Horizontal Gene Transfer Phylogenetics: A Random Walk Approach

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

[1] *Dept. of Evolutionary Biology, University of Haifa, Israel*

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

[3] *School of Mathematics and Statistics, University of Canterbury, NZ.*

[*] *Equal contribution*

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 resolution, the standard markers are too crude to provide a phylogenetic signal. Nevertheless, horizontal gene transfer (HGT) between organisms and gene loss provide far richer information. The *synteny index* (SI) between a pair of genomes combines gene order and gene content information, allowing comparison of unequal gene content genomes, together with order considerations of their common genes. Although this approach is useful for classifying close relatives, no rigorous statistical modelling for it has been suggested. Such modelling 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 synteny index measure. We model the *gene neighbourhood* 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. We apply this model to real biological data from the orthology database EggNOG and show that the average amount of HGT in bacterial genera is half the genome size. This result would seem difficult to achieve by other conventional approaches.

keywords: Gene Order, Horizontal Gene Transfer, Markovian Processes, Phylogenetics

1 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 astonishing 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. Such distances can come from a variety of sources, including measured distance, morphometric analysis or genetic distance derived from sequences, restriction fragments, or allozyme data. 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*.

Over the past few decades it has become apparent and accepted that statistical modelling, as opposed to parsimony (or combinatorial) approaches, is more accurate and hence is the preferred methodology. 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 simplest approach for inferring evolutionary distance based on genetic sequence is the Hamming distance [12], in which raw distance values can be calculated by simply counting the number of pairwise differences in character states. This approach does not take into consideration reversible mutations, so this approach is limited in its ability to estimate evolutionary distances and does not fit the additivity requirements of distance, as mentioned above. The Jukes–Cantor model of DNA substitution (JC69) [19] provides a simple mathematical correction to the Hamming distance that attempts to cope with the problem of unseen mutations and hence gives rise to distances that are additive for this simple model. Its simplicity, however, makes it less accurate with real biological data, since different mutations occur at different rates (both between different pairs of states and across sequence sites). Throughout the years, several refinements to the JC69 model have been proposed, considering rate differences and base frequencies. Among the most popular are Kimura 1980 [20], the F81 model [6], the HKY85 model [14] and the GTR model [32]. All these approaches, however accurate, are based on analysing a common ubiquitous gene, normally a housekeeping gene, shared by all the taxa under study. Notwithstanding, 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. Approaches taken to cope with this rely on the dynamics of prokaryotes' genes and are broadly divided into gene-order- and gene-content-based techniques. Under the order-based approach [24,13,36], 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 [29,33,9] ignores gene order entirely, and similarity is defined as the size of the set of shared genes. The *synteny index* (SI) [28,1] was suggested as an alternative method to the above techniques, allowing unequal gene content on one hand while accounting for the order among the shared genes.

Although a statistical framework has been devised for part of these models [26,34,3,25], to the best of our knowledge, no such framework has accounted for horizontal gene transfer (HGT). Such a model considers the sequence of events that may have led to the observed differences and selects the most likely (as opposed to the most parsimonious) explanation. This approach has acquired wide acceptance in the evolutionary community for its robustness and generality [15,16,17,5,6]. In this work, we provide for, the first time, such a model for the SI approach in which we model HGT events as a Markovian process. Based on this, we show that the gene neighbourhood 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 neighbourhood 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 crude SI [28,27]. We first demonstrate the accuracy of the model under a controlled setting in a simulation study. Applying this heuristic model to data from the orthology database EggNOG [23] yielded a set of 39 clusters of closely related taxa. Having a distance correction function at hand allows an approximate estimation of HGT activity inside these clusters, revealing that the average amount of HGT in bacterial genera is half the genome size.

2 Methods

Throughout the manuscript, for the sake of clarity, we will reserve the letter k to refer to the neighbourhood, 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 so . 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 (a) each of the possible $n-1$ positions between consecutive genes different from g_ℓ or at the start or end of the sequence and (b) with 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 $2k$–neighbourhood 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 $SI_\ell(t)$ as the relative intersection between $N_{2k}(g_\ell, \mathcal{G}^{(n)}(0))$ and $N_{2k}(g_\ell, \mathcal{G}^{(n)}(t))$ or formally

$$SI_\ell(t) = \frac{1}{2k}|N_{2k}(g_\ell, \mathcal{G}^{(n)}(0)) \cap N_{2k}(g_\ell, \mathcal{G}^{(n)}(t))| \tag{1}$$

(this is also called *the Jaccard index* between the two neighbourhoods [18]).
Let $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$ be the average of these $SI_\ell(t)$ values over all ℓ's between $k+1$ and $n-k$. That is,

$$\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)}) = \frac{1}{n-2k}\sum_{\ell=k+1}^{n-k} SI_\ell(t). \tag{2}$$

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 $\overline{SI}$ 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 1 *For an un-transferred 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 neighbourhood of g_ℓ at some time $t' < t$ and is still present there at time t.*

