Horizontal Gene Transfer Phylogenetics: A First Model-Based 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 deciphering the evolutionary history of strains of a species. Under this fine resolution, the standard markers are too crude to provide 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. While this approach was found useful for classifying close relatives, no rigorous statistical modelling for it was suggested. Such modelling is valuable as it allows observed measures to be transformed into estimates of time periods during evolution, yielding *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 neighborhood* as a *birth/death/immigration* process affected by the HGT activity over the genome, and analytically relate the HGT rate and time to expected SI. This model is asymptotic and so provides accurate results assuming infinite size genomes. Therefore, we also developed a heuristic model following an *exponential decay* function, accounting to real life values, that provided good performance 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 generas, 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. Leaves of that tree correspond to contemporary extant species and the tree edges (or branches) evolutionary relationship. Despite the astonishing advancement in the extraction of such molecular data, and of ever increasing quality, finding that 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 such sequences, restriction fragment or allozyme data. The tree sought should preserve the property that the length of the path between any two organisms at the its leaves, equals the inferred pairwise distance between these very organisms. When such a tree exists, these distances are called *additive* and so does the distance matrix storing them.

In the past few decades it became apparent and accepted that statistical modeling, as opposed to parsimony (or combinatorial) approaches, is the accurate and hence preferred method to follow. Consequently, vast efforts have been made, first to accurately model, and then for efficient inference from the model data.

Therefore, based on the discussion above, finding and demonstrating provable additivity of a distance measure, is an integral and central part of systematics.

The simplest approach for inferring evolutionary distance based on genetic sequence is the Hamming distance [11] 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 requirements of distance matrix mentioned above. The Jukes-Cantor model of DNA substitution (JC69) [17], is a simple mathematical correction to the Hamming distance which attempts to cope with the above problem of unseen mutations, and hence makes it additive based on that model. Its simplicity however, makes it less accurate with real biological data, since different mutations occur at different rates. Throughout the years, several refinements to the JC69 model were proposed, considering rate differences and base frequencies. Among the most popular are (Kimura 1980) [18], the F81 model (Felsenstein 1981) [6], the HKY85 model [12], and the GTR [30]. All these approaches, however accurate, are based on analysing a common ubiquitous gene, normally housekeeping, shared by all taxa under study. Notwithstanding, such a gene is highly conserved by definition and hence cannot provide a strong enough signal when sorting the shallow branches of the prokaryotic tree. Approaches to cope with this rely on the dynamics of prokaryotes genes and are majorly divided into gene-order- and gene-content- based techniques. Under the order-based-approach [21,32,23], 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 [27,31,9] ignores entirely gene order and similarity is defined as the size of the set of shared genes. The *synteny index* (SI)[26,1] was suggested as an alternative method to the above, allowing unequal gene content on one hand while accounting for the order among the shared genes.

Although statistical framework was devised for part of these models [25,33,3,24] to the best of our knowledge no such framework accounted for HGT. Such a model considers the sequence events that have led to the observed differences, and selects the most likely - as opposed to the shortest - explanation. This approach has acquired wide acceptance in the evolutionary community for its robustness and generality [13,14,15,5,6].

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 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 therefore 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 into account real life sizes, and relies on exponentially decaying functions. Applying this heuristic model to date from the orthology database EggNOG [22] reveals that the average amount of HGT in bacterial generas, is half the genome size.

2 Preliminaries

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 n is large enough allowing us to ignore the tips of $\mathcal{G}^{(n)}$ (or equivalently, $\mathcal{G}^{(n)}$ is cyclic and there are no tips). 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_i is independently subject to a Poisson process transfer event (at constant rate λ) in which g_i is moved to a different position in the sequence, with each of the possible $n-1$ positions (between consecutive genes different from g_i, or at the start or end of the sequence) and 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_i need not necessarily move to a position between two genes, it can also move the initial or the last position in the sequence.

Note that, by the definition of a Poisson process, the probability that g_i 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 $j \in k+1, \ldots, n-k$ the $2k$-neighborhood of gene g_j in a genome $\mathcal{G}^{(n)}$, $N_{2k}(g_j, \mathcal{G}^{(n)})$ is the set $2k$ genes (different from g_j) that have distance at most k from g_j in $\mathcal{G}^{(n)}$. We also define $SI_j(t)$ the relative intersection between $N_k(g_j, \mathcal{G}^{(n)}(0))$ and $N_k(g_j, \mathcal{G}^{(n)}(t))$ or formally $SI_j(t) = \frac{1}{2k} N_k(g_j, \mathcal{G}^{(n)}(0)) \cap N_k(g_j, \mathcal{G}^{(n)}(t))$ (this is also called *the Jaccard index* between the two neighborhoods [16])

Let $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$ be the average of these $SI_j(t)$ values over all j between $k+1$ and $n-k$. That is,

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

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 untransferred gene g_j, we define a* violation *as a gene g_i such that $g_i \in N_{2k}(g_j, \mathcal{G}^{(n)}(t))$ but $g_i \notin N_{2k}(g_j, \mathcal{G}^{(n)}(0))$, that is g_i entered to g_j's neighborhood at some time $t' < t$ and is still present there at time t.*

Lemma 1. *(the SI local lemma) A single transfer event affects at most $4k$ k-neighborhoods, and it increases $\overline{SI}$ by at most $\frac{6k}{2k(n-2k)}$ and for a constant k, is asymptotic to $3/n$.*

