Systematic Biology (2024), **0**, 0, *pp.* 1–24
doi:10.1093/sysbio/main

Gene Transfer-based Phylogenetics: Analytical Expressions and Additivity via Birth–Death Theory

Guy Katriel[1], Udi Mahanaymi[2], Shelly Brezner[2], Noor Kezel[3] Christoph Koutschan[4], Doron Zeilberger[5], Mike Steel[6], and Sagi Snir[2*]

[1]*Department of Mathematics, Ort Braude, Israel*
[2]*Department of Evolutionary and Environmental Biology, University of Haifa, Israel*
[3]*Department of Mathematics, University of Haifa, Israel*
[4]*RICAM, Austrian Academy of Sciences, Linz, Austria*
[5]*Department of Mathematics, Rutgers University, USA*
[6]*School of Mathematics and Statistics, University of Canterbury, NZ.*
**Corresponding author: ssagi@research.haifa.ac.il*

Abstract

The genomic era has opened up vast opportunities in molecular systematics, one of which is deciphering the evolutionary history in fine detail. Under this mass of data, analyzing the point mutations of standard markers is often too crude and slow for fine-scale phylogenetics. Nevertheless, genome dynamics (GD) events provide alternative, often richer information. The *synteny index* (SI) between a pair of genomes combines gene order and gene content information, allowing the comparison of genomes of unequal gene content, together with order considerations of their common genes. Recently, genome dynamics has been modelled as a continuous-time Markov process, and gene distance in the genome as a birth–death–immigration process. Nevertheless, due to complexities arising in this setting, no precise and provably consistent estimators could be derived, resulting in heuristic solutions.

Here, we extend this modelling approach by using techniques from birth–death theory to derive explicit expressions of the system's probabilistic dynamics in the form of rational functions of the model parameters. This, in turn, allows us to infer analytically accurate distances between organisms based on their SI. Subsequently, we establish additivity of this estimated evolutionary distance (a desirable property yielding phylogenetic consistency).

Applying the new measure in simulation studies shows that it provides accurate results in realistic settings and even under model extensions such as gene gain/loss or over a tree structure. In the real-data realm, we applied the new formulation to unique data structure that we constructed - the ordered orthology DB - based on a new version of the EggNOG database, to construct a tree with more than 4.5K taxa. To the best of our knowledge, this is the largest gene-order-based tree constructed and it overcomes shortcomings found in previous approaches. Constructing a GD-based tree allows to confirm and contrast findings based on other phylogenetic approaches, as we show.

Key words: **keywords:** Genome Dynamics, Prokaryotic Phylogenetics, Statistical Consistency, Synteny Index

Introduction

The genomic era has reached the point where tasks that seemed imaginary only a decade ago are now within reach. Among these tasks is the inference of the evolutionary history for tens of thousands of species, sometimes of very close origin. Such a history is depicted in a tree structure and is called a *phylogeny* or a *phylogenetic tree*. The leaves of that tree correspond to contemporary extant species, internal nodes correspond to ancestral species, and the tree's edges (or branches) between nodes correspond to evolutionary relationships. Despite the impressive advances in the extraction of molecular data, and of ever-increasing quality, finding a phylogenetic tree which accounts for the data in a satisfactory way is still a major challenge that requires reliable approaches for inferring the true evolutionary relationships between the species under study.

Statistical modelling in which the tree is a parameter of some assumed model is nowadays considered the method of choice for phylogenetic inference. Under this framework, vast efforts have been made, first to model data accurately, and then to draw inferences efficiently from the given data. One such approach is *maximum likelihood* (Felsenstein, 1978, 1981; Hasegawa et al., 1991; Yang, 1996), where the model (tree) selected is the one maximising the probability of observing the given data.

Standard phylogenetics, whether parsimony- or likelihood-based, analyses one or a few ubiquitous genes residing in all species under study, and uses the differences between respective gene copies i.e., orthologues, in order to infer evolution history. These genes are typically highly conserved by definition and therefore advantageous in certain settings such as very rapid viral evolution (Pybus and Rambaut, 2009) or long evolutionary distance where finer markers saturate (Ciccarelli et al., 2006). However, for the task of distinguishing the shallow branches of the prokaryotic tree, these genes often fail to provide a strong enough signal (Sevillya and Snir, 2019; Martinez-Gutierrez and Aylward, 2021; Rajendhran and Gunasekaran, 2011). In contrast, genome dynamics events (GDE's) are larger scale events compared to single nucleotide mutations, in which a complete gene or a sequence of genes, are involved. One such event is horizontal gene transfer (HGT), a mechanism by which organisms transfer genetic material to contemporaneous organisms rather than via vertical inheritance (Doolittle, 1999; Koonin et al., 2001; Ochman et al., 2000). Among prokaryotes, GDE's in the form of HGT and gene loss seem to provide far richer information, as indicated in (Puigbò et al., 2014): "The rates of genome change are remarkably high, typically tens of thousands of GDEs per nucleotide substitution per site, or tens to hundreds of GDEs per substitution per gene", and see also in e.g. (Schnknecht et al., 2014; Pang and Lercher, 2019; Koonin et al., 2021).) The latter fact calls for GD-based phylogenetic approaches, in particular when handling prokaryotes sharing a close origin. GD-based phylogenetics is mainly divided into gene-order-based and gene-content-based techniques. With the gene-order-based approach (Sankoff, 1992; Hannenhalli and Pevzner, 1999; Yancopoulos et al., 2005), two genomes are considered as permutations of the gene set, and distance is defined as the minimal number of operations needed to transform one genome to the other. In the other, the gene-content-based approach (Snel et al., 1999; Tekaia and Dujon, 1999; Fitz Gibbon and House, 1999) gene order is entirely ignored, and similarity is defined as the size of the set of shared genes. A statistical framework has been devised for part of both these models, the order- and content-based (Serdoz et al., 2017; Wang and Warnow, 2001; Biller et al., 2015; Sankoff and Nadeau, 1996; Lin et al., 2013; Zhao et al., 2021). The Jump operation studied here is accounted for by some these gene-order models; however to the best of our knowledge, a stochastic, rate-dependent framework accounting for HGT, has not been suggested.

Another and related line of works relies on gene tree amalgamation into a unified species tree, a task referred to as a *supertree construction* (Strimmer and Moulton, 2000; Bininda-Emonds, 2004; Baum, 1992; Ragan, 1992). Some of these works are

likelihood-based where GD events are accounted for by the differences in gene tree topologies (Morel et al., 2022, 2020). Nevertheless, these latter works ignore gene order, and in particular gene order likelihood.

Thus, devising a genuine evolutionary model, along with an estimator of the model parameters from observed data only, and an efficient inference method of this estimator, remains a challenging task.

A related task in this field is the *reconciliation* between a gene tree and the species tree. In this setting, a sequence of events acting on the species tree and yielding the given gene tree is sought. These events may contain events other than HGT which are commonly denoted *duplication, transfer* and *loss* (DTL) (Bansal et al., 2018; Stolzer et al., 2012). These works contain both parsimony-based approaches such as (Nakhleh et al., 2005; Doyon et al., 2010), and model/likelihood-based approaches (Szöllősi et al., 2013; Sjöstrand et al., 2014). Although it deals with the same objects, and the same events, as the tree is already given, the goal there is not tree reconstruction, and in particular not reconstruction based on gene order between multiple genes, as in our case here.

The *synteny index* (SI) (Shifman et al., 2013; Adato et al., 2015) was suggested as an alternative measure to the parsimony/statistical phylogenetics approaches mentioned above, allowing unequal gene content on one hand while accounting for the order among the shared genes. Here, the locality of a gene in the form of a "neighbourhood" is considered and compared with other genomes. Similarity between genomes is attained by averaging this local quantity over all the shared genes.

Aiming at a rigorous delineation of SI, in a recent paper (Sevillya et al., 2019), we defined an underlying simplistic model, the *Jump model*, to model genome dynamics, primarily HGT. Under this model, every gene stochastically (at some constant rate) "jumps" to a random location at the chromosome. Consequently distance between two genes along the genome, i.e., the number of genes separating between these two genes, can be described as a (critical) birth–death–immigration process. The setting poses intrinsic hurdles such as overlapping neighbourhoods, non-stationarity, confounding factors, and more. Consequently, precise quantities could not be obtained in this earlier work and were calculated heuristically. The Jump model consists of a Jump operation embedded within a stochastic framework. While the basic Jump operation is subsumed in some of the gene-order models mentioned above, to the best of our knowledge, no complete time-dependent framework accounting for HGT, has been suggested. (Dalevi and Eriksen, 2008) defines the single gene transposition model that is equivalent to a Jump, and expected distances are derived by a function of the number of breakpoints, however, the meaning of a model there is a type of operation as opposed to a stochastic, rate-dependent model considered here.

In this work, we take the Jump model and the SI a significant step further by deriving exact and invertible analytical expressions (Theorems 2, 3, and 4) that allow for the evolutionary distance between species to be inferred from the (averaged) SI values under the Jump model. By an earlier result (Theorem 1), this implies that the difference between these estimates of evolutionary distance and the true evolutionary distances converges to zero as the number of genes becomes large. Our results rely on techniques from the theory of birth-death processes. On the experimental side, we first show that the new expressions provide accurate reconstructions, even for real-life problem sizes although the theoretical underlying model on which these expressions are based assumes infinitely long genomes. We note here that, for the sake of rigorous analysis, the pure theoretical Jump model consists of only the Jump operation, and hence implies comparisons between equal content genomes - genomes over the same gene set. Such a model can can accommodate other, however restricted, evolutionary scenarios including gene gain/loss events, as we show in the Methods part. Nevertheless, to allow for a broader range of scenarios resulting in genomes with unequal content, as is the case in real data, we have

devised two heuristics - the *union* and the *intersection* gene set approaches described in the experimental section. Using the Jump model under these heuristics to simulated data including both jumps and gene gain and loss events (i.e., unequal gene content), shows robustness to such more diverse regimes.