Lemma 1. *(the SI local lemma) A single transfer event results in a new violation in each of at most $4k+1$ of the $2k$-neighbourhoods of genes in the sequence, and it decreases $\overline{SI}$ 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–neighbourhoods of the genes within distance k of p. Also, the removal of g_ℓ results in a single violation of at most $2k$ of the k–neighbourhoods of the genes that were in the k–neighbourhood of g_ℓ before the transfer (since other genes now move into the extremes of this neighbourhood). Finally, the k-neighbourhood 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 $2k+2k+1 = 4k+1$ k–neighbourhoods 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 Eqns. (1) and (2). ■

3 Results

3.1 Asymptotic Estimation of Divergence Times

We now introduce a random process, that will play a key role in the analysis of the random variable $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$. 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 position closer to g_ℓ. The above can be modelled as a continuous-time random walk on state space $1, 2, 3, \ldots$ with transitions from j to $j+1$ at rate $j\lambda$ (for all $j \geq 1$) and from j to $j-1$ at rate $(j-1)\lambda$ (for all $j \geq 2$), with all other transition rates 0. This is thus a (generalised linear) birth-death process, and the process is illustrated in Fig. 1. 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 affect 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, i.e., it can be seen that $p_{ij}(t) \neq p_{(i+r)(j+r)}(t)$. Formally,

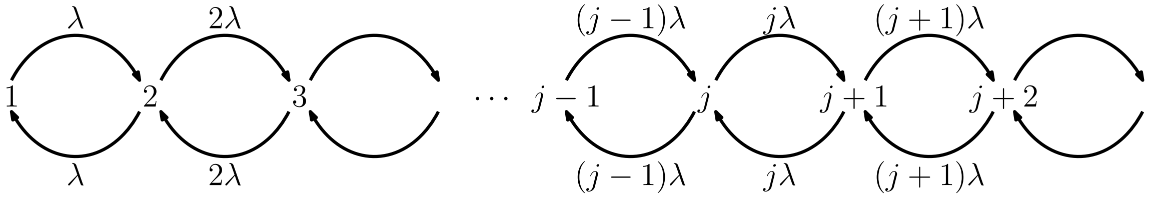


Fig. 1. Transitions for the process X_t

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

Lemma 2.

(a) The transition probabilities $p_{ij}(t)$ satisfy the following tri-diagonal differential system

$$\frac{1}{\lambda}\frac{dp_{ij}(t)}{dt} = -(2j-1)p_{ij}(t) + jp_{i(j+1)}(t) + (j-1)p_{i(j-1)}(t)$$

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 upwards.

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 Fig. 1. 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

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

Now the (tri-diagonal) system of differential equations for $\tilde{p}_{ij}$ is the well-known forward Kolmogorov differential equations (see supplementary text and for example Section 6.4.4. of [2]) and by translation these provide the equations in Part (a).

For Part (b) observe that:

$$\mathbb{E}[Y_t - \lambda t] = \mathbb{E}[Y_t] - \lambda t = \mathbb{E}[X_t - 1] - \lambda t = \mathbb{E}[X_t] - 1 - \lambda t. \tag{3}$$

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 Eqn. (3) gives $\mathbb{E}[X_t] = i + \lambda t$ as claimed.

That X_t has no stationary distribution follows from Theorem 6.1 of [2].

Part (c) is established using the Doob Martingale Inequality [10], which states that for a martingale process Z_t the following inequality holds for any $c > 0$:

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

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

$$\begin{aligned}
\mathbb{P}(\sup_{0\leq s\leq t} X_s > B) &= \mathbb{P}(\sup_{0\leq s\leq t} Y_s \geq B) && (4)\\
&\leq \mathbb{P}(\sup_{0\leq s\leq t} Y_s - \lambda t \geq B - \lambda t) && (5)\\
&\leq \mathbb{E}[Y_t - \lambda t]/(B - \lambda t) \qquad \text{(by Doob's inequality)} && (6)\\
&\leq \mathbb{E}[Y_0 - 0]/(B - \lambda t) \qquad \text{(by part b above)}, && (7)\\
& && (8)
\end{aligned}$$

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 "non jumping" gene stays in the k-neighbourhood 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,

$$q_{ik} = \sum_{j=1}^{k} p_{ij}(t). \tag{9}$$

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

$$q_k(t) := \frac{1}{k}\sum_{i=1}^{k} q_{ik}(t) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}(t). \tag{10}$$

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:*

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

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

Corollary 3 *Thus, if the function $t \mapsto \exp(-2\lambda t)q_k(t)$ has an inverse φ then*

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

.

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. We now return to prove Theorem 1.

Proof:

We first establish the following convergence in probability:

$$\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)}) \xrightarrow{p} \overline{\mu}_t, \tag{11}$$

where $\overline{\mu}_t$ is the expected value of $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$. We remark that this is necessary as the detailed analysis in the sequel, will pertain to the expected SI value of a single gene (denoted below μ_t). Nevertheless, we want to show that latter, converges tothe measured value $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$, as claimed in Eqn (19) below.