Proof. Let g_i be the transferred gene and for sake of simplicity we denote by $g_{i'}$ the new location of g_i. First, g_i causes a violation in at most $2k$ neighborhood with equality in the case it had not moved before. Next, $g_{i'}$ violates at most $2k$ neighborhoods with one violation, with equality when it caused a new violation in all these neighborhoods. Finally, $g_{i'}$ has at most $2k$ new neighbors. Summing these three contributions and dividing by the total number of neighborhoods in $\mathcal{G}^{(n)}$ yields $\frac{6k}{2k(n-2k)}$. □

3 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_i, not being transferred during time period t, with respect to another gene $g_{i'}$. W.l.g. assume $i > i'$ and let $j = i - i'$. Now, there are j "slots" between $g_{i'}$ and g_i 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_j one position away from g_i, and transfer from that interval, moves g_j closer to g_i. 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 (generalized linear) birth-death process, and the process is illustrated in Fig. 3 in the Appendix As the process is not affected by the specific values of i and j (rather by their difference), we can ignore them and let X_t denote the random variable that describes the state of this random walk (a number $1, 2, 3$ etc) at time t.

The process X_t is slightly different than the much-studied critical linear birth-death process, for which the rate 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 (here there are no absorbing states). 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. 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 time 0 it was at state i. Note that here i and j cannot be ignored as they don't 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,

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 $i/(B - \lambda t) \to 0$, as $B \to \infty$.

Proof. Consider a critical linear-birth death process with immigration Y_t, in which the birth rate and death rate are both equal to λ, and the immigration rate is also equal to λ. 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. 3. Thus, if we let $\tilde{p}_{ij} := \mathbb{P}(Y_t = j | Y_0 = i)$.

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

and so

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

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

For Part (b) observe that:

$$\mathbb{E}[Y_t - \lambda t] = \mathbb{E}[Y_t] - \lambda t = \mathbb{E}[X_t - 1] - \lambda t = \mathbb{E}[X_t] - 1 - \lambda t. \tag{2}$$

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. (2) 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) follows by applying the Doob Martingale inequality to the Martingale process $Y_t - \lambda t$. □

We now set to calculate the probability that a "non jumping" 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,

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

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

$$q_k(t) := \frac{1}{k} \sum_{i=1}^{k} q_{ik}(t) = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}(t). \tag{4}$$

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 $\lambda t << n$) one can use the expression on the left to estimate (an additive) evolutionary distance and hence construct a tree consistently.

Proof. We first apply the McDiarmid inequality [20] to establish that

$$\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)}) \xrightarrow{p} \overline{\mu}_t, \tag{5}$$

where $\overline{\mu}_t$ is the expected value of $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_t^{(n)})$.

To establish Eqn. (5) 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 to be made to (this is illustrated in Fig. 4 in the Appendix 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 others U_* values the same, then, by the SI local lemma, SI_j changes for $O(k)$ values of j (and by a maximum of $6k$ for each such value) and $\overline{SI}$ changes by at most $3/n$. Thus, conditional on N, the McDiarmid inequality 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:

$$2 \exp\left(-\frac{2n^{-2\beta}}{N(3/n)^2}\right).$$

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 $SI_j(t)$ for j selected uniformly at random between k and $n-k$. We have:

$$\overline{\mu}_t = \mu_t, \tag{6}$$

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

$$\mathbb{P}(F_j) = \exp(-\lambda t). \tag{7}$$

Moreover, we have:

$$\mathbb{E}(SI_j(t)|\overline{F_t}) = o(1), \tag{8}$$

where $\overline{F_j}$ is the complementary event to F_j. Essentially, Eqn. (8) 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 $SI_j(t)$, i.e. $\frac{1}{2k}$ times the number of the $2k$ genes (different from g_j) at distance at most k from g_j in $\mathcal{G}_0^{(n)}$ that also have distance at most k from g_j in $\mathcal{G}_t^{(n)}$. We now assume g_j has not moved (i.e. F_j), and claim that:

$$\mathbb{E}[W_t|F_j] = \exp(-\lambda t) q_k(t) + o(1). \tag{9}$$

We now establish Eqn. (9). For $0 < i \leq k$, observe that g_{j+i} is one of k genes to the right of g_j at time 0. Let $\mathcal{E}_i(t)$ be the event that g_{j+i} is within distance k of g_j at time t, and let F_{j+i} be the event that g_{j+i} has not been transferred during the interval $[0, t]$. By the law of total probability,

$$\mathbb{P}(\mathcal{E}_i(t)) = \mathbb{P}(\mathcal{E}_i(t)|F_{j+i})\mathbb{P}(F_{j+i}) + \mathbb{P}(\mathcal{E}_i(t)|\overline{F_{j+i}})\mathbb{P}(\overline{F_{j+i}}). \tag{10}$$

Now, $\mathbb{P}(F_{j+i}) = \exp(-\lambda t)$ and $\mathbb{P}(\mathcal{E}_i(t)|\overline{F_{j+i}}) = o(1)$ and so, from Eqn. (10), we have:

$$\mathbb{P}(\mathcal{E}_i(t)) = \exp(-\lambda t) \cdot \mathbb{P}(\mathcal{E}_i(t)|F_{j+i}) + o(1). \tag{11}$$

