2k-m)}{2kn}$. Now, in the new neighborhood of gene g_ℓ, there are 2*k* genes that are now having g_ℓ in their current 2*k*-neighborhood. Each of these 2*k* genes is already missing *m* original neighbors. For each of these genes, the probability that gene g_ℓ pushed out of the 2*k*-neighborhood an original neighbor is $\frac{2k-m}{2k}$. When this is the case, there is a contribution of $\frac{1}{2kn}$ to the d_{SI} score from this loss, and hence the expected contribution for a single 2*k*-neighborhood is $\frac{2k-m}{2k} \cdot \frac{1}{2kn}$. Therefore, for the new location of g_ℓ, the calculation of the expected contribution to d_{SI} is: The expected contribution for a single 2*k*-neighborhood times the number of affected 2*k*-neighborhoods,

$$2k\frac{2k-m}{2k}\frac{1}{2kn} = \frac{2k-m}{2kn}.$$

Summing the contribution from the old and new $2k$-neighborhoods and the jumping gene g_ℓ, we end with a contribution to the total d_{SI} score of $\frac{3(2k-m)}{2kn} = \frac{6k-3m}{2kn}$. ■

As shown, the expected contribution of the next event, after $m\frac{n}{6k}$ events, is $\frac{6k-3m}{2kn} = \frac{3}{n} - \frac{3m}{2kn}$. This means that the change in the contribution to d_{SI} for each period of $n/(6k)$ events is as follows:

$$\left(\frac{3}{n} - \frac{3m+3}{2kn}\right) - \left(\frac{3}{n} - \frac{3m}{2kn}\right) = -\frac{3}{2kn}.$$

Having proved that, we recall that the quantity B being measured is $\frac{d}{dt}\mathbb{E}[d_{SI}]$, so the expression $\frac{dB}{dt}$ that changes over time (or HGT events) is $\frac{d^2}{dt^2}\mathbb{E}[d_{SI}]$. We see that for a time period of $\frac{n}{6k}$ events, $\frac{d^2}{dt^2}\mathbb{E}[d_{SI}]) = -\frac{3}{2kn}$ and this is $\frac{dB}{dt}$ for this time period. Now, recall that under exponential decay, $\frac{dB}{dt} = \Lambda B$, so in order to find Λ, we write:

$$\Lambda = \frac{\frac{dB}{dt}}{B} = \frac{-\frac{3}{2kn}}{\frac{3}{n} - \frac{3m}{2kn}} \approx \frac{-\frac{3}{2kn}}{\frac{3}{n}} = -\frac{1}{2k}. \quad (4)$$

A similar procedure for the next time periods (i.e., for having expected violations 2 and 3) will yield, as long as $\frac{3}{n} \gg -\frac{3m}{2kn}$, the same $\Lambda = -\frac{1}{2k}$, as indeed required by such growth (exponential). As this derivation is involved, in the Supplementary Material online, we provide further details of the process.

Of course, the analysis above is crude and by no means provides a rigorous proof for the exponential decay, as this should be significantly harder even than the asymptotic case. Nevertheless, it provides us with intuition and insight of what are the parameters to the decay function as we next show. First, we note that the $\Lambda = -\frac{1}{2k}$ obtained is aggregated over an entire time period of $\frac{n}{6k}$ events, so in order to put it in the formula, we need to divide Λ by this factor, $\frac{n}{6k}$, yielding:

$$\Lambda^* = \frac{\Lambda}{\frac{n}{6k}} = \frac{-\frac{1}{2k}}{\frac{n}{6k}} = -\frac{3}{n}. \quad (5)$$

Now we want to plug it into the exponential decay function: $B_t = B_0 e^{\Lambda^* t}$. But for this, we first need to find B_0. As the first contribution to d_{SI} per event is $3/n$, we set $B_0 = 3/n$ yielding:

$$\frac{d}{dt}\mathbb{E}[d_{SI}] = B_t = B_0 e^{\Lambda^* t} = \frac{3}{n} e^{-\frac{3}{n}t}. \quad (6)$$

In order to obtain $\mathbb{E}[d_{SI}]$, we need to integrate equation (6). Specifically, we are interested in the definite integral in the interval $[0, t]$:

$$\int_0^t \frac{d}{dt}\mathbb{E}[d_{SI}] = \int_0^t \frac{3}{n} e^{-\frac{3}{n}t} = -e^{-\frac{3}{n}t}\Big|_0^t = 1 - e^{-\frac{3}{n}t}. \quad (7)$$