For real data, we created a new database of ordered orthology groups, based on the EggNOG (Huerta-Cepas et al., 2018) orthology database, encompassing over 4445 organisms spanning the entire prokaryotic phylogenetic spectrum. Applying the new measure to this database, produces a tree with very high agreement with the NCBI taxonomy (Federhen, 2011; Schoch et al., 2020). To the best of our knowledge, this is the largest genome-dynamics-based tree. In comparison with other SI-based trees, it is evident that the new technique reconstructs significantly more realistic distances, attesting to its capability as a distance measure in various other applications of genome dynamics (Che et al., 2006; Rogozin et al., 2002). Moreover, contrasting the SI-based trees with the NCBI taxonomy, suggests several incongruences that may be of independent, intrinsic interest.

Comment: For the sake of readability, technical proofs have been moved to the Appendix. Additionally, as both the theoretical and the experimental parts are technically involved, we provide in the Supplementary Text brief self-contained background to the theoretical material employed, as well as further details for the experimental parts. Finally, for the sake of reconstructability, we provide in the Supplementary Material data produced during this research. Supplementary Text and Material are found in the DRYAD link.

Materials and Methods

We start by defining a restricted model – *the Jump model* – which can be regarded as a transfer between genomes over the same gene set (*equal content*). Biologically, the *Jump* operation, in which a gene moves to another location, can account for several GDE's, such as a gene duplication and a subsequent loss, 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) as discussed in e.g. (Liu et al., 2004).

The Jump Model: Let $\mathcal{G}^n = (g_1, g_2, \ldots, g_n)$ be a sequence of n 'genes', and henceforth we remove the superscript n as it holds throughout. In the analysis, we will assume that n is large enough to allow us to ignore the tips of $\mathcal{G}$ (or, equivalently, $\mathcal{G}$ is cyclic and there are no tips). We now introduce a stochastic process operating on $\mathcal{G}$. Consider the following continuous-time Markovian process $\mathcal{G}(t), t \geqslant 0$ 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 a 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 that are 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 $\mathcal{G}(t) = (g_1, g_2, g_3, g_4, g_5)$, then g_4 might transfer to be inserted between g_1 and g_2 to give the sequence $\mathcal{G}(t+\delta) = (g_1, g_4, g_2, g_3, g_5)$. 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 $(g_1, g_2, g_3, g_5, g_4)$.

Since the model assumes 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 possibility of more than one transfer occurring in the time period δ (this possibility has probability of order δ^2 and so is asymptotically negligible compared with terms of order δ as $\delta \to 0$). Moreover, a single transfer event always results in a different sequence.

The Synteny Index: Let k be any constant positive integer (note that it may be possible to allow k to grow slowly with n, but we will not explore such an extension here). For $j \in k+1, \ldots, n-k$, the $2k$-neighbourhood of gene g_j in a genome $\mathcal{G}$, $N_{2k}(g_j, \mathcal{G})$ is the set of $2k$ genes (different from g_j) that have a distance of at most k from g_j in $\mathcal{G}$. We also define $SI_j(t)$ as the relative intersection size between $N_{2k}(g_j, \mathcal{G}(0))$ and $N_{2k}(g_j, \mathcal{G}(t))$, or formally, $SI_j(t) = \frac{1}{2k}|N_{2k}(g_j, \mathcal{G}(0)) \cap N_{2k}(g_j, \mathcal{G}(t))|$ (this is also called *the Jaccard index* between the two neighbourhoods (Jaccard, 1901)).

Let $\overline{SI}(\mathcal{G}(0), \mathcal{G}(t)$ be the average of these $SI_j(t)$ values over all j between $k+1$ and $n-k$. That is,

$$\overline{SI}\left(\mathcal{G}(0), \mathcal{G}(t)\right) = \frac{1}{n-2k} \sum_{j=k+1}^{n-k} SI_j(t). \tag{1}$$

Subsequently, when time t does not matter, we simply use $\overline{SI}$ or simply SI where it is clear from the context.

Phylogenetic Trees and Distances: For a set of species (denoted *taxa*) $\mathcal{X}$, a phylogenetic $\mathcal{X}$-tree T is a tree $T = (V, E)$ for which there is a one-to-one correspondence between $\mathcal{X}$ and the set $\mathcal{L}(T)$ of leaves of T. A tree T is *weighted* if there is a weight (or length) function associating non-negative weights (lengths) to the edges of T. In this paper, we will use the term length, as it corresponds to the number of events or the time span. Edge lengths are naturally extended to *paths*, where the path length is the sum of edge lengths along the path. Assume a model M operating on T by generating events, starting from the root down to the leaves, where edge lengths serve as expected number of events generated by M. The notion of additivity is classical in phylogenetics(Buneman, 1971; Semple and Steel, 2003). We briefly specialise it to our case. Let D be some *pairwise marker* or *a distance measure* between the outcome of M on the leaves of T. Then D is said to be *additive on the model* M if D can be transformed (or *corrected*) by applying a fixed function f to D (only), such that the corrected distance converges to the expected number of events under the model M, as the amount of data (e.g. sequence length, or in our case n, the number of genes), becomes large.

Gene Neighbourhood as a Markov Chain

We now introduce a local random process, induced by the Jump model defined above that operates on the entire genome level. This local model will play a key role in the analysis of the random variable $\overline{SI}(\mathcal{G}(0), \mathcal{G}(t))$. Consider the location of a gene g_i, that does not jump during time period t, with respect to another gene $g_{i'}$. Without loss of generality assume $i > i'$ and let $j = i - i'$. Now, there are j 'slots' between $g_{i'}$ and g_i into which a third jumping gene g_k can be inserted, but only $j-1$ genes in that interval can jump out. Obviously, a jump into that interval moves $g_{i'}$ one position away from g_i, and a jump from that interval, moves $g_{i'}$ closer to g_i. This local model, describing the distance between $g_{i'}$ and g_i, can be described by a continuous-time random walk on the state space $1, 2, 3, \ldots$ with transitions from j to $j+1$ at rate $j\lambda$ (for all $j \geqslant 1$) and from j to $j-1$ at rate $(j-1)\lambda$ (for all $j \geqslant 2$), with all other transition rates being 0. This, the distance between $g_{i'}$ and g_i, is thus a (generalised linear) birth–death process, illustrated in Fig. 1. We note though that this is not an independent model, occurring between individual genes, separated from the Jump model operating at the genome scale. Rather, the birth–death process modeling the distance between individual genes, is induced by that larger scale model of the Jump.

More formally, we will let X_t denote the random variable that describes the number of slots between two genes under this process described above. Then X_t is a continuous-time random walk on state space $1, 2, 3, \ldots$, with an arbitrary initial condition X_0 and transition probabilities of X_t defined as follows:

$$\mathbb{P}(X_{t+\delta} = j+1 | X_t = j) = j\lambda\delta + o(\delta), \quad j \geqslant 1, \tag{2}$$
$$\mathbb{P}(X_{t+\delta} = j-1 | X_t = j) = (j-1)\lambda\delta + o(\delta), \quad j \geqslant 1. \tag{3}$$

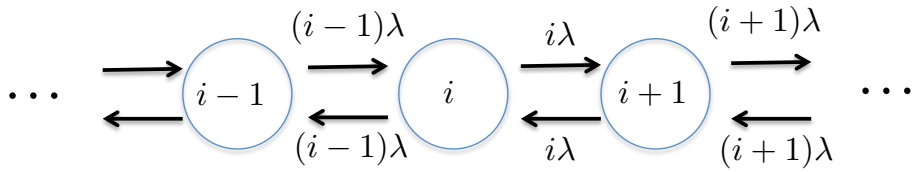

Fig. 1. Transitions for the process X_t

The process X_t is slightly different from 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 connection is key to the analysis of the divergence times that we establish below.

Results

Explicit Expressions for the Divergence Time

We now present the main theoretical contribution of this work, which is an analytical expression of divergence times. We first recall a result of (Sevillya et al., 2019) that links SI and the transition probabilities of the birth–death process X_t. This raises the need to obtain explicit expressions for these probabilities, which we do by making use of known results from the theory of birth–death processes. This theory also allows us to give a proof of the monotonicity of the SI as a function of time (in the limit of large n), a result that is crucial in order to ensure that we can use our explicit expressions to solve the divergence time in terms of the SI.

Let $p_{i,j}(t)$ be the transition probability for X_t to be at state j, given that at time 0 it was at state i:

$$p_{i,j}(t) = \mathbb{P}(X_t = j \mid X_0 = i), \quad i, j \geqslant 1.$$

We denote the conditional probability that $X_t \in [k]$ given that $X_0 = i$ by:

$$q_{i,k}(t) = \sum_{j=1}^{k} p_{i,j}(t). \tag{4}$$

Next, let

$$q_k(t) := \frac{1}{k}\sum_{i=1}^{k} q_{i,k}(t) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i,j}(t). \tag{5}$$

The quantity $q_k(t)$ is the probability that for a gene at an initial state i (i.e., at distance from a reference gene) chosen uniformly at random between 1 and k, the process X_* is still between 1 and k at time t. In (Sevillya et al., 2019) we proved the following result:

THEOREM 1 For any given value of t, as $n \to \infty$:

$$\overline{SI}(\mathcal{G}(0), \mathcal{G}(t)) \xrightarrow{p} \exp(-2\lambda t) q_k(t),$$

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

In the following we assume, without loss of generality, that $\lambda = 1$ (this is simply rescaling time). The functions $p_{i,j}(t)$ can be expressed as solutions of an infinite system of ordinary differential equations (Sevillya et al., 2019) (the Kolmogorov forward equations corresponding to the birth–death process), and these differential equations may be used to numerically approximate $p_{i,j}(t)$ and therefore the key quantity $q_k(t)$. However, in the present paper we will derive *explicit* algebraic expressions for $p_{i,j}(t)$ and thus $q_k(t)$. It thereby becomes possible to use Theorem 1 to solve for the divergence time t in terms of the SI.