We now calculate $\mathbb{P}(\mathcal{E}_i(t)|F_{j+i})$. Observe first that, conditional on events F_j and F_{j+i} holding, the gene g_{j+i} is always to the right of g_j during the interval $[0,t]$. Under the model, g_{j+i} moves one step closer or further from g_j or stays where it is at any given time. More precisely, let $r = r(t')$ denote the distance that g_{j+i} is to the right of g_j at time $t' \leq t$. Then with probability $1 - o(1)$, g_{j+i} moves one step to the left at time t' (towards g_j) if one of the $r-1$ genes between (but not including) g_j and g_{j+i} is transferred at time t', and this occurs at rate $(r-1)\lambda$. On the other hand, g_{j+i} moves one step to the right at time t' whenever some gene in the genome is transferred into one of the r places that available for insertion between g_j and g_{j+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_{j+i} is to the right of g_j 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_{j+i} has distance i to the right of j at time 0). In other words, $\mathbb{P}(\mathcal{E}_i(t)|F_{j+i})$ 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. (3). Thus,

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

and so, for each of the k genes to the right of g_j in $\mathcal{G}_0^{(n)}$ (namely $g_j, g_{j+1}, \ldots, g_{j+k}$), Eqn. (11) 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_j (i.e. of the form of the form g_{j-i}), leading to the same Equation for $\mathbb{P}(\mathcal{E}_i(t))$. Hence, by linearity of expectation we obtain:

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

where $q_k(t)$ is defined in Eqn. (4), thereby justifying Eqn. (9).

Next we show that

$$\mu_t = \exp(-2\lambda t) q_k(t) + o(1). \tag{12}$$

First, by the law of total expectation we have:

$$\mu_t = \mathbb{E}[SI_j(t)|F_t]\mathbb{P}(F_j) + \mathbb{E}[SI_j(t)|\overline{F_t}]\mathbb{P}(\overline{F_j}).$$

Next, if we apply Eqn. (7) and Eqn. (8) 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. (9), 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. (5), (6) we have:

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

Now $\mu_t = \exp(-2\lambda t) q_k(t) + o(1)$ (by Eqn. (12)) 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.

4 Analysis under Real Life 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 - $2k$ where k is constant or $o(n)$ - vanishes. However in real life bacterial genomes are of 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.
Nevertheless we start with some basic observations that are relevant to this part for the different settings than before. The following basic observation is given without proof.

Observation 4 *Under the uniform jump model, for genes g_i, g_j such that $g_i \in N_{2k}(g_j, \mathcal{G}^{(n)}(0))$,*

$$SI = \mathbb{P}(g_i \in N_{2k}(g_j, \mathcal{G}^{(n)}(t)).$$

The next simple lemma gives an upper bound on 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 $t \to \infty$ $SI = 1 - \frac{2k}{n-1}$*

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

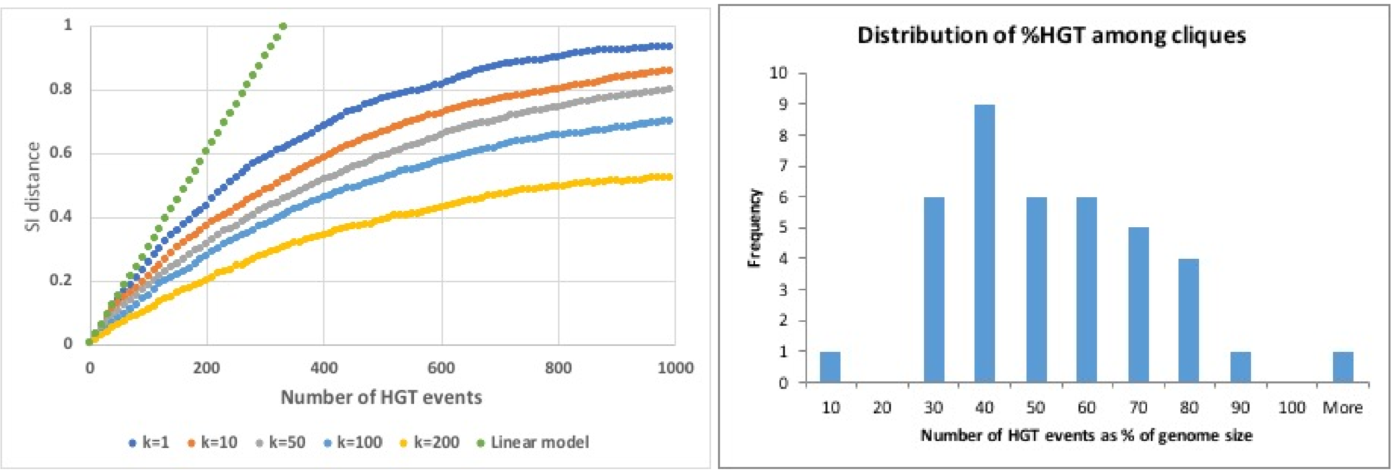
The Linear Model. We start with a simple case that will serve as the basis for the subsequent development. We first define the following.

Definition 5 *The* disjoint events assumption *(DEA) assumes that a transferred gene g_i 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, 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 SI is approximately $\frac{3}{n}$. Hence, under DEA, for relatively small number of HGT events N, the expected SI is $\frac{3N}{n}$.

The Expanded Model. As the DEA, and hence linearity of 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, rather to find a sound approximation to it. The first approach then is to obtain intuition via simulation study. To this end, 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, the SI distance between the two genomes was calculated. Figure 1(a) illustrates the outcome of our simulation by showing SI as a function of the number events, for various k's. 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 SI behaves



(a) (b)

Fig. 1. (a) **Results of pairwise simulation between two genomes:** SI as a function of number of HGT events. G_0 and G_1 were created identically with 1000 genes. Next, iteratively, a gene g_i was randomly chosen from G_1 and relocated to a random position in G_1. After each HGT event $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_1^{(n)})$ was calculated and is displayed as a function of number HGT events, for various sizes of k. Also, the theoretical linear model is presented in green. (b) **Distribution of %HGT (relative to genome size) among cliques**. In each clique 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 (Eqn. (16)), and averaged this value for each clique. Here we present the distribution of this value among cliques.

according to Lemma 3. In Figure 2(a) of Appendix B.1, we also present a pairwise simulation, but here we present the standard deviation (SD) of SI as a function of the number of HGT events. Here we can see that the first few events cause a rapid increase in SD, then SD starts to decrease in a logarithmic manner. The place where SD comes to its maximum value has interesting meaning which will be discussed later.