To establish Eqn. (11) let random variable N denote the total number of transfer events involving the n genes over the time period of duration t. Then N has a Poisson distribution with mean and variance equal to λnt and so N/n converges in probability to λt. Conditional on N, let $U_1, U_2, \ldots, U_N$ denote the actual sequence of transfer events that take place, regarding each of these as an arrow from n unlabelled ordered points on the line to the $n-1$ places that a transfer can be made to (this is illustrated in Fig. 2 for $n = 3$, where each of the six single transfers for U_i are indicated). These random variables are independent in this model, and $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$ is fully determined by $U_1, U_2, \ldots, U_N$. Moreover, if one of these transfer events – say U_i – were changed to a different transfer event (say U_i') while keeping the other U_* values the same, then, by the SI local lemma (Lemma 1) $\overline{SI}$ changes by at most α/n for a constant α.

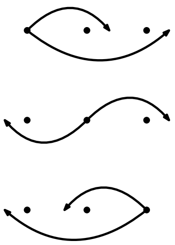


Fig. 2. The six possible values the random variable U_i can take when $n = 3$.

We now invoke the McDiarmid inequality [22], which states that if a random variable Y can be written as a function of N independent random variables, and if changing any one of those random variables (while holding the others fixed) never changes Y by more that c then

$$\mathbb{P}(|Y - \mathbb{E}[Y])| \geq \epsilon) \leq 2 \exp\left(\frac{2\epsilon^2}{Nc^2}\right).$$

Thus, conditional on N, the McDiarmid inequality (with $Y = \overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$, $\epsilon = n^{-\beta}$, and $c = \alpha/n$) implies that the probability that $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$ differs from its expected value $\overline{\mu}_t$ by more than $n^{-\beta}$ is bounded above by:

$$2\exp\left(-\frac{2n^{-2\beta}}{N(\alpha/n)^2}\right).$$

If we now select β strictly between 0 and $\frac{1}{2}$ and recall that N/n converges in probability to λt, and since $\lambda t = o(n)$, it follows that for all $\epsilon > 0$ the probability that $|\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)}) - \overline{\mu}_t| > \epsilon$ tends to zero as $n \to \infty$, in other words, $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)}) \xrightarrow{p} \overline{\mu}_t$, as claimed.

Next, let μ_t denote the expected value of $SI_\ell(t)$ for ℓ selected uniformly at random between k and $n-k$. We have:

$$\overline{\mu}_t = \mu_t, \tag{12}$$

by linearity of expectation. Also, for ℓ selected uniformly at random between k and $n-k$, let F_ℓ be the event that g_ℓ has not been transferred during the interval $[0,t]$. Under the assumptions of a Poisson process, we have:

$$\mathbb{P}(F_\ell) = \exp(-\lambda t). \tag{13}$$

Moreover, we have:

$$\mathbb{E}(SI_\ell(t)|\overline{F_\ell}) = o(1), \tag{14}$$

where $\overline{F_\ell}$ is the complementary event to F_ℓ. Essentially, Eqn. (14) says that if a gene in a large genome is transferred to a random position it is highly unlikely to have any overlap with the genes that it was previously within distance k of.

Next, let W_t be $SI_\ell(t)$, i.e. $\frac{1}{2k}$ times the number of the $2k$ genes (different from g_ℓ) at distance at most k from g_ℓ in $\mathcal{G}_0^{(n)}$ that also have distance at most k from g_ℓ in $\mathcal{G}_t^{(n)}$. We now assume g_ℓ has not moved (i.e., we consider an event from F_ℓ) and claim that:

$$\mathbb{E}[W_t|F_\ell] = \exp(-\lambda t)q_k(t) + o(1). \tag{15}$$

We now establish Eqn. (15). For $0 < i \leq k$, observe that $g_{\ell+i}$ is one of k genes to the right of g_ℓ at time 0. Let $\mathcal{E}_i(t)$ be the event that $g_{\ell+i}$ is within distance k of g_ℓ at time t, and let $F_{\ell+i}$ be the event that $g_{\ell+i}$ has not been transferred during the interval $[0,t]$. By the law of total probability,

$$\mathbb{P}(\mathcal{E}_i(t)) = \mathbb{P}(\mathcal{E}_i(t)|F_{\ell+i})\mathbb{P}(F_{\ell+i}) + \mathbb{P}(\mathcal{E}_i(t)|\overline{F_{\ell+i}})\mathbb{P}(\overline{F_{\ell+i}}). \tag{16}$$

Now, $\mathbb{P}(F_{\ell+i}) = \exp(-\lambda t)$ and $\mathbb{P}(\mathcal{E}_i(t)|\overline{F_{\ell+i}}) = o(1)$ and so, from Eqn. (16), we have:

$$\mathbb{P}(\mathcal{E}_i(t)) = \exp(-\lambda t) \cdot \mathbb{P}(\mathcal{E}_i(t)|F_{\ell+i}) + o(1). \tag{17}$$