Explicit expressions for $p_{i,j}(t)$

THEOREM 2

$$p_{i,j}(t) = \frac{1}{(t+1)^{i+j-1}} \cdot \sum_{\ell=1}^{\min(i,j)} \frac{(i+j-\ell-1)!}{(i-\ell)!(j-\ell)!(\ell-1)!} \left(1-t^2\right)^{\ell-1} t^{i+j-2\ell}. \tag{6}$$

This result follows from some general results for birth–death processes (see (Anderson, 2012) for more details). The full proof is given in the Appendix.

Explicit Expression for $q_k(t)$

As stated above, Theorem 1 (originally from (Sevillya et al., 2019)) gives an expression for the SI value between two genomes, $\mathcal{G}(0)$ and $\mathcal{G}(t)$. Nevertheless, in that paper, we could not derive an expression only in terms of the number of events that occurred during time t (or, alternatively, in a path along the tree of length λt "separating" genomes $\mathcal{G}_i$ and $\mathcal{G}_j$) as we could not arrive at an explicit expression for q_k. Now that we have obtained explicit expression for $p_{i,j}(t)$ in Theorem 2 we can explicitly describe q_k as follows.

THEOREM 3

$$q_k(t) = \frac{1}{k} \sum_{\ell=0}^{k-1} \sum_{i=0}^{k-\ell-1} \sum_{j=0}^{k-\ell-1} \frac{(i+j+\ell)!}{i!j!\ell!} t^{i+j} (t+1)^{-i-j-2\ell-1} \left(1-t^2\right)^{\ell}. \tag{7}$$

The proof is brought in the Appendix.

To give Theorem 3 an actual expression, we provide a few instances of the above formula:

$$q_2(t) = \frac{2t^2+2t+1}{(t+1)^3}$$

$$q_3(t) = \frac{3t^4 + 6t^3 + 8t^2 + 4t + 1}{(t+1)^5}$$

$$q_4(t) = \frac{4t^6 + 12t^5 + 26t^4 + 26t^3 + 18t^2 + 6t + 1}{(t+1)^7}$$

$$q_5(t) = \frac{5t^8 + 20t^7 + 60t^6 + 90t^5 + 102t^4 + 68t^3 + 32t^2 + 8t + 1}{(t+1)^9}.$$

In the supplementary text we provide the actual function for $q_{10}(t)$ that was used in the real data analysis.

Monotonicity of the SI Measure

Recall that we assumed, without loss of generality, that $\lambda = 1$, and so our goal now is to prove the monotonicity of the function, $h_k(t) = e^{-2t}q_k(t)$ and thus (by Theorem 1) the SI measure itself, in the limit of large n. In fact we will prove that $q_k(t)$ itself is monotone decreasing, which obviously implies that $h_k(t)$ is also monotone decreasing.

THEOREM 4 The function $q_k(t)$ is monotone decreasing on $[0, \infty)$.

The fact that $h_k(t) = \exp(-2t)q_k(t)$ is strictly monotone decreasing with t implies that $h_k(t)$ is a one-to-one–function (or *injective*) and hence the inverse function h_k^{-1} is well-defined. This allows us to use Theorem 1 to reconstruct the expected number of events (i.e. jump, or equivalently the time t) separating two sequences of n genes (where n is large) given their pairwise SI value. By applying h_k^{-1} to the $\overline{SI}$ for these two gene sequences, the estimated number of events is obtained. By Theorem 3, we have an explicit value for $h_k(t)$, so the value $h_k^{-1}(\overline{SI})$ can be calculated by numerically solving a simple equation.

Since the expected number of events is additive on the tree, that is, the sum of events along the tree edges separating two leaves equals the number of events occurred between these leaves, we can conclude the following corollary:

Corollary 1 Assume a genome with n genes at the root of an underlying tree T, is evolving according to the Jump model defined above. Then, as $n \to \infty$, the number of events per gene between the leaves of T can be reconstructed in a statistically consistent way from the $\overline{SI}$ values between the genomes at the leaves of T, by applying the transformation h_k^{-1}.

As a fully resolved (i.e. binary) tree T can be uniquely and consistently reconstructed from its pairwise distances by applying the distance-based reconstruction method, e.g. the Neighbour-Joining (NJ) algorithm, we obtain the following:

Corollary 2 Assume a genome is evolving on a tree binary tree T as in Corollary 1. Then T can be reconstructed in a statistically consistent way (as n grows) by transforming the $\overline{SI}$ values.

Experimental Results

In this section, we describe the experiments we conducted to demonstrate the applicability of the theoretical results described above. We begin with simulation results based on the Jump model and then move to an analysis of real genomic data.

Simulation Results, Single Edge, the Pure Jump Model: We simulated the Jump model over a single edge of length t, i.e. from $\mathcal{G}(0)$ to $\mathcal{G}(t)$, and for various values of the number n of genes. We set $k = 3$ (i.e. a neighbourhood of $2k = 6$), the rate was fixed at $\lambda = 1$ and time t varied over the interval $[0, 0.5]$. This has yielded a Jump probability that was applied to every gene in the initial (parent, $t = 0$) genome. For each value of t, the SI between the parent and the child genome was computed. The top part of Fig. 2 displays the value $e^{-2t}SI(t)$ (recall that $\lambda = 1$ and hence vanishes at the exponent) for each of 10 simulations, and the function $q_3(t)$ which is the limit to which $e^{-2t}SI(t)$ converges as $n \to \infty$. As can be seen, although there is some variability due to randomness, this variability decreases as n increases, and the agreement with the limiting curve $q_3(t)$ is clear.

In a related experiment, we checked how well the value $SI(t)$, computed using the simulated data, can be used to estimate the time t. For each value of t, we compute $SI(t)$ from the simulated data, and use this to estimate t by numerically solving the following equation:

$$e^{-2\hat{t}}q_3(\hat{t}) = SI(t), \tag{8}$$

In the lower part of Fig. 2 the true value of t is compared with the estimated values $\hat{t}$ for 10 simulations.

We note that the relevant values of λt as found in (Sevillya et al., 2019) are around 0.4 for distances within the phylogenetic rank of genus. We see that the error is almost insignificant even for realistic genome sizes, as we have here.

Simulation Results, Single Edge, Adding Gene Gain and Loss: Next, we extended the pure Jump model to include gain/loss events; still as above over a single edge: for each jump event, we also generate a gain/loss event with probability p, with equal probability for a gain or a loss, so that the expected genome length is fixed. However, here we face the problem that the gene content of pairs of genomes are not identical, a fact which needs to be accounted for when computing the SI of two genomes. We have devised two different approaches to computing the SI of two genomes with non-identical gene content.

In the first approach (I), *the union gene set approach*, we simply replace the sum in (1) by a sum only over the genes that are common to both genomes; however, the k neighborhoods whose intersection is used to define the quantities $SI_j(t)$ include also genes which are present in only one of the two genomes. We again used Eq. (7) to infer the distances. The results for the gain/loss probabilities $p = 0.1$ and $p = 0.2$ are shown in Fig. 3.

In a second approach (II), *the intersection gene set approach*, for computing the SI between two genomes, we first excise all the genes which are not common to both genomes from each of the genomes. This produces a pair of genomes of the same size (i.e., the size of the intersection of the genome contents of the original genomes), and the two genomes now have identical gene content, so that their SI can be computed in the standard way. The results for the gain/loss probabilities $p = 0.1$ and $p = 0.2$ are shown in Fig. 4.

Comparing Fig. 3 and Fig. 4, it is clear that the second approach for computing the SI of genomes with non-identical gene content is the more appropriate one for enabling accurate estimation in the presence of gene gain and loss; indeed in Fig. 3 we see that there is a systematic bias in the estimators, which is not present in Fig. 4. Of course the procedure of excising all genes which are not common to both genomes, performed in the second approach, entails some loss of information, which is responsible for the larger variance of the estimators as seen in Fig. 4 compared to the case $n = 2000$ shown at Fig. 2.