As can be perceived from Figure 1(a), all curves follow a logarithmic shape implying that the increase follows an *exponential decay*. A quantity is subjected to exponential decay (or growth) if it proceeds at a rate proportional to its current value, i.e., $\frac{dN}{dt} = \Lambda N$, where N is the quantity measured, t is time, and $\Lambda < 0$ is the decay constant [4]. Our goal is to have an expression for the expected change in SI after m HGT events. This is important since by integrating this expression we get the expected value of SI, $\mathbb{E}[SI]$, resulting from HGT events. We will denote this target expression as $\frac{d}{dt}\mathbb{E}[SI]$, since this is the derivative of $\mathbb{E}[SI]$. A further goal here is to develop more of an intuition and determine how to set the actual parameters tracing the simulation best.
We start by noting that the HGT events are distributed uniformly throughout the genome. Considering this, we start by finding, how many events are required for each gene g_i to have an SI score of $\frac{1}{2k}$ (and hence total SI of $\frac{1}{2kn}$). To tackle this question, we assume most of the events conform with DEA, as the events are uniformly distributed and genomes are relatively close - the process is at the beginning (see Figure 1(a)).

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

Proof. We can look at the genome as a sequence of $\frac{n}{2k}$ adjacent $2k$-neighborhoods. By Lemma 3, under DEA each event contributes $\frac{6k}{2kn} = 3/n$ to the total 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 $\frac{n}{6k}\frac{3}{n} = \frac{1}{2k}$ contribution to the total SI, and the lemma follows. □

Lemma 4. *If each gene contains m violations, the addition to the SI score resulting from the next event is $\frac{6k-3m}{2kn}$.*

A proof can be found in Appendix A.2.

We found that the 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 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 N being measured is $\frac{d}{dt}\mathbb{E}(SI)$, so the expression $\frac{dN}{dt}$ that changes over time (or HGT events), is $\frac{d^2}{dt^2}\mathbb{E}(SI)$. We found that for time period of $\frac{n}{6k}$ events, $\frac{d^2}{dt^2}\mathbb{E}(SI) = -\frac{3}{2kn}$ and this is $\frac{dN}{dt}$ for this time period. Now, recall that under exponential decay, $\frac{dN}{dt} = \Lambda N$ so in order to find Λ we write

$$\Lambda = \frac{\frac{dN}{dt}}{N} = \frac{-\frac{3}{2kn}}{\frac{3}{n} - \frac{3m}{2kn}} \approx \frac{-\frac{3}{2kn}}{\frac{3}{n}} = -\frac{1}{2k}. \tag{13}$$

A similar procedure for the next time periods (i.e. for having expected violations 2, 3, etc. will yield, as long as $\frac{3}{n} >> -\frac{3m}{2kn}$, the same $\Lambda = -\frac{1}{2k}$, as indeed required by such growth (exponential). As this derivation is involved, the table in Figure 5 (see Appendix) provides manual derivation for the initial steps of $1, 2, 3$ violations per neighborhoods, demonstrating the constancy of decay rate $\Lambda = -\frac{1}{2k}$. Of course this analysis is quite 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,

$$\Lambda^* = \frac{\Lambda}{\frac{n}{6k}} = \frac{-\frac{1}{2k}}{\frac{n}{6k}} = -\frac{3}{n}. \tag{14}$$

Now we can plug it into the exponential decay function: $N_t = N_0 e^{\Lambda^* t^*}$ where t^* in our case is the number of events, in our case λt.

$$\frac{d}{dt}\mathbb{E}[SI] = \frac{3}{n} e^{-\frac{3}{n}\lambda t}. \tag{15}$$

In order to obtain $\mathbb{E}[SI]$ we need to integrate Eqn. (15), yielding $\mathbb{E}[SI] = 1 - e^{-\frac{3}{n}\lambda t}$. Now, as n is finite here, the latter tends to one as $t \to \infty$. However, recall that by Lemma 3 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 neighborhoods (e.g. $k = 100, 200, 300, 400$) taking in consideration the case the gene transferred into its original neighborhood, as was shown above to be $\frac{3-\frac{5k}{n-1}}{n}$ instead of $3/n$. Therefore we obtain:

$$\mathbb{E}[SI] = \left(1 - \exp\left(-\frac{3 - \frac{5k}{n-1}}{n}\lambda t\right)\right)\left(1 - \frac{2k}{n-1}\right). \tag{16}$$

Although this study is not as rigorous as the asymptotic case, its strength is by considering practical values as found in nature. Indeed in Figure 2(b) we show the performance of this model compare to simulated data for various values of k. As can be seen, even for very large neighborhood size (k), prediction (of #HGTs) remains quite accurate and this is due to the refinement of incorporating k into the exponent.