We now calculate $\mathbb{P}(\mathcal{E}_i(t)|F_{\ell+i})$. Observe first that, conditional on events F_ℓ and $F_{\ell+i}$ holding, the gene $g_{\ell+i}$ is always to the right of g_ℓ during the interval $[0,t]$. Under the model, $g_{\ell+i}$ moves one step closer or further away from g_ℓ or stays where it is at any given time. More precisely, let $r = r(t')$ denote the distance that $g_{\ell+i}$ is to the right of g_ℓ at time $t' \leq t$. Then with probability $1-o(1)$, $g_{\ell+i}$ moves one step to the left at time t' (towards g_ℓ) if one of the $r-1$ genes between (but not including) g_ℓ and $g_{\ell+i}$ is transferred at time t', and this occurs at rate $(r-1)\lambda$. On the other hand, $g_{\ell+i}$ moves one step to the right at time t' whenever some gene in the genome is transferred into one of the r places available for insertion between g_ℓ and $g_{\ell+i}$ at time t'. Note that r can be larger than k; however, by Lemma 2(c), r is less than $\sqrt{n}$ with probability $1-o(1)$, and so for this second (right-move) case occurs with rate $\frac{r}{n} \times (n-r-1)\lambda = r\lambda + o(1)$.

Thus, the distance that $g_{\ell+i}$ is to the right of g_ℓ as time goes from 0 to t, behaves asymptotically (as $n \to \infty$) identically to the process X_t described above, conditioned on starting X_t at state i at time 0 (since $g_{\ell+i}$ has distance i to the right of ℓ at time 0). In other words, $\mathbb{P}(\mathcal{E}_i(t)|F_{\ell+i})$ is (asymptotically with n) the probability that $X_t \leq k$ conditional on $X_0 = i$, which is the quantity $q_{ik}(t)$ defined in Eqn. (9). Thus,

$$\mathbb{P}(\mathcal{E}_i(t)|F_{\ell+i}) = q_{ik}(t) + o(1),$$

and so, for each of the k genes to the right of g_ℓ in $\mathcal{G}_0^{(n)}$ (namely $g_\ell, g_{\ell+1}, \ldots, g_{\ell+k}$), Eqn. (17) gives:

$$\mathbb{P}(\mathcal{E}_i(t)) = \exp(-\lambda t) q_{ik}(t) + o(1).$$

An exactly analogous argument applies for the k genes to the left of g_ℓ (i.e., of the form $g_{\ell-i}$), leading to the same equation for $\mathbb{P}(\mathcal{E}_i(t))$. Hence, by linearity of expectation we obtain:

$$\mathbb{E}[W_t|F_\ell] = \frac{1}{2k} \sum_{-k \leq i \leq k} \mathbb{P}(\mathcal{E}_i(t)) = \exp(-\lambda t) q_k(t) + o(1),$$

where $q_k(t)$ is defined in Eqn. (10), thereby justifying Eqn. (15).

Next we show that

$$\mu_t = \exp(-2\lambda t) q_k(t) + o(1). \tag{18}$$

First, by the law of total expectation we have:

$$\mu_t = \mathbb{E}[SI_\ell(t)|F_\ell]\mathbb{P}(F_\ell) + \mathbb{E}[SI_\ell(t)|\overline{F_\ell}]\mathbb{P}(\overline{F_\ell}).$$

Next, if we apply Eqns. (13) and (14) to the first and second terms (respectively) on the right of this last equation we obtain:

$$\mu_t = \mathbb{E}[W_t|F_t] \exp(-\lambda t) + o(1)(1 - \exp(-\lambda t)),$$

and so, from Eqn. (15), we have:

$$\mu_t = \exp(-\lambda t) \exp(-\lambda t) q_k(t) + o(1) = \exp(-2\lambda t) q_k(t) + o(1),$$

as claimed. Combining the pieces from Eqns. (11) and (12) we have:

$$\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)}) \xrightarrow{p} \overline{\mu}_t = \mu_t. \tag{19}$$

Now $\mu_t = \exp(-2\lambda t)q_k(t) + o(1)$ (by Eqn. (18)) and so we obtain the required convergence in probability:

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

as $n \to \infty$.

This completes the proof of Theorem 1. ∎

3.2 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 neighbourhood 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 5000 genes and here many relaxations used above do not hold. Therefore, in order to analyse real data, we must find a realistic model that imitate 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 thereof the quantity $1 - SI$ that we denote d_{SI} to avoid confusion. Note that incontrary 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 3. *Under the uniform jump model, when $t \to \infty$, $d_{SI} = 1 - \frac{2k}{n-1}$.*

Proof: There are $2k$ genes in the original neighbourhood $N_{2k}(g, \mathcal{G}^{(n)}(0))$ 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. ∎