Finally, recall that we want to express d_{SI} with respect to the number of HGTs, that is under our model λt^*, where t^* is the real amount of time. Therefore, replacing the generic t used in the development above with λt^* results in:

$$\mathbb{E}[d_{SI}] = 1 - e^{-\frac{3}{n}\lambda t^*}. \quad (8)$$

Now, as n is finite here, the latter tends to one as $t \to \infty$. However, recall that by Lemma 2, d_{SI} is bounded from above by $1 - \frac{2k}{n-1}$ so we treat this as a scaling factor for our decay function. Our final refinement addresses cases of relatively large neighborhoods (e.g., $k = 100, 200, 300, 400$) taking into consideration the case of the gene being transferred back into its original neighborhood, as was shown above to be $\frac{3-\frac{5k}{n-1}}{n}$ instead of $3/n$. Therefore, we obtain:

$$\mathbb{E}[d_{SI}] = \left(1 - \exp\left(-\frac{3 - \frac{5k}{n-1}}{n}\lambda t\right)\right)\left(1 - \frac{2k}{n-1}\right). \quad (9)$$

Although this study is not as rigorous as the asymptotic case, rather simulation based, its strength is by considering practical values as found in nature. Moreover, equation (9) is invertible hence allows us to infer the expected distance (number of HGT events along a time period, denoted hereafter $\widehat{d} = \lambda t$) from a given SI between two genomes evolving through the jump model (see precise derivation in the supplementary text, Supplementary Material online). Indeed, in figure 2*b*, we show results from the same simulation study as described in figure 2*a*, however, here, we contrast the real, simulated HGT to the expected calculated HGTs under this expanded model, as obtained from equation (9). Model (expected) versus real (simulated) number of HGTs for various values of k are shown. As can be seen, even for very large neighborhood size k, reconstruction (of #HGTs) remains quite accurate and this is due to the refinement of incorporating k into the exponent.

In the supplementary text, Supplementary Material online, we provide further, more detailed simulation study of pairwise distances. This study shows that the expected number of events—the "distance," denoted $\widehat{d}$, which is the inverse of the measure defined in equation (9)—is within a small, constant fraction of 0.3 of the actual number of events generated. Furthermore, as this fraction is constant, it yields the additivity of the measure, meaning that a tree with distances between any two leaves is identical to distances between the genomes. Such a property is necessary for phylogenetic reconstruction and we confirm that in the phylogenetic simulations. We also extend the basic jump model to include "indels"—corresponding to events of gene gain and loss. Our simulation results (also in the supplementary text, Supplementary Material online) show that additivity is still maintained and even at a closer ratio of 0.8. We will return to this property of the accurate and additive measure under real life conditions, when analyzing real genomic data among genera.

Tree Reconstruction Experiments

Our final simulation study investigates the performance of the new measure in phylogenetic reconstruction. We provide a detailed description of the setting and the procedures taken in the supplementary text, Supplementary Material online, while here, we mostly report the results.

We used ALF (Dalquen et al. 2012) to generate genomes at the tips of a prespecified tree under a controlled regime of