4.1 Result on Real Microbial Data

Once we found a plausible model for relevant sizes, we aimed at using it to infer HGT activity in microbial data. We used the EggNOG database [22] that is the largest, unbiased, orthology database, containing protein sequences of 1133 species, most of them bacteria. In addition, this database clusters all proteins into COGs (Clusters of Orthologous Groups) [29]. This means that an organism is represented as a list of COG names ordered by their order of appearance in its genome. In order to work only in the valid region, i.e., avoid "1" entries in our matrix, we removed from the matrix all entries above a threshold (specifically SI ≥ 0.95) and in the resulted matrix searched for largest cliques. This has yielded 39 cliques (subsets) or bacteria, conforming with the conventional partition into *genera* (data not shown).

Figure 1(b) gives an overview on the distribution of the relative number of HGTs among the cliques. For a detailed account, we provide a table in Appendix B.4 (see Figure 6) which lists for each such clique, its corresponding genus, its average SI, and the average number (in terms of percentage of the average genome size (# genes)) of HGT events separating between each pair of species in that clique. We found that this parameter is normally distributed (Shapiro-Wilks test: $p = 0.238$) [8] with mean of 52.7%, median of 54.1% and SD of 23.78%. In other words, we found that the average number of HGT events between pairs of species inside a genera is about 50% (±20) of the genome size. We believe this is an interesting finding that cannot be readily obtained by standard HGT methods as constructing species trees within genera is not trivial. Moreover, the fact that the SI values themselves are not normally distributed (Shapiro-Wilks test: $p = 0.024$) intensifies the soundness of this finding.

5 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 not only a comparison between genomes over different gene sets, a pervasive phenomenon in prokaryotes, rather comparing the order of their shared, core, gene set, a signal that is ignored by 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 a gold standard also in phylogenetics and most popular packages [28,7,10] 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 to conserved and do not furnish a strong enough signal.

In contrast, gene-based (gene order and content) approaches were found to provide enough signal allowing more resolved trees [34]. Nevertheless, although such approaches exist for several decades already, to the best of our knowledge, no detailed model, showing additivity and consistency under HGT, was proposed. Similarly, while the usefulness of SI was proved empirically in previous works [26,1], no analytical proof for it correctness was shown. Therefore, the above discussion emphasises the importance of the direction we took in this work – laying the groundwork for gene based models phylogenetics.

As this work has just handled the most basic case – the jump model – extensions to 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 genome sizes is approximately fixed. Another natural extension is considering transfer of cluster of genes, forming *genomic islands* [19]. Therefore we believe that the tools developed here will serve as the basis for such further extensions.

References

1. O. Adato, N. Ninyo, U. Gophna, and S. Snir. Detecting horizontal gene transfer between closely related taxa. *PLOS Comput Biol*, 11:e1004408, 2015.
2. L. Allen. *An introduction to stochastic processes with applications to biology.* Boca Raton, FL : Chapman & Hall/CRC, 2nd edition, 2011.
3. P. Biller, L. Guéguen, and E. Tannier. Moments of genome evolution by Double Cut-and-Join. *BMC Bioinformatics*, 16(Suppl 14):S7, 2015.
4. R. Durrett. *Probability Models for DNA Sequence Evolution*, chapter 3, page 67. Probability and Its Applications. Springer New York, 2008.
5. J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*, 27(4):401–410, 1978.
6. J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 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 Res*, 27(21):4218–4222, 1999.
10. 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. *Syst Biol*, 59(3):307–321, 2010.
11. R. W. Hamming. Error detecting and error correcting codes. *Bell Syst Tech J*, 29(2):147–160, 1950.
12. M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.
13. M. D. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Syst Zool*, 38:297–309, 1989.
14. M. D. Hendy and D. Penny. Spectral analysis of phylogenetic data. *J Classif*, 10:5–24, 1993.
15. M. D. Hendy, D. Penny, and M. Steel. Discrete fourier analysis for evolutionary trees. *P Natl Acad Sci USA*, 91:3339–3343, 1994.
16. P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(142):547–579, 1901.
17. T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132, 1969.
18. M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, 1980.
19. K. Makarova, Y. Wolf, S. Snir, and E. Koonin. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J Bacteriol*, 193(21):6039–6056, 2011.
20. C. McDiarmid. *On the method of bounded differences*, pages 148–188. Long Math S. Cambridge University Press, 1989.
21. B. Moret, S. Wyman, D. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc of PSB*, volume 213, pages 583–94, 2001.
22. S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. Jensen, C. Mering, and P. Bork. EggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res*, 40:284–289, 2012.
23. D. Sankoff and N. El-Mabrouk. Genome rearrangement. *Current Topics in Computational Biology*, pages 135–155, 2002.
24. D. Sankoff and J. Nadeau. Conserved synteny as a measure of genomic distance. *Discrete Appl Math*, 71(1):247–257, 1996.
25. 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. *J Theor Biol*, 423:31–40, 2017.
26. 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.
27. B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *Nat Genet*, 21(1):108–110, 1999.
28. D. L. Swofford. *PAUP*: Phylogenetic analysis using parsimony (and other methods)*. Sinauer Associates, 2002.
29. R. Tatusov, D. Natale, I. Garkavtsev, T. Tatusova, U. Shankavaram, B. Rao, B. Kiryutin, M. Galperin, N. Fedorova, and E. Koonin. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29:22–28, 2001.
30. S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures Math Life Sci*, 17(2):57–86, 1986.
31. F. Tekaia and B. Dujon. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J Mol Evol*, 49:591–600, 1999.

32. G. Tesler. GRIMM: genome rearrangements web server. *Bioinformatics*, 18(3):492–493, 2002.
33. L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proc of STOC*, pages 637–646. ACM, 2001.
34. Y. Wolf, I. Rogozin, N. Grishin, R. Tatusov, and E. Koonin. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol*, 1(1):8, 2001.