Simulation Results over Tree Structure: Our last extension in this part is from a single edge (as reported above) to a tree structure. We describe this briefly here (a detailed

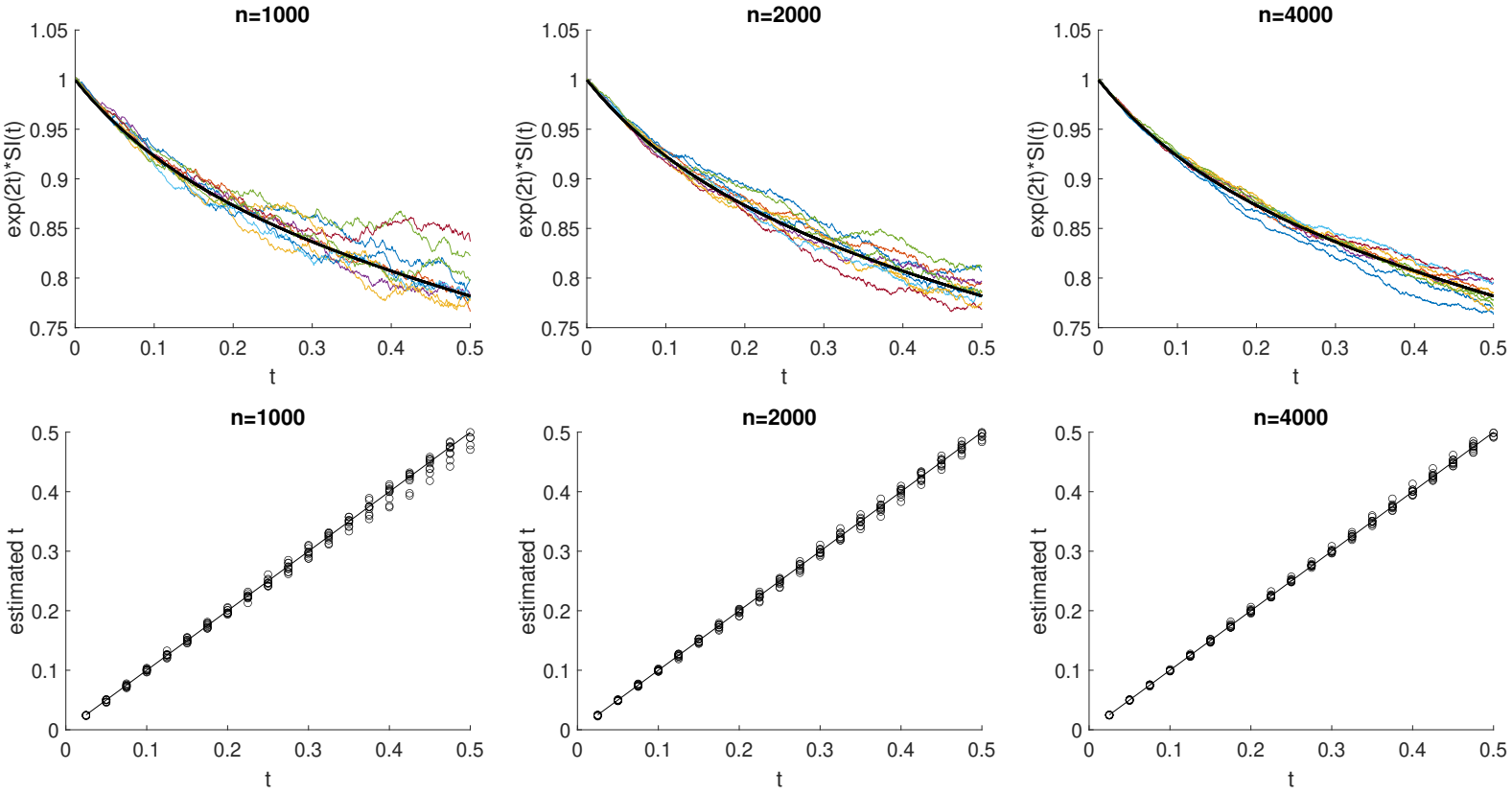

Fig. 2. **Simulation Results, Single Edge, the pure jump model:**. Genome sizes $n = 1000, 2000, 4000$. Top: Comparison of the curves $e^{\lambda t}SI(t)$ computed using simulation with the limiting curve $q_3(t)$, Bottom: Estimated vs effective t.

description of the procedure is provided in the supplementary text). We first draw a random tree with edge lengths distributed exponentially with mean l. A genome evolves recursively down this tree starting at the root with the identity permutation, via GD events such as Jump, gain, and loss, identically as the *single-edge* experiments described above. For each edge, the edge length e_l is the expected number of GD events. At the end of this procedure, we have genomes at the leaves over ordered subsets of the initial set at the root. We applied the intersection gene set approach (Approach II above) to cope with presence of gene gain/loss, in order to reconstruct a tree. The reconstructed tree was compared to the original model tree, using the Robinson–Foulds (RF) symmetric difference (Robinson and Foulds, 1981). The results, in terms of normalized error rate (incorrect edges) versus average edge length, are shown in Fig. 5 for a tree over 26 leaves. As can be seen, for small values of edge lengths, reconstruction quality is fairly high, almost perfect. Nevertheless tree distance rises (i.e. reconstruction quality falls) sharply initially and slowly levels off towards the value of 1. Note though that even at average jump rate of one, we still obaserve a reconstruction of 0.5 meaning half of the edges are correct.

Real Data Results: Here we report the real data results obtained using the new technique. Because of space limitations, and for the sake of reconstructability, fuller details and data are provided in the supplementary text and material respectively. We applied our method to real genomic data consisting of 4445 prokaryotes taken from the orthology data

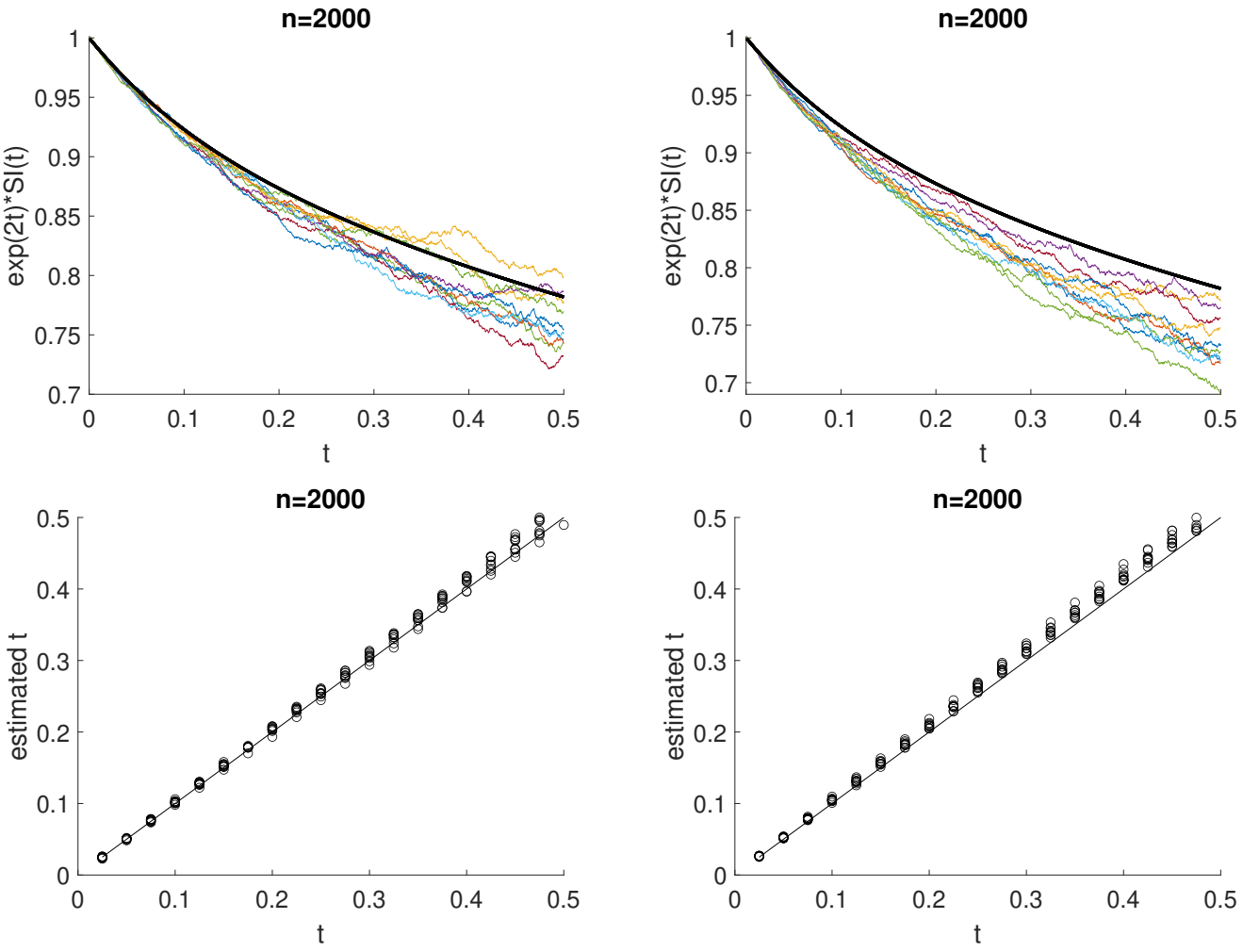

Fig. 3. **Simulation Results, Jump plus gain/loss:**. Genome size $n = 2000$ Top: Estimated vs effective $\hat{q}_k$ for gain/loss $p = 0.1, 0.2$. Bottom: Estimated vs effective $\hat{t}$ for gain/loss $p = 0.1, 0.2$. Here the computation of the SI is performed using approach I, as described in the text.

base EggNOG (Huerta-Cepas et al., 2018) with 4.4M clusters of orthologous groups (COGs) (Tatusov et al., 2001). For each COG, EggNOG provides a flat 'members' file indicating the organisms that harbour this gene, along with its location in the genome. This allowed us to sort the genes by location along the genome. Within this representation, a genome is simply a list of COGs sorted by genome location, where the COG names are universal across all organisms. Hence, we can infer neighbourhood similarities across genomes and therefore the pairwise SI values between any two genomes which we then store in an $n \times n$ SI matrix. We set $k = 10$ which was found to be informative for these data (Shifman et al., 2013; Sevillya et al., 2019) and computed SI for all pairs of taxa. The crude SI values are strongly concentrated around 0.02, as shown in Fig. 6(R). In order to convert the SI values to a dissimilarity measure, we set $d_{SI} = 1 - SI$. Once a (pairwise) dissimilarity D matrix has been computed, we can then apply a distance-based phylogenetic method to estimate a tree T in which the leaves are labelled by the organisms under study.