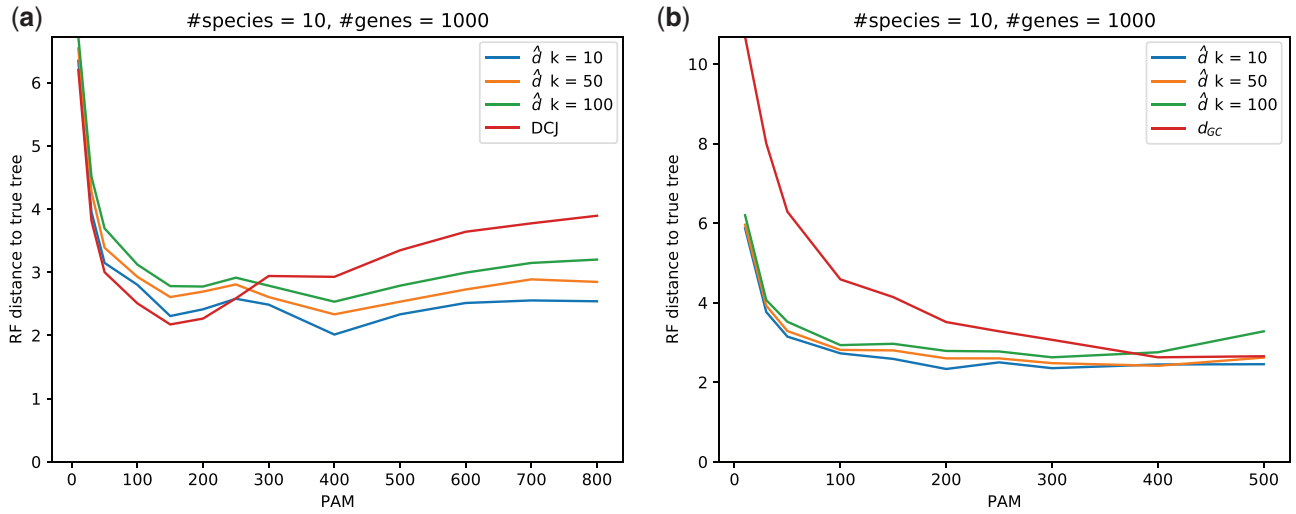

FIG. 3. (*a*) RF distances between model tree and inferred tree, as a function of events generated. Events are spread uniformly over a binary tree with ten leaves. All genomes at the leaves are of the same gene content. (*b*) Mean RF distances between the reconstructed and true trees of the genome evolution experiment with insertions and deletions.

mutational events. ALF is an overly powerful tool for our needs in this work. Here, ALF was merely used to generate model trees and then genome dynamics events along the branches of the tree. ALF controls the intensity of the mutational activity with the parameter PAM (point, or percent, accepted mutations). The greater the PAM, the more mutations are generated, producing stronger signal at the leaves. In our case, ALF generated trees over ten leaves, and subsequently gene sequences (i.e., genomes) at the leaves, which in turn serve as inputs to the reconstruction methods studied. By applying the inverse of equation (9) (see exact function in the Supplementary Material online) between any pair of these genomes, we obtain a distance matrix, to which a standard distance-based tree reconstruction is applied. This reconstructed tree is compared with ALF's model tree. We note that the trees generated by ALF are very challenging as they exhibit hard combinations of very short and long branches, and the results confirm this.

In the first experiment, plotted in figure 3*a*, the same jump model as discussed above is studied. Robinson–Foulds (RF) distance is measured between the model tree and the tree reconstructed by d_{SI}, as a function of the mutational activity along the tree, measured in PAM (see exact details in the supplementary text, Supplementary Material online). Three values of k are examined, 10, 50, and 100, where $k = 10$ achieves the best score. As demonstrated clearly, for low values of PAM, no information is produced at the leaves and reconstruction is very poor. Nevertheless, this improves sharply already at low PAM values, attesting to the power of gene order reconstruction. The fact that reconstruction maintains the same accuracy level even for high PAM values, where around 6,000 events are generated (these are spread over the entire tree), attests on the additivity and robustness of the measure, as also shown in supplementary figure 4*a*, Supplementary Material online. Also plotted in the figure is the double-cut-and-join (DCJ) distance (Bergeron et al. 2006) that under the jump model gives the same number of jumps. Notwithstanding, DCJ is a parsimony measure and as shown, saturates already at moderate levels of HGT, unlike d_{SI}. In figure 3*b*, we extended the jump model to include gain and loss events, where new genes are inserted into the genomes and others are lost. As can be seen, d_{SI} behaves similarly to the pure jump model, complying with the pairwise results shown in supplementary figure 5*a*, Supplementary Material online. Here, the DCJ approach cannot be employed and we used the GC approach depicted by the d_{GC} curve. d_{GC} performs similarly to d_{SI} at the top of the mutation spectrum, however, its sensitivity is significantly lower than d_{SI} at the more relevant values of PAM (agreeing with similar results obtained in Shifman et al. 2014, see fig. 8 thereof).