A Optional extra material

In this part we provide farther information and proofs to claims remained outside due to space considerations.

A.1 Describing $q_k(t)$

Recall that $q_{ik}(t)$ is the conditional probability $\mathbb{P}(X_t \in [k] | X_0 = i)$ (where $[k] = \{1, 2, \ldots, k\}$, and $p_{ij}(t)$ is the probability that the random walk X_t (described above) is in state j after duration t, conditional on $X_0 = i$. Thus, $q_{ik}(t) = \sum_{j=1}^{k} p_{ij}(t)$, and so:

$$\frac{dq_{ik}(t)}{dt} = \sum_{j=1}^{k} \frac{dp_{ij}(t)}{dt}.$$

Lemma 5.

$$\frac{dq_{ik}(t)}{dt} = k\lambda[p_{i(k+1)}(t) - p_{ik}(t)]. \tag{17}$$

Proof. Lemma 2(a) gives:

$$\frac{1}{\lambda}\frac{dq_{ik}(t)}{dt} = \sum_{j=1}^{k} -(2j-1)p_{ij}(t) + jp_{i(j+1)}(t) + (j-1)p_{i(j-1)}(t).$$

and we show that the term on the right is simply $k[p_{i(k+1)}(t) - p_{ik}(t)]$ by induction on k. The base case $k = 1$ gives $\frac{1}{\lambda}\frac{dq_{ik}(t)}{dt} = p_{i(k+1)}(t) - p_{ik}(t)$ as claimed. For the induction step, observe that:

$$\sum_{j=1}^{k+1} \frac{dp_{ij}(t)}{dt} = \sum_{j=1}^{k} \frac{dp_{ij}(t)}{dt} + \frac{dp_{i(k+1)}(t)}{dt},$$

and by the inductive hypothesis the first term equals $k\lambda[p_{i(k+1)}(t) - p_{ik}(t)]$ while by Lemma 2(a)) the second term equals λ times $-(2k+1)p_{i(k+1)}(t) + (k+1)p_{i(k+2)}(t) + kp_{ik}(t)$. Adding these two terms together gives $(k+1)\lambda[p_{i(k+2)}(t) - p_{i(k+1)}(t)]$, which establishes the induction step, thereby justifying Eqn. (17). □

Recall (from Eqn. (4)) that $q_k(t) = \frac{1}{k}\sum_{i=1}^{k} q_{ik}(t)$. It now follows that:

$$\frac{dq_k(t)}{dt} = \frac{1}{k}\sum_{i=1}^{k} \frac{dq_{ik}(t)}{dt} = \lambda \sum_{i=1}^{k} [p_{i(k+1)}(t) - p_{ik}(t)].$$

In particular,

$$\frac{d}{dt}\exp(-2\lambda t)q_k(t) = \lambda \exp(-2\lambda t)\left[-2q_k(t) + \sum_{i=1}^{k} [p_{i(k+1)}(t) - p_{ik}(t)]\right],$$

and so the claim that $\exp(-2\lambda t)q_k(t)$ is monotone decreasing (and so the function φ is well-defined) amounts to verifying that

$$\sum_{i=1}^{k}\left(p_{i(k+1)}(t) - p_{ik}(t) - 2\left(\sum_{j=1}^{k} p_{ij}(t)\right)\right) < 0. \tag{18}$$