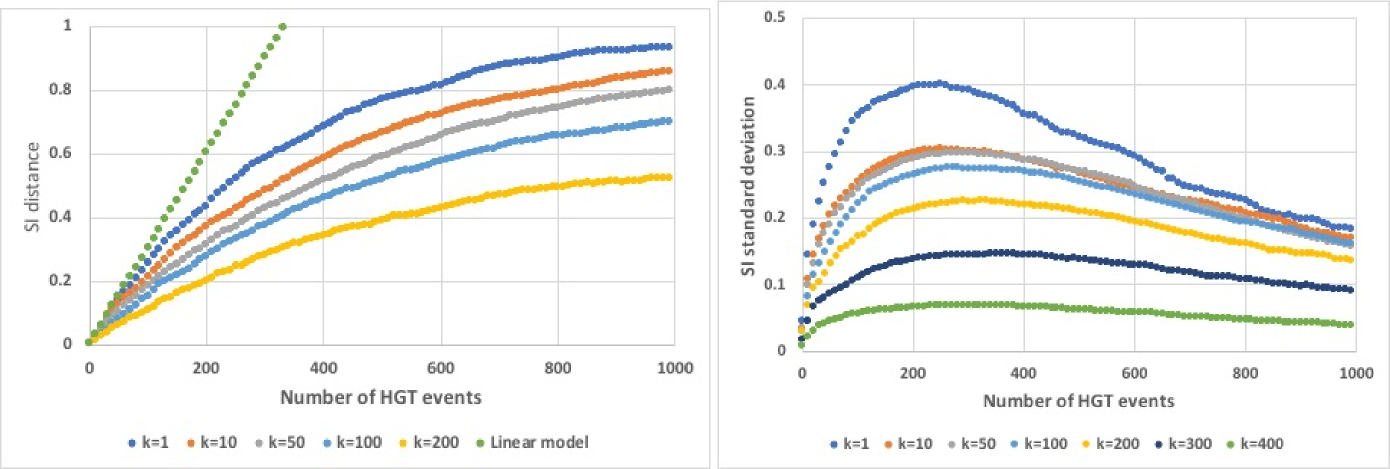
The Linear Model We start with a simple case that will serve as the basis for the subsequent development. We first define the following.

Definition 4 *The* disjoint events assumption *(DEA) assumes that a transferred gene g_ℓ leaves its original, unviolated neighbourhood and lands at a new, unviolated neighbourhood.*

In other words, under DEA, all neighbourhoods 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, i.e., between closely related species.

It is easy to see that under DEA, Lemma 1 – 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 non-disjoint events. As discussed above the goal here is not to find an exact model as in the asymptotic model of Section 3.1, rather to find a sound approximation to it. The first approach then is to obtain intuition via simulation study. Figure 3 depicts results of a simulation study between two genomes. Gur, this needs to move to suppl "At first, two identical genomes , G_1, G_2, were created. Then, iterations of HGT events were executed. In each iteration one gene was randomly chosen from G_2 and relocated to a random position in G_2. After each HGT event SI was calculated. Figure 3(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. Also, 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 (Lemma 1), this line is independent of k. Also, we can see that the maximum value of d_{SI} behaves according to Lemma 3. Figure 3(b) depicts the standard deviation of d_{SI} as a function of the number of HGT events. As shown the first few events cause a sharp increase in the standard deviation, then it starts to decrease in a logarithmic manner. The point where the standard deviation attains its maximum value has an interesting meaning which we discuss later.



(a) (b)

Fig. 3. Results of pairwise simulation between two genomes under realistic values: (a) d_{SI} as a function of number of HGT events. Simulations over 1000-gene genome sizes, under various k's ($k = 1, 10, 50, 100, 200$) (b) Standard deviation of d_{SI}.

Guided by the results depicted in Fig 3(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, i.e., $\frac{dB}{dt} = \Lambda B$, where B is the quantity measured, t is time, and $\Lambda < 0$ is the decay constant. Such a function behaves as $B_t = B_0 e^{\Lambda t}$ [4]. 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 $\mathbb{E}[d_{SI}]$. Finally, we determine how to set the actual parameters tracing the simulation best.

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. 3(a)).

Observation 5 *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 (non overlapping) $2k$-neighbourhoods. By Lemma 3, 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 neighbourhood 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 1 that a violation at a neighbourhood is a gene not originally from that neighbourhood.

Lemma 4. *If the neighbourhood 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 neighbourhood, 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 neighbours. Let us consider the next event, as some gene g_ℓ is jumping to a new neighbourhood. First, gene g_ℓ, as all other genes, is missing m old neighbours. That is, when making this jump, it loses its remaining $2k - m$ original neighbours, and also, these $2k - m$ genes are losing gene g_ℓ as their old neighbour. That is, its contribution to the d_{SI} score is $\frac{2(2k-m)}{2kn}$. Now, in the new neighbourhood of gene g_ℓ, there are $2k$ genes that are now having g_ℓ in their current $2k$-neighbourhood. Each of these $2k$ genes is already missing m original neighbours. For each of these genes, the probability that gene g_ℓ pushed out of the $2k$-neighbourhood an original neighbour 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 $2k$-neighbourhood is $\frac{2k-m}{2k} \cdot \frac{1}{2kn}$. So, for the new location of g_ℓ, the calculation of the expected contribution to d_{SI} is: The expected contribution for a single $2k$-neighbourhood times the number of affected $2k$-neighbourhoods,