Result on Real Microbial Data

Once a plausible model for biologically relevant sizes was found, we aimed to use it to infer genomic activity in microbial data. We used the EggNOG v3.0 database (Powell et al. 2012) which is the largest unbiased orthology database, containing protein sequences of 1,133 species, most of them bacteria. In addition, this database clusters all proteins into Clusters of Orthologous Groups (COGs) (Tatusov et al. 2001), information that is essential for the SI approach. This means that an organism is represented as a list of COG names ordered by their order of appearance in its genome. In Sevillya and Snir (2019), we partitioned the entire taxa set of 1,133 taxa into 39 clusters (subsets) of bacteria species, largely conforming with the conventional classification into genera. The exact procedure and definitions used in this partitioning are elaborated in Sevillya and Snir (2019) (specifically, Section 2.5 therein) and for the sake of completeness, we also provide a brief description of the process in the Supplementary Material online. As we now work with real genomic data, we need to relax the theoretical assumption of solely HGT events as implied by the jump model, and allow for other types of events, collectively denoted "genome dynamics events" (or GDEs) (Puigbò et al. 2014) which are largely dominated

by gene loss and gain (where gene gain is primarily HGTs and to a lesser extent gene duplications). As shown in the simulation part, the new measure $\widehat{d} = \lambda t$ accounts also for this type of events and even maintains additivity, hence we applied it to this data, aiming at inferring real biological insight. Details of this analysis are shown in a table format in supplementary table 2, Supplementary Material online, therein. Specifically, for each such cluster, we report its corresponding genus, its average d_{SI} (averaged overall pairs of genomes in that cluster), and the average number of GDEs "separating" each pair of species in that cluster (normalized by average genome size for that cluster). A histogram summarizing this data is also provided in the Supplementary Material online. We find that this parameter, the number of GDEs per gene, is normally distributed (Shapiro–Wilks test: $P = 0.238$), with a mean of 52.7%, a median of 54.1%, and a SD of 23.78%. In other words, we estimate the average number of GDEs between pairs of species inside a genus to be ~50% (±20) of the genome size. We find this result conforming and extending similar results, both in terms of intensity and deviation (Puigbò et al. 2014). For example, in Welch et al. (2002), it was found that even between three strains (i.e., subspecies) of *Escherichia coli*, the amount of genes shared between any two strains is ~40%, and similar results were also shown for the genus *Nautilia* (Smith et al. 2008). On the other hand, for the genus *Prochlorococcus*, Kettler et al. (2007) found that around 1,500 genes are shared in average between a pair of species, where the average genome size is around 1,700 genes. More examples to the above can be found in Bapteste et al. (2009).

Finally, we turned to compare the new, corrected, measure, to the old uncorrected SI. We used the same cluster set used in Sevillya and Snir (2019). In supplementary table 3, Supplementary Material online, we report the application of the two approaches to each of the clusters described earlier. Although we have no means to judge the correctness of the results, the small differences, as well as the results of the simulation study, and the comparison to previous experiment with the raw SI, suggest that both approaches perform satisfactorily. The trees and matrices are provided in Supplementary Material online.

Discussion

In this article, we have provided a first statistical modeling for the SI as a phylogenetic marker, a measure that was proved useful in prokaryotic genomics (Shifman et al. 2014; Adato et al. 2015; Sevillya and Snir 2019). The major advantage of SI is that it combines the evolutionary signal in both gene order and GC. The latter allows one to compare genomes over different gene sets (a pervasive phenomenon in prokaryotes). It permits also a comparison of the order of their shared core gene set, a signal that is ignored by purely content-based approaches.

Statistical parametric approaches are nowadays widely accepted as the method of choice in a host of applications in biology. Starting from Felsenstein's seminal demonstration of the statistical inconsistency of maximum parsimony (Felsenstein 1978) and his subsequent remedy to this flaw (Felsenstein 1981), maximum likelihood is now the gold standard in phylogenetics, and most popular packages (Felsenstein 1993) offer a likelihood-based solution. These approaches rely on one or several concatenated genes that are assumed to be shared among all taxa and hence are appropriate for comparison. Such genes however tend to be conserved and do not necessarily provide a sufficiently strong signal to trace subtle differences.