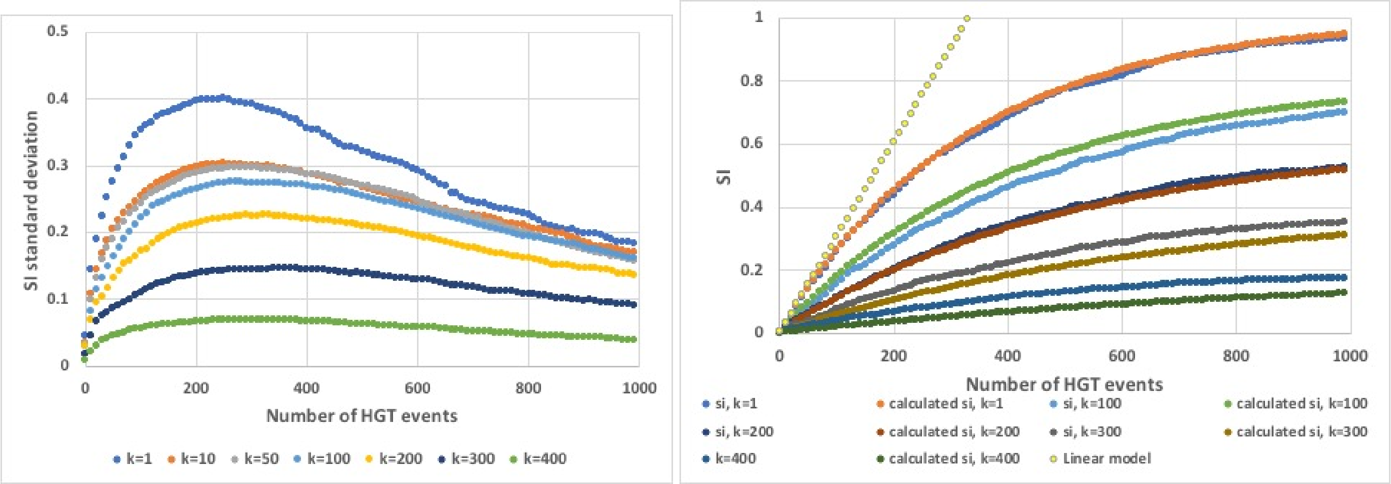
A.2 Proof of Lemma 4

Proof. We will show this holds for any $m < 2k$. Having m violation in each gene's neighborhood, means that each gene has 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_i is jumping to a new neighborhood. First, gene g_i, as all other genes, is missing m old neighbors. That is, when making this jump, it loses its remaining $2k - m$ neighbors, and also, these $2k - m$ genes are losing gene g_i as their old neighbor. That is, $\frac{2(2k-m)}{2kn}$ contributions to the SI score. Now, in the new neighborhood of gene g_i, there are $2k$ genes that are now having g_i in their current neighborhood. Each of these genes is already missing m old neighbors. For each of these genes, the probability that gene g_i pushed out of the neighborhood an old neighbor is $\frac{2k-m}{2k}$. When this is the case, there is a contribution of $\frac{1}{2kn}$ to the SI score. So, for the new neighborhood contribution calculation we have: the number of potential genes in the new neighborhood times the probability they lose a neighbor $\frac{2k-m)}{2k}$ times the change in the SI from this loss $\frac{1}{2kn}$. That is, $\frac{2k(2k-m)}{2k}\frac{1}{2kn} = \frac{2k-m}{2kn}$. Summing the contribution from the old and new neighborhoods and the jumping gene g_i we end with $\frac{3(2k-m)}{2kn} = \frac{6k-3m}{2kn}$ contribution to the total SI score. □

B Auxiliary Data and Figures

In this section of the Appendix we provide further details and illustrations to summaries provided in the main text.

B.1 Further results of pairwise simulation between two genomes



(a)

(b)

Fig. 2. (a) Standard deviation (SD) of $\overline{SI}(\mathcal{G}_0^{(n)}, \mathcal{G}_1^{(n)})$ is displayed as a function of number of HGT events, with different size of k. We can see that the first few events cause a rapid increase in SD, then SD decreases in a logarithmic manner. (b) Here we present the same simulation as in Figure 1(a), but we added the result of the theoretical SI calculated by our suggested model (Eqn. (16)). As can be seen, although the neighborhood size becomes a significant part of the genome size, performance is still fairly accurate.

B.2 The Gene Neighborhood as a Markov Chain

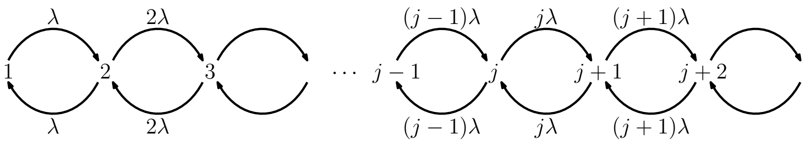


Fig. 3. Transitions for the process X_t

B.3 The State Space of the Random Walk

B.4 The increase in contribution to SI

We here illustrate in a table form the change in the contribution to SI at each iteration. As can be seen there is a constant change in the contribution. The table in Fig 5 summarized this finding.

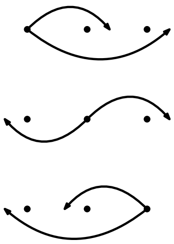
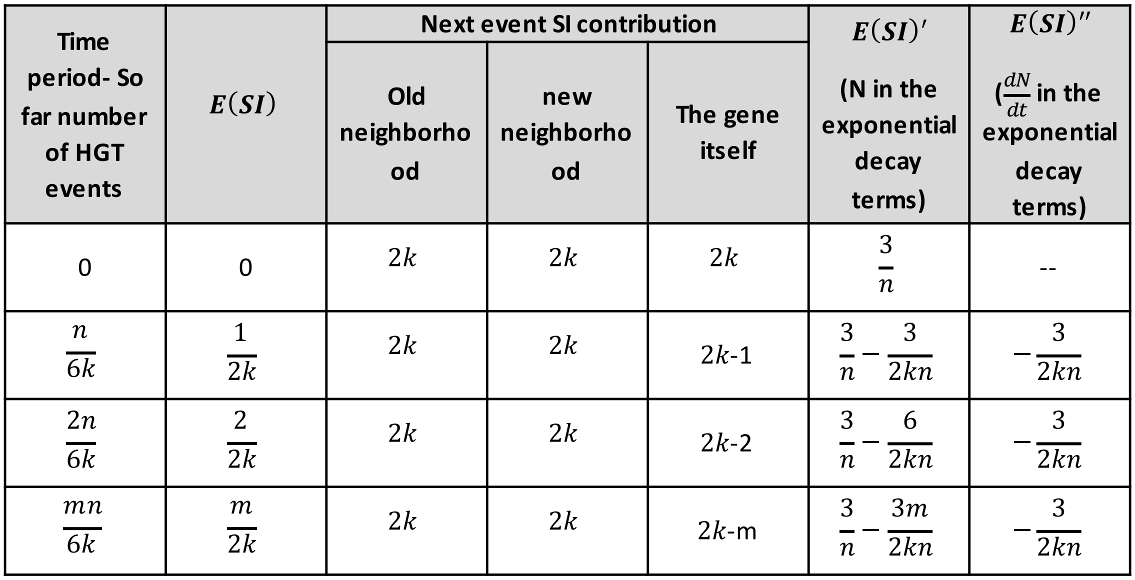


Fig. 4. The six possible values the random variable U_i can take when $n = 3$.