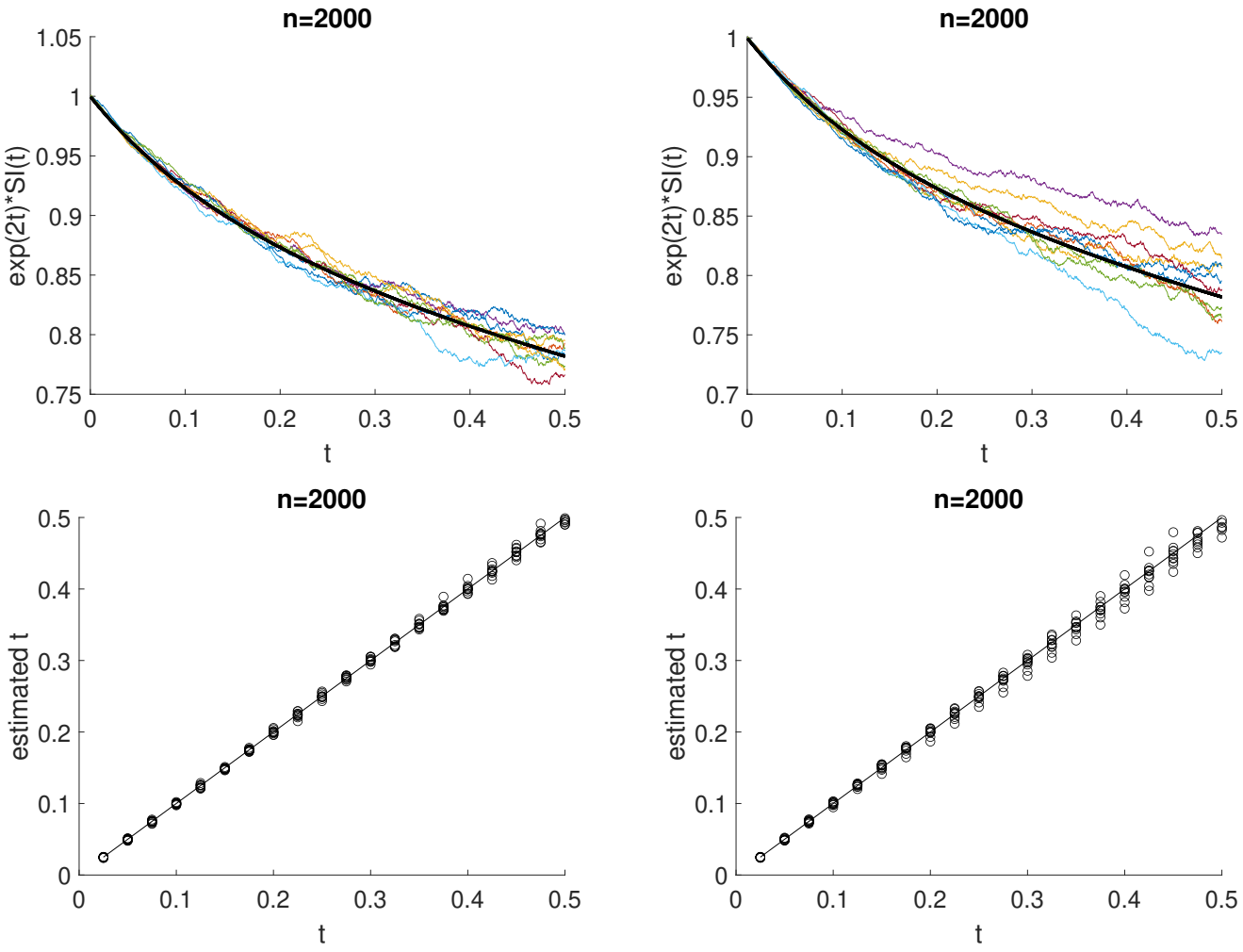

Fig. 4. **Simulation Results, Jump plus gain/loss:**. Genome size $n = 2000$ Top: Estimated vs effective $\hat{q}_k$ for gain/loss $p = 0.1, 0.2$. Bottom: Estimated vs effective $\hat{t}$ for gain/loss $p = 0.1, 0.2$. Here the computation of the SI is performed using approach II, as described in the text.

Path distances between the leaves of T, should approximate the distances in D. The most accurate algorithm for this task is the neighbour joining (NJ) algorithm (Saitou and Nei, 1987). Therefore, we used the program Neighbour from Phylip (Felsenstein, 1993) to construct a tree that we call the $1 - SI$ tree. Recall now that Eq. (8) was devised to "correct" the crude d_{SI} and provide a (provably) more reliable distance. Hence, we "corrected" the SI matrix accounting to Eq. (8) (specifically, finding $\hat{t}$ by solving Eq. (8) for the appropriate SI value in the matrix) and then applied Neighbour to this matrix, yielding what we denote the *exact tree*. Finally, as in (Sevillya et al., 2019), we did not have an explicit expression for distance and were forced to develop a simulation-based heuristic, we also constructed the *heuristic tree* by using Formula (9) from (Sevillya et al., 2019).

EggNOG labels its organisms with the same taxon ID used by the NCBI taxonomy database (Federhen, 2011). This database is also furnished with taxonomic ranks in a child-to-parent relationship that we can used for our task. We therefore constructed a tree from this child–parent relationship. This NCBI tree spans about 1.2M organisms with

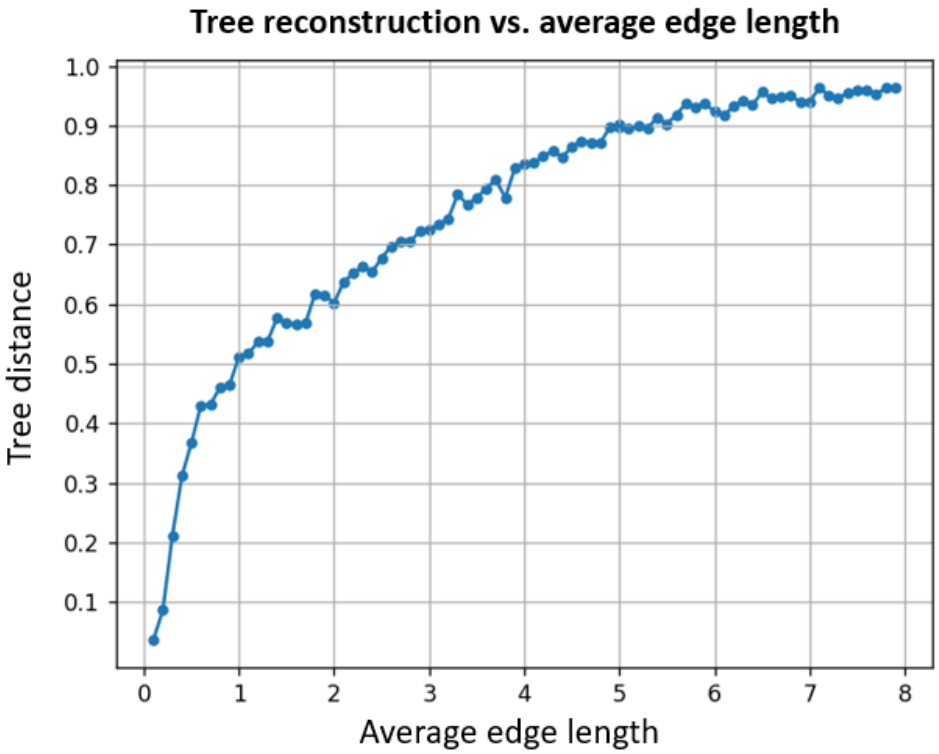

Fig. 5. **Simulation Results under a Tree Topology:**. Genomes of size $n = 2000$ are evolved down a tree over 26 taxa under the same regime as in the above trials. The x and y axes are average edge length and normalized tree distance between the model, original tree, and the reconstructed tree.

maximum depth (i.e. ranks) of 39. We extracted the tree induced by EggNOG's 4445 taxa (this is done by removing the leaves that were not in the selection, and the paths leading solely to them) and used this tree as a reference tree, dubbed *the NCBI tree* (or *taxonomy*). The four trees appear in Fig. 7 in two formats - rectangular (L) and polar(R). As can be seen, the $1 - SI$ and the heuristic trees exhibit serious flaws we will elaborate on later. We wanted to measure the distance from each of the three SI- (or GD-, genome dynamic) based reconstructed trees to the reference NCBI tree. Again we used the Robinson–Foulds (RF) symmetric difference (Robinson and Foulds, 1981) tree metric. In presence of a reference, or a model, tree, the RF-distance can also be used to derive the false positive and false negative (FP, FN) rates. The relevant distances are presented in Figure 6(L). As can be seen, the exact tree from Eq. (8) is the most similar to the NCBI tree and the heuristic tree is the least similar.

The RF distance is very sensitive and uninformative for large trees (Siu-Ting et al., 2014). Hence, we adopted a coarser compatibility measure to allow a more intuitive assessment that can also detect diferences between the two approaches, the sequence-based NCBI tree, and the other GD-based trees. We divided the NCBI reference tree into disjoint subtrees with sizes of between 80 and 800 taxa, resulting in 14 subtrees in total. This tree partitioning served as a *reference colouring* where each such NCBI subtree is mapped to a colour and all taxa (leaves) in this subtree attain that same colour (see the NCBI trees at the upper row of Fig. 7). As the NCBI Taxonomy provides classification to these families and genera represented by internal nodes in the NCBI tree, we could associate each such

subtree (or *clade*) with its corresponding evolutionary class. The classification is presented in Table 1.