In contrast, gene order- and content-based approaches were found to provide enough signal, allowing more resolved trees (Wolf et al. 2001). Nevertheless, although such approaches have existed for several decades, to the best of our knowledge, no detailed model showing additivity and consistency under HGT, has been proposed. Similarly, no analytical proof for SI correctness has been shown. The framework developed here, models HGT as a continuous-time Markov process affecting each gene in a genome independently. This in turn provides for modeling the gene neighborhood by a birth–death–immigration process and applying tools from this field in our specific setting. We accompanied this theoretical asymptotic study with a practical, simulation-based, study accounting for real life values such as genome and neighborhood sizes. Our experiments indicate that this model is sound and attains the desired property of additivity, both in pairwise distances and multigenome phylogenetics, and also under more realistic models of gene gain and loss. In particular, the experimental results presented here demonstrated that the new model-based approach which was devised, maintains additivity even for high level of genomic activity, whereas the raw prototypic approach, under the same conditions of unequal GC, does not (Shifman et al. 2014). As the core of this work has just handled the most basic case (the jump model), extensions are planned for studying more advanced models, such as the "indel model," which we examined in the simulation part, in which new genes are added to the genome, but at an equal rate of gene loss, so the genome size is approximately fixed. We comment that the jump model corresponds to several scenarios, such as a duplication or alien gain with a subsequent deletion of the old copy. The distinction between these cases can be done based on simultaneous multigenome analysis via phylogenetics, however, this is remained for further future research. We believe that the tools developed here, theoretical and experimental, will serve as the basis for further extensions.

Materials and Methods

Throughout the article, for the sake of clarity, we will reserve the letter k to refer to the neighborhood, ℓ to specific genes, and i and j as general indexes.

We start by defining a restricted model (the "jump model") that can be perceived as a transfer between genomes over the same gene set ("equal content").

The Jump Model

Let $\mathcal{G}^{(n)}(0) = (g_1, g_2, \ldots, g_n)$ be a sequence of "genes." In our analysis, we will assume that n is large. Consider the following continuous-time Markovian process $\mathcal{G}^{(n)}(t), t \geq 0$,

on the state space of all $n!$ permutations of $g_1, g_2, \ldots, g_n$. Each gene g_ℓ is independently subjected to a Poisson process of transfer events (at a constant rate λ) in which g_ℓ is moved to a different position (also denoted as a *slot*) in the sequence, with 1) each of the possible $n-1$ positions between consecutive genes different from g_ℓ or at the start or end of the sequence and 2) this target location for the transfer selected uniformly at random from these $n-1$ possibilities.

For example, if $\mathcal{G}^{(n)}(t) = (g_1, g_2, g_3, g_4, g_5)$, then g_4 might transfer to be inserted between g_1 and g_2 to give the sequence $\mathcal{G}^{(n)}(t+\delta) = (g_1, g_4, g_2, g_3, g_5)$. The other sequences that could arise by a single transfer of g_4 are $(g_4, g_1, g_2, g_3, g_5)$, $(g_1, g_2, g_4, g_3, g_5)$, and $(g_1, g_2, g_3, g_5, g_4)$. Note, in particular, that g_ℓ does not necessarily move to a position between two genes; it can also move to the initial or the last position in the sequence. A jump can also account for a "gene loss" in which a gene jumps outside of the genome, a "gene gain" when the jump is from an alien genome, or both (gain and subsequent loss, or vice versa).

Note that, by the definition of a Poisson process, the probability that g_ℓ is transferred to a different position between times t and $t+\delta$ is $\lambda\delta + o(\delta)$, where the $o(\delta)$ term accounts for the possibilities of more than one transfer occurring in the δ time period (these are of order δ^2 and so are asymptotically negligible compared to terms of order δ as $\delta \to 0$). Moreover, a single transfer event always results in a different sequence.

Let k be any constant positive integer (note it may be possible to allow k to grow slowly with n but we will ignore this for now). Then, for $\ell \in k+1, \ldots, n-k$, the $2k$-neighborhood of gene g_ℓ in a genome $\mathcal{G}^{(n)}$, $N_{2k}(g_\ell, \mathcal{G}^{(n)})$ is the set of $2k$ genes (different from g_ℓ) that have distance at most k from g_ℓ in $\mathcal{G}^{(n)}$. We also define $\mathrm{SI}_\ell(t)$ as the relative intersection between $N_{2k}(g_\ell, \mathcal{G}^{(n)}(0))$ and $N_{2k}(g_\ell, \mathcal{G}^{(n)}(t))$ or formally:

$$\mathrm{SI}_\ell(t) = \frac{1}{2k}\left|N_{2k}(g_\ell, \mathcal{G}^{(n)}(0)) \cap N_{2k}(g_\ell, \mathcal{G}^{(n)}(t))\right| \qquad (10)$$

(this is also called the "Jaccard index" between the two neighborhoods).