Time period- So far number of 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	$2k$	$2k$	$2k$	$\frac{3}{n}$	--
$\frac{n}{6k}$	$\frac{1}{2k}$	$2k$	$2k$	$2k$-1	$\frac{3}{n} - \frac{3}{2kn}$	$-\frac{3}{2kn}$
$\frac{2n}{6k}$	$\frac{2}{2k}$	$2k$	$2k$	$2k$-2	$\frac{3}{n} - \frac{6}{2kn}$	$-\frac{3}{2kn}$
$\frac{mn}{6k}$	$\frac{m}{2k}$	$2k$	$2k$	$2k$-m	$\frac{3}{n} - \frac{3m}{2kn}$	$-\frac{3}{2kn}$

Fig. 5. A manual calculation of the theoretical process for inferring the instantaneous increase at the first three "periods" .

Clique number	Genus list	Avg SI	Clique size	Estimated number of HGT events (average)	Genome size (average)	Number of HGT events as % of genome size
1	{'Borrelia'}	0.554	8	320.0	1158.8	27.6
2	{'unknow', 'Burkholderia', 'Ralstonia', 'Cupriavidus'}	0.819	25	3657.1	6367.9	57.4
3	{'Pelodictyon', 'Chlorobaculum', 'Chlorobium', 'Prosthecochloris'}	0.851	10	1478.8	2271.3	65.1
4	{'Shewanella'}	0.764	19	2072.1	4262.0	48.6
5	{'Streptococcus'}	0.856	10	1331.7	2000.9	66.6
6	{'Rickettsia'}	0.654	13	434.3	1192.7	36.4
7	{'Methanococcus'}	0.632	6	584.6	1721.0	34.0
8	{'Exiguobacterium', 'Oceanobacillus', 'Macrococcus', 'Bacillus', 'Geobacillus', 'Anoxybacillus', 'Staphylococcus', 'Listeria'}	0.877	25	2283.6	3202.6	71.3
9	{'Streptococcus', 'unknow'}	0.713	14	863.8	2037.3	42.4
10	{'unknow', 'Corynebacterium', 'Mycobacterium'}	0.86	10	1568.4	2331.7	67.3
11	{'Thermotoga'}	0.501	6	439.3	1867.8	23.5
12	{'Bartonella', 'Brucella'}	0.669	15	961.1	2572.6	37.4
13	{'Rhodopseudomonas', 'Nitrobacter', 'Bradyrhizobium', 'Oligotropha'}	0.855	11	3222.2	4947.1	65.1
14	{'Mycobacterium'}	0.817	19	2761.3	4827.3	57.2
15	{'unknow', 'Francisella'}	0.608	9	518.0	1627.3	31.8
16	{'Staphylococcus'}	0.226	12	228.0	2647.8	8.6
17	{'unknow', 'Shigella', 'Cronobacter', 'Serratia', 'Photorhabdus', 'Pectobacterium', 'Citrobacter', 'Klebsiella', 'Salmonella', 'Dickeya', 'Erwinia', 'Sodalis', 'Yersinia', 'Edwardsiella', 'Escherichia', 'Proteus', 'Enterobacter'}	0.796	85	2405.4	4492.3	53.5
18	{'Candidatus', 'Buchnera'}	0.938	8	605.9	486.0	124.7
19	{'Azotobacter', 'Pseudomonas'}	0.826	18	3167.8	5382.3	58.9
20	{'Clostridium'}	0.749	13	1630.0	3496.0	46.6
21	{'Photobacterium', 'Vibrio', 'Aliivibrio'}	0.797	14	2369.4	4410.4	53.7
22	{'Chlamydia', 'Chlamydophila', 'unknow'}	0.444	14	194.2	966.0	20.1
23	{'Lactobacillus', 'Pediococcus'}	0.884	12	1574.2	2122.0	74.2
24	{'Xanthomonas', 'Stenotrophomonas'}	0.717	10	1864.1	4391.6	42.4
25	{'Desulfotomaculum', 'Candidatus', 'Carboxydothermus', 'Pelotomaculum', 'Moorella', 'Ammonifex'}	0.923	6	2172.2	2447.6	88.7
26	{'Neisseria'}	0.524	7	518.1	2065.3	25.1
27	{'Agrobacterium', 'Rhizobium', 'Sinorhizobium', 'Ochrobactrum'}	0.866	12	4149.9	6132.2	67.7
28	{'Acidovorax', 'Variovorax', 'Delftia', 'Comamonas'}	0.88	6	3517.6	4910.2	71.6
29	{'Prochlorococcus', 'Synechococcus'}	0.706	16	904.7	2179.1	41.5
30	{'Bifidobacterium'}	0.787	8	981.5	1857.1	52.8
31	{'Bacillus', 'Geobacillus'}	0.612	15	1687.6	5315.8	31.7
32	{'Streptococcus'}	0.61	19	604.3	1893.1	31.9
33	{'Helicobacter'}	0.528	8	390.5	1531.4	25.5
34	{'Acinetobacter'}	0.616	7	1121.1	3481.7	32.2
35	{'Ehrlichia', 'Anaplasma'}	0.679	8	390.0	992.9	39.3
36	{'Geobacter'}	0.892	6	2926.0	3871.7	75.6
37	{'Campylobacter'}	0.76	8	839.4	1721.1	48.8
38	{'Methylobacterium'}	0.676	6	2147.7	5681.5	37.8
39	{'Sulfolobus'}	0.453	6	559.6	2756.5	20.3

Fig. 6. Real data results obtained for the 39 cliques derived by our technique. For every clique, we reported the genus corresponding to it, the size of the clique (#genomes), average SI and its corresponding #HGTs. The colors at the rightmost column signify distance fro average: In green- values fit the range of 1SD from the mean (i.e., > 28.92 and < 76.48). In blue values higher than 1SD from the mean (≥ 76.48). In yellow values lower more than 1SD from the mean (≤ 28.92).