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

Summing the contribution from the old and new $2k$-neighbourhoods 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

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

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

$$\Lambda = \frac{\frac{dB}{dt}}{B} = \frac{-\frac{3}{2kn}}{\frac{3}{n} - \frac{3m}{2kn}} \approx \frac{-\frac{3}{2kn}}{\frac{3}{n}} = -\frac{1}{2k}. \tag{20}$$

A similar procedure for the next time periods (i.e., for having expected violations 2, 3, etc.) 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, Table 1 provides a manual derivation for the initial steps of $1, 2, 3$ violations per neighbourhoods, demonstrating the constancy of decay rate $\Lambda = -\frac{1}{2k}$.

Time periods in units of n/(6k) HGT events	$E(SI)$	Next event SI contribution			$E(SI)'$ (N in the exponential decay terms)	$E(SI)''$ ($\frac{dN}{dt}$ in the exponential decay terms)
		Old neighborhood	new neighborhood	The gene itself		
0	0	$2k$	$2k$	$2k$	$\frac{3}{n}$	--
$\frac{n}{6k}$	$\frac{1}{2k}$	$2k-1$	$2k-1$	$2k$-1	$\frac{3}{n} - \frac{3}{2kn}$	$-\frac{3}{2kn}$
$\frac{2n}{6k}$	$\frac{2}{2k}$	$2k-2$	$2k-2$	$2k$-2	$\frac{3}{n} - \frac{6}{2kn}$	$-\frac{3}{2kn}$
$\frac{mn}{6k}$	$\frac{m}{2k}$	$2k$ −m	$2k$ −m	$2k$-m	$\frac{3}{n} - \frac{3m}{2kn}$	$-\frac{3}{2kn}$

Table 1. A manual calculation of the theoretical process for inferring the instantaneous increase at the first three "periods" . As can be seen there is a constant change in the contribution.

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 $\varLambda = -\frac{1}{2k}$ 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 $\varLambda$ by this factor, $\frac{n}{6k}$, yielding

$$\varLambda^* = \frac{\varLambda}{\frac{n}{6k}} = \frac{-\frac{1}{2k}}{\frac{n}{6k}} = -\frac{3}{n}. \tag{21}$$

Now we want to plug it into the exponential decay function: $B_t = B_0 e^{\varLambda^* t}$ ~~where t^* in our case is the number of events, in our case λt~~. 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:

$$\frac{d}{dt}\mathbb{E}[d_{SI}] = B_t = B_0 e^{\varLambda^* t} = \frac{3}{n} e^{-\frac{3}{n}t}. \tag{22}$$

In order to obtain $\mathbb{E}[d_{SI}]$ we need to integrate Eqn. (22). Specifically, we are interested in the definite integral in the interval $[0, t]$:

$$\int_0^t \frac{d}{dt}\mathbb{E}[d_{SI}] = \int_0^t \frac{3}{n} e^{-\frac{3}{n}t} = -e^{-\frac{3}{n}t}\Big|_0^t = 1 - e^{-\frac{3}{n}t}. \tag{23}$$

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. So replacing the generic t used in the development above with λt^* results in:

$$\mathbb{E}[d_{SI}] = 1 - e^{-\frac{3}{n}\lambda t^*}. \tag{24}$$

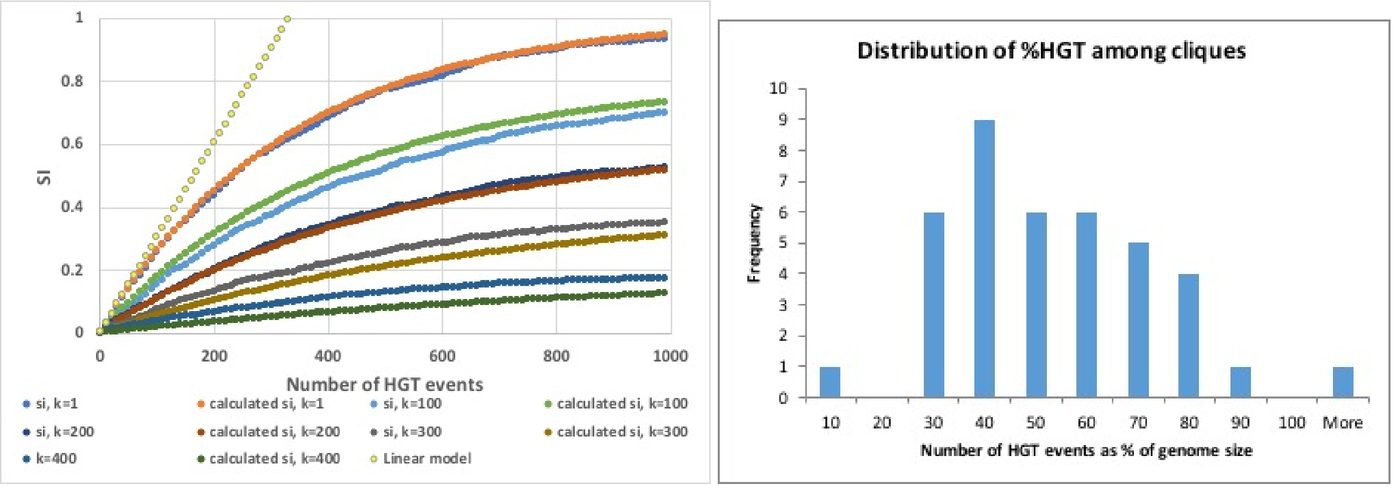
Now, as n is finite here, the latter tends to one as $t \to \infty$. However, recall that by Lemma 3 d_{SI} is bounded from above by $1 - \frac{2k}{n-1}$ so we treat this as a scaling factor for our decay function. Our final refinement concerns with cases of relatively large neighbourhoods (e.g., $k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