Let $\bar{\mathrm{SI}}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$ be the average of these $\mathrm{SI}_\ell(t)$ values overall ℓ's between $k+1$ and $n-k$. That is,

$$\bar{\mathrm{SI}}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)}) = \frac{1}{n-2k}\sum_{\ell=k+1}^{n-k} \mathrm{SI}_\ell(t). \qquad (11)$$

The assumption of a large n discards the effect of events at the tips of the genome, or the distinction between circular or linear genomes, or effects resulting from tiny genomes. We will refer to this question when it becomes relevant, in the practical part of the work.

In the sequel, when time t does not matter, we simply use $\bar{\mathrm{SI}}$ or simply SI, where it is clear from the context. We start with a rather simple, yet very central, lemma, that we denote the "SI local lemma." Before however, we need the following definition.

Definition 3. *For an untransferred gene g_ℓ, let us define a violation as a gene $g_{\ell'}$ such that $g_{\ell'} \in N_{2k}(g_\ell, \mathcal{G}^{(n)}(t))$ but $g_{\ell'} \notin N_{2k}(g_\ell, \mathcal{G}^{(n)}(0))$, that is $g_{\ell'}$ entered into the neighborhood of g_ℓ at some time $t' < t$ and is still present there at time t.*

Lemma 4. *(the SI local lemma) A single transfer event results in a new violation in each of at most $4k+1$ of the $2k$-neighborhoods of genes in the sequence, and it decreases $\bar{\mathrm{SI}}$ by at most $\frac{6k}{2k(n-2k)}$, which is asymptotic to $3/n$ for constant k as $n \to \infty$.*

Proof. Let g_ℓ be the gene transferred to a position p between two other genes. Then, g_ℓ results in a single violation of at most $2k$ of the k-neighborhoods of the genes within distance k of p. In addition, the removal of g_ℓ results in a single violation of at most $2k$ of the k-neighborhoods of the genes that were in the k-neighborhood of g_ℓ before the transfer (since other genes now move into the extremes of this neighborhood). Finally, the k-neighborhood of g_ℓ itself can change completely in the transfer, which results in $2k$ violations of the k-neighborhood of g_ℓ. In summary, a maximum of $2k+2$ $k+1 = 4k+1$ k-neighborhoods undergo one (or more) violations, and the total number of violations is at most $2k \cdot 1 + 2k \cdot 1 + 1 \cdot 2k = 6k$. The second part of the lemma now follows from equations (10) and (11). ■

Acknowledgment

We acknowledge the support of the Israeli Science Foundation (ISF) and the VolkswagenStiftung grant, project VWZN3157, for funding GS and YL.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

References

Adato O, Ninyo N, Gophna U, Snir S. 2015. Detecting horizontal gene transfer between closely related taxa. *PLoS Comput Biol.* 11(10):e1004408.

Allen LJ. 2010. An introduction to stochastic processes with applications to biology. Chapman and Hall/CRC.

Bansal MS, Kellis M, Kordi M, Kundu S. 2018. Ranger-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics* 34(18):3214–3216.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe F-J, Dupré J, Dagan T, Boucher Y, et al. 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct.* 4(1):34.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A.* 102(40):14332–14337.

Bergeron A, Mixtacki J, Stoye J. 2006. A unifying view of genome rearrangements. In: International Workshop on Algorithms in Bioinformatics. Springer. p. 163–173.

Biller P, Guéguen L, Tannier E. 2015. Moments of genome evolution by double cut-and-join. *BMC Bioinformatics* 16(S14):S7.

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311(5765):1283–1287.

Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C. 2012. Alf a simulation framework for genome evolution. *Mol Biol Evol.* 29(4):1115–1123.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284(5423):2124–2128.