Subtrees taxonomy

Color	Root ID	subtree description - NCBI Taxonomy
0	1117	Cyanobacteria (blue-green algae), phylum, cyanobacteria
1	72274	Pseudomonadales order, g-proteobacteria
2	91347	Enterobacterales order, g-proteobacteria
3	135614	Xanthomonadales order, g-proteobacteria
4	135622	Alteromonadales order, g-proteobacteria
5	28211	Alphaproteobacteria class, a-proteobacteria
6	28216	Betaproteobacteria class, b-proteobacteria
7	68525	delta/epsilon subdivisions subphylum, proteobacteria
8	91061	Bacilli class, firmicutes
9	186801	Clostridia class, firmicutes
10	909932	Negativicutes class, firmicutes
11	68336	Bacteroidetes/Chlorobi group clade, bacteria
12	2037	Actinomycetales order, high G+C Gram-positive bacteria
13	544448	Tenericutes phylum, bacteria

Table 1. Subtrees taxonomy as provided by NCBI Taxonomy

The reference colouring of the NCBI tree allows us to measure this coloring in the other three SI-based trees as we describe next. In particular, as shown below, it allows us to detect incongruences between the SI-based and the sequence-based trees, that may suggest either misclassification or significant evolutionary events. Recall that the leaves in all trees are colored with the original color from the NCBI tree. Now, for a colour c, the c-subtree is defined as the minimal connected graph (subtree, in our case) containing all c-coloured leaves. Given a coloured tree (i.e. with some of the nodes coloured), such a colouring is said to be *convex* on that tree, if for every two colours c and c', the c- and c'-coloured subtrees are disjoint (Moran and Snir, 2008, 2007). It is clear that the NCBI tree is convex, since the colouring is defined by this tree, i.e. for disjoint subtrees. Nevertheless, we aimed to test how far from convexity the NCBI colouring on the other trees is. There are rigorous definitions for the latter (the *recoloring distance (Moran and Snir, 2008)*); however, we used this approach to provide an intuitive and visual measure of compatibility, as demonstrated in Fig. 7.

As can be seen from the figure, all three trees maintained decent convexity under the NCBI colouring; however, it seems the exact tree has fewer violations than the heuristic and the $1 - SI$ trees. Fig. 7 also reveals major flaws in the heuristic and the $1 - SI$ approaches that are corrected by the exact approach. The $1 - SI$ approach takes crude values as the distances. These values are excessively concentrated around a tiny value of 0.02, causing severely distorted branch lengths, resulting in an artificially ultrametric tree with extremely short internal branches (third row in Fig. 7), which may disappear under bootstrapping, yielding a poorly resolved tree. Alternatively, the heuristic approach of (Sevillya et al., 2019), apart from achieving an outstandingly high RF distance, produces few exceptionally long branches non-proportional to the rest of the branches (left tree in the fourth row in Fig. 7). Hence, our real-data experiments showed that the theoretical conversion achieves its goal by producing a realistic distance, thereby

tree	Crude Robinson-Foulds				number of edges	common edges	% false positive	%false negative
NCBI Taxonomy	0	5516	4396	4358	1261			
Heuristic	5516	0	8884	8884	4443	94	98	92
$1 - SI$	4396	8884	0	2840	4443	654	85	48
Exact	4358	8884	2840	0	4443	673	85	47

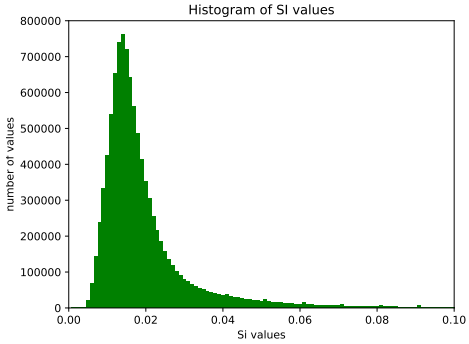

Fig. 6. **Left:** Robinson–Foulds distances. Pairwise RF distances between the four trees are depicted in columns 1–5. Column 6 contains the number of internal edges in each tree. Column 7 depicts the number of common edges with the NCBI tree, and Columns 8–9 depicts the rate of false positive and negative respectively (considering the NCBI tree as a model tree) **Right:** The distribution of pairwise SI values between all pairs from the 4445 EggNOG prokaryotic genomes.

correcting the severe flaws caused by the two more simplistic SI-based approaches.

As mentioned above, the incongruence in coloring between the two types of trees, GD-based (SI trees) and sequence-based (NCBI tree), may suggest closer scrutiny aiming at detecting genuine evolutionary phenomena. Thus, this procedure was followed. For each of the three SI-based trees and each of the 14 colors (corresponding to clades in the NCBI tree), we found the maximal subtree such that more than 90% of its leaves are colored in the desired color. For each such subtree, we counted the leaves colored in that color and their percentage, first of the total subtree size (number of leaves in the subtree), and next of the total number of leaves colored with that color. The exact numbers for each subtree appear in the table in supplementary text. While for some colors, all three SI-based tree are in nearly perfect agreement, conferring strong support in the NCBI classification, for other colors, all SI-based trees agree that the NCBI classification is incorrect. For example, the clades *Pseudomonadales*, *Alteromonadales*, the *delta/epsilon* subdivisions, and the *Bacilli* class, exhibit strong incongruence with the NCBI tree. In the supplementary text we provide details from the literature supporting the misclassification of these clades. Equivalently, this coloring provides hints to mislocation of specific taxa, a phenomenon referred to as *rogue taxa* (Smith, 2021). For example, all three SI-based trees allude to misclassification of taxa *gamma proteobacterium WG36, Gallaecimonas xiamenensis 3-C-1*, *Arsukibacterium perlucidum DSM 18276*, *Rheinheimera baltica DSM 14885*, and *Rheinheimera sp. A13L*, possibly mislocated based on their NCBI classification. Fuller details can be found in the supplementary text.

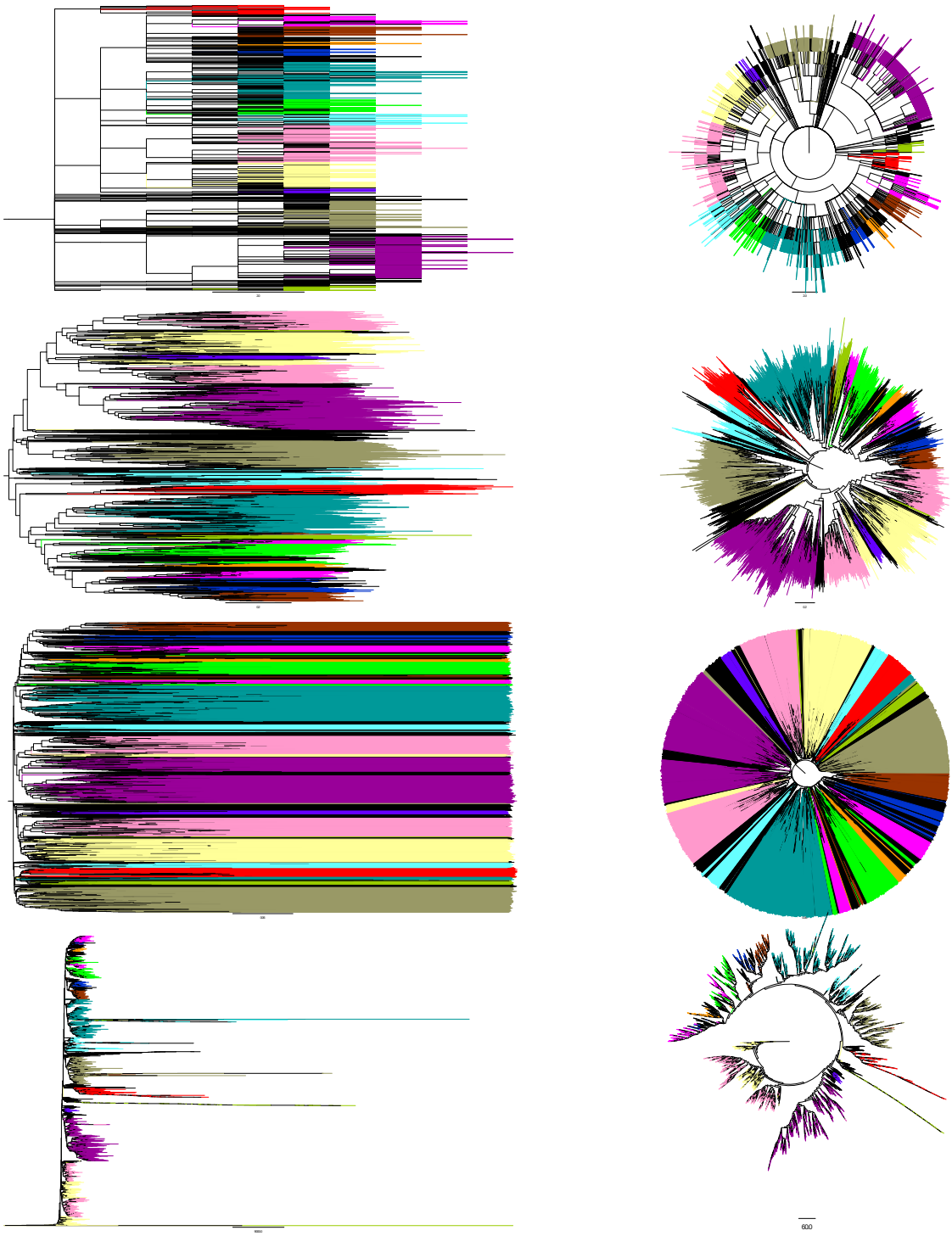

Fig. 7. **Coloured Trees:** Left: rectangular shape; right: polar shape. From the top: (1) The NCBI Taxonomy, (2) The exact SI tree, (3) the $1 - SI$ tree, (4) the heuristic exp. decay tree (the polar shape on right has log distances to accommodate the extremely long branches).

Discussion

In this paper, we explored the consequences of modeling genome organisation as a continuous-time Markov process. Although the initial modelling was suggested recently, fundamental problems were left open, making it impossible to formally answer basic questions such as the time since divergence on a tree or the additivity of the synteny index as a phylogenetic marker. Here, we have advanced this front by applying mathematical tools from analysis and algebra to arrive at a rational function describing the transition probabilities, and the use of spectral theory and orthogonal polynomials, to prove the measure's consistency.