$$\mathbb{E}[SI] = \left(1 - \exp\left(-\frac{3 - \frac{5k}{n-1}}{n}\lambda t\right)\right)\left(1 - \frac{2k}{n-1}\right). \tag{25}$$

Although this study is not as rigorous as the asymptotic case, its strength is by considering practical values as found in nature. Moreover, Eq (25) is invertible hence allows us to infer the expected distance (number of HGT events along a time period) from a given SI between two genomes (see precise derivation in supplementary text). Indeed, in Figure 4(a) we show results from the same simulation study as described in Fig 3, however here we contrast the real, simulated HGT to the expected calculated HGTs under this expanded model, as obtained from Eq (25). Model (expected) vs real (simulated) number of HGTs for various values of k are shown. As can be seen, even for very large neighbourhood size k, reconstruction (of #HGTs) remains quite accurate and this is due to the refinement of incorporating k into the exponent.

3.3 Result on Real Microbial Data

Once we found a plausible model for relevant sizes, we aimed to use it to infer HGT activity in microbial data. We used the EggNOG v3.0 database [23] which is the largest unbiased orthology



(a)

(b)

Fig. 4. (a) **Results of pairwise simulation between two genomes**. Here we present the same simulation as in Figure 3, but we added the result of the theoretical SI calculated by our suggested model (eq(25)). As can be seen, although the neighbourhood size becomes a significant part of the genome size, performance is still fairly accurate. (b) **Distribution of %HGT (relative to genome size) among clusters**. In each cluster of closely related species we calculate the SI between each pair of species. Then we calculate the predicted number of HGT events with the practical model we developed (eq(25)), and averaged this value for each cluster. Here we present the distribution of this value among clusters.

database, containing protein sequences of 1133 species, most of them bacteria. In addition, this database clusters all proteins into Clusters of Orthologous Groups (COGs) [31], 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. We used the partition of taxa from [27] where in order to work only in the valid region (i.e. avoid “1” entries in the SI matrix), all entries above a certain threshold (specifically SI ≥ 0.95) were removed. In the resulting matrix,we selected the largest cliques (see more details in [27]). This yielded 39 clusters (subsets) of bacteria species, largely conforming with the conventional classification into genera (data not shown).

Details of this analysis are shown in the table in the supplementary text. Specifically, for each such cluster, we report its corresponding genus, its average SI (averaged over all pairs of genomes in that cluster) and the average number (in terms of the percentage of the average genome size (number of genes)) of HGT events “separating” each pair of species in that cluster. A succinct representation of this data is presented graphically in Fig 4(b). We find that this parameter, the average number of HGT as a percentage of average genome size, is normally distributed (Shapiro–Wilks test: p=0.238) [8], with a mean of 52.7%, a median of 54.1% and a SD of 23.78%. In other words, we found that the average number of HGT events between pairs of species inside a genus is about 50% (±20) of the genome size.

We believe this is an interesting finding that would not have been readily obtained by standard HGT methods, as constructing species trees within genera is not trivial. Moreover, the fact that the SI values themselves, as opposed to the measure reported here, are not normally distributed (Shapiro–Wilks test: p=0.024) intensifies the soundness of this finding.

Finally, as we provide a more accurate model–based approach to construct trees based on SI data, it is important to apply the new measure to the same real data the original SI-based approach was applied in [27]. In the supplementary text, we report the application of the two approaches to each of the clusters described above. Although we have no means to judge the correctness of the results, the small differences, as well as the results of the simulation study, suggest that both approaches perform satisfactorily. The trees and matrices are provided in the supplementary data.

4 Discussion

In this paper, we have provided a first statistical modeling for the synteny index (SI) as a phylogenetic marker. The major advantage of SI is that it combines both gene order and gene content evolutionary signals. The latter allows one to compare genomes over different gene sets (a pervasive phenomenon in prokaryotes). It also 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 [5] and his subsequent remedy to this flaw [6], maximum likelihood is now the gold standard in phylogenetics, and most popular packages [30,7,11] offer a likelihood based solution. These approaches rely on a single gene that is conserved among all taxa and hence is appropriate for comparison. Such genes however are too conserved and do not supply a strong enough signal to catch subtle signals.