Doyon J-P, Scornavacca C, Gorbunov KY, Szöllősi GJ, Ranwez V, Berry V. 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: RECOMB International Workshop on Comparative Genomics. Springer. p. 93–108.
Durrett R. 2008. Probability models for DNA sequence evolution. Springer Science & Business Media.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*. 27(4):401–410.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.
Felsenstein J. 1993. Phylip (phylogeny inference package), version 3.5 c.
Grimmett G, Grimmett GR, Stirzaker D, et al. 2001. Probability and random processes. Oxford University Press.
Hannenhalli S, Pevzner PA. 1999. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J ACM*. 46:1–27.
Hendy MD, Penny D, Steel M. 1994. A discrete Fourier analysis for evolutionary trees. *Proc Natl Acad Sci U S A*. 91(8):3339–3343.
Huson DH, Steel M. 2004. Phylogenetic trees based on gene content. *Bioinformatics* 20(13):2044–2049.
Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, et al. 2007. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet*. 3(12):e231.
Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*. 55(1):709–742.
Martin W. 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* 21(2):99–104.
Nakhleh L, Ruths D, Wang L-S. 2005. Riata-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer. In: International Computing and Combinatorics Conference. p. 84–93.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.
Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, et al. 2012. eggnog v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res*. 40(D1):D284–D289.
Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol*. 12(1):66.
Puigbo P, Wolf YI, Koonin EV. 2010. The tree and net components of prokaryote evolution. *Genome Biol Evol*. 2:745–756.
Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459):431–437.
Sankoff D. 1992. Edit distance for genome comparison based on non-local operations. In: Annual Symposium on Combinatorial Pattern Matching. Springer. p. 121–135.
Sankoff D, Nadeau JH. 1996. Conserved synteny as a measure of genomic distance. *Discr Appl Math*. 71(1–3):247–257.
Semple C, Steel M, et al. 2003. Phylogenetics. Vol. 24. Oxford University Press on Demand.
Serdoz S, Egri-Nagy A, Sumner J, Holland BR, Jarvis PD, Tanaka MM, Francis AR. 2017. Maximum likelihood estimates of pairwise rearrangement distances. *J Theor Biol*. 423:31–40.
Sevillya G, Snir S. 2019. Synteny footprints provide clearer phylogenetic signal than sequence data for prokaryotic classification. *Mol Phylogenet Evol*. 136:128–137.
Shifman A, Ninyo N, Gophna U, Snir S. 2014. Phylo SI: a new genome-wide approach for prokaryotic phylogeny. *Nucleic Acids Res*. 42(4):2391–2404.
Sjöstrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B, Lagergren J. 2014. A Bayesian method for analyzing lateral gene transfer. *Syst Biol*. 63(3):409–420.
Smith JL, Campbell BJ, Hanson TE, Zhang CL, Cary SC. 2008. *Nautilia profundicola* sp. nov., a thermophilic, sulfur-reducing epsilonproteobacterium from deep-sea hydrothermal vents. *Int J Syst Evol Microbiol*. 58(7):1598–1602.
Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. *Nat Genet*. 21(1):108–110.
Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28(18):i409–i415.
Sumner JG, Jarvis PD, Francis AR. 2017. A representation-theoretic approach to the calculation of evolutionary distance in bacteria. *J Phys A Math Theor*. 50(33):335601.
Swingley WD, Blankenship RE, Raymond J. 2008. Integrating Markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. *Mol Biol Evol*. 25(4):643–654.
Szöllősi GJ, Tannier E, Lartillot N, Daubin V. 2013. Lateral gene transfer from the dead. *Syst Biol*. 62(3):386–397.
Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 29(1):22–28.
Tekaia F, Dujon B. 1999. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J Mol Evol*. 49(5):591–600.
Terauds V, Sumner J. 2019. Maximum likelihood estimates of rearrangement distance: implementing a representation-theoretic approach. *Bull Math Biol*. 81(2):535–567.
Thiergart T, Landan G, Martin WF. 2014. Concatenated alignments and the case of the disappearing tree. *BMC Evol Biol*. 14(1):266.
Wang L-S, Warnow T. 2001. Estimating true evolutionary distances between genomes. In: Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing. ACM. p. 637–646.
Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles E, Liou S-R, Boutin A, Hackett J, et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A*. 99(26):17020–17024.
Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. *Trends Genet*. 18(9):472–479.
Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol*. 1(1):8.
Woodhams M, Steane DA, Jones RC, Nicolle D, Moulton V, Holland BR. 2013. Novel distances for Dollo data. *Syst Biol*. 62(1):62–77.