In the experimental realm, we demonstrated the accurate results provided by the new analytic expressions for real-life genome sizes and event rates. We have also extended the analytic model to account for realistic phenomena other than the Jump, and also on a tree structure. For the real data analysis, we built an ordered database of orthologous groups across 4445 prokaryotes, to which we applied our measure. To the best of our knowledge, there is no such database of this size in terms of orthologous groups or the number of taxa. Such a database could have multiple uses, apart from phylogenetics.

Applying our new measure to this database produced a tree that was in high accordance with the NCBI taxonomy for these organisms. Importantly, the new measure reconstructed realistic distances, as opposed to the previous measures, even the heuristic measure that was developed based on simulations. Reconstructing accurate distances has prime importance for establishing the Jump model as an underlying model of genome dynamics. Our results demonstrate that developing a distance measure, complementary to existing ones, is important for the sake of validating existing knowledge.

We expect that both the technique developed here for the modelling and the data resources will be instrumental in further analyses of other genome architectures such as operon and pseudogene formation.

While the Jump model is far from a precise description of the likely actual genome dynamics, its simplicity provides for analytical tractability . Future extensions of the model will account for more realistic scenarios including inversions of blocks of genes, duplications, and other events

Disclosure Statement

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

References

Adato, O., N. Ninyo, U. Gophna, and S. Snir. 2015. Detecting horizontal gene transfer between closely related taxa. PLoS computational biology 11:e1004408.

Anderson, W. J. 2012. Continuous-time Markov chains: An applications-oriented approach. Springer Science & Business Media.

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

Baum, B. 1992. Combining trees as a way of combining data sets for phylogenetic inference. Taxon 41:3–10.

Biller, P., L. Guéguen, and E. Tannier. 2015. Moments of genome evolution by double cut-and-join. BMC bioinformatics 16:S7.

Bininda-Emonds, O. 2004. Phylogenetic supertrees: Combining information to reveal the Tree of Life. Springer.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. Pages 387–395 *in* Mathematics in the Archaeological and Historical Sciences (F. Hodson, D. Kendall, and P. Tautu, eds.). Edinburgh University Press, Edinburgh.

Che, D., G. Li, F. Mao, H. Wu, and Y. Xu. 2006. Detecting uber-operons in prokaryotic genomes. Nucleic acids research 34:2418–2427.

Ciccarelli, F. D., T. Doerks, C. Von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. science 311:1283–1287.

Dalevi, D. and N. Eriksen. 2008. Expected gene-order distances and model selection in bacteria. Bioinformatics 24:1332–1338.

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. Science 284:2124–2128.

Doyon, J.-P., C. Scornavacca, K. Y. Gorbunov, G. J. Szöllősi, V. Ranwez, and V. Berry. 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. Pages 93–108 *in* RECOMB International Workshop on Comparative Genomics Springer.

Federhen, S. 2011. The NCBI Taxonomy database. Nucleic Acids Research 40:D136–D143.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic zoology 27:401–410.

Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. Journal of molecular evolution 17:368–376.

Felsenstein, J. 1993. Phylip (phylogeny inference package), version 3.5 c .

Fitz Gibbon, S. T. and C. H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic acids research 27:4218–4222.

Hannenhalli, S. and P. A. Pevzner. 1999. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. Pages 1–27 vol. 46 ACM.

Hasegawa, M., H. Kishino, and N. Saitou. 1991. On the maximum likelihood method in molecular phylogenetics. J Mol Evol 32:443–445.

Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernndez-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. Jensen, C. vonMering, and P. Bork. 2018. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47:D309–D314.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat 37:547–579.

Koonin, E. V., K. S. Makarova, and L. Aravind. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annual Reviews in Microbiology 55:709–742.

Koonin, E. V., K. S. Makarova, and Y. I. Wolf. 2021. Evolution of microbial genomics: Conceptual shifts over a quarter century. Trends Microbiol 29:582–592.

Lin, Y., F. Hu, J. Tang, and B. M. E. Moret. 2013. Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. Pac Symp Biocomput Pages 285–296.

Liu, Y., P. Harrison, V. Kunin, and M. Gerstein. 2004. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol. 5:R64.

Martinez-Gutierrez, C. A. and F. O. Aylward. 2021. Phylogenetic signal, congruence, and uncertainty across bacteria and archaea. Mol Biol Evol 38:5514–5527.

Moran, S. and S. Snir. 2007. Efficient approximation of convex recolorings. Journal of Computer and System Sciences (JCSS) 73:1078–1089 earlier version appeared in APPROX/RANDOM 2005.

Moran, S. and S. Snir. 2008. Convex recolorings of strings and trees: Definitions, hardness results and algorithms. J. Comput. Syst. Sci. 74:850–869.

Morel, B., A. M. Kozlov, A. Stamatakis, and G. J. Szöllősi. 2020. Generax: A tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. Mol Biol Evol 37:2763–2774.

Morel, B., P. Schade, S. Lutteropp, T. A. Williams, G. J. Szllsi, and A. Stamatakis. 2022. SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. Molecular Biology and Evolution 39 msab365.

Nakhleh, L., D. Ruths, and L.-S. Wang. 2005. Riata-hgt: a fast and accurate heuristic for reconstructing horizontal gene transfer. Pages 84–93 *in* International Computing and Combinatorics Conference.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. nature 405:299.

Pang, T. Y. and M. J. Lercher. 2019. Each of 3,323 metabolic innovations in the evolution of E. coli arose through the horizontal transfer of a single DNA segment. Proceedings of the National Academy of Sciences of the United States of America 116:187–192.

Puigbò, P., A. E. Lobkovsky, D. M. Kristensen, Y. I. Wolf, and E. V. Koonin. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC biology 12:66.

Pybus, O. G. and A. Rambaut. 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics 10:540–550.

Ragan, M. 1992. Matrix representation in reconstructing phylogenetic-relationships among the eukaryotes. Biosystems 28:47–55.

Rajendhran, J. and P. Gunasekaran. 2011. Microbial phylogeny and diversity: small subunit ribosomal rna sequence analysis and beyond. Microbiol Res 166:99–110.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical biosciences 53:131–147.

Rogozin, I., K. Makarova, J. Murvai, E. Czabarka, Y. Wolf, R. Tatusov, L. Szekely, and E. Koonin. 2002. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30:2212–2223.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425.

Sankoff, D. 1992. Edit distance for genome comparison based on non-local operations. Pages 121–135 *in* Annual Symposium on Combinatorial Pattern Matching Springer.

Sankoff, D. and J. H. Nadeau. 1996. Conserved synteny as a measure of genomic distance. Discrete applied mathematics 71:247–257.

Schoch, C. L., S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. ONeill, B. Robbertse, S. Sharma, V. Soussov, J. P. Sullivan, L. Sun, S. Turner, and I. Karsch-Mizrachi. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020 baaa062.

Schönknecht, G., A. P. M. Weber, and M. J. Lercher. 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. BioEssays 36:9–20 iSBN: 1521-1878 (Electronic)\r0265-9247 (Linking).

Semple, C. and M. Steel. 2003. Phylogenetics vol. 24. Oxford University Press on Demand.

Serdoz, S., A. Egri-Nagy, J. Sumner, B. R. Holland, P. D. Jarvis, M. M. Tanaka, and A. R. Francis. 2017. Maximum likelihood estimates of pairwise rearrangement distances. Journal of theoretical biology 423:31–40.

Sevillya, G., D. Doerr, Y. Lerner, J. Stoye, M. Steel, and S. Snir. 2019. Horizontal Gene Transfer Phylogenetics: A Random Walk Approach. Molecular Biology and Evolution 37:1470–1479.

Sevillya, G. and S. Snir. 2019. Synteny footprints provide clearer phylogenetic signal than sequence data for prokaryotic classification. Molecular phylogenetics and evolution 136:128–137.

Shifman, A., N. Ninyo, U. Gophna, and S. Snir. 2013. Phylo si: a new genome-wide approach for prokaryotic phylogeny. Nucleic acids research 42:2391–2404.

Siu-Ting, K., D. Pisani, C. J. Creevey, and M. Wilkinson. 2014. Concatabominations: Identifying Unstable Taxa in Morphological Phylogenetics using a Heuristic Extension to Safe Taxonomic Reduction. Systematic Biology 64:137–143.

Sjöstrand, J., A. Tofigh, V. Daubin, L. Arvestad, B. Sennblad, and J. Lagergren. 2014. A bayesian method for analyzing lateral gene transfer. Systematic biology 63:409–420.

Smith, M. R. 2021. Using Information Theory to Detect Rogue Taxa and Improve Consensus Trees. Systematic Biology 71:1088–1094.

Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. Nature genetics 21:108.

Stolzer, M., H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28:i409–i415.

Strimmer, K. and V. Moulton. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol Biol Evol 17:875–881.

Szöllősi, G. J., E. Tannier, N. Lartillot, and V. Daubin. 2013. Lateral gene transfer from the dead. Systematic biology 62:386–397.

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. 2001. The cog database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic acids research 29:22–28.

Tekaia, F. and B. Dujon. 1999. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. Journal of molecular evolution 49:591–600.

Wang, L.-S. and T. Warnow. 2001. Estimating true evolutionary distances between genomes. Pages 637–646 *in* Proceedings of the thirty-third annual ACM symposium on Theory of computing ACM.

Yancopoulos, S., O. Attie, and R. Friedberg. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21:3340–3346.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution 42:587–596.

Zhao, T., A. Zwaenepoel, J.-Y. Xue, S.-M. Kao, Z. Li, M. E. Schranz, and Y. Van de Peer. 2021. Whole-genome microsynteny-based phylogeny of angiosperms. Nat Commun 12:3498.