In contrast, gene order– and content–based approaches were found to provide enough signal, allowing more resolved trees [35]. 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, although the usefulness of SI has been proved empirically in previous works [28,1], no analytical proof for its correctness has been shown. Therefore, this discussion emphasises the importance of the direction we took in this work, laying the groundwork for gene–based phylogenetics.

As this work has just handled the most basic case (the jump model) extensions are to be studied to arrive at more advanced models, such as *the indel model*, in which new genes are added to the genome, but at an equal rate of gene loss, so the genome sizes is approximately fixed. Another natural extension is to consider the transfer of clusters of genes, forming *genomic islands* [21]. Therefore, we believe that the tools developed here will serve as the basis for further extensions.

References

1. O. Adato, N. Ninyo, U. Gophna, and S. Snir. Detecting horizontal gene transfer between closely related taxa. *PLOS comp. Biol.*, 11:e1004408, 10 2015.
2. L.J.S. Allen. *An introduction to stochastic processes with applications to biology.* Boca Raton, FL : Chapman & Hall/CRC, 2nd ed edition, 2011.
3. P. Biller, L. Guéguen, and E. Tannier. Moments of genome evolution by double cut-and-join. *BMC Bioinformatics*, 16(14):S7, Oct 2015.
4. R. Durrett. *Probability Models for DNA Sequence Evolution*, chapter 3.
5. J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1978.
6. J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.
7. J. Felsenstein. Phylip - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.
8. A. Field. *Discovering Statistics using IBM SPSS Statistics.* Sage Publications Ltd, 4 edition, 2013.
9. F. Gibbon, T. Sorel, and H. Christopher. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27(21):4218–4222, 1999.
10. Geoffrey Grimmett and David Stirzaker. *Probability and random processes.* Oxford university press, 2001.
11. S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, 59(3):307–321, 2010.
12. R. W Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.
13. S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *STOC 1995*, pages 178–189, New York, NY, 1995. ACM Press.
14. M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22(2):160–174, 1985.
15. M. D. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Syst. Zool.*, 38:297–309, 1989.
16. M. D. Hendy and D. Penny. Spectral analysis of phylogenetic data. *J. Classif.*, 10:5–24, 1993.
17. M. D. Hendy, D. Penny, and M.A. Steel. Discrete fourier analysis for evolutionary trees. *Proc. Natl. Acad. Sci. USA.*, 91:3339–3343, 1994.
18. P. Jaccard. Etude comparative de la distribution florale dans une portion des alpe s et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37(142):547579, 1901.
19. T. Jukes and C. Cantor. Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132, 1969.
20. M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120, 1980.
21. K.S. Makarova, Y.I. Wolf, S. Snir, and E.V. Koonin. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *Journal of bacteriology*, 193(21):6039–6056, 2011.
22. C. McDiarmid. *On the method of bounded differences*, page 148188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.
23. S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L.J. Jensen, C.V. Mering, and P. Bork. Eggnog v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, 40:D284D289, 2012.
24. D. Sankoff. Edit distance for genome comparison based on non-local operations. In *CPM 1992*, volume 644 of *LNCS*, pages 121–135, Berlin, 1992. Springer Verlag.
25. D. Sankoff and J.H. Nadeau. Conserved synteny as a measure of genomic distance. *Discrete Applied Mathematics*, 71(1):247 – 257, 1996.
26. S. Serdoz, A. Egri-Nagy, J. Sumner, B. R. Holland, P. D. Jarvis, M. M. Tanaka, and A. R. Francis. Maximum likelihood estimates of pairwise rearrangement distances. *Journal of Theoretical Biology*, 423:31 – 40, 2017.
27. Gur Sevillya and Sagi Snir. Synteny footprints provide clearer phylogenetic signal than sequence data for prokayotic classification. *Molecular Phylogenetics and Evolution*, 2019.

28. A. Shifman, N. Ninyo, U. Gophna, and S. Snir. Phylo SI: a new genome-wide approach for prokaryotic phylogeny. *Nucleic Acids Res.*, 42(4):2391–404, 2014.
29. B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *Nat Genet*, 21(1):108–110, 1999.
30. D. L. Swofford. Paup*: Phylogenetic analysis using parsimony (*and other methods). 1981.
31. R. Tatusov, D. Natale, I. Garkavtsev, T. Tatusova, U. Shankavaram, B. Rao, B. Kiryutin, N. Fedorova M. Galperin, and E. Koonin. The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29:22–28, 2001.
32. S. Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.
33. F. Tekaia and B. Dujon. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *Journal of Molecular Evolution*, 49:591–600, 1999.
34. L. S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 637–646. ACM, 2001.
35. Y. Wolf, I. Rogozin, N. Grishin, R. Tatusov, and E. Koonin. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology*, 1(1):8, 2001.
36. S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.