Appendix: Mathematical proofs

Proof for Theorem 2

We now provide a proof for Theorem 2 provided in the main text. We first repeat the theorem.

Theorem 2:

$$p_{i,j}(t) = \frac{1}{(t+1)^{i+j-1}} \cdot \sum_{\ell=1}^{\min(i,j)} \frac{(i+j-\ell-1)!}{(i-\ell)!(j-\ell)!(\ell-1)!} \left(1-t^2\right)^{\ell-1} t^{i+j-2\ell}. \tag{9}$$

Proof: This result follows from some general results for birth–death processes (refer to (Anderson, 2012) for these). A simple change in notation will be needed, since the results of (Anderson, 2012) involve a birth–death process that is defined on the non-negative integers, whereas our process above is defined on the positive integers. We therefore define $Y_t = X_t - 1$, so that the process Y_t satisfies an instance of the general birth–death process described by:

$$\mathbb{P}(Y_{t+\delta} = i+1 | Y_t = i) = \lambda_i \delta + o(\delta), \quad i \geqslant 0$$
$$\mathbb{P}(Y_{t+\delta} = i-1 | Y_t = i) = \mu_i \delta + o(\delta), \quad i \geqslant 1$$

where in our case:

$$\lambda_i = i+1, \mu_i = i. \tag{10}$$

At the heart of the spectral theory of birth–death processes is the Karlin-McGregor representation of the state transition probabilities ((Anderson, 2012), Ch. 8, Theorem 2.1):

$$\mathbb{P}(Y_t = j | Y_0 = i) = \pi_j \int_0^\infty e^{-tx} Q_i(x) Q_j(x) d\psi(x), \tag{11}$$

where $d\psi(x)$ is a measure on $[0, \infty)$, known as the spectral measure, $\{Q_i(x)\}_{i=0}^\infty$ is a sequence of polynomials, orthogonal with respect to the measure $d\psi$, and $\pi_j = \prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}}$.

In the particular case where the birth–death process is given by (10), we have ((Anderson, 2012), Ch.8, Eq. 4.14):

$$d\psi(x) = e^{-x} dx, \tag{12}$$

$$\pi_j = 1, \quad j \geqslant 0, \tag{13}$$

and the polynomials $Q_i(x)$ are the Laguerre polynomials defined by ((Anderson, 2012), Ch.8, Eq. 4.12)

$$Q_i(x) = {}_1F_1(-m; 1, x) = \sum_{k=0}^{i} \frac{(-i)_k}{k!^2} \cdot x^k$$

where ${}_1F_1$ is the confluent hypergeometric function, and $(-i)_k = (-i)(-i+1)\cdots(-i+k-1)$. We also have the relation ((Anderson, 2012), Ch.8, eq. 4.15)

$$\int_0^\infty e^{-sx} Q_i(x) Q_j(x) dx = \frac{(i+j)!}{i!j!} \cdot \frac{(s-1)^{i+j}}{s^{i+j+1}} \cdot {}_2F_1\left(-i, -j; -i-j; \frac{s(s-2)}{(s-1)^2}\right), \tag{14}$$

where ${}_2F_1$ is the Gaussian hypergeometric function, defined by:

$${}_2F_1\left(-i, -j; -i-j; \frac{s(s-2)}{(s-1)^2}\right) = \sum_{k=0}^{\infty} \frac{(-i)_k (-j)_k}{(-i-j)_k} \cdot \left(\frac{s(s-2)}{(s-1)^2}\right)^k. \tag{15}$$

Using (12),(13),(14),(15) with $s = t + 1$, (11) leads to ((Anderson, 2012), Ch. 8, eq. 4.28):

$$\mathbb{P}(Y_t = j|Y_0 = i) = \pi_j \int_0^\infty e^{-tx} Q_i(x) Q_j(x) e^{-x} dx = \int_0^\infty e^{-x(t+1)} Q_i(x) Q_j(x) dx$$

$$= \frac{(i+j)!}{i!j!} \cdot \frac{t^{i+j}}{(t+1)^{i+j+1}} \cdot {}_2F_1\left(-i, -j; -i-j; \frac{t^2-1}{t^2}\right)$$

$$= \frac{(i+j)!}{i!j!} \cdot \frac{t^{i+j}}{(t+1)^{i+j+1}} \sum_{k=0}^{\infty} \frac{(-i)_k (-j)_k}{(-i-j)_k k!} \cdot \left(\frac{t^2-1}{t^2}\right)^k$$

$$= \frac{(i+j)!}{i!j!} \cdot \frac{t^{i+j}}{(t+1)^{i+j+1}} \sum_{k=0}^{\min(i,j)} \frac{i!j!(i+j-k)!(-1)^k}{(i-k)!(j-k)!(i+j)!k!} \cdot \left(\frac{t^2-1}{t^2}\right)^k$$

$$= \frac{t^{i+j}}{(t+1)^{i+j+1}} \sum_{k=0}^{\min(i,j)} \frac{(i+j-k)!}{(i-k)!(j-k)!k!} \cdot \left(\frac{1-t^2}{t^2}\right)^k.$$

Therefore, going back from the process Y_t to the process X_t, we have

$$p_{i,j}(t) = p_{ij}(t) = \mathbb{P}(X_t = j|X_0 = i) = p_{ij}(t) = \mathbb{P}(Y_t = j+1|Y_0 = i+1)$$

$$= \frac{t^{i+j-2}}{(t+1)^{i+j-1}} \sum_{k=0}^{\min(i,j)-1} \frac{(i+j-k-2)!}{(i-1-k)!(j-1-k)!k!} \cdot \left(\frac{1-t^2}{t^2}\right)^k$$

$$= \frac{1}{(t+1)^{i+j-1}} \cdot \sum_{\ell=1}^{\min(i,j)} \frac{(i+j-\ell-1)!}{(i-\ell)!(j-\ell)!(\ell-1)!} \left(1-t^2\right)^{\ell-1} t^{i+j-2\ell}.$$

□

Proof for Theorem 3

Theorem 3:

$$q_k(t) = \frac{1}{k} \sum_{\ell=0}^{k-1} \sum_{i=0}^{k-\ell-1} \sum_{j=0}^{k-\ell-1} \frac{(i+j+\ell)!}{i!j!\ell!} t^{i+j} (t+1)^{-i-j-2\ell-1} \left(1-t^2\right)^{\ell}. \tag{16}$$

Proof. Summing the expressions for $p_{ij}(t)$ we get

$$q_k(t) = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}(t) = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{1}{(t+1)^{i+j-1}} \cdot \sum_{\ell=1}^{\min(i,j)} \frac{(i+j-\ell-1)!}{(i-\ell)!(j-\ell)!(\ell-1)!} \left(1-t^2\right)^{\ell-1} t^{i+j-2\ell}$$

$$= \frac{1}{k}(t+1) \sum_{\ell=1}^{k} \frac{1}{(\ell-1)!} t^{-2\ell} \left(1-t^2\right)^{\ell-1} \sum_{i=\ell}^{k} \frac{t^i}{(t+1)^i} \frac{1}{(i-\ell)!} \sum_{j=\ell}^{k} \frac{t^j}{(t+1)^j} \frac{(i+j-\ell-1)!}{(j-\ell)!}$$

$$= \frac{1}{k}(t+1) \sum_{\ell=1}^{k} \frac{1}{(\ell-1)!} t^{-2\ell} \left(1-t^2\right)^{\ell-1} \sum_{i=\ell}^{k} \frac{t^i}{(t+1)^i} \frac{1}{(i-\ell)!} \sum_{j=0}^{k-\ell} \frac{t^{j+\ell}}{(t+1)^{j+\ell}} \frac{(i+j-1)!}{j!}$$

$$= \frac{1}{k}\frac{1}{(t+1)^{2k-1}}\sum_{\ell=0}^{k-1}\frac{1}{\ell!}\left(1-t^2\right)^{\ell}\sum_{i=0}^{k-\ell-1}\sum_{j=0}^{k-\ell-1}\frac{(i+j+\ell)!}{i!j!}t^{i+j}(t+1)^{2k-i-j-2\ell-2}$$

$$= \frac{1}{k}\sum_{\ell=0}^{k-1}\sum_{i=0}^{k-\ell-1}\sum_{j=0}^{k-\ell-1}\frac{(i+j+\ell)!}{i!j!\ell!}t^{i+j}(t+1)^{-i-j-2\ell-1}\left(1-t^2\right)^{\ell}.$$

□

Proof for Theorem 4

Theorem 4: The function $q_k(t)$ is monotone decreasing on $[0,\infty)$.

Proof. Using the representation given by Eq. (11) we have:

$$p_{ij}(t) = \int_0^{\infty} e^{-tx}Q_{i-1}(x)Q_{j-1}(x)d\psi(x).$$

This implies that

$$q_k(t) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}p_{ij}(t) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}\int_0^{\infty}e^{-tx}Q_{i-1}(x)Q_{j-1}(x)d\psi(x)$$

$$= \frac{1}{k}\int_0^{\infty}e^{-tx}\sum_{i=1}^{k}\sum_{j=1}^{k}Q_{i-1}(x)Q_{j-1}(x)d\psi(x)$$

$$= \frac{1}{k}\int_0^{\infty}e^{-tx}\left(\sum_{i=1}^{k}Q_{i-1}(x)\right)^2 d\psi(x)$$

Therefore, by differentiating the above with respect to t we obtain:

$$q_k'(t) = -\frac{1}{k}\int_0^{\infty}e^{-tx}x\left(\sum_{i=1}^{k}Q_{i-1}(x)\right)^2 d\psi(x) < 0,$$

since the integrand is positive. This establishes that $q_k(t)$ is monotone decreasing. □